INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was microfilmed as received.

17

This reproduction is the best copy available.

UMI

A STATISTICAL STUDY OF CANADIAN UNIVERSITY
ENROLMENTS, GRADUATES, AND THE FUTURE PROFESSORIATE:
ANALYSIS OF AVAILABLE DATA SOURCES

by

Catherine Fontaine

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Science

School of Mathematics and Statistics

Carleton University

July 2000

© Catherine Fontaine

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-57770-8

Canada

The undersigned recommended to
the Faculty of Graduate Studies and Research
acceptance of the thesis

**A Statistical Study of Canadian University
Enrolments, Graduates, and the Future Professoriate:
Analysis of Different Data Sources**

submitted by Catherine Fontaine
in partial fulfilment of the requirements for
the degree of Master of Science

Thesis Supervisor

Director, School of Mathematics and Statistics

Carleton University
2000

Abstract

The use of methods for the analysis of categorical data, in order to predict the number of full-time university teachers in Canada from 1998 to 2018, are studied. In Chapter 1, the problem, i.e. the impact of the production of Canadian doctoral graduates from 1998 to 2018 and the demand for university education from 1976 to 1998 on the professoriate from 1998 to 2018, is presented. Chapter 2 contains a review of literature on some aspects of university education, in addition to a presentation of the surveys that will be used for the analysis. In Chapter 3, a review of literature on methods for analyzing categorical data is presented. The proportional odds model and Binder's method for taking account of the sample design when estimating the variance of a parameter are described. Chapter 4 contains the results of a proportional odds model used to characterize the university level of certification obtained at graduation. In Chapter 5, factors describing the university enrolment are examined. Then estimated numbers of Canadian full-time university teachers are obtained from 1998 to 2018. Chapter 6 discusses these results and contains suggestions for further research.

Acknowledgements

I would like to thank my supervisor, Dr. Shirley E. Mills for her excellence in guiding me through all aspects of my research. Her patience, consideration, and constant follow-up of my progress were deeply appreciated. I have thoroughly enjoyed working under her supervision.

As well, thanks to the Canadian Association of University Teachers for its financial support. I also thank Rosalind Riseborough for her assistance and for providing me the suggestion of this research topic.

I would also like to acknowledge the Centre for Education Statistics at Statistics Canada for giving me access to the data files and other facilities required for my research. I am also indebted to the members of the Centre for Education Statistics for answering my many questions related to my topic and the surveys I used.

I am grateful to Dr. Norminton from the School of Mathematics and Statistics for putting his work aside and helping me with the setup of my thesis.

I wish to thank Dr. Georgia Roberts and Dr. Roland Thomas for their useful correspondence with me on some aspects of my research topic.

I would like to give special thanks to my parents for their unconditional support and for passing on me their love of knowledge. I also wish to thank my three sisters for the joy they bring to my life. And finally, I want to thank to my beloved fiancé Eric for all his encouragement, patience, and for taking care of me during my masters.

Table of Contents

	Page
ACCEPTANCE SHEET	
ABSTRACT	i
ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	ix
LIST OF FIGURES	xiii
LIST OF APPENDICES	xx
CHAPTER 1: INTRODUCTION	1
1.0 Problem	1
 CHAPTER 2: LITERATURE AND SURVEYS	 4
2.1 Previous studies	4
2.2 Surveys on Education from Statistics Canada	6
2.3 Comparison of variables	13
 CHAPTER 3: ANALYSIS OF CATEGORICAL DATA FROM COMPLEX SURVEYS	 18
3.1 Categorical variables	18
3.2 Review of some methods of analysing categorical data	19
3.3 Proportional odds model	24
3.4 Assumptions of the model	26
3.5 Estimation of the parameters and variance	26
3.5.1 Generalized linear models	26
3.5.2 Estimation for generalized linear models	27

3.5.3	Note about the existence of maximum likelihood estimates	28
3.5.4	Inferences	28
3.6	Binder's method	30
3.7	Goodness-of-fit and residuals	33
CHAPTER 4: PROPORTIONAL ODDS ANALYSIS OF CANADIAN		
UNIVERSITY GRADUATES 1986-1995		38
4.1	Proportional odds model in SUDAAN	38
4.1.1	Estimation of model parameters	39
4.1.2	Estimating variances	41
4.2	Implementation of a proportional odds model using SUDAAN	42
4.2.1	Subpopulation under study	42
4.2.2	Model definition	43
4.2.3	Conditional cumulative probabilities	48
4.3	Validation of the model	51
4.4	Results	61
4.4.1	Design effects	62
4.4.2	Comparison between the two 1997 models	62
4.4.3	Testing of the parameter estimates	63
4.4.4	Testing of the main effects	66
4.4.5	Results of odds ratio analyses	67
4.5	Problems	76

CHAPTER 5: CANADIAN UNIVERSITY ENROLMENT AND PROJECTIONS

FOR FULL-TIME CANADIAN UNIVERSITY TEACHERS	78
5.1 Canadian university enrolment 1976-1998	78
5.1.1 Enrolment profile	78
5.1.2 Factors affecting enrolment	81
5.2 Estimated proportions of Canadian university enrolment 1994-1998	85
5.2.1 Patterns over 1994-1998	85
5.2.2 Modelling proportions	88
5.3 Predictors of the university level of certification obtained at graduation	90
5.3.1 Proportional odds model	90
5.3.2 Validation of the model	91
5.3.3 Design effects	98
5.3.4 Comparison between the two 1997 models	98
5.3.5 Testing of the parameter estimates	100
5.3.6 Testing of the main effects	100
5.3.7 Results of odds ratio analyses	101
5.4 The effect of the region of institution	103
5.4.1 Proportional odds model	103
5.4.2 Validation of the model	104
5.4.3 Design effects	112
5.4.4 Comparison between the two 1997 models	112
5.4.5 Testing of the parameter estimates	113
5.4.6 Testing of the main effects	114
5.4.7 Results of odds ratio analyses	115
5.5 Comparison of two proportional odds models	116

5.6	Estimation of the number of Canadian university teachers 1998-2018	118
5.6.1	The number of university teachers after retirement and attrition	118
5.6.2	Estimated number of potential university teachers in the 1996 Canadian population	123
5.6.3	Estimation of the number of Canadian doctoral graduates 1998-2018	126
5.6.3.1	Estimated proportion of university graduates 1986-1995	126
5.6.3.2	Estimated number of university graduates from 1998 to 2018 in the population	128
5.6.3.3	Estimated number of doctoral graduates from 1998 to 2018 in the population	130
5.6.3.4	Comparison between different estimates of the number of doctoral graduates	134
5.6.4	Estimated number of full-time university teachers 1998-2018	134
CHAPTER 6: SUMMARY RESULTS AND FUTURE RESEARCH		137
6.1	Discussion about the proportional odds model (chapter 4)	137
6.1.1	Justification for the interactions included in the model	137
6.1.2	Design effects	138
6.1.3	Population counts	138
6.1.4	Interpretation of main effects	139
6.1.4.1	First language spoken and still understood	139
6.1.4.2	Status of registration	140
6.1.4.3	Region of residence	141
6.1.5	Interpretation of the interactions	142
6.1.5.1	Gender and age	142
6.1.5.2	Gender and field of study	143

6.1.5.3 Field of study and age	144
6.1.6 Summary	147
6.2 Discussion about the university enrolment	148
6.3 Discussion about the estimates of the full-time university teachers in Canada	150
6.3.1 The effect of the region of institution in the model	150
6.3.2 Design effects	152
6.3.3 Interpretation of the interactions	152
6.3.3.1 Gender and region of residence	152
6.3.3.2 Gender and age	153
6.3.3.3 Region of residence and age	154
6.3.4 Problems with the validation	157
6.3.5 Estimated number of full-time university teachers 1998-2018	157
6.3.5.1 The number of full-time university teachers after retirement and attrition	157
6.3.5.2 Estimated number of potential university teachers in the 1996 Canadian population	158
6.3.5.3 Estimated number of university graduates from 1998 to 2018 in the population	162
6.3.5.4 Estimated number of doctoral graduates from 1998 to 2018 in the population	164
6.3.5.5 Final estimates	165
6.3.5.6 Conclusion	166
6.4 Further research	166

BIBLIOGRAPHY	168
Appendix A: Field of study categories	172
Appendix B: Creation of the status of registration's variable for the NGS 1988	175
Appendix C: Assumptions of Binder's approach	178
Appendix D: SUDAAN's output from the proportional odds model (chapter 4)	180
Appendix E: SUDAAN's output from the proportional odds model (chapter 5, without region of institution)	198
Appendix F: SUDAAN's output from the proportional odds model (chapter 5, with region of institution)	211
Appendix G: Graphs	222

List of Tables

Table 4.1:	Pearson chi-squared statistics	52
Table 4.2:	Deviance statistics	53
Table 4.3:	Wald chi-square statistics	55
Table 4.4:	Frequencies of cell residuals by group	55
Table 4.5:	Proportions of cell residuals (excluding missing residuals) between -2 and 2	55
Table 4.6:	Maximum and minimum cell residuals	55
Table 4.7:	Frequencies of cumulative cell residuals by group	56
Table 4.8:	Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2	57
Table 4.9:	Absolute differences, by group, of the robust and the model-based covariance matrices	61
Table 4.10:	Parameter estimates determined to be equal to zero by the t -tests	65
Table 4.11:	Estimated odds of having response $> j$ ($j = 1, 2$) (status of registration)	70
Table 4.12:	Estimated odds of having response $> j$ ($j = 1, 2$) (region of residence)	71
Table 4.13:	Estimated odds of having response $> j$ ($j = 1, 2$) (language)	72
Table 4.14:	Estimated odds of having response $> j$ ($j = 1, 2$) (interaction between gender and field of study for the 15 to 24 age category)	73
Table 4.15:	Estimated odds of having response $> j$ ($j = 1, 2$) (interaction between gender and field of study for the 25 to 29 age category)	74
Table 4.16:	Number of observations in the subpopulation by cell	76
Table 5.1:	Observed p-values and coefficient of determination ("undergraduate" model)	84

Table 5.2: Observed p-values and coefficient of determination ("graduate" model)	85
Table 5.3: Pearson chi-squared statistics	92
Table 5.4: Deviance statistics	92
Table 5.5: Wald chi-square statistics	93
Table 5.6: Frequencies of cell residuals by group	93
Table 5.7: Proportions of cell residuals (excluding missing residuals) between -2 and 2	93
Table 5.8: Maximum and minimum cell residuals	94
Table 5.9: Frequencies of cumulative cell residuals by group	94
Table 5.10: Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2	95
Table 5.11: Absolute differences, by group, of the robust and the model-based covariance matrices	97
Table 5.12: Number of observations in the subpopulation by cell	98
Table 5.13: Parameter estimates determined to be equal to zero by the t -tests	100
Table 5.14: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women	102
Table 5.15: Pearson chi-squared statistics	104
Table 5.16: Deviance statistics	105
Table 5.17: Wald chi-square statistics	106
Table 5.18: Frequencies of cell residuals by group	106
Table 5.19: Proportions of cell residuals (excluding missing residuals) between -2 and 2	107
Table 5.20: Maximum and minimum cell residuals	107
Table 5.21: Frequencies of cumulative cell residuals by group	108

Table 5.22: Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2	108
Table 5.23: Absolute differences, by group, of the robust and the model-based covariance matrices	111
Table 5.24: Number of observations in the subpopulation by cell	112
Table 5.25: Parameter estimates determined to be equal to zero by the t -tests	114
Table 5.26: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women for a fixed region of institution	115
Table 5.27: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women	117
Table 5.28: Year of retirement of 1998 full-time university teachers	120
Table 5.29: Number of full-time university teachers after retirement	120
Table 5.30: Number of full-time university teachers after retirement from 1998 to 2018	121
Table 5.31: Number of full-time university teachers affected by the attrition from 1998 to 2018	122
Table 5.32: Number of full-time university teachers after retirement and attrition from 1998 to 2018	123
Table 5.33: Estimated number of Canadians with a doctorate who were not full-time university teachers in 1996 (15-24)	124
Table 5.34: Estimated number of Canadians with a doctorate who were not full-time university teachers in 1996 (25-34)	124
Table 5.35: Proportion of people with a doctoral degree who are full-time university teachers in 1996	125
Table 5.36: Number of Canadians with a doctorate in 1996 who could become full-time university teachers by the end of 2000	125

Table 5.37: Proportion of university graduates by region and gender in 1986, 1990, and 1995	127
Table 5.38: Population projections from 1998 to 2018 (34 years old and less)	129
Table 5.39: Estimated number of university graduates from 1998 to 2018	130
Table 5.40: Estimated number of doctoral graduates from 1998 to 2018 aged 34 years old and less	132
Table 5.41: Estimated number of doctoral graduates aged 34 years old and less who would become full-time university teachers from 1998 to 2018	133
Table 5.42: Estimated number of full-time university teachers from 1998 to 2018	136

List of Figures

Figure 4.1: Deviance residual by its index for $X = x_i$: $Y = 1$ (1988)	222
Figure 4.2: Deviance residual by its index for $X = x_i$: $Y = 2$ (1988)	222
Figure 4.3: Deviance residual by its index for $X = x_i$: $Y = 3$ (1988)	222
Figure 4.4: Deviance residual by its index for $X = x_i$: $Y = 1$ (1992)	222
Figure 4.5: Deviance residual by its index for $X = x_i$: $Y = 2$ (1992)	223
Figure 4.6: Deviance residual by its index for $X = x_i$: $Y = 3$ (1992)	223
Figure 4.7: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997)	223
Figure 4.8: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997)	223
Figure 4.9: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997)	224
Figure 4.10: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997, administrative counts)	224
Figure 4.11: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997, administrative counts)	224
Figure 4.12: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997, administrative counts)	224
Figure 5.1: Canadian undergraduate enrolment by year for several age categories	225
Figure 5.2: Canadian graduate enrolment by year for several age categories	225
Figure 5.3: Enrolment totals by year of enrolment for different categories of age (undergraduates) in the Atlantic provinces	225
Figure 5.4: Enrolment totals by year of enrolment for different categories of age (undergraduates) in Quebec	225
Figure 5.5: Enrolment totals by year of enrolment for different categories of age (undergraduates) in Ontario	226

Figure 5.6: Enrolment totals by year of enrolment for different categories of age (undergraduates) in the Prairies	226
Figure 5.7: Enrolment totals by year of enrolment for different categories of age (undergraduates) in British Columbia	226
Figure 5.8: Enrolment totals by year of enrolment for different categories of age (graduates) in the Atlantic provinces	226
Figure 5.9: Enrolment totals by year of enrolment for different categories of age (graduates) in Quebec	227
Figure 5.10: Enrolment totals by year of enrolment for different categories of age (graduates) in Ontario	227
Figure 5.11: Enrolment totals by year of enrolment for different categories of age (graduates) in the Prairies	227
Figure 5.12: Enrolment totals by year of enrolment for different categories of age (graduates) in British Columbia	227
Figure 5.13: Enrolment totals by year of enrolment for regions of enrolment (undergraduates)	228
Figure 5.14: Enrolment totals by year of enrolment for regions of enrolment (graduates)	228
Figure 5.15: Enrolment totals by year of enrolment for different categories of age (men) in the Atlantic provinces	228
Figure 5.16: Enrolment totals by year of enrolment for different categories of age (men) in Quebec	228
Figure 5.17: Enrolment totals by year of enrolment for different categories of age (men) in Ontario	229
Figure 5.18: Enrolment totals by year of enrolment for different categories of age (men) in the Prairies	229

Figure 5.19: Enrolment totals by year of enrolment for different categories of age (men) in British Columbia	229
Figure 5.20: Enrolment totals by year of enrolment for different categories of age (women) in the Atlantic provinces	229
Figure 5.21: Enrolment totals by year of enrolment for different categories of age (women) in Quebec	230
Figure 5.22: Enrolment totals by year of enrolment for different categories of age (women) in Ontario	230
Figure 5.23: Enrolment totals by year of enrolment for different categories of age (women) in the Prairies	230
Figure 5.24: Enrolment totals by year of enrolment for different categories of age (women) in British Columbia	230
Figure 5.25: Male enrolment by year for regions of residence	231
Figure 5.26: Female enrolment by year for regions of residence	231
Figure 5.27: Residuals by predicted values (ANOVA model for undergraduate enrolment)	231
Figure 5.28: Residuals by predicted values (ANOVA model for graduate enrolment)	231
Figure 5.29: Residuals by predicted values (ANOVA model for undergraduate enrolment, with transformed response)	232
Figure 5.30: Residuals by predicted values (ANOVA model for graduate enrolment, with transformed response)	232
Figure 5.31: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 15 to 24 years of age)	232

Figure 5.32: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 25 to 29 years of age)	232
Figure 5.33: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 30 to 34 years of age)	233
Figure 5.34: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 35 years of age and over)	233
Figure 5.35: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 15 to 24 years of age)	233
Figure 5.36: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 25 to 29 years of age)	233
Figure 5.37: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 30 to 34 years of age)	234
Figure 5.38: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 35 years of age and over)	234
Figure 5.39: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 15 to 24 years of age)	234

Figure 5.40: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 25 to 29 years of age)	234
Figure 5.41: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 30 to 34 years of age)	235
Figure 5.42: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 35 years of age and over)	235
Figure 5.43: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 15 to 24 years of age)	235
Figure 5.44: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 25 to 29 years of age)	235
Figure 5.45: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 30 to 34 years of age)	236
Figure 5.46: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 35 years of age and over)	236
Figure 5.47: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 15 to 24 years of age)	236

Figure 5.48: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 25 to 29 years of age)	236
Figure 5.49: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 30 to 34 years of age)	237
Figure 5.50: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 35 years of age and over)	237
Figure 5.51: Deviance residual by its index for $X = x_i$: $Y = 1$ (1988)	237
Figure 5.52: Deviance residual by its index for $X = x_i$: $Y = 2$ (1988)	237
Figure 5.53: Deviance residual by its index for $X = x_i$: $Y = 3$ (1988)	238
Figure 5.54: Deviance residual by its index for $X = x_i$: $Y = 1$ (1992)	238
Figure 5.55: Deviance residual by its index for $X = x_i$: $Y = 2$ (1992)	238
Figure 5.56: Deviance residual by its index for $X = x_i$: $Y = 3$ (1992)	238
Figure 5.57: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997)	239
Figure 5.58: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997)	239
Figure 5.59: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997)	239
Figure 5.60: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997, administrative counts)	239
Figure 5.61: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997, administrative counts)	240
Figure 5.62: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997, administrative counts)	240
Figure 5.63: Deviance residual by its index for $X = x_i$: $Y = 1$ (1988)	240
Figure 5.64: Deviance residual by its index for $X = x_i$: $Y = 2$ (1988)	240

Figure 5.65: Deviance residual by its index for $X = x_i$: $Y = 3$ (1988)	241
Figure 5.66: Deviance residual by its index for $X = x_i$: $Y = 1$ (1992)	241
Figure 5.67: Deviance residual by its index for $X = x_i$: $Y = 2$ (1992)	241
Figure 5.68: Deviance residual by its index for $X = x_i$: $Y = 3$ (1992)	241
Figure 5.69: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997)	242
Figure 5.70: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997)	242
Figure 5.71: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997)	242
Figure 5.72: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997, administrative counts)	242
Figure 5.73: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997, administrative counts)	243
Figure 5.74: Deviance residual by its index for $X = x_i$: $Y = 3$ (1997, administrative counts)	243
Figure 5.75: Canadian male enrolment by year for several age categories	243
Figure 5.76: Canadian female enrolment by year for several age categories	243

List of Appendices

Appendix A: Field of study categories	172
Appendix B: Creation of the status of registration's variable for the NGS 1988	175
Appendix C: Assumptions of Binder's approach	178
Appendix D: SUDAAN's output from the proportional odds model (chapter 4)	180
Appendix E: SUDAAN's output from the proportional odds model (chapter 5, without region of institution)	198
Appendix F: SUDAAN's output from the proportional odds model (chapter 5, with region of institution)	211
Appendix G: Graphs	222

CHAPTER 1: INTRODUCTION

1.0 Problem

Several organizations are interested in university faculty renewal in Canada. Due to retirement and normal attrition (such as deaths or leave), the number of university teachers decreases each year. Fortunately, new university teachers are hired to replace these departures. But one question arises from this: is the number of university teachers produced high enough to replace all of these departures? Will there be a shortage of university teachers in the future? The issue of faculty renewal must take into account the aging of professors, early retirement, normal attrition, and the number of people who can become potential university teachers in the future (generally, people with a doctorate). Furthermore, high demand for a university education in the future is a factor (in addition to a low faculty renewal) that can create shortages of university professors. Thus the demand for a university education and the renewal of faculty are closely related.

When it comes time to study faculty renewal, different data sources can be used. Data about doctoral graduates, current university teachers, enrolment in university programs, and population projections are needed. But there is no unique data source with information on all of these topics. So selected data sources must be combined. Data about the graduates were found from the National Graduates Survey (and the follow-up of this survey) for 1986, 1990, and 1995 graduates. When the enrolment was studied, administrative data (from the University Student Information System) from 1976 to 1998 were used. Census estimates from 1994 to 1998 were used (in addition to USIS data from 1994 to 1998) to produce estimated proportions of Canadians enrolled in university programs for certain characteristics. Data about university teachers were obtained from the University and College Academic Staff System for the years 1996 and 1998. Data about university teachers

and about the highest level of education of Canadians were available through Census CD-ROM products for the year 1996. Finally, projections of the number of Canadians from 1998 to 2018 were used in the analysis. These data are all produced by Statistics Canada. The combination of these different data sources must, however, be done with care, and the population of interest of different data sources must be comparable. And when survey data are used, the sampling plan must be taken into account when the statistical analyses are performed.

This research has produced estimates of the number of Canadian full-time university teachers from 1998 to 2018. This was done in three major steps. The first step was the estimation of the number of actual university teachers lost to retirement or other reasons in the coming years; from this information we obtained an estimate the number of university teachers that would be left. The second step was the estimation of the number of Canadians already having a doctorate who could become university teachers. Finally, the number of students that will obtain a doctorate in the future (and who could thus become full-time university teachers) has also been estimated. The second and third steps were performed assuming that only people with a doctoral degree at certain ages may become university teachers and assuming that they enter the professoriate within four or five years after their doctoral graduation. From the combination of these three steps arose projections for faculty renewal.

Statistical methods have been used to estimate the number of graduate students that will obtain a doctorate in the future, given some characteristics (such as the age, the gender, and the region of residence). A proportional odds model has been used for this purpose, since the response was an ordinal variable. This gives cumulative conditional probabilities of having obtained a certain level of university education at graduation. The method of taking into account the sampling plan in those statistical analyses occupied the main part of this

research. The statistical software SUDAAN was used to fit this type of model. Following this analysis, comments on results encountered and future areas of research are discussed.

CHAPTER 2: LITERATURE AND SURVEYS

In this chapter, we will review the literature on some aspects of university education. The surveys and the variables in these surveys used for this research work are presented.

2.1 Previous studies

The first task of this project was to find relevant sources of data. These sources of data should include variables that would allow the examination of the demand for university education and the prediction of the number of university graduates having certain characteristics. The following are some studies that helped in choosing sources of data with pertinent variables for the current research.

Butlin (1999) studied the results of the 1995 School Leavers Follow-up Survey (conducted at Statistics Canada). He used a multinomial logistic regression model to assess how social-demographic and school-related factors affect the odds of: participating in university education versus no postsecondary education; participating in community college education versus no postsecondary education; and participating in trade-vocational education versus no postsecondary education. He examined in depth the results of the multinomial logistic regression allowing the enumeration of some determinants of postsecondary participation. Some of the results he obtained are: high school graduates with university-educated parents had higher odds of attending university (controlling for other factors); high average grades in high school, the absence of difficulties with high school mathematics and English, and not failing a grade in elementary school are strong predictors of university participation versus no postsecondary participation (while holding constant parent's level of education, and other predictors); women are more likely than men to participate in university and community college (controlling for other factors); students whose most often used language is French had slightly lower odds of university participation (controlling for other factors); high school

graduates from rural areas are less likely to participate in university education (controlling for other factors); high school graduates born outside Canada are more likely to attend university (controlling for other factors). His conclusions were useful in researching sources of data with possible determinants of participating in university education.

The issue of people returning to school full-time must also be taken into account in the prediction of the future demand for university education. Rose (1994) presented a portrait of the adult Canadians who chose to return to school. He discussed such characteristics as gender, age, income, and distribution by academic level and field of study. His analysis is based on the results of the 1992 Adult Education and Training Survey (AETS).

As stated in the introduction, the ability to hire future university teachers will be impacted by the number of doctoral graduates in the coming years. Gonzalez (1996) examined the graduation rates and the time needed to obtain the doctoral degree among four cohorts. The graduation rates and the times to completion were examined by gender, field of study, immigration status, discipline and status at the beginning of the doctoral program (part-time or not). The data for this study were derived from the University Student Information System (from Statistics Canada). Her article was useful for identifying sources of data providing information about some factors affecting the completion of doctoral programs. Another aspect under study was the status of registration of the graduates during their studies (full-time only, part-time only, etc.). Bélanger & Omiecinski (1987) studied the part-time university enrolment from 1965 to 1985 in Canada. The Enrolment and Degrees Survey (from Statistics Canada) was used. They studied the part-time enrolment versus the full-time enrolment at some periods between 1965 and 1985. The part-time enrolment by gender and field of study, in addition to the academic background of part-time students and the reasons for part-time enrolment, were also investigated. Their results were helpful for the determination of factors affecting part-time enrolment versus full-time enrolment.

2.2 Surveys on Education from Statistics Canada

The next task was to select surveys with pertinent information about the topic of interest, according to the studies made previously. Statistics Canada produces a certain number of surveys on education. Many of them are conducted by the Education, Culture and Tourism Division. The two surveys selected come from this Division.

1) National Graduates Survey

The first survey selected is called the National Graduates Survey. It targeted graduates from Canadian university, college and trade/vocational programs (excluding apprenticeships and continuing education programs). The main survey (also called the NGS) is conducted two years after graduation and a follow-up survey of these same graduates (FOG) is conducted five years after graduation. The primary objective of the NGS was to obtain information on the relationship between education/training and labour market experiences, including: characteristics of programs of study, satisfaction with education, additional training after graduation, activities before completing studies, activities after completing post-secondary studies, jobs held since graduation, finances and loans, and socio-economic background. The FOG survey provides an opportunity to analyze employment, occupational and geographic shifts longitudinally over the first five years after graduation. For the purpose of this project, three classes have been chosen: the class of the 1986 graduates; the class of the 1990 graduates; and the class of the 1995 graduates. Here are the NGS and FOG surveys selected for the research:

Class of the 1986 graduates: 1988 NGS (May 1988) and 1991 FOG (March 1991-covers the period between June 1988 and March 1991)

Class of the 1990 graduates: 1992 NGS (June 1992) and 1995 FOG (June 1995)

Class of the 1995 graduates: 1997 NGS (May-June 1997)

The 2000 FOG had not been conducted yet at the time of this research, explaining why it could not be used. The determination of the graduates implied some specific inclusions and exclusions. These specifications are described here for the class of the 1986 graduates and the class of the 1990 graduates (the specifications for the class of the 1995 graduates are similar to the specifications for the class of the 1990 graduates).

Specifications for the class of 1986:

A “graduate” was defined as someone who had received, or who was eligible to receive, a degree, diploma or certificate from a recognized university, college, trade/vocational school or similar institution in the calendar year 1986. This included:

1. graduates of university programs leading to bachelor’s, master’s or doctoral degrees, or specialized certificates or diplomas. Excluded were those from continuing education courses and military colleges.
2. graduates of postsecondary programs (i.e., programs of one year duration or longer which normally require high school completion or its equivalent for admission) in community colleges, Colleges of Applied Arts and Technology (CAAT), Collèges d’enseignement général et professionnel (CEGEP), technical institutes, hospital schools of nursing or radiology or similar institutions. Graduates from CEGEP general programs (i.e., pre-university CEGEP programs in Quebec) were also surveyed.
3. graduates of skilled trades programs (i.e. pre-employment) excluding programs of less than three months duration, basic training or skill development programs, and apprenticeship programs.

Graduates from private postsecondary institutions were excluded (i.e. institutions operating

on a for profit basis and not usually receiving public funding (e.g., commercial business schools)).

Specifications for the class of 1990:

The base population encompassed graduates from Canadian postsecondary education institutions who completed the requirements for degrees, diplomas, or certificates during the calendar year 1990. More specifically, these include:

- 1.** graduates of university programs leading to bachelor's, master's or doctoral degrees, or specialized certificates or diplomas.
- 2.** graduates of postsecondary programs (i.e., programs of one year duration or longer which normally require high school completion or its equivalent for admission) in the CAAT's, CEGEP's, community colleges, technical schools, and similar institutions.
- 3.** graduates of skilled trades (i.e. pre-employment) programs which normally were 3 months or more in duration. A trade/vocational school is a public educational institution that offers courses to prepare people for employment in a specific occupation such as heavy equipment operator, automotive mechanic or upholsterer. Many community colleges and technical institutes offer certificates or diplomas at the trade level.

The population excludes:

- a)** graduates from private postsecondary education institutions (e.g. commercial secretarial or computer programming schools);
- b)** those who completed "continuing education" courses (unless these led to degrees or regular diplomas or certificates);
- c)** those who took part-time trade courses (e.g. adult education evening courses) while employed full-time;

- d)** persons who completed vocational programs
 - i) lasting less than three months, or
 - ii) other than in the skilled trades (e.g. basic training and skill development);
 and
- e)** persons in apprenticeship programs.

Specifications for the class of 1995:

Graduates from the class of 1995 are students who completed the requirements for a degree, a diploma, or certificate during the 1995 calendar year in trade/vocational, college, or university programs. For the inclusions and exclusions, see “Specifications for the class of 1990”.

Once the population was well defined, the frame had to be determined. To obtain the 1986 frame, a list as complete as possible of all 1986 graduates was compiled from lists provided by individual institutions and provincial Ministries of Education including names, permanent addresses and telephone numbers for the 1988 National Graduates Survey. The list or “frame” of 1990 graduates was provided by Canada’s universities and colleges. For the trade/vocational frame, a supplementary list was obtained from Human Resources Development Canada (HRDC) of those who had taken trades training arranged for by the Department. The list or “frame” of 1995 graduates was created, starting with a list of all universities, colleges and trade-vocational schools in Canada provided by Statistics Canada’s Centre for Education Statistics. For the trade-vocational frame, a list was also obtained from HRDC of those who had taken trades training arranged for by the Department.

For each NGS survey, a sample was selected from the frame. The National Graduates Survey was based on a stratified systematic random sample design. The population was first stratified by province of institution. Within each province, the graduates were stratified by

level of study and major field of study groups. This three-way stratification has been applied to the 1988 and 1992 NGS. The stratification for the 1997 NGS will be described later. The three-way stratification was defined from the following three variables.

1) Province of institution

Newfoundland; Prince-Edward Island; Nova Scotia; New Brunswick; Quebec; Ontario; Manitoba; Saskatchewan; Alberta; British Columbia; Yukon; Northwest Territories.

2) Level of study (certification) at graduation

a) Bachelor's degrees including first professional degrees

(M.D., D.D.S., LL.B., D.V.M., O.D., etc.) and undergraduate diplomas

b) Master's degrees including graduate diplomas

c) Earned doctorates

d) Postsecondary level programs from community colleges, CEGEPs, CAATs, technical institutes, hospital schools of nursing and radiology and similar institutions

e) Trade/vocational programs

3) For the class of 1986 graduates, the nine university major field of study groups, the nine college field of study groups, and the ten trade/vocational field of study groups are listed in Appendix A.

For the class of 1990 graduates, the nine fields of study for university programs and career/technical programs and the ten fields of study for trade/vocational were based on the 5-digit USIS (University Student Information System) and CCSIS (Community College Student Information System) major field of study codes. The field of study categories were the same as the field of study categories of the class of 1986 graduates.

For the class of 1995 graduates, the nine fields of study for university programs, the nine

fields of study for career/technical programs, and the eight fields of study for trade/vocational programs were based on the 5-digit USIS and CCSIS major field of study codes. The university and college field of study categories were the same as the field of study categories of the class of 1986 graduates. The eight field of study categories for trade/vocational programs are also listed in Appendix A.

For the 1997 NGS, the sample design was slightly different. A “base” sample was first selected according to a stratified systematic random sample design (with the three-way stratification described above, i.e. province of institution by level of certification and by field of study). Then an additional sample of graduates was selected. These additional sample units had been purchased by several institutions and provincial agencies in order to supplement their base sample so that better levels of reliability could be attained for their particular province or institution. Therefore, the stratification was in fact a four-way stratification: province of institution by level of certification by field of study by an indicator saying if the unit came from the base sample or from a particular institution who bought an additional sample. This indicator had two categories (unit selected from the base sample or not) but the details can not be revealed for confidentiality reasons.

In each of the three NGS surveys, a weight was assigned to each respondent in the sample. The weighting is done so that each person in the sample represents, besides himself or herself, several other persons not in the sample. The weighting and the sample design must be taken into account in the estimations and statistical analyses. This will be discussed later in the thesis.

This survey has been chosen mainly because of the large number of variables available about the graduates, and because it was possible to make a comparison of these variables at three different years. We can characterize the graduates by, for example, their gender, level of

certification at graduation, major field of study and examine the relationships existing between some variables of interest (such as the language first spoken versus the level of certification of the graduate at graduation) and see the changes across time. This will allow the determination of predictors of the demand for university education.

II) University Student Information System (USIS) - Enrolment and Degrees Awarded

The second survey selected is in fact two surveys grouped under a system called the University Student Information System. This is a national database containing pertinent information on student participation in Canadian degree-granting institutions. Data is available on enrolment and on graduation. The enrolment and degrees annual surveys (conducted by Statistics Canada) are the data source. Topics such as enrolment and degrees granted; university participation rates; foreign student participation in Canadian education; discipline profile trends; and sex distribution of Canadian students and graduates are available through the USIS database.

University enrolment statistics are available on students enrolled in Canadian degree-granting institutions leading towards a degree, diploma or certificate. These statistics include students enrolled in courses and students who have completed their course requirements and who are engaged in thesis writing or research. Also included are students taking courses eligible for credit but who are not seeking a degree, diploma or certificate, e.g. auditors. The enrolment survey collects information on student counts as of December 1st in all provinces except Ontario, where the reference date is November 1st. This means that each student who attends university in the fall session is counted only once annually, even though the student may be enrolled in more than one program during the academic year. The student count is used as a proxy for the total number of students enrolled during a complete academic year. The degrees survey collects information on all students who have received a degree, diploma or certificate during the calendar year ending in December. It is a

count of the number of degrees, diplomas and certificates awarded, not the number of individual students who receive them. Through record linkage with the enrolment file, much information on the enrolment file is also available on the degrees file. Record linkage is possible for most institutions, accounting for approximately 70% of the student records for the 1993 survey year.

The USIS database was chosen mainly because of the information about enrolment for several years. Several socio-demographic variables are available on this database, which will allow the characterization of the students enrolled in university programs.

2.3 Comparison of variables

The three NGS data files contain a large number of variables. Some of them are identical (i.e. same question asked on the questionnaire and same response categories) for each of the three surveys. But it occurred that sometimes either the question itself on the questionnaire or the response categories were different for two or more surveys. So a comparison of the variables of interest for this research was necessary. A recoding of the categories of some variables had to be done since the variables that were used in the analysis had to be comparable. Here are the variables that were compared. (Note that the 1986 file refers to the data file with the graduates selected for the 1988 NGS, with all the variables on the 1988 NGS questionnaire and all the variables on the 1991 FOG questionnaire; the 1990 file refers to the data file with the graduates selected for the 1992 NGS, with all the variables on the 1992 NGS questionnaire and all the variables on the 1995 FOG questionnaire; and finally the 1995 file refers to the data file with the graduates selected for the 1997 NGS, with all the variables on the 1997 NGS questionnaire.)

A) Region of residence

This variable gives the region of residence of the graduate at the time on

the interview (two years after their graduation). The categories of “Province of residence” for the 1986 and the 1990 files were converted according to the categories of the same variable on the 1995 file. Then some provinces and territories have been collapsed into larger regions. A list of the regions of residence that have been created are: the Atlantic provinces (Newfoundland, Prince-Edward Island, Nova Scotia, New Brunswick); Quebec; Ontario; the Prairies (Manitoba, Saskatchewan, Alberta); British Columbia; Yukon and Northwest Territories.

Another category encompassed either the categories “Not stated”, “Unknown” or the name of a province or a country outside Canada. Respondents who belong to this category were deleted from the analysis..

B) Gender

This variable describes the gender of the graduate. On the 1986 file, there were three categories: “Male”, “Female” and “Unknown”. On the 1990 file, only the “Male” and the “Female” categories were available. Then on the 1995 file, there were also three categories: “Male”, “Female” and “Not stated”. Records with gender either “Not stated” or “Unknown” were deleted from the analysis..

C) Level of certification

This is the level of certification attained by the graduate for the program he/she graduated either in 1986, 1990 or 1995. This is the stratification variable (with five levels) described in section 2.2..

D) Major field of study

This is the major field of study of the program that the graduate graduated from in either 1986, 1990 or 1995. Categories have been created for each survey from the 5-digit USIS field of study codes:

00000-09999 = General arts and science

10000-19999 = Education

20000-39999 = Fine arts and Humanities

412xx, 427xx, 433xx = Commerce, law, economics

40000-49999, except 412xx, 427xx, 433xx = Other Social sciences

50000-59999 = Agriculture and biological sciences

60000-69999 = Engineering and applied sciences

70000-79999 = Health professions

80000-89999 = Mathematics and physical sciences

There was also a category created for the fields of study “Not stated”, “Unknown”, or “Uncodable”. Records from this category were deleted from the analysis.

E) Status of registration

On the 1995 file, there was a variable with five categories created from a question of the questionnaire asking if the graduate took his/her program:

1) Full-time only

2) Part-time only

3) Combination of full-time and part-time

4) Refused

5) Not stated

On the 1992 NGS questionnaire, the same question was asked but only 4 categories were possible: the category “Refused” was not a possible answer. On the 1988 NGS questionnaire, such a question was not asked. So the variable had to be derived (see Appendix B for further details). Records with the status of registration that was either “Refused” or “Not stated” were deleted from the analysis.

F) First spoken language and still understood

This variable was derived from three questions from the questionnaire.

It was used to identify graduates whose first language spoken and still understood was other than English or French, and those whose first language spoken and still understood was either English or French, but no other. This variable had two categories:

- 1) First language spoken and still understood was other than English or French
- 2) First language spoken and still understood was either English or French, but no other

The undefinable cases were deleted from the analysis.

G) Categories of age (at the time of the NGS)

This variable groups the graduates’ ages into four groups, following the Census grouping.

- 1) 15 to 24 years old
- 2) 25 to 29 years old
- 3) 30 to 34 years old
- 4) 35 years old and over

CHAPTER 3: ANALYSIS OF CATEGORICAL DATA FROM COMPLEX SURVEYS

We will begin by a review of methods for analyzing categorical data. The textbook from Agresti (1990) was used to review these statistical methods (see the bibliography for a detailed reference of this textbook). Then the proportional odds model, the generalized linear models and Binder's method for taking account of the sample design when estimating the variance of a parameter are described. The chapter ends with a presentation of methods that may be used for the validation of the proportional odds model.

3.1 Categorical variables

Once research objectives have been determined, the data must be examined in order to choose the most appropriate statistical methods to attain these objectives. Chapter 2 presented the variables of interest selected for the analysis. They are all categorical. A categorical variable is one for which the measurement scale consists of a set of categories. For example, the variable "Gender" has two categories: male and female. There are many types of categorical variables. The variables that will be used for the analysis are either ordinal, nominal or interval. An ordinal variable is one for which levels have a natural ordering. For example, the level of certification is an ordinal variable since we can order the levels from the lower (bachelor's degree) to the higher (doctorate). The gender is a nominal variable since there is no natural ordering between "Male" and "Female". An interval variable does have numerical distances between any two levels of the scale. For ordinal variables, even if they clearly order the categories, the absolute distances between categories are unknown. Sometimes, continuous variables are treated as categorical; this is called categorization or discretization of the continuous variable. In practice, categorization is often necessary because either the substantive meaning or the actual measurement of a continuous

variable is categorical. For example, age is usually reported in single-year or five-year intervals. A list of the categorical variables used for the analysis is:

Nominal variables:

Region of residence; gender; field of study; status of registration; first language spoken and still understood

Ordinal variable:

Level of certification

Interval variable:

Categories of age at the time of the NGS

Distinction between dependent and independent variables is important at this point. A dependent (also called response or outcome) variable represents a population characteristic of interest being explained in a study. Independent (also called explanatory) variables are variables that are used to explain the variation in the dependent variable. Including categorical variables as independent variables in regression-type models does not present any particular difficulties. However, methods for analyzing categorical dependent variables are much less widely known. The type of dependent variable must be checked carefully because some methods of analysis are appropriate for one type of categorical dependent variable, but not for another.

3.2 Review of some methods of analysing categorical data

When we have two categorical variables with respectively I and J modalities (I and J greater than 2), we can form a two-way contingency table by crossing these two variables. Since most of the variables of interest have more than two categories, the 2×2 contingency tables will not be discussed here. We can describe a two-way contingency table $I \times J$ by using some measures of association. Some of them will be more appropriate in the case of ordinal variables and some will be more appropriate in the case of nominal variables. In the case of

ordinal variables, the notion of linearity is not meaningful because ordinal variables do not have a defined metric. As the Pearson correlation coefficient describes the degree to which Y has a linear relationship with X , measures for ordinal variables that are analogous to this coefficient describe the degree to which the relationship is monotone. The natural ordering of categories allows consideration of monotonicity- for instance, whether Y tends to increase as X does. Some measures of ordinal association for $I \times J$ tables include the Goodman and Kruskal's γ (based on the notion of concordance and discordance); the Kendal's tau-b (similar to Goodman and Kruskal's γ except that tau-b uses a correction for ties); the Spearman rank correlation coefficient (computed by using rank scores); and the Pearson chi-square (used to test the null hypothesis of no association between the row variable and the column variable).

When we want to analyze the relationships between more than two variables, incorrect conclusions can arise if we study the relationships between only two variables at a time. Loglinear models can be used to describe association patterns among categorical variables (two or more). This approach models the cell counts in a contingency table in terms of association among the variables. For example, in a three-way contingency table ($X \times Y \times Z$), the loglinear model

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

represents the case where the three variables are mutually independent; the model

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

represents the case where the Y is jointly independent of X and Z ; the model

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

represents the case where X and Y are conditionally independent, given Z ; and the model

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$$

represents the case where each pair of variables may be conditionally dependent. This last

model is also called a saturated model. In the models above, we define m_{ijk} as the expected cell frequencies; λ_i^X as the parameter associated with the main effect of X ; λ_j^Y as the parameter associated with the main effect of Y ; λ_k^Z as the parameter associated with the main effect of Z ; λ_{ik}^{XZ} is the parameter pertaining to the X - Z partial association; λ_{jk}^{YZ} is the parameter pertaining to the Y - Z partial association; λ_{ij}^{XY} is the parameter pertaining to the X - Y partial association; and λ_{ijk}^{XYZ} is the parameter allowing the three-factor interaction between X , Y and Z . Note that all of these loglinear models are subject to constraints. The limitation of loglinear models is that they treat all classifications as nominal, in the sense that parameter estimates and chi-squared statistics are invariant to orderings of categories. Important information is ignored when at least one variable is ordinal. However, some loglinear models are adapted to exploit ordinality. One of the advantages of such models over those mentioned earlier is that the tests based on ordinal models have improved power for detecting certain types of association and interaction (see Agresti (1990) for more details).

All of the methods described in the paragraphs above do not distinguish between response and explanatory variables. The following paragraphs discuss models for a nominal dependent variable and models for a ordinal dependent variable. First we will begin by a discussion about models for nominal dependent variable. Generalized logit models for nominal responses can contain continuous or discrete explanatory variables. Let Y be a categorical dependent variable with J categories and $\pi_j(\mathbf{x}_i)$ denote the probability of response j , $j = 1, \dots, J$, at the i^{th} setting of values of k explanatory variables $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$. In terms of the response probabilities, the generalized logit model is

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\beta_j' \mathbf{x}_i)}{\sum_{l=1}^J \exp(\beta_l' \mathbf{x}_i)}. \text{ It is assumed that } \beta_J = 0. \text{ Then, } \log\left(\frac{\pi_j(\mathbf{x}_i)}{\pi_J(\mathbf{x}_i)}\right) = \beta_j' \mathbf{x}_i, j = 1, \dots, J-1.$$

This model is called the “standard” multinomial logit model. A different form of the multinomial logit model is known as the discrete-choice model or conditional logit model. In the standard multinomial logit model, explanatory variables are invariant with outcome categories, but their parameters vary with outcome. In the conditional logit model,

explanatory variables vary by outcome as well as by the individual, whereas their parameters are assumed constant over all the outcome categories. Let $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijk})'$ denotes the vector of the k explanatory variables associated with response choice j pertaining to the i^{th} individual. Conditional on the set of response choices C_i for subject i , the model is $\pi_j(\mathbf{x}_{ij}) = \frac{\exp(\beta' \mathbf{x}_{ij})}{\sum_{h \in C_i} \exp(\beta' \mathbf{x}_{ih})}$. For each pair of alternatives a and b , this model has the logit form $\log \left[\frac{\pi_a(\mathbf{x}_{ia})}{\pi_b(\mathbf{x}_{ib})} \right] = \beta'(\mathbf{x}_{ia} - \mathbf{x}_{ib})$. Conditional on a subject's choice being a or b , the influence of a particular explanatory variable depends on the distance between the subject's values of that variable for those alternatives.

