Modelling Resource Intensity Weights (RIWs) at the Canadian Institute for Health Information (CIHI)

by

Fares Said

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

Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada
January 13, 2013

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-94312-0

Our file Notre référence
ISBN: 978-0-494-94312-0

Abstract

This thesis aims to assess alternative models to the current cost model used to predict health care costs in Canada. By using statistical models, it will be shown how we can stabilise the cost model to an acceptable and efficient level. This thesis outlines the work conducted at CIHI and its relevance in Canada. Chapters 2 through 5 explain, analyse and compare the Current model, the Linear Mixed Model (LMM), the Generalized Linear Model (GLM), the Heteroscedastic Regression Model (HER), and the Heteroscedastic, Random and Fixed Effects Model (HEREM). The resource indicator: Resource Intensity Weight (RIW) is the cost measure analysed with the main focus being on the typical RIW. Chapter 6 presents the results of the simulations conducted, where we conclude that modifying the current model to the HEREM model did not yield the results we had initially hoped for.

Acknowledgments

Many thanks to Jeff Hatcher - Senior Consultant Case Mix and Activity-based Funding, at the Canadian Institute of Health Information (CIHI) to have made this learning opportunity available. My sincere gratitude goes to Winfried Jakob, SAS administrator, for all his help in solving coding problems and resolving bugs, as well as the IT department for providing all the necessary software and tools to conduct my work. Finally, I would also like to thank Professor Jason Nielsen, from Carleton University, for taking the time out of his busy schedule to meet with me over the past year and to provide his input and insight into the work I was conducting. His assistance and guidance were integral in helping me develop the ideas you will see here.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronyms	Definition
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Prediction
CCI	Canadian Classification of Health Interventions
CIHI	Canadian Institute for Health Information
CL	Comorbidity level
CMG	Case Mix Group
CMG+	Case Mix Group plus extra factors
CV	Coefficient of Variation
DAD	Discharge Abstract Database
df	Degree of Freedom
EBLUE	Empirical Best Linear Unbiased Estimator
EBLUP	Empirical Best Linear Unbiased Prediction
EF	Exponential Family

FI	Flagged Intervention
GLM	Generalized Linear Model
GLS	Generalized Least Squares
HER	Heteroscedastic Regression
HEREM	Heteroscedastic, Random and Fixed Effects
HID	Health Care Facility Identification
ICD-10-CA	International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Canada
IE	Intervention Event
LMM	Linear Mixed Model
LOS	Length of Stay
LS	Long-Stay
MCC	Major Clinical Categories
mgf	Moment Generating Function
MIS	Management Information Standards
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MSE	Mean Square Error
MVQUE	Minimum Variance Quadratic Unbiased Estimation

MRDx	Most Responsible Diagnosis
NA	Not Applicable
NB	Newborn
NR	Newton-Raphson
OLS	Ordinary Least Square
OOH	Out-of-Hospital Intervention
pdf	Probability Density Function
REML	Restricted Maximum Likelihood
RIW	Resource Intensity Weight
SD	Standard Deviation
SE	Standard Error
SS	Sum of Squares
SSE	Sum Square Error
VC	Variance Component
WHO	World Health Organizations
WLS	Weighted Least Square

Chapter 1

CIHI

1.1 Canadian Institute for Health Information

The Canadian Institute for Health Information (CIHI) is a not-for-profit organisation that collects and analyzes data and then publishes its findings as public information so that private and public bodies can use the information for the benefit of Canadians. CIHIs mandate and vision, simply put, is to improve the health of Canadians through the development and maintenance of reliable, quality data that can be used by health leaders to make better-informed decisions. CIHI's information is also used to create sound and actionable health policies with concrete and tangible results.

All areas within CIHI work collaboratively internally as well as with external partners in order to enhance data holdings and continue to foster an understanding of the data and the results of various studies conducted at CIHI. The work conducted at CIHI is done in an impartial and focused manner. As a not-for-profit organisation, CIHI has loyalty to its work and mandate. It provides information to public and private sector bodies but does not report to either. This is one of the reasons CIHI's information holdings are considered reliable and impartial. Moreover, CIHI aims to provide unbiased, reliable and comparable data from various Canadian health

systems. For this reason, data and results are treated in the same way throughout all the provinces and territories; it maintains consistency and allows for various uses of data and results. [3][1]

1.2 Case Mix Classification Systems

The case mix system used at CIHI is a system that groups together medical cases similar in nature for the purpose of categorizing patients into statistically and clinically homogeneous groups based on the collection of clinical and administrative data. [4] Case mix has two parts: the grouping methodologies and the resource indicators. As stated earlier, patients are grouped based on clinical and administrative factors. Resource indicators are what identify the costs of individuals in each group. Resource Intensity Weights (RIWs) are "a relative cost weight value derived from case-cost data submitted to CIHI's Management Information Standards (MIS) and Costing department. All RIW cost weights are relative to the average typical inpatient case, such that the sum of typical cases is equal to the sum of the typical weighted cases." [1] Health leaders use this information to benchmark their funding allocations for each health care facility among its other counterparts. Moreover, health care facilities also use this information to manage their allocated funding. This information is used to make payments to health care centres for each case they treat.

1.3 Discharge Abstract Database

The Discharge Abstract Database (DAD) is a database containing administrative, demographic and clinical information on outgoing acute patients (this includes deaths,

[1]The information found in this section was obtained from various sections (including "Vision and Mandate" and "Corporate Strategies") on the organisation's official web site at `www.cihi.ca`

transfers and sign-out patients). This information is collected from each patient abstract across Canada. All provinces take part in sharing this information with CIHI for research purposes.[2] Once a patient is admitted and discharged, the patient abstract is completed with all the necessary information using a classification system called the International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Canada (ICD-10-CA), and is forwarded to CIHI for entry into the DAD. Some of the information is classified using various codes (type 1, type 2, type 3, Most Responsible Diagnosis (MRDx), among others). More plainly, type 1 is assigned when the condition is significant and present during admission. Patients classified as type 1 require treatment. Type 2 differs slightly in that it is a condition identified after a patient is admitted, and is a condition for which the patient is provided with treatment. The type 2 condition must prolong the length of stay by a minimum of 24 hours. A type 3 is assigned, optionally, when a condition is identified and is either treated in a manner that does not affect the length of stay or is not treated at all, thereby also not affecting the length of stay. Since a type 3 assignment is optional, it may or may not appear on a patient abstract. Every abstract will contain between 1 and 25 diagnoses. An abstract contains between 0 and 20 interventions. [5]

1.3.1 MRDx, Type 6 and Asterisk Code

A patient is assigned an MRDx code (diagnosis type M) upon admission to a hospital based on a determination that this condition is most responsible for the cost of care and length of stay. World Health Organization (WHO) coding rules specify that where conditions consisting of etiology (e.g. Diabetes) and a manifestation (e.g. foot ulcer), the etiology has to be coded as the MRDx. When the second line of the

[2]Quebec submits their information separately from other provinces. This information is appended to the DAD and forms part of another database.

patient's abstract indicates that this manifestation is the most responsible for the patient's stay in the hospital, this condition is assigned a Diagnosis type 6. However, when the etiology is the most responsible for the length of stay in the hospital or when it cannot be determined which of the two, the manifestation is assigned a Diagnosis type 3, meaning secondary diagnosis [1].

1.3.2 ICD-10-CA

ICD-10-CA is a Canadian enhanced classification based on the WHO publication of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision. It stems from an international standard for reporting clinical diagnoses. By using a modified version of an international classification system, CIHI maintains the consistency of the data collected with that of international data. This method allows CIHI to not only collect, store and use data locally, provincially and nationally but to also compare the data on an international level should the opportunity arise.

1.3.3 Canadian Classification of Interventions

The Canadian Classification of Health Interventions (CCI) consists of several groups of intervention codes, totalling approximately 17,845 codes in 2012, used to categorize procedures and interventions that patients undergo. A patient can be assigned more than one intervention code and up to 25 intervention codes.

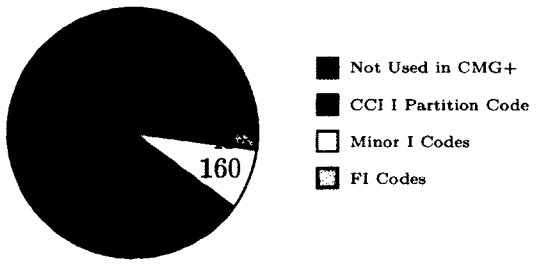


Figure 1.1: CCI Code List - scaled by 10

1.4 Case Mix Groups+ Methodology

The CMG+ methodology is used to group acute patients into similar groups based on clinical conditions or procedures in order to better determine their resource consumption and the length of their stay. Currently, the CMG+ is based on the ICD-10-CA diagnosis and CCI intervention activity and cost data. The CMG+ grouping methodology is based on the codes (i.e. Conditions and Interventions) that exist on a patient abstract. By using these codes, patients are grouped into similar clinical categories, called Major Clinical Categories (MCCs) (see Appendix A), which are mostly divided based on body system. If the MRDx is invalid, the patient is assigned to MCC99, Miscellaneous CMGs and Ungroupable Data. MCCs are further divided into Diagnosis partitions and Intervention partitions based on presence or absence of specific diagnosis and intervention partition codes. The minor interventions are listed in the 2012 Directory, under Other Interventions Used in CMG Assignment. If the patient is less than 29 days old, they are automatically assigned to MCC14, Newborns and Neonates With Conditions Originating in the Perinatal Period. However, in general, the MRDx, the asterisk code (see section 1.3.1), and the intervention partition code determine to which MCC a patient is assigned. [1]

1.4.1 Major Clinical Categories

Major Clinical Categories are, in general, identified and named based on either a major body part or a clinical problem of a body system (see Appendix A). If a patient is already assigned to an MCC based on either an MRDx or an asterisk code/type 6 code, then they are run through the CCI intervention partition code to determine which partition they must be assigned to: diagnosis partition (if a CCI intervention partition code is not found) or the intervention partition (if a CCI intervention partition code is found) (see Figure 1.2, on page 12). There are exceptions

to this method, more specifically in MCC14 and in cases where there are gender splits being conducted. [6] Each of the 562 CMGs is, in general, assigned to one MCC only, meaning that CMGs assigned to one MCC are mutually exclusive from other CMGs assigned to another MCC; this however, is with the exception of a few CMGs called Crossover CMGs (see section 1.4.5).

1.4.2 Intervention Partition

The CCI intervention partition code list contains approximately 11,369 CCI codes, which determine whether or not the patient is placed in the Intervention partition. Patients assigned to the intervention partition may have more than one intervention partition code. If so, the patient is assigned to a CMG based on an Intervention hierarchy (codes representing interventions organised from most to least expensive). A code is selected using the grouping methodology application which loops through all the interventions on a patient abstract in order to select the most expensive CMG. If the intervention is not assigned to a CMG according to the MCC-specific hierarchy, it is put into the MCC-specific unrelated intervention CMG (901 to 918). [6]

1.4.3 Diagnosis Partition

If a patient does not have an intervention code, they are assigned to a diagnosis partition instead of the intervention partition. A patient within the diagnosis partition is assigned to a CMG based on the MRDx. Regardless of which partition a patient is assigned, in some cases, diagnoses, gestational age, entry code and interventions not on the intervention partition code list may also be used in CMG assignment. [6]

1.4.4 CMG+ Factors

A patient is assigned to a CMG based on the MRDx and intervention partition codes. However, the assignment of resource indicators, such as the RIW, to a specific case is based on the base CMG value combined with the five factors:

1. Age
2. Comorbidity level (CL)
3. Flagged Intervention (FI)
4. Intervention Event (IE)
5. Out-of-Hospital Intervention (OOH)

Age Category

The age category is independent from the other factors. It is combined with the CMG to calculate a base RIW value, that is, a base RIW value is assigned for each CMG/age category combination. These base values represent the RIW of the CMG/age category when there is no other factor present. All other factors can then be used to adjust these base values. See table 1.1 which outlines the age categories as grouped by the CMG+ methodology.

Table 1.1: Age Categories [1]

Category	Description	Ages
	Newborn and Neonate MCC	
A	Newborn	0 days
B	Neonate	1 to 7 days
C	Neonate	8 to 28 days
	Pediatric age groups across most MCCs	
F	Pediatric	29 to 364 days
G	Pediatric	1 to 7 years
H	Pediatric	8 to 17 years
	Adult age groups across most MCCs	
R	Adult	18 to 59 years
S	Adult	60 to 79 years
T	Adult	80+ years

Comorbidity Level

A patient is assigned to type MRDx, and other significant condition(s) identified are coded type 1 or type 2. If a patient is assigned a diagnosis code from the Diagnosis code list, and the code belongs to Diagnosis type 1 or type 2, then this patient is assigned a CL based on the comorbidity factor. Should a comorbidity be identified pre-admission, a type 1 code is applied. Should a comorbidity be identified post-admission, a type 2 code is applied. It is important to note, however, that the CL is calculated using type 1 and type 2; type 3 is not used for this purpose. The cumulative comorbidity factor is what we use to determine which CL a patient is assigned. The CL factor explains the impact comorbid conditions in-patients have on the consumption of resources. Table 1.2 shows the percentage increase in resource use for comorbid patients. The levels are ranked from 0 to 4. Each MCC has its own list of comorbid possibilities, which have been assigned their potential effect within that MCC. [1]

Table 1.2: Definition of Comorbidity Levels [1]

Comorbidity Levels	Impact for Comorbidity Conditions
0	No Significant Comorbidity
1	Increase the Case Resources by 25% to 49%
2	Increase the Case Resources by 50% to 74%
3	Increase the Case Resources by 75% to 24%
4	Increase the Case Resources by at Least 125%
8	Not Applicable: Comorbidity Not Applied (For Example, Normal Newborns)

Flagged Intervention

Flagged Interventions (FIs) are high-cost interventions and are identified by CCI codes. FIs are not used for CMG assignments and are not included on the CCI intervention partition code list. They are applied to a patient after the initial CMG assignment and are factored into the RIW methodology. There are 16 categories of possible FIs (Categories A to P), which represent approximately 433 CCI codes. Within each of these categories, there can be multiple intervention codes. If a patient has one or more intervention code that corresponds to the category, then the FI-Category value = 1. Otherwise, the FI-Category value = 0. This means that there are 16 dummy variables (A to P). [1]

The 16 categories of flagged interventions are the following:

(A) Non-invasive biopsy;

(B) Cardioversion;

(C) Cell saver;

(D) Chemotherapy;

(E) Dialysis;

(F) Per-orifice endoscopy;

(G) Feeding tubes (PEG);

(H) Heart resuscitation;

(I) Mechanical ventilation greater than 96 hours;

(J) Mechanical ventilation less than 96 hours;

(K) Paracentesis;

(L) Parenteral nutrition;

(M) Pleurocentesis;

(N) Radiotherapy;

(O) Tracheostomy; and

(P) Vascular access device.

Intervention Event

Throughout the admission, a patient may need other interventions or may return to the operation room for the same intervention at a different time than the initial intervention used for CMG assignment. Presence of a CCI code from the intervention partition code list is used initially to assign the patient to an intervention CMG (at this point, IE = 1); if the same patient returns to a surgical suite one more time after the initial intervention or requires another intervention (then IE = 2). If more than two interventions, then the IE is set to 3+. If a patient is assigned to the CMG according to the diagnosis code, then the IE is set to zero (see Table 1.3). Multiple IE for one patient means more complicated treatments and as a result, higher incurred costs. IE is independent from all the other factor effects. [1]

Table 1.3: Intervention Event Codes and Descriptions [1]

Intervention Event Code	Code Description
1	1 Intervention Event
2	2 Intervention Event
3	3 Intervention Events or more
8	Not Applicable: Diagnosis Partition (0 Intervention Events)

Out-of-Hospital Intervention

The CMGs: 161, 174, 175, 176, 193, 195, 201, 203 and 207, contain interventions that may be performed out of hospital (OOH). It is important to distinguish an OOH patient from other patients since the intervention of this patient does not happen in

the facility he/she is admitted to. OOH interventions are split into three categories: A (Cardiac Catheter), B (Pacemaker) and C (Percutaneous Coronary Intervention). The OOH factor is used only in the calculation of RIW. Performing an intervention on an OOH basis has an expected reduction in costs to the facility where the patient resides as an inpatient, and increased length of stay (LOS). [1]

1.4.5 Crossover CMG

We have seen that the CMG is assigned based on the MCC in which a patient is grouped. Generally, the CMG has a home MCC, meaning that the CMG is specific to one MCC, based on the major body system in which the intervention is performed. However, in some instances, the CMG may be assigned based on another MCC. This type of CMG is called a crossover CMG because it may be assigned through more than one MCC. See Table 1.4 for a list of all crossover CMGs. [1]

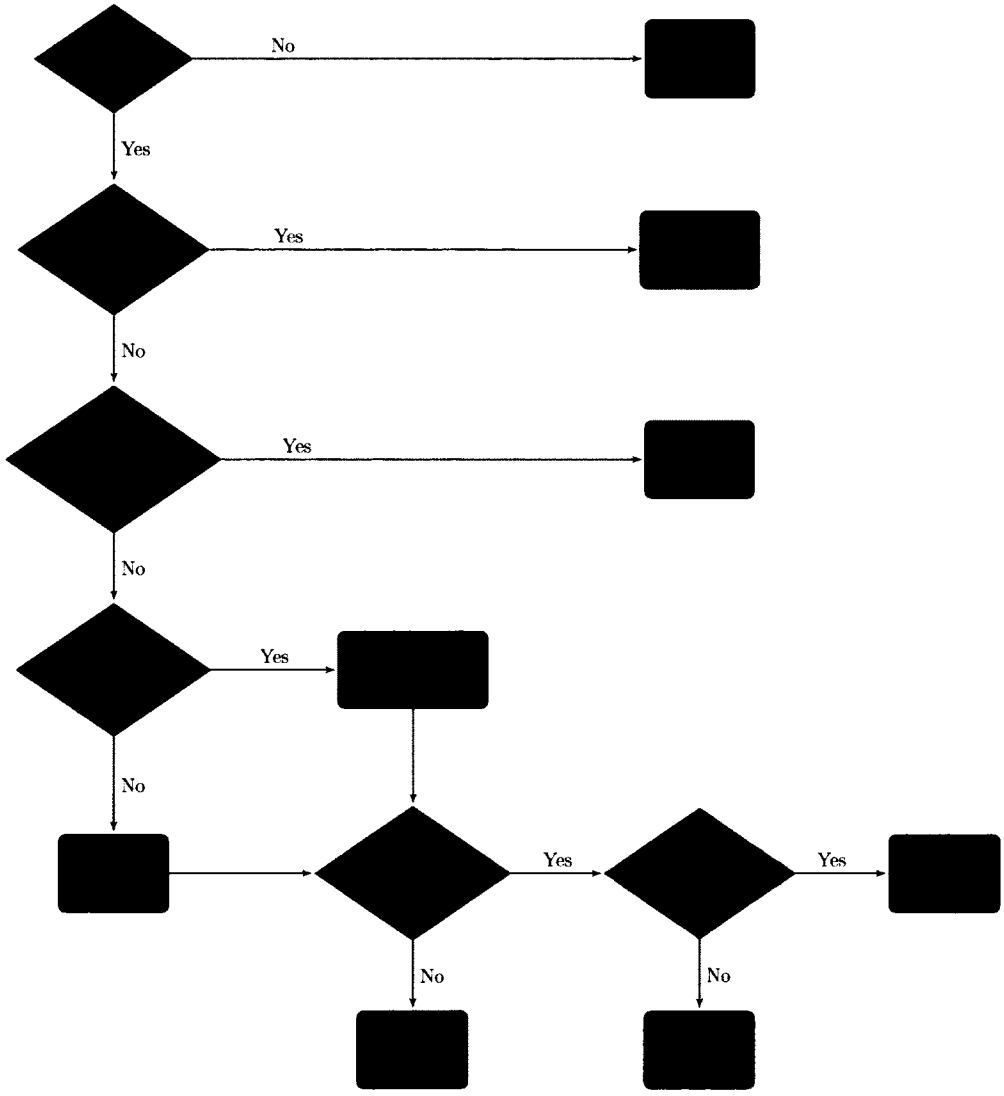
Table 1.4: Crossover CMGs

Crossover Intervention CMG			
CMG	**Description**	**Home MCC**	**Other MCC**
110	Lung Transplant	4	10
270	Liver/Pancreas/Duodenum Transplant	7	10
271	Excision Pancreas with Duodenum	7	10
272	Drainage/Biopsy of Pancreas	7	10
273	Bypass/Excision of Pancreas	7	10
Crossover Diagnosis CMG			
CMG	**Description**	**Home MCC**	**Other MCC**
999	Ungroupable	99	14

1.5 Summary

Now that we have reviewed all the relevant aspects of the CMG+ methodology, we can summarise the relationship between them all (see Figure 1.2). The figure below outlines a patient's path when being assigned a CMG.

Figure 1.2: Patient Grouping and Coding. [2]



Chapter 2

The Current Model and HER Model

2.1 Introduction

This chapter explains the methodology for fitting the CIHI model and the Heteroscedastic Regression (HER) model. We will also discuss the differences in how these models are used to calculate the weight and how they handle the heteroscedasticity differently. The current model uses the ordinary least square (OLS) and weighted least square (WLS) to estimate the parameters while the HER model uses the optimization algorithms (Maximum Likelihood Estimators (MLE)). We expect the HER model to have less significant covariates because the SE for the HER model would be smaller than that of the current model.

2.2 Current Production Model at CIHI

The resource intensity weight is predicted using previous year costs. The RIW is then used to predict the cost of interventions for the upcoming year based on the CMG+ factors.

The current model uses the log cost because the cost is log-normally distributed and because a multiplicative relationship (rather than an additive one) exists between

the cost and the independent effects. The log cost is derived in order to meet the assumption of the least square and a normally distributed residual. Moreover, the current model estimates are on a log scale that needs to be retransformed to the dollar scale. This retransformation generates a bias that is then fixed using:

$$\hat{y}_i = exp\left(\hat{z}_i + \lambda \times \frac{\hat{\sigma}_i^2}{2}\right) \tag{2.1}$$

where,

- $\hat{z}_i$ is the predicted log of RIW
- $\hat{\sigma}_i^2$ is the predicted variance (see section 2.2.2)
- λ is an adjustment factor to ensure that there is no bias

The bias correction factor used will still yield biases on the dollar scale for many effects. The mean model uses WLS to account for non-constant variances across the independent effects and to resolve the heteroscedasticity. The variance model provides weights for use in WLS estimation of the mean model. The weights used for the WLS model does not take into account that the weights are estimated; they assume they are fixed. Due to this degree of freedom being used to estimate the variance parameters are ignored. This results in the underestimation of variance estimates of the mean parameters.

The Model

CIHI uses two types of models in three steps for predicting costs. The first model is a *mean model* and the second model is a *variance model.* The first step uses the OLS to regress the dependent log cost variable against the covariates. The residual value is obtained without the removal of insignificant effects. The second step takes the residual square of the first model as the dependent variable and regresses them against the covariates. In this step, the insignificant effects will be removed. The

third step uses the WLS model where the log cost is the dependent variable that will be regressed against the covariates using the prediction from the second step (variance model). The weight for this model is the reciprocals of these predictions. In this step, the insignificant effects will be removed based on 5% significance level.

2.2.1 Mean Model

Let $\mathbf{Y}$ be an $(n \times 1)$ vector, where the i^{th} element is the i^{th} patient's log cost. The model is represented by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

where,

- $\mathbf{X}$ is an $(n \times p)$ design matrix that contains covariate information for the i^{th} patient in the i^{th} row, $i = (1, \ldots, n)$. These covariates are dummy variables created from the CMGs and the five-factor effects as seen in section 5.3
- $\boldsymbol{\beta}$ is a $(p \times 1)$ parameter vector
- $\boldsymbol{\epsilon}$ is an $(n \times 1)$ error vector that is assumed to be normally distributed with mean zero and common variance σ^2 under the OLS model.

2.2.2 Variance Model

Let $\mathbf{Z}$ be the residual square obtained from the OLS model defined $\mathbf{Z} = (Z_1, \ldots, Z_n)'$, where $Z_i = \epsilon_i^2$ and ϵ_i^2 is the i^{th} element in the residual vector $\hat{\boldsymbol{\epsilon}}$. Note $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the OLS estimate for $\boldsymbol{\beta}$ obtained by the mean model. We consider the following model:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{e} \tag{2.3}$$

where,

- $\mathbf{X}$ represents the same design matrix of the OLS model above
- γ is a $(p \times 1)$ parameter vector
- $\mathbf{e}$ is an $(n \times 1)$ error vector that is assumed to be normally distributed with mean $\mathbf{0}$ and common variance σ^2

From this model, the weights are calculated using the reciprocal predicted value of $\mathbf{Z}$. This is denoted as:

$$w_i = \frac{1}{\hat{Z}_i} \tag{2.4}$$

If the weight is less than or equal to 0.5, we set it to 0.5; if the weight is greater than or equal to 200, we set it to 200. If the predicted values $\hat{\mathbf{Z}}$ of $\mathbf{Z}$ are less than 0, they are set to 0 as variance can not be negative.

2.2.3 Weighted Least Square

The purpose of the first two models was to calculate the weights to fit a weighted least square regression represented by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\nu} \tag{2.5}$$

where,

- $\mathbf{X}$ represents the same design matrix of the OLS model above
- $\boldsymbol{\alpha}$ is a $(p \times 1)$ parameter vector
- $\boldsymbol{\nu}$ is an $(n \times 1)$ error vector that is assumed to be normally distributed with mean $\mathbf{0}$ and heteroscedastic variance σ_i^2

Here, $\text{Var}(Y_i) = \sigma_i^2$. We identify observations with higher variance and less weight in the regression fitting.

The WLS estimate for $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = (\mathbf{XWX}')^{-1}\mathbf{X}'\mathbf{WY}$ where, $\mathbf{W}$ is an $(n \times n)$ matrix with diagonal equal to w_i obtained by equation (2.4) and zero off-diagonal.

2.3 The Heteroscedastic Regression (HER) model

Let $\mathbf{Y}$ be an $(n \times 1)$ vector, where the i^{th} element is the i^{th} patient's log cost. The model is represented by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.6}$$

where,

- $\mathbf{X}$ is an $(n \times p)$ design matrix that contains covariate information for the i^{th} patient in the i^{th} row, $i = (1, \ldots, n)$. These covariates are dummy variables created from the CMGs and the five-factor effects as seen in section 5.3
- $\boldsymbol{\beta}$ is a $(p \times 1)$ parameter vector
- $\boldsymbol{\epsilon}$ is an $(n \times 1)$ error vector that is assumed to be independent Gaussian random variables with mean $\mathbf{0}$ and variance covariance matrix $\mathbf{V}$ is an $(n \times n)$ matrix with diagonal equal to v_i obtained by equation (2.7) and zero off-diagonal

As the residual variance is not homogeneous we will model it as

$$v_i = \exp\{\mathbf{z}_i'\boldsymbol{\gamma}\} \tag{2.7}$$

where,

- $\mathbf{z}_i'$ is the i^{th} row of an $(n \times q)$ matrix $\mathbf{Z}$ that contains covariate information for the i^{th} patient, $i = (1, \ldots, n)$
- $\boldsymbol{\gamma}$ is a $(q \times 1)$ vector of the associated regression parameters

- $\mathbf{V} = \begin{pmatrix} \exp(\mathbf{z}_1'\boldsymbol{\gamma}) & 0 & \dots & 0 \\ 0 & \exp(\mathbf{z}_2'\boldsymbol{\gamma}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \exp(\mathbf{z}_n'\boldsymbol{\gamma}) \end{pmatrix}$

A log-linear relationship between the diagonal of $\mathbf{V}$ and the covariates of $\mathbf{Z}$ is used to ensure that estimates of the variance are never negative. Note that the covariates in $\mathbf{X}$ are used for the mean estimates and the covariates for $\mathbf{Z}$ are used for the variance estimates; where each row of these matrices contain a subset of covariate information for each individual i and matrix $\mathbf{Z}$ could be a subset of $\mathbf{X}$. The estimation of the mean ($\boldsymbol{\beta}$) and variance ($\boldsymbol{\gamma}$) parameters are carried out by restricted maximum likelihood estimation.

Remark 1 *If a nonsingular matrix A depends on vector x then for any element in x, we determine: $\frac{\partial A^{-1}}{\partial x_i} = -A^{-1}\left(\frac{\partial A}{\partial x_i}\right)A^{-1}$. If A is also positive definite then for any element in x, we determine $\frac{\partial}{\partial x_i} ln(|A|) = tr\left(A^{-1}\frac{\partial A}{\partial x_i}\right)$.*

2.3.1 Estimation Methods

Let $\mathbf{x}_i' = (x_{i1}, \dots, x_{ip})$ be a vector of covariates for the i^{th} patient, then the likelihood function for model (2.6) is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v_i}} \exp\left\{-\frac{(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2v_i}\right\} \propto \frac{1}{\prod_{i=1}^{n}\sqrt{v_i}} \exp\left\{-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2}\right\} \quad (2.8)$$

and the log-likelihood by

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\Big\{\sum_{i=1}^{n}\log(v_i) - (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\Big\} \\ &= -\frac{1}{2}\Big\{\sum_{i=1}^{n}\mathbf{z}_i'\boldsymbol{\gamma} - (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\Big\} \\ &= -\frac{1}{2}\Big\{\mathbf{1}'\mathbf{Z}\boldsymbol{\gamma} - (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\Big\} \end{aligned} \quad (2.9)$$

where $\boldsymbol{\theta} = \binom{\boldsymbol{\beta}}{\boldsymbol{\gamma}}$ is an $(m \times 1)$ vector for mean and variance parameters with $m = p+q$, $\mathbf{1}$ is a vector of n ones. The first term of the equation (2.9) is the log of the variance residual. See equation (2.7). Let $\hat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$. Since the MLE is the global maximum of the log-likelihood function $l(\boldsymbol{\theta})$, the solution for the system (2.10) may or may not be the MLE. Usually, the best method to find the MLE is the "fine grid" method. This method uses grids (subspaces) of the parameter space to determine the value of the log-likelihood for a specific point within each grid. From that value, we can then obtain an approximate global maximum of the log-likelihood at each of those specific points. This value becomes increasingly accurate as the grids become smaller. However, when the parameter is multi-dimensional, we may use another method to obtain the MLE. For this method, we can look at the stationary points, which are the solutions for the Maximum Likelihood (ML) equations, and then we verify the solution of these stationary points to identify either the (a) global maximum, (b) local maximum or (c) local minimum or saddle points. Since we wish to find the global maximum, we determine the Hessian matrix (2^{nd} derivatives for the ML equations) of the log-likelihood, which is positive definite at the local maximum. In order to ensure that the solution is the global maximum, we conduct a complete and successful implementation by finding all the solutions for the ML equations, by then comparing the values of the log-likelihood of the stationary points with the values on the boundary of the parameter space. This process allows us to identify the global maximum. The same process may also be applied for Restricted Maximum Likelihood (REML) estimation [7]. To obtain the maximum likelihood estimator (MLE) required for inference, we need to find the maximum of (2.8) in $\boldsymbol{\theta}$, which is equivalent to the maximum of (2.9), since log is a strictly monotone transform function of (2.8). The

maximum can be obtained by solving the following system of normal equations:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \boldsymbol{\gamma}} l(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{W}_j\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \sum_{i=1}^{n} z_{ij} \right\}_{q \times 1} \end{pmatrix} = \mathbf{0} \quad (2.10)$$

where $\mathbf{W}_j = \frac{\partial \mathbf{V}}{\partial \gamma_j} = \text{diag}\left\{ z_{ij} e^{\mathbf{z}_i'\boldsymbol{\gamma}} \right\}$ and $j = 1, \ldots, q$. The 2^{nd} equation in (2.10) does not have closed form solution; it can be solved using many different methods. Here we will use Fisher Scoring (FS). See Appendix (B) on page (83) for more information. Note that there is a closed form solution to $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$ in $\boldsymbol{\beta}$. For simplicity, we assume the matrix $\mathbf{X}$ is of full rank, where rank $(\mathbf{X}) = p$. The optimal value of equation (2.11) is then obtained by $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$, where $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$ is the **Generalized Least Squares** (GLS) estimator which is denoted by:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{V}^{-1}\mathbf{Y} \quad (2.11)$$

Since we have an estimator for $\boldsymbol{\beta}$ as a function of $\boldsymbol{\gamma}$ we can derive the MLE $\hat{\boldsymbol{\theta}}$ by maximizing the *profile log-likelihood*

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\beta}(\boldsymbol{\gamma}), \boldsymbol{\beta}) = -\frac{1}{2}\left\{\mathbf{1}'\mathbf{Z}\boldsymbol{\gamma} - \mathbf{Y}'\mathbf{M}\mathbf{Y}\right\} \quad (2.12)$$

with respect to $\boldsymbol{\gamma}$ where

$$\mathbf{M} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\Big(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad (2.13)$$

and

$$\mathbf{Y}'\mathbf{M}\frac{\partial \mathbf{V}}{\partial \gamma_i}\mathbf{M}\mathbf{Y} = tr\Big(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_i}\Big) \quad (2.14)$$

which is generally numerically more efficient and stable. The maximum of $l_p(\gamma)$ is the solution to

$$\frac{\partial}{\partial \gamma} l_p(\gamma) = \left\{ \mathbf{Y}'\mathbf{M}\mathbf{W}_j\mathbf{M}\mathbf{Y} - \sum_{i=1}^{n} z_{ij} \right\}_{q\times 1} = \mathbf{0} \tag{2.15}$$

which is the MLE $\hat{\gamma}$ for γ. By substituting $\hat{\gamma}$ into (2.11), i.e. $\hat{\beta} = \hat{\beta}(\hat{\gamma})$, we can obtain the MLE for β.

On the other hand, we could use the REML, which is favoured over ML estimation since the REML approach corrects for degrees of freedom used in estimating β. In REML estimation, we maximize the log-likelihood function for the residual vector $\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$, where $\hat{\beta}$ is the GLS in equation (2.11). $\hat{\beta}$ and $\hat{\epsilon}$ are independent (see equation (2.16)) and have a normal distribution because they are linear functions of a normal distribution vector $\mathbf{Y}$.

$$\begin{aligned} Cov\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \hat{\epsilon}\right) &= \mathbf{X}'\mathbf{V}^{-1}\mathbf{V}[\mathbf{I} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'] \\ &= \mathbf{X}' - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}' = 0. \end{aligned} \tag{2.16}$$

The log-likelihood function for the residual vector $\hat{\epsilon}$ is:

$$\begin{aligned} l\left(\hat{\epsilon}, \theta\right) &= l\left(\mathbf{Y}, \theta\right) - l(\hat{\beta}, \theta) \\ &= -\frac{1}{2}\Big\{ \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \ln|\mathbf{V}| \\ &\quad + \Big(\mathbf{Y} - \mathbf{X}\beta\Big)'\mathbf{V}^{-1}\Big(\mathbf{Y} - \mathbf{X}\beta\Big) - \Big(\hat{\beta} - \beta\Big)'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big(\hat{\beta} - \beta\Big)\Big\} \end{aligned} \tag{2.17}$$

The log-likelihood function for the residual vector $\hat{\epsilon}$ can be rewritten as:

$$l_{REML}(\gamma) = l(\beta(\gamma), \gamma) = -\frac{1}{2}\Big\{ \mathbf{1}'\mathbf{Z}\gamma + \log\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| + \mathbf{Y}'\mathbf{M}\mathbf{Y} \Big\} \tag{2.18}$$

The (2.18) function is called a residual or restricted log-likelihood function, different from a profile log-likelihood function (2.12) by the extra term $-\frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$. As we

can see, the restricted log-likelihood function (2.18) is a function of $\boldsymbol{\gamma}$ only, we could solve

$$\frac{\partial}{\partial \boldsymbol{\gamma}} l_{REML}(\boldsymbol{\gamma}) = \{\mathbf{Y}'\mathbf{M}\mathbf{W}_j\mathbf{M}\mathbf{Y} - \mathrm{tr}\,(\mathbf{M}\mathbf{W}_j)\}_{q\times 1} = \mathbf{0} \tag{2.19}$$

and obtain the REML estimator $\hat{\boldsymbol{\gamma}}$, a bias corrected version of the MLE. REML is a method to correct for the fact that our variance estimator is not using the true mean but an estimate of it.

2.3.2 Inference

We can also find the Hessian matrix for the model by taking the 2^{nd} derivative with respect to $\boldsymbol{\theta}$ for the log-likelihood function (2.9). And by MLE theory, we know that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges weakly (in distribution) to a multivariate Gaussian with zero mean and variance-covariance $\mathbf{G} = \mathrm{I}(\boldsymbol{\theta})^{-1}$ where

$$\mathrm{I}(\boldsymbol{\theta}) = \mathrm{E}\left(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} l(\boldsymbol{\theta})\right) = \mathrm{E}\begin{pmatrix} -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} l(\boldsymbol{\theta}) & -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}'} l(\boldsymbol{\theta}) \\ -\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}'} l(\boldsymbol{\theta}) & -\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} l(\boldsymbol{\theta}) \end{pmatrix} \tag{2.20}$$

is Fisher's Information. For our model the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$:

$$\mathbf{G} = \begin{pmatrix} \mathrm{E}\left(-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} l(\boldsymbol{\theta})\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}\left(-\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} l(\boldsymbol{\theta})\right)^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{-1} \end{pmatrix} \tag{2.21}$$

where $\mathbf{R} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ and $\mathbf{S} = \{s_{jk}\}_{q\times q}$ with

$$s_{jk} = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{W}_j\mathbf{V}^{-1}\mathbf{W}_k\right) = \frac{1}{2}\sum_{i=1}^{n} z_{ij}z_{ik}$$

if using full maximum likelihood with REML and $s_{jk} = \frac{1}{2}\mathrm{tr}\,(\mathbf{M}\mathbf{W}_j\mathbf{M}\mathbf{W}_k)$ then for large n, the MLE's $\hat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \mathbf{R}^{-1})$ and $\hat{\boldsymbol{\gamma}} \dot{\sim} N(\boldsymbol{\gamma}, \mathbf{S}^{-1})$ are independent. By the assumption of normality for the residuals ϵ_i's the MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is normally distributed,

i.e. $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{R}^{-1})$ for any n. Inference can then be carried out as usual. For example a $100(1-\alpha)\%$ confidence interval for γ_j is given by

$$\hat{\gamma}_j \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{\gamma}_j)}$$

where $\text{Var}(\hat{\gamma}_j)$ is the j^{th} diagonal element of $\mathbf{S}^{-1}$.

Chapter 3

Linear Mixed Model

3.1 Introduction

There are three categories of linear models: *fixed*, *random* and *mixed effects*. The type of linear model depends on whether the equation $\boldsymbol{\beta}$ vector in the $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ is fixed, random, or has both fixed and random (i.e. mixed). In fact, a random model is considered mixed because a fixed overall mean for observations is always assumed to exist [8].

When the distribution of the residuals are normal, but may not be independent or have a constant variance, a linear mixed model (LMM) for the continuous response variable can be used. Because of its flexibility, the mixed model can handle complex, multi-level and hierarchical data. Observations within the same cluster or level are dependent as they belong to the same subpopulation induced in the model by shared random effects, while the observations between clusters or levels are independent. [9] Mixed models have many uses, including but not limited to: modelling complex clustered or longitudinal data, modelling variation with multiple sources, displaying data of variance heterogeneity, dealing with missing data, and combining the Classical and Bayesian models, etc. [8].

3.1.1 Factors and Effects

Factors

We define a fixed factor as a categorical or classification variable for which the statistician will include all relevant levels in a study. These factors may inlcude qualitative covariate or ordinal classification variables. We define a random factor as a classification variable for which the statistician will include a random sample from a population of levels in a study. This random choice of levels is done to allow the statistician to make inference about the entire population of levels. While fixed factors only represent themselves, random factors are representative of the entire population of levels being studied.

Effects

We define fixed effects as regression coefficients or fixed effects parameters. The relationship between the dependent variable and the predictor variables is described by fixed effects. We assume that the fixed effects are an unknown fixed quantity in the LMM. We define random effects as the random values specified to given levels of a random factor in the LMM. These values normally represent the random deviations of the relationships that fixed effects describe. Random effects are introduced to model dependence within a cluster, which allows for borrowing strength. While fixed effects are represented as fixed variables in a LMM, random effects are represented as random variables.

3.2 Mathematical Representation of LMM

We can use the information mentioned above to fit the LMM, which is represented by

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i}_{\text{random}} \tag{3.1a}$$

$$\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \mathbf{D}_i) \tag{3.1b}$$

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i) \tag{3.1c}$$

where $i = 1, \ldots, m$ and,

$$\mathbf{Y}_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{pmatrix}, \quad \mathbf{X}_i = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i 1} & X_{n_i 2} & \cdots & X_{n_i p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\mathbf{Z}_i = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1q} \\ Z_{21} & Z_{22} & \cdots & Z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i 1} & Z_{n_i 2} & \cdots & Z_{n_i q} \end{pmatrix}, \quad \boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \vdots \\ \gamma_{qi} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \vdots \\ \epsilon_{n_i i} \end{pmatrix}.$$

Where in equation (3.1a),

- $\mathbf{Y}_i$ is an $(n_i \times 1)$ vector of continuous response for the i^{th} subject/CMG.
- n_i is the number of elements that may vary in each subject.
- $\mathbf{X}_i$ is an $(n_i \times p)$ fixed-effects design matrix, representing the known values of the p covariates, $X^{(1)}, \ldots, X^{(p)}$ for the i^{th} subject.

- $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients (or fixed-effect parameters) associated with the p covariates in $\mathbf{X}_i$.
- $\mathbf{Z}_i$ is an $(n_i \times q)$ random design matrix, representing the known values of the q covariates, $Z^{(1)}, \ldots, Z^{(q)}$, for the i^{th} subject.
- $\boldsymbol{\gamma}_i$ vector is a $(q \times 1)$ vector of random effects associated with the q covariates in the $\mathbf{Z}_i$ matrix for the i^{th} subject.
- $\boldsymbol{\epsilon}_i$ is an $(n_i \times 1)$ vector of residuals, with each element in $\boldsymbol{\epsilon}_i$ denoting the residual associated with an observed response at occasion t for the i^{th} subject.

We assume that the $\mathbf{X}_i$ matrices are of full rank; that is, none of the columns (or rows) is a linear combination of the other columns (or rows). In some cases where $\mathbf{X}_i$ matrices are not a full rank, some aliasing (or parameter identifiability) problems may occur for the fixed effects vector $\boldsymbol{\beta}$. The $\mathbf{Z}_i$ matrix is similar to the $\mathbf{X}_i$ matrix since it represents the observed values of covariates; however, it normally has fewer columns than the $\mathbf{X}_i$ matrix.

The random covariate in the $\mathbf{Z}_i$ matrix will have effects on the continuous response variable, which vary randomly, across different i^{th} subjects. In many cases, predictors with effects that vary randomly across subjects are represented in both $\mathbf{X}_i$ and $\mathbf{Z}_i$.

Assumptions:

Since random effects are random variables, we assume that the $\boldsymbol{\gamma}_i$ vector follows a multivariate normal distribution of dimension q, with mean vector $\mathbf{0}$ and variance covariance matrix denoted by $\mathbf{D}_i$:

$$\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \mathbf{D}_i) \text{ where } \mathbf{D}_i = \text{Cov}(\boldsymbol{\gamma}_i) = \begin{pmatrix} \text{Var}(\gamma_{1i}) & \text{Cov}(\gamma_{1i}, \gamma_{2i}) & \cdots & \text{Cov}(\gamma_{1i}, \gamma_{qi}) \\ \text{Cov}(\gamma_{1i}, \gamma_{2i}) & \text{Var}(\gamma_{2i}) & \cdots & \text{Cov}(\gamma_{2i}, \gamma_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\gamma_{1i}, \gamma_{qi}) & \text{Cov}(\gamma_{2i}, \gamma_{qi}) & \cdots & \text{Var}(\gamma_{qi}) \end{pmatrix}. \quad (3.2)$$

The main diagonal in a $(q \times q)$ matrix $\mathbf{D}_i$, which is symmetric and positive definite, is the variance of each random effect in $\boldsymbol{\gamma}_i$ that is associated to i^{th} subject. The off-diagonal elements in this matrix are the covariances between two corresponding random effects. We place the elements (variances and covariances) of the matrix $\mathbf{D}_i$ in the vector $\theta_{\mathbf{D}i}$, which imposes a structure (or constraints) on the elements of $\mathbf{D}_i$ matrix. These elements are defined as functions of a set of unique covariance parameters. Vector $\theta_{\mathbf{D}i}$ has different structures that we will cover in section 3.3.
We also assume that:

- Residuals associated with different subjects are independent of each other.
- All random vectors $\{\boldsymbol{\epsilon}_i, \boldsymbol{\gamma}_i, i = 1, \cdots, m\}$ are mutually independent.
- The residual vector $\boldsymbol{\epsilon}_i$ is a random vector that follows a multivariate normal distribution with a mean vector $\mathbf{0}$ and a positive definite symmetric covariance matrix $\mathbf{R}_i$.
- The residuals associated with repeated observations on the same subject in a LMM can be correlated.