Now let us turn to models for an ordinal dependent variable. When response categories have a natural ordering, some logit models utilize that ordering. Three types of logits for ordered response categories are:

(1) The adjacent-categories logits

Let $\{\pi_1(\mathbf{x}), \dots, \pi_J(\mathbf{x})\}$ denote response probabilities at value $\mathbf{x}$ for a set of explanatory variables. The adjacent-categories logits are:

$$L_j = \log \left(\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} \right), \quad j = 1, \dots, J-1.$$

These logits are a basic set equivalent to the baseline-category logits

$L_j^* = \log \left(\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} \right)$, $j = 1, \dots, J-1$. Either set determines logits for all $\binom{J}{2}$ pairs of response categories. The adjacent-categories logit model is then $L_j = \alpha_j + \beta' \mathbf{x}$, $j = 1, \dots, J-1$.

(2) The continuation-ratio logits

They are defined as:

$$L_j = \log \left(\frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \dots + \pi_J(\mathbf{x})} \right), \quad j = 1, \dots, J-1$$

or as

$$L_j^* = \log \left(\frac{\pi_{j+1}(\mathbf{x})}{\pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})} \right), \quad j = 1, \dots, J-1.$$

Let $\rho_j(\mathbf{x})$ denote the probability of response j , given response j or higher; that is, $\rho_j(\mathbf{x}) = \frac{\pi_j(\mathbf{x})}{\pi_j(\mathbf{x}) + \dots + \pi_J(\mathbf{x})}$, $j = 1, \dots, J-1$. The continuation-ratio logits $\{L_j\}$ are ordinary logits of these conditional probabilities, namely

$$L_j = \log \left(\frac{\rho_j(\mathbf{x})}{1 - \rho_j(\mathbf{x})} \right), \quad j = 1, \dots, J-1.$$

(3) The third way of using ordered response categories is by forming logits of cumulative probabilities,

$$F_j(\mathbf{x}) = \pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x}), \quad j = 1, \dots, J$$

The cumulative logits are defined as

$$L_j = \text{logit}[F_j(\mathbf{x})] = \log \left(\frac{F_j(\mathbf{x})}{1 - F_j(\mathbf{x})} \right) = \log \left(\frac{\pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \dots + \pi_J(\mathbf{x})} \right), \quad j = 1, \dots, J-1.$$

Each cumulative logit uses all J response categories. Cumulative logit models can be formed by using the cumulative logit link. For a J -category response, the $J-1$ cumulative logits can be incorporated into a single model. Let $\{L_j(\mathbf{x}) = \text{logit}[F_j(\mathbf{x})], j = 1, \dots, J-1\}$, where $F_j(\mathbf{x}) = P(Y \leq j | \mathbf{x})$ is the cumulative probability for response category j , when the explanatory variables take value $\mathbf{x}$. The cumulative logit model $L_j = \alpha_j$, $j = 1, \dots, J-1$ implies the response variable is simultaneously independent of all explanatory variables and the $\{\alpha_j\}$ are called cutpoint parameters. These cutpoint parameters are nondecreasing in j , since the cumulative logit is an increasing function of $F_j(\mathbf{x})$, which is itself increasing in j for fixed $\mathbf{x}$. To include effects of explanatory variables, we use model $L_j = \alpha_j + \beta' \mathbf{x}$, $j = 1, \dots, J-1$. It is called a proportional odds model. Other cumulative logit models, such as the uniform association model for cumulative odds ratios, can be formed.

Other types of links can be used to link the cumulative distribution function (cdf) of the response to the explanatory variables. Let G denote the cdf of a continuous random variable

having positive density over the real line. Then G^{-1} is a strictly monotone function from $(0, 1)$ onto the real line. The cumulative link model $G^{-1}[F_j(\mathbf{x})] = \alpha_j - \beta' \mathbf{x}$ links the cumulative probabilities $\{F_j(\mathbf{x}), j = 1, \dots, J\}$ to the real line, using the link function G^{-1} . The use of the standard normal cdf Φ for G gives the cumulative probit model, a generalization of the binary probit model to ordered response categories. The link $G^{-1}(u) = \log[-\log(1 - u)]$ is called the complementary log-log link and it is the inverse cdf for the extreme-value distribution. The ordinal model using this link is called a proportional hazards model. The cumulative probit model is appropriate when the underlying response distribution is normal whereas the proportional hazards model is appropriate when the underlying distribution is exponential or a type used in survival analysis.

(4) The last model presented is the mean response model. This regression-type model applies directly to an observed ordinal response, rather than to an assumed underlying scale. For scores $v_1 \leq v_2 \leq \dots \leq v_J$ and fixed value $\mathbf{x}$ for explanatory variables, let $M(\mathbf{x}) = \sum_j v_j \pi_j(\mathbf{x})$ denote the mean response. The model $M(\mathbf{x}) = \alpha + \beta' \mathbf{x}$ assumes a linear relationship between the conditional mean and the explanatory variables. Mean response models are primarily appropriate when we want to make inferences about underlying continuous variables. Fitting them is justified if the categorical nature of the response reflects crude measurement of an inherently continuous variable (Agresti, 1990).

3.3 Proportional odds model

The appropriate model to be selected to analyze the categorical variables of interest should satisfy some constraints. First, it must distinguish between the dependent and the explanatory variables; it should allow an ordinal response and categorical (either nominal or ordinal) explanatory variables; the interpretation of the parameters must be simple. McCullagh (1980) said that the proportional odds model and the proportional hazards model are useful

in practice because of the simplicity of their interpretation and because they are shown to be multivariate extensions of generalized linear models. However, the proportional hazards model is more appropriate for survival data, which is not the case here. That is why these considerations lead to the choice of the proportional odds model for the analysis of the NGS data.

In the previous section, we saw that to include the effect of the explanatory variables, we use model

$$L_j(\mathbf{x}) = \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, \dots, J-1 \quad (3.1)$$

This model assumes that the effect of a variable on the odds of response below category j is the same for all j . It satisfies

$$L_j(\mathbf{x}_1) - L_j(\mathbf{x}_2) = \log \left[\frac{P(Y \leq j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1))}{P(Y \leq j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} \right] = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2) \quad (3.2)$$

We call a cumulative odds ratio the odds ratio of cumulative probabilities in expression (3.2). The log of the cumulative odds ratio is proportional to the distance between the values of the explanatory variables, with the same proportionality constant applying to each cutpoint. It is because of this property that (3.1) is called a proportional odds model. Its interpretation is that the odds of observing response $\leq j$ are $\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$ times higher at $\mathbf{x} = \mathbf{x}_1$ than $\mathbf{x} = \mathbf{x}_2$. For J ordered categories and a single covariate, the fitted logits correspond to $J - 1$ parallel lines. Note that the explanatory variables in a cumulative logit model can be continuous or categorical.

In the preceding parametrization for a given covariate x_i , when $\beta_i > 0$, the cumulative logits increase as x_i increases. This means that Y tends to be smaller for higher values of x_i . Similarly, when β_i is negative, increases in x_i are associated with higher levels of Y . Sometimes, to make $\beta_i > 0$ have the more usual meaning of Y tending to be larger at higher values of x_i , β is replaced in (3.1) by $-\beta$. For this parametrization,

$$L_j(\mathbf{x}) = \alpha_j - \beta' \mathbf{x}, \quad j = 1, \dots, J-1 \quad (3.3)$$

Walker and Duncan (1967) and McCullagh (1980) gave Fisher scoring algorithms for iterative calculation of maximum likelihood estimates of parameters. The algorithms use expected values in the second derivative matrix, instead of observed values.

McCullagh (1980) discusses the invariance of the proportional odds model under a reversal of category order. He said that it is invariant since the parameter β merely changes sign and the $\{\alpha_j\}$ reverse sign and order. He also stated that the parameter or parameters of interest should not depend for their interpretation on the actual response categories involved although the estimate will in general be affected.

3.4 Assumptions of the model

- 1) This model assumes the effect of a variable on the odds of response below category j is the same for all j .
- 2) The response has a multinomial distribution.

3.5 Estimation of the parameters and variance

3.5.1 Generalized linear models

Generalized linear models (GLMs) are an extension of classical linear models. They are described in Dobson (1990) and summarized in section 3.5.1, section 3.5.2 and section 3.5.4. GLMs are specified by three components:

- 1) The random component

This identifies the probability distribution of the response variable. The random component of a GLM consists of independent observations $Y_1, \dots, Y_N$ from a distribution in the natural exponential family. That is, each observation Y_i has probability density function or mass function of the form

$$f(y_i; \theta_i) = a(\theta_i)e(y_i) \exp[y_i b(\theta_i)].$$

2) The systematic component

This specifies a linear function of explanatory variables that is used as a predictor. The systematic component relates a vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)'$ to a set of explanatory variables through a linear model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. Here, $\mathbf{X}$ is a model matrix (or a design matrix) consisting of values of explanatory variables for the N observations, and $\boldsymbol{\beta}$ is a vector of model parameters. The vector $\boldsymbol{\eta}$ is called the linear predictor.

3) The link function

Describes the functional relationship between the systematic component and the expected value of the random component. A monotonic differentiable link function g such that $g(\mu_i) = \eta_i$, $i = 1, \dots, N$ where $\mu_i = E(Y_i)$ links μ_i to η_i . Thus, the model links expected values of observations to explanatory variables through the formula $g(\mu_i) = \sum_j \beta_j x_{ij}$, $i = 1, \dots, N$.

3.5.2 Estimation for generalized linear models

We define the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^N y_i b(\theta_i) + \sum_{i=1}^N c(\theta_i) + \sum_{i=1}^N d(y_i)$

where $E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i)$ and $g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta} = \eta_i$.

Then, $\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$.

In general, the equations $U_j = 0$ are non-linear and they have to be solved by numerical iteration. If the Newton-Raphson method is used, then the m^{th} approximation is given by:

$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}^{-1} \mathbf{U}^{(m-1)}$ where $\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\boldsymbol{\beta} = \mathbf{b}^{(m-1)}}$ is the matrix of second derivatives of l evaluated at $\boldsymbol{\beta} = \mathbf{b}^{(m-1)}$ and $\mathbf{U}^{(m-1)}$ is the vector of first derivatives $U_j = \frac{\partial l}{\partial \beta_j}$ evaluated at $\boldsymbol{\beta} = \mathbf{b}^{(m-1)}$.

An alternative procedure is the method of scoring. We replace the matrix of second

derivatives in

$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta=\mathbf{b}^{(m-1)}}^{-1} \mathbf{U}^{(m-1)}$ by $E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]$. The information matrix $\mathbf{F} = E[\mathbf{U}\mathbf{U}']$

has the elements $F_{jk} = E[U_j U_k] = E \left[\frac{\partial l}{\partial \beta_j} \frac{\partial l}{\partial \beta_k} \right] = -E \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]$.

Thus we have $\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + [\mathbf{F}^{(m-1)}]^{-1} \mathbf{U}^{(m-1)}$.

For the generalized linear model, the (j, k) element of $\mathbf{F}$ is $F_{jk} = \sum_{i=1}^N \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$.

Thus $\mathbf{F} = \mathbf{X}'\mathbf{W}\mathbf{X}$ where $\mathbf{W}_{N \times N}$ is the diagonal matrix with elements

$w_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$. $\mathbf{F}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$ is the vector with elements

$\sum_k \sum_i \frac{x_{ij}x_{ik}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$ evaluated at $\mathbf{b}^{(m-1)}$.

Thus $\mathbf{F}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} = \mathbf{X}'\mathbf{W}\mathbf{z}$ where $\mathbf{z}$ has elements

$z_i = \sum_k x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$ with μ_i and $\frac{\partial \mu_i}{\partial \eta_i}$ evaluated at $\mathbf{b}^{(m-1)}$.

The iterative equation for the method of scoring can be written as $\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}'\mathbf{W}\mathbf{z}$.

3.5.3 Note about the existence of maximum likelihood estimates

McCullagh (1980) states that in practice, problems occasionally arise with infinite parameter values associated with patterns of zeros in sparse data arrays. Furthermore, in small samples, convergence problems not associated with patterns of zeros can arise when non-linear models are used. In this case, non-convergence is usually a reliable indication that the model being fitted is inappropriate.

3.5.4 Inferences

Testing hypotheses about the parameters can be done knowing the sampling distribution of

the maximum likelihood estimates $\hat{\beta}_j$. Suppose that the log-likelihood function has a unique maximum at $\mathbf{b}$ (near the true value of β). The first-order Taylor approximation for $U(\beta)$ (when $U(\beta)$ is the score statistic at the point β) about the point $\beta = \mathbf{b}$ is given by:

$U(\beta) \approx U(\mathbf{b}) + H(\mathbf{b})(\beta - \mathbf{b})$ where $H(\mathbf{b})$ denotes the matrix of second derivatives of the log-likelihood function evaluated at $\beta = \mathbf{b}$. Asymptotically, H is equal to its expected value which is related to the information matrix by $\mathcal{I} = E[\mathbf{U}\mathbf{U}'] = E[-H]$. For large samples, $U(\beta) \approx U(\mathbf{b}) - \mathcal{I}(\beta - \mathbf{b})$.

But $U(\mathbf{b}) = 0$ because $\mathbf{b}$ is the point at which the log-likelihood function is maximal and its derivatives are 0. Therefore, $(\mathbf{b} - \beta) \approx \mathcal{I}^{-1}\mathbf{U}$. If $\mathcal{I}$ is regarded as constant, then $E(\mathbf{b} - \beta) \approx \mathcal{I}^{-1}E(\mathbf{U}) = 0$ where $\mathbf{b}$ is an unbiased (at least asymptotically) estimator of β .

We have that $E[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] \approx \mathcal{I}^{-1} \underbrace{E[\mathbf{U}\mathbf{U}']}_{\mathcal{I}} \mathcal{I}^{-1} = \mathcal{I}^{-1}$ and $(\mathcal{I}^{-1})' = \mathcal{I}^{-1}$.

For large samples, $(\mathbf{b} - \beta)' \mathcal{I}(\mathbf{b} - \beta) \sim \chi_p^2$ or $(\mathbf{b} - \beta) \sim N(\mathbf{0}, \mathcal{I}^{-1})$. Sometimes, $(\mathbf{b} - \beta)' \mathcal{I}(\mathbf{b} - \beta)$ is called the Wald statistic. The Wald statistic can be used to test several constraints. We can define $W = (\mathbf{C}\hat{\beta})'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}(\mathbf{C}\hat{\beta})$ when $\mathbf{C}$ represents a linear contrast coefficient matrix; $\mathbf{V}$ is the design-consistent variance-covariance matrix of $\hat{\beta}$; and $\hat{\beta}$ is the vector of the estimated parameters. The Wald statistic has r degrees of freedom ($r = \text{rank}(\mathbf{C})$). More specifically, we can test $\beta_j = 0$ by using $\mathbf{C} = (0, \dots, 0, 1, 0, \dots, 0)$ and the corresponding Wald statistic has one degree of freedom.

To assess the adequacy of the model, we can compare the likelihood under this model with the likelihood under the maximal or saturated model. The maximal model is a GLM using the same distribution and same link function as the model of interest with a number of parameters equal to the total number of observations N . We obtain the MLE $\mathbf{b}_{\max}$ and $\mathbf{b}$ to obtain values $L(\mathbf{b}_{\max}; \mathbf{y})$ and $L(\mathbf{b}; \mathbf{y})$. If the model of interest describes the data well then $L(\mathbf{b}; \mathbf{y})$ should be equal to $L(\mathbf{b}_{\max}; \mathbf{y})$. The generalized ratio statistic $\lambda = L(\mathbf{b}_{\max}; \mathbf{y}) / L(\mathbf{b}; \mathbf{y})$ or $\log(\lambda)$ is then obtained. If $\log(\lambda)$ is too large, we can say that the model of interest is a

poor description of the data.

Suppose that the model of interest involves p parameters. Taylor series approximation for $l(\boldsymbol{\beta}; y)$ can be obtained:

$l(\boldsymbol{\beta}; y) \approx l(\mathbf{b}; y) + (\boldsymbol{\beta} - \mathbf{b})' U(\mathbf{b}) + (1/2)(\boldsymbol{\beta} - \mathbf{b})' H(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})$. For large samples, $-H(\mathbf{b})$ can be approximated by $\boldsymbol{\Sigma}$ because $\boldsymbol{\Sigma} = E(-H)$. So

$$l(\mathbf{b}; y) - l(\boldsymbol{\beta}; y) = (1/2) \underbrace{(\mathbf{b} - \boldsymbol{\beta})' \boldsymbol{\Sigma} (\mathbf{b} - \boldsymbol{\beta})}_{\sim \chi_p^2}$$

Therefore,

$2[l(\mathbf{b}; y) - l(\boldsymbol{\beta}; y)] \sim \chi_p^2$. The deviance statistic is defined by

$D = 2 \log(\lambda) = 2[l(\mathbf{b}_{\max}; y) - l(\mathbf{b}; y)]$. The model is good if $D \sim \chi_{N-p}^2$. The sampling distribution of the log-likelihood ratio statistic can be used to investigate the adequacy of a model by estimating D from the data and comparing the value with the appropriate chi-squared distribution. If a model with p parameters provides a good description for a data set of N observations, we can expect $D \approx N - p$ (which is $E(\chi_{N-p}^2)$).

3.6 Binder's method

Since the NGS data come from a complex survey design, the weighting and the sample design must be taken into account when the parameters and the variance of these parameters are estimated. Binder (1983) considered the problem of specifying and estimating the variance of estimated parameters based on complex sample designs from finite populations. First he distinguished between the concepts of implicit and explicit parameters. For example, in the linear regression context, the implicit regression parameter $\mathbf{B}$ is defined by $\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y}$ where $\mathbf{X}$ is an $N \times p$ matrix and $\mathbf{Y}$ is an $N \times 1$ vector. It is true to say that $\mathbf{B}$ could be explicitly defined by $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (assuming that $\mathbf{X}$ is of full rank). But in his paper, Binder has shown that the implicit definition gives more tractable results for the variances of the estimated regression coefficients.

He defined the generalized linear model as follows: given parameters θ and ϕ , the density function for observation y from an infinite population is given by $p(y; \theta; \phi) = \exp[\alpha(\phi)\{y\theta - g(\theta) + h(y)\} + \gamma(\phi, y)]$, where $\alpha(\phi) > 0$. Note that $E(Y) = g'(\theta) = \mu(\theta)$ and $V(Y) = \mu'(\theta)/\alpha(\phi)$. For a given vector $\mathbf{x} = (x_1, \dots, x_p)$, we assume that $\theta = f(\sum x_i \beta_i)$ where $f(\cdot)$ is a known differentiable function and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is unknown. For observations $(y_k, \mathbf{x}_{1k}, \dots, \mathbf{x}_{pk})$, where $k = 1, \dots, N$, the maximum likelihood solution for $\boldsymbol{\beta}$ is the solution of

$$\sum_{k=1}^N [y_k - \mu\{f(\mathbf{x}'_k \boldsymbol{\beta})\}] f'(\mathbf{x}'_k \boldsymbol{\beta}) x_{ik} = 0 \quad (i = 1, \dots, p).$$

If $\mu(\cdot)$ and $f'(\cdot)$ are strictly monotone functions then the solution to $\boldsymbol{\beta}$ is unique, whenever the $N \times p$ matrix of x values is of full rank. In the finite population context, he assumes that for the k^{th} individual we observe $(Y_k, X_{1k}, \dots, X_{pk})$. He defines the population parameter $\mathbf{B}$ as the solution of

$$\sum_{k=1}^N [Y_k - \mu\{f(\mathbf{X}'_k \mathbf{B})\}] f'(\mathbf{X}'_k \mathbf{B}) X_{ik} = 0 \quad (i = 1, \dots, p).$$

Although the parameter $\mathbf{B}$ has been motivated by infinite population concepts, we are only concerned with the design-based sampling distributions of $\hat{\boldsymbol{\beta}}$, defined in terms of the finite population values.

Now consider the finite population parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ defined by an expression of the form

$$\mathbf{W}_N(\boldsymbol{\theta}) = \sum_{k=1}^N \mathbf{u}(\mathbf{Z}_k; \boldsymbol{\theta}) - \mathbf{v}(\boldsymbol{\theta}) = 0, \quad (3.4)$$

where $\mathbf{Z}_k = (Z_{1k}, \dots, Z_{qk})$, the data values for the k^{th} unit, and $\mathbf{Z}$ is the $N \times q$ matrix of the data values for all units in the finite population. The term $\mathbf{v}(\boldsymbol{\theta})$ is included to allow for explicitly defined parameters. The distinction is necessary because for a given $\boldsymbol{\theta}$, we know the value of $\mathbf{v}(\boldsymbol{\theta})$, but $\mathbf{u}(\mathbf{Z}_k; \boldsymbol{\theta})$ is only known for those units in the sample.

Binder then gave an estimator for θ defined by (3.4). For any given value of θ , the term $\sum_{k=1}^N \mathbf{u}(\mathbf{Z}_k; \theta)$ is the population total of functions of the data values $\mathbf{Z}_1, \dots, \mathbf{Z}_k$. The estimator for these totals is represented by $\widehat{\mathbf{U}}(\theta)$ and it is simply the estimator of a total based on data values $\mathbf{u}(\mathbf{Z}_1; \theta), \dots, \mathbf{u}(\mathbf{Z}_n; \theta)$. It is assumed to be asymptotically normal with mean $\mathbf{U}(\theta) = \sum_{k=1}^N \mathbf{u}(\mathbf{Z}_k; \theta)$ and variance $\sum_U(\theta)$. A consistent estimator for $\sum_U(\theta)$ is denoted by $\widehat{\sum_U}(\theta)$. Now let

$$\widehat{\mathbf{W}}(\theta) = \widehat{\mathbf{U}}(\theta) - \mathbf{v}(\theta) \quad (3.5)$$

and $\widehat{\theta}$ is defined by the equations $\widehat{\mathbf{W}}(\widehat{\theta}) = \mathbf{0}$.

Finally, he obtained the variance of $\widehat{\theta}$ by taking the Taylor expansion of $\widehat{\mathbf{W}}(\widehat{\theta})$ at $\widehat{\theta} = \theta_0$, where θ_0 is the population parameter value. Then we obtain

$$0 = \widehat{\mathbf{W}}(\widehat{\theta}) \simeq \widehat{\mathbf{W}}(\theta_0) + \frac{\partial \widehat{\mathbf{W}}(\theta_0)}{\partial \theta_0} (\widehat{\theta} - \theta_0), \quad \text{so that} \quad \widehat{\mathbf{W}}(\theta_0) \simeq -\frac{\partial \widehat{\mathbf{W}}(\theta_0)}{\partial \theta_0} (\widehat{\theta} - \theta_0).$$

Taking variances of both sides, we obtain in the limit

$$\sum_U(\theta_0) = \left[\frac{\partial \mathbf{W}_N(\theta_0)}{\partial \theta_0} \right] \mathbf{V}(\widehat{\theta}) \left[\frac{\partial \mathbf{W}_N(\theta_0)}{\partial \theta_0} \right]' \quad \text{or equivalently}$$

$$\mathbf{V}(\widehat{\theta}) = \left[\frac{\partial \mathbf{W}_N(\theta_0)}{\partial \theta_0} \right]^{-1} \sum_U(\theta_0) \left[\frac{\partial \mathbf{W}_N(\theta_0)}{\partial \theta_0} \right]^{-1}, \quad (3.6)$$

providing $\frac{\partial \mathbf{W}_N}{\partial \theta_0}$ is of full rank. An estimator of $\mathbf{V}(\widehat{\theta})$ is given by

$$\widehat{\mathbf{V}}(\widehat{\theta}) = \left[\frac{\partial \widehat{\mathbf{W}}(\widehat{\theta})}{\partial \widehat{\theta}} \right]^{-1} \widehat{\sum_U}(\widehat{\theta}) \left[\frac{\partial \widehat{\mathbf{W}}(\widehat{\theta})}{\partial \widehat{\theta}} \right]^{-1}. \quad (3.7)$$

The required assumptions of the approach presented above for the estimation of θ and the variance of $\widehat{\theta}$ are listed in Appendix C.

3.7 Goodness-of-fit and residuals

When a model is fitted to data, a measurement of the discrepancy between the observed and fitted values must be done as an assessment of the goodness-of-fit. Two measures are presented here. The first one is called the deviance statistic and has been presented in section 3.5.4. McCullagh and Nelder (1989) give the deviance function for models with polytomous data. Let $\pi_{ij} = \frac{y_{ij}}{m_i}$ be the point at which the maximum achievable log likelihood occurs, when y_{ij} is the total number of observations having response j ($1, \dots, J$) and explanatory variable i ($1, \dots, I$), and m_i is the number of observations having explanatory variable i . Note that when there is more than one explanatory variable in the model, I represents the number of combinations of the levels of these explanatory variables available in the sample. Therefore, the deviance function is:

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\pi}}) &= 2l(\boldsymbol{\pi}, \mathbf{y}) - 2l(\hat{\boldsymbol{\pi}}, \mathbf{y}) \\ &= 2 \sum_{ij} y_{ij} \log(\pi_{ij}) - 2 \sum_{ij} y_{ij} \log(\hat{\pi}_{ij}) \end{aligned}$$

where $\hat{\pi}_{ij}$ is the fitted probability of having response j and explanatory variable i . If $m_i \hat{\pi}_{ij}$ are sufficiently large and there is no over-dispersion (i.e. the variance of the response Y does not exceed the nominal variance), the deviance function has an approximate χ^2 distribution (McCullagh & Nelder, 1989).

An other important measure of discrepancy is the generalized Pearson chi-squared statistic, which is defined by:

$$X^2 = n \sum_i \hat{W}_i \sum_j (\pi_{ij} - \hat{\pi}_{ij})^2 / \hat{\pi}_{ij}$$

where $\hat{W}_i = m_i/n$. Under independent multinomial sampling in each of the domains, the X^2 has an asymptotically χ^2 distribution with $I(J-1) - p$ degrees of freedom (p is the number of parameters in the model). For further details, see Rao, Kumar, and Roberts (1989).

When sparseness occurs in a sample, the deviance function and the Pearson's statistic are

poor indicators of the goodness-of-fit of the model. Sparseness here means that a sizeable proportion of the observed counts are small. Examples of the effect of sparseness on the deviance function and the Pearson's statistic is given in McCullagh & Nelder (1989). It is preferable, when there is presence of sparseness, not to rely only on the deviance function or the Pearson's chi-square statistic as absolute measures of goodness-of-fit.

These two measures of goodness-of-fit are not corrected for the sample design used to obtain the survey data. Roberts (1985) provided (in her Ph.D. thesis) adjustments to the Pearson chi-squared statistic that take into account the sampling plan for polytomous models (which include proportional odds models). These corrected tests are also given in Rao, Kumar, and Roberts (1989), where we find two corrections for general polytomous models. Suppose that the population of interest is divided into I cells (or domains) according to the levels of one or more factors. Let $P_{j(i)}$ be the population proportion in the i^{th} cell ($i = 1, \dots, I$) having response j ($j = 1, \dots, J$) so that $\sum_j P_{j(i)} = 1$. Define a general polytomous model as:

$$P_{i(j)} = F_{ij}(\theta), \quad i = 1, \dots, I; \quad j = 1, \dots, J-1 \quad (3.8)$$

where θ is a p -vector of unknown parameters ($p \leq I(J-1)$) and $F_{ij}(\theta)$ is a function of known form.

Now let $\hat{\mathbf{P}} = (\widehat{P_{1(1)}}, \dots, \widehat{P_{J(I)}})'$ be the survey estimate of $\mathbf{P}$ and $\hat{\mathbf{F}} = F(\hat{\theta}) = (\widehat{F'_1}, \dots, \widehat{F'_I})'$,

where $\widehat{F'_i} = (\widehat{F_{i1}}, \dots, \widehat{F_{iJ}})'$, and $\widehat{F}_{ij} = F_{ij}(\hat{\theta})$, be the fitted response proportions. The survey estimate of W_i (the domain proportion) is $\widehat{W}_i$.

To test a nested hypothesis $\theta_2 = 0$, given the model (3.8), we obtain the reduced model $P_{i(j)} = F_{ij}(\theta_1)$, where $F_{ij}(\theta_1) = F_{ij}(\theta_1, 0)$. Let $\hat{\theta}_1$ be the pseudo maximum likelihood estimate of θ_1 and $\hat{\mathbf{F}} = F\left(\begin{smallmatrix} \hat{\theta}_1 \\ 0 \end{smallmatrix}\right)$ be the corresponding vector of fitted response proportions, where $\theta' = (\theta'_1, \theta'_2)$, θ_1 is $q \times 1$ and θ_2 is $u \times 1$ ($q + u = p$). Let

$X^2(2|1) = n \sum_i W_i \sum_j \frac{(\hat{F}_{ij} - \hat{\bar{F}}_{ij})^2}{\hat{\bar{F}}_{ij}}$ be the Pearson chi-squared test of the nested hypothesis,

i.e. the test of the hypothesis that $\theta_2 = 0$ given that θ_1 is known (2|1). A first-order correction to $X^2(2|1)$ is obtained by:

$X_c^2(2|1) = X^2(2|1) / \hat{\delta} \cdot (2|1)$ as χ^2 with u degrees of freedom. We obtain $\hat{\delta} \cdot (2|1)$ by replacing θ' by $(\hat{\theta}'_1, 0')$ and V_p (the covariance matrix of $\hat{\mathbf{P}}$) by $\hat{V}_p$ in the following definition:

$$u\delta \cdot (2|1) = \sum_{i=1}^u \delta_i(2|1) = \text{trace}(D(2|1)).$$

The matrix $D(2|1)$ is a generalized design effects matrix given by

$$D(2|1) = (H'_2 \nabla H_2)^{-1} (H'_2 \nabla V_p \nabla' H_2),$$

where $\nabla = (D(W) \otimes I)Q^{-1}$, Q is the block diagonal matrix with $Q_i = \text{diag}(F_i) - F_i F'_i$, $F_i = F_i(\theta)$, and $H_2 = [I - M_1(M'_1 \nabla M_1)^{-1} M'_1 \nabla] M_2$, where $M_1 = \frac{\partial F}{\partial \theta'_1}$ and $M_2 = \frac{\partial F}{\partial \theta'_2}$.

A second-order correction to $X^2(2|1)$, based on the Satterthwaite approximation, is obtained by

$$X_s^2(2|1) = X_c^2(2|1) / [1 + \hat{a}(2|1)^2] \text{ as } \chi^2 \text{ with } u/[1 + \hat{a}(2|1)^2] \text{ degrees of freedom.}$$

The term $\hat{a}(2|1)^2$ is obtained by replacing θ' by $(\hat{\theta}'_1, 0')$ in the definition of $a(2|1)^2$ given below:

$$a(2|1)^2 = \left\{ \sum_{i=1}^u \delta_i(2|1)^2 - u\delta_1(2|1)^2 \right\} / u\delta \cdot (2|1)^2,$$

where $\sum_i \delta_i(2|1)^2 = \text{trace}(D(2|1)^2)$.

These two corrections imply knowledge of the covariance matrix of $\hat{\mathbf{P}}$ to obtain $\hat{\delta}$. When the estimated covariance matrix of $\hat{\mathbf{P}}$ is not available, Roberts (1985) states that it is worthwhile to obtain an upper bound on δ . which depends only on the cell design effects or on the

generalised design effects.

Other general methods of measure of discrepancy are residuals, often used to detect outliers. McCullagh (1980) states that the analysis of residuals from multinomial response models is complicated by several factors:

1. The residuals, however standardized, must take one of a limited number of possible values which causes problems in rankit or related plots when the cell counts are small.
2. It is far from clear in general that cell residuals are the relevant quantities to examine. What is important in the data is how the count for a cell or group of cells varies relative to another cell or group of cells. For binary data, the usual standardized residual is $(p - \hat{p})/(\hat{p}\hat{q}/n)^{1/2}$ which is the signed square root of the contribution to Pearson's chi-squared statistic. An analogous residual could be defined by using the contribution to the likelihood ratio statistic (deviance) instead.

He proposed to use as a residual the contribution to the likelihood ratio statistic from each multinomial sample, which is always positive. At a second stage in an analysis of residuals, cell residuals can be computed to examine the precise nature of the observed discrepancy. Different cell residual patterns in a single multinomial sample indicate different inadequacies in the model. A plot of each of the contributions to the deviance function against its index (i, j) can help detecting outlying residuals.

McCullagh and Nelder (1989) obtained in their example 5.6.1 the cell residuals from a proportional odds model. They compare the cell residuals, i.e. $(y_{ij} - \hat{y}_{ij})/[m_i \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]^{1/2}$ to the value 2.0, in order to detect too large residuals. They also based other residual calculations on the cumulative totals $z_{ij} = y_{i1} + \dots + y_{ij}$. But using residuals based on z_{ij} raises the problem of correlation among these residuals.

A last method of testing the adequacy of the model is the Wald chi-square statistic, used to test that all betas in the model are equal to zero. Accepting this hypothesis would imply

really bad predictors of the logit of the cumulative conditional probabilities.

All of these methods are attempts at assessing the goodness-of-fit of any proportional odds model. In fact, other methods or types of residuals exist for logistic regression models or generalized linear models. But the focus of this research was not on the extensive development or the adjustment of these methods or residuals for the proportional odds model, but only on having a general idea of the goodness-of-fit of the model, using existing methods with little modifications. Therefore, graphical methods from Landwehr and al. (1984) or Atkinson (1981) have not been implemented.

CHAPTER 4: PROPORTIONAL ODDS ANALYSIS OF CANADIAN UNIVERSITY GRADUATES 1986-1995

This chapter describes the software that was used to fit a proportional odds model using complex survey data and presents the results of a proportional odds model used to characterize the university level of certification obtained at graduation, based on data from Canadian university graduates 1986 - 1995.

4.1 Proportional odds model in SUDAAN

SUDAAN (release 7.5.2) was the software used to fit a proportional odds model to the NGS data, via the MULTLOG procedure. SUDAAN implements the proportional odds model (McCullagh, 1980) with cumulative logit link for ordinal responses (with two or more categories). The MULTLOG procedure estimates model parameters using generalized estimating equations (GEE). For estimating robust variances of the parameter estimates MULTLOG implements two methods described in section 3.6 (Binder's method) and in Zeger and Liang (1986). In addition, SUDAAN estimates the model-based, or naïve, variance of the parameter estimates. All three variance methods allow a choice of independent vs. exchangeable working correlations for describing the dependence of responses within clusters.

The proportional odds model used in SUDAAN is described as follows:

$L_j(\mathbf{x}_i) = \text{logit}[F_j(\mathbf{x}_i)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i$, $j = 1, \dots, J-1$ and $i = 1, \dots, N$ (denotes the subject) and $\mathbf{x}_i$ denotes the vector of the p explanatory variables for the subject i . This model is the same as the one described in section 3.3 Note that the sign between α_j and $\boldsymbol{\beta}'\mathbf{x}_i$ is a plus and not a minus. This will have to be taken into account for the interpretation of the parameters.

In the MULTLOG procedure, the sample design can be specified. In this analysis, the option “STRWOR” is used since the NGS sample design is a stratified systematic random sample design. The weight assigned to each graduate in the sample was also used for the fitting using the statement “WEIGHT”. Then, the total number of graduates in the population by stratum had to be known in order to be included in the statement “TOTCNT”. The statement “SAMCNT” determined simply the number of graduates in the sample by stratum.

4.1.1 Estimation of model parameters

The method of Generalized Estimating Equations (GEEs) used by SUDAAN’s MULTLOG procedure for fitting and evaluating multinomial logit models bears a close resemblance to the survey regression techniques proposed by Binder (1983). In fact, the estimating equations for both of the methods are identical under the assumption of independent working correlations. It should be noted that the estimating equations are the derivatives of the log likelihood (or weighted log likelihood for survey data) under the assumption of independent working correlations.

The SUDAAN User’s manual (release 7.5.2) gives a brief overview of the estimation method in SUDAAN. Let i index the clusters, t index the observational units within clusters, and k index the level of response within clusters with $i = 1, \dots, n$; $t = 1, \dots, T_i$; $k = 1, \dots, K$. Define K indicator random variables Y_{ikt} , where $Y_{ikt} = 1$ if the t -th unit from cluster i has response k , and 0 otherwise. The response vector for each observational unit within each cluster is a $(K - 1) \times 1$ vector $\mathbf{Y}_{it} = [Y_{i1t}, \dots, Y_{i,K-1,t}]'$. Grouping the response vectors together yields $\mathbf{Y}_i = [\mathbf{Y}'_{i1}, \dots, \mathbf{Y}'_{iT_i}]'$. Each cluster has also T_i $P \times 1$ covariate vectors $\mathbf{x}_{it}$, where P is the number of regression coefficients for the covariate vector $\mathbf{x}_{it}$. Let $\pi_{ikt} = \pi_{ikt}(\boldsymbol{\beta}) = E(Y_{ikt}|\mathbf{x}_{it}, \boldsymbol{\beta}) = \text{prob}(Y_{ikt} = 1|\mathbf{x}_{it}, \boldsymbol{\beta})$ be the probability of response k for unit t . MULTLOG allows the user to make inferences about the parameters $\boldsymbol{\beta}$ associated with the

vector of marginal probabilities $\pi_{it}(\boldsymbol{\beta}) = [\pi_{i1t}, \dots, \pi_{i,K-1,t}]'$. We can group these vectors together to form the $T_i(K-1) \times 1$ vector $\boldsymbol{\pi}_i = [\boldsymbol{\pi}'_{i1}, \dots, \boldsymbol{\pi}'_{iT_i}]'$.

Let $\mathbf{A}_i$ be a diagonal matrix with diagonal elements equal to the marginal variances of observational units within clusters, $\text{var}(\mathbf{Y}_{it})$, where the marginal distribution of $\mathbf{Y}_{it}$ is multinomial. Following this, $V_i(\alpha) = \mathbf{A}_i^{1/2} R_i(\alpha) \mathbf{A}_i^{1/2}$ is the working covariance matrix for the vector of responses within each cluster specified in terms of the pairwise correlations between elements of the vectors $\mathbf{Y}_{is}$ and $\mathbf{Y}_{it}$ and the marginal variance matrices. Let $R_i(\alpha)$ be the working correlation matrix, which describes the dependence of responses between pairs of observations within the cluster and α represents the unknown parameters associated with a specified model for the pairwise correlations.

The GEE estimate of $\boldsymbol{\beta}$ is the solution to the generalized estimating equation

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\partial \boldsymbol{\pi}'_i}{\partial \boldsymbol{\beta}} W_i^{1/2} \mathbf{V}_i(\hat{\alpha})^{-1} W_i^{1/2} (\mathbf{Y}_i - \boldsymbol{\pi}_i) = \mathbf{0},$$

where $\boldsymbol{\pi}_i = E(\mathbf{Y}_i) = (\pi'_{i1}, \dots, \pi'_{iT_i})'$ is the mean response vector for observations and responses within a cluster, $\frac{\partial \boldsymbol{\pi}'_i}{\partial \boldsymbol{\beta}}$ is the vector of first partial derivatives of the mean response $\boldsymbol{\pi}_i$ with respect to the regression coefficients $\boldsymbol{\beta}$; $\hat{\alpha}$ is a consistent estimate of α ; and W_i is the diagonal matrix of sampling design weights (which equal 1 for most non-survey designs). We can see $(\mathbf{Y}_i - \boldsymbol{\pi}_i)$ as the vector of observed minus expected values within each cluster, and the working covariance matrix $\mathbf{V}_i(\hat{\alpha})$ is inserted only to weight the data efficiently. The goal is to find the estimates of the regression coefficients that make $(\mathbf{Y}_i - \boldsymbol{\pi}_i)$ as small as possible. A modified Newton-Raphson method is used to simultaneously solve the estimating equations, updating $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ at each step of the iterative process. As long as the regression for the marginal responses is correctly specified, the parameter estimates are consistent and asymptotically normally distributed, regardless of the true correlation structure.

The SUDAAN User's manual also states that SUDAAN currently implements two models for the working correlations: independence and exchangeable correlations. For the independence working assumption, the correlation matrix is the identity matrix, denoting zero correlation between pairs of observations within the cluster. Under exchangeability for binary responses ($K = 2$), only a single pairwise correlation parameter ρ is needed to describe the dependence of responses within clusters. For $K > 2$ response levels, correlations may potentially depend on the observation times s and t as well as on response levels j and k . Under exchangeability for $K > 2$, the correlations do not depend on observational units s and t , but they do depend on the response levels j and k , that is, $\text{Corr}(Y_{ijs}, Y_{ikt}) = \rho_{jk}$. The complete correlation matrix $\mathbf{R}_i(\alpha)$ is of dimension $(K-1)T_i$ by $(K-1)T_i$. Under exchangeability, and following Lipsitz *et al* (1994), SUDAAN estimates a $(K-1)$ by $(K-1)$ symmetric correlation matrix which holds for all pairs of times $s \neq t$. The number of unique correlation parameters to be estimated is $K(K-1)/2$. The symmetric nature of the correlation matrix implies that the correlation for being in category j at time s and category k at time t is equivalent to the correlation for being in category k at time s and category j at time t . The estimates of the variance-covariance matrix of β can be made robust to the misspecification of the working correlation structure (Diggle *et al.*, 1994). However, efficiency is improved when the working correlation structure is consistent with empirical correlations.

4.1.2 Estimating variances

In the MULTLOG procedure, the variance estimates for regression parameters can be either robust (Binder or Zeger) or model-based. The sampling errors produced by the two robust methods are different when the outcome variable has more than two levels. There is no research to favor either robust method. Binder's method as described in chapter 3 was the one chosen for the analysis. The model-based variance estimate assumes that the working correlation structure was correct. But under cluster sampling, the model-based variance

estimate is sensitive to departures from the assumed correlation structure.

When there is only one sample unit within a stratum, we can add the option “MISSUNIT” in the statement “NEST” so that the variance contribution of that unit is estimated using the difference in that unit’s value and the overall mean value for the population. This option has been included since there were some strata for which the sample count was one.

4.2 Implementation of a proportional odds model using SUDAAN

4.2.1 Subpopulation under study

For each of the three NGSs, not all of the respondents on the three files have been retained. Only the graduates from the university strata were kept (those who graduated either from a trade/vocational program or a college program were excluded). Since one goal of the research was to characterize the production within Canada of university teachers, only Canadian residents at the time of the interview (NGS) were kept. It was desired to know the number of potential university teachers already living in Canada. Since it was not known from the questionnaire if the graduates who lived outside Canada at the time of the survey were planning on returning to Canada, they have been excluded of the analysis. Also, the graduates having for at least one of the explanatory variables the value “Unknown” or “Not stated” were not retained in the analysis. No interesting conclusions could be drawn for the respondents who refused either to answer a question or who did not know the answer. The explanatory variables are listed in section 4.2.2.

For the 1986 class, 15,466 graduates were kept from the original sample; this is equivalent to 115,125 graduates in the population of interest. For the 1990 class, 15,653 graduates were kept from the original sample; this is equivalent to 118,070 graduates in the population of interest. For the 1995 class, 21,317 graduates were kept from the original sample; this is equivalent to 153,829 graduates in the population of interest.

4.2.2 Model definition

The dependent variable of the proportional odds model fitted in SUDAAN is the level of certification (CERST) and the modalities are ordered from the lower level to the higher. This is one of the stratification variables and its three categories are:

- 1) Bachelor's degree; first professional degree; university diploma below bachelor
- 2) Master's degree; university diploma above bachelor
- 3) Earned Ph.D.

The explanatory variables are all categorical. They are either nominal, ordinal, or interval.