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i) \text{ where } \mathbf{R}_i = \text{Cov}(\boldsymbol{\epsilon}_i) = \begin{pmatrix} \text{Var}(\epsilon_{1i}) & \text{Cov}(\epsilon_{1i}, \epsilon_{2i}) & \cdots & \text{Cov}(\epsilon_{1i}, \epsilon_{n_i i}) \\ \text{Cov}(\epsilon_{1i}, \epsilon_{2i}) & \text{Var}(\epsilon_{2i}) & \cdots & \text{Cov}(\epsilon_{2i}, \epsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\epsilon_{1i}, \epsilon_{n_i i}) & \text{Cov}(\epsilon_{2i}, \epsilon_{n_i i}) & \cdots & \text{Var}(\epsilon_{n_i i}) \end{pmatrix}. \quad (3.3)$$

The main diagonal in an $(n_i \times n_i)$ matrix $\mathbf{R}_i$, which is symmetric and positive definite, is the variance of each residual ϵ_{ti}. The off-diagonal elements of this matrix are the covariances between two corresponding residuals. We place the elements (variances

and covariances) of the matrix $\mathbf{R}_i$ in the vector $\theta_{\mathbf{R}i}$, which imposes a structure (or constraints) on the elements of $\mathbf{R}_i$ matrix. These elements are defined as functions of a set of unique covariance parameters. Vector $\theta_{\mathbf{R}i}$ has different structures that we will cover in section 3.3.

3.3 Covariance Structure for Matrices R_i and D_i

The covariance matrices $\mathbf{D}_i$ and $\mathbf{R}_i$ are not only positive definite and symmetric, but they are also commonly assumed to take on specific structures such as ***variance components, unstructured, compound symmetry, Toeplitz and autoregressive, etc.*** representing the type of assumed dependence within the cluster. Next, we will focus on the variance component, unstructured and compound symmetry structures of the covariance matrices.

3.3.1 Variance Component Covariance Matrix

The variance component (VC) structure (or diagonal) is a commonly used structure for matrix $\mathbf{D}_i$ because all off-diagonal covariances in the matrix are defined as zero. Each random effect in γ_i has its own variance (the diagonals of matrix $\mathbf{D}_i$). We use the diagonal elements to make up the vector $\theta_{\mathbf{D}i}$ of length q, where q is the number of covariance parameters.

If the LMM has two random effects for the i^{th} subject, then matrix $\mathbf{D}_i$ has this form: $\begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{12}^2 \end{pmatrix}$, which gives a two-parameter vector $\theta_{\mathbf{D}i} = \begin{pmatrix} \sigma_{11}^2 \\ \sigma_{12}^2 \end{pmatrix}$.

When the uncorrelated and equal variance residuals are associated with the observations on the same subject, the diagonal structure is the simplest structure used for the $\mathbf{R}_i$ covariance matrix, of this form: $\mathbf{R}_i = \text{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2\mathbf{I}$ for each subject i. It can then be deduced that $\theta_{\mathbf{R}i} = (\sigma^2)$ has one parameter that defines the constant variance. Other structures can be used but may not follow the same simplicity of

the diagonal structure. For example, the Toeplitz structure may be used; however, even though it allows more flexibility with the correlations, it requires the use of more covariance parameters in the $\theta_{\mathbf{R}i}$ vector.

3.3.2 Unstructured D_i Covariance Matrix

The unstructured $\mathbf{D}_i$ matrix is imposed in the random coefficient LMM. The off-diagonals are the covariances of two different random effects and the diagonal is the variance of each random effect in γ_i. Since matrix $\mathbf{D}_i$ is symmetric, it is implied that vector $\theta_{\mathbf{D}i}$ has parameters $(q \times (q+1))/2$.

If the LMM has two random effects for the i^{th} subject, then matrix $\mathbf{D}_i$ has this form: $\begin{pmatrix} \sigma_{11}^2 & \sigma_{11,12} \\ \sigma_{11,12} & \sigma_{12}^2 \end{pmatrix}$, which gives a three-parameter vector $\theta_{\mathbf{D}i} = \begin{pmatrix} \sigma_{11}^2 \\ \sigma_{11,12} \\ \sigma_{12}^2 \end{pmatrix}$.

3.3.3 Compound Symmetry R_i Covariance Matrix

When the assumption that the equal correlation of residuals for each subject i is true, the compound symmetry structure is used for the $\mathbf{R}_i$ covariance matrix:

$$\mathbf{R}_i = \mathrm{Cov}\left(\boldsymbol{\epsilon}_i\right) = \begin{pmatrix} \sigma^2+\sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2+\sigma_1 \end{pmatrix}.$$

The compound symmetry structure has two parameters in vector $\theta_{\mathbf{R}i}$ that define a variance and covariance parameter for matrix $\mathbf{R}_i$. So, $\theta_{\mathbf{R}i} = \begin{pmatrix} \sigma^2 \\ \sigma_1 \end{pmatrix}$ has two parameters, one that defines the constant variance and the other that defines the covariance.

3.4 General Matrix Form of LMM

In reference to section 3.2, we have seen m occurrences of the equation (3.1a) that can be stacked into general vector and matrix form.

$$\mathbf{Y} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}}_{\text{random}} \tag{3.4a}$$

$$\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{D}) \tag{3.4b}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}) \tag{3.4c}$$

where we stack as follows,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{Z}_m \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}.$$

Where in equations (3.4a),(3.4b) and (3.4c),

- $n = \sum_{i=1}^{m}(n_i)$ where n_i is the number of elements that may vary in each subject
- $\mathbf{Y}$ is an $(n \times 1)$ vector of the continuous response

- $\mathbf{X}$ is an $(n \times p)$ fixed-effects design matrix, representing the known values of the p covariates, $X^{(1)}, \ldots, X^{(p)}$
- $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients (or fixed-effect parameters) associated with the p covariates in $\mathbf{X}$
- $\mathbf{Z}$ is an $(n \times mq)$ random design matrix, representing the known values of the mq covariates, $Z^{(1)}, \ldots, Z^{(mq)}$
- $\boldsymbol{\gamma}$ is an $(mq \times 1)$ vector of random effects associated with the mq covariates in the $\mathbf{Z}$ matrix
- $\mathbf{D}$ is an $(mq \times mq)$ covariance matrix for $\boldsymbol{\gamma}$, with $\mathbf{D} = \text{diag}(\mathbf{D}_1, \cdots, \mathbf{D}_m)$
- $\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of residuals
- $\mathbf{R}$ is an $(n \times n)$ covariance matrix for the residuals, with $\mathbf{R} = \text{diag}(\mathbf{R}_1, \cdots, \mathbf{R}_m)$

3.5 Estimation Methods

Before we can estimate our unknown fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{\gamma}$ and the variance vector $\boldsymbol{\theta}_i = \{\theta_{\mathbf{R}i}, \theta_{\mathbf{D}i}\}$, we must make the LMM identifiable for: $\boldsymbol{\beta}$, σ^2 and $\mathbf{D}_i$. In order to do so, we assume that:

- matrix $\sum \mathbf{X}_i' \mathbf{X}_i$ is nonsingular and $\sum_{i=1}^{m} n_i > p$; and
- at least one matrix $\mathbf{Z}_i' \mathbf{Z}_i$ is positive definite and $\sum_{i=1}^{m} (n_i - q) > 0$.

For more details on identifiability, see section 3.2 in Demidenko (2004). [9]

In section 3.2, we assumed that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\gamma}_i$ are normally distributed. Equations (3.1b) and (3.1c) imply that the response vector $\mathbf{Y}_i$, where $i = 1, \cdots, m$ is normally distributed as follows:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{V}_i) \tag{3.5}$$

Since,

$$\begin{pmatrix} \gamma_i \\ \epsilon_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} \mathbf{D}_i & 0 \\ 0 & \mathbf{R}_i \end{bmatrix} \right)$$

where,

$$\mathrm{E}\left(\mathbf{Y}_i\right) = \mathrm{E}\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i\right) = \mathbf{X}_i\boldsymbol{\beta}$$

$$\mathbf{V}_i = \mathrm{Cov}\left(\mathbf{Y}_i\right) = \mathrm{Cov}\left(\mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i\right) = \mathbf{Z}_i\mathrm{Cov}\left(\boldsymbol{\gamma}_i\right)\mathbf{Z}_i' + \mathrm{Cov}\left(\boldsymbol{\epsilon}_i\right) = \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i' + \mathbf{R}_i$$

and $\mathbf{V}_i$ is assumed to be a positive definite nonsingular matrix.

In section 3.4 we assumed that $\boldsymbol{\epsilon}$ and $\boldsymbol{\gamma}$ are normally distributed. Equations (3.4b) and (3.4c) imply that the response vector $\mathbf{Y}$ is normally distributed as follows:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \tag{3.6}$$

where,

$$\mathrm{E}\left(\mathbf{Y}\right) = \mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\right) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{V} = \mathrm{Cov}\left(\mathbf{Y}\right) = \mathrm{Cov}\left(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\right) = \mathbf{Z}\mathrm{Cov}\left(\boldsymbol{\gamma}\right)\mathbf{Z}' + \mathrm{Cov}\left(\boldsymbol{\epsilon}\right) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$$

and $\mathbf{V}$ is assumed to be a positive definite nonsingular matrix.

We can express the LMM as a hierarchical model as follows:

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \mathbf{R}) \quad \text{with } \mathbf{R} = \mathbf{R}(\boldsymbol{\theta}) \\ \boldsymbol{\gamma}|\boldsymbol{\theta} &\sim N(\mathbf{0}, \mathbf{D}) \qquad\quad \text{with } \mathbf{D} = \mathbf{D}(\boldsymbol{\theta}) \end{aligned} \tag{3.7}$$

where $\boldsymbol{\theta} = \{\theta_{\mathbf{R}}, \theta_{\mathbf{D}}\}$ and $\theta_{\mathbf{R}}$ and $\theta_{\mathbf{D}}$ are the variance-covariance component vectors for matrix $\mathbf{R}$ and $\mathbf{D}$ respectively.

We write:

$$\mathrm{E}\left(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \mathrm{E}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

In order to find the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ in models (3.1a) and (3.4a), we must estimate the values of $\mathbf{V}_i$ and $\mathbf{V}$. These estimations must then be used as the real values for $\mathbf{V}_i$ and $\mathbf{V}$.

The following is an explanation of the estimation methods that will be used to estimate our unknown fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{\gamma}$ and the variance vector $\boldsymbol{\theta_V}$. The methods include: Maximum Likelihood (ML), Restricted/Residual Maximum Likelihood (REML), Minimum Variance Quadratic Unbiased Estimation (MIVQUE) and Type 1 to Type 3. Only the ML and REML will be covered in section 3.5.1.

3.5.1 Maximum and Restricted Maximum Likelihood

Obtaining the estimates of unknown parameters by maximizing a likelihood function is called maximum likelihood estimation. The distribution of the vector $\mathbf{Y}_i$ has a multivariate normal probability density function (pdf).

$$f\left(\mathbf{Y}_i|\boldsymbol{\beta},\boldsymbol{\theta}_i\right)=\frac{1}{(2\pi)^{\frac{n_i}{2}}\left|\mathbf{V}_i\right|^{\frac{1}{2}}}exp\Big\{-\frac{1}{2}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)'\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)\Big\} \tag{3.8}$$

Thus, the likelihood function is given by:

$$\mathcal{L}\left(\boldsymbol{\beta},\boldsymbol{\theta}_i\right)=\prod_{i=1}^{m}\frac{1}{(2\pi)^{\frac{n_i}{2}}\left|\mathbf{V}_i\right|^{\frac{1}{2}}}exp\Big\{-\frac{1}{2}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)'\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)\Big\} \tag{3.9}$$

Similarly to the derivation shown in chapter 2 in section 2.3.1, using the ML methods discussed, we obtain the maximum by solving the following equations:

$$\frac{\partial l}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{m}\Big\{\mathbf{X}_i'\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\Big)\Big\} \tag{3.10a}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\theta}_{ij}} &= -\frac{1}{2}\Big\{\sum_{i=1}^{m}\frac{\partial}{\partial \boldsymbol{\theta}_{ij}}\Big\{\ln\left(|\mathbf{V}_i|\right)\Big\} + \sum_{i=1}^{m}\frac{\partial}{\partial \boldsymbol{\theta}_{ij}}\Big\{\Big(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\Big)'\mathbf{V}_i^{-1}\Big(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\Big)\Big\}\Big\} \\ &= -\frac{1}{2}\Big\{\sum_{i=1}^{m} tr\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_{ij}}\right) - \sum_{i=1}^{m}\Big(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\Big)'\Big\{\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\theta}_{ij}}\mathbf{V}_i^{-1}\Big\}\Big(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\Big)\Big\}\end{aligned} \tag{3.10b}$$

where $\boldsymbol{\theta}_{ij}$ is the j^{th} element of the variance vector $\boldsymbol{\theta}_i$, where $j = 1, \cdots, r$.

The optimal value of equation (3.10a) is then obtained by $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the GLS estimator:

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\Big)^{-1}\Big(\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i\Big) \tag{3.11}$$

When the covariance matrix $\mathbf{V}_i$ is known, $\hat{\boldsymbol{\beta}}$ is BLUE of $\boldsymbol{\beta}$, but if $\mathbf{V}_i$ is unknown, we replace it with the estimation matrix $\hat{\mathbf{V}}_i$=$\mathbf{Z}_i\hat{\mathbf{D}}_i\mathbf{Z}_i' + \hat{\mathbf{R}}_i$, where the estimator $\hat{\boldsymbol{\beta}}$ is called Empirical Best Linear Unbiased Estimator (EBLUE) of $\boldsymbol{\beta}$. To estimate the covariance parameters in $\boldsymbol{\theta}_i$, we replace the expression of $\hat{\boldsymbol{\beta}}$ in equation (3.11). We do this in order to construct the *profile log-likelihood function* $l_p\left(\boldsymbol{\theta}_i\right)$. We notice that this function is similar to chapter 2, section 2.3.1; however, the first term $\mathbf{1}'\mathbf{Z}\boldsymbol{\gamma}$ will change to $\sum_{i=1}^{m}\ln(|\mathbf{V}_i|)$.

The distribution of the vector $\mathbf{Y}$ in the equation (3.6) has the multivariate normal pdf.

$$f\left(\mathbf{Y}|\boldsymbol{\beta},\boldsymbol{\theta}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{V}|^{\frac{1}{2}}}exp\Big\{-\frac{1}{2}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big)'\mathbf{V}^{-1}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big)\Big\} \tag{3.12}$$

Thus, the log-likelihood function is given by:

$$l_y\left(\boldsymbol{\beta},\boldsymbol{\theta}\right) = -\frac{1}{2}\Big\{n\ln\left(2\pi\right) + \ln\left(|\mathbf{V}|\right) + \Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big)'\mathbf{V}^{-1}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big)\Big\} \tag{3.13}$$

We differentiate the log-likelihood with respect to the parameters to obtain the following equations:

$$\frac{\partial l_y}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big) \tag{3.14a}$$

$$\frac{\partial l_y}{\partial \boldsymbol{\theta}_k} = -\frac{1}{2}tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_k}\right) + \frac{1}{2}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big)'\Big\{\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_k}\mathbf{V}^{-1}\Big\}\Big(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\Big) \tag{3.14b}$$

Similarly, we set the equations (3.14a) and (3.14b) to zero and solve them in order to determine the ML estimator, where $\boldsymbol{\theta}_k$ is the k^{th} element of the variance vector $\boldsymbol{\theta}$, where $k = 1, \cdots, s$. Let $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\}$ be the MLE for $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$. The optimal value of equation (3.14a) occurs at $\boldsymbol{\beta}$ if the matrix $\mathbf{X}$ is of full rank. This is denoted by:

$$\hat{\boldsymbol{\beta}} = \Big(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{3.15}$$

Since the MLE is consistent and asymptotically normal with an asymptotic covariance matrix, it will be equal to the inverse of the Fisher information matrix

$$\mathrm{I}^{-1}(\boldsymbol{\beta}) = -\mathrm{E}\left(\frac{\partial^2 l_y}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)$$

$$E\left(\frac{\partial^2 l_y}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right) = -\Big(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big). \tag{3.16}$$

The variance of $\hat{\boldsymbol{\beta}}$ is $\mathrm{Var}\Big(\hat{\boldsymbol{\beta}}\Big) = \Big(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\Big)^{-1}$ where $\hat{\mathbf{V}}$ is the estimate of $\mathbf{V}$. The diagonal elements on $\mathrm{Var}\Big(\hat{\boldsymbol{\beta}}\Big)$ are biased since the uncertainty is not considered when replacing the $\mathbf{V}$ with $\hat{\mathbf{V}}$. This bias is added to the bias in estimation of $\boldsymbol{\theta}$ in the ML method. In order to account for these biases, we approximate degree of freedom (df) for the t-test and F-test for fixed effects. The approximation methods that apply to these tests consider the presence of random and correlated residuals. The MLE for equation (3.14b) occurs at $\boldsymbol{\theta}$ and is obtained using the same equation (2.13), where

the derivative with respect to γ_i will be replaced with the derivative with respect to $\boldsymbol{\theta}_k$, and equation (2.14) as done in chapter 2, section 2.3.1.

Alternatively, we can derive the REML function similarly to the function (2.18), with the term $\mathbf{1}'\mathbf{Z}\boldsymbol{\gamma}$ changed to $\ln|\mathbf{V}|$. The restricted log-likelihood function is a function of $\boldsymbol{\theta}$ only, so the REML method is a method of estimating $\boldsymbol{\theta}$ because we replaced $\boldsymbol{\beta}$ before estimating. The covariance matrix is easily determined because the REML estimator is consistent and asymptotically normal. The covariance matrix is equal to the inverse of the restricted Fisher information matrix:

$$\mathrm{Var}\left(\frac{\partial l_R}{\partial \boldsymbol{\theta}}\right) = -\mathrm{E}\left(\frac{\partial^2 l_R}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right) \tag{3.17}$$

If $\mathbf{V}$ is twice differentiable with respect to the components of $\boldsymbol{\theta}$, then:

$$\mathrm{E}\left(\frac{\partial^2 l_R}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l}\right) = -\frac{1}{2} tr\left(M\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_k} M\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_l}\right) \ \textit{where } k, l = 1, \cdots, s \tag{3.18}$$

The maximum likelihood estimation of variances is biased for finite samples; so the unbiased estimator of the variance is the residual of sum of squares (SS) divided by the df $= (n - p)$, where n represents the number of observations and p represents the number of the coefficient parameters.

3.6 Prediction of Random Effects

The values in the random vector $\boldsymbol{\gamma}_i$ are predicted rather than estimated because they are random variables. In fixed effects, we are interested in estimating the mean; however, in the multivariate normal distribution of random effects, we assume that the expected value of these random effects is zero mean vector. Assuming the normality

of the data, we have:

$$\begin{pmatrix} \gamma_i \\ \mathbf{Y}_i \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}_i\boldsymbol{\beta} \end{bmatrix} \quad \begin{bmatrix} \mathbf{D}_i & \mathbf{D}_i\mathbf{Z'}_i \\ \mathbf{Z}_i\mathbf{D}_i & \mathbf{V}_i \end{bmatrix} \right) \tag{3.19}$$

because the covariance of $(\gamma_i, \mathbf{Y}_i)$ is

$$\text{Cov}(\gamma_i, \mathbf{Y}_i) = \text{Cov}(\gamma_i, \mathbf{X}_i\boldsymbol{\beta}) + \text{Cov}(\gamma_i, \mathbf{Z}_i\gamma_i) + \text{Cov}(\gamma_i, \boldsymbol{\epsilon}_i) = \mathbf{D}_i\mathbf{Z'}_i$$

and the covariance of $(\mathbf{Y}_i, \gamma_i)$ is

$$\text{Cov}(\mathbf{Y}_i, \gamma_i) = \text{Cov}(\mathbf{X}_i\boldsymbol{\beta}, \gamma_i) + \text{Cov}(\mathbf{Z}_i\gamma_i, \gamma_i) + \text{Cov}(\boldsymbol{\epsilon}_i, \gamma_i) = \mathbf{Z}_i\mathbf{D}_i$$

When the fixed effects $\boldsymbol{\beta}$ and the variance components of the vector $\boldsymbol{\theta}_i$ are known, the conditional expectation of the random effect given data is the Best Linear Unbiased Prediction (BLUP) of the random effects in γ_i for the i^{th} subject, see equation (3.20). We use the property of bivariate normal distribution as seen in equation (C.2) in Appendix C.

$$\tilde{\gamma}_i = \text{E}\left(\gamma_i|\mathbf{Y}_i\right) = \mathbf{D}_i\mathbf{Z'}_i\mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right) \tag{3.20}$$

Using the same derivation method as above, the BLUP of the random effect vector γ in general matrix form (3.4a) is:

$$\tilde{\gamma} = \text{E}\left(\gamma|\mathbf{Y}\right) = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \tag{3.21}$$

If we replace the covariance matrices: $\mathbf{V}_i$, $\mathbf{D}_i$, $\mathbf{V}$ and $\mathbf{D}$, as well as $\boldsymbol{\beta}$, for equations (3.20) and (3.21) with their estimates, we then obtain Empirical Best Linear Unbiased Prediction (EBLUP)

$$\tilde{\gamma}_i = \text{E}\left(\gamma_i|\mathbf{Y}_i\right) = \hat{\mathbf{D}}_i\mathbf{Z'}_i\hat{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right) \tag{3.22}$$

The EBLUP of the random effect vector $\boldsymbol{\gamma}$ in general matrix form (3.4a) is:

$$\tilde{\boldsymbol{\gamma}} = \mathrm{E}(\boldsymbol{\gamma}|\mathbf{Y}) = \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{3.23}$$

3.7 Variance-Covariance Matrices

The predictor for $\boldsymbol{\gamma}$ has the form $\hat{\boldsymbol{\gamma}} = \mathbf{CY}$ where:

$$\mathbf{C} = \mathbf{D}\mathbf{Z}'\left(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\right) \tag{3.24}$$

There is a covariance matrix for the BLUEs, BLUPs, EBLUEs and EBLUPs based on whether or not the covariance matrices $\mathbf{R}_i, \mathbf{D}_i, \mathbf{R}$ and $\mathbf{D}$ are known.
When matrix $\mathbf{R}$ and $\mathbf{D}$ are known, then the covariance matrix of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is:

$$\mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \\ \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix}^{-} \tag{3.25}$$

When matrix $\mathbf{R}$ and $\mathbf{D}$ are not known, then the covariance matrix of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is:

$$\mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \\ \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{D}}^{-1} \end{pmatrix}^{-} \tag{3.26}$$

The equation (3.26) can be written as follows:

$$\mathrm{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \\ \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-} & (-\hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-})' \\ \\ (-\hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}) & (\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{D}}^{-1})^{-1} + \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Z}\hat{\mathbf{D}} \end{pmatrix} \tag{3.27}$$

The covariance matrix of the EBLUPs of the i^{th} subject can be written as follows:

$$\text{Cov}(\tilde{\gamma}_i) = \text{Cov}\left(\hat{\mathbf{D}}_i\mathbf{Z}'_i\hat{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right) = \hat{\mathbf{D}}_i\mathbf{Z}'_i\left(\hat{\mathbf{V}}_i^{-1} - \hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i\hat{\mathbf{V}}_i^{-1}\right)\mathbf{Z}_i\hat{\mathbf{D}}_i \quad (3.28)$$

3.8 Residuals: Marginal and Conditional

The marginal residual is expressed as:

$$\hat{\boldsymbol{\epsilon}}_m = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \quad (3.29)$$

Where the variance for the marginal residual is:

$$\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_m) = \hat{\mathbf{V}} - \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}' \quad (3.30)$$

The conditional residual is expressed as:

$$\hat{\boldsymbol{\epsilon}}_c = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\gamma}} \quad (3.31)$$

Where the variance for the conditional residual is:

$$\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_c) = (\mathbf{I} - \mathbf{Z}\hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1})(\hat{\mathbf{V}} - \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}')(\mathbf{I} - \mathbf{Z}\hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1})' \quad (3.32)$$

The studentized marginal residuals can be calculated by dividing the raw residuals $\hat{\boldsymbol{\epsilon}}_m$ by their estimated standard deviations $\sqrt{\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_m)}$. They are denoted as:

$$\hat{\boldsymbol{\epsilon}}_m^{student} = \frac{\hat{\boldsymbol{\epsilon}}_m}{\sqrt{\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_m)}} \quad (3.33)$$

If we divide the raw residuals by their true standard deviations, we obtain the marginal standardized residuals.

The studentized conditional residuals can be calculated by dividing the raw residuals $\hat{\boldsymbol{\epsilon}}_c$ by their estimated standard deviations $\sqrt{\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_c)}$. They are denoted as:

$$\hat{\boldsymbol{\epsilon}}_c^{student} = \frac{\hat{\boldsymbol{\epsilon}}_c}{\sqrt{\hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}_c)}} \tag{3.34}$$

If we divide the conditional raw residuals by their true standard deviations, we obtain the conditional standardized residuals.

Chapter 4

Generalized Linear Model

4.1 Introduction

The Generalized Linear Model (GLM) is used to unify various statistical models in order to yield more favourable results. When a GLM is used, the distribution of the response variable ($\mathbf{Y}$) must belong to an exponential family (EF). For more details on the EF and its properties, see Appendix C.2.

4.2 Generalized Linear Model

The generalized linear model extends the linear model by $\eta_i = g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$ and $Y_i \sim EF(\mu_i, \phi)$ where ϕ is a scale parameter and g is a monotone link function. The GLM is related to the expected value of the response $E(Y_i) = \mu_i$ to a linear prediction η_i via the link function $g(.)$

Let $g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) = \mu_i$ be the inverse link function. For any random variable $Y|X$ with a pdf $f_Y\Big(y; g^{-1}(\mathbf{x}'\boldsymbol{\beta}), \phi\Big)$, which depends on a canonical parameter $\mu = g^{-1}(\mathbf{x}'\boldsymbol{\beta})$, the pdf is then considered a member of the EF if:

$$f_Y\Big(y; \mu, \phi\Big) = exp\Big\{\frac{s(y)\mu - a(\mu)}{b(\phi)} + c(y, \phi)\Big\}. \tag{4.1}$$

Using the nice property of the EF distribution, we can easily obtain the mean and the variance for the random variable Y. To obtain the mean and the variance, we use the log-likelihood of $f_Y\Big(y;\mu,\phi\Big)$ denoted below:

$$l(\mu,\phi,y) = log(f_Y(y;\mu,\phi) = \frac{s(y)\mu - a(\mu)}{b(\phi)} + c(y,\phi) \tag{4.2}$$

We know that $\mathrm{E}(\frac{\partial l}{\partial \mu}) = 0$ and the expected value of Y is:

$$E(Y) = \mu = a'(\mu) = a'(g^{-1}(\mathbf{x}'\boldsymbol{\beta})). \tag{4.3}$$

We also know that $\mathrm{E}(\frac{\partial^2 l}{\partial \mu^2}) = -\mathrm{E}(\frac{\partial l}{\partial \mu})^2$ because the variance of Y and hence,

$$\mathrm{Var}(Y) = a''(\mu)b(\phi) = a''(g^{-1}(\mathbf{x}'\boldsymbol{\beta}))b(\phi). \tag{4.4}$$

The link function in a GLM relates the expected value Y to the covariates. Each member of the EF has a different canonical link function as a result of its specific distribution. Below, we list some of these distributions, their link functions as well as the mean functions.

Table 4.1: Exponential Distribution

Exponential Distribution	Link Name	Link Function	Mean Function
Bernoulli	Logit	$\mathbf{x}'\boldsymbol{\beta} = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+exp(-\mathbf{x}'\boldsymbol{\beta})}$
Binomial	Logit	$\mathbf{x}'\boldsymbol{\beta} = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+exp(-\mathbf{x}'\boldsymbol{\beta})}$
Exponential	Inverse	$\mathbf{x}'\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{x}'\boldsymbol{\beta})^{-1}$
Gamma	Inverse	$\mathbf{x}'\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{x}'\boldsymbol{\beta})^{-1}$
Inverse Gaussian	Inverse squared	$\mathbf{x}'\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{x}'\boldsymbol{\beta})^{-\frac{1}{2}}$
Normal	Identity	$\mathbf{x}'\boldsymbol{\beta} = \mu$	$\mu = \mathbf{x}'\boldsymbol{\beta}$
Multinomial	Logit	$\mathbf{x}'\boldsymbol{\beta} = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+exp(-\mathbf{x}'\boldsymbol{\beta})}$
Poisson	Log	$\mathbf{x}'\boldsymbol{\beta} = ln(\mu)$	$\mu = exp(\mathbf{x}'\boldsymbol{\beta})$

In GLM, the unknown parameter $\boldsymbol{\beta}$, is commonly estimated using the methods of MLE. Here we will fit a gamma GLM with log link function to predict the cost based on the RIW. We fit the GLM to predict the response variable by replacing the expected value of $\mathbf{Y}$ with $\hat{\boldsymbol{\mu}}$ and we measure the discrepancy of this fitted model, the goodness of fit and the deviation from the saturated model (full model) by the deviance. The *residual deviance* function is defined as twice the log likelihood ratio, and is denoted as follows:

$$D(c, s) = -2log\Big(\frac{L_c}{L_s}\Big) \tag{4.5}$$

where L_c is the likelihood for the current model and L_s is the likelihood for the saturated model. The random variable $D(c, s)$ is asymptotically distributed as χ^2_{n-p}, where p is the number of fitted parameters and n is the sample size. Whereas the residual sum of squares is used in the linear model to check the goodness of fit, the deviance is used for this purpose in the GLM. The scale deviance $D^*(c, s)$ is obtained by dividing the $D(c, s)$ by the estimated dispersion parameter.

Chapter 5

Data Analysis

5.1 Introduction

This analysis discusses various models in an attempt to identify the most efficient estimation method. We also briefly discuss the resource intensity weights and how data is organized based on one MCC. The models being tested and analysed are the Current model, the Heteroscedastic Regression (HER) model and the Heteroscedastic, Random and Fixed Effects (HEREM) model and gamma Generalized Linear Model (GLM) with log link function. All models, except the current, use the optimization algorithms discussed in Appendix B to estimate the parameters.

5.2 Resource Intensity Weights

Patient data is coded based on atypical codes, which include typical and atypical cases. As the name implies, atypical cases are cases where patients experience unusual or exceptional circumstances such as death, transfer to and/or from other acute care institutions, or sign-out. Each atypical case/category is treated using a different methodology. For this reason, this thesis touches on typical cases only as they make up the majority of cases. Typical cases are expected cases that fall within the normal

parameters of a medical condition or situation. Typical cases fall within the same code list as atypical codes; they are coded as 00 (see Table 5.1). Furthermore, typical cases include both factor and non-factor cases without the presence of unusual circumstances that would otherwise define the case as atypical. The typical data also defines the average inpatient RIW, which is set to a value of 1. The RIW is a relative cost weight value derived from case-cost data submitted to CIHI's MIS and Costing department. All RIW cost weights are relative to the average typical inpatient case such that the sum of typical cases is equal to the sum of the typical weighted cases. CIHI receives cost data from provincial and regional case-costing jurisdictions. The organisation then uses the last two years' information to predict the new cost weight for each individual case; this information is later used when various stakeholders allocate funding. With regard to costing, two types of cases may be considered: factor cases and non-factor cases. [1] The RIW estimates are adjusted beyond the observed base CMG and age-category values to adjust for the observed factors. The base table CMG and age category-specific values define the RIW for all non-factor cases, and if appropriate, the value is adjusted for an atypical code. [6]

Table 5.1: CMG+ Atypical Code List [1]

Atypical code	Definition	Atypical code	Definition
00	Typical	11	LS Transfer In
01	Transfer In	12	LS Transfer Out
02	Transfer Out	13	LS Sign-Out
03	Sign-Out	14	LS Death
04	Death	15	LS Transfer In and Transfer Out
05	Transfer In and Transfer Out	16	LS Transfer In and Sign-Out
06	Transfer In and Sign-Out	17	LS Transfer In and Death
07	Transfer In and Death	97	Invalid Length of Stay
08	CMG > 989	98	Not Applicable (Day Surgery Record and Unusual CMG)
10	Long-Stay (LS)		

5.3 MCC5 Data

CIHI's data is organized in a hierarchical manner. Codes are assigned based on the MRDx, then on intervention type, then on CMG and finally on MCC. For the purpose of this thesis, MCC 5 data from 2007 and 2008 is the only data used. The typical cases are used as the reference data set. This data totals approximately 86,759 observations and 377 dummy variables, including the main effects. These observations account for all sample sizes above the minimum sample size of 30+ patients for each category of the main covariates and interactions. All five factors are categorical variables and so, we parameterize them in different ways, as outlined below. The IE and CL are parameterized as ordinal variables. All other categoric variables are parameterized by using a reference category.

- The reference category for Age is category (R).
- The reference category for all FIs is category (0).
- The reference category for OOH is category (No).
- The reference category for the CMGs is category (CMG 905).
- The reference category for the health care facilities is category (90604).
- The reference category for the fiscal year is category (2007).

Note: All statistical outliers were not included in the RIW models due to the extreme values (greater than 3* SD) when compared with similar CMG cases.

5.4 Current Model

The Current model aims to estimate the parameters and calculate the weights for the heteroscedasticity of the residuals. The model is represented in section 2.2. The

following histogram and Q-Q normal plot demonstrates how the residuals follow a normal distribution.

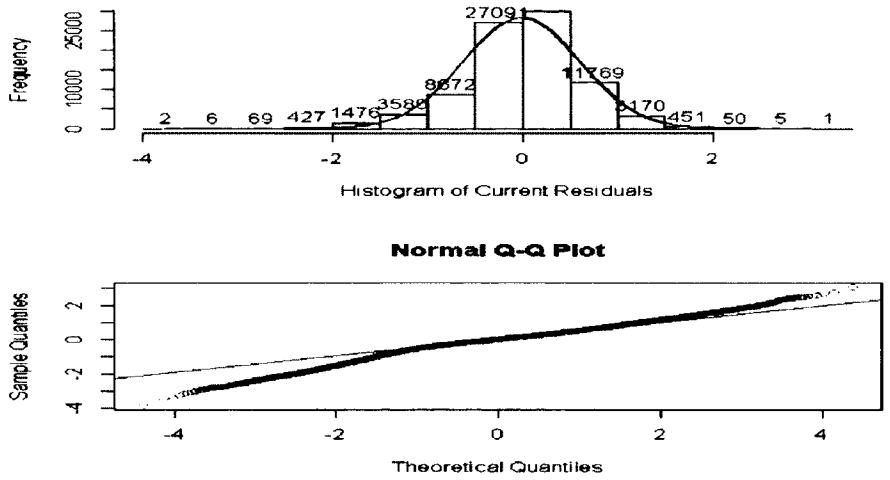


Figure 5.1: Histogram of Current Model Residuals. Normal Q-Q Plot

5.4.1 Current Model Findings

The coefficients of the covariates in the design matrix **X** for the Current model represent changes in the respond variable (log cost) that can be associated with a given predictor for fixed values of other predictors; these coefficients will be called net effects. The intercept represents the overall mean of the log cost and the log dollars for the reference categories. On the average, each additional significant IE_3 case was associated with an additional 0.1934 log dollars above the reference category (i.e the overall mean) on the log cost, if we held all other factors constant. We also found that, on the average, each additional significant CL_4 case was associated with an additional 0.2821 log dollars above the reference category on the log cost, if we held all other factors constant. We also found that, on the average, each additional significant CMG (160, 161, 166 and 171) case was associated with an additional 1.4850, 1.1246, 1.2373

and 1.1324 log dollars respectively above the reference category. Other additional significant CMG (184, 198, 205 and 208) cases were each associated with a reduced $-1.0636, -1.4250, -1.7388$ and -1.9358 log dollars respectively below the reference category on the log cost, if we held all other factors constant. Also, on average, each additional OOH (OOhCath_Y,OOhImplant_Y and OOhPtca_Y) case was associated with a reduced $-0.0751, -0.4669$ and -0.4047 log dollars respectively below the reference category on the log cost, if we held all other factors constant. Each additional FI out of the 16 FI cases was also associated with an additional range between 0.4788 to 0.7762 log dollars above the reference category on the log cost, if we held all other factors constant. For more information about the significant main effects, see Table D.1 in Appendix D on page 91. For more details on insignificant main effects, see Table D.2 in Appendix D on page 94.

Table 5.2: Significant Current Beta (Mean Estimate)

Covariate	Estimate	SE	t-value	p-value
Intercept	8.8304	0.0518	170.5322	0.0000
IE_3	0.1934	0.0326	5.9379	0.0000
CL_4	0.2821	0.0260	10.8505	0.0000
CMG_160	1.4850	0.1456	10.2009	0.0000
CMG_161	1.1246	0.0416	27.0115	0.0000
CMG_166	1.2373	0.0601	20.5991	0.0000
CMG_171	1.1324	0.2699	4.1954	0.0000
CMG_184	-1.0636	0.0704	-15.1004	0.0000
CMG_198	-1.4250	0.1010	-14.1049	0.0000
CMG_205	-1.7388	0.0897	-19.3936	0.0000
CMG_208	-1.9358	0.0890	-21.7429	0.0000
OOhCath_Y	-0.0751	0.0138	-5.4388	0.0000
OOhImplant_Y	-0.4669	0.0325	-14.3823	0.0000
OOhPtca_Y	-0.4047	0.0141	-28.6047	0.0000
fiBio_1	0.4788	0.0349	13.7367	0.0000
fiMeG96_1	0.7762	0.0346	22.4558	0.0000

The coefficients of the covariates in the heteroscedastic matrix **Z** for the Current model show that the variance of the heteroscedasdicity residuals is impacted by these covariates when they are significant. We obtain the parameter estimates of the heteroscedastic covariates using the OLS estimator. If the coefficient is negative, then each additional significant covariate (e.g CL_1) is associated with a reduced -0.2965 percent point from the overall variance. If the coefficient is positive, then each additional significant covariate (e.g OOhCath_Y and OOhImplant_Y) is associated with an additional 0.0710 percent point from the overall variance. For a detailed table on significant heteroscedastic effects, see Table D.3 in Appendix D on page 95. For a quick glance, see Table 5.3.

Table 5.3: Significant Current Gamma (Variance Estimate)

Covariate	Estimate	SE	t-value	p-value
Intercept	0.6010	0.0478	12.5635	0.0000
CL_1	-0.1080	0.0500	-2.1617	0.0306
CMG_162	-0.5365	0.0409	-13.1020	0.0000
CMG_169	-0.5259	0.1627	-3.2334	0.0012
CMG_170	-0.4979	0.1598	-3.1153	0.0018
CMG_172	-0.5647	0.0383	-14.7484	0.0000
CMG_181	-0.3906	0.0358	-10.9158	0.0000
CMG_184	-0.3401	0.0961	-3.5407	0.0004
CMG_193	-0.4328	0.1036	-4.1793	0.0000
CMG_207	-0.3802	0.1032	-3.6848	0.0002
OOhCath_Y	0.0682	0.0167	4.0861	0.0000
OOhImplant_Y	0.0710	0.0341	2.0801	0.0375
HID_51199	-0.4020	0.0716	-5.6102	0.0000
HID_51213	-0.3102	0.1018	-3.0489	0.0023
fiNut_1	0.1488	0.0698	2.1308	0.0331
fiPar_1	0.1566	0.0791	1.9789	0.0478

The following statistical measurements indicate how much of the variability of log cost is explained by the covariates and the goodness of fit in the Current model:

$R^2 = 0.6530248$, *adjusted* $R^2 = 0.6515105$, $R^2_{WLS} = 0.7723774$ and *adjusted* $R^2_{WLS} = 0.771384$. As you see, 65% was explained by pseudo R-square and 77% was explained by R^2_{WLS}, where weighted R-square is the coefficient of determination of the transformed data that measure the proportion of variation on a weighted Y that can be accounted for by weighted X. For more information about evaluation of R^2 and weighted R^2, refer to [10].

5.5 HER Model

The HER model aims to estimate the parameters and calculate the weights for the heteroscedasticity of the residuals. The model is represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.1}$$

where,

- $\mathbf{Y}$ represents an $n \times 1$ vector of the log of C, where C is the cost
- $\mathbf{X}$ represents an $n \times p$ design matrix of the covariates
- $\boldsymbol{\beta}$ represents a $p \times 1$ vector of regression parameters
- $\boldsymbol{\epsilon}$ represents an $n \times 1$ vector of the residuals with Gaussian random variables with mean $\mathbf{0}$ and variance $\mathbf{V}$ (refer to Figure 5.2 on page 52)

The following histogram and Q-Q normal plot demonstrates how the residuals follow a normal distribution.

Since the variance of the residuals are heteroscedastic (refer to Figure 5.3 on page 52), we model them as shown in equation (2.7) in chapter 2. In reference to equation (2.7) and Figure 5.4 on page 52, we see that the standardized residuals are demonstrating a homogeneous distribution (therefore, no pattern is identified).

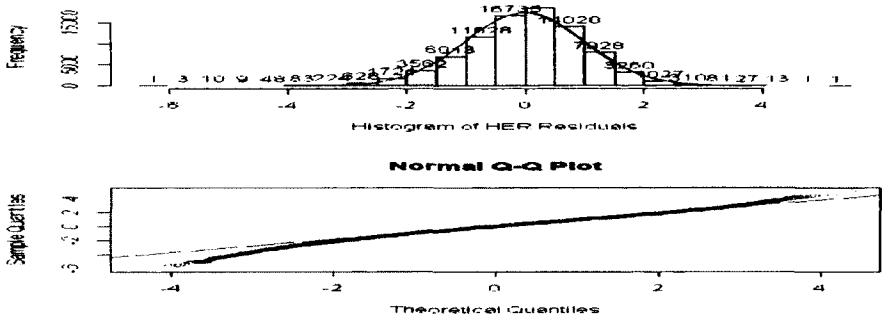


Figure 5.2: Histogram of HER Residuals with a Normal Curve. Normal Q-Q Plot

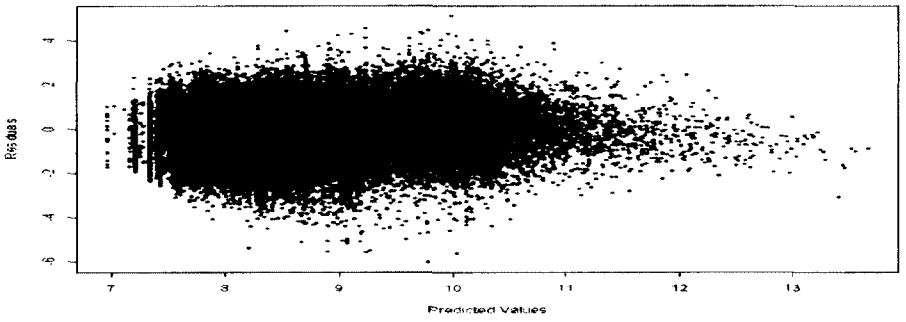


Figure 5.3: HER Residuals VS Predicted Values

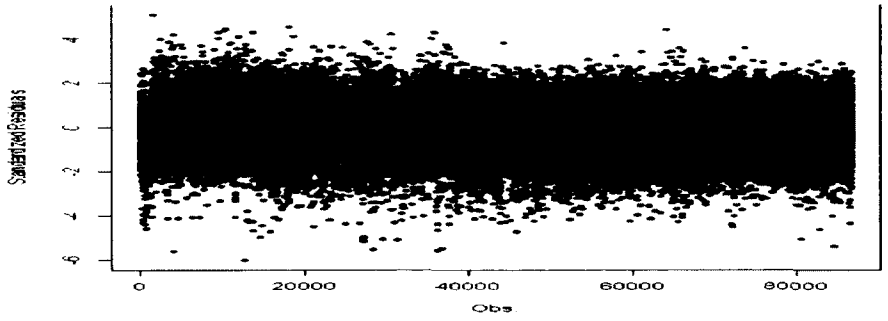


Figure 5.4: Standardized HER Residuals

5.5.1 HER Model Findings

The coefficients of the covariates in the design matrix **X** for HER model represent changes in the respond variable (log cost) that can be associated with a given predictor for fixed values of other predictors; these coefficients will be called net effects. The intercept represents the overall mean of the log cost and the log dollars for the reference categories. On the average, each additional significant IE_3 case was associated with an additional 0.2343 log dollars above the reference category (i.e the overall mean) on the log cost, if we held all other factors constant. We also found that, on the average, each additional significant CL_4 case was associated with an additional 0.267 log dollars above the reference category on the log cost, if we held all other factors constant. We also found that, on the average, each additional significant CMG (160, 161, 166 and 171) case was associated with an additional $1.5716, 1.1511, 1.2492$ and 1.0391 log dollars respectively above the reference category. Other additional significant CMG (184, 198, 205 and 208) cases were each associated with a reduced $-1.0692, -1.3632, -1.6914$ and -1.8881 log dollars respectively below the reference category on the log cost, if we held all other factors constant. Also, on average, each additional OOH (OOhCath_Y,OOhImplant_Y and OOhPtca_Y) case was associated with a reduced $-0.0632, -0.4631$ and -0.3931 log dollars respectively below the reference category on the log cost, if we held all other factors constant. Each additional FI out of the 16 FI cases was also associated with an additional range between 0.4495 to 0.7789 log dollars above the reference category on the log cost, if we held all other factors constant. For more information about the significant main effects, see Table D.5 in Appendix D on page 100. The following main effects of the mean estimate are insignificant CMGs (177 and 178), HID (51199, 51213, 51444, 51748, 51754, 51982, 53917, 53932, 53988, 80070, 80150, 90130 and 90136) and fiscal year effects at 5% significant level. For more details on insignificant main effects, see

Table D.6 in Appendix D on page 103.