The list is:

Gender (GNDN):

- 1: Men
- 2: Women

Field of study (FOS):

- 1) No specialization/ general arts & science/ unknown
- 2) Education
- 3) Fine arts & Humanities
- 4) Commerce, law and economics
- 5) Other Social sciences
- 6) Agricultural and biological sciences
- 7) Engineering and applied sciences
- 8) Medical and health professions
- 9) Mathematics and physical sciences

Status of registration (FTPT):

- 1) Full-time only
- 2) Part-time only

3) Combination of full-time and part-time

First spoken language and still understood (LANG):

- 1) Other than English or French
- 2) English or French

Region of residence at the time of the NGS (REG):

- 1) Atlantic provinces
- 2) Quebec
- 3) Ontario
- 4) Prairies
- 5) British Columbia
- 6) Yukon and Northwest Territories

Age at the time of the NGS (CATAGE):

- 1) From 15 to 24 years old
- 2) From 25 to 29 years old
- 3) From 30 to 34 years old
- 4) 35 years old and over

In SUDAAN, it is possible to specify a subsample of the original sample for analysis purpose. This is done with the statement "SUBPOPN". The subpopulation was defined as described in section 4.2.1.

The input file had to include the total population counts by stratum and the sample counts, also by stratum. The sample counts by stratum are in fact the number of observations on the original data file in each stratum. The total population counts by stratum correspond to the total number of graduates on the survey frame, from which the sample of graduates was drawn. When the sampling was done, the population counts were based on administrative records. So these administrative records should be used as the population counts by stratum.

However, the population estimates based on the weights are more accurate when we talk about the population of interest (e.g. the graduates) because the weighting takes account of “out-of-scope” units (i.e. deaths, duplicates,...). We know that the weights are available on the data file for each of the three NGS surveys (1988, 1992, 1997). It has been possible to obtain the population counts based on the administrative records for the NGS 1997 only. So the analysis will be performed twice for this survey: once with the population counts based on the administrative records and once with the population estimates based on the weights. We will be able to assess if there is an impact on the results on using either type of population counts.

In PROC MULTLOG, the strata are identified by the variable STRA in the statement “NEST”. The variable STRA goes from 1 to the number of strata by input file (1986 file, 1990 file or 1997 file). The method chosen for estimating the variance of a parameter is Binder’s method, described in chapter 3 (SEMETHOD=BINDER). The independence correlation matrix between observations is selected and under this assumption, the estimating equations using the Liang & Zeger method are equal to the estimating equations using Binder’s method. The option DESIGN=STRWOR was selected since the sample design is a stratified and systematic random one, without replacement.

Models allowing different interaction terms and main effects have been fitted for each of the three surveys (including two fits with different population counts for NGS 1997). Then a comparison of the results of each of these four models has been performed. The following model has been fitted, using the CUMLOGIT option in the MODEL statement (to define the proportional odds model):

$$\begin{aligned}
\log \left(\frac{F_j(\mathbf{x})}{1 - F_j(\mathbf{x})} \right) = & \alpha_{j=1} + \alpha_{j=2} + \beta_1 FOS_1 + \beta_2 FOS_2 + \beta_3 FOS_3 \\
& + \beta_4 FOS_4 + \beta_5 FOS_5 + \beta_6 FOS_6 + \beta_7 FOS_7 + \beta_8 FOS_8 + \beta_9 FOS_9 \\
& + \beta_{10} GNDN_1 + \beta_{11} GNDN_2 + \beta_{12} REG_1 + \beta_{13} REG_2 + \beta_{14} REG_3 \\
& + \beta_{15} REG_4 + \beta_{16} REG_5 + \beta_{17} REG_6 + \beta_{18} FTPT_1 + \beta_{19} FTPT_2 \\
& + \beta_{20} FTPT_3 + \beta_{21} LANG_1 + \beta_{22} LANG_2 + \beta_{23} CATAGE_1 \\
& + \beta_{24} CATAGE_2 + \beta_{25} CATAGE_3 + \beta_{26} CATAGE_4 + \beta_{27} FOS_1 GNDN_1 \\
& + \beta_{28} FOS_1 GNDN_2 + \beta_{29} FOS_2 GNDN_1 + \beta_{30} FOS_2 GNDN_2 \\
& + \beta_{31} FOS_3 GNDN_1 + \beta_{32} FOS_3 GNDN_2 + \beta_{33} FOS_4 GNDN_1 \\
& + \beta_{34} FOS_4 GNDN_2 + \beta_{35} FOS_5 GNDN_1 + \beta_{36} FOS_5 GNDN_2 \\
& + \beta_{37} FOS_6 GNDN_1 + \beta_{38} FOS_6 GNDN_2 + \beta_{39} FOS_7 GNDN_1 \\
& + \beta_{40} FOS_7 GNDN_2 + \beta_{41} FOS_8 GNDN_1 + \beta_{42} FOS_8 GNDN_2 \\
& + \beta_{43} FOS_9 GNDN_1 + \beta_{44} FOS_9 GNDN_2 + \beta_{45} FOS_1 CATAGE_1 \\
& + \beta_{46} FOS_1 CATAGE_2 + \beta_{47} FOS_1 CATAGE_3 + \beta_{48} FOS_1 CATAGE_4 \\
& + \beta_{49} FOS_2 CATAGE_1 + \beta_{50} FOS_2 CATAGE_2 + \beta_{51} FOS_2 CATAGE_3 \\
& + \beta_{52} FOS_2 CATAGE_4 + \beta_{53} FOS_3 CATAGE_1 + \beta_{54} FOS_3 CATAGE_2 \\
& + \beta_{55} FOS_3 CATAGE_3 + \beta_{56} FOS_3 CATAGE_4 + \beta_{57} FOS_4 CATAGE_1 \\
& + \beta_{58} FOS_4 CATAGE_2 + \beta_{59} FOS_4 CATAGE_3 + \beta_{60} FOS_4 CATAGE_4 \\
& + \beta_{61} FOS_5 CATAGE_1 + \beta_{62} FOS_5 CATAGE_2 + \beta_{63} FOS_5 CATAGE_3 \\
& + \beta_{64} FOS_5 CATAGE_4 + \beta_{65} FOS_6 CATAGE_1 + \beta_{66} FOS_6 CATAGE_2 \\
& + \beta_{67} FOS_6 CATAGE_3 + \beta_{68} FOS_6 CATAGE_4 + \beta_{69} FOS_7 CATAGE_1 \\
& + \beta_{70} FOS_7 CATAGE_2 + \beta_{71} FOS_7 CATAGE_3 + \beta_{72} FOS_7 CATAGE_4 \\
& + \beta_{73} FOS_8 CATAGE_1 + \beta_{74} FOS_8 CATAGE_2 + \beta_{75} FOS_8 CATAGE_3 \\
& + \beta_{76} FOS_8 CATAGE_4 + \beta_{77} FOS_9 CATAGE_1 + \beta_{78} FOS_9 CATAGE_2 \\
& + \beta_{79} FOS_9 CATAGE_3 + \beta_{80} FOS_9 CATAGE_4 + \beta_{81} GNDN_1 CATAGE_1 \\
& + \beta_{82} GNDN_1 CATAGE_2 + \beta_{83} GNDN_1 CATAGE_3 \\
& + \beta_{84} GNDN_1 CATAGE_4 + \beta_{85} GNDN_2 CATAGE_1 \\
& + \beta_{86} GNDN_2 CATAGE_2 + \beta_{87} GNDN_2 CATAGE_3 \\
& + \beta_{88} GNDN_2 CATAGE_4
\end{aligned} \tag{4.1}$$

The choice of the six main effects (i.e. field of study, gender, region of residence, first language spoken and still understood, status of registration, and age) was made according to

the results of several papers on the subject of this analysis. First, Gonzalez (1996) has examined the effect of the sex and the field of study, among other things, on the graduation rate and the time needed to complete a doctoral degree. She stated that overall graduation rates for men were 5 percentage points higher than those for women. Also, she obtained as a result that the scientific fields showed consistent patterns with graduation rates between 70% and 80%, and times to completion of between four or five years. The non-scientific fields showed more variations which shows large variations among cohorts possibly because there were low numbers of students in each of them. So it made sense to keep the gender and the field of study as potential explanatory variables.

Then Butlin (1999) stated that women are more likely to participate in university and community college (controlling for other factors) and that a higher percentage of high school graduates born outside Canada (53%) attended university compared to Canadian born graduates (41%). So we know that gender and place of birth seemed to be two determinants of university education. Then it was interesting to see if the gender had also an effect on the university level of certification obtained at graduation. Unfortunately, the place of birth was not available through the NGS. A language variable was introduced as an attempt to distinguish graduates who spoke English or French (Canada's two official language) from those whose first language spoken is neither English nor French.

Finally, Bélanger and Omiecinski (1987) obtained results about the participation in part-time (or full-time) university studies by age, sex, and level of studies (undergraduate or graduate studies) from 1965 to 1985. As a result, they found that women made up the majority of part-time undergraduate students in all fields in 1985, except Business, Engineering, and Mathematics and the Physical Sciences. Also in this case it was interesting to see if the gender, age, and status of registration had an effect on the level of studies attained at graduation.

Note that the region of residence at the time of the NGS was also kept because comparisons

of the results by region was desired.

It has been impossible to include all of the two-way interactions (interaction between two variables) in the model in addition to the six main effects. When too many two-way interaction terms were included in the model, SUDAAN provided a warning in the output, saying that the matrix for estimable parameters is singular and the model may be overspecified. Then only the three more important two-way interaction terms have been included. The first one is the interaction between the field of study and gender. It was interesting to compare the men's probability of obtaining a bachelor's or a master's degree (or the equivalent) and the women's probability of obtaining a bachelor's or a master's degree (or the equivalent), in particular in the scientific fields of study. The second interaction retained was the interaction between the field of study and age. It was desired to see how does the probability of obtaining a bachelor's or a master's degree in a certain field of study differs for different categories of age. The last interaction included in the model is the one between gender and age. It was desired to see how does the probability of a male (or female) obtaining a bachelor's or a master's degree (or the equivalent) differs for different categories of age. Other interaction terms would have been interesting to be included in the model, but the model chosen with the six main effects and the three interaction terms was run and no warning message from SUDAAN about the singularity of the matrix for estimable parameters was provided. And this model is still simple, which is important for interpretation.

4.2.3 Conditional cumulative probabilities

Once the model (4.1) was fitted, it was possible to obtain the fitted conditional cumulative probabilities (or the fitted conditional probabilities). The fitted model becomes:

$$\begin{aligned}
\log \left(\frac{\widehat{F}_j(\mathbf{x})}{1 - \widehat{F}_j(\mathbf{x})} \right) = & \widehat{\alpha}_{j=1} + \widehat{\alpha}_{j=2} + \widehat{\beta}_1 FOS_1 + \widehat{\beta}_2 FOS_2 + \widehat{\beta}_3 FOS_3 \\
& + \widehat{\beta}_4 FOS_4 + \widehat{\beta}_5 FOS_5 + \widehat{\beta}_6 FOS_6 + \widehat{\beta}_7 FOS_7 + \widehat{\beta}_8 FOS_8 + \widehat{\beta}_9 FOS_9 \\
& + \widehat{\beta}_{10} GNDN_1 + \widehat{\beta}_{11} GNDN_2 + \widehat{\beta}_{12} REG_1 + \widehat{\beta}_{13} REG_2 + \widehat{\beta}_{14} REG_3 \\
& + \widehat{\beta}_{15} REG_4 + \widehat{\beta}_{16} REG_5 + \widehat{\beta}_{17} REG_6 + \widehat{\beta}_{18} FTPT_1 + \widehat{\beta}_{19} FTPT_2 \\
& + \widehat{\beta}_{20} FTPT_3 + \widehat{\beta}_{21} LANG_1 + \widehat{\beta}_{22} LANG_2 + \widehat{\beta}_{23} CATAGE_1 \\
& + \widehat{\beta}_{24} CATAGE_2 + \widehat{\beta}_{25} CATAGE_3 + \widehat{\beta}_{26} CATAGE_4 + \widehat{\beta}_{27} FOS_1 GNDN_1 \\
& + \widehat{\beta}_{28} FOS_1 GNDN_2 + \widehat{\beta}_{29} FOS_2 GNDN_1 + \widehat{\beta}_{30} FOS_2 GNDN_2 \\
& + \widehat{\beta}_{31} FOS_3 GNDN_1 + \widehat{\beta}_{32} FOS_3 GNDN_2 + \widehat{\beta}_{33} FOS_4 GNDN_1 \\
& + \widehat{\beta}_{34} FOS_4 GNDN_2 + \widehat{\beta}_{35} FOS_5 GNDN_1 + \widehat{\beta}_{36} FOS_5 GNDN_2 \\
& + \widehat{\beta}_{37} FOS_6 GNDN_1 + \widehat{\beta}_{38} FOS_6 GNDN_2 + \widehat{\beta}_{39} FOS_7 GNDN_1 \\
& + \widehat{\beta}_{40} FOS_7 GNDN_2 + \widehat{\beta}_{41} FOS_8 GNDN_1 + \widehat{\beta}_{42} FOS_8 GNDN_2 \\
& + \widehat{\beta}_{43} FOS_9 GNDN_1 + \widehat{\beta}_{44} FOS_9 GNDN_2 + \widehat{\beta}_{45} FOS_1 CATAGE_1 \\
& + \widehat{\beta}_{46} FOS_1 CATAGE_2 + \widehat{\beta}_{47} FOS_1 CATAGE_3 + \widehat{\beta}_{48} FOS_1 CATAGE_4 \\
& + \widehat{\beta}_{49} FOS_2 CATAGE_1 + \widehat{\beta}_{50} FOS_2 CATAGE_2 + \widehat{\beta}_{51} FOS_2 CATAGE_3 \\
& + \widehat{\beta}_{52} FOS_2 CATAGE_4 + \widehat{\beta}_{53} FOS_3 CATAGE_1 + \widehat{\beta}_{54} FOS_3 CATAGE_2 \\
& + \widehat{\beta}_{55} FOS_3 CATAGE_3 + \widehat{\beta}_{56} FOS_3 CATAGE_4 + \widehat{\beta}_{57} FOS_4 CATAGE_1 \\
& + \widehat{\beta}_{58} FOS_4 CATAGE_2 + \widehat{\beta}_{59} FOS_4 CATAGE_3 + \widehat{\beta}_{60} FOS_4 CATAGE_4 \\
& + \widehat{\beta}_{61} FOS_5 CATAGE_1 + \widehat{\beta}_{62} FOS_5 CATAGE_2 + \widehat{\beta}_{63} FOS_5 CATAGE_3 \\
& + \widehat{\beta}_{64} FOS_5 CATAGE_4 + \widehat{\beta}_{65} FOS_6 CATAGE_1 + \widehat{\beta}_{66} FOS_6 CATAGE_2 \\
& + \widehat{\beta}_{67} FOS_6 CATAGE_3 + \widehat{\beta}_{68} FOS_6 CATAGE_4 + \widehat{\beta}_{69} FOS_7 CATAGE_1 \\
& + \widehat{\beta}_{70} FOS_7 CATAGE_2 + \widehat{\beta}_{71} FOS_7 CATAGE_3 + \widehat{\beta}_{72} FOS_7 CATAGE_4 \\
& + \widehat{\beta}_{73} FOS_8 CATAGE_1 + \widehat{\beta}_{74} FOS_8 CATAGE_2 + \widehat{\beta}_{75} FOS_8 CATAGE_3 \\
& + \widehat{\beta}_{76} FOS_8 CATAGE_4 + \widehat{\beta}_{77} FOS_9 CATAGE_1 + \widehat{\beta}_{78} FOS_9 CATAGE_2 \\
& + \widehat{\beta}_{79} FOS_9 CATAGE_3 + \widehat{\beta}_{80} FOS_9 CATAGE_4 + \widehat{\beta}_{81} GNDN_1 CATAGE_1 \\
& + \widehat{\beta}_{82} GNDN_1 CATAGE_2 + \widehat{\beta}_{83} GNDN_1 CATAGE_3 \\
& + \widehat{\beta}_{84} GNDN_1 CATAGE_4 + \widehat{\beta}_{85} GNDN_2 CATAGE_1 \\
& + \widehat{\beta}_{86} GNDN_2 CATAGE_2 + \widehat{\beta}_{87} GNDN_2 CATAGE_3 \\
& + \widehat{\beta}_{88} GNDN_2 CATAGE_4
\end{aligned} \tag{4.2}$$

From the fitted model, if we want to obtain the probability that $Y \leq 2$ ($Y \leq 2$ means that the level of certification at graduation is the bachelor's degree, first professional degree, university diploma below bachelor, master's degree, or university diploma above bachelor) given that:

- a) The field of study is Education (FOS=2)
- b) Gender is male (GNDN=1)
- c) The region of residence at the time of the interview is Ontario (REG=3)
- d) The registration status is full-time only (FTPT=1)
- e) The first language learned and still understood in English or French (LANG=2)
- f) The age is between 25 and 29 (CATAGE=2)

then the terms FOS_2 , $GNDN_1$, REG_3 , $FTPT_1$, $LANG_2$, $CATAGE_2$, FOS_2GNDN_1 , $FOS_2CATAGE_2$, and $GNDN_1CATAGE_2$ are all equal to 1 and all of the other terms ($FOS_{\neq 2}$, $GNDN_{\neq 1}$, $REG_{\neq 3}$, $FTPT_{\neq 1}$, $LANG_{\neq 2}$, $CATAGE_{\neq 2}$, $FOS_{\neq 2}GNDN_{\neq 1}$, $FOS_{\neq 2}CATAGE_{\neq 2}$, and $GNDN_{\neq 1}CATAGE_{\neq 2}$) of the model are equal to 0. So to obtain the logit of the fitted cumulative conditional probabilities, we have the equation:

$$\begin{aligned} \log\left(\frac{\hat{F}_2(\mathbf{x})}{1 - \hat{F}_2(\mathbf{x})}\right) &= \hat{\alpha}_{j=2} + \hat{\beta}_2 FOS_2 + \hat{\beta}_{10} GNDN_1 + \hat{\beta}_{14} REG_3 + \hat{\beta}_{18} FTPT_1 + \hat{\beta}_{22} LANG_2 \\ &\quad + \hat{\beta}_{24} CATAGE_2 + \hat{\beta}_{29} FOS_2 GNDN_1 + \hat{\beta}_{50} FOS_2 CATAGE_2 \\ &\quad + \hat{\beta}_{82} GNDN_1 CATAGE_2 \\ &= \hat{\alpha}_{j=2} + \hat{\beta}_2 + \hat{\beta}_{10} + \hat{\beta}_{14} + \hat{\beta}_{18} + \hat{\beta}_{22} + \hat{\beta}_{24} + \hat{\beta}_{29} + \hat{\beta}_{50} + \hat{\beta}_{82} \end{aligned}$$

So

$$\left(\frac{\hat{F}_2(\mathbf{x})}{1 - \hat{F}_2(\mathbf{x})}\right) = \exp(\hat{\alpha}_{j=2} + \hat{\beta}_2 + \hat{\beta}_{10} + \hat{\beta}_{14} + \hat{\beta}_{18} + \hat{\beta}_{22} + \hat{\beta}_{24} + \hat{\beta}_{29} + \hat{\beta}_{50} + \hat{\beta}_{82})$$

Then

$$\widehat{F_2(\mathbf{x})} = \left(\exp(\widehat{\alpha}_{j=2} + \widehat{\beta}_2 + \widehat{\beta}_{10} + \widehat{\beta}_{14} + \widehat{\beta}_{18} + \widehat{\beta}_{22} + \widehat{\beta}_{24} + \widehat{\beta}_{29} + \widehat{\beta}_{50} + \widehat{\beta}_{82}) \right) * \left(1 - \widehat{F_2(\mathbf{x})} \right)$$

$$\Rightarrow \widehat{F_2(\mathbf{x})} = \frac{\exp(\widehat{\alpha}_{j=2} + \widehat{\beta}_2 + \widehat{\beta}_{10} + \widehat{\beta}_{14} + \widehat{\beta}_{18} + \widehat{\beta}_{22} + \widehat{\beta}_{24} + \widehat{\beta}_{29} + \widehat{\beta}_{50} + \widehat{\beta}_{82})}{1 + \exp(\widehat{\alpha}_{j=2} + \widehat{\beta}_2 + \widehat{\beta}_{10} + \widehat{\beta}_{14} + \widehat{\beta}_{18} + \widehat{\beta}_{22} + \widehat{\beta}_{24} + \widehat{\beta}_{29} + \widehat{\beta}_{50} + \widehat{\beta}_{82})}$$

This is the estimated probability that the response takes a value less than or equal to 2 (i.e. that the level of certification is the bachelor's degree, first professional degree, university diploma below bachelor, master's degree, or university diploma above bachelor) when the graduate is male, he studies in Education, he took his program full-time only, his first language learned and still understood is English or French, his region of residence at the time of the interview is Ontario, and his age is between 25 and 29.

Now the conditional probability that Y takes the value 1 is equal to $F_1(\mathbf{x})$; the conditional probability that Y takes the value 2 is equal to $F_2(\mathbf{x}) - F_1(\mathbf{x})$; and the conditional probability that Y takes the value 3 is equal to $1.0 - F_2(\mathbf{x})$. The "PREDICTED" option of the "OUTPUT" statement offers the possibility of getting these fitted conditional probabilities in a file.

4.3 Validation of the model

The first goodness-of-fit statistic that was computed was the Pearson chi-squared statistic, for the four models. In the following formula

$$X^2 = n \sum_i \widehat{W}_i \sum_j (\widetilde{\pi}_{ij} - \widehat{\pi}_{ij})^2 / \widehat{\pi}_{ij},$$

the $\widehat{\pi}_{ij}$ were obtained as an output from SUDAAN; the $\widetilde{\pi}_{ij}$ were computed from the subpopulation described in section 4.2.1; in the term $\widehat{W}_i$, m_i was the weighted number of observations for explanatory variable i ; and n was the weighted number of observations in the subpopulation. Recall that i identifies a *combination* of the six explanatory variables of the model (field of study, gender, region of residence, first language spoken and still

understood, categories of age, and status of registration). For example, $i = 1$ could identify the combination where field of study takes the value 1; the gender takes the value 1; the region of residence takes the value 2; the first language spoken and still understood takes the value 1; the categories of age takes the value 3; and the status of registration takes the value 2. Section 4.2.2 details the levels of these variables with their corresponding level number.

The chi-squared statistics for each of the surveys are:

Table 4.1: Pearson chi-squared statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
16,548.58	16,548.58	18,289.13	18,289.13

These four chi-squared statistics take large values. If we compare the 1988 statistic to $\chi^2_{2684-90(0.95)} = 2,713.33$, when 90 is the number of parameter estimates in the model and 2,684 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,342$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1992 statistic to $\chi^2_{2782-90(0.95)} = 2,813.54$, when 90 is the number of parameter estimates in the model and 2,782 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,391$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Finally, if we compare the 1997 statistics to $\chi^2_{2712-90(0.95)} = 2,741.97$, when 90 is the number of parameter estimates in the model and 2,712 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,356$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Recall that this is not a test corrected for the sampling

plan and note that when $\widehat{\pi}_{ij} = 0$, the calculation of the contribution to the Pearson's statistic could not be performed, since there was a zero in the denominator. Large values of the contributions to the Pearson's chi-squared statistic occur often when $\widehat{W}_i$ was very large and when $\widehat{\pi}_{ij}$ was very small.

Then four deviance functions have also been computed. Here the y_{ij} are the weighted number of observations having response j and combination of explanatory variables i .

The deviance statistics for each of the surveys are:

Table 4.2: Deviance statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
13,727.09	12,919.99	17,663.52	17,663.52

These four deviance functions take large values. If we compare the 1988 statistic to $\chi^2_{2684-90(0.95)} = 2,713.33$, when 90 is the number of parameter estimates in the model and 2,684 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,342$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1992 statistic to $\chi^2_{2782-90(0.95)} = 2,813.54$, when 90 is the number of parameter estimates in the model and 2,782 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,391$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Finally, if we compare the 1997 statistics to $\chi^2_{2712-90(0.95)} = 2,741.97$, when 90 is the number of parameter estimates in the model and 2,712 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 1,356$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is

correct, at the significance level 5%. When $\widehat{\pi}_{ij} = 0$ or $\widetilde{\pi}_{ij} = 0$, the calculation of the contribution to the deviance function could not be performed, since $\log(0)$ does not exist. Large values of the contributions to the deviance function occur often when $\widetilde{\pi}_{ij}$ is greater than $\widehat{\pi}_{ij}$.

So according to these two statistics, the proportional odds model fitted is not correct. But since there is sparseness in the data and these tests have not been corrected to take account of the sample design, we cannot strongly rely on these statistics as the only validations of this model.

SUDAAN's output provided an approximate chi-square statistic with its number of degrees of freedom. If the clustering or the stratification effects are negligible, then the difference between the likelihoods for the overall model and the model with all of betas equal to zero is an approximate chi-square. The primary use of this number is to compare several models used to predict the same variable with the same data but with different set of independent variables. However, if the clustering or stratification effects are significant, this is not a reliable test. A better test is the test of significance that all betas are zero. Since the design effects for the parameter estimates are almost all lower than one, we can say that the stratification had a significant effect. So the Wald chi-square has been used to test the null hypothesis that all of the parameter estimates in the model are equal to zero. This will test simultaneously that every of the 90 parameters are equal to zero. For all of the four models, this null hypothesis is strongly rejected, indicating that the proportional odds model fits better than a model with all coefficients equal to zero. See results below:

Table 4.3: Wald chi-square statistics

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Wald statistic	2,150,317.71	322,723.83	910,333.79	818,293.67
<i>p</i> -value	0.0000	0.0000	0.0000	0.0000

The cell residuals, defined by $(y_{ij} - \hat{y}_{ij}) / [m_i \hat{\pi}_{ij} (1 - \hat{\pi}_{ij})]^{1/2}$ have been computed. Note that in this formula, $\hat{y}_{ij} = m_i \hat{\pi}_{ij}$. Once they have been computed, four groups have been created and the number of cells included in each of these groups was calculated.

Table 4.4: Frequencies of cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	2,330(58%)	2,553(61%)	2,360(58%)	2,360(58%)
$-10 \leq \text{cell res.} < -2$ or	896	954	955	955
$2 < \text{cell res.} \leq 10$	—	—	—	—
cell res. < -10 or	34	20	48	48
cell res. > 10	—	—	—	—
cell res. is missing	766	646	705	705

Approximately 60% of the cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not be computed (either when $\hat{\pi}_{ij} = 1$ or $\hat{\pi}_{ij} = 0$). If we delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 4.5: Proportions of cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
$-2 \leq \text{cell res.} \leq 2$	71.5%	67.9%	63.4%	63.4%

These proportions seem reasonable. There does not appear to be a cell residual with an abnormally large value (see table below). If we compare the maximum and minimum values of the four models, we see that they are roughly in the same range of values

Table 4.6: Maximum and minimum cell residuals

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Max. cell residual	30.56	18.54	26.44	26.44
Min. cell residual	-17.11	-15.82	-23.95	-23.95

The cumulative cell residuals have also been computed, using cumulative totals and cumulative fitted probabilities. Once they have been computed, five groups have been created and the number of cells included in each of these groups was calculated.

Table 4.7: Frequencies of cumulative cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin)
$-2 \leq \text{cum. cell res.} \leq 2$	1,513(38%)	1,678(40%)	1,533(38%)	1,533(38%)
$-10 \leq \text{cum. cell res.} < -2$	458	531	558	558
or $2 < \text{cum. cell res.} \leq 10$	—	—	—	—
$-100 \leq \text{cum. cell res.} < -10$	8	3	21	21
or $10 < \text{cum. cell res.} \leq 100$	—	—	—	—
cum. cell res. < -100 or	284	261	306	306
cum. cell res. > 100	—	—	—	—
cum. cell res. is missing	1,763	1,700	1,650	1,650

Approximately 50% of the cumulative cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not been computed (either when the cumulative probability was equal to zero or one). If we

delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 4.8: Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	66.9%	67.9%	63.4%	63.4%

These proportions are about 25% to 30% higher than when the missing cumulative cell residuals were included in the counts. These counts are comparable to those obtained for the cell residuals. However, some really large cumulative cell residuals occurred (larger than 10^6). This occurred when the cumulative probability has a value near one and that m_i was close to zero. If there would not have been sparseness in the data, the m_i s would have had a value greater than one, at least.

A plot of each of the contributions to the deviance function against its index (i, j) was obtained in order to detect large deviance contributions (see Appendix G). For each model, a plot of the deviance contributions corresponding to response level $j, j = 1, 2, 3$, for all of the I combinations of explanatory variables has been drawn. On Figure 4.1, for $Y = 1$ (i.e. the level of certification at graduation is a bachelor's degree, a first professional degree, or an university diploma below bachelor), in 1988, there were two deviance contributions that seemed to be large. One has the deviance contribution equal to 300 (it represents the female graduates aged 35 years old and over living in Quebec who studied in Education part-time only and whose first language spoken is English or French) and the other as a deviance contribution equal to -300 (it represents the female graduates aged 35 years old and over living in Ontario who studied in Education part-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -100 and 200. On Figure 4.2, for $Y = 2$ (i.e. the level of certification at graduation is a master's

degree or an university diploma above bachelor), in 1988, there was one deviance contribution that seemed to be large. This deviance contribution equal to 500 approximately (it represents the female graduates aged 35 years old and over living in Ontario who studied in Education part-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -100 and 200 . On Figure 4.3, for $Y = 3$ (i.e. the level of certification at graduation is an earned Ph.D.), in 1988, there was one deviance contribution that seemed to be large. This deviance contribution was approximately equal to 100 (it represents the female graduates aged 35 years old and over living in Ontario who studied in Other Social sciences full-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -10 and 70 . So for the 1988 model, only three deviance contributions (for $Y = 1$ and $Y = 2$) seem to be large.

On Figure 4.4, for $Y = 1$, in 1992, there does not seem to be any large deviance contribution. The deviance contributions on the graph seem to be between -200 and 200 . On Figure 4.5, for $Y = 2$, in 1992, there was one deviance contribution that seemed to be large. This deviance contribution was approximately equal to 300 (it represents the male graduates aged 25 to 29 years old living in Ontario who studied in Commerce, law, and economics full-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -100 and 200 . On Figure 4.6, for $Y = 3$, in 1992, there was one deviance contribution that seemed to be large. This deviance contribution was approximately equal to 90 (it represents the male graduates aged 30 to 34 years old living in Ontario who studied in Engineering and applied sciences full-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -20 and 70 . So for the 1992 model, only one deviance contribution (for $Y = 2$) seems to be large.

On Figure 4.7, for $Y = 1$, in 1997, there were three deviance contributions that seemed to be

large. The first two have a deviance contribution equal to approximately 500 (they represent the male graduates aged 35 years old and over living in Ontario who studied in Education full-time only and whose first language spoken is English or French and also the female graduates aged 35 years old and over living in Ontario who studied in Education full-time only and whose first language spoken is English or French) and the other has a value of -300 (it represents the female graduates aged 35 years old and over living in Ontario who studied in Education part-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be roughly between -200 and 200. On Figure 4.8, for $Y = 2$, in 1997, there were five deviance contributions that seemed to be large. One has a value of approximately 900 (it represents the female graduates aged 35 years old and over living in the Atlantic provinces who studied in Education part-time only and whose first language spoken is English or French) and the four others have a value of 300 to 400 (they represent the female graduates aged 35 years old and over living in Ontario who studied in Education part-time only and whose first language spoken is English or French; and the female graduates aged 25 to 29 years old living in Quebec who studied in Fine arts and Humanities full-time only and whose first language spoken is English or French; and the female graduates aged 25 to 29 years old living in Quebec who studied in Other Social sciences full-time only and whose first language spoken is English or French; and the female graduates aged 25 to 29 years old living in Quebec who studied in Medical and health professions full-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -200 and 300. On Figure 4.9, for $Y = 3$, in 1997, there were three deviance contribution that seemed to be large. They have a deviance contribution between 100 and 200 (they represent the female graduates aged 35 years old and over living in Ontario who studied in Education full-time and part-time and whose first language spoken is English or French; and the male graduates aged 35 years old and over living in Ontario who studied in Fine arts and Humanities full-time only and whose

first language spoken is English or French; and the male graduates aged 30 to 34 years old living in Ontario who studied in Mathematics and physical sciences full-time only and whose first language spoken is English or French). The other deviance contributions on the graph seem to be between -30 and 100 . So for the 1997 model, eight deviance contributions (for $Y = 1$ and $Y = 2$) seem to be large. The same results applied for the 1997 model using administrative counts (on figures 4.10, 4.11, and 4.12).

So it seems that the 1997 models have a higher number of deviance contributions that are large than the 1988 model and the 1992 model. We also saw previously that the deviance statistics for the 1997 models were higher than the deviance statistics for the 1988 model and the 1992 model. This can be explained by the highest number of large contributions to the deviance in these two models.

The comparison between the robust variance-covariance matrix and the model-based variance-covariance matrix is a way of assessing that the true correlation structure is indeed the one defined by the statement "R" in PROC MULTLOG. For the model fitted, we chose the independence working correlation structure. If there is little difference between a pair of covariance term (on the upper diagonal of the robust variance-covariance matrix and the same term but on the upper diagonal of the model-based variance-covariance matrix), for all pairs of terms, this will indicate that the choice of the independent correlation structure was correct. The following table provides a grouping of the absolute differences in ten categories.

Table 4.9: Absolute differences, by group, of the robust and the model-based covariance matrices

Percentages	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$0 \leq \text{abs. diff.} \leq 0.01$	80.5	79.3	79.2	78.1
$0.01 < \text{abs. diff.} \leq 0.02$	11.6	2.6	14.0	14.1
$0.02 < \text{abs. diff.} \leq 0.03$	4.0	0.2	3.0	3.9
$0.03 < \text{abs. diff.} \leq 0.04$	1.1	1.5	1.2	1.1
$0.04 < \text{abs. diff.} \leq 0.05$	0.9	12.5	0.4	0.5
$0.05 < \text{abs. diff.} \leq 0.1$	1.7	3.2	0.4	0.5
$0.1 < \text{abs. diff.} \leq 0.2$	0.2	0.7	1.4	0.8
$0.2 < \text{abs. diff.} \leq 0.3$	0.0	0.0	0.3	0.8
$0.3 < \text{abs. diff.} \leq 0.4$	0.0	0.0	0.1	0.1
$\text{abs. diff.} > 0.4$	0.0	0.0	0.1	0.1

Approximately 80% of the absolute differences are between 0 and 0.01. From 0.2% to 1.9% of the absolute differences are above 0.1. This seems to indicate that a great proportion of the variance-covariance values on the upper triangle of the robust matrix have a value near the value of the reciprocal term on the upper triangle of the model-based matrix. Thus we can conclude that the correlation structure was correctly specified.

In summary, the sparseness in the data caused some problems when it came time to assess the validity of it (see section 4.5). But we know that the parameters in the model are significant; that the correlation structure can be assumed to be true; and that if we exclude the cell residuals affected by sparseness, there does not seem to be too many large cell residuals

4.4 Results

In the output, SUDAAN provides the estimated parameters, the standard error of each of these parameter estimates, the design effects, the t -test values (to test the hypothesis that the parameter estimate is equal to 0) and p -values of these tests. The complete list of the

parameter estimates, the standard errors, the design effects, the t -test values, and the p -values is in Appendix D for the 1988 model, the 1992 model, the 1997 model using weighted counts for the population counts, and the 1997 model using the administrative counts for the population counts.

4.4.1 Design effects

When the design effect is less than 1, we know that the sample design chosen was more appropriate than simple random sampling. Ignoring the sample design in this case would have increased the variance of the parameter estimate. The design effects of the 1988 and the 1992 models were all less than one. But for the two 1997 models, there were some design effects greater than one. The greatest design effect was 1.65 for the 1997 model using weighted counts and 1.67 for the 1997 model using the administrative counts, which are not too large design effects. The interpretation of a design effect of 1.65 (or 1.67) is that the variance of the parameter estimate is 1.65 times (or 1.67 times) higher when the stratified systematic random sampling is assumed than when the simple random sampling is assumed. But since only two parameter estimates have a design effect greater than one, we can conclude, in general, that the sample design chosen (stratified systematic random sampling) improved the precision of the parameter estimates.

4.4.2 Comparison between the two 1997 models

The model for the 1997 NGS data has been fitted twice: once using the population counts obtained from the weights and once using the administrative population counts. This has been done to verify if there was a difference in the precision of the estimates in using one or the other type of population counts. If there was a difference, it was interesting to know how different the standard errors from the parameter estimates from the weighted counts' model were from the administrative counts' model.

The standard errors of the parameter estimates from the model using the administrative counts were always greater or equal than the standard errors of the parameter estimates from the model using the weighted counts. So we can say that the use of the weighted counts gave a better precision of the estimates. A test was performed in order to verify if any of the estimated variances that differ from one model to another were statistically different. This F-test is highly sensitive to departures from the assumption of normality. A similar test from Levene (less sensitive) could have been performed. The test statistic for the following hypothesis

$$H_0 : \sigma_{1i}^2 = \sigma_{2i}^2$$

$$H_0 : \sigma_{1i}^2 \neq \sigma_{2i}^2$$

(when σ_{1i}^2 is the variance of the i^{th} parameter estimate from the model with weighted counts and σ_{2i}^2 is the variance of the i^{th} parameter estimate from the model with administrative counts, $i = 1, \dots, 34$ when 34 is the number of unequal standard errors) is the ratio of the sample variances

$$F_0 = \frac{S_{1i}^2}{S_{2i}^2}.$$

The null hypothesis is rejected if $F_0 > F_{0.025,33,33}$ or if $F_0 < F_{0.975,33,33}$. It turned out that the null hypothesis has been accepted for all of the 34 unequal standard errors. In conclusion, it seems that the use of weights to obtain population counts by stratum did not provide a significant improvement in the standard errors of the parameter estimates than the use of the original administrative population counts. But we know that when two standard errors of a given parameter estimate are unequal, it is always the one from the model using administrative counts that is the greatest.

4.4.3 Testing of the parameter estimates

At the 5% significance level, each p -value over 0.05 will mean that the corresponding

parameter is not significantly different from zero and has no effect on the estimated cumulative logit, $L_j(\mathbf{x})$. The parameter estimate of this term can therefore be considered as being zero and thus not included in the final model retained.

Table 4.10: Parameter estimates determined to be equal to zero by the *t*-tests

Levels of the parameter...	Model 1988	Model 1992	Model 1997
INTERCEPT	-	1	-
FIELD OF STUDY	4 and 8	2 and 5	2, 3, 5, 7, and 8
GENDER	-	-	-
REGION OF RESIDENCE	1 to 5	1, 2, 3, and 5	2, 4, and 5
STATUS OF REGISTRATION	2	2	2
FIRST LANGUAGE SPOKEN	-	-	-
AGE	-	-	3
FIELD OF STUDY BY GENDER	1 by 1	2 by 1	1 by 1
	4 by 1	3 by 1	3 by 1
	5 by 1	4 by 1	4 by 1
	7 by 1	5 by 1	5 by 1
	-	7 by 1	7 by 1
	-	8 by 1	-
FIELD OF STUDY BY AGE	3 by 1	1 by 2	1 by 2
	3 by 2	1 by 3	1 by 3
	3 by 3	3 by 1	3 by 1
	4 by 2	3 by 2	3 by 2
	4 by 3	4 by 1	3 by 3
	5 by 1	4 by 2	4 by 1
	5 by 2	4 by 3	4 by 2
	5 by 3	5 by 1	4 by 3
	8 by 1	5 by 2	5 by 1
	8 by 2	5 by 3	5 by 2
	8 by 3	8 by 2	5 by 3
	-	8 by 3	6 by 3
	-	-	7 by 3
	-	-	8 by 1
	-	-	8 by 2
	-	-	8 by 3
GENDER BY AGE	1 by 1	1 by 1	-

The preceding table gives the levels of the parameters in the four models that are determined to be equal to zero.

4.4.4 Testing of the main effects

The Wald chi-square statistic was used to test different hypotheses about the main effects. For example, for the variable FTPT, the Wald chi-square statistic can be used to test if the parameter estimates at the three levels of this variable are equal to zero simultaneously. The null hypothesis is then:

$$H_0 : \begin{pmatrix} \beta_{18} \\ \beta_{19} \\ \beta_{20} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This statistic can also be used to test different hypotheses about the interaction terms of the model. For example, the Wald chi-square statistic can be used to test if the parameter estimates of the interaction terms defined by the combination of each level of GNDN and each level of CATAGE are equal to zero simultaneously. The null hypothesis was then:

$$H_{0:} \begin{pmatrix} \beta_{81} \\ \beta_{82} \\ \beta_{83} \\ \beta_{84} \\ \beta_{85} \\ \beta_{86} \\ \beta_{87} \\ \beta_{88} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

When the degrees of freedom associated with the test is large compared to the degrees of freedom available for estimating the variance of regression coefficients (number of observations minus the number of strata), the Wald chi-square is too liberal (Thomas and

Rao, 1987). It does not happen here since the number of degrees of freedom available for estimating the variance of regression coefficients is greater than 30,000 and the highest number of degrees of freedom for a test is 57. This is true for the four models. All of the six main effects and the three two-way interaction terms of the model have been tested. No three-way interaction could be included in the model because it caused problems with the singularity of the matrix of estimable parameters. At the 5% level of significance, none of the six main effects were statistically equal to 0. At the same significance level, none of the three interaction terms were statistically equal to 0. Consequently, the effect of the gender must be assessed taking account of the field of study and the age; the effect of the field of study must be assessed taking account of the gender and the age; the effect of the age must be assessed taking account of the field of study and the gender. See chapter 6 for further interpretation of these main effects and interactions.

4.4.5 Results of odds ratio analyses

As described in Chapter 3, the odds ratio of cumulative probabilities in the expression below is called a cumulative odds ratio and its interpretation is that the odds of making response $\leq j$ are $\exp[\beta'(\mathbf{x}_1 - \mathbf{x}_2)]$ times higher at $\mathbf{x} = \mathbf{x}_1$ than $\mathbf{x} = \mathbf{x}_2$.

$$L_j(\mathbf{x}_1) - L_j(\mathbf{x}_2) = \log \left[\frac{P(Y \leq j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1))}{P(Y \leq j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} \right] = \beta'(\mathbf{x}_1 - \mathbf{x}_2).$$

We defined $L_j(\mathbf{x}_1)$ as being $\alpha_j + \beta' \mathbf{x}_1$, when $\mathbf{x}_1$ is a vector having a particular combination of one level of each explanatory variable, for $j = 1, 2$. The procedure to follow in order to obtain odds ratios for the explanatory variables of this model is detailed below.

For the explanation, we will consider GNDN, FTPT, LANG, CATAGE and FOS as having a fixed level. In $\mathbf{x}_1$, REG takes value 2 and in $\mathbf{x}_2$, REG takes value 4. In both explanatory vectors, we find the same combination of the fixed levels of GNDN, FTPT, LANG,

CATAGE and FOS. The odds of making response $\leq j$ at REG = 2 (Quebec) compared to REG = 4 (the Prairies) will be found. We have then:

$$\begin{aligned} L_j(\mathbf{x}_1) &= L_j(\text{REG} = 2 \text{ and other variables fixed}) \\ &= \hat{\alpha}_j + \hat{\beta}_{a'}FOS_a + \hat{\beta}_{b'}GNDN_b + \hat{\beta}_{13}REG_2 + \hat{\beta}_{c'}FTPT_c + \hat{\beta}_{d'}LANG_d \\ &\quad + \hat{\beta}_{e'}CATAGE_e + \hat{\beta}_{a'b'}FOS_aGNDN_b + \hat{\beta}_{a'e'}FOS_aCATAGE_e \\ &\quad + \hat{\beta}_{b'e'}GNDN_bCATAGE_e, \end{aligned}$$

where $a = 1, \dots, 9$; $b = 1, 2$; $c = 1, \dots, 3$; $d = 1, 2$; $e = 1, \dots, 4$ and a' corresponds to the parameter number that goes with the level a of FOS; b' corresponds to the parameter number that goes with the level b of GNDN; c' corresponds to the parameter number that goes with the level c of FTPT; d' corresponds to the parameter number that goes with the level d of LANG; e' corresponds to the parameter number that goes with the level e of CATAGE; $a'b'$ corresponds to the parameter number that goes with the interaction between FOS = a and GNDN = b ; $a'e'$ corresponds to the parameter number that goes with the interaction between FOS = a and CATAGE = e ; $b'e'$ corresponds to the parameter number that goes with the interaction between GNDN = b and CATAGE = e .