Table 5.4: Significant HER Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	8.8265	0.0517	8.7252	8.9278
IE_3	0.2343	0.0366	0.1624	0.3061
CL_4	0.267	0.0305	0.2073	0.3268
CMG_160	1.5716	0.1274	1.3218	1.8213
CMG_161	1.1511	0.0405	1.0717	1.2304
CMG_166	1.2492	0.062	1.1277	1.3706
CMG_171	1.0391	0.2629	0.5238	1.5543
CMG_184	-1.0692	0.0712	-1.2088	-0.9297
CMG_198	-1.3632	0.0979	-1.555	-1.1713
CMG_205	-1.6914	0.0871	-1.8622	-1.5206
CMG_208	-1.8881	0.0865	-2.0576	-1.7186
OOhCath_Y	-0.0632	0.0139	-0.0905	-0.036
OOhImplant_Y	-0.4631	0.0324	-0.5266	-0.3997
OOhPtca_Y	-0.3931	0.0143	-0.4211	-0.3651
fiBio_1	0.4495	0.045	0.3614	0.5376
fiMeG96_1	0.7789	0.0379	0.7047	0.8531

The coefficients of the covariates in the heteroscedastic matrix **Z** for HER model show that the variance of the heteroscedasdicity residuals is impacted by these covariates when they are significant. By using the Fisher-Scoring algorithm, we obtain the parameter estimates of the heteroscedastic covariates. If the coefficient is negative, then each additional significant covariate (e.g CL_1) is associated with a reduced -0.2965 percent point from the overall variance. If the coefficient is positive, then each additional significant covariate (e.g OOhCath_Y, OOhImplant_Y and OOhPtca_Y) is associated with an additional 0.3462 percent point from the overall variance. For a detailed table on significant heteroscedastic effects, see Table D.7 in Appendix D on page 104. For a quick glance, see Table 5.5.

Table 5.5: Significant HER Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	-0.4235	0.1154	-0.6498	-0.1973
CL_1	-0.2965	0.113	-0.5179	-0.0751
CMG_162	-2.2665	0.1119	-2.4858	-2.0472
CMG_169	-2.4209	0.3774	-3.1606	-1.6811
CMG_170	-2.0549	0.2811	-2.6058	-1.5041
CMG_172	-2.4418	0.0945	-2.6271	-2.2566
CMG_181	-1.1415	0.083	-1.3043	-0.9787
CMG_184	-1.0578	0.2417	-1.5315	-0.5841
CMG_193	-1.5221	0.2059	-1.9258	-1.1185
CMG_207	-1.3714	0.2045	-1.7722	-0.9705
OOhCath_Y	0.34620	0.0422	0.2634	0.429
OOhImplant_Y	0.3375	0.0773	0.1861	0.489
OOhPtca_Y	0.4043	0.0441	0.3178	0.4907
HID_51199	-1.2636	0.131	-1.5203	-1.0068
HID_51423	0.564	0.1019	0.3642	0.7637
fiNut_1	0.3991	0.1484	0.1083	0.6899
fiPlr_1	0.2282	0.0871	0.0574	0.3989

The following statistical measurements indicate how much of the variability of log cost is explained by the covariates and the goodness of fit in the HER model: $R^2 = 0.652196$, $adjusted R^2 = 0.650678$, $R^2_{WLS} = 0.932956$ and $adjusted R^2_{WLS} = 0.9326634$. As you see, 65% was explained by pseudo R-square and 93% was explained by R^2_{WLS}, where weighted R-square is the coefficient of determination of the transformed data that measure the proportion of variation on a weighted Y that can be accounted for by weighted X. For more information about evaluation of R^2 and weighted R^2, refer to [10].

5.6 HEREM Model

The HEREM model aims to estimate the parameters and calculate the weights for the heteroscedasticity of the residuals. The model is represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{b} + \boldsymbol{\epsilon} \tag{5.2}$$

where,

- $\mathbf{Y}$ represents an $n \times 1$ vector of the log of C, where C is the cost
- $\mathbf{X}$ represents an $n \times p$ matrix of fixed covariates
- $\boldsymbol{\beta}$ represents a $p \times 1$ vector of fixed effects
- $\mathbf{W}$ represents an $n \times q$ matrix of random covariates
- $\mathbf{b}$ represents a $q \times 1$ vector of random effects
- $\boldsymbol{\epsilon}$ represents an $n \times 1$ vector of the residuals with Gaussian random variables with mean $\mathbf{0}$ and variance $\mathbf{V}$ (refer to Figure 5.5 on page 57).

The following histograms and Q-Q normal plots demonstrate how the conditional and unconditional residuals follow a normal distribution.
Since the variance of the conditional and unconditional residuals are heteroscedastic (refer to Figure 5.6 on page 57), we model it as shown in equation (2.7) in chapter 2. In reference to equation (2.7) and Figure 5.7 on page 57, we see that the conditional and unconditional standardized residuals are demonstrating a homogeneous distribution (therefore, no pattern is identified).

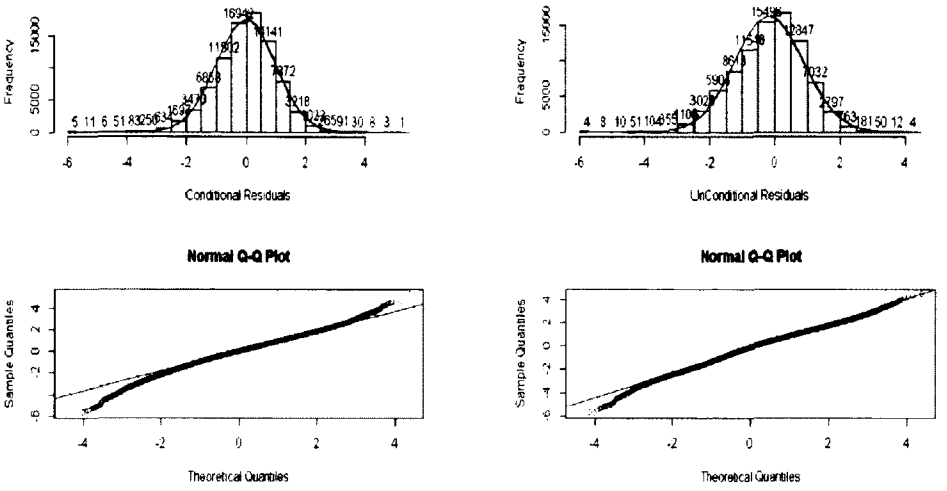


Figure 5.5: Histogram and Q-Q Plot of HEREM Conditional and UnConditional Residuals

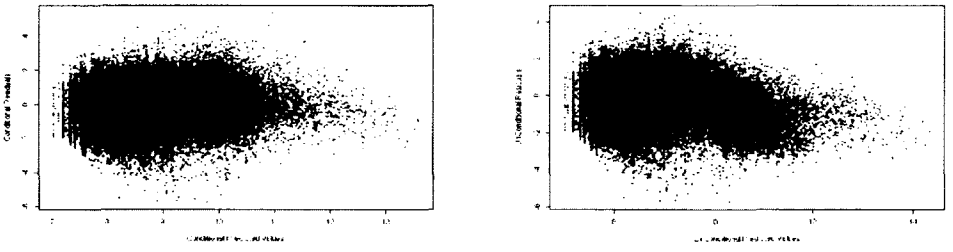


Figure 5.6: HEREM Conditional and UnConditional Residuals VS Predicted Values

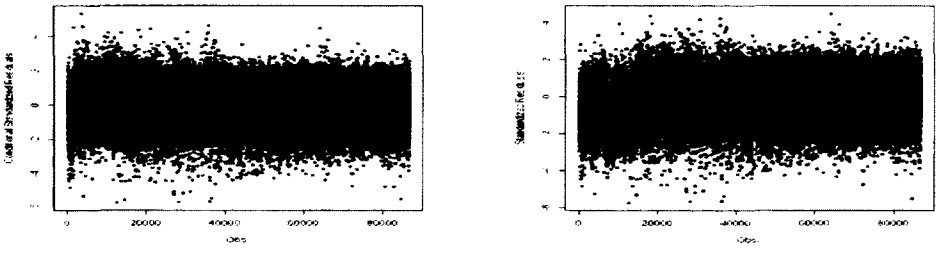


Figure 5.7: Standardized HEREM Conditional and UnConditional Residuals VS Observed Values

5.6.1 The HEREM Model Findings

The coefficients of the fixed effects in matrix **X** for the HEREM model represent the change in the respond variable (log cost) that can be associated with a given predictor for fixed values of other predictors, meaning net effects. The intercept represents the overall mean of the log cost and the log dollars for the reference categories. On the average, each additional significant IE_3 case was associated with an additional 0.2688 log dollars above the reference category (i.e the overall mean) on the log cost, if we held all other factors constant. We also found that, on the average, each additional significant CL_4 case was associated with an additional 0.2919 log dollars above the reference category on the log cost, if we held all other factors constant. On the average, each additional significant CMG (160, 161, 166 and 171) case was associated with an additional 1.327, 1.0751, 1.1259 and 0.7569 log dollars respectively above the reference category. Other additional significant CMG (184, 198, 205 and 208) cases were each associated with a reduced $-1.1488, -1.5106, -1.8498$ and -2.0491 log dollars respectively below the reference category on the log cost, if we held all other factors constant. We also found that, on the average, each additional OOH (OOhCath_Y,OOhImplant_Y and OOhPtca_Y) case was associated with a reduced $-0.0642, -0.454$ and -0.3802 log dollars respectively below the reference category on the log cost, if we held all other factors constant. Each additional FI out of the 16 FI cases was also associated with an additional range between 0.1282 to 0.6999 log dollars above the reference category on the log cost, if we held all other factors constant. For more information about the significant main effects, see Table D.9 in Appendix D on page 109. Moreover, we found that the following main effects of the mean estimate are insignificant CMGs (177 and 178), HID (51199, 51213, 51444, 51748, 51754, 51982, 53917, 53932, 53988, 80070, 80150, 90130 and 90136) and fiscal year effects at 5% significant level. For more details on insignificant main effects, see

Table D.10 in Appendix D on page 112.

Table 5.6: Significant HEREM Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	8.9042	0.0407	8.8245	8.984
IE_3	0.2688	0.0365	0.1973	0.3403
CL_4	0.2919	0.0287	0.2357	0.3481
CMG_160	1.327	0.1072	1.1168	1.5371
CMG_161	1.0751	0.0344	1.0078	1.1425
CMG_166	1.1259	0.0609	1.0066	1.2452
CMG_171	0.7569	0.1596	0.444	1.0697
CMG_184	-1.1488	0.0683	-1.2826	-1.0149
CMG_198	-1.5106	0.0766	-1.6607	-1.3605
CMG_205	-1.8498	0.0602	-1.9679	-1.7318
CMG_208	-2.0491	0.0593	-2.1653	-1.9329
OOhCath_Y	-0.0642	0.014	-0.0916	-0.0368
OOhImplant_Y	-0.454	0.0327	-0.5181	-0.3898
OOhPtca_Y	-0.3802	0.0144	-0.4085	-0.3519
fiBio_1	0.4695	0.0392	0.3927	0.5463
fiMeG96_1	0.6999	0.0388	0.6238	0.776

The coefficients of the covariates in the heteroscedastic matrix **Z** for HEREM model shows that the variance of the heteroscedasticity residuals is impacted by these covariates when they are significant. By using the Newton-Raphson (NR) algorithm, we obtain the parameter estimates of the heteroscedastic covariates. The intercept represents the overall variance of the reference categories. So, if the coefficient is negative, then each additional significant covariate (e.g CL_1) is associated with an additional 0.0441 percent point from the overall variance. If the coefficient is positive, then each additional significant covariate (e.g OOhCath_Y, OOhImplant_Y and OOhPtca_Y) is associated with an additional 0.3692 percent point from the overall variance. For a detailed table on significant heteroscedastic effects, see Table D.11 in Appendix D on page 113. For a quick glance, see Table 5.7.

Table 5.7: Significant HEREM Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	-0.4475	0.0789	-0.6021	-0.293
CL_1	0.0441	0.0155	0.0137	0.0746
CMG_162	-2.0173	0.0641	-2.143	-1.8916
CMG_169	-1.9707	0.1216	-2.209	-1.7324
CMG_170	-1.8506	0.0945	-2.0358	-1.6653
CMG_172	-2.1741	0.063	-2.2975	-2.0507
CMG_181	-1.1291	0.0643	-1.2551	-1.003
CMG_184	-0.9541	0.2119	-1.3694	-0.5387
CMG_193	-1.3359	0.0998	-1.5315	-1.1403
CMG_207	-1.2024	0.0989	-1.3962	-1.0086
OOhCath_Y	0.3506	0.0368	0.2785	0.4227
OOhImplant_Y	0.3015	0.075	0.1544	0.4486
OOhPtca_Y	0.3692	0.0402	0.2904	0.448
HID_51199	-1.1923	0.1511	-1.4884	-0.8961
HID_51423	0.3831	0.0645	0.2568	0.5095
HID_51444	0.8523	0.0595	0.7356	0.969
HID_51994	0.4379	0.0732	0.2946	0.5813

The table below (Table 5.8) shows the variance component for each group of the following interactions:

Table 5.8: Variance Component Estimate

Covariate	Estimate	SE	95% LB	95% UB
CmL*CMG	0.025644452	0.1749	-4.0061	-3.3207
IE*CMG	0.039439652	0.3128	-3.846	-2.6199
FI*CMG	0.046853088	0.1519	-3.3585	-2.763
HID*FsYr	0.003915397	0.3975	-6.3218	-4.7638

The following statistical measurements indicate how much of the variability of log cost is explained by the covariates and the goodness of fit in the HEREM model: $R^2 = 0.6073918$, *adjusted* $R^2 = 0.6059136$, $R^2_{WLS} = 0.9300158$ and *adjusted* $R^2_{WLS} = 0.9297523$. As you see, 60% was explained by pseudo R-square and 93% was explained by R^2_{WLS}, where weighted R-square is the coefficient of determination of the

transformed data that measure the proportion of variation on a weighted Y that can be accounted for by weighted X. For more information about evaluation of R^2 and weighted R^2, refer to [10].

5.7 Generalized Linear Model

Using McCullagh and Nelders (refer to [11]) point that it is common for data in the form of continuous measurements to have variance positively correlated with the mean, we fitted the cost using the generalized linear model since the patient cost Y is a continuous positive random variable and is skewed to the right. In this instance then, we can say that the coefficient of variation (mean/standard deviation) (CV) is a more realistic assumption than constant variance. Then Y is a gamma distribution with a shape parameter α and a rate parameter ν, where the inverse of the rate parameter is called a scale parameter, and the inverse of the shape parameter is called the dispersion ($\phi = CV^2$). The mean parameter $\mu = \frac{\alpha}{\nu}$ and using this parametrization, the pdf is as follows:

$$f(y) = \frac{\nu^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\nu y} \tag{5.3}$$

where $\alpha > 0, \nu > 0, y > 0$ and $Y \sim Gamma(\alpha, \nu)$.

Gamma Model:

Here $Y_1, \ldots, Y_n$ are assumed independent and $Y_i \sim Gamma(\frac{\mu_i}{\alpha}, \alpha)$ where $\mu_i = e^{\eta_i}$ with $\eta_i = \mathbf{x_i'}\boldsymbol{\beta}$ the linear predictor for the i^{th} patient, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ a vector of regression coefficients and $\mathbf{x_i'} = (x_{i1}, \ldots, x_{ip})$ is a vector of covariates for the i^{th} patient.

5.7.1 The GLM Findings

The coefficients of the covariates in the design matrix **X** for GLM represent changes in the expected value for the respond variable (cost) that can be associated with a given predictor for fixed values of other predictors. The intercept represents the overall mean of the expected cost and the log dollars for the reference categories. On the average, each additional significant IE_3 case was associated with an additional 0.2090 log dollars above the reference category (i.e the overall mean) on the expected value for cost, if we held all other factors constant. Also, on average, each additional significant CL_4 case was associated with an additional 0.2793 log dollars above the reference category on the expected value for cost, if we held all other factors constant. We also found that, on the average, each additional significant CMG (160, 161, 166, and 171) case was associated with an additional $1.3851, 0.9965, 1.0956$, and 0.8432 log dollars respectively above the reference category. Other additional significant CMG (184, 198, 205 and 208) cases were each associated with a reduced $-1.2248, -1.5559, -1.9390$ and -2.1496 log dollars respectively below the reference category on the expected value for cost, if we held all other factors constant. We also found that, on the average, each additional OOH (OOhCath_Y,OOhImplant_Y and OOhPtca_Y) case was associated with a reduced $-0.0329, -0.3617$ and -0.3624 log dollars respectively below the reference category on the expected value for cost, if we held all other factors constant. On average, each additional FI out of the 16 FI cases was associated with an additional range between 0.1513 to 0.9063 log dollars above the reference category on the expected value for cost, if we held all other factors constant. For more information about the significant main effects, see Table D.13 in Appendix D on page 118.

Table 5.9: Significant GLM Beta

Covariate	Estimate	SE	t-value	p-value
Intercept	9.1387	0.0445	205.3137	0.0000
IE_3	0.2090	0.0549	3.8091	0.0001
CL_4	0.2793	0.0373	7.4892	0.0000
CMG_160	1.3851	0.1854	7.4719	0.0000
CMG_161	0.9965	0.0346	28.7682	0.0000
CMG_166	1.0956	0.1092	10.0338	0.0000
CMG_171	0.8432	0.4149	2.0323	0.0421
CMG_184	-1.2248	0.0894	-13.7029	0.0000
CMG_198	-1.5559	0.1028	-15.1343	0.0000
CMG_205	-1.9390	0.0957	-20.2565	0.0000
CMG_208	-2.1496	0.0953	-22.5589	0.0000
OOhCath_Y	-0.0329	0.0155	-2.1153	0.0344
OOhImplant_Y	-0.3617	0.0318	-11.3901	0.0000
OOhPtca_Y	-0.3624	0.0170	-21.2536	0.0000
fiBio_1	0.4369	0.0519	8.4150	0.0000
fiMeG96_1	0.9063	0.0437	20.7521	0.0000

Moreover, the following table (Table 5.10) identifies the insignificant mean estimates for the main effects.

Table 5.10: Insignificant GLM Beta

Covariate	Estimate	SE	t-value	p-value
CMG_175	0.0366	0.0305	1.2011	0.2297
CMG_178	-0.0247	0.0368	-0.6710	0.5022
CMG_182	-0.0137	0.0338	-0.4042	0.6861
HID_51444	-0.0153	0.0357	-0.4286	0.6682
HID_53850	0.0490	0.0353	1.3881	0.1651
HID_53910	0.0381	0.0350	1.0886	0.2763
HID_53917	0.0050	0.0352	0.1419	0.8872
HID_53932	-0.0659	0.0378	-1.7447	0.0810
HID_53994	0.0946	0.0490	1.9301	0.0536
HID_80150	-0.0961	0.0782	-1.2280	0.2194
HID_90130	0.0761	0.0391	1.9453	0.0517
HID_90136	-0.0851	0.0441	-1.9270	0.0540

The following statistical measurements indicate the goodness of fit in the GLM model: *null deviance* = 2043804616, *deviance* = 29473.26, $R^2_{WLS} = 0.9999856$ and

adjusted $R^2_{WLS} = 0.9999855$. As you see, 99.99% was explained by R^2_{WLS}, where weighted R-square is the coefficient of determination of the transformed data that measure the proportion of deviance $1 - \frac{deviance}{null\ deviance}$.

5.8 Log-normal vs. Gamma with Log-link

Normally, we analyze Y_i with a constant CV by obtaining the log of Y_i as $Z_i = \log(Y_i)$. This transformation stabilizes the variance (see Sections 5.5 and 5.6). We assume that the observations are now normalized and then analyze Z_i with a linear normal model. If the Z_i are normally distributed, then we can say that the Y_i are log-normally distributed. The CV is constant in both a log-normal distribution and a gamma distribution. The log-normal distribution differs from the gamma distribution in that it cannot be modelled as a generalized linear model.
We use both approaches to analyze the cost data.

- Log-transforming the cost Y where the variance is proportional to its mean squared as $(e^{\sigma_i} - 1)\mu_i^2$:
  - systematic: $\mathrm{E}(\log(Y_i)) = \tilde{\eta}_i = \mathbf{x}_i'\boldsymbol{\beta}$
  - random: $\log(Y_i) \sim N(\tilde{\eta}_i, \sigma_i)$
- Gamma GLM for cost Y with a log-link where the variance is also proportional to its mean squared as $\frac{\mu_i^2}{\alpha}$:
  - systematic: $\log(\mathrm{E}(Y_i)) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$
  - random: $Y_i \sim Gamma(\frac{\mu_i}{\alpha}, \alpha)$

In the gamma model, the mean $\mathrm{E}(Y_i) = \mu_i$ of the cost for patient i is estimated by:

$$\hat{\mu}_i = \exp(\hat{\eta}_i) \tag{5.4}$$

but for the log-normal approach one has to also use the variance to get an unbiased estimate

$$\hat{\mu}_i = \exp(\hat{\hat{\eta}}_i + \frac{\sigma_i^2}{2}) \tag{5.5}$$

where $\frac{\sigma_i^2}{2}$ is called a volatility adjustment factor.

The table below shows that both distributions produce virtually a small difference in results.

Table 5.11: Comparison

	Log-normal						Gamma	
	Current		HER		HEREM		GLM	
Covariates	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	8.8304	0.0518	8.8265	0.0517	8.9042	0.0407	9.1387	0.0445
IE_3	0.1934	0.0326	0.2343	0.0366	0.2688	0.0365	0.2090	0.0549
CL_4	0.2821	0.0260	0.267	0.0305	0.2919	0.0287	0.2793	0.0373
CMG_160	1.4850	0.1456	1.5716	0.1274	1.327	0.1072	1.3851	0.1854
CMG_161	1.1246	0.0416	1.1511	0.0405	1.0751	0.0344	0.9965	0.0346
CMG_162	1.0078	0.0377	1.0155	0.0384	0.947	0.0324	0.8636	0.0381
CMG_166	1.2373	0.0601	1.2492	0.062	1.1259	0.0609	1.0956	0.1092
CMG_167	1.0514	0.1182	1.0092	0.1407	0.8505	0.1301	0.7946	0.3097
CMG_168	1.1001	0.0440	1.0672	0.0543	0.9958	0.0465	0.8764	0.0761
CMG_171	1.1324	0.2699	1.0391	0.2629	0.7569	0.1596	0.8432	0.4149
CMG_184	-1.0636	0.0704	-1.0692	0.0712	-1.1488	0.0683	-1.2248	0.0894
CMG_194	-1.1517	0.0896	-1.1027	0.087	-1.2674	0.0601	-1.3629	0.0958
CMG_196	-1.1876	0.0890	-1.1412	0.0865	-1.3025	0.0593	-1.3632	0.0952
CMG_197	-1.2719	0.1245	-1.2285	0.1224	-1.3995	0.093	-1.4245	0.1166
CMG_198	-1.4250	0.1010	-1.3632	0.0979	-1.5106	0.0766	-1.5559	0.1028
CMG_199	-1.0842	0.0977	-1.0531	0.0948	-1.2042	0.0701	-1.2159	0.0995
CMG_205	-1.7388	0.0897	-1.6914	0.0871	-1.8498	0.0602	-1.9390	0.0957
CMG_208	-1.9358	0.0890	-1.8881	0.0865	-2.0491	0.0593	-2.1496	0.0953
CMG_213	-1.4640	0.0944	-1.4162	0.0921	-1.5706	0.0666	-1.5162	0.0976
OOhCath_Y	-0.0751	0.0138	-0.0632	0.0139	-0.0642	0.014	-0.0329	0.0155
OOhImplant_Y	-0.4669	0.0325	-0.4631	0.0324	-0.454	0.0327	-0.3617	0.0318
OOhPtca_Y	-0.4047	0.0141	-0.3931	0.0143	-0.3802	0.0144	-0.3624	0.0170
fiBio_1	0.4788	0.0349	0.4495	0.045	0.4695	0.0392	0.4369	0.0519
fiMeG96_1	0.7762	0.0346	0.7789	0.0379	0.6999	0.0388	0.9063	0.0437

5.9 Summary

Presented is a brief overview of the number of significant and insignificant covariates for each of the models discussed. Table 5.12 will show the number of these covariates.

Table 5.12: Number of Significant and Insignificant Covariates

Covariate	Current	HER	HEREM	GLM
Significant	218	197	103	185
Insignificant	159	180	14	192

Table 5.13 compares the HER, HEREM and GLM models to the Current model and identifies the number of common significant and insignificant covariates among these models. It shows that the significant covariates in one model could be insignificant in another. Note that the HEREM model differs in that we only compared the fixed effects to the covariates of the Current, HER and GLM models.

Table 5.13: Comparing HER, HEREM and GLM Models to the Current Model

		Current	
		Significant	Insignificant
HER	Significant	185	12
	Insignificant	33	147
HEREM	Significant	98	5
	Insignificant	3	11
GLM	Significant	156	29
	Insignificant	62	130

Here are the 12 covariates that are significant in HER but not in the Current model (CL_2, HID_53936, CL_CMG_2_168, IE_CMG_2_178, IE_CMG_3_180, CMG_fiCel_163_1, CMG_fiDia_180_1, CMG_fiDia_182_1, CMG_fiDia_209_1, CMG_fiEnd_208_1, CMG_fiVad_163_1 and HID_FsYr_53932_2008).

Here are the 33 covariates that are significant in the Current but not in the HER model (HID_51748, HID_51982, CL_CMG_1_199, CL_CMG_1_205, CL_2 interact with CMG (166, 167, 194, 196, 200, 209), CL_3 interact with CMG (162, 193, 196, 200), IE_2 interact with CMG (164,170) , IE_3 interact with CMG (162, 172, 174,181), fiCel_1 interact with CMG (162, 166_1, 167, 169, 171, 172) CMG_fiDia_175_1, CMG_fiEnd_202_1, CMG_fiMeG96_172_1, CMG_fiMeL96_164_1, CMG_fiMeL96_171_1, CMG_fiPlr_162_1 and CMG_fiPlr_202_1). Here are the 5 covariates that are significant in HEREM but not in the Current model (CL_2, CMG_178, HID_53936, HID_80070 and FsYr_2008).

Here are the 3 covariates that are significant in the Current but not in the HEREM model (CMG_182, HID_51748 and HID_51982).

Here are the 29 covariates that are significant in the GLM but not in the Current model (CL_2, CMG_177, HID (51199, 51213, 51754, 53936, 53988, 80070), FsYr_2008, CL_CMG_1_202, IE_2 interact with CMG (161, 162, 165, 169, 172, 174 and 181), CMG_fiBio_200_1, CMG_fiCel_163_1, CMG_fiDia_180_1, CMG_fiMeG96_163_1, CMG_fiPar_196_1, CMG_fiVad_161_1, CMG_fiVad_163_1, CMG_fiVad_180_1, CMG_fiVad_185_1, HID_FsYr (51444_2008,53917_2008 and 90109_2008)).

Here are the 62 covariates that are significant in the Current but not in the GLM model (CMG (175, 182), HID (53850, 53910, 53994), CL_1 interact with CMG (164, 182, 193, 194, 195, 197, 199, 205, 207), CL_2 interact in CMG (166, 167, 175, 181, 194, 196, 200, 209), CL_3 interact with CMG (162, 193, 196, 200), CL_CMG_4_175, CL_CMG_4_194, IE_2 interact with CMG (164, 176, 180, 182, 185), IE_3 interact with CMG (162, 172, 174, 181, 182), CMG_fiCel (162_1, 164_1, 166_1, 167_1, 169_1, 171_1, 172_1), CMG_fiCrd_196_1, CMG_fiDia (162_1, 172_1, 175_1 193_1), CMG_fiEnd_202_1, CMG_fiMeG96_175_1, CMG_fiMeL96 (167_1, 171_1, 194_1), CMG_fiPlr (196_1, 200_1, 202_1), CMG_fiTub_163_1, CMG_fiVad (175_1, 194_1 and 209_1)).

Table 5.14 compares the HEREM and GLM models to the HER model and identifies the number of common significant and insignificant covariates among these models. It shows that the significant covariates in one model could be insignificant in another.

Table 5.14: Comparing HEREM and GLM Models to the HER Model

		HER	
		Significant	Insignificant
HEREM	Significant	100	3
	Insignificant	1	13
GLM	Significant	155	30
	Insignificant	42	150

Here are the 3 covariates that are significant in HEREM but not in the HER model (CMG_178, HID_80070 and FsYr_2008). One covariate (CMG_182) is significant in the HER but not in the HEREM model.

Here are the 30 covariates that are significant in the GLM but not in the HER model (CMG_177, HID (51199, 51213, 51748, 51754, 51982, 53988, 80070), FsYr_2008, CL_CMG_1_202, IE_2 interact with CMG (161, 162, 165, 169, 170, 172, 174, 181), CMG_fiBio_200_1, CMG_fiMeG96 (163_1, 172_1), CMG_fiMeL96_164_1, CMG_fiPar_196_1, CMG_fiPlr_162_1, CMG_fiVad (161_1, 180_1, 185_1), HID_FsYr (51444_2008, 53917_2008, and 90109_2008)).

Here are the 42 covariates that are significant in HER but not in the GLM model (CMG (175, 182), HID (53850, 53910, 53994), CL_1 interact with CMG (164, 182, 193, 194, 195, 197, 207), CL_2 interact with CMG (168, 175, 181), CL_4 interact with CMG (175, 194), IE_2 interact with CMG (176, 178, 180, 182, 185), IE_3 interact with CMG (180, 182), CMG_fiCel_164_1, CMG_fiCrd_196_1, CMG_fiDia (162_1, 172_1, 182_1, 193_1, 209_1, CMG_fiEnd_208_1, CMG_fiMeG96_175_1, CMG_fiMeL96 (167_1, 194_1), CMG_fiPlr(196_1, 200_1), CMG_fiTub_163_1, CMG_fiVad (175_1, 194_1, 209_1) and HID_FsYr_53932_2008).

Table 5.15 compares the HEREM to the GLM model and identifies the number of common significant and insignificant covariates among these models. It shows that the significant covariates in one model could be insignificant in another.

Table 5.15: Comparing HEREM Model to the GLM Model

		HEREM	
		Significant	Insignificant
GLM	Significant	98	7
	Insignificant	5	7

Here are the 7 covariates that are significant in the GLM but not in the HEREM model (CMG_177, HID (51199, 51213, 51748, 51754, 51982 and 53988).
Here are the 5 covariates that are significant in the HEREM but not in the GLM model (CMG (175, 178), HID (53850, 53910 and 53994)).
Table 5.16 below shows the CL coefficients of value to CIHI regarding the CL methodology.

Table 5.16: Coefficients for Comorbidity Level

	Current		HER		HEREM		GLM	
Covariates	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
CL_1	0.6621	0.0527	0.597	0.0518	0.4156	0.0239	0.5299	0.0465
CL_2	0.0647*	0.0395*	0.1654	0.0375	0.1645	0.0254	0.1443	0.0460
CL_3	0.1239	0.0369	0.1496	0.0374	0.1645	0.0268	0.1631	0.0458
CL_4	0.2821	0.0260	0.267	0.0305	0.2919	0.0287	0.2793	0.0373

**The covariate is insignificant*

CIHI classifies comorbidity levels into five levels. The current model shows that CL_2 is insignificant (see Table 5.13) as opposed to CL_1, which is significant. This implies that CIHI could change the CL methodology to only four levels. However HER,

HEREM and GLM show that the five levels are all significant. *This observation is of value to CIHI as it may prompt the organisation to modify its CL methodology.*

The following tables present the descriptive statistics about the parameter estimates and their standard errors for the above four models.

Table 5.17: Descriptive Statistics for Significant Beta's in each Model

Model	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Current	-1.9358	-0.3952	10.7662	-0.1425	-0.0612	0.8416	0.6789	0.2837	8.8304
HER	-1.8881	-0.3808	10.7146	-0.1261	-0.047	0.8711	0.703	0.3222	8.8265
HEREM	-2.0491	-0.4286	10.9533	0.1588	-0.0336	1.1844	0.841	0.4125	8.9042
GLM	-2.1496	-0.4146	11.2883	-0.2141	-0.1422	0.9323	0.6292	0.2146	9.1387

Table 5.17 shows that the averages for the parameter estimates are slightly different.

Table 5.18: Descriptive Statistics of SE for Significant Beta's in each Model

Model	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Current	0.013	0.0436	0.2569	0.0621	0.0696	0.0339	0.0461	0.0896	0.2699
HER	0.0139	0.044	0.249	0.0636	0.0699	0.032	0.0432	0.0872	0.2629
HEREM	0.014	0.0316	0.1456	0.0394	0.0505	0.0279	0.0296	0.0612	0.1596
GLM	0.0155	0.046	0.3993	0.072	0.0791	0.0469	0.0511	0.0971	0.4149

Table 5.18 shows that the averages for SE of the parameter estimates of HER, HEREM and GLM show a difference from the Current by 0.0003, -0.0191 and 0.0095 respectively.

The following tables present the descriptive statistics for the difference in common significant parameter estimates and the difference in their standard errors in each paired model below.

Table 5.19: Descriptive Statistics for the difference among the pairs of models for the Common Significant Beta's

Model	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Current/HER	-0.1763	-0.0485	0.3599	-0.0141	-0.0156	0.0523	0.0552	0.0067	0.1836
Current/HEREM	-0.0916	-0.0127	0.4671	0.0297	0.0457	0.0765	0.1200	0.1072	0.3755
Current/GLM	-0.3082	-0.0452	0.6115	0.0523	0.0454	0.1292	0.1776	0.1324	0.3033
HER/HEREM	-0.0777	-0.0014	0.3599	0.0151	0.0554	0.0780	0.1403	0.1389	0.2822
HER/GLM	-0.3122	-0.0433	0.6701	0.0671	0.0614	0.1284	0.1984	0.1551	0.3579
HEREM/GLM	-0.1992	-0.0826	0.4337	-0.0442	-0.0184	0.0956	0.1389	0.0563	0.2345

Table 5.19 shows that the averages of the difference of the parameter estimates are different.

Table 5.20: Descriptive Statistics for the difference in SE among the pairs of models for Common Significant Beta's

Model	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Current/HER	-0.0298	-0.0024	0.0637	0.0000	-0.0009	0.0067	0.0042	0.0018	0.0339
Current/HEREM	-0.0119	0.0037	0.1222	0.0080	0.0125	0.0159	0.0223	0.0260	0.1103
Current/GLM	-0.1915	-0.0190	0.2446	-0.0057	-0.0114	0.0282	0.0212	0.0022	0.0531
HER/HEREM	-0.0022	0.0058	0.1055	0.0083	0.0119	0.0133	0.0147	0.0205	0.1033
HER/GLM	-0.1690	-0.0174	0.2198	-0.0050	-0.0112	0.0277	0.0195	0.0021	0.0508
HEREM/GLM	-0.0469	0.0044	0.3022	0.0084	0.0185	0.0354	0.0252	0.0296	0.2553

Table 5.20 shows that the averages for the difference in SE of the parameter estimates of HER, HEREM and GLM are different.

Chapter 6

Simulation

6.1 Introduction

In statistics, a simulation is a numerical technique used to evaluate models under various scenarios in order to ascertain their performance. Since exact analysis is often impossible, using simulations will give us confidence in our statistical methods and will identify whether or not the models under consideration have desirable properties under certain conditions (small sample size, low/high variance, etc.) Here, we will investigate the performance of the models considered in this thesis. Statistical techniques are often used in order to predict the cost of the health care system. This is done by modelling and analyzing cost data. The cost data includes 86,759 observations, the intercept and 218 dummy variables from the factors within the CMG+ methodlgy, which include the main effects and the interactions of the factors (between: HID and Fiscal Year, IE and CMG, CL and CMG, and FI and CMG). Since the relationship between the cost and these factors is multiplicative, we proposed the four models (Current, HER, HEREM, and GLM), where we assumed that the first three models follow a log-normal distribution while the GLM follows a gamma distribution with a log link. For more details, see chapter 5 on page 45.

6.2 Simulation Procedure

To conduct our simulation, we sampled the rows using the bootstrap method. Each row represents a patient to which a cost and effect is identified; we assume that these rows are independent. This procedure is repeated 1000 times to generate new datasets. For each of these new datasets, we choose a random 18% of it for the Test Data (15617 observations) and the remainder of the 86579 observations are grouped into the Train Data (71142 observations). This was done to validate the prediction on the Test Data after we estimated the parameter coefficients from the Train Data by fitting these four models using the effects as shown in table 6.1.

Table 6.1: Covariates Used in the Models

Model	Fixed Effects	Random Effects	Heteroscedastic Effects
Current	IE, CL, CMG, OOH, HID, FI, FsYr,CL*CMG, FI*CMG, IE*CMG, HID*FsYr		IE, CL, CMG, OOH, HID, FI, FsYr
HER	IE, CL, CMG, OOH, HID, FI, FsYr,CL*CMG, FI*CMG, IE*CMG, HID*FsYr		IE, CL, CMG, OOH, HID, FI, FsYr
HEREM	IE, CL, CMG, OOH, HID, FI, FsYr	CL*CMG, FI*CMG, IE*CMG, HID*FsYr	IE, CL, CMG, OOH, HID, FI, FsYr
GLM	IE, CL, CMG, OOH, HID, FI, FsYr, CL*CMG, FI*CMG, IE*CMG, HID*FsYr		

In each regression, we calculate the mean and heteroscedastic coefficients of the dummy variables created by the factors above and the EBLUPS for the random effects in the HEREM model. We also calculate the prediction cost value or log cost value on the Train and Test Data using the models discussed in chapters 2 and 3. At each iteration, we calculate the sum square error (SSE) and the mean square error (MSE) for the Test and the Train data and we calculate the R^2 and R^2_{adj} for

the Train data only. We will briefly cover some common methods for evaluating the performance of a simulation; however, we only use the SSE, MSE, the R^2 and the R^2_{adj}.

6.3 Stored Estimates for Each Simulation and Summary Measures Calculated Over all Simulations

Storing estimates after each simulation is necessary as it prevents the need to repeat simulations in the event errors occur or results are required. This storage acts as a backup of the estimates being generated and permits the researcher to review for consistency, to identify errors or outlying values and to define any pattern within each simulation. The average of all simulation estimates, such that $\bar{\hat{\beta}} = \sum_{i=1}^{B} \hat{\beta}_i / B$, where B is the number of simulations, is taken as a measure of the true estimate of the interest. For every simulation, we obtain a SE for the estimate of interest, then the average for all the SEs, obtained by $\sum_{i=1}^{B} SE(\hat{\beta}_i)/B$, could be used as an estimated standard deviation. The empirical SE, defined as $\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}})^2}$, should be close to the average SE if the estimates are unbiased.

6.4 Performance Evaluation Criteria for Statistical Methods

To determine the validity of a study's results, one must conduct an evaluation of the simulation performance. It is highly recommended and advisable, however, to assess this validity by using more than one performance indicator or evaluation criteria. Table 6.2 will outline some of the most common evaluation criteria. For our study,

we will focus our evaluation criteria on the Bias, the Mean Square Error, and the Coverage rate of the Confidence Interval.

Table 6.2: Evaluation Criteria for Different Methods

Evaluation Criteria	Formula
Bias	$\bar{\hat{\beta}} - \beta$
Percentage bias	$\left(\frac{\bar{\hat{\beta}}-\beta}{\beta}\right) \times 100$
Standardised bias	$\left(\frac{\bar{\hat{\beta}}-\beta}{\mathrm{SE}(\hat{\beta})}\right) \times 100$
MSE (Accuracy)	$\left(\bar{\hat{\beta}} - \beta\right)^2 + \left(\mathrm{SE}(\hat{\beta})\right)^2$
Confidence Interval	$\hat{\beta}_i \pm z_{\alpha/2}\mathrm{SE}(\hat{\beta}_i)$
Average length of Confidence Interval	$\frac{\sum_{i=1}^{B} 2z_{\alpha/2}\mathrm{SE}(\hat{\beta}_i)}{B}$

6.4.1 Assessment of Bias

The bias is a performance indicator that measures the deviation in an estimate from the true value. Various biases may indicate an overestimation or an underestimation of the true value. If the magnitude of the bias is small, it may go unnoticed and may prove to be insignificant in comparison to the main objective of the study; however, if the magnitude of the bias is significant, it may alter results. For this reason, we must ensure that the bias is as close to zero as possible. If the estimator is biased, then the effect of this bias on the estimator distribution must be considered. There are several approaches to calculating the bias which result in: the Bias, the Percentage Bias and the Standardised Bias. We will be using the Bias approach as outlined in Table 6.2. The resulting bias between $\frac{1}{2}\mathrm{SE}(\hat{\beta})$ to $2\mathrm{SE}(\hat{\beta})$ must be avoided as it may alter results in a study. [12]

6.4.2 Assessment of Mean Square Error

The MSE is a measure of the accuracy of the simulation. It incorporates the bias and variability measures and can be transformed to the same scale as the parameter by taking its square root. The MSE of the estimator is a measurement of the difference between the values obtained through an estimator and the true value of the quantity being estimated.

6.4.3 Assessment of Confidence Interval Coverage

The coverage of a confidence interval (CI) is a performance measure of the success of the CI in containing the true specified parameter value. In order to control the type I error rate, the coverage should be close to the same value as the nominal coverage (also the average coverage); this is often set to 95 CI. There are two issues that must be avoided in order to consider a simulation successful: over-coverage and under-coverage. When the coverage rate is over 95, over-coverage occurs and this leads to too many type II errors. When the coverage rate is less than 95, under-coverage occurs and this leads to too many type I errors. Over-coverage results in a loss of statistical power and unreliable results, and under-coverage is unacceptable since more simulations will highlight the error rate. An acceptable coverage rate is one that falls within approximately two SEs of the average coverage rate (p), where $\mathrm{SE}(p) = \sqrt{p(1-p)/\mathrm{B}}$. Most studies use the average length of the CI for the parameter estimate as an evaluation tool. More precise estimates are characterized by narrower CI. The narrower the CI, the more unbiased a parameter estimator is considered.

6.5 Results

The following tables will outline the simulation results for the four models: Current, HER, HEREM and GLM.

Table 6.3: Current Model Simulation Results

Statistics	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Train SSE	26361.83	26873.56	1243.052	26998.21	27000.18	197.2996	255.328	27128.89	27604.88
Train MSE	0.3717	0.3789	0.0175	0.3807	0.3807	0.0028	0.0036	0.3825	0.3892
Train R^2	0.6425	0.6485	0.0135	0.65	0.6499	0.0022	0.0029	0.6514	0.656
Train R^2_{adj}	0.6414	0.6474	0.0135	0.6489	0.6488	0.0022	0.0029	0.6503	0.6549
Test SSE	5627.773	5899.994	604.3249	5955.671	5957.581	86.3423	112.0277	6012.022	6232.098
Test MSE	0.3655	0.3832	0.0392	0.3868	0.3869	0.0056	0.0073	0.3905	0.4048

Table 6.4: HER Model Simulation Results

Statistics	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Train SSE	26383.38	26901.95	1217.685	27019.03	27024.39	196.6839	252.3413	27154.3	27601.06
Train MSE	0.372	0.3793	0.0172	0.381	0.381	0.0028	0.0036	0.3829	0.3892
Train R^2	0.6424	0.6481	0.0132	0.6496	0.6496	0.0022	0.0029	0.651	0.6556
Train R^2_{adj}	0.6413	0.647	0.0132	0.6486	0.6485	0.0022	0.0029	0.6499	0.6546
Test SSE	5631.882	5905.19	616.6319	5959.332	5962.534	87.1533	112.7209	6017.911	6248.514
Test MSE	0.3658	0.3835	0.04	0.387	0.3873	0.0057	0.0073	0.3908	0.4058

Table 6.5: HEREM Model Simulation Results

Statistics	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Train SSE	26393.31	26912.5	1227.08	27030.46	27037.1	196.8913	255.2562	27167.76	27620.39
Train MSE	0.311	0.3794	0.0843	0.3811	0.3811	0.0036	0.0036	0.383	0.3953
Train R^2	0.6422	0.648	0.0133	0.6495	0.6494	0.0022	0.0029	0.6508	0.6555
Train R^2_{adj}	0.6339	0.6469	0.0781	0.6484	0.6484	0.0031	0.0029	0.6498	0.7121
Test SSE	5633.341	5906.516	611.6918	5960.262	5963.669	87.0872	112.3804	6018.896	6245.033
Test MSE	0.1889	0.3835	0.2902	0.387	0.3871	0.0095	0.0075	0.391	0.4791

Table 6.6: GLM Model Simulation Results

Statistics	Min	1st Qu	Range	Median	Mean	Std.Err	IQR	3rd Qu	Max
Trrain SSE	58545073	60422719	4953270	61010206	60999031	847357.6614	1098569.752	61521289	63498343
Train MSE	825.4854	851.9602	69.8411	860.2437	860.0862	11.947741	15.48983	867.45	895.3265
Train R^2	0.999985	0.999985	0.000001	0.999985	0.999985	0.000001	0.000001	0.999986	0.999986
Train R^2_{adj}	0.999985	0.999985	0.000001	0.999985	0.999985	0.000001	0.000001	0.999986	0.999986
Test SSE	1.25E+12	2.35E+12	1.14E+13	3.02E+12	3.27E+12	1.34762E+12	1.48879E+12	3.83E+12	1.26E+13
Test MSE	81424252	1.52E+08	7.38E+08	1.96E+08	2.12E+08	87524973.66	96693226.6	2.49E+08	8.19E+08

Chapter 7

Conclusion

7.1 Introduction

The Canadian Institute for Health Information conducts applied research and analysis that helps policy developers make informed decisions about health care spending and resource allocation. My work on assessing the various models to identify one that can predict health care costs more efficiently than the current model is only the beginning of a process to bring more precision and accuracy to the estimation of health care costs.