We obtain $L_j(\mathbf{x}_1) = L_j(\text{REG} = 4 \text{ and other variables fixed})$ the same way. Thus we obtain:

$$\begin{aligned} L_j(\mathbf{x}_1) - L_j(\mathbf{x}_2) &= (\hat{\alpha}_j + \hat{\beta}_{a'}FOS_a + \hat{\beta}_{b'}GNDN_b + \hat{\beta}_{13}REG_2 + \hat{\beta}_{c'}FTPT_c + \hat{\beta}_{d'}LANG_d \\ &\quad + \hat{\beta}_{e'}CATAGE_e + \hat{\beta}_{a'b'}FOS_aGNDN_b + \hat{\beta}_{a'e'}FOS_aCATAGE_e \\ &\quad + \hat{\beta}_{b'e'}GNDN_bCATAGE_e) \\ &\quad - (\hat{\alpha}_j + \hat{\beta}_{a'}FOS_a + \hat{\beta}_{b'}GNDN_b + \hat{\beta}_{15}REG_4 + \hat{\beta}_{c'}FTPT_c + \hat{\beta}_{d'}LANG_d \\ &\quad + \hat{\beta}_{e'}CATAGE_e \\ &\quad + \hat{\beta}_{a'b'}FOS_aGNDN_b + \hat{\beta}_{a'e'}FOS_aCATAGE_e + \hat{\beta}_{b'e'}GNDN_bCATAGE_e) \\ &= \hat{\beta}_{13}REG_2 - \hat{\beta}_{15}REG_4 \\ &= \hat{\beta}_{13} - \hat{\beta}_{15} \end{aligned}$$

So, for example, the odds of obtaining response $\leq j$ are $\exp(\hat{\beta}_{13} - \hat{\beta}_{15})$ times higher when the region of residence is Quebec than when the region of residence is the Prairies (for other variables fixed).

We can now obtain the odds of having response $> j$ by noting that

$$L_j(\mathbf{x}_i) = \log \left[\frac{P(Y \leq j | \mathbf{x}_i)}{P(Y > j | \mathbf{x}_i)} \right] = \log \left[\frac{1 - P(Y > j | \mathbf{x}_i)}{P(Y > j | \mathbf{x}_i)} \right] = \alpha_j + \beta' \mathbf{x}_i .$$

Therefore,

$$L_j(\mathbf{x}_1) - L_j(\mathbf{x}_2) = \log \left[\frac{1 - P(Y > j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{1 - P(Y > j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} \right] = \beta'(\mathbf{x}_1 - \mathbf{x}_2)$$

and

$$\left[\frac{1 - P(Y > j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{1 - P(Y > j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} \right] = \exp(\beta'(\mathbf{x}_1 - \mathbf{x}_2))$$

which implies that

$$\begin{aligned} \left[\frac{1 - P(Y > j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{1 - P(Y > j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} \right] &= \left[\frac{(1 - P(Y > j | \mathbf{x}_1)) \times (P(Y > j | \mathbf{x}_2))}{(1 - P(Y > j | \mathbf{x}_2)) \times (P(Y > j | \mathbf{x}_1))} \right] \\ &= \left[\frac{P(Y > j | \mathbf{x}_2) / 1 - P(Y > j | \mathbf{x}_2)}{P(Y > j | \mathbf{x}_1) / 1 - P(Y > j | \mathbf{x}_1)} \right] \\ &= \exp(\beta'(\mathbf{x}_1 - \mathbf{x}_2)) \end{aligned}$$

and consequently

$$\left[\frac{P(Y > j | \mathbf{x}_1) / 1 - P(Y > j | \mathbf{x}_1)}{P(Y > j | \mathbf{x}_2) / 1 - P(Y > j | \mathbf{x}_2)} \right] = (\exp(\beta'(\mathbf{x}_1 - \mathbf{x}_2)))^{-1}.$$

So the odds of obtaining response $> j$ are $(\exp(\beta'(\mathbf{x}_1 - \mathbf{x}_2)))^{-1}$ times higher at $\mathbf{x} = \mathbf{x}_1$ than $\mathbf{x} = \mathbf{x}_2$ ($j = 1, 2$).

Some odds ratios that have been estimated are:

1) Full-time/part-time status: estimated odds ratio of obtaining response $> j$ for ...

Table 4.11: Estimated odds of having response $> j$ ($j = 1, 2$) (status of registration)

Odds ratios	1988	1992	1997
Full-time over combination of full-time/part-time	1.51	1.16	1.36
Part-time over combination of full-time/part-time	0.90	1.00	0.86
Full-time over part-time	1.68	1.16	1.58

for a given combination of field of study, gender, region of residence, first spoken language and still understood, and category of age. So for all of the three years, graduates who studied full-time only have higher odds of having a master's degree (in this section, when we mention master's degree, we mean master's degree and university diploma above bachelor) or a doctoral degree compared to the graduates who studied full-time and part-time. This seems to indicate that those who studied full-time only were more likely to pursue higher levels of education than those who studied full-time and part-time.

In 1988 and 1997, graduates who studied part-time only have reduced odds of having a master's degree or a doctoral degree compared to the graduates who studied full-time and part-time. This seems to indicate that graduates who studied part-time only were less likely to pursue higher levels of education than those who studied full-time and part-time. In 1992, the odds of having a master's degree or a doctoral degree are equal for graduates who studied part-time only and those who studied full-time and part-time.

Finally, for all of the three years, graduates who studied full-time only have higher odds of having a master's degree or a doctoral degree compared to graduates who studied part-time only. This seems to indicate that graduates who studied full-time only were more likely to pursue higher levels of education than graduates who studied part-time only.

2) Region of residence at the time of the interview: estimated odds ratio of obtaining response $> j$ for ...

Table 4.12: Estimated odds of having response $> j$ ($j = 1, 2$) (region of residence)

Odds ratios	1988	1992	1997
Ontario over the Prairies	1.36	1.32	1.12
Ontario over Quebec	1.90	0.84	1.27
Ontario over British Columbia	1.27	1.0	1.11

for a given combination of field of study, full-time/part-time status, gender, first spoken language and still understood, and category of age. So for all of the three years, the graduates who lived in Ontario at the time of the interview have higher odds of having obtained a master's degree or a doctoral degree at graduation compared to the graduates who lived in the Prairies. This seems to indicate that graduates who lived in Ontario two years after graduation were more likely to have pursued higher levels of education than graduates who lived in the Prairies two years after their graduation. In 1988 and 1997, the graduates who lived in Ontario at the time of the interview had higher odds of having obtained a master's degree or a doctoral degree compared to the graduates who lived in Quebec. This seems to indicate that the 1988 and 1997 graduates who lived in Ontario were more likely to have pursued higher levels of education than graduates who lived in Quebec two years after their graduation. However, in 1992, the graduates who lived in Ontario at the time of the interview had reduced odds of having obtained a master's degree or a doctoral degree compared to the graduates who lived in Quebec. This seems to indicate that in 1992, graduates who lived in Ontario were less likely to have pursued higher levels of education than graduates who lived in Quebec two years after their graduation. In 1988 and 1997, the graduates who lived in Ontario at the time of the interview had higher odds of having obtained a master's degree or a doctoral degree compared to the graduates who lived in British Columbia. This seems to indicate that the 1988 and 1997 graduates who lived in Ontario were more likely to have pursued higher levels of education than graduates who lived in British Columbia two years after their graduation. However, in 1992, the graduates

who lived in Ontario at the time of the interview had the same odds of having obtained a master's degree or a doctoral degree compared to the graduates who lived in British Columbia.

3) First language spoken and still understood: estimated odds ratio of obtaining response $> j$ for ...

Table 4.13: Estimated odds of having response $> j$ ($j = 1, 2$) (language)

Odds ratios	88	92	97
Language other than English or French over English or French	1.26	1.39	1.57

for a given combination of field of study, full-time/part-time status, gender, region of residence, and category of age. So for all of the three years, graduates whose first language is other than English or French had higher odds of having a master's degree or a doctoral degree compared to the graduates who spoke English or French. This seems to indicate that graduates whose first language is other than English or French were more likely to pursue higher levels of education than graduates who spoke English or French.

4) Field of study and gender (for age from 15 to 24): estimated odds ratio of obtaining response $> j$ for ...

Table 4.14: Estimated odds of having response $> j$ ($j = 1, 2$) (interaction between gender and field of study for the 15 to 24 age category)

Odds ratios	88	92	97
Men in Other Soc. sc. over women in Other Soc. sc.	1.21	0.96	0.86
Men in Ag. and bio. sc. over women in Ag. and bio. sc.	0.85	1.84	1.58
Men in Eng. and app. sc. over women in Eng. and app. sc.	1.11	0.76	0.89
Men in Med. and health prof. over women in Med. and health prof.	0.96	1.06	1.48

for a given combination of region of residence, first spoken language and still understood, category of age, and status of registration.

In 1988 and 1992, male graduates in “Other Social Sciences” had higher odds of having a master’s degree or a doctoral degree compared to female graduates in “Other Social Sciences”; male graduates in Agricultural and biological sciences had higher odds of having a master’s degree or a doctoral degree compared to female graduates in Agricultural and biological sciences; male graduates in Engineering and applied sciences had higher odds of having a master’s degree or a doctoral degree compared to female graduates in Engineering and applied sciences; and male graduates in Medical and health professions had higher odds of having a master’s degree or a doctoral degree compared to female graduates in Medical and health professions. This seems to indicate that in 1988 and in 1992, the male graduates in these fields of study were more likely to pursue higher levels of education than female graduates studying in the same fields of study. However, in 1997, male graduates in “Other Social Sciences” had reduced odds of having a master’s degree or a doctoral degree compared to female graduates in “Other Social Sciences”; male graduates in Agricultural and biological sciences had reduced odds of having a master’s degree or a doctoral degree

compared to female graduates in Agricultural and biological sciences; male graduates in Engineering and applied sciences had reduced odds of having a master's degree or a doctoral degree compared to female graduates in Engineering and applied sciences; and male graduates in Medical and health professions had reduced odds of having a master's degree or a doctoral degree compared to female graduates in Medical and health professions. This seems to indicate that in 1997, male graduates in these fields of study were less likely to pursue higher levels of education than female graduates studying in the same fields of study. Recall that these conclusions are for graduates between 15 and 24 years old at the time of the survey.

5) Field of study and gender (for age from 25 to 29): estimated odds ratio of obtaining response $\rightarrow j$ for

Table 4.15: Estimated odds of having response $> j$ ($j = 1, 2$) (interaction between gender and field of study for the 25 to 29 age category)

Odds ratios	88	92	97
Men in Other Soc. sc. over women in Other Soc. sc.	1.21	0.96	0.86
Men in Ag. and bio. sc. over women in Ag. and bio. sc.	0.85	1.84	1.58
Men in Eng. and app. sc. over women in Eng. and app. sc.	1.11	0.76	0.89
Men in Med. and health prof. over women in Med. and health prof.	0.96	1.06	1.48

for a given combination of region of residence, first spoken language and still understood, category of age, and status of registration.

In 1988, the male graduates in Other Social sciences had higher odds of having a master's

degree or a doctoral degree compared to the female graduates in Other social sciences and the male graduates in Engineering and applied sciences had higher odds of having a master's degree or a doctoral degree compared to the female graduates in Engineering and applied sciences. This seems to indicate that in 1988, the male graduates in these fields of study were more likely to pursue higher levels of education than female graduates studying in the same fields of study. However, in 1992 and in 1997, the male graduates in Other Social sciences had reduced odds of having a master's degree or a doctoral degree compared to the female graduates in Other social sciences and the male graduates in Engineering and applied sciences had reduced odds of having a master's degree or a doctoral degree compared to the female graduates in Engineering and applied sciences. This seems to indicate that in 1992 and in 1997, the male graduates in these fields of study were less likely to pursue higher levels of education than female graduates studying in the same fields of study.

In 1988, male graduates in Agricultural and biological sciences had reduced odds of having a master's degree or a doctoral degree compared to female graduates in Agricultural and biological sciences and male graduates in Medical and health professions had higher odds of having a master's degree or a doctoral degree compared to female graduates in Medical and health professions. This seems to indicate that in 1988, the male graduates in these fields of study were less likely to pursue higher levels of education than female graduates studying in the same fields of study. However, in 1992 and in 1997, male graduates in Agricultural and biological sciences had higher odds of having a master's degree or a doctoral degree compared to female graduates in Agricultural and biological sciences and male graduates in Medical and health profession had higher odds of having a master's degree or a doctoral degree compared to female graduates in Medical and health professions. This seems to indicate that in 1992 and in 1997, male graduates in these fields of study were more likely to pursue higher levels of education than female graduates studying in the same fields of study. Recall that these conclusions apply for graduates who were between 25 and 29 years old at

the time of the NGS.

4.5 Problems

One of the problems encountered during this analysis was the fact that the total population counts by stratum were not available for the 1988 NGS and the 1992 NGS. So the total population counts by stratum had to be estimated from the weights. This was the only way of getting an estimate of those totals, since the original population counts were not available. Fortunately, the administrative records from which the 1997 sample was drawn were available; that allowed the calculation of the population counts by stratum from this database. It was also possible to estimate the population counts from the weights on the sample file. This allowed a comparison of the results between the model using the weighted counts and the model using the administrative counts. We saw that the precision of the parameter estimates was better, but not significantly better, for the model using the weighted counts.

Another problem was the sparseness in the data. In the table below, we see that almost half of the cells defined by response j and combination of explanatory variables i have less than one observation by cell.

Table 4.16: Number of observations in the subpopulation by cell

Number of cells with	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
< 1 observation in the subpopulation	1,753	1,710	1,658	1,658
$1 \leq \text{observations} < 5$	691	2,463	786	786
$5 \leq \text{observations} < 25$	867	0	816	816
$25 \leq \text{observations}$	715	0	808	808
TOTAL	4,026	4,173	4,068	4,068

Infinite parameter values can arise from patterns of zeros in sparse data arrays (McCullagh, 1980) and the deviance function and the Pearson's statistic are poor indicators of the goodness-of-fit of the model in the presence of sparseness.

One predictor that was interesting to include in the model was the level of parental education (using the highest level of education of the parents). Butlin (1999) stated that high school graduates with university educated parents had higher odds of attending university (controlling for other factors). So this factor could also have had an effect on the university level of certification obtained at the graduation. But this factor and other interaction terms had to be removed from the analysis because it caused problems with the singularity of the matrix of estimable parameters. In this case, no conclusion could have been drawn. Also, it occurred that the standard errors of the parameters were zero for many parameter estimates and therefore no design effect or p -value was computed. As a result the most important effects and interaction terms have been retained in the analysis.

The corrected test for Pearson's chi-squared statistic (see Roberts, 1985 or Rao, Kumar, and Roberts, 1989) has not been computed for the purpose of assessing the goodness-of-fit of the model. The estimated covariance matrix of $\hat{\mathbf{P}}$ was not available through SUDAAN's output. In this case, Roberts (1985) stated that it is worthwhile to obtain an upper bound on δ , which depends only on the cell design effects or on the generalised design effects. The cell design effects were not available either. So this method of validating the model has not been performed. It would have been interesting to compare the value of the corrected Pearson's statistic to the standard Pearson's statistic.

CHAPTER 5: CANADIAN UNIVERSITY ENROLMENT AND PROJECTIONS FOR FULL-TIME CANADIAN UNIVERSITY TEACHERS

5.1 Canadian university enrolment 1976-1998

Using the USIS database, we examine a profile of Canadian university students and factors affecting graduate and undergraduate enrolment.

5.1.1 Enrolment profile

The USIS database has been described in chapter 2. A special file from this database has been created on which enrolment totals by province, gender, status of registration and undergraduate/graduate status from 1972-73 to 1998-99 were displayed. A graphical profile of the enrolments was obtained. These graphics are in Appendix G. The enrolment totals for the Census years (from 1976-77 to 1996-97) have been obtained, in addition to the 1997-98 and the 1998-99 years (two intercensal years). Note that when we refer to the year 1976, we mean the survey year 1976-77, and so on for the other years. The regions of institution have been created as follows:

- 1)** the Atlantic provinces (e.g. Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick)
- 2)** Quebec
- 3)** Ontario
- 4)** the Prairies (Manitoba, Saskatchewan, Alberta)
- 5)** British Columbia

Furthermore, four categories of age have been created: 15-24 years old; 25-29 years old; 30-34 years old; and 35 years old and over.

(See figures 5.1 to 5.26 in Appendix G)

At the national level, the undergraduate enrolment totals from 1976 to 1998 were always the highest for the students aged 15 to 24 years old, followed by the students in the 35 year-old and over category, then by the students in the 25 to 29 year-old category, and finally by the students in the 30 to 34 year-old category. The undergraduate enrolment totals in the 15 to 24 year-old category increased from 1991 to 1998, while the enrolment totals in the three other categories decreased slightly from 1991 to 1998.

Now the undergraduate enrolment totals were the highest for the students in the 15 to 24 year-old category, in every region of Canada (i.e. the Atlantic provinces, Quebec, Ontario, the Prairies, and British Columbia). The undergraduate enrolment totals were the lowest for the students in the 30 to 34 year-old category, in every region of Canada. In every region, the undergraduate enrolment totals increased from 1976 to 1998, for the students in the 15 to 24 year-old category. In all of the regions except Quebec and Ontario, the undergraduate enrolment totals in the three other categories of age remain more or less constant from 1976 to 1998, with some small increases or small decreases. In Quebec and Ontario, the undergraduate enrolment totals in the 35 year-old and over category increased from 1976 to 1991, and then decreased slightly until 1998.

Still at the national level, the graduate enrolment totals from 1976 to 1998 were always the highest for the students aged 25 to 29 years old, and almost always the lowest for the students aged 15 to 24 years old. The graduate enrolment totals have had a large increase from 1976 to 1991 for the students 35 years old and over, and have started to be higher than the graduate enrolment totals for the students 30-34 years old between 1981 and 1986, and until 1998. The graduate enrolment totals for the students in the 25-29 year-old category increased from 1976 to 1998.

The graduate enrolments totals were the lowest for the students in the 15 to 24 year-old category, in every region of Canada from 1981 to 1998. In 1976, the graduate enrolments

totals were the lowest for the students in the 35 year-old and over category, in every region. But in every region, the graduate enrolment totals from 1976 to 1991 have had a large increase for students in the 35 year-old and over category. In Quebec, the Prairies and British Columbia, the graduate enrolment totals were the highest for the students in the 35 year-old and over category from 1991 to 1998. In Ontario, the the graduate enrolment totals were the highest for the students in the 25-29 year-old category from 1976 to 1998.

It is Ontario that had the highest undergraduate enrolment totals from 1976 to 1998. It is followed (in decreasing order) by Quebec, then the Prairies, the Atlantic provinces, and British Columbia. Ontario had the highest graduate enrolment totals from 1976 to 1981 but from 1986 to 1998, it is Quebec that had the highest graduate enrolment totals. The number of students enrolled in a graduate program in Quebec increased from 1976 to 1998. The number of students enrolled in a graduate program in the Atlantic provinces, the Prairies, or British Columbia slightly increased from 1976 to 1998.

At the national level, the male enrolment totals (for undergraduates and graduates) from 1976 to 1998 were always the highest for the students aged 15 to 24 years old, and it increased a little. The male enrolment totals from 1976 to 1998 remained more or less constant in the three other categories of age. Still at the national level, the female enrolment totals from 1976 to 1998 were always the highest for the students aged 15 to 24 years old, and it increased by about 150,000. The female enrolment totals from 1976 to 1998 were always the lowest for the students aged 30 to 34 years old.

In every region of Canada, the male enrolment totals (for undergraduates and graduates) were always the highest for the students aged 15 to 24 years old. In every region, except Quebec, the male enrolment totals increased slightly for the students in the 15 to 24 year-old category, while the male enrolment totals remained more or less constant in the three other age categories. In Quebec, the male enrolment totals increased for the students in the four age categories from 1976 to 1986, then decreased for the students in the 15 to 24 year-old

category, the 25 to 29 year-old category and the 30 to 34 year-old category until 1998.

In every region of Canada, the female enrolment totals (for undergraduates and graduates) were always the highest for the students aged 15 to 24 years old. In every region, the female enrolment totals increased for the students in the 15 to 24 year-old category from 1976 to 1998. In the Atlantic provinces, the Prairies and British Columbia, the female enrolment totals remained more or less constant in the three older age categories from 1976 to 1998. In Quebec, the female enrolment totals increased for the students in the three older age categories, from 1976 to 1986. Then they remained more or less constants until 1998. In Ontario, the female enrolment totals increased for the students in the three older age categories, from 1976 to 1991. Then they remained more or less constants until 1998.

It is Ontario that had the highest male enrolment totals from 1976 to 1998. It is followed (in decreasing order) by Quebec, then the Prairies, the Atlantic provinces, and British Columbia. The same conclusions apply for the female enrolment totals. In all regions, except British Columbia, the male enrolment totals decreased slightly from 1991 to 1998. In the Atlantic provinces, the Prairies, and British Columbia, the female enrolment totals increased from 1976 to 1998. In Quebec and Ontario, the female enrolment totals had a large increase from 1976 to 1991 and then decreased from 1991 to 1998.

5.1.2 Factors affecting enrolment

In order to study the factors affecting the undergraduate and graduate enrolment, two ANOVA models have been fitted. The response variable is the number of enrolments. The four predictors are the year of survey (1976, 1981, 1986, 1991, 1996, 1997, and 1998); the region of enrolment (the Atlantic provinces, Quebec, Ontario, the Prairies, British Columbia); the age category (15-24, 25-29, 30-34, 35 and over); and the gender (male or female). For the "undergraduate" model, the response is the number of undergraduate enrolments and for the "graduate" model, the response is the number of graduate enrolments.

The statistical model for this design is:

$$y_{ijkl} = \mu + \tau_i + \beta_j + \delta_k + \lambda_l + \epsilon_{ijkl} \{i = 1, \dots, 7; j = 1, \dots, 5; k = 1, 2; l = 1, \dots, 4\}$$

where μ is the overall mean, τ_i is the effect of the i^{th} survey year, β_j is the effect of the j^{th} region of enrolment, δ_k is the effect of the k^{th} gender, λ_l is the effect of the l^{th} age category, and ϵ_{ijkl} is the usual normally and independently distributed random error term $(0, \sigma^2)$.

For both models, the hypothesis of normality of the errors and the hypothesis of constant variance of the errors were rejected. On both graphs (see figures 5.27 and 5.28 in Appendix G) of the residuals versus the predicted values, the plot has a pattern: it looks like the letter “V”. That is why the hypothesis of constant variance of the residuals has been rejected (and so was the hypothesis of a linear relationship between the factors and the response). The hypothesis of normality of the residuals has been tested with the Shapiro-Wilk statistic. At the significance level 5%, this hypothesis is rejected for both models (p -value lower than 0.05). A normal probability plot has been obtained in order to assess the normality of the residuals too. The two normal probability plots seem to have positive skewness so we can conclude that the hypotheses of normality of the errors are also rejected for both models.

So a transformation of the response variable has been performed as an attempt to solve the nonconstant variance and the nonnormality problems. Instead of using as the response the number of enrolments, the natural logarithm of the number of enrolments was used for both models. The same explanatory variables were used for this fitting.

For both models, the hypothesis of normality of the errors was rejected. But the hypotheses of constant variances were accepted. On both graphs (see figures 5.29 and 5.30 in Appendix G) of the residuals versus the predicted values, residuals are mostly between -1 and 1 . There is no specific pattern, except that a small group of residuals are a little higher than 1 . There is no specific pattern on these plots. That is why the hypothesis of constant variance of the residuals has been accepted for both models. The hypothesis of normality of the residuals has been tested with the Shapiro-Wilk statistic. At the significance level 5%, this hypothesis is

rejected for both models (p -value lower than 0.05). A normal probability plot has been obtained in order to assess the normality of the residuals too. The two normal probability plots seem to have positive skewness. We can conclude that the hypotheses of normality of the errors are also rejected for both models.

Since the hypothesis of constant variances is verified but the hypothesis of a normal distribution of the residuals is rejected, a nonparametric ANOVA (based on ranks) was performed for both models. If the conclusions drawn from the nonparametric ANOVA are the same than the ones obtained from the parametric ANOVA (with transformed response), we can conclude that the conclusions drawn from the latter model are correct.

In the table below, we see the p -values obtained for the testing that the treatment effects are zero in the “undergraduate” model (year, region of enrolment, age, gender). The null hypotheses are:

- 1) $H_0 : \mu_1 = \dots = \mu_7 = 0$
 $H_1 : \text{at least one } \mu_i \neq 0$
- 2) $H_0 : \beta_1 = \dots = \beta_5 = 0$
 $H_1 : \text{at least one } \beta_i \neq 0$
- 3) $H_0 : \delta_1 = \delta_2 = 0$
 $H_1 : \text{at least one } \delta_i \neq 0$
- 4) $H_0 : \lambda_1 = \dots = \lambda_4 = 0$
 $H_1 : \text{at least one } \lambda_i \neq 0$

If we compare the p -values, for each factor, obtained from the parametric model (with transformed response) and the nonparametric model, we see that they are all lower than 0.05. So at the 5% significance level, we reject the null hypotheses. So for the four variables, at least one level of each factor has an effect on the response. We obtain these conclusions from the parametric and the nonparametric models. We also see that the coefficient of determination (i.e. the proportion of variability explained by the model) is approximately the

same for both models.

Table 5.1: Observed p -values and coefficient of determination (“undergraduate” model)

Undergraduate	Parametric ANOVA	Nonparametric ANOVA
Year	0.0001	0.0001
Region of enrolment	0.0001	0.0001
Age	0.0001	0.0001
Gender	0.0005	0.0004
R-square	0.9435	0.9470

In the table below, we see the p -values obtained for the testing that the treatment effects are zero in the “graduate” model (year, region of enrolment, age, gender). The null hypotheses are:

- 1) $H_0 : \mu_1 = \dots = \mu_7 = 0$
 $H_1 : \text{at least one } \mu_i \neq 0$
- 2) $H_0 : \beta_1 = \dots = \beta_5 = 0$
 $H_1 : \text{at least one } \beta_i \neq 0$
- 3) $H_0 : \delta_1 = \delta_2 = 0$
 $H_1 : \text{at least one } \delta_i \neq 0$
- 4) $H_0 : \lambda_1 = \dots = \lambda_4 = 0$
 $H_1 : \text{at least one } \lambda_i \neq 0$

If we compare the p -values, for each factor, obtained from the parametric model (with transformed response) and the nonparametric model, we see that they are all lower than 0.05. So at the 5% significance level, we reject the null hypotheses. So for the four variables, at least one level of each factor has an effect on the response. We obtain these conclusions from the parametric and the nonparametric models. We also see that the coefficient of determination (i.e. the proportion of variability explained by the model) is a little lower for

the nonparametric model but both are high.

Table 5.2: Observed p -values and coefficient of determination (“graduate” model)

Graduates	Parametric ANOVA	Nonparametric ANOVA
Year	0.0001	0.0001
Region of enrolment	0.0001	0.0001
Age	0.0001	0.0001
Gender	0.0001	0.0001
R-square	0.9268	0.8975

From the ANOVA model for the undergraduates and the ANOVA model for the graduates, we can conclude that the year of enrolment, the region of enrolment, the age and the gender have an effect on, respectively, the natural logarithm of the total number of undergraduate enrolments and on the natural logarithm of the total number of graduate enrolments.

5.2 Estimated proportions of Canadian university enrolment 1994-1998

In this section, we combine Census population estimates with USIS enrolment information to analyze patterns in proportions of Canadian university enrolment.

5.2.1 Patterns over 1994-1998

Statistics Canada, through the Statistical Reference Centre, provided population estimates, based on Census estimates, from 1971 to 1998, by single year of age, gender and province of residence. Since the population estimates by single year of age, gender, and province of residence were known, it was possible to estimate the proportion of students enrolled at university by region of institution, gender, and age category, for several years. From section 5.1.1, we obtained the number of enrolments from the USIS from 1994 to 1998, by region of institution, gender, and age category. Age categories and region of residence have been

created on the population estimates file, following, respectively, the creation of the age categories and the region of institution in section 5.1.1. Then the population estimates from 1994 to 1998 have been obtained by region of residence, gender, and categories of age. It was then possible to obtain the proportion of students enrolled at university by region of institution, gender, and age category from 1994 to 1998.

When these proportions have been obtained, we made some assumptions. The first one is that the region of the institution corresponds to the region of residence. It may not be always the case (for example if a student is enrolled in Ottawa (Ontario) but lives in Aylmer (Quebec)). To explain the second assumption, recall that the enrolment survey collects information on students count as of December 1st in all provinces except Ontario, where the reference date is November 1st. The student count is used as a proxy for the total number of students enrolled during a complete academic year. So since the population estimates are obtained for July 1st of each year, and the enrolment counts are obtained in the Fall of each year, we have to keep in mind that the proportions obtained will not be exactly the proportions that would have been obtained if the enrolment counts for a civil year (and not only at the Fall term of each year). The last thing to notice is that no proportions in the Yukon or Northwest Territories have been obtained since no enrolment counts are available in these provinces at the university level.

(See figures 5.31 to 5.50 in Appendix G)

In every region and for every age category, the proportion of women enrolled at university is always higher than the proportion of men enrolled at university, from 1994 to 1998. In the Atlantic provinces and Quebec, the proportion of men enrolled at university is fairly stable from 1994 to 1998, for every age category except the 30-34 year-old category, where this proportion decreased a little from 1994 to 1998. In these same regions, the proportion of women enrolled at university increased slightly from 1997 to 1998, for women aged 15 to 29 years old. The proportion of women aged 30 to 34 years old enrolled at university decreased

a little from 1994 to 1998. The proportion of women enrolled at university is fairly stable from 1994 to 1998, for women aged 35 years old and over. The highest proportion of men enrolled at university is for men aged 15 to 24 years old. The highest proportion of women enrolled at university is for women aged 15 to 24 years old.

In Ontario, the proportion of men enrolled at university is fairly stable from 1994 to 1998, for men aged 35 years old and over. The proportion of men enrolled at university decreased a little from 1994 to 1998 in the three lower age categories. In this same region, the proportion of women enrolled at university decreased a little from 1994 to 1998 aged 25 to 29 and 30 to 34 years old. The proportion of women enrolled at university is fairly stable from 1994 to 1998, for women aged 15 to 24 years old and 35 years old and over. The highest proportion of men enrolled at university is for men aged 15 to 24 years old. The highest proportion of women enrolled at university is for women aged 15 to 24 years old.

In the Prairies, the proportion of men enrolled at university decreased a little from 1994 to 1998 aged 15 to 24 years old. The proportion of men enrolled at university is fairly stable from 1994 to 1998, for men in the three older age categories. In this same region, the proportion of women enrolled at university is fairly stable from 1994 to 1998, for women aged 15 to 24 years old, 30 to 34 years old, and 35 years old and over. The proportion of women enrolled at university increased a little from 1995 to 1998 for women aged 25 to 29 years old. The highest proportion of men enrolled at university is for men aged 15 to 24 years old. The highest proportion of women enrolled at university is for women aged 15 to 24 years old.

In British Columbia, the proportion of men enrolled at university increased a little from 1994 to 1998 aged 15 to 24 years old. The proportion of men enrolled at university is fairly stable from 1994 to 1998, for men aged 25 years old and over. In this same region, the proportion of women enrolled at university increased a little from 1994 to 1998 for women aged 15 to 24 years old, and 25 to 29 years old. The proportion of women enrolled at university is fairly

stable from 1994 to 1998, for women in the two older age categories. The highest proportion of men enrolled at university is for men aged 15 to 24 years old. The highest proportion of women enrolled at university is for women aged 15 to 24 years old.

5.2.2 Modelling proportions

When we looked at the variations in the proportions of Canadian university enrolment in section 5.2.1, we sometimes saw some increasing trends or decreasing trends. In order to predict if these proportions would increase through time, decrease, or remain constant, we could have modelled these proportions. The proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia seemed to increase from 1994 to 1998. In order to assess the effect of the year of enrolment on this proportion, a simple linear regression analysis was performed. The response was the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia. The explanatory variable was the year of enrolment. We assumed the year of enrolment to be a continuous variable. For the purpose of the analysis, the variable “year” had to be recoded as follows:

- 1) 1: 1994
- 2) 2: 1995
- 3) 3: 1996
- 4) 4: 1997
- 5) 5: 1998

The model used was:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

when Y_i is the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia for the i^{th} year ($i = 1, 2, 3, 4, 5$); β_0 is the intercept term to be estimated; β_1 is the parameter associated to the explanatory variable “year” to be estimated; X_i is the value of the i^{th} year; and ϵ_i is the normally and independently distributed random

error term ($0, \sigma^2$). Note that only five observations were used to model the response and therefore we do not rely on a long period of time to establish the predictions.

For this model, the hypothesis of normality of the errors and the hypothesis of constant variance of the errors were accepted. The graph of the residuals versus the predicted values has no specific pattern so we can conclude that the assumption of constant variance of the errors is accepted. The hypothesis of normality of the residuals has been tested with the Shapiro-Wilk statistic. At the significance level of 5%, this hypothesis is accepted (the p -value of the test is equal to 0.82). A normal probability plot has been obtained in order to assess the normality of the residuals too. On the normal probability plot, the residuals seem aligned on a straight line. We can conclude that the hypothesis of normality of the errors is accepted for this model. It seems that the observation corresponding to the year 1997 might be an outlier. The studentized residual corresponding to this observation is a greater than two and the Cook distance of this observation is a little greater compared to the Cook distances of the other observations.

In order to assess if this observation is influential, it has been deleted from the analysis. The parameter estimates are not highly affected by the deletion of this possible outlying observation. So we can keep this observation in the analysis. The graph of the response versus the explanatory variable does not show a specific pattern, nor the graph of the residuals versus the explanatory variable (the time). So we can conclude that there is a linear relationship between the response and the explanatory variable and there is no correlation between the observations.

The coefficient of determination is very high (0.9778) and we can conclude that almost 98% of the variability in the data is explained by the model. The hypothesis that a parameter was equal to zero was tested with a t -test. Both the intercept and the parameter associated with the year were highly significant (p -value less than 5%). So at significance level 5%, we can conclude that these two parameters must be retained in the model.

Since the assumptions of the model are satisfied, we can now give the estimated regression equation:

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ &= 8.9025 + 0.3469 * X_i\end{aligned}$$

If, for example, we want to estimate the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia in 2002, we replace X_i in the equation above by 9 (since the year was recoded). We obtain that the estimated proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia in 2002 is 12.02%. In conclusion, we saw that on the graph the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia seemed to increase a little through time and the regression analysis confirmed that the year of enrolment was significant. This regression analysis gives an indication of the type of analysis that could be performed to study the more in depth the Canadian university enrolment. But the results of this analysis must be used with care since they are only based on five observations.

5.3 Predictors of the university level of certification obtained at graduation

5.3.1 Proportional odds model

In order to obtain estimates of the number of doctoral graduates (given some characteristics), the proportional odds model fitted in chapter 4 could not be used because some of the explanatory variables were not available on the population projections data file. The way that the proportional odds model has been used to estimate the number of doctoral graduates is explained in section 5.6.3.3, later in this chapter. So a first proportional odds model was fitted using only the age, the region of residence and the gender as the three explanatory variables. The response was still the level of certification at the graduation. The levels of each of these variables are the same as the levels of the variables of the proportional odds model fitted in chapter 4, except for the region of residence where British Columbia, Yukon,

and Northwest Territories have been collapsed into the same category. When British Columbia and Yukon/Northwest Territories were in two distinct categories, we got some parameter estimates of an interaction involving British Columbia or Yukon/NWT (not equal to zero) their with standard error equal to zero. It seemed that collapsing these two regions into one category solved the problem. The estimated model is defined as:

$$\begin{aligned}
 \log \left(\frac{\widehat{F}_j(\mathbf{x})}{1 - \widehat{F}_j(\mathbf{x})} \right) = & \widehat{\alpha}_{j=1} + \widehat{\alpha}_{j=2} + \widehat{\beta}_1 GNDN_1 + \widehat{\beta}_2 GNDN_2 + \widehat{\beta}_3 REG_1 + \widehat{\beta}_4 REG_2 \\
 & + \widehat{\beta}_5 REG_3 + \widehat{\beta}_6 REG_4 + \widehat{\beta}_7 REG_5 + \widehat{\beta}_8 CATAGE_1 + \widehat{\beta}_9 CATAGE_2 + \widehat{\beta}_{10} CATAGE_3 \\
 & + \widehat{\beta}_{11} CATAGE_4 + \widehat{\beta}_{12} GNDN_1 REG_1 + \widehat{\beta}_{13} GNDN_1 REG_2 + \widehat{\beta}_{14} GNDN_1 REG_3 \\
 & + \widehat{\beta}_{15} GNDN_1 REG_4 + \widehat{\beta}_{16} GNDN_1 REG_5 + \widehat{\beta}_{17} GNDN_2 REG_1 + \widehat{\beta}_{18} GNDN_2 REG_2 \\
 & + \widehat{\beta}_{19} GNDN_2 REG_3 + \widehat{\beta}_{20} GNDN_2 REG_4 + \widehat{\beta}_{21} GNDN_2 REG_5 + \widehat{\beta}_{22} GNDN_1 CATAGE_1 \\
 & + \widehat{\beta}_{23} GNDN_1 CATAGE_2 + \widehat{\beta}_{24} GNDN_1 CATAGE_3 + \widehat{\beta}_{25} GNDN_1 CATAGE_4 \\
 & + \widehat{\beta}_{26} GNDN_2 CATAGE_1 + \widehat{\beta}_{27} GNDN_2 CATAGE_2 + \widehat{\beta}_{28} GNDN_2 CATAGE_3 \\
 & + \widehat{\beta}_{29} GNDN_2 CATAGE_4 + \widehat{\beta}_{30} REG_1 CATAGE_1 + \widehat{\beta}_{31} REG_1 CATAGE_2 \\
 & + \widehat{\beta}_{32} REG_1 CATAGE_3 + \widehat{\beta}_{33} REG_1 CATAGE_4 + \widehat{\beta}_{34} REG_2 CATAGE_1 \\
 & + \widehat{\beta}_{35} REG_2 CATAGE_2 + \widehat{\beta}_{36} REG_2 CATAGE_3 + \widehat{\beta}_{37} REG_2 CATAGE_4 \\
 & + \widehat{\beta}_{38} REG_3 CATAGE_1 + \widehat{\beta}_{39} REG_3 CATAGE_2 + \widehat{\beta}_{40} REG_3 CATAGE_3 \\
 & + \widehat{\beta}_{41} REG_3 CATAGE_4 + \widehat{\beta}_{42} REG_4 CATAGE_1 + \widehat{\beta}_{43} REG_4 CATAGE_2 \\
 & + \widehat{\beta}_{44} REG_4 CATAGE_3 + \widehat{\beta}_{45} REG_4 CATAGE_4 + \widehat{\beta}_{46} REG_5 CATAGE_1 \\
 & + \widehat{\beta}_{47} REG_5 CATAGE_2 + \widehat{\beta}_{48} REG_5 CATAGE_3 + \widehat{\beta}_{49} REG_5 CATAGE_4
 \end{aligned} \tag{5.1}$$

5.3.2 Validation of the model

The first goodness-of-fit statistic that was computed was the Pearson chi-squared statistic, for the four models.

Table 5.3: Pearson chi-squared statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
530.80	522.59	1,188.16	1,188.16

These four chi-squared statistics take large values. If we compare the 1988, 1992, and 1997 statistics to $\chi^2_{80-51(0.95)} = 39.08$, when 51 is the number of parameter estimates in the model and 80 is the number of levels of the response minus one ($= 2$) times the number of combinations of levels of the explanatory variables ($= 40$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Recall that this is not a test corrected for the sampling plan and note that when $\widehat{\pi}_{ij} = 0$, the calculation of the contribution to the Pearson's statistic could not be performed, since there was a zero in the denominator.

The four deviance functions have also been computed:

Table 5.4: Deviance statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
365.80	132.13	1,440.77	1,440.77

These four deviance functions take large values. If we compare the 1988, 1992, and 1997 statistics to $\chi^2_{80-51(0.95)} = 39.08$, we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the 5% significance level. When $\widehat{\pi}_{ij} = 0$ or $\widetilde{\pi}_{ij} = 0$, the calculation of the contribution to the deviance function could not be performed, since $\log(0)$ does not exist.

The Wald chi-square has been used to test the null hypothesis that all of the parameter estimates in the model are equal to zero. This will test simultaneously that every of the 51 parameters are equal to zero. For all of the four models, this null hypothesis is strongly rejected, indicating that the proportional odds model fits better than a model with all coefficients equal to zero. See results below:

Table 5.5: Wald chi-square statistics

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Wald statistic	669,248.73	903,046.48	905,232.58	915,814.82
<i>p</i> -value	0.0000	0.0000	0.0000	0.0000

The cell residuals have been computed. Once they have been computed, four groups have been created and the number of cells included in each of these groups was calculated.

Table 5.6: Frequencies of cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	50(42%)	62(52%)	44(37%)	44(37%)
$-10 \leq \text{cell res.} < -2$ or $2 < \text{cell res.} \leq 10$	41	30	45	45
cell res. < -10 or cell res. > 10	0	0	1	1
cell residual is missing	29	28	30	30

Approximately 40 to 50% of the cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not be computed (either when $\hat{\pi}_{ij} = 1$ or $\hat{\pi}_{ij} = 0$). If we delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 5.7: Proportions of cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
$-2 \leq \text{cell res.} \leq 2$	55%	67%	49%	49%

These proportions seem reasonable. There does not appear to be a cell residual with an abnormally large value (see table below). If we compare the maximum and minimum values of the four models, we see that they are roughly in the same range of values.

Table 5.8: Maximum and minimum cell residuals

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Max. cell residual	7.75	5.05	9.06	9.06
Min. cell residual	-8.56	-8.22	-12.62	-12.62

The cumulative cell residuals have also been computed, using cumulative totals and cumulative fitted probabilities. Once they have been computed, five groups have been created and the number of cells included in each of these groups was calculated.

Table 5.9: Frequencies of cumulative cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cum. cell res.} \leq 2$	34(28%)	36(30%)	30(25%)	30(25%)
$-10 \leq \text{cum. cell res.} < -2$ or $2 < \text{cum. cell res.} \leq 10$	26	28	31	31
$-100 \leq \text{cum. cell res.} < -10$ or $10 < \text{cum. cell res.} \leq 100$	1	0	1	1
$\text{cum. cell res.} < -100$ or $\text{cum. cell res.} > 100$	1	0	0	0
cum. cell res. is missing	58	56	58	58

Approximately 25% to 30% of the cumulative cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not been computed (either when the cumulative probability was equal to zero or one). If we delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 5.10: Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	55%	56%	48%	48%

These proportions are about 25% higher than when the missing cumulative cell residuals were included in the counts. These counts are comparable to those obtained for the cell residuals.

A plot of each of the contributions to the deviance function against its index (i, j) was obtained in order to detect large deviance contributions (see Appendix G). For each model, a plot of the deviance contributions corresponding to response level $j, j = 1, 2, 3$, for all of the I combinations of explanatory variables has been drawn. On figure 5.51, for $Y = 1$, in 1988, there does not seem to have any deviance contributions that is larger than the others. The deviance contributions on the graph seem to be between -100 and 100 . On figure 5.52, for $Y = 2$, in 1988, there was one deviance contribution that seemed to be large. This deviance contribution was approximately equal to 250 (it represents the female graduates aged 25 to 29 years old living in Ontario). The other deviance contributions on the graph seem to be between -150 and 150 . On figure 5.53, for $Y = 3$, in 1988, there were three deviance contributions that seemed to be large. These deviance contributions are between 100 and 250 (they represent the male graduates aged 30 to 34 years old living in Ontario; and it represents the male graduates aged 35 years old and over living in Ontario; and the female graduates aged 35 years old and over living in Ontario). The other deviance contributions on the graph seem to be between -50 and 50 . So for the 1988 model, only four deviance contributions (for $Y = 2$ and $Y = 3$) seem to be large.

On figure 5.54, for $Y = 1$, in 1992, there was one deviance contribution that was large. It was approximately equal to -200 (it represents the male graduates aged 25 to 29 years old living in Ontario). The deviance contributions on the graph seem to be between -100 and 100 . On figure 5.55, for $Y = 2$, in 1992, there were two deviance contributions that seemed to be

large. These deviance contributions were approximately equal to 250 and 350 (they represent the male graduates aged 25 to 29 years old living in Ontario; and the female graduates aged 25 to 29 years old living in Quebec). The other deviance contributions on the graph seem to be between -150 and 100 . On figure 5.56, for $Y = 3$, in 1992, there does not seem to have any large contribution. The deviance contributions on the graph seem to be between -100 and 150 . So for the 1992 model, only three deviance contributions (for $Y = 1$ and $Y = 2$) seem to be large.

On figure 5.57, for $Y = 1$, in 1997, there were two deviance contributions that seemed to be large. The first had a deviance contribution that was approximately equal to -200 (it represents the male graduates aged 25 to 29 years old living in Ontario) and the other has a value that was approximately of 300 (it represents the male graduates aged 35 years old and over living in Ontario). The other deviance contributions on the graph seem to be roughly between -100 and 150 . On figure 5.58, for $Y = 2$, in 1997, there were two deviance contributions that seemed to be large. One had a value approximately equal to -450 (it represents the male graduates aged 35 years old and over living in Ontario) and the other had a value approximately equal to 450 (it represents the male graduates aged 25 to 29 years old living in Ontario). The other deviance contributions on the graph seem to be between -200 and 200 . On figure 5.59, for $Y = 3$, in 1997, there was no deviance contribution that seemed to be large. The deviance contributions on the graph seem to be between -100 and 250 . So for the 1997 model, four deviance contributions seem to be large. Same results applied for the 1997 model using administrative counts (on figures 5.60, 5.61, and 5.62).

So it seems that the four models have approximately the same number of large deviance contributions. But the 1997 models have the largest deviance contributions, which could explain the fact the the deviance statistic was higher for the two 1997 models than the 1988 and 1992 models.

For the model fitted, we chose the independence working correlation structure. If there is

little difference between a pair of covariance terms (on the upper diagonal of the robust variance-covariance matrix and the same term but on the upper diagonal of the model-based variance-covariance matrix), for all pairs of terms, this will indicate that the choice of the independent correlation structure was correct. The following table provides a grouping of the absolute differences in eight categories.

Table 5.11: Absolute differences, by group, of the robust and the model-based covariance matrices

Percentages	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$0 \leq \text{abs. diff.} \leq 0.01$	90.5	96.1	96.5	96.2
$0.01 < \text{abs. diff.} \leq 0.02$	7.7	2.2	2.0	2.2
$0.02 < \text{abs. diff.} \leq 0.03$	0.4	0.5	0.2	0.2
$0.03 < \text{abs. diff.} \leq 0.04$	0.4	0.1	0.1	0.2
$0.04 < \text{abs. diff.} \leq 0.05$	0.3	0.0	0.2	0.2
$0.05 < \text{abs. diff.} \leq 0.1$	0.8	0.8	0.2	0.8
$0.1 < \text{abs. diff.} \leq 0.2$	0.0	0.4	1.0	0.3
$0.2 < \text{abs. diff.} \leq 0.3$	0.0	0.1	0.0	0.0

Approximately 90% to 96% of the absolute differences are between 0 and 0.01. This seems to indicate that a great proportion of the variance-covariance values on the upper triangle of the robust matrix have a value near the value of the reciprocal term on the upper triangle of the model-based matrix. Thus we can conclude that the correlation structure was correctly specified.

Sparseness in the data is less of a problem when this model was fitted. In the table below, we see that approximately 85% of the cells defined by response j and combination of explanatory variables i have more than five observations per cell.

Table 5.12: Number of observations in the subpopulation by cell

Number of cells with	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
< 1 observation	10	10	11	11
$1 \leq \text{observations} < 5$	6	5	6	6
$5 \leq \text{observations} < 25$	17	14	14	14
$25 \leq \text{observations}$	87	91	89	89
TOTAL	120	120	120	120

In summary, the sparseness in the data caused some problems when it came time to assess the validity of the model. But this problem was less important than when the model in chapter 4 was fitted. We know that the parameters in the model are significant; that the correlation structure can be assumed to be true; and that if we exclude the cell residuals affected by sparseness, there does not seem to be too many large cell residuals. So we can conclude that we can rely on this model.

5.3.3 Design effects

The design effects of the 1988 and the 1992 models were all less than one. In the 1992 model, however, the design effect of the parameter estimate of the interaction between the men aged 30 to 34 years old is 0.99. But for the two 1997 models, there were three design effects greater than one. The greatest design effect was 1.75 for the 1997 model using weighted counts and 1.77 for the 1997 model using the administrative counts, which are not too large design effects. We can conclude, in general, that the sample design chosen improved the precision of the parameter estimates.

5.3.4 Comparison between the two 1997 models

The same test as the one performed in the section 4.4.2 in chapter 4 was applied to the two 1997 models. Recall that the test statistic for the following hypothesis

$$H_0 : \sigma_{1i}^2 = \sigma_{2i}^2$$

$$H_0 : \sigma_{1i}^2 \neq \sigma_{2i}^2$$

(when σ_{1i}^2 is the variance of the i^{th} parameter estimate from the model with weighted counts and σ_{2i}^2 is the variance of the i^{th} parameter estimate from the model with administrative counts, $i = 1, \dots, 11$ when 11 is the number of unequal standard errors) is the ratio of the sample variances

$$F_0 = \frac{S_{1i}^2}{S_{2i}^2}.$$

The null hypothesis is rejected if $F_0 > F_{0.025,10,10}$ or if $F_0 < F_{0.975,10,10}$. It turned out that the null hypothesis has been accepted for all of the 11 unequal standard errors. In conclusion, it seems that the use of weights to obtain population counts by stratum did not provide a significant improvement in the standard errors of the parameter estimates than the use of the original administrative population counts. But we know that when two standard errors of a given parameter estimate are unequal, it is always the one from the model using administrative counts that is the greatest.

5.3.5 Testing of the parameter estimates

The following table gives the levels of the parameters in the four models that are determined to be equal to zero.

Table 5.13: Parameter estimates determined to be equal to zero by the *t*-tests

Levels of the parameter...	Model 1988	Model 1992	Model 1997
INTERCEPT	-	-	-
GENDER	-	-	-
REGION OF RESIDENCE	3, 4	1, 3, 4	3
AGE	-	3	3
GENDER BY REGION	1 by 1	1 by 1	1 by 2
	1 by 2	1 by 2	1 by 3
	1 by 3	1 by 3	1 by 4
	1 by 4	1 by 4	-
GENDER BY AGE	1 by 1	1 by 1	-
REGION BY AGE	1 by 1	1 by 1	1 by 1
	1 by 3	1 by 3	2 by 1
	3 by 1	2 by 1	3 by 1
	4 by 1	2 by 3	3 by 2
	4 by 3	3 by 1	3 by 3
		3 by 2	4 by 1
		3 by 3	4 by 3
		4 by 1	-
		4 by 2	-

5.3.6 Testing of the main effects

The Wald chi-square statistic was used to test different hypotheses about the main effects and the two-way interaction terms of the model. No three-way interaction could be included in the model because it caused problems with the singularity of the matrix of estimable parameters. All of the three main effects and the three interaction terms of the model have

been tested. Recall that when the degrees of freedom associated with the test is large compared to the degrees of freedom available for estimating the variance of regression coefficients (number of observations minus the number of strata), the Wald chi-square is too liberal (Thomas and Rao, 1987). It does not happen here since the number of degrees of freedom available for estimating the variance of regression coefficients is greater than 30,000 and the highest number of degrees of freedom for a test is 29. This is true for the four models. For the 1988 model, at the significance level 5%, none of the interaction terms and none of the main effects were statistically equal to zero. For the 1992 and the 1997 models, at the significance level 5%, two interaction terms and none of the main effects were statistically equal to zero. For the 1992 model, the p -value of the interaction between gender and region of residence is 0.0565. So at 5% significance level, we would accept the assumption that there is no significant interaction between the gender and the region of residence for the 1992 model. For the 1997 model, the p -value of the interaction between gender and region of residence is 0.0589. For the 1997 model (using administrative counts), the p -value of the interaction between gender and region of residence is 0.0703. So at 5% significance level, we would accept the assumption that there is no significant interaction between the gender and the region of residence for the 1997 models. See chapter 6 for further interpretation.

5.3.7 Results of odds ratio analyses

Some odds ratios that have been estimated are:

Table 5.14: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women

REGION	AGE	1988	1992	1997
Atlantic provinces	15-24	1.27	2.86	0.55
	25-29	1.03	1.34	0.95
Quebec	15-24	1.67	2.25	0.65
	25-29	1.35	1.05	1.13
Ontario	15-24	1.49	3.00	0.73
	25-29	1.21	1.40	1.26
The Prairies	15-24	1.16	2.97	0.66
	25-29	0.94	1.39	1.15

For graduates aged from 15 to 24 years old, in every region, the odds of having a master's degree (in this section, when we mention master's degree, we mean master's degree and university diploma above bachelor) or a doctoral degree at graduation for men over women increased from 1988 to 1992 and decreased from 1992 to 1997. In 1992, in every region, the odds are at least two times higher for men of having a master's degree or a doctoral degree at graduation than women (at the age 15-24 years old). But in 1997, again in every region, the men have reduced odds of having a master's degree or a doctoral degree at graduation compared to women (also in the 15 to 24 year-old category). So it seems that the men who graduated in 1986 or in 1990 in any of the Canadian regions were more likely to pursue higher levels of education than women. But the men who graduated in 1995, in any Canadian region, were less likely to pursue higher levels of education than women.

For graduates aged from 25 to 29 years old, in the Atlantic provinces, Ontario and the Prairies, the odds of having a master's degree or a doctoral degree at graduation for men over women increased from 1988 to 1992 and decreased from 1992 to 1997. For graduates aged from 25 to 29 years old, in Quebec, the odds of having a master's degree or a doctoral degree at graduation for men over women decreased from 1988 to 1992 and increased from 1992 to 1997. Except in the Prairies, in 1988, the odds of having a master's degree or a doctoral

degree at graduation in the other regions are always higher than for men (in the 25 to 29 year-old category). Except in the Atlantic provinces, in 1997, the odds of having a master's degree or a doctoral degree at graduation in the other regions are always higher than for men (in the 25 to 29 year-old category). So it seems that in general, the male graduates aged 25 to 29 years old were more likely to attain high levels of education than female graduates aged 25 to 29 years old.

5.4 The effect of the region of institution

A second proportional odds model was fitted using the same explanatory variables as in section 5.3.1, in addition to the region of the institution as a main effect. The response was still the level of certification at the graduation. The levels of each of these variables are the same as the levels of the variables of the proportional odds model fitted in chapter 4. The levels of the region of institution are:

- 1) the Atlantic provinces (Newfoundland; Prince Edward Island; Nova Scotia; New Brunswick)
- 2) Quebec
- 3) Ontario
- 4) the Prairies (Manitoba; Saskatchewan; Alberta)
- 5) British Columbia

5.4.1 Proportional odds model

The estimated model is defined as:

$$\begin{aligned}
\log \left(\frac{\widehat{F}_j(\mathbf{x})}{1 - \widehat{F}_j(\mathbf{x})} \right) = & \widehat{\alpha}_{j=1} + \widehat{\alpha}_{j=2} + \widehat{\beta}_1 GNDN_1 + \widehat{\beta}_2 GNDN_2 + \widehat{\beta}_3 REG_1 + \widehat{\beta}_4 REG_2 \\
& + \widehat{\beta}_5 REG_3 + \widehat{\beta}_6 REG_4 + \widehat{\beta}_7 REG_5 + \widehat{\beta}_8 CATAGE_1 + \widehat{\beta}_9 CATAGE_2 \\
& + \widehat{\beta}_{10} CATAGE_3 + \widehat{\beta}_{11} CATAGE_4 + \widehat{\beta}_{12} REGIN_1 + \widehat{\beta}_{13} REGIN_2 \\
& + \widehat{\beta}_{14} REGIN_3 + \widehat{\beta}_{15} REGIN_4 + \widehat{\beta}_{16} REGIN_5 + \widehat{\beta}_{17} GNDN_1 REG_1 \\
& + \widehat{\beta}_{18} GNDN_1 REG_2 + \widehat{\beta}_{19} GNDN_1 REG_3 + \widehat{\beta}_{20} GNDN_1 REG_4 \\
& + \widehat{\beta}_{21} GNDN_1 REG_5 + \widehat{\beta}_{22} GNDN_2 REG_1 + \widehat{\beta}_{23} GNDN_2 REG_2 \\
& + \widehat{\beta}_{24} GNDN_2 REG_3 + \widehat{\beta}_{25} GNDN_2 REG_4 + \widehat{\beta}_{26} GNDN_2 REG_5 \\
& + \widehat{\beta}_{27} GNDN_1 CATAGE_1 + \widehat{\beta}_{28} GNDN_1 CATAGE_2 + \widehat{\beta}_{29} GNDN_1 CATAGE_3 \\
& + \widehat{\beta}_{30} GNDN_1 CATAGE_4 + \widehat{\beta}_{31} GNDN_2 CATAGE_1 + \widehat{\beta}_{32} GNDN_2 CATAGE_2 \\
& + \widehat{\beta}_{33} GNDN_2 CATAGE_3 + \widehat{\beta}_{34} GNDN_2 CATAGE_4 + \widehat{\beta}_{35} REG_1 CATAGE_1 \\
& + \widehat{\beta}_{36} REG_1 CATAGE_2 + \widehat{\beta}_{37} REG_1 CATAGE_3 + \widehat{\beta}_{38} REG_1 CATAGE_4 \\
& + \widehat{\beta}_{39} REG_2 CATAGE_1 + \widehat{\beta}_{40} REG_2 CATAGE_2 + \widehat{\beta}_{41} REG_2 CATAGE_3 \\
& + \widehat{\beta}_{42} REG_2 CATAGE_4 + \widehat{\beta}_{43} REG_3 CATAGE_1 + \widehat{\beta}_{44} REG_3 CATAGE_2 \\
& + \widehat{\beta}_{45} REG_3 CATAGE_3 + \widehat{\beta}_{46} REG_3 CATAGE_4 + \widehat{\beta}_{47} REG_4 CATAGE_1 \\
& + \widehat{\beta}_{48} REG_4 CATAGE_2 + \widehat{\beta}_{49} REG_4 CATAGE_3 + \widehat{\beta}_{50} REG_4 CATAGE_4 \\
& + \widehat{\beta}_{51} REG_5 CATAGE_1 + \widehat{\beta}_{52} REG_5 CATAGE_2 + \widehat{\beta}_{53} REG_5 CATAGE_3 \\
& + \widehat{\beta}_{54} REG_5 CATAGE_4
\end{aligned} \tag{5.2}$$

5.4.2 Validation of the model

The first goodness-of-fit statistic that was computed was the Pearson chi-squared statistic, for the four models.

Table 5.15: Pearson chi-squared statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
2,754.69	3,796.70	3,261.54	3,261.54

These four chi-squared statistics take large values. If we compare the 1988 statistic to $\chi^2_{380-56(0.95)} = 366.70$, when 56 is the number of parameter estimates in the model and 380 is

the product of the number of levels of the response minus one ($= 2$) with the number of combinations of levels of the explanatory variables ($= 190$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1992 statistic to $\chi^2_{392-56(0.95)} = 379.46$, when 56 is the number of parameter estimates in the model and 392 is the product of the number of levels of the response minus one ($=2$) by the number of combinations of levels of the explanatory variables ($= 196$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1997 statistics to $\chi^2_{366-56(0.95)} = 351.78$, when 56 is the number of parameter estimates in the model and 366 is the product of the number of levels of the response minus one ($= 2$) by the number of combinations of levels of the explanatory variables ($= 183$), we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Recall that this is not a test corrected for the sampling plan and note that when $\hat{\pi}_{ij} = 0$, the calculation of the contribution to the Pearson's statistic could not be performed, since there was a zero in the denominator.

The four deviance functions have also been computed:

Table 5.16: Deviance statistics

NGS 88	NGS 92	NGS 97	NGS 97 (admin. counts)
1,828.32	2,650.24	2,998.18	2,998.18

These four deviance functions take large values. If we compare the 1988 statistic to $\chi^2_{380-56(0.95)} = 366.70$, we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1992 statistic to $\chi^2_{392-56(0.95)} = 379.46$, we reject the null hypothesis, i.e. the hypothesis that the fitted proportional odds model is correct, at the significance level 5%. Then if we compare the 1997 statistics to $\chi^2_{366-56(0.95)} = 351.78$, we reject the null hypothesis, i.e. the

hypothesis that the fitted proportional odds model is correct, at the significance level 5%. When $\widehat{\pi}_{ij} = 0$ or $\widetilde{\pi}_{ij} = 0$, the calculation of the contribution to the deviance function could not be performed, since $\log(0)$ does not exist.

The Wald chi-square has been used to test the null hypothesis that all of the parameter estimates in the model are equal to zero. This will test simultaneously that every of the 56 parameters are equal to zero. For all of the four models, this null hypothesis is strongly rejected, indicating that the proportional odds model fits better than a model with all coefficients equal to zero. See results below:

Table 5.17: Wald chi-square statistics

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Wald statistic	674,137.32	866,030.67	905,232.58	711,488.84
<i>p</i> -value	0.0000	0.0000	0.0000	0.0000

The cell residuals have been computed. Once they have been computed, four groups have been created and the number of cells included in each of these groups was calculated.

Table 5.18: Frequencies of cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	246(43%)	274(47%)	253(46%)	253(46%)
$-10 \leq \text{cell res.} < -2$ or $2 < \text{cell res.} \leq 10$	183	179	169	169
cell res. < -10 or cell res. > 10	2	8	4	4
cell residual is missing	139	127	123	123

Approximately 45% of the cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not be

computed (either when $\hat{\pi}_{ij} = 1$ or $\hat{\pi}_{ij} = 0$). If we delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 5.19: Proportions of cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.counts)
$-2 \leq \text{cell res.} \leq 2$	57.1%	59.4%	59.4%	59.4%

These proportions seem reasonable. There does appear to be a cell residual with an abnormally large value (see table below). If we compare the maximum and minimum values of the four models, we see that they are roughly in the same range of values.

Table 5.20: Maximum and minimum cell residuals

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
Max. cell residual	11.16	13.85	11.07	11.07
Min. cell residual	-8.68	-13.22	-12.46	-12.46

The cumulative cell residuals have also been computed, using cumulative totals and cumulative fitted probabilities. Once they have been computed, five groups have been created and the number of cells included in each of these groups was calculated.

Table 5.21: Frequencies of cumulative cell residuals by group

	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$-2 \leq \text{cum.cell res.} \leq 2$	169(30%)	197(34%)	193(35%)	193(35%)
$-10 \leq \text{cum. cell res.} < -2$ or $2 < \text{cum. cell res.} \leq 10$	108	112	89	89
$-100 \leq \text{cum. cell res.} < -10$ or $10 < \text{cum. cell res.} \leq 100$	1	3	3	3
$\text{cum. cell res.} < -100$ or $\text{cum. cell res.} > 100$	8	11	14	14
cum. cell res. is missing	284	265	250	250

Approximately 30% of the cumulative cell residuals for each models are between -2 and 2 . But we see that there is a relatively large number of cells for which the cell residual could not been computed (either when the cumulative probability was equal to zero or one). If we delete all of the cells for which the cell residual is missing, the proportion of cells with cell residuals between -2 and 2 is:

Table 5.22: Proportions of cumulative cell residuals (excluding missing residuals) between -2 and 2

	NGS 88	NGS 92	NGS97	NGS 97 (admin.)
$-2 \leq \text{cell res.} \leq 2$	59.1%	61.0%	64.5%	64.5%

These proportions are about 30% higher than when the missing cumulative cell residuals were included in the counts. These counts are comparable to those obtained for the cell residuals.

A plot of each of the contributions to the deviance function against its index (i, j) was obtained in order to detect large deviance contributions (see Appendix G). For each model, a

plot of the deviance contributions corresponding to response level $j, j = 1, 2, 3$, for all of the I combinations of explanatory variables has been drawn. On Figure 5.63, for $Y = 1$, in 1988, there does not seem to be any deviance contributions that is large. The deviance contributions on the graph seem to be between -50 and 150 . On Figure 5.64, for $Y = 2$, in 1988, there were four deviance contributions that seemed to be large. These deviance contributions are approximately equal to 200 and -200 (they represent the female graduates aged 25 to 29 years old living in Ontario who studied in Ontario; and the male graduates aged 30 to 34 years old living in Ontario who studied in Ontario; and it represents the male graduates aged 35 years old and over living in Ontario who studied in Ontario; and the female graduates aged 35 years old and over living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -100 and 100 . On Figure 5.65, for $Y = 3$, in 1988, there was one deviance contribution that seemed to be large. This deviance contribution is approximately equal to 200 (it represents the male graduates aged 35 years old and over living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -50 and 80 . So for the 1988 model, only five deviance contributions (for $Y = 2$ and $Y = 3$) seem to be large.

On Figure 5.66, for $Y = 1$, in 1992, there was one deviance contribution that seemed to be large. This deviance contribution is approximately equal to 300 (it represents the female graduates aged 25 to 29 years old living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -100 and 200 . On Figure 5.67, for $Y = 2$, in 1992, there were two deviance contributions that seemed to be large. These deviance contributions are between -250 and -200 (they represent the male graduates aged 30 to 34 years old living in Ontario who studied in Ontario; and the male graduates aged 35 years old and over living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -100 and 150 . On Figure 5.68, for $Y = 3$, in 1992, there does not seem to have any large deviance contribution. The deviance contributions on the graph seem to be between -100 and 100 . So for the 1992 model, only

three deviance contributions (for $Y = 1$ and $Y = 2$) seem to be large.

On Figure 5.69, for $Y = 1$, in 1997, there was one deviance contribution that seemed to be large. This deviance contribution was approximately equal to 300 (it represents the male graduates aged 35 years old and over living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be roughly between -100 and 200 . On Figure 5.70, for $Y = 2$, in 1997, there were three deviance contributions that seemed to be large. One had a value approximately equal to -450 (it represents the male graduates aged 35 years old and over living in Ontario who studied in Ontario), the other had a value approximately equal to -300 (it represents the male graduates aged 30 to 34 years old living in Ontario who studied in Ontario) and the last one had a value approximately equal to 350 (it represents the male graduates aged 25 to 29 years old living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -200 and 200 . On Figure 5.71, for $Y = 3$, in 1997, there were three deviance contributions that seemed to be large. They have a deviance contribution between 200 and 300 (they represent the male graduates aged 35 years old and over living in Quebec who studied in Quebec; and the male graduates aged 30 to 34 years old living in Ontario who studied in Ontario; and the male graduates aged 35 years old and over living in Ontario who studied in Ontario). The other deviance contributions on the graph seem to be between -100 and 150 . So for the 1997 model, seven deviance contributions seem to be large. The same results applied for the 1997 model using administrative counts (on figures 5.72, 5.73, and 5.74).

So it seems that the 1997 models had a larger number of deviance contributions that are large than the 1988 model and the 1992 model. We also saw previously that the deviance statistics for the 1997 models were higher than the deviance statistics for the 1988 model and the 1992 model. This can be explained by the high number of large contributions to the deviance in these two models.

For the model fitted, we chose the independence working correlation structure. If there is

little difference between a pair of covariance term (on the upper diagonal of the robust variance-covariance matrix and the same term but on the upper diagonal of the model-based variance-covariance matrix), for all pairs of terms, this will indicate that the choice of the independent correlation structure was correct. The following table provides a grouping of the absolute differences in eight categories.

Table 5.23: Absolute differences, by group, of the robust and the model-based covariance matrices

Percentages	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
$0 \leq \text{abs. diff.} \leq 0.01$	87.3	94.7	95.4	95.1
$0.01 < \text{abs. diff.} \leq 0.02$	10.1	3.6	2.3	2.8
$0.02 < \text{abs. diff.} \leq 0.03$	1.1	0.6	0.3	0.3
$0.03 < \text{abs. diff.} \leq 0.04$	0.6	0.1	0.5	0.5
$0.04 < \text{abs. diff.} \leq 0.05$	0.1	0.0	0.3	0.3
$0.05 < \text{abs. diff.} \leq 0.1$	0.8	0.6	0.1	0.7
$0.1 < \text{abs. diff.} \leq 0.2$	—	0.3	1.0	0.4
$0.2 < \text{abs. diff.} \leq 0.3$	—	0.1	—	—

Approximately 87% to 95% of the absolute differences are between 0 and 0.01. This seems to indicate that a great proportion of the variance-covariance values on the upper triangle of the robust matrix have a value near the value of the reciprocal term on the upper triangle of the model-based matrix. Thus we can conclude that the correlation structure was correctly specified.

Sparseness in the data caused some problems when this model was fitted. In the table below, we see that approximately 30% of the cells defined by response j and combination of explanatory variables i have less than one observation by cell.

Table 5.24: Number of observations in the subpopulation by cell

Number of cells with	NGS 88	NGS 92	NGS 97	NGS 97 (admin.)
< 1 observation	160	162	188	188
$1 \leq \text{observations} < 5$	114	95	99	99
$5 \leq \text{observations} < 25$	120	130	124	124
$25 \leq \text{observations}$	176	201	138	138
TOTAL	570	588	549	549

In summary, the sparseness in the data caused some problems when it came time to assess the validity of it. But this problem was less important than when the model in chapter 4 was fitted. We know that the parameters in the model are significant; that the correlation structure can be assumed to be true; and that, if we exclude the cell residuals affected by sparseness, there does not seem to be too many large cell residuals. So we can rely on these results.

5.4.3 Design effects

The design effects of the 1988 model were all less than one. In the 1992 model, the design effect of the parameter estimate of the interaction between the men aged 30 to 34 years old is 1.02. But for the two 1997 models, there were five design effects greater than one. The greatest design effect was 1.76 for the 1997 model using weighted counts and 1.78 for the 1997 model using the administrative counts, which are not too large design effects. We can conclude, in general, that the sample design chosen improved the precision of the parameter estimates.

5.4.4 Comparison between the two 1997 models

The same test as the one performed in the section 4.4.2 in chapter 4 was applied to the two 1997 models. Recall that the test statistic for the following hypothesis

$$H_0 : \sigma_{1i}^2 = \sigma_{2i}^2$$

$$H_0 : \sigma_{1i}^2 \neq \sigma_{2i}^2$$

(for $i = 1, \dots, 13$) is the ratio of the sample variances

$$F_0 = \frac{S_{1i}^2}{S_{2i}^2}.$$

The null hypothesis is rejected if $F_0 > F_{0.025,12,12}$ or if $F_0 < F_{0.975,12,12}$. It turned out that the null hypothesis has been accepted for all of the 13 unequal standard errors. In conclusion, it seems that the use of weights to obtain population counts by stratum did not provide a significant improvement in the standard errors of the parameter estimates than the use of the original administrative population counts. But we know that when two standard errors of a given parameter estimate are unequal, it is always the one from the model using administrative counts that is the greatest.

5.4.5 Testing of the parameter estimates

The following table gives the levels of the parameters in the four models that are determined to be equal to zero.

Table 5.25: Parameter estimates determined to be equal to zero by the *t*-tests

Levels of the parameter...	Model 1988	Model 1992	Model 1997
INTERCEPT	-	-	-
GENDER	-	-	-
REGION OF RESIDENCE	1, 3, 4	1, 2, 4	3
AGE	-	3	3
REGION OF THE INSTITUTION	2, 3	1	2, 3, 4
GENDER BY REGION	1 by 1	1 by 1	1 by 2
	1 by 2	1 by 2	1 by 3
	1 by 3	1 by 3	1 by 4
	1 by 4	1 by 4	-
GENDER BY AGE	1 by 1	1 by 1	-
REGION BY AGE	1 by 1	1 by 1	1 by 1
	1 by 3	1 by 3	2 by 1
	3 by 1	2 by 1	3 by 1
	4 by 1	2 by 2	3 by 2
	4 by 2	2 by 3	3 by 3
	4 by 3	3 by 1	4 by 1
	-	3 by 2	4 by 3
	-	3 by 3	-
	-	4 by 1	-
	-	4 by 2	-

5.4.6 Testing of the main effects

The Wald chi-square statistic was used to test different hypotheses about the main effects and the two-way interaction terms of the model. All of the four main effects and the three two-way interaction terms of the model have been tested. Recall that when the degrees of freedom associated with the test is large compared to the degrees of freedom available for estimating the variance of regression coefficients (number of observations minus the number of strata), the Wald chi-square is too liberal (Thomas and Rao, 1987). It does not happen here since the number of degrees of freedom available for estimating the variance of

regression coefficients is greater than 30,000 and the highest number of degrees of freedom for a test is 33. This is true for the four models. For the 1988 model, at the 5% level of significance, none of the three interaction terms and none of the main effects were statistically equal to zero. For the 1992 and the 1997 models, at the significance level 5%, two interaction terms and none of the main effects were statistically equal to zero. For the 1992 model, the p -value of the interaction between gender and region of residence is 0.0724. So at 5% significance level, we would accept the assumption that there is no significant interaction between the gender and the region of residence for the 1992 model. For the 1997 model, the p -value of the interaction between gender and region of residence is 0.0868. For the 1997 model (using administrative counts), the p -value of the interaction between gender and region of residence is 0.1005. So at 5% significance level, we would accept the assumption that there is no significant interaction between the gender and the region of residence for the 1997 models.

5.4.7 Results of odds ratio analyses

Some odds ratios that have been estimated are:

Table 5.26: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women for a fixed region of institution

REGION	AGE	1988	1992	1997
Atlantic provinces	15-24	1.25	2.86	0.55
	25-29	1.02	1.32	0.96
Quebec	15-24	1.67	2.27	0.65
	25-29	1.36	1.05	1.13
Ontario	15-24	1.49	3.03	0.73
	25-29	1.22	1.40	1.26
The Prairies	15-24	1.16	2.94	0.66
	25-29	0.95	1.36	1.14

For graduates aged from 15 to 24 years old, in every region, the odds of having a master's degree (in this section, when we mention master's degree, we mean master's degree and university diploma above bachelor) or a doctoral degree at graduation for men over women increased from 1988 to 1992 and decreased from 1992 to 1997. In 1992, in every region, the odds are at least two times higher for men of having a master's degree or a doctoral degree at graduation than women (at the age 15-24 years old). But in 1997, again in every region, the men have reduced odds of having a master's degree or a doctoral degree at graduation compared to women (also in the 15 to 24 year-old category). So it seems that the men who graduated in 1986 or in 1990 in any of the Canadian region were more likely to pursue higher levels of education than women. But the men who graduated in 1995, in any Canadian region, were less likely to pursue higher levels of education than women.

For graduates aged from 25 to 29 years old, in the Atlantic provinces, Ontario and the Prairies, the odds of having a master's degree or a doctoral degree at graduation for men over women increased from 1988 to 1992 and decreased from 1992 to 1997. For graduates aged from 25 to 29 years old, in Quebec, the odds of having a master's degree or a doctoral degree at graduation for men over women decreased from 1988 to 1992 and increased from 1992 to 1997. Except in the Prairies, in 1988, the odds of having a master's degree or a doctoral degree at graduation in the other regions are always higher than for men (in the 25 to 29 year-old category). Except in the Atlantic provinces, in 1997, the odds of having a master's degree or a doctoral degree at graduation in the other regions are always higher than for men (in the 25 to 29 year-old category). So it seems that in general, the male graduates aged 25 to 29 years old were more likely to attain high levels of education than female graduates aged 25 to 29 years old.

5.5 Comparison of two proportional odds models

The model in section 5.3.1 was fitted in order to obtain estimated conditional probabilities

that will be used in the projections. The model in section 5.4.1 was fitted in order to see if there would be much difference in the odds ratios if the region of the institution would be added to the model. The estimated odds ratios have been obtained in sections 5.3.7 and 5.4.7. Model 1 refers to the model defined in section 5.3.1 and model 2 refers to the model defined in section 5.4.1 (with the region of institution fixed).

Table 5.27: Estimated odds of having response $> j$ ($j = 1, 2$) for men over women

REGION	AGE	1988	1988	1992	1992	1997	1997
		Model 1	Model2	Model 1	Model 2	Model 1	Model 2
Atlantic prov.	15-24	1.27	1.25	2.86	2.86	0.55	0.55
	25-29	1.03	1.02	1.34	1.32	0.95	0.96
Quebec	15-24	1.67	1.67	2.25	2.27	0.65	0.65
	25-29	1.35	1.36	1.05	1.05	1.13	1.13
Ontario	15-24	1.49	1.49	3.00	3.03	0.73	0.73
	25-29	1.21	1.22	1.40	1.40	1.26	1.26
The Prairies	15-24	1.16	1.16	2.97	2.94	0.66	0.66
	25-29	0.94	0.95	1.39	1.36	1.15	1.14

For both models, we draw the same conclusions, i.e. if there is an increase in the odds ratio for model 1 between two years, there will also be an increase in the same odds ratio for model 2 between the same two years (same thing for a decrease in an odds ratio). If an odds ratio has a value greater than one for model 1, the corresponding odds ratio for model 2 will also be greater than one (same thing for an odds ratio lower than one). In fact, the corresponding odds ratios for model 1 and model 2 have almost exactly the same value. Sometimes they have the exact same value; sometimes the odds ratio for model 1 is a little lower than the corresponding odds ratio for model 2 (mostly in 1988); and sometimes the odds ratio for model 1 is a little greater than the corresponding odds ratio for model 2 (mostly in 1997). But the maximum absolute difference between two corresponding odds ratios is 0.03, which is low.

From the Wald chi-squared statistics that tested the main effects and the interaction effects, we can also say that the effects of the region of residence, the gender and the age are significant in both models, in 1988, 1992 and 1997. For both models, the three interactions were significant in 1988. In 1992 and 1997, the interaction between the region of residence and gender was determined to be nonsignificant, for model 1 and model 2. But in model 2, the p -values of this interaction in 1992 and 1997 were a little higher than the p -values for model 1. So we can say that the addition of the region of institution in the proportional odds model makes us accept the assumption of no interaction between the region of residence and gender a little more strongly. But the conclusions did not change.

So we can conclude that even if the region of the institution is included in the model, the value of the odds ratios are not highly affected. We draw the same conclusions about the testing of the main effects and the three interactions of the model, whether or not the region of institution is included in the model.

In general, students enrolled in one province will also live in the same province. But it can happen that, for example, a student is enrolled in Ontario (Ottawa) but lives in Aylmer (Quebec). For these particular cases, it was important to know if the region of the institution has an effect on the logit of the cumulative conditional probability of obtaining a bachelor's degree or a master's degree (or the equivalent). We know from the Wald chi-squared test that the region of institution has an effect but we also saw that the odds ratios are not highly affected when this variable is used as an explanatory variable in the model. So it is right not to include this variable in the proportional odds model, and consequently, not use it for the projections.

5.6 Estimation of the number of Canadian university teachers 1998-2018

5.6.1 The number of university teachers after retirement and attrition

The estimation of the number of university teachers will be done in three main steps. In the

first step, the University and College Academic Staff System (UCASS) is used. This is an annual survey, designed to collect information on the characteristics (e.g. salaries, age and social characteristics) of full-time teaching staff employed in degree-granting institutions in Canada. Full-time includes:

1. Staff appointed on a full-time basis whose term of appointment is not less than twelve months (include any staff member on leave).
2. New appointees hired on a full-time basis whose term of contract is twelve months but for the first year he/she can only be at the institution for less than twelve months.
3. Staff who were appointed to teach full-time (12 months or more) and at a later date have entered into a formal agreement with the institution to work on a reduced load basis.

The situation usually arises with staff member who are approaching retirement.

Then the teaching staff includes:

1. All teachers within faculties, whether or not they hold an academic rank.
2. Academic staff in teaching hospitals.
3. Visiting academic staff in faculties.
4. Research staff who have an academic rank and a salary scale similar to teaching staff.

The number of full-time university teachers by age category is known for 1998. These staff members are not in a faculty of medicine or dentistry. Under the assumption that normal retirement will occur at 65 years old, we can know the year at which these teachers will retire.

Table 5.28: Year of retirement of 1998 full-time university teachers

Age in 1998	Year of retirement
Under 30	2034 and over
30-34	2029-2033
35-39	2024-2028
40-44	2019-2023
45-49	2014-2018
50-54	2009-2013
55-59	2004-2008
60 and over	2003 and before

If we want the estimated number of full-time university teachers that will be left in Canada after the retirement, for the single years from 1998 to 2018, here is what was done. We know the number of teachers that will be left for certain years after normal (i.e. age 65) retirement (see table below).

Table 5.29: Number of full-time university teachers after normal retirement

Year	Number of full-time university teachers after retirement
1998	26,763
2004	23,658
2009	18,252
2014	12,822
2019	8,346

Source: UCASS, Statistics Canada

Then we assume a constant retirement rate between two of these years that are consecutive. For example, between 1998 and 2004, we subtract 23,658 from 26,763 and we divide the difference by 6 in order to obtain the retirement rate per year between 1998 and 2004. We then obtain:

Table 5.30: Number of full-time university teachers after normal retirement from 1998 to 2018

Year	Number of full-time university teachers after retirement
1998	26,763
1999	26,246
2000	25,728
2001	25,211
2002	24,693
2003	24,176
2004	23,658
2005	22,577
2006	21,496
2007	20,414
2008	19,333
2009	18,252
2010	17,166
2011	16,080
2012	14,994
2013	13,908
2014	12,822
2015	11,927
2016	11,032
2017	10,136
2018	9,241

Source: UCASS, Statistics Canada

Then from this number, we can subtract a certain number of teachers, due to what we call attrition (deaths, left teaching for another job, etc). We will make the estimations using an attrition rate of 4%.

Table 5.31: Number of full-time university teachers affected by the attrition from 1998 to 2018

Year	Attrition
1998	1,071
1999	1,050
2000	1,029
2001	1,008
2002	988
2003	967
2004	946
2005	903
2006	860
2007	817
2008	772
2009	730
2010	687
2011	643
2012	600
2013	556
2014	513
2015	477
2016	441
2017	405
2018	370

Source: Table 5.30

We can now obtain, at a 4% attrition rate, the number of full-time teachers that will be left, after normal retirement and attrition.

Table 5.32: Number of full-time university teachers after retirement and attrition from 1998 to 2018

Year	Number of full-time university teachers after retirement and attrition	Year	Number of full-time university teachers after retirement and attrition
1998	25,692	2009	17,522
1999	25,196	2010	16,479
2000	24,699	2011	15,437
2001	24,203	2012	14,394
2002	23,705	2013	13,352
2003	23,209	2014	12,309
2004	22,712	2015	11,450
2005	21,674	2016	10,591
2006	20,636	2017	9,731
2007	19,597	2018	8,871
2008	18,561		

Source: Tables 5.30 and 5.31

5.6.2 Estimated number of potential university teachers in the 1996 Canadian population

From the 1996 Census, we know the number of people in Canada whose highest degree, certificate or diploma obtained was the doctorate. These counts are available by gender and age categories (#1). For people having a doctorate in 1996, for a given age category, we make the assumption that if they become full-time university teachers, it is going to be by the end of 2000. But in 1996, there were a certain number of full-time university teachers in the same age group (#2). If we make the assumption that all of the university teachers in this age group had a doctorate, we can subtract (#2) from (#1) in order to obtain the number of people, in 1996, in Canada, who had a doctorate but were not university teachers, for a given age group. The approximate number of people with a doctorate in 1996 who are not full-time

university teachers was obtained for two age categories: the 15 to 24 year-old category and the 25 to 34 year-old category. This number is an approximation since the number of full-time university teachers in 1996 was unknown for the 15 to 24 year-old category and the 25 to 34 year-old category. The number of full-time university teachers in 1996 was known for those aged 29 years old and less and those in the 30 to 34 year-old category. So we made the assumption that the number of people who are full-time university teachers in 1996 (who were 15-24 years old) is not so different than the number of people who are full-time university teachers in 1996 (who were 29 years old and less). And we assumed that the number of people who are full-time university teachers in 1996 (who were 25-34 years old) is not very different from the number of people who are full-time university teachers in 1996 (who were 30-34 years old).

Table 5.33: Estimated number of Canadians with a doctorate who were not full-time university teachers in 1996 (15-24)

In 1996...	Count
Canadians with a doctorate (15 to 24 years old)	275
Full-time university teachers in Canada (29 years old and less)	163
Canadians with a doctorate not full-time univ. teachers (15-24)	112

Source: 1996 Census CD-ROM Products and UCASS, Statistics Canada

Table 5.34: Estimated number of Canadians with a doctorate who were not full-time university teachers in 1996 (25-34)

In 1996	Count
Canadians with a doctorate (25 to 34 years old)	12,375
Full-time university teachers in Canada (30 to 34 years of age)	1,441
Canadians with a doctorate not full-time univ. teachers (25-34)	10,934

Source: 1996 Census CD-ROM Products and UCASS, Statistics Canada

But not all of these 112 people aged 15 to 24 years old and 10,934 people aged 25 to 34 years old in 1996 will become university teachers. So, also from the 1996 Census data, we computed the proportion of people (for all ages) having a doctorate in Canada who are a full-time university teacher. The number of full-time university teachers from the 1996 Census was obtained for the year 1995 and not the year 1996. We assume that at the time of the 1996 Census, there was not a large difference in the number of full-time university teachers in 1995 and the same number in 1996.

Table 5.35: Proportion of people with a doctoral degree who are full-time university teachers in 1996

Proportion of people with a doctorate who are full-time univ. teach.	30.2%
--	-------

Source: 1996 Census CD-ROM Products, Statistics Canada

Now, we can apply this proportion to the number of people in Canada having a doctorate but who are not university teachers. We obtain the estimated number of people in Canada having a doctorate and who are 15 to 24 years of age or 25 to 34 years of age, and who could become university teachers by the end of 2000.

Table 5.36: Number of Canadians with a doctorate in 1996 who could become full-time university teachers by the end of 2000

Number of Canadians (15-24) who could become university teachers by the end of 2000	34
Number of Canadians (25-34) who could become university teachers by the end of 2000	3,302
TOTAL	3,336

5.6.3 Estimation of the number of Canadian doctoral graduates 1998-2018

From the proportional odds model in section 5.3.1, we obtained probabilities that the university level of certification obtained at graduation in 1995 is either :

- 1) Bachelor's degree; first professional degree; university diploma below bachelor
- 2) Master's degree; university diploma above bachelor
- 3) Earned Ph.D.

given the age of the graduate at the time of the NGS; the Canadian region of residence of the graduate at the time of the NGS; and the graduate's gender. These conditional probabilities are used to estimate the number of doctoral graduates from 1998 to 2018. Only the number of doctoral graduates aged 34 years old and less is estimated because we made the assumption that the doctoral graduates in this age category are more likely to be hired as university teachers for the first time than older doctoral graduates. A comparison of the estimates of the doctoral graduates aged 34 years old and less and the doctoral graduates (for all age categories) was made in section 5.6.3.4.

5.6.3.1 Estimated proportion of university graduates 1986-1995

The first thing that was done has been to obtain the proportion of university graduates in 1986, 1990, and 1995, in Canada, by region, age (34 years old and less), and gender. From the 1988, 1992, and 1997 NGS, the number of Canadian university graduates has been estimated by region, age, and gender. Then, the population estimates of people in Canada by region, age, and gender was obtained from Statistics Canada. If we divide the number of Canadian university graduates by the Canadian population estimates by region, age, and gender, in 1986, 1990, and 1995 respectively, we obtain the proportion of university graduates in 1986, 1990, and 1995 in the population.

Table 5.37: Proportion of university graduates by region and gender in 1986, 1990, and 1995

Region	Gender	Age	1986	1990	1995
Atlantic provinces	Male	15-24	0.007248	0.007143	0.007472
Atlantic provinces	Male	25-29	0.013226	0.016705	0.026898
Atlantic provinces	Male	30-34	0.004173	0.004214	0.005218
Atlantic provinces	Female	15-24	0.011239	0.012251	0.015114
Atlantic provinces	Female	25-29	0.011571	0.019174	0.030617
Atlantic provinces	Female	30-34	0.003600	0.005687	0.005928
Quebec	Male	15-24	0.005203	0.003913	0.005738
Quebec	Male	25-29	0.024233	0.021013	0.027866
Quebec	Male	30-34	0.008234	0.005602	0.008722
Quebec	Female	15-24	0.008420	0.008018	0.012893
Quebec	Female	25-29	0.020061	0.023618	0.043944
Quebec	Female	30-34	0.010454	0.005579	0.010539
Ontario	Male	15-24	0.007050	0.005677	0.005709
Ontario	Male	25-29	0.026901	0.026313	0.034604
Ontario	Male	30-34	0.005543	0.005911	0.005779
Ontario	Female	15-24	0.011361	0.009882	0.012911
Ontario	Female	25-29	0.022703	0.026883	0.041012
Ontario	Female	30-34	0.005236	0.005490	0.005870
Prairies	Male	15-24	0.005832	0.006148	0.005705
Prairies	Male	25-29	0.014829	0.017209	0.026229
Prairies	Male	30-34	0.005530	0.005924	0.006553
Prairies	Female	15-24	0.009291	0.008709	0.010809
Prairies	Female	25-29	0.012895	0.015374	0.027606
Prairies	Female	30-34	0.004465	0.005723	0.006364
BC/Yukon/NWT	Male	15-24	0.003701	0.004622	0.003706
BC/Yukon/NWT	Male	25-29	0.013176	0.016370	0.019934
BC/Yukon/NWT	Male	30-34	0.004730	0.005134	0.005251
BC/Yukon/NWT	Female	15-24	0.005187	0.006680	0.007379
BC/Yukon/NWT	Female	25-29	0.012052	0.017658	0.022991
BC/Yukon/NWT	Female	30-34	0.003981	0.004514	0.004667

Source: 1988 NGS, 1992 NGS, 1997 NGS, and Statistical Reference Centre, Statistics Canada

5.6.3.2 Estimated number of university graduates from 1998 to 2018 in the population

The proportions of university graduates by region, age (34 years of age and less), and gender from section 5.6.3.1 have been multiplied by the population projections by region, age (34 years of age and less), and gender, from 1998 to 2018. There are four population projections assumptions and the number of university graduates from 1998 to 2018 by region, age, and gender have been obtained for these four assumptions.