7.2 Overview

Chapter 1 of this thesis covered the work conducted at CIHI and its relevance in Canada. Chapter 2 explained the Current model at CIHI and introduced the HER model. Chapter 3 and Chapter 4 detailed the LMM and the GLM respectively. Chapter 5 briefly discussed how data is organised at CIHI and highlighted the RIW as a measure for costs. Chapter 5 also presented the research findings, covering all four models: the Current, HER, HEREM and GLM, while drawing a comparison between them all. Chapter 6 outlined the simulation procedure followed and the covariates

used in all four models while explaining the criteria for assessing the performance of a simulation and presenting the results of the simulations conducted for this research.

7.3 Recommendation

This thesis aimed to identify a more efficient costing model. We modified the current model to the HER model and thought that by assuming the interaction covariates are random effects, that follows a normal distribution and that is fitted using the HEREM model, we could borrow strength for low volumes. Unfortunately, using the random effect was not as successful as we had hoped. For future research and work in this field, it is suggested that a gamma model at the CMG level is expected to yield more favourable results.

List of References

[1] CIHI. *Case Mix Decision-Support Guide: CMG+.* Canadian Institute for Health Information. Ottawa, Ont. (2009).

[2] CIHI. *2012 CMG+ Directory.* Canadian Institute for Health Information. Ottawa, Ont. (2012).

[3] CIHI. *Strategic Directions, 2008-2009 to 2011-2012.* Canadian Institute for Health Information. Ottawa, Ont. (2008).

[4] CIHI. "Standards and Data Submission - Case Mix." `http://www.cihi.ca`.

[5] CIHI. *DAD Abstracting Manual.* Canadian Institute for Health Information. Ottawa, Ont. (2010).

[6] CIHI. *DAD Resource Intensity Weights and Expected Length of Stay for CMG+ 2011.* Canadian Institute for Health Information. Ottawa, Ont. (2011).

[7] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications.* Springer. Davis, California (2007).

[8] R. Christensen. *Plane Answers to Complex Questions.* Springer. Fourth Edition (2011).

[9] E. Demidenko. *Mixed Models, Theory and Applications.* Wiley Interscience. Hoboken, New Jersey (2004).

[10] J. B. Willett and J. D. Singer. *Another Cautionary Note about R2: Its Use in Weighted Least-Squares Regression Analysis.* The American Statistical Association. The American Statistician, Vol. 42, No. 3 (Aug. 1988), pp. 236-238 (1988).

[11] P. McCullagh and J. Nelder. *Generalized Linear Model.* Chapman and Hall. 2nd edition (1989).

[12] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. *The Design of Simulation Studies in Medical Statistics.* Wiley InterScience. Statistics in Medicine (2006).

[13] J. Lindsey. *Applying Generalized Linear Models.* Limburgs Universitair Centrum, Diepenbeek (2007).

Appendix A

Major Clinical Categories

Table A.1: List of Major Clinical Categories (MCCs)

MCC	MCC Description
MCC 1	Diseases and Disorders of the Nervous System
MCC 2	Diseases and Disorders of the Eye
MCC 3	Diseases and Disorders of the Ea r, Nose, Mouth and Throat
MCC 4	Diseases and Disorders of the Respiratory System
MCC 5	Diseases and Disorders of the Circulatory System
MCC 6	Diseases and Disorders of the Digestive System
MCC 7	Diseases and Disorders of the Hepatobiliary System and Pancreas
MCC 8	Diseases and Disorders of the Musculoskeletal System and Connective Tissue
MCC 9	Diseases and Disorders of the Skin, Subcutaneous Tissue and Breast
MCC 10	Diseases and Disorders of the Endocrine System, Nutrition and Metabolism
MCC 12	Diseases and Disorders of the Kidney, Urinary Tract and Male Reproductive System
MCC 13	Pregnancy and Childbirth
MCC 14	Newborns and Neonates With Conditions Originating in the Perinatal Period
MCC 15	Diseases and Disorders of the Blood and Lymphatic System
MCC 16	Multisystemic or Unspecified Site Infections
MCC 17	Mental Diseases and Disorders
MCC 18	Burns
MCC 19	Significant Trauma, Injury, Poisoning and Toxic Effects of Drugs
MCC 20	Other Reasons for Hospitalization
MCC 99	Miscellaneous CMGs and Ungroupable Data

Appendix B

Optimization Algorithms

Let $\hat{\boldsymbol{\theta}}$ be the MLE that can be computed by numerical procedures (computing iteratively until convergence) since there is no closed form solution for optimal $\boldsymbol{\theta}$. There are three types of iterative alogrithms: the **Expectation Maximization (EM), Newton-Raphson (NR) and Fisher Scoring (FS)**. These algorithms differ by a negative Hessian matrix $\mathbf{H}$, which is the second derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$. NR and FS are the prefered alogrithms for ML and REML estimation because they have a quadratic convergence and produce an asymptotic covariance matrix $\mathbf{H}^{-1}$ of the estimated parameter. Statistically, we prefer to use the FS because it uses the expected negative Hessian matrix to estimate the covariance parameter while the NR only uses the observed Hessian matrix.

To find the minimum or maximum of a function, we often use the Newton-Raphson (NR) algorithm. This thesis focuses on the HER and HEREM models where we find the maximum of $l(\theta)$ which occurs when the gradient of $l(\theta)$ is equal to the zero vector, meaning when $\frac{\partial}{\partial\theta}l(\theta) = 0$. This means that the maximum is attained when the gradient points are equal to zero. The NR method brings us to the nearest point to the maximum; this method is called local optimization. To obtain good results using local optimization, we must choose a good initial starting value θ_0. In reference to the HER and HEREM models, we use the NR algorithm in this fashion:

- For $i = 0, \ldots$ update the current position using

$$\theta_{i+1} = \theta_i + C_i s_i \tag{B.1}$$

where s_i represents the gradient points at position i and their position in relation to the maximum. C_i is evaluated at θ_i and gives information about the curvature of the log-likelihood, thereby identifying the rate at which one could reach the maximum.

- The NR algorithm is repeated until s_i is as close as possible to 0.

Appendix C

Statistical Distribution

C.1 Multivariate Normal Distribution

If X_1 and X_2 are normally distributed and independent, this implies they are "jointly normally distributed", i.e., the vector $X = (X_1, X_2)'$ must have multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$X \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{C.1}$$

if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ such that } \begin{pmatrix} q \times 1 \\ (p-q) \times 1 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{ such that } \begin{pmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{pmatrix}$$

then the condition distribution of $X_1|X_2 = x_2$ is multivariate normal distribution

with mean vector μ and variance-covariance matrix Σ, where

$$\mu = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{C.2}$$

and

$$\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{C.3}$$

The mean of the conditional distribution is on a straight line called the regression line. In other words, the mean of the distribution of X_1 conditionally to $X_2 = x_2$ is a linear function of x_2. The equation of this line is

$$E(X_1|X_2 = x_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$$

and

$$Var(X_1|X_2 = x_2) = (1 - \rho^2)\sigma_1^2$$

where, ρ is the correlation of (X_1, X_2)

C.2 Exponential Family

A class of distributions, ranging between both continuous and discrete random variables, with the following form are part of the one-parameter exponential family (EF):

$$f(y_i;\eta) = h(y_i)s(\eta_i)exp\{T(y_i)u(\eta_i)\} \tag{C.4}$$

where Y_i is an independent response variable and η_i is a location parameter, which indicates the location of the distribution in the range of possible response values, and $i = 1, \ldots, n$. To simplify the above, we obtain the family distribution, the parameter and the canonical form by performing a one-to-one transformation $x = t(y)$ and

$\theta = u(\eta)$.

This can be rewritten in the following canonical form, where $a(\theta_i)$ is a normalizing constant distribution:

$$f(x_i;\theta_i) = exp\{x_i\theta_i - a(\theta_i) + c(x_i)\} \tag{C.5}$$

If we then add a scale parameter, ϕ, to the above, we can generalize the exponential dispersion family to give the following:

$$f(x_i;\theta_i,\phi) = exp\{\frac{x_i\theta_i - a(\theta_i)}{b_i(\phi)} + c(x_i,\phi)\} \tag{C.6}$$

where θ_i remains the canonical form of η_i, a function of the mean μ_i.

The mean and variance of the exponential and exponential dispersion families hold a special relationship. The likelihood function $L(\theta,\phi;x) = \prod_{i=1}^{n} f(x_i;\theta_i,\phi)$ is one method in which we can obtain the variance and the mean, for which the first derivative of the log likelihood $l(\theta,\phi;x) = log\Big(L(\theta,\phi;x)\Big)$ is obtained by

$$U = \frac{\partial l}{\partial \theta} \tag{C.7}$$

If we set the equation (C.7) to zero, then the MLE is derived.

From the standard inference theory, we can show that

$$\mathrm{E}(U) = 0 \tag{C.8}$$

and

$$Var(U) = \mathrm{E}(U^2) = \mathrm{E}(-\frac{\partial U}{\partial \theta}) \tag{C.9}$$

The log likelihood for a particular observation $l(\theta_i, \phi; x_i) = \frac{x_i\theta_i - a(\theta_i)}{b_i(\phi)} + c(x_i, \phi)$ Then for each θ_i,

$$U_i = \frac{x_i - \frac{\partial a(\theta_i)}{\partial \theta_i}}{b_i(\phi)} \tag{C.10}$$

From equation (C.8),

$$\mathrm{E}(U_i) = \mathrm{E}\Big(\frac{x_i - \frac{\partial a(\theta_i)}{\partial \theta_i}}{b_i(\phi)}\Big) = 0 \tag{C.11}$$

so that

$$\mathrm{E}(x_i) = \frac{\partial a(\theta_i)}{\partial \theta_i} = \mu_i \tag{C.12}$$

From equation (C.8), $U_i' = -\frac{\frac{\partial^2 a(\theta_i)}{\partial \theta_i^2}}{b_i(\phi)}$. Then, the variance of μ_i is obtained using the formula (C.10):

$$Var(U_i) = \frac{Var(X_i)}{b_i^2(\phi)} = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} b_i(\phi) \tag{C.13}$$

Then, we rearrange the above equation, which yields $Var(X_i) = \frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} b_i(\phi)$. This can be further simplified by taking $b_i(\phi) = \frac{\phi}{w_i}$ where w_i represents prior weights. If we then let the variance function (a function of μ_i or θ_i only) $\frac{\partial^2 a(\theta_i)}{\partial \theta_i^2} = \tau_i^2$, we obtain the product of the dispersion parameter and a function of the mean.[1]

$$Var(X_i) = b_i(\phi)\tau_i^2 = \frac{\phi \tau_i^2}{w_i} \tag{C.14}$$

The EF members share the same properties.

- The product of the pdf for two random variables (X, Y), or more, will belong to an EF if the pdf for each of the random variables belongs to an EF.
- Bayesian estimation is easy to calculate because every EF distribution has a conjugate prior.

[1]The derivation method described in this appendix for the EF was obtained largely from J.K. Lindsey [13].

- For the modelling purpose, if Y is from an EF, then $var(Y) = V(\mu)\phi$ where V is a known function of $\mu = \mathrm{E}(Y)$, and ϕ is a scale parameter.

Table C.1 lists distributions that are members of the EF:

Table C.1: Exponential Distribution

Exponential Distribution	**Domain**
Bernoulli	binary{0,1}
Beta	(0,1)
Binomial	counts of success or failure
Dirichlet	(Simplex)
Exponential	$\mathbb{R}^+$
Gamma	$\mathbb{R}^+$
Gaussian	$\mathbb{R}^p$
Laplace	$\mathbb{R}^+$
Multinomial	*categorical*
Poisson	$\mathbb{N}^+$
Von mises	sphere
Weibull	$\mathbb{R}^+$
Weishart	symmetric positive definite matrices

The lognormal and Pareto distributions are not in the exponential family.

Appendix D

Data Analysis Results

As seen in chapter 5, the data used for the analysis was varied. Data has been organised in this appendix by significant and insignificant main covariates for the mean estimate and the heteroscedastic variance estimate respectively, where the significant and insignificant covariates are obtained by four models: the Current Model (see section D.1), the HER model (see section D.2), the HEREM model (see section D.3) and the GLM (see section D.4).

D.1 Current Model Results

Table D.1: Significant Current Beta (Mean Estimate)

Covariate	Estimate	SE	t-value	p-value
Intercept	8.8304	0.0518	170.5322	0.0000
IE_2	0.4520	0.0780	5.7950	0.0000
IE_3	0.1934	0.0326	5.9379	0.0000
CmL_1	0.6621	0.0527	12.5689	0.0000
CmL_3	0.1239	0.0369	3.3533	0.0008
CmL_4	0.2821	0.0260	10.8505	0.0000
CMG_160	1.4850	0.1456	10.2009	0.0000
CMG_161	1.1246	0.0416	27.0115	0.0000
CMG_162	1.0078	0.0377	26.7081	0.0000
CMG_163	0.6779	0.0423	16.0261	0.0000
CMG_164	0.2417	0.0589	4.1054	0.0000
CMG_165	0.6972	0.0506	13.7909	0.0000
CMG_166	1.2373	0.0601	20.5991	0.0000
CMG_167	1.0514	0.1182	8.8941	0.0000
CMG_168	1.1001	0.0440	24.9952	0.0000
CMG_169	0.9994	0.0610	16.3875	0.0000
CMG_170	0.8502	0.1152	7.3790	0.0000
CMG_171	1.1324	0.2699	4.1954	0.0000
CMG_172	0.8363	0.0396	21.1383	0.0000
CMG_173	0.2211	0.0598	3.6968	0.0002
CMG_174	0.3320	0.0382	8.6798	0.0000
CMG_175	0.1981	0.0370	5.3505	0.0000
CMG_176	-0.1342	0.0371	-3.6175	0.0003
CMG_179	-0.5450	0.0398	-13.7000	0.0000
CMG_180	0.5119	0.0546	9.3710	0.0000
CMG_181	0.8064	0.0380	21.2082	0.0000

Continued on next page

Table D.1 – continued Significant Current Beta

Covariate	Estimate	SE	t-value	p-value
CMG_182	0.0802	0.0388	2.0663	0.0388
CMG_183	-0.2515	0.0839	-2.9985	0.0027
CMG_184	-1.0636	0.0704	-15.1004	0.0000
CMG_185	-0.0995	0.0415	-2.3960	0.0166
CMG_193	-0.8060	0.0893	-9.0274	0.0000
CMG_194	-1.1517	0.0896	-12.8525	0.0000
CMG_195	-0.4975	0.0916	-5.4320	0.0000
CMG_196	-1.1876	0.0890	-13.3411	0.0000
CMG_197	-1.2719	0.1245	-10.2165	0.0000
CMG_198	-1.4250	0.1010	-14.1049	0.0000
CMG_199	-1.0842	0.0977	-11.1012	0.0000
CMG_200	-1.3386	0.0905	-14.7875	0.0000
CMG_201	-0.8345	0.0927	-9.0021	0.0000
CMG_202	-1.6873	0.0894	-18.8825	0.0000
CMG_203	-1.1415	0.0899	-12.6978	0.0000
CMG_204	-1.6343	0.0899	-18.1710	0.0000
CMG_205	-1.7388	0.0897	-19.3936	0.0000
CMG_206	-1.7324	0.0936	-18.5089	0.0000
CMG_207	-1.2138	0.0892	-13.6055	0.0000
CMG_208	-1.9358	0.0890	-21.7429	0.0000
CMG_209	-1.4488	0.0897	-16.1487	0.0000
CMG_210	-1.2474	0.1163	-10.7281	0.0000
CMG_211	-1.3112	0.0957	-13.6952	0.0000
CMG_212	-1.4323	0.1125	-12.7297	0.0000
CMG_213	-1.4640	0.0944	-15.5041	0.0000
OOhCath_Y	-0.0751	0.0138	-5.4388	0.0000
OOhImplant_Y	-0.4669	0.0325	-14.3823	0.0000
OOhPtca_Y	-0.4047	0.0141	-28.6047	0.0000

Continued on next page

Table D.1 – continued Significant Current Beta

Covariate	Estimate	SE	t-value	p-value
HID_51406	0.5151	0.0404	12.7556	0.0000
HID_51423	-0.1626	0.0470	-3.4607	0.0005
HID_51597	-0.3628	0.0623	-5.8208	0.0000
HID_51748	-0.1508	0.0567	-2.6593	0.0078
HID_51982	-0.0840	0.0372	-2.2590	0.0239
HID_51983	-0.2974	0.0434	-6.8463	0.0000
HID_51994	-0.4322	0.0481	-8.9809	0.0000
HID_52003	-0.2669	0.0428	-6.2331	0.0000
HID_52038	-0.0990	0.0372	-2.6631	0.0077
HID_52046	-0.1850	0.0409	-4.5226	0.0000
HID_53850	0.1917	0.0376	5.1030	0.0000
HID_53910	0.1520	0.0374	4.0623	0.0000
HID_53992	-0.3698	0.0490	-7.5493	0.0000
HID_53994	0.2127	0.0515	4.1320	0.0000
HID_54048	-0.1212	0.0457	-2.6503	0.0080
HID_80015	0.2082	0.1028	2.0251	0.0429
HID_80016	0.2992	0.0375	7.9803	0.0000
HID_80020	0.3783	0.0400	9.4682	0.0000
HID_80033	-0.5026	0.1478	-3.4015	0.0007
HID_80041	0.3931	0.0441	8.9194	0.0000
HID_80042	0.3276	0.0405	8.0908	0.0000
HID_80043	0.2687	0.0397	6.7700	0.0000
HID_80044	0.2886	0.0381	7.5689	0.0000
HID_80052	-0.3397	0.0825	-4.1180	0.0000
HID_80120	0.1863	0.0431	4.3216	0.0000
HID_80122	-0.5734	0.1086	-5.2772	0.0000
HID_80148	0.2927	0.0397	7.3784	0.0000
HID_90102	0.2649	0.0382	6.9403	0.0000

Continued on next page

Table D.1 – continued Significant Current Beta

Covariate	Estimate	SE	t-value	p-value
HID_90109	0.1686	0.0379	4.4436	0.0000
FIcnt2_1	-0.0692	0.0130	-5.3267	0.0000
FIcnt3_1	-0.1870	0.0280	-6.6873	0.0000
fiBio_1	0.4788	0.0349	13.7367	0.0000
fiCel_1	0.3250	0.0397	8.1930	0.0000
fiCmo_1	0.4281	0.1180	3.6271	0.0003
fiCrd_1	0.2089	0.0332	6.2839	0.0000
fiDia_1	0.3261	0.0234	13.9613	0.0000
fiEnd_1	0.3627	0.0406	8.9372	0.0000
fiHrt_1	0.1124	0.0330	3.4053	0.0007
fiMeG96_1	0.7762	0.0346	22.4558	0.0000
fiMeL96_1	0.6019	0.0405	14.8685	0.0000
fiNut_1	0.3089	0.0675	4.5764	0.0000
fiPar_1	0.3232	0.0849	3.8077	0.0001
fiPlr_1	0.3232	0.0244	13.2629	0.0000
fiRad_1	0.4037	0.1547	2.6098	0.0091
fiTra_1	0.5038	0.0619	8.1407	0.0000
fiTub_1	0.5573	0.0425	13.1135	0.0000
fiVad_1	0.4266	0.0364	11.7246	0.0000

Table D.2: Insignificant Current Beta (Mean Estimate)

Covariate	Estimate	SE	t-value	p-value
CmL_2	0.0647	0.0395	1.6388	0.1013
CMG_177	-0.0512	0.0565	-0.9047	0.3657
CMG_178	-0.0452	0.0439	-1.0285	0.3037
HID_51199	0.0015	0.0562	0.0275	0.9781

Continued on next page

Table D.2 – continued Insignificant Current Beta

Covariate	Estimate	SE	t-value	p-value
HID_51213	-0.0794	0.0770	-1.0313	0.3024
HID_51444	-0.0691	0.0397	-1.7403	0.0818
HID_51754	-0.0446	0.0434	-1.0295	0.3033
HID_53917	0.0368	0.0374	0.9819	0.3262
HID_53932	0.0147	0.0397	0.3707	0.7108
HID_53936	-0.0695	0.0371	-1.8728	0.0611
HID_53988	-0.0014	0.0420	-0.0339	0.9729
HID_80070	-0.1201	0.0674	-1.7815	0.0748
HID_80150	-0.0495	0.1049	-0.4721	0.6369
HID_90130	0.0415	0.0421	0.9841	0.3251
HID_90136	-0.0393	0.0451	-0.8719	0.3833
FsYr_2008	0.0831	0.0504	1.6478	0.0994

Table D.3: Significant Current Gamma (Variance Estimate)