- 1) Low growth, low fertility (1.3 by 2026), low life expectancy m: 78.5/ f: 83.0 by 2026, low immigration (150,000 in 2026), central internal migration.
- 2) Medium growth, medium fertility(1.5 by 2026), medium life expectancy m: 80.0/f: 84.0 by 2026, medium immigration(210,000 in 2026), medium internal migration.
- 3) High growth,high fertility (1.8 by 2026), high life expectancy m: 81.5/f:85.0, high immigration (270,000 in 2026),west internal migration
- 4) High growth, high fertility (1.8 by 2026), high life expectancy m: 81.5/f:85.0, high immigration (270,000 in 2026), central internal migration

**Table 5.38: Population projections from 1998 to 2018
(34 years old and less)**

Year	No assumption	Ass. 1	Ass. 2	Ass. 3	Ass. 4
1998	8,636,988	–	–	–	–
1999	8,565,032	–	–	–	–
2000	–	8,518,845	8,523,122	8,529,503	8,529,473
2001	–	8,502,815	8,514,787	8,530,560	8,530,485
2002	–	8,502,648	8,525,569	8,553,606	8,553,568
2003	–	8,507,939	8,545,093	8,588,243	8,588,157
2004	–	8,523,338	8,577,785	8,638,780	8,638,707
2005	–	8,546,812	8,621,581	8,703,037	8,702,995
2006	–	8,568,779	8,666,750	8,771,210	8,771,171
2007	–	8,605,503	8,725,941	8,852,709	8,852,714
2008	–	8,639,191	8,781,461	8,929,863	8,929,848
2009	–	8,672,499	8,835,906	9,005,248	9,005,268
2010	–	8,692,500	8,876,429	9,066,134	9,066,156
2011	–	8,707,889	8,911,805	9,121,343	9,121,354
2012	–	8,698,607	8,921,943	9,150,855	9,150,915
2013	–	8,682,993	8,925,400	9,173,321	9,173,443
2014	–	8,646,507	8,907,735	9,174,434	9,174,630
2015	–	8,592,393	8,876,272	9,173,325	9,173,570
2016	–	8,530,278	8,838,288	9,168,908	9,169,197
2017	–	8,468,229	8,801,846	9,167,975	9,168,330
2018	–	8,401,483	8,761,957	9,165,586	9,166,060

Source: Demography Division, Statistics Canada

Before 2000, the population estimates were available, and not the population projections. That is why the population estimates are not given for four assumptions. The 1995 proportions obtained in section 5.6.3.1 have been retained to estimate the number of university graduates in future. We assume that the proportion of university graduates by region, age, and gender will be the same from 1998 to 2018. The estimated number of

university graduates from 1998 to 2018, by assumption, are in the table below.

Table 5.39: Estimated number of university graduates from 1998 to 2018

Year	No assumption	Ass. 1	Ass. 2	Ass. 3	Ass. 4
1998	120,882	–	–	–	–
1999	119,861	–	–	–	–
2000	–	119,329	119,378	119,461	119,496
2001	–	119,045	119,187	119,389	119,473
2002	–	119,287	119,558	119,906	120,065
2003	–	119,638	120,082	120,614	120,846
2004	–	120,224	120,882	121,634	121,938
2005	–	120,732	121,652	122,664	123,024
2006	–	121,231	122,452	123,757	124,170
2007	–	121,720	123,224	124,806	125,270
2008	–	122,280	124,052	125,900	126,410
2009	–	122,775	124,805	126,907	127,462
2010	–	123,063	125,344	127,694	128,290
2011	–	123,252	125,780	128,374	129,009
2012	–	123,207	125,978	128,816	129,488
2013	–	122,997	126,013	129,094	129,800
2014	–	122,841	126,102	129,427	130,169
2015	–	122,848	126,389	130,061	130,841
2016	–	122,818	126,654	130,702	131,521
2017	–	122,864	127,009	131,449	132,310
2018	–	122,711	127,177	132,028	132,931

Source: Based on tables 5.37 and 5.38

5.6.3.3 Estimated number of doctoral graduates from 1998 to 2018 in the population

The number of of university graduates from 1998 to 2018 by region, age, and gender

obtained for the four assumptions has been multiplied by the 1995 conditional probabilities (described at the beginning of the section 5.6.3.) of obtaining a doctorate at the university graduation. These probabilities have been estimated from the subpopulation of Canadian university graduates, and not the canadian population. For comparison purposes, the number of graduates obtaining a doctorate have also been obtained from 1992 to 1997. Note that from 1992 to 1999, the population estimates were not given in the file for different assumptions (since they were not projections but estimates). It is interesting to note that from the 1997 NGS, we estimated that there were 1,053 doctoral graduates at the 1995 graduation, and from the table below, we estimated that there were 1,347 doctoral graduates in 1995, which is close

Table 5.40: Estimated number of doctoral graduates from 1998 to 2018
aged 34 years old and less

Year	No assumption	Ass. 1	Ass. 2	Ass. 3	Ass. 4
1992	1,419	—	—	—	—
1993	1,395	—	—	—	—
1994	1,371	—	—	—	—
1995	1,347	—	—	—	—
1996	1,321	—	—	—	—
1997	1,297	—	—	—	—
1998	1,267	—	—	—	—
1999	1,238	—	—	—	—
2000	—	1,220	1,220	1,221	1,222
2001	—	1,210	1,212	1,214	1,215
2002	—	1,208	1,211	1,214	1,217
2003	—	1,207	1,212	1,218	1,221
2004	—	1,207	1,215	1,223	1,228
2005	—	1,208	1,218	1,229	1,235
2006	—	1,208	1,222	1,237	1,244
2007	—	1,212	1,228	1,246	1,254
2008	—	1,217	1,237	1,257	1,266
2009	—	1,223	1,246	1,270	1,279
2010	—	1,227	1,253	1,279	1,289
2011	—	1,231	1,259	1,288	1,298
2012	—	1,234	1,264	1,295	1,306
2013	—	1,237	1,269	1,303	1,314
2014	—	1,242	1,277	1,313	1,325
2015	—	1,251	1,288	1,326	1,339
2016	—	1,260	1,300	1,340	1,353
2017	—	1,270	1,312	1,354	1,368
2018	—	1,276	1,320	1,365	1,380

Source: Based on table 5.39

From these estimated numbers of doctoral graduates, some will become full-time university

teachers and some others will not. We multiplied every cell from the table above with 30.2%, which is the proportion of people with a doctorate and who are university teachers in 1996 (this is a result from section 5.6.2) and then we obtain the estimated number of doctoral graduates who may become university teachers in the future.

Table 5.41: Estimated number of doctoral graduates aged 34 years old and less who would become full-time university teachers from 1998 to 2018

Year	No assumption	Ass. 1	Ass. 2	Ass. 3	Ass. 4
1998	383	–	–	–	–
1999	374	–	–	–	–
2000	–	368	369	369	369
2001	–	366	366	367	367
2002	–	365	366	367	368
2003	–	365	366	368	369
2004	–	365	367	369	371
2005	–	365	368	371	373
2006	–	365	369	373	376
2007	–	366	371	376	379
2008	–	367	373	380	382
2009	–	369	376	383	386
2010	–	371	378	386	389
2011	–	372	380	389	392
2012	–	373	382	391	394
2013	–	374	383	393	397
2014	–	375	386	396	400
2015	–	378	389	401	404
2016	–	381	393	405	409
2017	–	383	396	409	413
2018	–	385	399	412	417

Source: Based on table 5.40

5.6.3.4 Comparison between different estimates of the number of doctoral graduates

The estimated number of doctoral graduates aged 34 years old and less from 1998 to 2018 who could become full-time university teachers varies from approximately 365 to 415. The estimated number of doctoral graduates (for all age categories) from 1998 to 2018 who could become full-time university teachers varies from approximately 780 to 950. So when the doctoral graduates aged 35 years old and over are not included as potential full-time university teachers, approximately 400 to 500 doctoral graduates are excluded.

5.6.4 Estimated number of full-time university teachers 1998-2018

From the results of 5.6.1, 5.6.2, and 5.6.3, we can establish the estimated number of university teachers from 1998 to 2018, according to the assumptions made throughout the process in adding:

- 1) The number of full-time university teachers in table 5.32 for each single year from 1998 to 2018.
- 2) The estimated number of people with a doctoral degree in 1996 that we assume will become full-time university teachers by the end of 2000 (table 5.36). We divided 3,336 by 5 to separate the number of people with a doctoral degree into the five years that we think it will take for them to become university teachers. So we added the number 667 to the years 1999 and 2000 as potential university professors. We added 2,001 to the year 1998 (to take account of the university teachers hired from 1996 to 1998 inclusively).
- 3) The estimated number of doctoral graduates, from table 5.41, of the year before the one for which we want to obtain an estimated number of full-time university teachers.

Here is an example: the estimated number of full-time university teachers in 1999 is the result of the sum of:

- 1)** the number of full-time university teachers in 1999 (from table 5.32);
- 2)** the number of people with a doctoral degree in 1996 that we assume will become university teachers in 1999 (from table 5.36);
- 3)** the number of doctoral graduates produced in 1998 (from table 5.41).

The estimated number of full-time university teachers in 1998 is only the result of the sum of steps 1 and 2 above (since we did not get an estimate of the number of doctoral graduates produced in 1997).

Table 5.42: Estimated number of full-time university teachers
from 1998 to 2018

Year	No assumption	Ass. 1	Ass. 2	Ass. 3	Ass. 4
1998	27,693	–	–	–	–
1999	26,246	–	–	–	–
2000	–	25,740	25,740	25,740	25,740
2001	–	24,571	24,572	24,572	24,572
2002	–	24,071	24,071	24,072	24,072
2003	–	23,574	23,575	23,576	23,577
2004	–	23,077	23,078	23,080	23,081
2005	–	22,039	22,041	22,043	22,045
2006	–	21,001	21,004	21,007	21,009
2007	–	19,962	19,966	19,970	19,973
2008	–	18,927	18,932	18,937	18,940
2009	–	17,889	17,895	17,902	17,904
2010	–	16,848	16,855	16,862	16,865
2011	–	15,808	15,815	15,823	15,826
2012	–	14,766	14,774	14,783	14,786
2013	–	13,725	13,734	13,743	13,746
2014	–	12,683	12,692	12,702	12,706
2015	–	11,825	11,836	11,846	11,850
2016	–	10,969	10,980	10,992	10,995
2017	–	10,112	10,124	10,136	10,140
2018	–	9,254	9,267	9,280	9,284

Source: Tables 5.32, 5.36 and 5.41.

CHAPTER 6: SUMMARY RESULTS AND FUTURE RESEARCH

This chapter discusses the results obtained in chapter 4 and chapter 5. It also gives suggestions for further research on the same topic.

6.1 Discussion about the proportional odds model (Chapter 4)

The fitting of the proportional odds model in section 4.2 of chapter 4 has been done to try to explain the university level of certification obtained at graduation. This model used the following explanatory variables: field of study, gender, region of residence (at the time of the NGS), the status of registration, the first spoken language and the age category (at the time of the NGS). The following interaction terms were also included in the model:

- 1) field of study by gender
- 2) field of study by age category
- 3) gender by age category

6.1.1 Justification for the interactions included in the model

As stated in Chapter 4, these three two-way interactions were assumed to be the more important to be tested and therefore have been included in the model. A limited number of explanatory variables and interaction terms could be included in the model, without having a problem with singularity of the matrix of estimable parameters. Therefore, an important interaction between two (or more) terms may have been missed and if it would have been included, the results could have changed. There were good reasons to keep these three interaction terms in the model. It was interesting to see if the gender and the field of study have an interaction at some levels. Especially, it was desired to see if there was an interaction between the gender and the scientific fields of study. For the interaction between the field of study and age, it was also interesting to see if there was an interaction between

older graduates who studied in a certain field of study (scientific or not). Then for the interaction between the gender and age, it was desired to see if there was an interaction between gender and older graduates. The changes over the years (1986, 1990, and 1995) were also important to verify.

6.1.2 Design effects

One result obtained from the fitting of this model is that according to the design effects of the parameter estimates that were obtained, we concluded that the use of the sampling design when the variances were estimated provided a great improvement in these estimated variances. Almost every design effect was less than one. If the proportional odds model would have been fitted assuming simple random sampling instead of stratified systematic random sampling, the standard errors would have been higher and this would have probably have affected the p-values of the t-test of the parameter estimates.

6.1.3 Population counts

When SUDAAN was used to fit the proportional odds model, recall that we used estimated (from the weights) population counts by stratum instead of the population counts by stratum obtained from the administrative records (used for the sampling). Since the population counts from the administrative records were not available for the 1988 and 1992 models, the estimated population counts by stratum were used for the 1988, the 1992, and the 1997 models. But the population counts were available from the administrative records for the 1997 NGS. So from the 1997 NGS data, two proportional odds models have been fitted: one using estimated population counts by stratum and one using administrative population counts by stratum. We saw that when the standard error of a parameter estimate was different from both 1997 models, it was always the standard error of the parameter estimate from the model using administrative population counts that was the highest. But the conclusion from the

testing of the equality of the unequal standard errors was that all of the unequal standard errors were determined to be statistically equal, at the 5% significance level. We assume that the same conclusions would have been obtained from the 1988 and the 1992 models, and so the use of estimated population counts by stratum was correct.

6.1.4 Interpretation of main effects

From the Wald chi-square statistics observed, we concluded that the three interaction terms had a significant effect. But there were some levels of the explanatory variables that were determined to be equal to zero. Therefore, they can be excluded from the computations of the estimated logit of the cumulative conditional probabilities ($L_j(\mathbf{x}_i)$). There were also some interactions between two levels of the variables forming an interaction that were also determined to be zero and could also be excluded from the computations of the estimated $L_j(\mathbf{x}_i)$. When the parameter estimate of a level of an explanatory variable was determined to be zero, we concluded that graduates having this level did not get significantly higher (or lower) levels of certification at graduation than the graduates in the reference level (the last level for each explanatory variable). And when the parameter estimate of an interaction involving the level of an explanatory variable and the level of another explanatory variable was determined to be zero, we concluded that graduates having these levels for these two variables did not get significantly higher (or lower) levels of certification at graduation than graduates having the level of the first explanatory variable of the interaction and the last level (the reference level) of the second variable in the interaction. These results are in table 4.10 in Chapter 4.

6.1.4.1 First language spoken and still understood

For the variable “First language spoken and still understood”, graduates whose first language spoken and still understood was neither of Canada’s official languages had significantly

higher levels of certification at graduation than graduates who spoke either one of Canada's official languages. This is reflected in the negative parameter estimate for the first language spoken and still understood being other than English or French, for all of the three models. This result is also reflected in the odds ratios obtained in section 4.4.5 in chapter 4. From 1986 to 1995, the odds of obtaining either a master's degree (or a university diploma above bachelor) or a doctorate degree for graduates who spoke neither of Canada's official languages (compared to graduates who spoke English or French) increased (and were greater than one). We might attempt to conclude that the chances of attaining high levels of education are higher for graduates who were born outside Canada than for graduates who were born in Canada but this, of course, is if graduates who were born outside Canada are correctly identified by their first language spoken and still understood being other than English or French.

6.1.4.2 Status of registration

The status of registration "Part-time only" was not significant for the three models (1988, 1992, and 1997). So we can say that graduates who studied part-time only did not have significantly higher (or lower) levels of certification at graduation than graduates who studied full-time and part-time. However, graduates who studied full-time only had significantly higher levels of certification at graduation than graduates who studied full-time and part-time, for the three models. This is reflected in the negative parameter estimate for the graduates who studied full-time only, for all of the three models. This result is also reflected in the odds ratios obtained in section 4.4.5, chapter 4. From 1986 to 1995, the odds of obtaining either a master's degree (or a university diploma above bachelor) or a doctorate degree for graduates who studied full-time only were always higher than for graduates who studied full-time and part-time. But the odds decreased from 1988 to 1992, and even if it they increased from 1992 to 1997, they remained lower than the odds in 1988. We might

then think that this result can indicate decreasing chances for graduates who studied full-time only of attaining high levels of education compared to graduates who studied full-time and part-time.

6.1.4.3 Region of residence

Even if the region of residence was determined to be significant in 1988, 1992, and 1997 from the overall Wald chi-square test, many regions did not have significantly higher (or lower) levels of certification at graduation than graduates who lived in the Yukon or Northwest Territories two years after graduation. In 1988, graduates who lived in any Canadian region have not had significantly higher (or lower) levels of certification at graduation (two years earlier) than graduates who lived in the Yukon or Northwest Territories; in 1992, graduates who lived in the Prairies have had significantly lower levels of certification at graduation (two years earlier) than graduates who lived in the Yukon and Northwest Territories; and in 1997, graduates who lived in the Atlantic provinces or in Ontario have had significantly higher levels of certification at graduation (two years earlier) than graduates who lived in the Yukon or Northwest Territories. These conclusions are also reflected in the positive parameter estimate for the graduates who lived in the Prairies in 1992, and in the negative parameter estimates for the students who lived in the Atlantic provinces or in Ontario in 1997. These results are also reflected in the odds ratios. The odds of having obtained either a master's degree (or a university diploma above bachelor) or a doctorate degree in 1990 for graduates who lived in the Prairies in 1992 were reduced by 42% compared to graduates who lived in the Yukon or Northwest Territories. The odds of having obtained either a master's degree (or a university diploma above bachelor) or a doctorate degree in 1995 for graduates who lived in the Atlantic provinces in 1997 were almost 2.5 times higher than for graduates who lived in the Yukon or Northwest Territories; and the odds of having obtained either a master's degree (or a university diploma above

bachelor) or a doctorate degree in 1995 for graduates who lived in Ontario in 1997 were almost 2.5 times higher than for graduates who lived in the Yukon or Northwest Territories. This might indicate that the likelihood of obtaining a master's degree (or a university diploma above bachelor) or a doctoral degree in 1990 was higher for graduates who lived (and maybe worked) in the Prairies in 1992 than for graduates who lived in the Yukon or Northwest Territories; the likelihood of obtaining a master's degree (or a university diploma above bachelor) or a doctoral degree in 1995 was higher for graduates who lived (and maybe worked) in the Atlantic provinces or in Ontario in 1997 than for graduates who lived in the Yukon or Northwest Territories.

6.1.5 Interpretation of the interactions

6.1.5.1 Gender and age

The first interaction of the model was between gender and age. The interaction between male graduates and graduates aged 15 to 24 years old (two years after graduation) was not significant for the 1988 and 1992 models (but was significant for the 1997 model). So we can say that male graduates aged 15 to 24 years old (in 1988 or 1992) did not have significantly higher (or lower) levels of certification at graduation than male graduates aged 35 years old and over (in 1988 or 1992). However, male graduates aged 15 to 24 years old (in 1997) had significantly lower levels of certification at graduation than male graduates aged 35 years old and over (in 1997). This is reflected in the positive parameter estimate for the male graduates aged 15 to 24 years old (in 1997). This might indicate that a shift occurred between the 1990 graduation and the 1995 graduation. Male younger graduates (represented by the 15 to 24 year-old category) obtained approximately the same levels of education than male older graduates (represented by the 35 years old and over category) in 1986 and 1990 but male younger graduates obtained low levels of education, compared to male older graduates in 1995. This might indicate a tendency for older male graduates to get high levels

of education compared to young male graduates.

6.1.5.2 Gender and field of study

The interaction between male graduates and graduates who studied in Agricultural and biological sciences was significant for the three models. So we can say that male graduates in Agricultural and biological sciences had significantly lower levels of certification at graduation than male graduates in Mathematics and physical sciences in 1986. However, male graduates in Agricultural and biological sciences had significantly higher levels of certification at graduation than male graduates in Mathematics and physical sciences in 1990 and 1995. This is reflected in the positive parameter estimate for the male graduates in Agricultural and biological sciences (1986) and in the negative parameter estimates for the male graduates in Agricultural and biological sciences (1990 and 1995). This seems to indicate that a shift occurred between the 1986 graduation and the 1990 graduation. Male graduates in Agricultural and biological sciences seem, in the recent years, to obtain higher levels of education than male graduates in Mathematics and applied sciences. Is there an increasing demand on the market for high educated graduates in Agricultural and biological sciences compared to the demand for high educated graduates in Mathematics and physical sciences? Maybe the type of work of the graduates in Agricultural and biological sciences requires now more educated workers than the type of work of the graduates in Mathematics and physical sciences.

The interaction between male graduates and graduates who studied in Medical and health professions was significant for the 1988 and 1997 models, but not for the 1992 model. We can say that male graduates in Medical and health professions had significantly lower levels of certification at graduation than male graduates in Mathematics and physical sciences in 1986. However, male graduates in Medical and health professions did not have significantly

higher (or lower) levels of certification at graduation than male graduates in Mathematics and physical sciences in 1990. Then, male graduates in Medical and health professions had significantly higher levels of certification at graduation than male graduates in Mathematics and physical sciences in 1995. This is reflected in the positive parameter estimate for the male graduates in Medical and health professions (1986) and in the negative parameter estimate for the male graduates in Medical and health professions (1995). This seems to indicate that a shift occurred between the 1986 graduation and the 1990 graduation, and between the 1990 graduation and the 1995 graduation. Male graduates in Medical and health professions seem, in the past years, to have gone from having lower levels of education than male graduates in Mathematics and applied sciences to having higher levels of education than male graduates in Mathematics and applied sciences, with a time where there was not a significant difference between levels of certification obtained at graduation for male graduates in Medical and health professions or in Mathematics and physical sciences.

The interaction between male graduates and graduates who studied in Engineering and applied sciences was not significant for the three models. We can say that male graduates in Engineering and applied sciences did not have significantly higher (or lower) levels of certification at graduation than male graduates in Mathematics and physical sciences. This seems to indicate that from 1986 to 1995, there was not a significant difference between levels of certification obtained at graduation for male graduates in Engineering and applied sciences and levels of certification obtained at graduation for male graduates in Mathematics and physical sciences.

6.1.5.3 Field of study and age

The last interaction term in the model was the interaction between the field of study and the age. The interaction for graduates in Medical and health professions aged 15 to 24 years old

was significant only in the 1992 model. So we can say that graduates in Medical and health professions aged 15 to 24 years old had significantly lower levels of certification at graduation than graduates in Medical and health professions aged 35 years old and over in 1990. This is reflected in the positive parameter estimate for the 1990 graduates in Medical and health professions aged 15 to 24 years old. So in 1986, graduates in Medical and health professions aged 15 to 24 years old did not get significantly higher (or lower) levels of university education than graduates in Medical and health professions aged 35 years old and over; but did get significantly lower levels of certification at graduation than graduates in Medical and health professions aged 35 years old and over in 1990; and come back in 1995 to no significantly higher (or lower) levels of university education than graduates in Medical and health professions aged 35 years old and over. So we can not say if it indicates a trend for graduates in Medical and health professions aged 15 to 24 years old to have significantly higher (or lower) levels of certification at graduation than graduates in Medical and health professions aged 35 years old and over.

The interaction for graduates in Agricultural and biological sciences aged 30 to 34 years old was significant for the 1988 and 1992 models, but not for the 1997 model. So we can say that graduates in Agricultural and biological sciences aged 30 to 34 years old had significantly lower levels of certification at graduation than graduates in Agricultural and biological sciences aged 35 years old and over in 1986 and 1990. But in 1995, graduates in Agricultural and biological sciences aged 30 to 34 years old did not have significantly higher (or lower) levels of certification at graduation than graduates in Agricultural and biological sciences aged 35 years old and over. This is reflected by the positive parameter estimates for graduates in Agricultural and biological sciences aged 30 to 34 years old in 1986 and 1990. This might indicate that a few years ago, older graduates in Agricultural and biological sciences got high levels of certification, compared to the graduates a little younger. But in

1995, this result was not valid anymore.

The interaction for graduates in Engineering and applied sciences aged 30 to 34 years old was significant for the 1988 and 1992 models, but not for the 1997 model. So we can say that graduates in Engineering and applied sciences aged 30 to 34 years old had significantly lower levels of certification at graduation than graduates in Engineering and applied sciences aged 35 years old and over in 1986 and 1990. But in 1995, graduates in Engineering and applied sciences aged 30 to 34 years old did not have significantly higher (or lower) levels of certification at graduation than graduates in Engineering and applied sciences aged 35 years old and over. This is reflected by the positive parameter estimates for graduates in Engineering and applied sciences aged 30 to 34 years old in 1986 and 1990. This might indicate that a few years ago, older graduates in Engineering and applied sciences got high levels of certification, compared to the graduates a little younger. But in 1995, this result was not valid anymore.

The interactions for graduates who studied in Other Social sciences and aged either 15 to 24 years old; 25 to 29 years old; or 30 to 34 years old were not significant for the three models. We can say that graduates in Other Social sciences sciences aged 15 to 24 years old did not have significantly higher (or lower) levels of certification at graduation than graduates in Other Social sciences sciences aged 35 years old and over; graduates in Other Social sciences sciences aged 25 to 29 years old did not have significantly higher (or lower) levels of certification at graduation than graduates in Other Social sciences sciences aged 35 years old and over ; and that graduates in Other Social sciences sciences aged 30 to 34 years old did not have significantly higher (or lower) levels of certification at graduation than graduates in Other Social sciences sciences aged 35 years old and over. This seems to indicate that from 1986 to 1995, there was not a significant difference between levels of certification obtained at graduation for graduates in Other Social sciences in any of the age

category compared to graduates in Other Social sciences aged 35 years old and over.

6.1.6 Summary

This summarizes results obtained from the proportional odds model from section 4.2 in chapter 4. This can help explaining some factors affecting the university level of certification attained at graduation, from 1986 to 1995. But we should not forget that even if the model was determined to be acceptable, problems were caused in the validation due to the sparseness in the data. The effect of the sparseness was not measured but some large parameter estimates may have been caused by sparseness. Consequently, the odds ratios calculated from these parameter estimates may also have been affected by the sparseness. So despite the large subsample size and subpopulation size, the number of cells (determined by the combinations of the levels of the response and the levels of the explanatory variables) was not small enough to avoid the sparseness problem. Sparseness also caused problems in the validation of the model. Another problem with the validation of the model was the fact that the sample design was not taken into account. A corrected Pearson's chi-squared statistic was described, which take into account the sample design. But since the estimated covariance matrix of $\hat{\mathbf{P}}$ (the population proportions) was unknown, it has not been implemented. Then all of the other methods used to assess the validity of this model were not based on the sample design. For generalized linear models, methods of assessing goodness-of-fit or residuals to be used are well-documented. But when it comes time to take into account the sample design when using those methods, that was much less well-documented. In fact, I did not find anything about the assessment of the goodness-of-fit of the proportional odds model using survey data (except the corrected Pearson's chi-squared statistic from Roberts (1985)). Even though the model was declared acceptable, sparseness in data and the use of non-survey methods in the model validation imply that the use of the

results must be done with care.

6.2 Discussion about the university enrolment

A study of university enrolment was done but the results were not used to predict supply of university teachers. This is left to further research. But interesting facts were noted. From the ANOVA models, we noted that the region of institution, the gender, the age category and the year of enrolment have an effect on the undergraduate enrolment and on the graduate enrolment. So these four variables must be taken into account when the future enrolment is studied. It is also possible to fit other ANOVA models using different interactions of interest between some of these variables.

About the university enrolment profile obtained, we note that the undergraduate enrolment totals in the 15 to 24 year-old category increased from 1991 to 1998, while the enrolment totals in the three other categories decreased slightly from 1991 to 1998. The graduate enrolment totals have had a large increase from 1976 to 1991 for the students 35 years of age and over, and have started to be higher than the graduate enrolment totals for the students 30 to 34 years of age between 1981 and 1986, and until 1998. This result might indicate an increasing number of adults returning to school at the graduate level. The graduate enrolment totals for the students in the 25 to 29 year-old category increased from 1976 to 1998. So it seems that the number of students enrolled in an undergraduate program and aged 15 to 24 increased in the past years, and that the number of students enrolled in a graduate program for any age category increased in the past years. At the national level, the number of males enrolled in an university program and aged either 15 to 24, 25 to 29, or 35 and over increased a little from 1976 to 1991 and decreased a little from 1991 to 1998. But these variations were not very large. Still at the national level, the number of females enrolled in an university program and aged 15 to 24 has had a large increase from 1976 to 1998. In the three other age categories, the number of females enrolled in an university program

increased from 1976 to 1991 and then decreased from 1991 to 1998. These variations are larger than the variations in the number of males enrolled in an university program for the same age categories.

Before summarizing results about the proportions of students enrolled in a university program, recall that we assumed that the student's region of institution was his region of residence. The region of residence of the students enrolled in an university program was not available on the enrolment data file. That is why we had to assume that these two regions were the same for the students. We do not know how many students study in a region but live in another so we can not say how this has affected the proportions obtained. The second assumption was about the year used for the population estimates and the survey year used for the enrolment data. The population estimates provide an estimate of the number of people in a civil year (as of July 1st of each year). The USIS counts are obtained in the Fall of each year, as a proxy of the number of students in a complete academic year (September to August). So the population counts and the enrolment counts are not totally comparable. The last thing to remember is that no proportions could be obtained in the Yukon and Northwest Territories, since there is no university in these two regions.

The proportion of females in the population enrolled in an university program was always higher than the male proportion, in every region and for every age category. The Atlantic provinces had the highest proportion of males enrolled in an university program, in the 15 to 24 year-old category (around 14%). British Columbia had the lowest proportion of males enrolled in an university program for the same age category (around 7%). The proportions of males enrolled in an university program by region and age category did not vary very much (less than 1% difference between 1994 to 1998). The Atlantic provinces had the highest proportion of females enrolled in an university program, in the 15 to 24 year-old category (around 19%). British Columbia had the lowest proportion of females enrolled in an university program for the same age category (around 9%). The proportions of females

enrolled in an university program by region and age category did not vary very much (less than 1% difference between 1994 to 1998).

Since we saw that the proportions of males or females enrolled in an university program did not vary very much from 1994 to 1998, it might be because the number of males or females enrolled in an university program increased (or decreased) as the number of males or females in the population increased (or decreased). This should be taken into account for a further study of the factors affecting university enrolment.

We also concluded from the regression analysis that the year of enrolment had a significant effect on the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia. According to the estimated regression equation obtained, the proportion of women aged 15 to 24 years old enrolled in an university program from British Columbia will increase as year will increase. For all the other combinations of region of institution, gender, and age category, a regression analysis could be performed in order to assess if the year had a significant effect on the proportion of people enrolled in an university program (according to some specific characteristics).

6.3 Discussion about the estimates of the full-time university teachers in Canada

6.3.1 The effect of the region of institution in the model

The fitting of the proportional odds model in chapter 5, in section 5.3, was done mainly to use the fitted conditional probabilities when we want to estimate the number of doctoral graduates in the future. But even though only three explanatory variables were used (and we saw in chapter 4 that six explanatory variables were used and were significant), a certain interpretation of the model was done. First thing to say is that when the region of institution is added to the proportional odds model from section 5.3, chapter 5, the odds ratios calculated from both models (the one in section 5.3 and the one in section 5.4) did not vary much. Even if the difference was not tested (to be determined statistically significant or not),

we saw that the conclusions drawn from the odds ratios did not change from one model to another. The testing of the main effects and the interactions was done for both models, and again the conclusions drawn from one model to another did not change. There was only one interaction that was determined to be not significant in both models, and when the region of institution was included in the model, it made this interaction a little more nonsignificant. Lastly, the inclusion of the region of institution in the model brought some changes in the results from the testing of the levels of the parameter estimates (to be determined equal to zero or not). A greater number of levels of parameter estimates or combination of levels for an interaction term were determined to be equal to zero by the t-tests. In the 1988 model, the parameter estimate corresponding to the Atlantic provinces (as the region of residence) was determined to be equal to zero in the model including the region of institution (and not in the model that did not include the region of institution); and the parameter estimate corresponding to the graduates living in the Prairies two years after their graduation in the 25 to 29 year-old category was determined to be equal to zero in the model including the region of institution (and not in the model that did not include the region of institution). For the 1992 model, the parameter estimate corresponding to Quebec (as the region of residence) was determined to be equal to zero in the model including the region of institution (and not in the model that did not include the region of institution); the parameter estimate corresponding to Ontario (as the region of residence) was determined not to be equal to zero in the model including the region of institution (and it was in the model that did not include the region of institution); the parameter estimate corresponding to the graduates who lived in Quebec two years after graduation in the 25 to 29 year-old category was determined to be equal to zero in the model including the region of institution (and not in the model that did not include the region of institution). In conclusion, there has not been too many changes in the conclusions from the t-tests of the parameter estimates from the two proportional odds models fitted in chapter 5. So even if the main effect of the region of institution was determined to be significant, it did not bring a totally different interpretation of the model

without the region of institution. So the omission of the region of institution from the model, and consequently from the fitted conditional probabilities, was justified. This is a good thing because this variable could not have been used to estimate the number of doctoral graduates in the future since this variable was not available on the population estimates file.

6.3.2 Design effects

Now if we discuss the results of the proportional odds model with the age category, the gender, and the region of residence as the explanatory variables, first recall that by the design effects of the parameter estimates obtained, we concluded that the use of the sampling design when the variances were estimated provided a great improvement in these estimated variances. We also concluded that the use of estimated population counts by stratum was correct for the three models.

6.3.3 Interpretation of the interactions

6.3.3.1 Gender and region of residence

The first interaction was between the gender and the region of residence. The interaction for male graduates living in the Atlantic provinces (two years after graduation) was not significant for the 1988 and 1992 models (but was significant for the 1997 model). So we can say that male graduates living in the Atlantic provinces (in 1988 or 1992) did not have significantly higher (or lower) levels of certification at graduation than male graduates living either in British Columbia, Yukon, or Northwest Territories. However, male graduates living in the Atlantic provinces (in 1997) had significantly lower levels of certification at graduation than male graduates living either in British Columbia, Yukon, or Northwest Territories (in 1997). This is reflected in the positive parameter estimate for the male graduates living in the Atlantic provinces (in 1997). This might indicate that a shift occurred between the 1990 graduation and the 1995 graduation. Male graduates living in the Atlantic

provinces two years after their graduation in 1988 and 1992 obtained approximately the same levels of education than male graduates living either in British Columbia, Yukon, or Northwest Territories in 1988 and 1992 but male graduates living in the Atlantic provinces two years after their graduation in 1997 obtained low levels of education, compared to male graduates living either in British Columbia, Yukon, or Northwest Territories in 1997. This might indicate a tendency in future for male graduates with low levels of education to go live (and possibly work) in the Atlantic provinces instead of British Columbia, Yukon, or Northwest Territories.

6.3.3.2 Gender and age

The second interaction of the model was between gender and age. The interaction for male graduates aged 15 to 24 years old (two years after graduation) was not significant for the 1988 and 1992 models (but was significant for the 1997 model). So we can say that male graduates aged 15 to 24 years old (in 1988 and 1992) did not have significantly higher (or lower) levels of certification at graduation than male graduates aged 35 years old and over. However, male graduates 15 to 24 years of age (in 1997) had significantly lower levels of certification at graduation than male graduates aged 35 years old and over (in 1997). This is reflected in the positive parameter estimate for the male graduates 15 to 24 years of age (in 1997). This might indicate that a shift occurred between the 1990 graduation and the 1995 graduation. Male graduates aged 15 to 24 years old two years after their graduation obtained approximately the same levels of education than male graduates aged 35 years old and over in 1988 and 1992 but male graduates aged 15 to 24 years old two years after their graduation obtained low levels of education, compared to male graduates aged 35 years old and over in 1997. This might indicate a tendency in future for young male graduates to get lower levels of education at graduation compared to older male graduates.

6.3.3.3 Region of residence and age

The third and last interaction of the model was between the region of residence and the age two years after graduation. The interaction for graduates living in the Atlantic provinces and aged 30 to 34 years old (two years after graduation) was not significant for the 1988 and 1992 models (but was significant for the 1997 model). So we can say that graduates living in the Atlantic provinces and aged 30 to 34 years old (in 1988 and 1992) did not have significantly higher (or lower) levels of certification at graduation than graduates living in the Atlantic provinces and aged 35 years old and over. However, graduates living in the Atlantic provinces and aged 30 to 34 years old (in 1997) had significantly lower levels of certification at graduation than graduates living in the Atlantic provinces and aged 35 years old and over (in 1997). This is reflected in the positive parameter estimate for the graduates living in the Atlantic provinces and aged 30 to 34 years of age (in 1997). This might indicate that a shift occurred between the 1990 graduation and the 1995 graduation. Graduates living in the Atlantic provinces and aged 30 to 34 years old two years after their graduation obtained approximately the same levels of education as graduates living in the Atlantic provinces and aged 35 years old and over in 1988 and 1992 but graduates living in the Atlantic provinces and aged 30 to 34 years old two years after their graduation obtained low levels of education, compared to graduates living in the Atlantic provinces and aged 35 years old and over in 1997. This might indicate a tendency in future for older graduates living in the Atlantic provinces two years after graduation to get higher levels of education at graduation compared to graduates living in the same region who are a little younger.

The interaction for graduates living in Quebec and aged 30 to 34 years old (two years after graduation) was not significant for the 1992 model (but was significant for the 1988 and 1997 models). So we can say that graduates living in Quebec and aged 30 to 34 years old (in 1992) did not have significantly higher (or lower) levels of certification at graduation than

graduates living in Quebec and aged 35 years old and over. However, graduates living in Quebec and aged 30 to 34 years old (in 1988 and 1997) had significantly higher levels of certification at graduation than graduates living in Quebec and aged 35 years old and over (in 1988 and 1997). This is reflected in the negative parameter estimate for the graduates living in Quebec and aged 30 to 34 years of age (in 1988 and 1997). This might indicate that a shift occurred between the 1986 graduation and the 1990 graduation, and then between the 1990 graduation and the 1995 graduation. Graduates living in Quebec and aged 30 to 34 years old two years after their graduation obtained approximately the same levels of education than graduates living in Quebec and aged 35 years old and over in 1992 but graduates living in Quebec and aged 30 to 34 years old two years after their graduation obtained high levels of education, compared to graduates living in Quebec and aged 35 years old and over in 1988 and 1997.

The interaction for graduates living in Ontario and aged 25 to 29 years old (two years after graduation) was not significant for the 1992 and the 1997 models (but was significant for the 1988 model). So we can say that graduates living in Ontario and aged 25 to 29 years old (in 1992 and 1997) did not have significantly higher (or lower) levels of certification at graduation than graduates living in Ontario and aged 35 years old and over. However, graduates living in Ontario and aged 25 to 29 years old (in 1988) had significantly higher levels of certification at graduation than graduates living in Ontario and aged 35 years old and over (in 1988). This is reflected in the negative parameter estimate for the graduates living in Ontario and aged 25 to 29 years of age (in 1988). This might indicate that a shift occurred between the 1986 graduation and the 1990 graduation. Graduates living in Ontario and aged 25 to 29 years old two years after their graduation obtained higher levels of education than graduates living in Ontario and aged 35 years old and over in 1988 but graduates living in Ontario and aged 25 to 29 years old two years after their graduation obtained approximately the same levels of education, compared to graduates living in

Ontario and aged 35 years old and over in 1992 and 1997. The same conclusions apply for graduates living in Ontario and aged 30 to 34 years old compared to graduates living in Ontario and aged 35 years old and over.

The interaction for graduates living in the Prairies and aged 25 to 29 years old (two years after graduation) was not significant for the 1992 model (but was significant for the 1988 and 1997 models). So we can say that graduates living in the Prairies and aged 25 to 29 years old (in 1992) did not have significantly higher (or lower) levels of certification at graduation than graduates living in the Prairies and aged 35 years old and over. However, graduates living in the Prairies and aged 25 to 29 years old (in 1988) had significantly higher levels of certification at graduation than graduates living in the Prairies and aged 35 years old and over (in 1988). And graduates living in the Prairies and aged 25 to 29 years old (in 1997) had significantly lower levels of certification at graduation than graduates living in the Prairies and aged 35 years old and over (in 1997). This is reflected in the negative parameter estimate for the graduates living in the Prairies and aged 25 to 29 years of age (in 1988) and in the positive parameter estimate for the graduates living in the Prairies and aged 25 to 29 years of age (in 1997). This might indicate that a shift occurred between the 1986 graduation and the 1990 graduation, and then between the 1990 graduation and the 1995 graduation. Graduates living in the Prairies and aged 25 to 29 years old two years after their graduation obtained higher levels of education than graduates living in the Prairies and aged 35 years old and over in 1988, then these same graduates obtained approximately the same levels of education as graduates living in the Prairies and aged 35 years old and over in 1992, and then graduates living in the Prairies and aged 25 to 29 years old two years after their graduation obtained lower levels of education, compared to graduates living in the Prairies and aged 35 years old and over in 1997. This can indicate there is a tendency for graduates aged 25 to 29 years old living in the Prairies two years after their graduation to get low levels of education, through the

years, compared to graduates aged 35 years old and over living the Prairies.

6.3.4 Problems with the validation

Some problems were also caused in the validation of this model due to the sparseness in the data. But among the three proportional odds models fitted through chapter 4 and chapter 5, it is the model with the age, the gender, and the region of residence as the explanatory variables that had the largest number of cells (defined by the response level and the combination of the levels of the explanatory variables) with more than one observation by cell. So the sparseness seemed to be less of a problem for this model. The Pearson's chi-squared statistics and the deviance statistics took the lowest values (by year) compared to the same statistics for the two other models. The maximum cell residual and the minimum cell residual were lower than the same cell residuals for the two other models. However, it is the model from chapter 4 that had the largest proportion of cell residuals and cumulative cell residuals between -2 and 2, once the missing residuals were deleted. We can conclude that the model with age, gender, and region of residence as the explanatory variables was better than the two others, based on some aspects of the validation, but again validation methods based on survey data would have given a more accurate assessment of goodness-of-fit.

6.3.5 Estimated number of full-time university teachers 1998-2018

6.3.5.1 The number of full-time university teachers after retirement and attrition

The estimates of the number of full-time university teachers from 1998 to 2018 were obtained based on several assumptions. The starting point of the process is the number of full-time teachers in 1998. It does not include the staff members in faculties of medicine or dentistry. In section 6.3.5.2, we will see that these excluded members might have been included in the proportion of Canadians with a doctorate who are full-time university teachers in 1996. Then we assumed that the full-time university teachers in 1998 will retire

at age 65. In fact, some may continue teaching after 65 years old and some may quit teaching before 65 years old. The early retirement (before 65 years old) was taken into account in the attrition rate. Then, after retirement at 65 years old, we knew the number of full-time university teachers that would be left in Canada for some years only between 1988 and 2018. That is why we assumed a constant retirement rate between two years (two consecutive) for which the number of full-time teachers after retirement was known. In fact, the retirement rate between two specific years will not be necessarily constant but since we wanted to obtain the number of full-time university teachers after retirement for each single year from 1998 to 2018, this assumption was necessary. The attrition rate (of 4% per year) takes into account professors who leave their job, either for other jobs, for health reasons, early retirement, or deaths. It was based on an estimate of the attrition rate at Carleton University per year. Then based on these assumptions, the number of full-time university teachers in 1998 that will be left in Canada after retirement and attrition for each single year from 1998 and 2018 was obtained. Approximately 16,800 full-time teachers in 1998 will have left their job by 2018. This is not surprising since in 1998, approximately 40% of the full-time university teachers in Canada (before retirement and attrition) were 50 to 59 years of age. It seems that between 2004 and 2014 the number of retired full-time university teachers per year will be the highest (approximately 1,000).

6.3.5.2 Estimated number of potential university teachers in the 1996 Canadian population

Once the number of full-time university teachers that will be left (after retirement and attrition) was known, we wanted to add to it a number of potential full-time university teachers by single year (from 1998 to 2018). The number of potential full-time university teachers is defined in two ways: those who will get a doctorate degree in future and those in the population who we know already have a doctorate and are not full-time university teachers. We assume that the doctorate degree is a prerequisite to be hired as a full-time

university teacher. In fact, some may have a master's degree and become university teachers too.

Among people in Canada who already have a doctorate degree but who are not full-time university teachers, only those aged either 15 to 24 or 25 to 34 in 1996 were kept. We first assume that if a person in 1996 has a doctorate (and is 34 years old and less) but is not a full-time university teacher, he will become a full-time university teacher within five years (by the end of 2000). After this period, we assume that if this person is still not a full-time university teacher, it is because he got another job. Only young (15 to 34 years old) people with a doctorate were kept because we assume that older people (age greater than 34 years old) with a doctorate have already gotten another job (other than university teaching). And the year 1996 was retained because data available about people with a doctorate degree in Canada was only for 1991 and 1996. And we assumed that people with a doctorate in 1991, if they became full-time university teachers, must have been included in the counts of the full-time university teachers in 1998.

So the number of Canadians aged 15 to 24 (or 25 to 34) whose highest degree, diploma, or certificate ever obtained was the doctorate in 1996 was compared to the number of full-time university teachers (see definition of the UCASS full-time teaching staff in section 5.6.1, chapter 5) in 1996 aged, respectively, 29 and less and 30 to 34. We made the assumption that the the number of full-time university teachers 29 years of age and less in 1996 and those 15 to 24 years of age is not too different. It would have been better using the number of full-time university teachers 15 to 24 years of age in 1996 but it was not available on the data file we had (from UCASS). We also made the assumption that the the number of full-time university teachers 25 to 34 years old in 1996 and those 30 to 34 years of age is not too different. It would have been better using the number of full-time university teachers aged 25 to 34 in 1996 but it was not available on the data file we had (from UCASS). So the number of people with a doctorate in 1996 but who were not full-time university teachers is an

approximation for these two age categories.

From this number, we know that only a given proportion will become full-time university teachers. The proportion of people with a doctorate in Canada who are full-time university teachers was estimated in 1996. Note that the number of full-time university teachers from the UCASS data we had were not available by gender. And later in the process, the same proportion by gender was needed. To be comparable, these proportions had to come from the same data source. This is why the Census counts (and not the UCASS counts) were used to obtain the proportion of people with a doctorate in Canada who are full-time university teachers in 1996. The number of full-time university teachers in 1995 was divided by the number of Canadians in 1996 whose highest degree, diploma, or certificate ever obtained was the doctorate to obtain the required proportion (data obtained from the 1996 Census CD-ROM Products). We assumed that at the time of the 1996 Census, there was not a big difference in the number of full-time university teachers in 1995 and the same number in 1996.

The full-time university teachers from the 1996 Census were targeted according to their detailed occupation (based on 1991 Standard Occupation Classification). This refers to the kind of work persons were doing during the reference week (week before May 16th, 1996), as determined by their kind of work and the description of the most important duties in their job. Data are available for persons 15 years of age and over, excluding institutional residents. If the person did not have a job during the week prior to enumeration, the data relate to the job of longest duration since January 1, 1995. Persons with two or more jobs were to report the information for the job at which they worked the most hours. There is no mention that the medical or dental staff members are excluded. However, we know that from the UCASS data, the number of full-time university teachers excluded these staff members. The term “full-time”, for the full-time university teachers from the 1996 Census data, does not have the exact same meaning as the definition of “full-time” for the UCASS data. It refers to persons 15 years of age and over, excluding institutional residents, who worked 49-52 weeks

on a full-time basis in the reference year (1995) for pay or in self-employment.

Once the proportion of people with a doctorate who are also full-time university teachers was obtained, it was multiplied to the number of people with a doctorate 15 to 24 years of age (or 25 to 34 years old) in 1996 but who were not full-time university teachers to obtain a number of potential full-time university teachers by the end of 2000 (from those who had a doctorate in 1996 but who were not full-time university teachers). But we do not know exactly the year before the end of 2000 at which they would become university teachers. So a same fixed number of new university teachers was assumed for each year between 1996 and 2000. This number (667) was added to the number of full-time teachers after retirement in 1999 and 2000. In 1998, we added the number of potential university teachers for the years 1996, 1997 and 1998. This is only an assumption and it may happen that the 3,336 potential full-time university teachers were hired all the same year.

In assuming that a certain proportion of 1996 people with a doctorate degree as their highest degree, diploma, or certificate would become university teachers, we added approximately 3,336 new university teachers from 1988 to 2000. If we estimated the number of people in 1996 with a master's degree as their highest degree, diploma, or certificate who could become full-time university teachers, we would have made the assumption that they would have first obtained a doctorate. Based on the assumption that it takes in average five years to obtain a doctorate degree (see Gonzalez, 1996), these people with a master's degree in 1996 would have obtained their doctorate degree approximately in 2001 (and we already obtained an estimate of the number of doctoral graduates in 2001 in section 5.6.3.3 in chapter 5). So these people would have been counted twice. The same remark applies for the people with a bachelor's degree as their highest degree, diploma, or certificate in 1996. Finally, note that if the number of people with a doctorate degree as their highest degree, diploma, or certificate would have been available for a year more recent than 1996 (such as 1997 or 1998), it would have been better to use it.

6.3.5.3 Estimated number of university graduates from 1998 to 2018 in the population

The number of doctoral graduates produced in the future has to be taken into account when we want to estimate the potential number of full-time university teachers in the future. Recall that the number of doctoral graduates estimated is for those aged 34 or less at the time they obtain their doctorate degree. We assumed that the doctoral graduates aged 34 years old and less were more likely than doctoral graduates aged 35 years old and over to become full-time university teachers after their graduation. It is at this step that the fitted conditional probabilities obtained from the proportional odds model in chapter 5, section 5.3 were used. First, recall that the subpopulation of interest (for the fitting of this model) encompassed the university graduates (from 1986, 1990, and 1995) who were Canadian residents two years after their graduation, and for which the gender, the age two years after graduation and the region of residence two years after their graduation were known. So for each graduation year (1986, 1990, and 1995), the estimated probabilities of having obtained a doctorate at university graduation given the graduate's age two years after graduation, the graduate's gender and the graduate's region of residence two years after graduation were obtained. We assumed that if we multiplied these estimated probabilities (for each age category, gender, and region of residence) by the number of Canadian university graduates at the same year (for each respective age category, gender, and region of residence), we would obtain an estimate of the number of Canadian doctoral graduates by age category, gender, and region of residence for the year of graduation.

Since the estimated number of Canadian university graduates from 1998 to 2018 (aged 34 years old and less) was not known, it had to be estimated. From the 1988 NGS, the 1992 NGS, and the 1997 NGS data, the estimated proportions of university graduates by age category, gender, and region of residence were obtained. But only the 1995 proportions of university graduates by age category, region, and gender were used to estimate the number of

university graduates from 1998 to 2018. In doing this, we assumed that the proportions of university graduates by age category, region, and gender will remain constant from 1998 to 2018, and will have the same values as the 1995 proportions. It is expected that these proportions, by age category, gender, and region of residence, will either increase from 1998 to 2018, or decrease, or will have some other pattern. The prediction of how these proportions would fluctuate in the coming years would have required a deeper analysis. But we assumed that if the proportion of university graduates do not differ too much from the 1995 proportions, the estimated number of graduates from 1998 to 2018 would be not too far from reality. The 1995 proportions were chosen as opposed to the 1988 and the 1992 proportions because they were the most recent available.

So the estimated number of university graduates from 1998 to 2018 by age category at graduation, gender, and region of residence at graduation were obtained from the product of the 1995 proportions and the estimated number of people in population, by age category, gender, and region of residence. It must be noted that the proportions obtained are the proportions of having obtained a doctorate degree in 1995, given the graduate's age two years after graduation, the graduate's gender, and the graduate's region of residence two years after graduation. These are not exactly the proportions given the age at graduation, the gender, and the region of residence at graduation. So we assume that the number of university graduates given the age at graduation, the gender, and the region of residence at graduation from 1998 to 2018 was well estimated even though the proportions were not exactly comparable to those we should have used if the age and region of residence at graduation would have been used.

The population projections used to obtain the estimated number of university graduates were obtained from the Demography Division at Statistics Canada. For each year from 2000 to 2018, by age category (except the 35 year-old and over category), gender, and region of

residence, four different projections were provided according to four different assumptions, based on growth, fertility, life expectancy, and immigration. We decided to provide the estimated number of university graduates in future for these four assumptions. In 1998 and 1999, population estimates were provided, but not for different assumptions (since they were estimates and not projections).

6.3.5.4 Estimated number of doctoral graduates from 1998 to 2018 in the population

Once the numbers of university graduates aged 34 years old and less, by age category, gender, and region of residence, were estimated from 1998 to 2018, they could be multiplied by the fitted probabilities of having obtained a doctorate degree in 1995, given the graduate's age category two years after graduation, the graduate's gender, and the graduate's region of residence two years after graduation. Again in this case, the age categories and the regions of residence of the two terms of the product are not exactly comparable (as discussed in a previous paragraph). The 1995 conditional probabilities were used, because they were the most recent probabilities available. And again, these probabilities were assumed to be constant (and equal to the 1995 probabilities) from 1998 to 2018. A prediction on how these probabilities would fluctuate in future, by age category, gender, and region of residence, would be an interesting study. It was assumed that the probabilities of obtaining a doctorate when one is a Canadian university graduate from 1998 to 2018 will not vary too much from the 1995 conditional probabilities, given the age category, the gender, and the region of residence.

Finally, the estimated number of Canadian doctoral graduates aged 34 years old and less by age category, gender, and region of residence from 1998 to 2018 were obtained. But we know that it is not all of these doctoral graduates that may become full-time university

teachers. So the proportion of people with a doctorate in Canada who are full-time university teachers in 1996 (as used before, i.e. 30.2%) was again multiplied to the number of doctoral graduates produced per year from 1998 to 2018. So we assumed that at each year, from 1998 to 2018, the same proportion of doctoral graduates will become full-time university teachers at the year following the year they obtained their doctoral degree. It might happen that doctoral graduates, if they become university teachers, do not become university teachers the year following the year they obtained their doctoral degree. It was one of the ways of adding these potential university teachers to the existing number of full-time university teachers (after retirement and attrition).

6.3.5.5 Final estimates

Finally, after many assumptions and different estimations obtained, the estimated numbers of full-time university teachers in Canada (non-medical/dental only) from 1998 to 2018 were obtained. These estimates decreased from approximately 28,000 in 1998 to approximately 9,250 in 2018. In this analysis, a certain proportion of doctoral graduates was assumed to become full-time university teachers. If it decreases, from 1998 to 2018, these estimates would be even lower. This research did not take account of the fact that people with a doctorate could become part-time university teachers, instead of full-time university teachers. So the number of university teachers, full-time and part-time, might not decrease that much. But if most of the future university teachers are hired as part-time teachers, other problems would be caused. A last thing to take into consideration when using these numbers is that a certain proportion of doctoral graduates are assumed to become Canadian full-time university teachers. This study did not analyze the effect on this proportion of doctoral graduates who leave the country after their graduation. Consequently, if too many doctoral graduates leave the country after their graduation, the proportion of doctoral graduates who

are assumed to become Canadian full-time university teachers could decrease.

6.5.3.6 Conclusion

In conclusion, these estimated numbers of full-time university teachers by year from 1998 to 2018 are based on several assumptions and must be used with caution. Given the data we had, the best was done to provide the most accurate profile of full-time university teachers in the future. But other estimates may be obtained using other assumptions. Further areas of research are given in the next section, that could help to improve these results.

6.4 Further research

Methods for validating a proportional odds model using survey data, such as general goodness-of-fit methods and residuals analysis, should be developed. This would help in improving the confidence in the results that are obtained when the fitting of such a model is done. Different combinations of predictors could be used as explanatory variables when fitting a proportional odds models with response being the university level of certification obtained at graduation. Then the best set of explanatory variables, in the sense that they explain better what can affect the university level of certification obtained at graduation, should be determined. The parental level of education should be included as one of the explanatory variables to assess its effect.

Estimates of the number of full-time university teachers from 1998 to 2018 could be obtained using different assumptions. The fluctuations of the proportions of university graduates by age category, gender, and region of residence in future could be studied and used to estimate the number of university graduates from 1998 to 2018. Similarly, the changes in the probabilities of obtaining a doctorate at university graduation, given the graduate's age category, gender, and region of residence could also be studied and used to

estimate the number of doctoral graduates by age category, gender, and region of residence from 1998 to 2018. This would make the estimated number of full-time university teachers from 1998 to 2018 more accurate. It would also be important to use the age of the graduate and the region of residence of the graduate at the time of the graduation, and not at the time of the survey, to ensure a better comparability between data about graduates and population projections. A study on the part-time university teachers that could become full-time professors should be done in order to add these people to the number of potential full-time university teachers in the future. And finally, if data are available, the estimated number of full-time university teachers in the future by field of study taught could be obtained. This would allow the identification of field of study that might have shortages of full-time university teachers in the future. But the prediction of shortages of full-time university teachers would have to be studied jointly with the demand for a university education. Some factors affecting enrolment have been identified in this research but enrolment trends in the future should be obtained. And when the proportions of Canadians enrolled in a university program, for some characteristics, are estimated, we must make sure that the data are comparable by year (using either the civil year for enrolment data and population estimates or school year for enrolment data and population estimates).

BIBLIOGRAPHY

- Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons, New York.
- Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*. 68, 13-20.
- Bélanger, R. and Omiecinski, T. (1987). Part-time university enrolment. *Canadian Social Trends*, summer 1987, 22-25.
- Binder, D.A., Gratton, M., Hidioglou, M.A., Kumar, S. and Rao, J.N.K. (1983). Analysis of categorical data from surveys with complex designs: some canadian experiences. Revised and expanded version of that presented at the Seminar on Recent Developments in the Analysis of Large Scale Data Sets sponsored by Statistical Office of European Communities, Luxembourg.
- Binder, D. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279-292.
- Butlin, G (1999). Determinants of postsecondary participation. *Education Quarterly Review*, 5, no. 3, 9-35.
- Cohen, S.B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician*, 51, 285-292.
- Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford Science Publications, Oxford.
- Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*, Chapman & Hall, London.
- Full-time University Teaching Staff Data Elements Manual (1998). Culture, Tourism and the Centre for Education Statistics, Statistics Canada, Ottawa.
- Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya*, C 37, 117-132.

- Gonzalez, A. (1996). Graduation rates and times to completion for doctoral programs in Canada. *Education Quarterly Review*, 3, no. 2, 45-56.
- Hajék, J. (1960). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Pub. Math. Inst. Hung. Acad. Sci.*, 35, 1491-1523.
- Krewski, D. and Rao J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.
- Landis and al. (1987). Cumulative logit models for weighted data from complex sample surveys. *Proceedings of the Sociological Statistics Section (ASA)*, 165-170.
- Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*. 79, 61-83.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Madow, W.G. (1948). On the limiting distribution of estimates based on samples from finite universes. *Ann. Math. Statist.*, 19, 535-545.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society (B)*, 42, 109-142.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models (second edition)*. Chapman & Hall, London.
- Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models (fourth edition)*. Irwin, United States of America.
- Pierce, D.A. and Schafer, D.W. (1986). Residuals in Generalized Linear Models. *Journal of the American Statistical Association*, 81, no. 396, 977-986.

Powers, D.A. and Xie, Y. (1999). *Statistical Methods for Categorical Data Analysis*.

Academic Press, New York.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*. 9, 705-724.

Rao, J.N.K., Kumar, S. and Roberts, G. (1989). Analysis of sample survey data involving categorical response variables: methods and software. *Survey Methodology*, 15, no. 2, 161- 186.

Roberts, G. (1985). Contributions to Chi-Squared Tests with Survey Data. Unpublished Ph.

D. Thesis, Carleton University, Department of Mathematics and Statistics, Ottawa.

Rose, A. (1994). Returning to school full-time. *Education Quarterly Review*, 1, no. 2, 33-40.

Rosen, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I and II. *Ann. Math. Statist.*, 43, 373-397 and 748-776.

Shah, B.V., Folsom, R.E., LaVange, L.M., Wheelless, S.C., Boyle, K.E., Williams, R.L.

(1993). Statistical Methods and Mathematical Algorithms used in SUDAAN. Research triangle Institute, North Carolina.

SUDAAN User's Manual (Release 7.5), Research Triangle Institute, North Carolina.

Survey of 1986 Graduates (1990). Survey methodology report and user's guide. Household Surveys Division, Statistics Canada, Ottawa.

Survey of 1990 Graduates (1992). Microdata User's Guide. Special Surveys Group, Statistics Canada, Ottawa.

Survey of 1995 Graduates (1997). Survey methodology report and user's guide. Special Surveys Division, Statistics Canada, Ottawa.

The Class of 86 Revisited (1996). A compendium of findings of the 1991 Follow-up of 1986 Graduates Survey with comparisons to the 1988 National Graduates Survey, Human Resources Development Canada and Statistics Canada.

The Class of 90 (1996). A compendium of findings from the 1992 National Graduates

Survey of 1990 Graduates, Human Resources Development Canada and Statistics Canada.

The Class of '95 (1999). Report of the 1997 National Survey of 1995 Graduates, Human Resources Development Canada and Statistics Canada, Statistics Canada Catalogue 81-584-XPB.

Thomas, D.R. and Rao, J.N.K. (1987). Small-Sample Comparisons of Level and Power for Simple Goodness-of-fit Statistics Under Cluster Sampling. *Journal of the American Statistical Association*, 82, 630-636.

University Student Information System: User Guide (1995). Postsecondary Education Section, Education, Culture and Tourism Division, Statistics Canada, Ottawa.

Von Bahr, B. (1972). On sampling from a finite set of independent random variables. *Z. Wahr. verw. Geb.*, 24, 279-286.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54, 167-179.

Zeger, S.L., Liang, K.-Y., Albert, P.S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, 44, 1049-1060.

Appendix A: Field of study categories

University field of study categories (1986):

1. No specialization/ Specialization unknown
2. Education
3. Fine arts & humanities
4. Commerce, law and economics
5. Other social sciences
6. Agriculture and biological sciences
7. Engineering
8. Medical and health professionals
9. Mathematics and physical sciences

College field of study categories (1986):

1. No specialization, arts, humanities, unknown
2. Health and related sciences
3. Chemical technologies, transportation technologies, general engineering, aeronautical engineering, industrial engineering
4. Electrical and electronic technologies, mathematics and computer science
5. Mechanical engineering, architectural and construction engineering
6. Natural sciences and primary industries
7. Social sciences and services
8. Secretarial sciences, merchandising and sales, service industry technologies, miscellaneous

9. Management and administration

Trade/vocational field of study categories (1986):

1. No specialization, arts, arts and science, transportation technology, merchandising and sales, service industry technology
2. Health services and related, social sciences and services
3. Electrical and electronic technologies
4. Automotive mechanics
5. Other mechanical
6. General engineering, architectural and construction engineering
7. Engineering technologies, chemical technology, architectural design, drafting technology, industrial engineering
8. Natural sciences and primary industries
9. Journalism, secretarial science
10. Mathematics and computer science, business and commerce, management and administration

Trade/vocational field of study categories (1995):

1. No specialization, unknown, arts, transportation technology, merchandising and sales, service industry technology
2. Health services and related, social sciences and services
3. Electrical and electronic technologies
4. Automotive mechanics, other mechanical
5. General engineering, architectural and construction engineering, engineering

technologies, chemical technology, architectural design, drafting technology, industrial engineering

6. Natural sciences and primary industries
7. Journalism, secretarial science
8. Mathematics and computer science, business and commerce, management and administration

Appendix B: Creation of the status of registration variable for the NGS 1988

A) Here are the SAS commands which created the derived variable about the status of registration for the 1986 graduates. But first is a description of the variables used:

Q11: Were you ever registered as a part-time student for the ... program? (1=Yes, 2=No)

Q14: Check with values 1 if Q11=1 or Q13=1 (Q13 is not related to the status of registration)

Q16A: How much of this time (when "this time" refers to the length of the program) were you enrolled full-time (in months)? (99=not stated)

Q16B: How much of this time were you enrolled full-time (in years)? (99=not stated)

Q16C: How much of this time were you enrolled part-time (in months)? (99=not stated)

Q16D: How much of this time were you enrolled part-time (in years)? (99=not stated)

```
IF Q14='1' THEN DO;
```

```
IF NOT (Q16C='99' OR Q16C=' ' OR Q16C='00') AND (Q16A='99' OR Q16A=' ' OR  
Q16A='00') AND (Q16B='99' OR Q16B=' ' OR Q16B='00') AND (Q16D='99' OR  
Q16D=' ' OR Q16D='00') THEN DO;
```

```
FTPT='02';
```

```
END;
```

```
ELSE IF NOT (Q16D='99' OR Q16D=' ' OR Q16D='00') AND (Q16A='99' OR Q16A=' '  
OR Q16A='00') AND (Q16B='99' OR Q16B=' ' OR Q16B='00') AND (Q16C='99' OR  
Q16C=' ' OR Q16C='00') THEN DO;
```

```
FTPT='02';
```

```
END;
```

ELSE IF NOT (Q16A='99' OR Q16A=' ' OR Q16A='00') AND NOT (Q16C='99' OR Q16C=' ' OR Q16C='00') AND (Q16D='99' OR Q16D=' ' OR Q16D='00') AND (Q16B='99' OR Q16B=' ' OR Q16B='00') THEN DO;

FTPT='03';

END;

ELSE IF NOT (Q16A='99' OR Q16A=' ' OR Q16A='00') AND (Q16C='99' OR Q16C=' ' OR Q16C='00') AND NOT (Q16D='99' OR Q16D=' ' OR Q16D='00') AND (Q16B='99' OR Q16B=' ' OR Q16B='00') THEN DO;

FTPT='03';

END;

ELSE IF NOT (Q16B='99' OR Q16B=' ' OR Q16B='00') AND NOT (Q16C='99' OR Q16C=' ' OR Q16C='00') AND (Q16D='99' OR Q16D=' ' OR Q16D='00') AND (Q16A='99' OR Q16A=' ' OR Q16A='00') THEN DO;

FTPT='03';

END;

ELSE IF NOT (Q16B='99' OR Q16B=' ' OR Q16B='00') AND (Q16C='99' OR Q16C=' ' OR Q16C='00') AND NOT (Q16D='99' OR Q16D=' ' OR Q16D='00') AND (Q16A='99' OR Q16A=' ' OR Q16A='00') THEN DO;

FTPT='03';

END;

else IF NOT (Q16A='99' OR Q16A=' ' OR Q16A='00') AND (Q16B='99' OR Q16B=' ' OR Q16B='00')

AND (Q16C='99' OR Q16C=' ' OR Q16C='00') AND (Q16D='99' OR Q16D=' ' OR Q16D='00') AND

Q11^='1' THEN DO;

FTPT='01';

```

END;

ELSE IF NOT (Q16B='99' OR Q16B=' ' OR Q16B='00') AND (Q16A='99' OR Q16A=' '
OR Q16A='00') AND (Q16C='99' OR Q16C=' ' OR Q16C='00') AND (Q16D='99' OR
Q16D=' ' OR Q16D='00') AND Q11^='1' THEN DO;
  FTPT='01';
END;

ELSE IF Q14='1' AND Q11='2' THEN FTPT='01';
ELSE IF Q14='1' AND Q11='1' AND (NOT (Q16B='99' OR Q16B=' ' OR Q16B='00')
OR
NOT (Q16A='99' OR Q16A=' ' OR Q16A='00')) THEN FTPT='97';
ELSE FTPT='97';
end;

ELSE IF Q14='2' THEN DO;
  IF Q11='2' THEN FTPT='01';
  ELSE IF Q11='9' THEN FTPT='97';
  ELSE IF Q11=' ' THEN FTPT='97';
  ELSE FTPT='97';
END;

ELSE DO;
  FTPT='97';
END;

```

The variable FTPT takes the value '01' when the graduate was enrolled full-time only; '02' when the graduate was enrolled part-time only; '03' when the graduate was enrolled both part-time and full-time; and '97' when it is impossible to determine the status of registration.

Appendix C: Assumptions of Binder's approach

First, suppose there is a sequence of populations, indexed by $t = 1, 2, 3, \dots$. The population sizes are N_t . The function $W_N(\theta)$ is defined for all $\theta \in \Theta$, the parameter space.

Condition 1. We assume $\lim_{t \rightarrow \infty} \frac{W_{N_t}(\theta)}{N_t}$ exists and equals $\Omega(\theta)$.

Condition 2. We assume $\Omega(\theta)$ is a one-to-one function, so that $\Omega^{-1}(\cdot)$ exists.

Condition 3. There exists a $\theta_0 \in \Theta$ such that $\Omega(\theta_0) = 0$.

Condition 4. The parameter space Θ contains a neighbourhood of θ_0 .

We can now suppose that a sequence of sample of sizes $n_1, \dots$ are taken and that for each sample we construct the function $\widehat{W}(\theta)$, defined in (3.1).

Condition 5 We assume $\frac{n^{1/2}}{N_t} [\widehat{W}(\theta) - W_{N_t}(\theta)]$ converges in distribution to the normal law with mean 0 and positive-definite variance matrix $\Phi(\theta)$ as $t \rightarrow \infty$, for all θ in a neighbourhood of θ_0 .

Conditions required for condition 5 to be satisfied are given in the literature; see, for example, Madow (1948), Hajék (1960, 1964), Rosen (1972), von Bahr (1972), Fuller (1975), Krewski & Rao (1981).

Condition 6. We assume $\widehat{W}(\cdot)$ is totally differentiable in a neighbourhood of θ_0 .

Condition 7. We assume $\lim_{\theta \rightarrow \theta_0, t \rightarrow \infty} \left[N_t^{-1} \left(\frac{\partial \widehat{W}(\theta)}{\partial \theta} \right) \right]$ exists and equals

$$\frac{\partial \Omega(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = J(\theta_0).$$

Condition 8. The matrix $J(\theta_0)\Phi(\theta_0)J(\theta_0)$ is full rank.

Condition 9. We assume $\Phi(\theta)$ is a continuous function of θ .

Condition 10. We assume $\frac{\partial \Omega(\theta)}{\partial \theta}$ is a continuous function of θ .

Condition 11. A consistent estimator for $\Phi(\theta)$ is $\widehat{\Phi}(\theta)$, assumed to exist under the sample design.

Lemma 1. The asymptotic distribution of $n_t^{1/2}(\widehat{\theta} - \theta_0)$ is the same as the

asymptotic distribution of $\mathbf{J}^{-1}(\theta_0) \frac{n_t^{1/2}}{N_t} \widehat{\mathbf{W}}(\theta_0)$ (the proof is in Binder's article)

Corollary 1. The asymptotic distribution of $n_t^{1/2} (\widehat{\theta} - \theta_0)$ is the normal law with mean 0 and variance matrix $[\mathbf{J}^{-1}(\theta_0)][\Phi(\theta_0)][\mathbf{J}^{-1}(\theta_0)]'$. This is the limiting value of (3.3).

Corollary 2. Let F_t be the distribution function for $n_t^{1/2} (\widehat{\theta} - \theta_0)$, based on the t^{th} sample and let G_t be the distribution function of a multivariate normal distribution with mean zero and variance matrix

$[\widehat{\mathbf{J}}^{-1}(\widehat{\theta})][\widehat{\Phi}(\widehat{\theta})][\widehat{\mathbf{J}}^{-1}(\widehat{\theta})]'$, where $\widehat{\mathbf{J}}(\theta) = N_t^{-1} \frac{\partial \widehat{\mathbf{W}}(\theta)}{\partial \theta}$. Then, by virtue of conditions 9 to 11,

$$\lim_{t \rightarrow \infty} \sup |F_t - G_t| = 0.$$

Appendix D: SUDAAN's output from the proportional odds model (Chapter 4)

CERST: level of certification

GNDN =Gender

CATAGE = Age category

REG = Region of residence

FOS = Field of study

LANG =First language spoken and still understood

1) 1988 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

CERST (cum-logit)				
Intercept 1:	2.27	0.40	0.39	5.66
Intercept 2:	5.04	0.40	0.38	12.55
FOS				
1:	-0.47	0.22	0.36	-2.09
2:	-1.19	0.21	0.35	-5.71
3:	-0.85	0.22	0.37	-3.91
4:	-0.28	0.24	0.41	-1.16
5:	-0.71	0.22	0.36	-3.26
6:	-1.61	0.33	0.43	-4.86
7:	-1.48	0.37	0.48	-4.01
8:	-0.31	0.21	0.29	-1.47
9:	0.00	0.00	.	.
GNDN				
1:	-1.25	0.17	0.39	-7.31
2:	0.00	0.00	.	.
REG				
1:	-0.18	0.34	0.39	-0.52
2:	0.24	0.34	0.40	0.71
3:	-0.40	0.34	0.40	-1.17
4:	-0.09	0.34	0.40	-0.28
5:	-0.16	0.35	0.40	-0.48
6:	0.00	0.00	.	.

```

FTPT
  1:  -0.41   0.06   0.70  -6.48
  2:   0.11   0.08   0.73   1.35
  3:   0.00   0.00   .     .
LANG
  1:  -0.23   0.07   0.69  -3.31
  2:   0.00   0.00   .     .
CATAGE
  1:   3.61   0.42   0.25   8.51
  2:   0.53   0.20   0.39   2.60
  3:  -0.62   0.26   0.49  -2.40
  4:   0.00   0.00   .     .
FOS, GNDN
  1,  1:   0.27   0.20   0.42   1.34
  1,  2:   0.00   0.00   .     .
  2,  1:   0.67   0.20   0.44   3.40
  2,  2:   0.00   0.00   .     .
  3,  1:   0.55   0.19   0.42   2.96
  3,  2:   0.00   0.00   .     .
  4,  1:   0.10   0.19   0.44   0.52
  4,  2:   0.00   0.00   .     .
  5,  1:   0.26   0.18   0.42   1.39
  5,  2:   0.00   0.00   .     .
  6,  1:   0.61   0.20   0.30   3.06
  6,  2:   0.00   0.00   .     .
  7,  1:   0.35   0.25   0.39   1.41
  7,  2:   0.00   0.00   .     .
  8,  1:   0.49   0.19   0.31   2.63
  8,  2:   0.00   0.00   .     .
  9,  1:   0.00   0.00   .     .
  9,  2:   0.00   0.00   .     .
FOS, CATAGE
  1,  1:   2.17   0.49   0.15   4.42
  1,  2:   1.41   0.24   0.40   5.95
  1,  3:   0.64   0.31   0.56   2.05
  1,  4:   0.00   0.00   .     .
  2,  1:   2.83   0.87   0.23   3.25
  2,  2:   2.09   0.24   0.41   8.80
  2,  3:   1.16   0.28   0.53   4.08
  2,  4:   0.00   0.00   .     .
  3,  1:   0.81   0.52   0.29   1.57
  3,  2:   0.42   0.23   0.45   1.80
  3,  3:   0.49   0.29   0.56   1.66
  3,  4:   0.00   0.00   .     .
  4,  1:   1.07   0.43   0.21   2.47
  4,  2:   0.16   0.23   0.48   0.66
  4,  3:   0.28   0.30   0.61   0.93
  4,  4:   0.00   0.00   .     .
  5,  1:   0.88   0.51   0.30   1.73
  5,  2:   0.18   0.23   0.44   0.81
  5,  3:   0.16   0.29   0.56   0.55
  5,  4:   0.00   0.00   .     .
  6,  1:  26.72   0.44   0.00  60.60
  6,  2:   1.14   0.34   0.47   3.35
  6,  3:   0.92   0.40   0.54   2.27
  6,  4:   0.00   0.00   .     .
  7,  1:   2.33   0.61   0.23   3.83
  7,  2:   1.35   0.33   0.55   4.12
  7,  3:   1.01   0.39   0.61   2.62
  7,  4:   0.00   0.00   .     .
  8,  1:   0.41   0.42   0.11   0.97

```

8,	2:	0.25	0.23	0.32	1.09
8,	3:	0.24	0.29	0.42	0.83
8,	4:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
9,	3:	0.00	0.00	.	.
9,	4:	0.00	0.00	.	.
GNDN, CATAGE					
1,	1:	0.56	0.33	0.31	1.70
1,	2:	0.80	0.12	0.73	6.64
1,	3:	0.44	0.14	0.85	3.19
1,	4:	0.00	0.00	.	.
2,	1:	0.00	0.00	.	.
2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
FOS
1: 0.0363
2: 0.0000
3: 0.0001
4: 0.2444
5: 0.0011
6: 0.0000
7: 0.0001
8: 0.1429
9: .
GNDN
1: 0.0000
2: .
REG
1: 0.6037
2: 0.4752
3: 0.2428
4: 0.7825
5: 0.6329
6: .
FTPT
1: 0.0000
2: 0.1758
3: .
LANG
1: 0.0009
2: .
CATAGE
1: 0.0000
2: 0.0093

```

```

3: 0.0166
4: .
FOS, GNDN
1, 1: 0.1802
1, 2: .
2, 1: 0.0007
2, 2: .
3, 1: 0.0031
3, 2: .
4, 1: 0.5996
4, 2: .
5, 1: 0.1644
5, 2: .
6, 1: 0.0022
6, 2: .
7, 1: 0.1590
7, 2: .
8, 1: 0.0085
8, 2: .
9, 1: .
9, 2: .
FOS, CATAGE
1, 1: 0.0000
1, 2: 0.0000
1, 3: 0.0402
1, 4: .
2, 1: 0.0012
2, 2: 0.0000
2, 3: 0.0000
2, 4: .
3, 1: 0.1167
3, 2: 0.0717
3, 3: 0.0978
3, 4: .
4, 1: 0.0137
4, 2: 0.5093
4, 3: 0.3539
4, 4: .
5, 1: 0.0829
5, 2: 0.4197
5, 3: 0.5819
5, 4: .
6, 1: 0.0000
6, 2: 0.0008
6, 3: 0.0233
6, 4: .
7, 1: 0.0001
7, 2: 0.0000
7, 3: 0.0088
7, 4: .
8, 1: 0.3297
8, 2: 0.2743
8, 3: 0.4049
8, 4: .
9, 1: .
9, 2: .
9, 3: .
9, 4: .
GNDN, CATAGE
1, 1: 0.0895
1, 2: 0.0000

```

1, 3: 0.0014
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .

2) 1992 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

CERST (cum-logit)					
Intercept	1:	0.66	0.37	0.41	1.79
Intercept	2:	3.23	0.37	0.41	8.76
FOS					
1:	1.40	0.27	0.40	5.16	
2:	-0.28	0.26	0.63	-1.08	
3:	0.12	0.27	0.63	0.45	
4:	0.36	0.28	0.66	1.30	
5:	0.40	0.26	0.63	1.50	
6:	-0.72	0.34	0.65	-2.15	
7:	-1.00	0.38	0.70	-2.65	
8:	0.54	0.27	0.54	2.01	
9:	0.00	0.00	.	.	
GNDN					
1:	-0.91	0.18	0.52	-5.17	
2:	0.00	0.00	.	.	
REG					
1:	0.31	0.26	0.30	1.18	
2:	0.09	0.26	0.30	0.36	
3:	0.27	0.26	0.30	1.02	
4:	0.55	0.26	0.30	2.10	
5:	0.27	0.26	0.30	1.01	
6:	0.00	0.00	.	.	
FTPT					
1:	-0.15	0.06	0.75	-2.65	
2:	-0.00	0.08	0.87	-0.04	
3:	0.00	0.00	.	.	
LANG					
1:	-0.33	0.07	0.83	-4.50	
2:	0.00	0.00	.	.	
CATAGE					
1:	4.71	0.53	0.36	8.88	
2:	0.96	0.25	0.71	3.83	
3:	-0.67	0.30	0.80	-2.26	
4:	0.00	0.00	.	.	
FOS, GNDN					
1, 1:	0.49	0.20	0.22	2.41	
1, 2:	0.00	0.00	.	.	
2, 1:	0.36	0.20	0.57	1.83	
2, 2:	0.00	0.00	.	.	
3, 1:	0.11	0.19	0.54	0.56	
3, 2:	0.00	0.00	.	.	
4, 1:	-0.22	0.18	0.54	-1.20	
4, 2:	0.00	0.00	.	.	

5,	1:	0.13	0.19	0.54	0.72
5,	2:	0.00	0.00	.	.
6,	1:	-0.52	0.20	0.42	-2.61
6,	2:	0.00	0.00	.	.
7,	1:	0.36	0.23	0.46	1.57
7,	2:	0.00	0.00	.	.
8,	1:	0.03	0.19	0.39	0.17
8,	2:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
FOS, CATAGE					
1,	1:	-2.35	0.46	0.18	-5.09
1,	2:	-0.20	0.28	0.30	-0.69
1,	3:	0.52	0.38	0.45	1.36
1,	4:	0.00	0.00	.	.
2,	1:	1.52	0.56	0.13	2.72
2,	2:	1.82	0.28	0.68	6.60
2,	3:	1.56	0.32	0.81	4.91
2,	4:	0.00	0.00	.	.
3,	1:	0.01	0.58	0.36	0.01
3,	2:	0.29	0.27	0.73	1.07
3,	3:	0.65	0.32	0.84	1.99
3,	4:	0.00	0.00	.	.
4,	1:	-0.14	0.50	0.34	-0.27
4,	2:	0.35	0.27	0.77	1.29
4,	3:	0.50	0.32	0.87	1.54
4,	4:	0.00	0.00	.	.
5,	1:	1.42	0.85	0.38	1.67
5,	2:	0.05	0.27	0.73	0.19
5,	3:	0.32	0.32	0.85	0.98
5,	4:	0.00	0.00	.	.
6,	1:	3.41	0.84	0.15	4.08
6,	2:	1.28	0.34	0.69	3.78
6,	3:	1.26	0.39	0.76	3.20
6,	4:	0.00	0.00	.	.
7,	1:	2.87	0.80	0.13	3.57
7,	2:	1.11	0.36	0.82	3.12
7,	3:	1.29	0.41	0.88	3.17
7,	4:	0.00	0.00	.	.
8,	1:	1.34	0.67	0.07	2.00
8,	2:	-0.00	0.27	0.55	-0.01
8,	3:	0.34	0.33	0.65	1.02
8,	4:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
9,	3:	0.00	0.00	.	.
9,	4:	0.00	0.00	.	.
GNDN, CATAGE					
1,	1:	-0.07	0.40	0.46	-0.17
1,	2:	0.82	0.12	0.81	6.78
1,	3:	0.50	0.14	0.96	3.54
1,	4:	0.00	0.00	.	.
2,	1:	0.00	0.00	.	.
2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0736
Intercept 2: 0.0000
FOS
1: 0.0000
2: 0.2781
3: 0.6503
4: 0.1925
5: 0.1346
6: 0.0312
7: 0.0082
8: 0.0445
9: .
GNDN
1: 0.0000
2: .
REG
1: 0.2372
2: 0.7160
3: 0.3068
4: 0.0356
5: 0.3116
6: .
FTPT
1: 0.0081
2: 0.9700
3: .
LANG
1: 0.0000
2: .
CATAGE
1: 0.0000
2: 0.0001
3: 0.0240
4: .
FOS, GNDN
1, 1: 0.0162
1, 2: .
2, 1: 0.0670
2, 2: .
3, 1: 0.5733
3, 2: .
4, 1: 0.2314
4, 2: .
5, 1: 0.4709
5, 2: .
6, 1: 0.0089
6, 2: .
7, 1: 0.1154
7, 2: .
8, 1: 0.8639
8, 2: .
9, 1: .

```

```

9, 2: .
FOS, CATAGE
1, 1: 0.0000
1, 2: 0.4898
1, 3: 0.1742
1, 4: .
2, 1: 0.0066
2, 2: 0.0000
2, 3: 0.0000
2, 4: .
3, 1: 0.9891
3, 2: 0.2865
3, 3: 0.0471
3, 4: .
4, 1: 0.7847
4, 2: 0.1957
4, 3: 0.1232
4, 4: .
5, 1: 0.0953
5, 2: 0.8524
5, 3: 0.3247
5, 4: .
6, 1: 0.0000
6, 2: 0.0002
6, 3: 0.0014
6, 4: .
7, 1: 0.0004
7, 2: 0.0018
7, 3: 0.0015
7, 4: .
8, 1: 0.0460
8, 2: 0.9939
8, 3: 0.3076
8, 4: .
9, 1: .
9, 2: .
9, 3: .
9, 4: .
GNDN, CATAGE
1, 1: 0.8657
1, 2: 0.0000
1, 3: 0.0004
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .

```

3) 1997 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

CERST (cum-logit)					
Intercept	1:	1.78	0.50	0.35	3.57
Intercept	2:	4.32	0.50	0.35	8.67
FOS					
1:	2.21	0.19	0.22	11.77	
2:	-0.22	0.19	0.63	-1.16	
3:	0.25	0.20	0.61	1.27	
4:	0.60	0.20	0.66	2.94	
5:	0.16	0.20	0.63	0.84	
6:	-0.57	0.28	0.70	-2.01	
7:	-0.44	0.26	0.67	-1.72	
8:	0.28	0.19	0.52	1.48	
9:	0.00	0.00	.	.	
GNDN					
1:	-0.66	0.17	0.69	-3.97	
2:	0.00	0.00	.	.	
REG					
1:	-0.91	0.46	0.33	-1.97	
2:	-0.70	0.46	0.33	-1.51	
3:	-0.94	0.46	0.33	-2.03	
4:	-0.83	0.46	0.33	-1.81	
5:	-0.84	0.46	0.33	-1.81	
6:	0.00	0.00	.	.	
FTPT					
1:	-0.31	0.07	1.37	-4.11	
2:	0.15	0.09	1.55	1.61	
3:	0.00	0.00	.	.	
LANG					
1:	-0.45	0.10	1.40	-4.46	
2:	0.00	0.00	.	.	
CATAGE					
1:	4.58	0.60	0.29	7.63	
2:	1.03	0.18	0.63	5.72	
3:	-0.27	0.24	0.86	-1.14	
4:	0.00	0.00	.	.	
FOS, GNDN					
1, 1:	-0.40	0.22	0.23	-1.78	
1, 2:	0.00	0.00	.	.	
2, 1:	0.55	0.21	0.89	2.69	
2, 2:	0.00	0.00	.	.	
3, 1:	-0.05	0.19	0.77	-0.28	
3, 2:	0.00	0.00	.	.	
4, 1:	-0.22	0.19	0.77	-1.18	
4, 2:	0.00	0.00	.	.	

5,	1:	0.04	0.19	0.80	0.20
5,	2:	0.00	0.00	.	.
6,	1:	-0.57	0.21	0.58	-2.71
6,	2:	0.00	0.00	.	.
7,	1:	0.01	0.20	0.63	0.05
7,	2:	0.00	0.00	.	.
8,	1:	-0.50	0.19	0.63	-2.55
8,	2:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
FOS, CATAGE					
1,	1:	24.00	0.61	0.00	39.14
1,	2:	-0.11	0.22	0.16	-0.49
1,	3:	-0.25	0.30	0.26	-0.82
1,	4:	0.00	0.00	.	.
2,	1:	1.97	0.84	0.24	2.35
2,	2:	1.40	0.22	0.78	6.32
2,	3:	0.85	0.27	1.00	3.12
2,	4:	0.00	0.00	.	.
3,	1:	0.64	0.79	0.34	0.81
3,	2:	0.20	0.21	0.72	0.95
3,	3:	0.26	0.29	1.02	0.92
3,	4:	0.00	0.00	.	.
4,	1:	-1.12	0.71	0.38	-1.59
4,	2:	-0.22	0.21	0.77	-1.04
4,	3:	-0.28	0.28	1.04	-1.02
4,	4:	0.00	0.00	.	.
5,	1:	0.53	0.74	0.35	0.71
5,	2:	0.36	0.21	0.74	1.71
5,	3:	0.37	0.29	1.09	1.26
5,	4:	0.00	0.00	.	.
6,	1:	3.14	0.84	0.11	3.71
6,	2:	1.25	0.30	0.76	4.20
6,	3:	0.11	0.38	0.92	0.30
6,	4:	0.00	0.00	.	.
7,	1:	26.39	0.66	0.00	39.74
7,	2:	0.51	0.24	0.76	2.12
7,	3:	0.32	0.30	0.93	1.07
7,	4:	0.00	0.00	.	.
8,	1:	-0.25	0.90	0.38	-0.28
8,	2:	-0.03	0.21	0.58	-0.14
8,	3:	0.25	0.27	0.79	0.93
8,	4:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
9,	3:	0.00	0.00	.	.
9,	4:	0.00	0.00	.	.
GNDN, CATAGE					
1,	1:	1.27	0.44	0.49	2.87
1,	2:	0.77	0.13	1.32	6.09
1,	3:	0.45	0.16	1.65	2.89
1,	4:	0.00	0.00	.	.
2,	1:	0.00	0.00	.	.
2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.

First column: **parameter level**

Second column: **p -value**

```

CERST (cum-logit)
Intercept 1: 0.0004
Intercept 2: 0.0000
FOS
1: 0.0000
2: 0.2448
3: 0.2039
4: 0.0033
5: 0.4020
6: 0.0443
7: 0.0857
8: 0.1402
9: .
GNDN
1: 0.0001
2: .
REG
1: 0.0491
2: 0.1319
3: 0.0428
4: 0.0710
5: 0.0706
6: .
FTPT
1: 0.0000
2: 0.1076
3: .
LANG
1: 0.0000
2: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.2559
4: .
FOS, GNDN
1, 1: 0.0749
1, 2: .
2, 1: 0.0071
2, 2: .
3, 1: 0.7799
3, 2: .
4, 1: 0.2371
4, 2: .
5, 1: 0.8438
5, 2: .
6, 1: 0.0068
6, 2: .
7, 1: 0.9639
7, 2: .
8, 1: 0.0107
8, 2: .
9, 1: .

```

```
9, 2: .  
FOS, CATAGE  
1, 1: 0.0000  
1, 2: 0.6224  
1, 3: 0.4104  
1, 4: .  
2, 1: 0.0187  
2, 2: 0.0000  
2, 3: 0.0018  
2, 4: .  
3, 1: 0.4158  
3, 2: 0.3440  
3, 3: 0.3588  
3, 4: .  
4, 1: 0.1128  
4, 2: 0.2991  
4, 3: 0.3080  
4, 4: .  
5, 1: 0.4751  
5, 2: 0.0880  
5, 3: 0.2071  
5, 4: .  
6, 1: 0.0002  
6, 2: 0.0000  
6, 3: 0.7640  
6, 4: .  
7, 1: 0.0000  
7, 2: 0.0341  
7, 3: 0.2830  
7, 4: .  
8, 1: 0.7829  
8, 2: 0.8866  
8, 3: 0.3531  
8, 4: .  
9, 1: .  
9, 2: .  
9, 3: .  
9, 4: .  
GNDN, CATAGE  
1, 1: 0.0041  
1, 2: 0.0000  
1, 3: 0.0039  
1, 4: .  
2, 1: .  
2, 2: .  
2, 3: .  
2, 4: .
```

4) 1997 model (administrative counts)

First column: parameter level

Second column: parameter estimate

Third column: standard error

Fourth column: design effect

Fifth column: t: beta=0

CERST (cum-logit)					
Intercept	1:	1.78	0.51	0.36	3.52
Intercept	2:	4.32	0.51	0.36	8.53
FOS					
1:	2.21	0.19	0.23	11.44	
2:	-0.22	0.20	0.65	-1.14	
3:	0.25	0.20	0.64	1.25	
4:	0.60	0.21	0.69	2.88	
5:	0.16	0.20	0.66	0.82	
6:	-0.57	0.29	0.73	-1.97	
7:	-0.44	0.26	0.70	-1.69	
8:	0.28	0.19	0.55	1.44	
9:	0.00	0.00	.	.	
GNDN					
1:	-0.66	0.17	0.71	-3.89	
2:	0.00	0.00	.	.	
REG					
1:	-0.91	0.47	0.34	-1.94	
2:	-0.70	0.47	0.34	-1.49	
3:	-0.94	0.47	0.34	-2.00	
4:	-0.83	0.47	0.34	-1.78	
5:	-0.84	0.47	0.34	-1.78	
6:	0.00	0.00	.	.	
FTPT					
1:	-0.31	0.08	1.39	-4.09	
2:	0.15	0.09	1.56	1.60	
3:	0.00	0.00	.	.	
LANG					
1:	-0.45	0.10	1.42	-4.42	
2:	0.00	0.00	.	.	
CATAGE					
1:	4.58	0.62	0.31	7.38	
2:	1.03	0.18	0.66	5.61	
3:	-0.27	0.24	0.90	-1.11	
4:	0.00	0.00	.	.	
FOS, GNDN					
1, 1:	-0.40	0.23	0.25	-1.72	
1, 2:	0.00	0.00	.	.	
2, 1:	0.55	0.21	0.92	2.66	
2, 2:	0.00	0.00	.	.	
3, 1:	-0.05	0.20	0.80	-0.27	
3, 2:	0.00	0.00	.	.	
4, 1:	-0.22	0.19	0.80	-1.16	
4, 2:	0.00	0.00	.	.	

5,	1:	0.04	0.20	0.83	0.19
5,	2:	0.00	0.00	.	.
6,	1:	-0.57	0.22	0.61	-2.64
6,	2:	0.00	0.00	.	.
7,	1:	0.01	0.21	0.65	0.04
7,	2:	0.00	0.00	.	.
8,	1:	-0.50	0.20	0.66	-2.51
8,	2:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
FOS, CATAGE					
1,	1:	24.00	0.63	0.00	37.82
1,	2:	-0.11	0.23	0.17	-0.48
1,	3:	-0.25	0.31	0.28	-0.80
1,	4:	0.00	0.00	.	.
2,	1:	1.97	0.86	0.26	2.29
2,	2:	1.40	0.23	0.80	6.23
2,	3:	0.85	0.28	1.04	3.07
2,	4:	0.00	0.00	.	.
3,	1:	0.64	0.81	0.36	0.79
3,	2:	0.20	0.21	0.75	0.93
3,	3:	0.26	0.29	1.05	0.90
3,	4:	0.00	0.00	.	.
4,	1:	-1.12	0.73	0.40	-1.54
4,	2:	-0.22	0.21	0.80	-1.02
4,	3:	-0.28	0.28	1.08	-1.00
4,	4:	0.00	0.00	.	.
5,	1:	0.53	0.76	0.37	0.70
5,	2:	0.36	0.21	0.77	1.68
5,	3:	0.37	0.29	1.13	1.24
5,	4:	0.00	0.00	.	.
6,	1:	3.14	0.91	0.13	3.46
6,	2:	1.25	0.30	0.79	4.13
6,	3:	0.11	0.38	0.96	0.29
6,	4:	0.00	0.00	.	.
7,	1:	26.39	0.69	0.00	38.45
7,	2:	0.51	0.25	0.78	2.08
7,	3:	0.32	0.31	0.97	1.05
7,	4:	0.00	0.00	.	.
8,	1:	-0.25	0.93	0.40	-0.27
8,	2:	-0.03	0.21	0.60	-0.14
8,	3:	0.25	0.28	0.82	0.91
8,	4:	0.00	0.00	.	.
9,	1:	0.00	0.00	.	.
9,	2:	0.00	0.00	.	.
9,	3:	0.00	0.00	.	.
9,	4:	0.00	0.00	.	.
GNDN, CATAGE					
1,	1:	1.27	0.45	0.50	2.82
1,	2:	0.77	0.13	1.34	6.05
1,	3:	0.45	0.16	1.67	2.87
1,	4:	0.00	0.00	.	.
2,	1:	0.00	0.00	.	.
2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0004
Intercept 2: 0.0000
FOS
1: 0.0000
2: 0.2542
3: 0.2130
4: 0.0039
5: 0.4111
6: 0.0484
7: 0.0917
8: 0.1491
9: .
GNDN
1: 0.0001
2: .
REG
1: 0.0525
2: 0.1375
3: 0.0459
4: 0.0752
5: 0.0748
6: .
FTPT
1: 0.0000
2: 0.1092
3: .
LANG
1: 0.0000
2: .
medskip CATAGE
1: 0.0000
2: 0.0000
3: 0.2667
4: .
FOS, GNDN
1, 1: 0.0855
1, 2: .
2, 1: 0.0079
2, 2: .
3, 1: 0.7834
3, 2: .
4, 1: 0.2451
4, 2: .
5, 1: 0.8461
5, 2: .
6, 1: 0.0082
6, 2: .
7, 1: 0.9646
7, 2: .
8, 1 0.0122
8, 2: .
9, 1: .

```

```
9, 2: .
FOS, CATAGE
1, 1: 0.0000
1, 2: 0.6346
1, 3: 0.4260
1, 4: .
2, 1: 0.0219
2, 2: 0.0000
2, 3: 0.0021
2, 4: .
3, 1: 0.4293
3, 2: 0.3521
3, 3: 0.3670
3, 4: .
4, 1: 0.1234
4, 2: 0.3070
4, 3: 0.3166
4, 4: .
5, 1: 0.4864
5, 2: 0.0931
5, 3: 0.2145
5, 4: .
6, 1: 0.0005
6, 2: 0.0000
6, 3: 0.7691
6, 4: .
7, 1: 0.0000
7, 2: 0.0374
7, 3: 0.2934
7, 4: .
8, 1: 0.7877
8, 2: 0.8887
8, 3: 0.3633
8, 4: .
9, 1: .
9, 2: .
9, 3: .
9, 4: .
GNDN, CATAGE
1, 1: 0.0047
1, 2: 0.0000
1, 3: 0.0041
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
```

APPENDIX E: SUDAAN's output from the proportional odds model (Chapter 5, without region of institution)

CERST: level of certification

GNDN=Gender

CATAGE= Age category

REGCO= Region of residence, with BC and Yukon/NWT collapsed

1) 1988 model

First column: parameter level

Second column: parameter estimate

Third column standard error

Fourth column: design effect

Fifth column: t: beta=0

CERST (cum-logit)					
Intercept	1:	0.75	0.10	0.37	7.18
Intercept	2:	3.51	0.10	0.31	33.51
GNDN					
1:	-0.83	0.14	0.51	-5.87	
2:	0.00	0.00	.	.	
REGCO					
1:	0.27	0.13	0.25	2.11	
2:	1.33	0.12	0.39	11.13	
3:	0.16	0.13	0.46	1.28	
4:	0.22	0.12	0.31	1.84	
5:	0.00	0.00	.	.	
CATAGE					
1:	5.22	0.50	0.20	10.50	
2:	1.68	0.14	0.40	11.60	
3:	0.38	0.17	0.52	2.20	
4:	0.00	0.00	.	.	
GNDN, REGCO					
1, 1:	0.09	0.16	0.31	0.59	
1, 2:	-0.18	0.15	0.49	-1.20	
1, 3:	-0.07	0.15	0.52	-0.45	
1, 4:	0.18	0.14	0.37	1.25	
1, 5:	0.00	0.00	.	.	
2, 1:	0.00	0.00	.	.	

2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.
2,	5:	0.00	0.00	.	.
GNDN, CATAGE					
1,	1:	0.50	0.27	0.26	1.87
1,	2:	0.71	0.10	0.74	6.75
1,	3:	0.28	0.13	0.88	2.19
1,	4:	0.00	0.00	.	.
2,	1:	0.00	0.00	.	.
2,	2:	0.00	0.00	.	.
2,	3:	0.00	0.00	.	.
2,	4:	0.00	0.00	.	.
REGCO, CATAGE					
1,	1:	-0.53	0.51	0.15	-1.04
1,	2:	-0.42	0.18	0.29	-2.36
1,	3:	-0.07	0.21	0.34	-0.33
1,	4:	0.00	0.00	.	.
2,	1:	-1.29	0.50	0.17	-2.56
2,	2:	-1.36	0.17	0.45	-8.23
2,	3:	-0.88	0.20	0.59	-4.47
2,	4:	0.00	0.00	.	.
3,	1:	-0.50	0.50	0.19	-0.99
3,	2:	-0.49	0.17	0.48	-2.91
3,	3:	-0.49	0.21	0.66	-2.35
3,	4:	0.00	0.00	.	.
4,	1:	0.34	0.60	0.18	0.56
4,	2:	-0.35	0.16	0.34	-2.16
4,	3:	-0.16	0.19	0.43	-0.86
4,	4:	0.00	0.00	.	.
5,	1:	0.00	0.00	.	.
5,	2:	0.00	0.00	.	.
5,	3:	0.00	0.00	.	.
5,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST      (cum-logit)
  Intercept 1: 0.0000
  Intercept 2: 0.0000
GNDN
  1: 0.0000
  2: .
REGCO
  1: 0.0345
  2: 0.0000
  3: 0.1988
  4: 0.0661
  5: .
CATAGE
  1: 0.0000
  2: 0.0000
  3: 0.0279
  4: .
GNDN, REGCO
  1, 1: 0.5549
  1, 2: 0.2282
  1, 3: 0.6528
  1, 4: 0.2114
  1, 5: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
  2, 5: .
GNDN, CATAGE
  1, 1: 0.0620
  1, 2: 0.0000
  1, 3: 0.0286
  1, 4: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
REGCO, CATAGE
  1, 1: 0.2980
  1, 2: 0.0183
  1, 3: 0.7393
  1, 4: .
  2, 1: 0.0105
  2, 2: 0.0000
  2, 3: 0.0000
  2, 4: .
  3, 1: 0.3241
  3, 2: 0.0037
  3, 3: 0.0186
  3, 4: .
  4, 1: 0.5722
  4, 2: 0.0311
  4, 3: 0.3917

```

```

4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

2) 1992 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST (cum-logit)
Intercept 1: 1.04 0.08 0.40 12.56
Intercept 2: 3.55 0.08 0.33 42.38
GNDN
1: -0.75 0.12 0.57 -6.14
2: 0.00 0.00 . .
REGCO
1: -0.12 0.11 0.28 -1.12
2: -0.45 0.11 0.48 -4.10
3: 0.10 0.11 0.54 0.91
4: 0.11 0.10 0.33 1.16
5: 0.00 0.00 . .
CATAGE
1: 5.37 0.50 0.21 10.65
2: 1.33 0.12 0.45 10.88
3: 0.12 0.16 0.60 0.74
4: 0.00 0.00 . .
GNDN, REGCO
1, 1: -0.12 0.14 0.35 -0.83
1, 2: 0.12 0.13 0.56 0.91
1, 3: -0.17 0.14 0.64 -1.22
1, 4: -0.16 0.13 0.42 -1.23
1, 5: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: -0.18 0.32 0.35 -0.56
1, 2: 0.58 0.10 0.81 5.64
1, 3: 0.26 0.13 0.99 2.05
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .

```

REGCO,	CATAGE				
1,	1:	0.17	0.56	0.17	0.29
1,	2:	0.36	0.15	0.30	2.37
1,	3:	0.20	0.19	0.40	1.05
1,	4:	0.00	0.00	.	.
2,	1:	1.00	0.75	0.27	1.33
2,	2:	0.28	0.15	0.51	1.89
2,	3:	-0.03	0.19	0.66	-0.16
2,	4:	0.00	0.00	.	.
3,	1:	-0.94	0.53	0.24	-1.77
3,	2:	-0.09	0.15	0.57	-0.58
3,	3:	-0.32	0.19	0.78	-1.64
3,	4:	0.00	0.00	.	.
4,	1:	0.15	0.60	0.21	0.26
4,	2:	0.13	0.14	0.37	0.91
4,	3:	0.35	0.18	0.48	1.98
4,	4:	0.00	0.00	.	.
5,	1:	0.00	0.00	.	.
5,	2:	0.00	0.00	.	.
5,	3:	0.00	0.00	.	.
5,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST      (cum-logit)
  Intercept 1: 0.0000
  Intercept 2: 0.0000
GNDN
  1: 0.0000
  2: .
REGCO
  1: 0.2633
  2: 0.0000
  3: 0.3622
  4: 0.2462
  5: .
CATAGE
  1: 0.0000
  2: 0.0000
  3: 0.4564
  4: .
GNDN,      REGCO
  1, 1: 0.4092
  1, 2: 0.3619
  1, 3: 0.2237
  1, 4: 0.2181
  1, 5: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
  2, 5: .
GNDN,      CATAGE
  1, 1: 0.5776
  1, 2: 0.0000
  1, 3: 0.0403
  1, 4: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
REGCO,      CATAGE
  1, 1: 0.7695
  1, 2: 0.0177
  1, 3: 0.2959
  1, 4: .
  2, 1: 0.1820
  2, 2: 0.0586
  2, 3: 0.8742
  2, 4: .
  3, 1: 0.0773
  3, 2: 0.5643
  3, 3: 0.1012
  3, 4: .
  4, 1: 0.7970
  4, 2: 0.3646
  4, 3: 0.0474

```

```

4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

3) 1997 model

First column: parameter level

Second column: parameter estimate

Third column: standard error

Fourth column: design effect

Fifth column: t: beta=0

```

CERST (cum-logit)
Intercept 1: 0.93 0.07 0.39 12.58
Intercept 2: 3.42 0.07 0.34 45.92
GNDN
1: -0.83 0.12 0.63 -7.11
2: 0.00 0.00 . .
REGCO
1: -0.43 0.09 0.28 -4.54
2: 0.58 0.09 0.51 6.23
3: -0.09 0.12 0.82 -0.71
4: -0.20 0.09 0.33 -2.38
5: 0.00 0.00 . .
CATAGE
1: 5.10 0.25 0.05 20.55
2: 1.50 0.11 0.42 14.13
3: 0.06 0.15 0.65 0.42
4: 0.00 0.00 . .
GNDN, REGCO
1, 1: 0.35 0.13 0.34 2.82
1, 2: 0.18 0.13 0.69 1.40
1, 3: 0.07 0.14 0.78 0.50
1, 4: 0.16 0.11 0.41 1.37
1, 5: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: 1.08 0.38 0.39 2.83
1, 2: 0.53 0.12 1.46 4.59
1, 3: 0.30 0.15 1.75 2.07
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .

```

REGCO,	CATAGE				
1,	1:	0.60	0.45	0.09	1.33
1,	2:	0.37	0.13	0.28	2.92
1,	3:	0.50	0.19	0.45	2.68
1,	4:	0.00	0.00	.	.
2,	1:	-0.86	0.53	0.19	-1.61
2,	2:	-1.01	0.13	0.58	-7.51
2,	3:	-0.41	0.18	0.85	-2.22
2,	4:	0.00	0.00	.	.
3,	1:	-0.53	0.42	0.13	-1.27
3,	2:	-0.05	0.15	0.74	-0.34
3,	3:	-0.18	0.21	1.07	-0.87
3,	4:	0.00	0.00	.	.
4,	1:	0.14	0.30	0.05	0.49
4,	2:	0.39	0.12	0.33	3.26
4,	3:	0.26	0.16	0.52	1.61
4,	4:	0.00	0.00	.	.
5,	1:	0.00	0.00	.	.
5,	2:	0.00	0.00	.	.
5,	3:	0.00	0.00	.	.
5,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST      (cum-logit)
  Intercept 1: 0.0000
  Intercept 2: 0.0000
GNDN
  1: 0.0000
  2: .
REGCO
  1: 0.0000
  2: 0.0000
  3: 0.4759
  4: 0.0172
  5: .
CATAGE
  1: 0.0000
  2: 0.0000
  3: 0.6739
  4: .
GNDN,      REGCO
  1, 1: 0.0048
  1, 2: 0.1601
  1, 3: 0.6140
  1, 4: 0.1699
  1, 5: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
  2, 5: .
GNDN,      CATAGE
  1, 1: 0.0047
  1, 2: 0.0000
  1, 3: 0.0388
  1, 4: .
  2, 1: .
  2, 2: .
  2, 3: .
  2, 4: .
REGCO,      CATAGE
  1, 1: 0.1844
  1, 2: 0.0036
  1, 3: 0.0073
  1, 4: .
  2, 1: 0.1074
  2, 2: 0.0000
  2, 3: 0.0263
  2, 4: .
  3, 1: 0.2056
  3, 2: 0.7362
  3, 3: 0.3855
  3, 4: .
  4, 1: 0.6256
  4, 2: 0.0011
  4, 3: 0.1065

```

```

4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

4) 1997 (administrative counts)

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST      (cum-logit)
Intercept  1:  0.93   0.07   0.41  12.41
Intercept  2:  3.42   0.08   0.35  45.29
GNDN
 1:  -0.83   0.12   0.65  -6.99
 2:   0.00   0.00   .      .
REGCO
 1:  -0.43   0.10   0.29  -4.46
 2:   0.58   0.09   0.53   6.17
 3:  -0.09   0.12   0.83  -0.71
 4:  -0.20   0.09   0.34  -2.35
 5:   0.00   0.00   .      .
CATAGE
 1:   5.10   0.30   0.07  17.27
 2:   1.50   0.11   0.44  13.85
 3:   0.06   0.15   0.68   0.41
 4:   0.00   0.00   .      .
GNDN,      REGCO
 1,  1:   0.35   0.13   0.36   2.75
 1,  2:   0.18   0.13   0.71   1.38
 1,  3:   0.07   0.14   0.80   0.50
 1,  4:   0.16   0.12   0.43   1.34
 1,  5:   0.00   0.00   .      .
 2,  1:   0.00   0.00   .      .
 2,  2:   0.00   0.00   .      .
 2,  3:   0.00   0.00   .      .
 2,  4:   0.00   0.00   .      .
 2,  5:   0.00   0.00   .      .
GNDN,      CATAGE
 1,  1:   1.08   0.39   0.41   2.77
 1,  2:   0.53   0.12   1.47   4.56
 1,  3:   0.30   0.15   1.77   2.06
 1,  4:   0.00   0.00   .      .
 2,  1:   0.00   0.00   .      .
 2,  2:   0.00   0.00   .      .
 2,  3:   0.00   0.00   .      .
 2,  4:   0.00   0.00   .      .

```

REGCO,	CATAGE				
1,	1:	0.60	0.49	0.11	1.23
1,	2:	0.37	0.13	0.30	2.83
1,	3:	0.50	0.19	0.47	2.63
1,	4:	0.00	0.00	.	.
2,	1:	-0.86	0.56	0.21	-1.53
2,	2:	-1.01	0.14	0.60	-7.40
2,	3:	-0.41	0.19	0.88	-2.19
2,	4:	0.00	0.00	.	.
3,	1:	-0.53	0.45	0.15	-1.18
3,	2:	-0.05	0.15	0.75	-0.33
3,	3:	-0.18	0.21	1.10	-0.86
3,	4:	0.00	0.00	.	.
4,	1:	0.14	0.34	0.07	0.42
4,	2:	0.39	0.12	0.34	3.18
4,	3:	0.26	0.16	0.54	1.58
4,	4:	0.00	0.00	.	.
5,	1:	0.00	0.00	.	.
5,	2:	0.00	0.00	.	.
5,	3:	0.00	0.00	.	.
5,	4:	0.00	0.00	.	.

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
GNDN
1: 0.0000
2: .
REGCO
1: 0.0000
2: 0.0000
3: 0.4790
4: 0.0189
5: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.6792
4: .
GNDN, REGCO
1, 1: 0.0060
1, 2: 0.1663
1, 3: 0.6187
1, 4: 0.1792
1, 5: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
2, 5: .
GNDN, CATAGE
1, 1: 0.0056
1, 2: 0.0000
1, 3: 0.0399
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
REGCO, CATAGE
1, 1: 0.2194
1, 2: 0.0046
1, 3: 0.0086
1, 4: .
2, 1: 0.1255
2, 2: 0.0000
2, 3: 0.0284
2, 4: .
3, 1: 0.2398
3, 2: 0.7391
3, 3: 0.3906
3, 4: .
4, 1: 0.6742
4, 2: 0.0015
4, 3: 0.1134

```

4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

APPENDIX F: SUDAAN's output from the proportional odds model (Chapter 5, with region of institution)

CERST: level of certification

GNDN=Gender

CATAGE= Age category

REGCO= Region of residence, with BC and Yukon/NWT collapsed

REGIN=Resion of institution

1) 1988 model

First column: parameter level
Second column: parameter estimate
Third column standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST(cum-logit)
Intercept 1: 0.72 0.10 0.36 6.90
Intercept 2: 3.48 0.10 0.31 33.37
GNDN
1: -0.84 0.14 0.50 -5.96
2: 0.00 0.00 . .
REGCO
1: -0.34 0.23 0.40 -1.52
2: 1.06 0.24 0.63 4.35
3: 0.14 0.21 0.62 0.64
4: -0.26 0.19 0.43 -1.34
5: 0.00 0.00 . .
CATAGE
1: 5.21 0.50 0.20 10.46
2: 1.67 0.15 0.41 11.35
3: 0.35 0.18 0.54 1.98
4: 0.00 0.00 . .
REGIN
1: 0.68 0.17 0.40 3.97
2: 0.31 0.20 0.63 1.54
3: 0.05 0.15 0.57 0.30
4: 0.52 0.13 0.39 3.90
5: 0.00 0.00 . .

```

```

GNDN, REGCO
1, 1: 0.13 0.16 0.31 0.81
1, 2: -0.16 0.15 0.49 -1.0 6
1, 3: -0.05 0.15 0.52 -0.3 7
1, 4: 0.20 0.14 0.38 1.37
1, 5: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: 0.49 0.27 0.27 1.85
1, 2: 0.69 0.11 0.75 6.59
1, 3: 0.27 0.13 0.89 2.12
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
REGCO, CATAGE
1, 1: -0.54 0.51 0.15 -1.0 4
1, 2: -0.38 0.18 0.30 -2.1 2
1, 3: -0.04 0.21 0.35 -0.1 9
1, 4: 0.00 0.00 . .
2, 1: -1.27 0.50 0.17 -2.5 2
2, 2: -1.34 0.17 0.45 -8.0 2
2, 3: -0.85 0.20 0.61 -4.2 4
2, 4: 0.00 0.00 . .
3, 1: -0.50 0.51 0.20 -0.9 9
3, 2: -0.48 0.17 0.49 -2.8 0
3, 3: -0.47 0.21 0.68 -2.2 0
3, 4: 0.00 0.00 . .
4, 1: 0.34 0.60 0.18 0.57
4, 2: -0.29 0.16 0.35 -1.7 7
4, 3: -0.12 0.19 0.45 -0.6 1
4, 4: 0.00 0.00 . .
5, 1: 0.00 0.00 . .
5, 2: 0.00 0.00 . .
5, 3: 0.00 0.00 . .
5, 4: 0.00 0.00 . .

```

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
GNDN
1: 0.0000
2: .
REGCO
1: 0.1296
2: 0.0000

```

```

3: 0.5219
4: 0.1787
5: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.0476
4: .
REGIN
1: 0.0001
2: 0.1229
3: 0.7673
4: 0.0001
5: .
GNDN, REGCO
1, 1: 0.4187
1, 2: 0.2906
1, 3: 0.7101
1, 4: 0.1715
1, 5: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
2, 5: .
GNDN, CATAGE
1, 1: 0.0643
1, 2: 0.0000
1, 3: 0.0342
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
REGCO, CATAGE
1, 1: 0.2961
1, 2: 0.0342
1, 3: 0.8479
1, 4: .
2, 1: 0.0116
2, 2: 0.0000
2, 3: 0.0000
2, 4: .
3, 1: 0.3241
3, 2: 0.0051
3, 3: 0.0277
3, 4: .
4, 1: 0.5698
4, 2: 0.0764
4, 3: 0.5402
4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

2) 1992 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST (cum-logit)
Intercept 1: 1.04 0.08 0.36 12.81
Intercept 2: 3.56 0.08 0.31 43.32
GNDN
1: -0.75 0.12 0.59 -5.98
2: 0.00 0.00 . .
REGCO
1: -0.19 0.17 0.40 -1.07
2: -0.02 0.20 0.69 -0.08
3: 0.50 0.18 0.74 2.81
4: -0.28 0.16 0.48 -1.72
5: 0.00 0.00 . .
CATAGE
1: 5.38 0.51 0.21 10.63
2: 1.37 0.13 0.47 10.82
3: 0.12 0.16 0.62 0.73
4: 0.00 0.00 . .
REGIN
1: 0.09 0.14 0.40 0.65
2: -0.44 0.16 0.62 -2.75
3: -0.42 0.13 0.64 -3.09
4: 0.42 0.12 0.42 3.47
5: 0.00 0.00 . .
GNDN, REGCO
1, 1: -0.11 0.14 0.36 -0.75
1, 2: 0.12 0.14 0.58 0.91
1, 3: -0.17 0.14 0.66 -1.26
1, 4: -0.14 0.13 0.44 -1.01
1, 5: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: -0.19 0.32 0.35 -0.58
1, 2: 0.58 0.10 0.83 5.54
1, 3: 0.26 0.13 1.02 1.99
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
REGCO, CATAGE
1, 1: 0.16 0.57 0.17 0.29
1, 2: 0.37 0.15 0.30 2.38
1, 3: 0.23 0.20 0.42 1.13
1, 4: 0.00 0.00 . .
  
```

```

2, 1: 0.98 0.75 0.27 1.31
2, 2: 0.25 0.15 0.52 1.64
2, 3: -0.03 0.19 0.68 -0.16
2, 4: 0.00 0.00 . .
3, 1: -0.96 0.54 0.24 -1.78
3, 2: -0.12 0.15 0.59 -0.77
3, 3: -0.33 0.20 0.81 -1.64
3, 4: 0.00 0.00 . .
4, 1: 0.18 0.60 0.21 0.30
4, 2: 0.15 0.15 0.39 1.00
4, 3: 0.37 0.18 0.51 2.06
4, 4: 0.00 0.00 . .
5, 1: 0.00 0.00 . .
5, 2: 0.00 0.00 . .
5, 3: 0.00 0.00 . .
5, 4: 0.00 0.00 . .

```

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
GNDN
1: 0.0000
2: .
REGCO
1: 0.2861
2: 0.9355
3: 0.0050
4: 0.0863
5: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.4682
4: .
REGIN
1: 0.5139
2: 0.0060
3: 0.0020
4: 0.0005
5: .
GNDN, REGCO
1, 1: 0.4544
1, 2: 0.3645
1, 3: 0.2092
1, 4: 0.3106
1, 5: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .

```

```

2, 5: .
GNDN, CATAGE
1, 1: 0.5646
1, 2: 0.0000
1, 3: 0.0469
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
REGCO, CATAGE
1, 1: 0.7717
1, 2: 0.0171
1, 3: 0.2602
1, 4: .
2, 1: 0.1899
2, 2: 0.1004
2, 3: 0.8753
2, 4: .
3, 1: 0.0743
3, 2: 0.4385
3, 3: 0.1003
3, 4: .
4, 1: 0.7650
4, 2: 0.3196
4, 3: 0.0398
4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

3) 1997 model

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST (cum-logit)
Intercept 1: 0.92 0.07 0.39 12.58
Intercept 2: 3.41 0.07 0.34 46.10
GNDN
1: -0.83 0.12 0.63 -7.11
2: 0.00 0.00 . .
REGCO
1: -0.86 0.25 0.54 -3.49
2: 0.72 0.36 1.12 2.01
3: 0.10 0.24 0.75 0.40
4: -0.56 0.22 0.55 -2.55

```

```

5: 0.00 0.00 . .
CATAGE
1: 5.11 0.25 0.05 20.31
2: 1.50 0.11 0.42 14.13
3: 0.05 0.15 0.66 0.36
4: 0.00 0.00 . .
REGIN
1: 0.46 0.22 0.53 2.12
2: -0.13 0.34 1.13 -0.40
3: -0.18 0.20 0.64 -0.91
4: 0.36 0.20 0.58 1.79
5: 0.00 0.00 . .
GNDN, REGCO
1, 1: 0.34 0.13 0.35 2.67
1, 2: 0.18 0.13 0.69 1.42
1, 3: 0.07 0.14 0.79 0.49
1, 4: 0.17 0.12 0.41 1.48
1, 5: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: 1.08 0.38 0.39 2.83
1, 2: 0.53 0.12 1.46 4.56
1, 3: 0.30 0.15 1.76 2.05
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
REGCO, CATAGE
1, 1: 0.59 0.45 0.09 1.30
1, 2: 0.40 0.13 0.27 3.12
1, 3: 0.56 0.19 0.47 2.91
1, 4: 0.00 0.00 . .
2, 1: -0.86 0.53 0.19 -1.61
2, 2: -1.01 0.13 0.57 -7.49
2, 3: -0.40 0.18 0.86 -2.18
2, 4: 0.00 0.00 . .
3, 1: -0.53 0.42 0.13 -1.27
3, 2: -0.05 0.15 0.73 -0.32
3, 3: -0.17 0.21 1.08 -0.84
3, 4: 0.00 0.00 . .
4, 1: 0.15 0.30 0.05 0.50
4, 2: 0.41 0.12 0.32 3.51
4, 3: 0.30 0.16 0.52 1.84
4, 4: 0.00 0.00 . .
5, 1: 0.00 0.00 . .
5, 2: 0.00 0.00 . .
5, 3: 0.00 0.00 . .
5, 4: 0.00 0.00 . .

```

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
GNDN
1: 0.0000
2: .
REGCO
1: 0.0005
2: 0.0446
3: 0.6915
4: 0.0109
5: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.7157
4: .
REGIN
1: 0.0340
2: 0.6926
3: 0.3653
4: 0.0727
5: .
GNDN, REGCO
1, 1: 0.0077
1, 2: 0.1549
1, 3: 0.6248
1, 4: 0.1380
1, 5: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
2, 5: .
GNDN, CATAGE
1, 1: 0.0047
1, 2: 0.0000
1, 3: 0.0402
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
REGCO, CATAGE
1, 1: 0.1936
1, 2: 0.0018
1, 3: 0.0036
1, 4: .
2, 1: 0.1075
2, 2: 0.0000
2, 3: 0.0296
2, 4: .
3, 1: 0.2033

```



```

3, 2: 0.7452
3, 3: 0.3989
3, 4: .
4, 1: 0.6152
4, 2: 0.0004
4, 3: 0.0656
4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .

```

4) 1997 model (administrative counts)

First column: parameter level
Second column: parameter estimate
Third column: standard error
Fourth column: design effect
Fifth column: t: beta=0

```

CERST (cum-logit)
Intercept 1: 0.92 0.07 0.40 12.41
Intercept 2: 3.41 0.08 0.35 45.49
GNDN
1: -0.83 0.12 0.65 -6.99
2: 0.00 0.00 . .
REGCO
1: -0.86 0.25 0.57 -3.41
2: 0.72 0.37 1.15 1.98
3: 0.10 0.24 0.78 0.39
4: -0.56 0.22 0.57 -2.49
5: 0.00 0.00 . .
CATAGE
1: 5.11 0.30 0.07 17.11
2: 1.50 0.11 0.43 13.83
3: 0.05 0.15 0.68 0.36
4: 0.00 0.00 . .
REGIN
1: 0.46 0.22 0.56 2.06
2: -0.13 0.35 1.17 -0.39
3: -0.18 0.20 0.67 -0.88
4: 0.36 0.20 0.61 1.76
5: 0.00 0.00 . .
GNDN, REGCO
1, 1: 0.34 0.13 0.37 2.60
1, 2: 0.18 0.13 0.71 1.40
1, 3: 0.07 0.14 0.81 0.48
1, 4: 0.17 0.12 0.43 1.45
1, 5: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .

```

```

2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
2, 5: 0.00 0.00 . .
GNDN, CATAGE
1, 1: 1.08 0.39 0.41 2.77
1, 2: 0.53 0.12 1.48 4.54
1, 3: 0.30 0.15 1.78 2.04
1, 4: 0.00 0.00 . .
2, 1: 0.00 0.00 . .
2, 2: 0.00 0.00 . .
2, 3: 0.00 0.00 . .
2, 4: 0.00 0.00 . .
REGCO, CATAGE
1, 1: 0.59 0.49 0.11 1.20
1, 2: 0.40 0.13 0.29 3.03
1, 3: 0.56 0.20 0.49 2.85
1, 4: 0.00 0.00 . .
2, 1: -0.86 0.56 0.21 -1.53
2, 2: -1.01 0.14 0.59 -7.38
2, 3: -0.40 0.19 0.88 -2.15
2, 4: 0.00 0.00 . .
3, 1: -0.53 0.45 0.15 -1.18
3, 2: -0.05 0.15 0.75 -0.32
3, 3: -0.17 0.21 1.10 -0.83
3, 4: 0.00 0.00 . .
4, 1: 0.15 0.35 0.07 0.43
4, 2: 0.41 0.12 0.33 3.43
4, 3: 0.30 0.17 0.54 1.81
4, 4: 0.00 0.00 . .
5, 1: 0.00 0.00 . .
5, 2: 0.00 0.00 . .
5, 3: 0.00 0.00 . .
5, 4: 0.00 0.00 . .

```

First column: parameter level

Second column: p -value

```

CERST (cum-logit)
Intercept 1: 0.0000
Intercept 2: 0.0000
GNDN
1: 0.0000
2: .
REGCO
1: 0.0007
2: 0.0478
3: 0.6974
4: 0.0128
5: .
CATAGE
1: 0.0000
2: 0.0000
3: 0.7204

```

```
4: .
REGIN
1: 0.0389
2: 0.6969
3: 0.3768
4: 0.0792
5: .
GNDN, REGCO
1, 1: 0.0093
1, 2: 0.1610
1, 3: 0.6293
1, 4: 0.1465
1, 5: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
2, 5: .
GNDN, CATAGE
1, 1: 0.0056
1, 2: 0.0000
1, 3: 0.0413
1, 4: .
2, 1: .
2, 2: .
2, 3: .
2, 4: .
REGCO, CATAGE
1, 1: 0.2289
1, 2: 0.0024
1, 3: 0.0043
1, 4: .
2, 1: 0.1256
2, 2: 0.0000
2, 3: 0.0317
2, 4: .
3, 1: 0.2372
3, 2: 0.7481
3, 3: 0.4039
3, 4: .
4, 1: 0.6645
4, 2: 0.0006
4, 3: 0.0709
4, 4: .
5, 1: .
5, 2: .
5, 3: .
5, 4: .
```

APPENDIX G: Figures

Figure 4.1: Deviance residual by its index for $X = x_i$: $Y = 1$ (1983)

Figure 4.2: Deviance residual by its index for $X = x_i$: $Y = 2$ (1988)

Figure 4.3: Deviance residual by its index for $X = x_i$: $Y = 3$ (1983)

Figure 4.4: Deviance residual by its index for $X = x_i$: $Y = 1$ (1992)

Figure 4.5: Deviance residual by its index for $X = x_i$: $Y = 2$ (1992)

Figure 4.6: Deviance residual by its index for $X = x_i$: $Y = 3$ (1992)

Figure 4.7: Deviance residual by its index for $X = x_i$: $Y = 1$ (1997)

Figure 4.8: Deviance residual by its index for $X = x_i$: $Y = 2$ (1997)

Figure 4.9: Deviance residual by its index for $X = x_i: Y = 3$ (1997)

Figure 4.10: Deviance residual by its index for $X = x_i: Y = 1$ (1997, administrative counts)

Figure 4.11: Deviance residual by its index for $X = x_i: Y = 2$ (1997, administrative counts)

Figure 4.12: Deviance residual by its index for $X = x_i: Y = 3$ (1997, administrative counts)

Figure 5.1: Canadian undergraduate enrolment by year for several age categories

Figure 5.2: Canadian graduate enrolment by year for several age categories

Figure 5.3: Enrolment totals by year of enrolment for different categories of age (undergraduates) in the Atlantic provinces

Figure 5.4: Enrolment totals by year of enrolment for different categories of age (undergraduates) in Quebec

Figure 5.5: Enrolment totals by year of enrolment for different categories of age (undergraduates) in Ontario

Figure 5.6: Enrolment totals by year of enrolment for different categories of age (undergraduates) in the Prairies

Figure 5.7: Enrolment totals by year of enrolment for different categories of age (undergraduates) in British Columbia

Figure 5.8: Enrolment totals by year of enrolment for different categories of age (graduates) in the Atlantic provinces

Figure 5.9: Enrolment totals by year of enrolment for different categories of age (graduates) in Quebec

Figure 5.10: Enrolment totals by year of enrolment for different categories of age (graduates) in Ontario

Figure 5.11: Enrolment totals by year of enrolment for different categories of age (graduates) in the Prairies

Figure 5.12: Enrolment totals by year of enrolment for different categories of age (graduates) in British Columbia

Figure 5.13: Enrolment totals by year of enrolment for regions of enrolment (undergraduates)

Figure 5.14: Enrolment totals by year of enrolment for regions of enrolment (graduates)

Figure 5.15: Enrolment totals by year of enrolment for different categories of age (men) in the Atlantic provinces

Figure 5.16: Enrolment totals by year of enrolment for different categories of age (men) in Quebec

Figure 5.17: Enrolment totals by year of enrolment for different categories of age (men) in Ontario

Figure 5.18: Enrolment totals by year of enrolment for different categories of age (men) in the Prairies

Figure 5.19: Enrolment totals by year of enrolment for different categories of age (men) in British Columbia

Figure 5.20: Enrolment totals by year of enrolment for different categories of age (women) in the Atlantic provinces

Figure 5.21: Enrolment totals by year of enrolment for different categories of age (women) in Quebec

Figure 5.22: Enrolment totals by year of enrolment for different categories of age (women) in Ontario

Figure 5.23: Enrolment totals by year of enrolment for different categories of age (women) in the Prairies

Figure 5.24: Enrolment totals by year of enrolment for different categories of age (women) in British Columbia

Figure 5.25: Male enrolment by year for the regions of residence

Figure 5.26: Female enrolment by year for the regions of residence

Figure 5.27: Residuals by predicted values (ANOVA model for undergraduate enrolment)

Figure 5.28: Residuals by predicted values (ANOVA model for graduate enrolment)

Figure 5.29: Residuals by predicted values (ANOVA model for undergraduate enrolment, with transformed response)

Figure 5.30: Residuals by predicted values (ANOVA model for graduate enrolment, with transformed response)

Figure 5.31: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 15 to 24 years of age)

Figure 5.32: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 25 to 29 years of age)

Figure 5.33: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 30 to 34 years of age)

Figure 5.34: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Atlantic provinces, 35 years of age and over)

Figure 5.35: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 15 to 24 years of age)

Figure 5.36: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 25 to 29 years of age)

Figure 5.37: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 30 to 34 years of age)

Figure 5.38: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Quebec, 35 years of age and over)

Figure 5.39: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 15 to 24 years of age)

Figure 5.40: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 25 to 29 years of age)

Figure 5.41: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 30 to 34 years of age)

Figure 5.42: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (Ontario, 35 years of age and over)

Figure 5.43: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 15 to 24 years of age)

Figure 5.44: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 25 to 29 years of age)

Figure 5.45: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 30 to 34 years of age)

Figure 5.46: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (the Prairies, 35 years of age and over)

Figure 5.47: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 15 to 24 years of age)

Figure 5.48: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 25 to 29 years of age)

Figure 5.49: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 30 to 34 years of age)

Figure 5.50: Proportion of Canadian enrolled at university from 1994 to 1998 by gender (British Columbia, 35 years of age and over)

Figure 5.51: Deviance residual by its index for $X = x_i$; $Y = 1$ (1988)

Figure 5.52: Deviance residual by its index for $X = x_i$; $Y = 2$ (1988)

Figure 5.53: Deviance residual by its index for $X = x_i$: $Y = 3$ (1988)

Figure 5.54: Deviance residual by its index for $X = x_i$: $Y = 1$ (1992)

Figure 5.55: Deviance residual by its index for $X = x_i$: $Y = 2$ (1992)

Figure 5.56: Deviance residual by its index for $X = x_i$: $Y = 3$ (1992)

Figure 5.57: Deviance residual by its index for $X = x_i$; $Y = 1$ (1997)

Figure 5.58: Deviance residual by its index for $X = x_i$; $Y = 2$ (1997)

Figure 5.59: Deviance residual by its index for $X = x_i$; $Y = 3$ (1997)

Figure 5.60: Deviance residual by its index for $X = x_i$; $Y = 1$ (1997, administrative counts)

Figure 5.61: Deviance residual by its index for $X = x_i$; $Y = 2$
(1997, administrative counts)

Figure 5.62: Deviance residual by its index for $X = x_i$; $Y = 3$
(1997, administrative counts)

Figure 5.63: Deviance residual by its index for $X = x_i$; $Y = 1$ (1988)

Figure 5.64: Deviance residual by its index for $X = x_i$; $Y = 2$ (1988)

Figure 5.65: Deviance residual by its index for $X = x_i$; $Y = 3$ (1988)

Figure 5.66: Deviance residual by its index for $X = x_i$; $Y = 1$ (1992)

Figure 5.67: Deviance residual by its index for $X = x_i$; $Y = 2$ (1992)

Figure 5.68: Deviance residual by its index for $X = x_i$; $Y = 3$ (1992)

Figure 5.69: Deviance residual by its index for $X = x_i$; $Y = 1$ (1997)

Figure 5.70: Deviance residual by its index for $X = x_i$; $Y = 2$ (1997)

Figure 5.71: Deviance residual by its index for $X = x_i$; $Y = 3$ (1997)

Figure 5.72: Deviance residual by its index for $X = x_i$; $Y = 1$ (1997, administrative counts)

Figure 5.73: Deviance residual by its index for $X = x_i$; $Y = 2$
(1997, administrative counts)

Figure 5.74: Deviance residual by its index for $X = x_i$; $Y = 3$
(1997, administrative counts)

Figure 5.75: Canadian male enrolment by year for several age categories

Figure 5.76: Canadian female enrolment by year for several age categories