Covariate	Estimate	SE	t-value	p-value
Intercept	0.6010	0.0478	12.5635	0.0000
CmL_1	-0.1080	0.0500	-2.1617	0.0306
CMG_161	-0.1287	0.0372	-3.4557	0.0005
CMG_162	-0.5365	0.0409	-13.1020	0.0000
CMG_163	-0.4125	0.0507	-8.1294	0.0000
CMG_164	-0.1627	0.0601	-2.7061	0.0068
CMG_165	-0.4054	0.0743	-5.4550	0.0000
CMG_166	-0.4818	0.1174	-4.1057	0.0000
CMG_168	-0.4649	0.0818	-5.6805	0.0000
CMG_169	-0.5259	0.1627	-3.2334	0.0012
CMG_170	-0.4979	0.1598	-3.1153	0.0018

Continued on next page

Table D.3 – continued Significant Current Gamma

Covariate	Estimate	SE	t-value	p-value
CMG_172	-0.5647	0.0383	-14.7484	0.0000
CMG_173	-0.2099	0.0629	-3.3349	0.0009
CMG_174	-0.2371	0.0340	-6.9793	0.0000
CMG_175	-0.3652	0.0328	-11.1478	0.0000
CMG_176	-0.3187	0.0324	-9.8314	0.0000
CMG_177	0.1897	0.0445	4.2673	0.0000
CMG_178	-0.1393	0.0396	-3.5212	0.0004
CMG_180	-0.2430	0.0576	-4.2181	0.0000
CMG_181	-0.3906	0.0358	-10.9158	0.0000
CMG_182	-0.3451	0.0364	-9.4884	0.0000
CMG_184	-0.3401	0.0961	-3.5407	0.0004
CMG_185	-0.1360	0.0368	-3.6977	0.0002
CMG_193	-0.4328	0.1036	-4.1793	0.0000
CMG_195	-0.3579	0.1064	-3.3629	0.0008
CMG_201	-0.3219	0.1073	-3.0011	0.0027
CMG_207	-0.3802	0.1032	-3.6848	0.0002
CMG_210	0.2569	0.1138	2.2585	0.0239
OOhCath_Y	0.0682	0.0167	4.0861	0.0000
OOhImplant_Y	0.0710	0.0341	2.0801	0.0375
HID_51199	-0.4020	0.0716	-5.6102	0.0000
HID_51213	-0.3102	0.1018	-3.0489	0.0023
HID_51423	0.3378	0.0418	8.0864	0.0000
HID_51444	0.3420	0.0384	8.9154	0.0000
HID_51597	-0.2561	0.0689	-3.7155	0.0002
HID_51748	-0.3352	0.0742	-4.5176	0.0000
HID_51754	-0.3169	0.0477	-6.6451	0.0000
HID_51994	0.3393	0.0402	8.4483	0.0000
HID_52003	0.1218	0.0388	3.1376	0.0017

Continued on next page

Table D.3 – continued Significant Current Gamma

Covariate	Estimate	SE	t-value	p-value
HID_52038	-0.0903	0.0368	-2.4525	0.0142
HID_53910	-0.1218	0.0376	-3.2376	0.0012
HID_53917	-0.1551	0.0379	-4.0979	0.0000
HID_53932	-0.1452	0.0406	-3.5756	0.0003
HID_53936	-0.1126	0.0362	-3.1120	0.0019
HID_53988	-0.2089	0.0435	-4.7989	0.0000
HID_53992	-0.2662	0.0552	-4.8213	0.0000
HID_53994	-0.1466	0.0527	-2.7825	0.0054
HID_54048	0.1239	0.0426	2.9074	0.0036
HID_80016	-0.1675	0.0380	-4.4103	0.0000
HID_80020	-0.1343	0.0399	-3.3610	0.0008
HID_80033	0.2420	0.1061	2.2802	0.0226
HID_80043	-0.0847	0.0408	-2.0756	0.0379
HID_90102	-0.0869	0.0407	-2.1357	0.0327
fiBio_1	-0.1260	0.0558	-2.2587	0.0239
fiCmo_1	0.2296	0.0872	2.6344	0.0084
fiDia_1	-0.0740	0.0360	-2.0568	0.0397
fiNut_1	0.1488	0.0698	2.1308	0.0331
fiPar_1	0.1566	0.0791	1.9789	0.0478
fiRad_1	0.3010	0.1092	2.7562	0.0058
fiTra_1	0.1763	0.0711	2.4784	0.0132

Table D.4: Insignificant Current Gamma (Variance Estimate)

Covariate	Estimate	SE	t-value	p-value
IE_2	0.0600	0.0839	0.7160	0.4740
IE_3	0.0053	0.0590	0.0891	0.9290

Continued on next page

Table D.4 – continued Insignificant Current Gamma

Covariate	Estimate	SE	t-value	p-value
CmL_2	-0.0528	0.0494	-1.0682	0.2854
CmL_3	0.0014	0.0492	0.0281	0.9776
CmL_4	-0.0028	0.0401	-0.0693	0.9448
CMG_160	-0.3222	0.1992	-1.6170	0.1059
CMG_167	-0.4701	0.3328	-1.4125	0.1578
CMG_171	-0.2929	0.4459	-0.6569	0.5113
CMG_179	-0.0207	0.0340	-0.6089	0.5426
CMG_183	-0.1028	0.0779	-1.3197	0.1869
CMG_194	-0.1890	0.1030	-1.8348	0.0665
CMG_196	-0.1145	0.1024	-1.1180	0.2636
CMG_197	-0.0317	0.1253	-0.2528	0.8004
CMG_198	-0.0571	0.1105	-0.5164	0.6056
CMG_199	-0.0103	0.1070	-0.0967	0.9230
CMG_200	-0.1546	0.1036	-1.4923	0.1356
CMG_202	-0.0287	0.1025	-0.2801	0.7794
CMG_203	-0.1815	0.1030	-1.7610	0.0782
CMG_204	-0.1151	0.1030	-1.1170	0.2640
CMG_205	-0.1464	0.1029	-1.4227	0.1548
CMG_206	0.0096	0.1048	0.0918	0.9268
CMG_208	-0.1440	0.1024	-1.4063	0.1596
CMG_209	0.0138	0.1027	0.1339	0.8935
CMG_211	-0.0760	0.1069	-0.7108	0.4772
CMG_212	0.1244	0.1137	1.0941	0.2739
CMG_213	0.1021	0.1049	0.9732	0.3305
OOhPtca_Y	-0.0185	0.0183	-1.0107	0.3122
HID_51406	-0.0578	0.0436	-1.3267	0.1846
HID_51982	-0.0143	0.0357	-0.4005	0.6888
HID_51983	0.0558	0.0395	1.4131	0.1576

Continued on next page

Table D.4 – continued Insignificant Current Gamma

Covariate	Estimate	SE	t-value	p-value
HID_52046	0.0226	0.0381	0.5944	0.5523
HID_53850	-0.0403	0.0380	-1.0613	0.2886
HID_80015	-0.0079	0.0898	-0.0882	0.9297
HID_80041	-0.0795	0.0464	-1.7142	0.0865
HID_80042	-0.0507	0.0424	-1.1963	0.2316
HID_80044	-0.0092	0.0391	-0.2352	0.8141
HID_80052	0.0078	0.0750	0.1042	0.9170
HID_80070	0.0185	0.0624	0.2965	0.7669
HID_80120	-0.0529	0.0452	-1.1696	0.2422
HID_80122	0.1095	0.0871	1.2574	0.2086
HID_80148	-0.0766	0.0404	-1.8965	0.0579
HID_80150	0.1450	0.0841	1.7242	0.0847
HID_90109	-0.0617	0.0399	-1.5469	0.1219
HID_90130	-0.0424	0.0421	-1.0080	0.3135
HID_90136	-0.0914	0.0474	-1.9269	0.0540
FIcnt2_1	0.0241	0.0214	1.1257	0.2603
FIcnt3_1	0.0707	0.0434	1.6288	0.1034
FsYr_2008	0.0521	0.0488	1.0689	0.2851
fiCel_1	0.0181	0.0688	0.2640	0.7918
fiCrd_1	0.0010	0.0562	0.0181	0.9856
fiEnd_1	-0.0180	0.0535	-0.3366	0.7364
fiHrt_1	0.0380	0.0488	0.7783	0.4364
fiMeG96_1	-0.0532	0.0469	-1.1343	0.2567
fiMeL96_1	0.0157	0.0465	0.3385	0.7350
fiPlr_1	0.0103	0.0411	0.2518	0.8012
fiTub_1	-0.0530	0.0665	-0.7971	0.4254
fiVad_1	0.0432	0.0445	0.9697	0.3322

D.2 HER Model Results

This section will detail the significant and insignificant mean and variance of main effects of chapter 5, section 5.5

Table D.5: Significant HER Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	8.8265	0.0517	8.7252	8.9278
IE_2	0.3565	0.0737	0.212	0.501
IE_3	0.2343	0.0366	0.1624	0.3061
CmL_1	0.597	0.0518	0.4954	0.6986
CmL_2	0.1654	0.0375	0.0918	0.2389
CmL_3	0.1496	0.0374	0.0762	0.223
CmL_4	0.267	0.0305	0.2073	0.3268
CMG_160	1.5716	0.1274	1.3218	1.8213
CMG_161	1.1511	0.0405	1.0717	1.2304
CMG_162	1.0155	0.0384	0.9402	1.0908
CMG_163	0.7001	0.0426	0.6167	0.7835
CMG_164	0.2059	0.0595	0.0894	0.3225
CMG_165	0.7121	0.0513	0.6116	0.8127
CMG_166	1.2492	0.062	1.1277	1.3706
CMG_167	1.0092	0.1407	0.7335	1.285
CMG_168	1.0672	0.0543	0.9608	1.1736
CMG_169	0.9589	0.0695	0.8227	1.0951
CMG_170	0.9594	0.0813	0.8001	1.1187
CMG_171	1.0391	0.2629	0.5238	1.5543
CMG_172	0.7369	0.0378	0.6628	0.8109
CMG_173	0.2795	0.0605	0.1609	0.3982
CMG_174	0.3346	0.0382	0.2597	0.4095
CMG_175	0.1993	0.037	0.1267	0.2719

Continued on next page

Table D.5 – continued Significant HER Beta

Covariate	Estimate	SE	95% LB	95% UB
CMG_176	-0.1261	0.0371	-0.1988	-0.0535
CMG_179	-0.4502	0.0387	-0.526	-0.3744
CMG_180	0.4962	0.0553	0.3878	0.6046
CMG_181	0.8107	0.0385	0.7353	0.886
CMG_182	0.0792	0.0392	0.0024	0.1559
CMG_183	-0.2738	0.0849	-0.4402	-0.1075
CMG_184	-1.0692	0.0712	-1.2088	-0.9297
CMG_185	-0.1034	0.0417	-0.1852	-0.0217
CMG_193	-0.7474	0.0869	-0.9177	-0.5772
CMG_194	-1.1027	0.087	-1.2733	-0.9321
CMG_195	-0.4517	0.0888	-0.6257	-0.2777
CMG_196	-1.1412	0.0865	-1.3107	-0.9718
CMG_197	-1.2285	0.1224	-1.4685	-0.9886
CMG_198	-1.3632	0.0979	-1.555	-1.1713
CMG_199	-1.0531	0.0948	-1.2388	-0.8673
CMG_200	-1.2885	0.088	-1.461	-1.116
CMG_201	-0.7844	0.0901	-0.961	-0.6079
CMG_202	-1.6323	0.0868	-1.8024	-1.4622
CMG_203	-1.0793	0.0871	-1.2499	-0.9086
CMG_204	-1.5853	0.0874	-1.7565	-1.4141
CMG_205	-1.6914	0.0871	-1.8622	-1.5206
CMG_206	-1.6839	0.0909	-1.8621	-1.5057
CMG_207	-1.164	0.0867	-1.3339	-0.9941
CMG_208	-1.8881	0.0865	-2.0576	-1.7186
CMG_209	-1.3971	0.0872	-1.5679	-1.2262
CMG_210	-1.2035	0.1123	-1.4236	-0.9833
CMG_211	-1.2673	0.0933	-1.4502	-1.0844
CMG_212	-1.4219	0.1098	-1.6371	-1.2068

Continued on next page

Table D.5 – continued Significant HER Beta

Covariate	Estimate	SE	95% LB	95% UB
CMG_213	-1.4162	0.0921	-1.5968	-1.2357
OOhCath_Y	-0.0632	0.0139	-0.0905	-0.036
OOhImplant_Y	-0.4631	0.0324	-0.5266	-0.3997
OOhPtca_Y	-0.3931	0.0143	-0.4211	-0.3651
HID_51406	0.5058	0.0409	0.4256	0.5859
HID_51423	-0.1806	0.0463	-0.2713	-0.0899
HID_51597	-0.3313	0.0638	-0.4564	-0.2061
HID_51983	-0.2825	0.0433	-0.3674	-0.1977
HID_51994	-0.4057	0.0476	-0.499	-0.3125
HID_52003	-0.2667	0.0424	-0.3499	-0.1836
HID_52038	-0.0828	0.0374	-0.1561	-0.0095
HID_52046	-0.1785	0.0406	-0.2581	-0.0989
HID_53850	0.2006	0.0373	0.1275	0.2738
HID_53910	0.1511	0.0372	0.0783	0.2239
HID_53936	-0.0857	0.0373	-0.1588	-0.0125
HID_53992	-0.3345	0.0495	-0.4315	-0.2375
HID_53994	0.2142	0.0513	0.1136	0.3148
HID_54048	-0.1043	0.0451	-0.1927	-0.0158
HID_80015	0.2136	0.1041	0.0097	0.4176
HID_80016	0.2906	0.0373	0.2176	0.3637
HID_80020	0.3867	0.0397	0.3089	0.4644
HID_80033	-0.5016	0.1443	-0.7843	-0.2188
HID_80041	0.3945	0.044	0.3082	0.4807
HID_80042	0.3317	0.0401	0.2531	0.4102
HID_80043	0.2673	0.0394	0.1901	0.3446
HID_80044	0.3068	0.0377	0.2329	0.3807
HID_80052	-0.3282	0.082 -	0.489	-0.1674
HID_80120	0.1802	0.0428	0.0963	0.264

Continued on next page

Table D.5 – continued Significant HER Beta

Covariate	Estimate	SE	95% LB	95% UB
HID_80122	-0.569	0.1072	-0.7792	-0.3588
HID_80148	0.278	0.0396	0.2003	0.3557
HID_90102	0.2633	0.038	0.1887	0.3378
HID_90109	0.1996	0.0381	0.125	0.2742
FIcnt2_1	-0.1033	0.0144	-0.1316	-0.0751
FIcnt3_1	-0.2839	0.0297	-0.3422	-0.2257
fiBio_1	0.4495	0.045	0.3614	0.5376
fiCel_1	0.1414	0.0422	0.0587	0.2241
fiCmo_1	0.4373	0.1144	0.213	0.6615
fiCrd_1	0.223	0.0348	0.1547	0.2912
fiDia_1	0.3753	0.0276	0.3211	0.4294
fiEnd_1	0.4466	0.0456	0.3573	0.536
fiHrt_1	0.1302	0.0344	0.0629	0.1976
fiMeG96_1	0.7789	0.0379	0.7047	0.8531
fiMeL96_1	0.6125	0.0423	0.5296	0.6954
fiNut_1	0.328	0.0636	0.2033	0.4526
fiPar_1	0.2923	0.0757	0.1439	0.4407
fiPlr_1	0.3623	0.0281	0.3072	0.4173
fiRad_1	0.3826	0.1524	0.0839	0.6813
fiTra_1	0.5448	0.0629	0.4216	0.6679
fiTub_1	0.5825	0.0478	0.4887	0.6762
fiVad_1	0.4258	0.0353	0.3565	0.4951

Table D.6: Insignificant HER Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
CMG_177	-0.0228	0.0555	-0.1317	0.0861

Continued on next page

Table D.6 – continued Insignificant HER Beta

Covariate	Estimate	SE	95% LB	95% UB
CMG_178	-0.0374	0.0439	-0.1234	0.0486
HID_51199	0.0079	0.0554	-0.1008	0.1166
HID_51213	-0.0122	0.0855	-0.1797	0.1553
HID_51444	0.0028	0.039	-0.0737	0.0792
HID_51748	-0.0387	0.0625	-0.1611	0.0838
HID_51754	-0.0117	0.0445	-0.0989	0.0755
HID_51982	-0.0456	0.037	-0.1182	0.027
HID_53917	0.0555	0.0374	-0.0177	0.1288
HID_53932	0.0295	0.0399	-0.0487	0.1076
HID_53988	0.0033	0.0421	-0.0791	0.0858
HID_80070	-0.1162	0.0656	-0.2448	0.0124
HID_80150	-0.0371	0.1013	-0.2357	0.1615
HID_90130	0.0438	0.0426	-0.0397	0.1273
HID_90136	-0.0242	0.0457	-0.1137	0.0654
FsYr2008	0.0814	0.0504	-0.0175	0.1802

Table D.7: Significant HER Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	-0.4235	0.1154	-0.6498	-0.1973
CL_1	-0.2965	0.113	-0.5179	-0.0751
CMG_160	-1.5935	0.4384	-2.4527	-0.7342
CMG_161	-0.5599	0.1024	-0.7607	-0.3591
CMG_162	-2.2665	0.1119	-2.4858	-2.0472
CMG_163	-1.6354	0.1161	-1.8629	-1.408
CMG_164	-0.4945	0.1675	-0.8229	-0.1661
CMG_165	-1.6418	0.1579	-1.9514	-1.3323

Continued on next page

Table D.7 – continued Significant HER Gamma

Covariate	Estimate	SE	95% LB	95% UB
CMG_166	-1.9551	0.3023	-2.5477	-1.3626
CMG_167	-2.4185	0.4242	-3.25	-1.5871
CMG_168	-1.8187	0.1838	-2.1789	-1.4586
CMG_169	-2.4209	0.3774	-3.1606	-1.6811
CMG_170	-2.0549	0.2811	-2.6058	-1.5041
CMG_172	-2.4418	0.0945	-2.6271	-2.2566
CMG_173	-0.5969	0.2167	-1.0217	-0.1721
CMG_174	-0.6871	0.0823	-0.8484	-0.5258
CMG_175	-1.4774	0.0773	-1.6288	-1.326
CMG_176	-1.2401	0.0777	-1.3924	-1.0879
CMG_177	0.2552	0.0942	0.0705	0.4398
CMG_178	-0.3479	0.1071	-0.5577	-0.1381
CMG_179	-0.5526	0.0803	-0.7099	-0.3952
CMG_180	-0.6089	0.1213	-0.8466	-0.3711
CMG_181	-1.1415	0.083	-1.3043	-0.9787
CMG_182	-0.9183	0.0963	-1.107	-0.7296
CMG_184	-1.0578	0.2417	-1.5315	-0.5841
CMG_185	-0.2648	0.092	-0.4452	-0.0844
CMG_193	-1.5221	0.2059	-1.9258	-1.1185
CMG_194	-0.6312	0.2044	-1.0318	-0.2307
CMG_195	-1.3736	0.2182	-1.8014	-0.9459
CMG_196	-0.4515	0.2019	-0.8472	-0.0558
CMG_199	-0.4397	0.2157	-0.8624	-0.017
CMG_200	-0.5165	0.2052	-0.9187	-0.1142
CMG_201	-1.1339	0.2141	-1.5536	-0.7141
CMG_203	-0.9752	0.2049	-1.3769	-0.5735
CMG_204	-0.477	0.2038	-0.8764	-0.0776
CMG_205	-0.5191	0.2029	-0.9167	-0.1216

Continued on next page

Table D.7 – continued Significant HER Gamma

Covariate	Estimate	SE	95% LB	95% UB
CMG_207	-1.3714	0.2045	-1.7722	-0.9705
CMG_208	-0.5239	0.2017	-0.9193	-0.1285
OOhCath_Y	0.3462	0.0422	0.2634	0.429
OOhImplant_Y	0.3375	0.0773	0.1861	0.489
OOhPtca_Y	0.4043	0.0441	0.3178	0.4907
HID_51199	-1.2636	0.131	-1.5203	-1.0068
HID_51213	-0.8205	0.2469	-1.3044	-0.3366
HID_51423	0.564	0.1019	0.3642	0.7637
HID_51444	1.0304	0.0942	0.8459	1.2149
HID_51597	-0.684	0.181	-1.0387	-0.3293
HID_51748	-0.9411	0.1742	-1.2826	-0.5996
HID_51754	-0.9129	0.1062	-1.121	-0.7049
HID_51994	0.3608	0.0842	0.1958	0.5257
HID_53850	-0.202	0.094	-0.3862	-0.0179
HID_53910	-0.3658	0.0933	-0.5488	-0.1829
HID_53917	-0.4191	0.0934	-0.6022	-0.236
HID_53932	-0.3652	0.1027	-0.5666	-0.1638
HID_53936	-0.2369	0.0798	-0.3933	-0.0805
HID_53988	-0.5422	0.1041	-0.7463	-0.3381
HID_53992	-0.7411	0.1323	-1.0004	-0.4819
HID_53994	-0.3622	0.1233	-0.6037	-0.1206
HID_80016	-0.5075	0.094	-0.6917	-0.3232
HID_80020	-0.3722	0.0967	-0.5619	-0.1826
HID_80041	-0.2785	0.1167	-0.5072	-0.0498
HID_80042	-0.2629	0.1106	-0.4798	-0.0461
HID_80043	-0.2198	0.1026	-0.4208	-0.0188
HID_80148	-0.2402	0.0987	-0.4336	-0.0468
HID_90102	-0.2548	0.1006	-0.4519	-0.0577

Continued on next page

Table D.7 – continued Significant HER Gamma

Covariate	Estimate	SE	95% LB	95% UB
fiBio_1	-0.3404	0.1206	-0.5769	-0.104
fiNut_1	0.3991	0.1484	0.1083	0.6899
fiPar_1	0.3427	0.1639	0.0215	0.6639
fiPlr_1	0.2282	0.0871	0.0574	0.3989
fiTra_1	0.6284	0.2412	0.1558	1.1011

Table D.8: Insignificant HER Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
IE_2	0.2924	0.1569	-0.0151	0.5999
IE_3	0.0516	0.1242	-0.1918	0.2951
CmL_2	-0.1049	0.104	-0.3087	0.0989
CmL_3	-0.0471	0.1076	-0.2581	0.1639
CmL_4	-0.0042	0.0859	-0.1725	0.164
CMG_171	-0.4196	0.6007	-1.5969	0.7578
CMG_183	-0.1688	0.2133	-0.587	0.2494
CMG_197	-0.2777	0.2633	-0.7938	0.2384
CMG_198	-0.4167	0.2284	-0.8643	0.0309
CMG_202	-0.3242	0.2021	-0.7203	0.072
CMG_206	-0.2541	0.207	-0.6597	0.1516
CMG_209	-0.2233	0.2027	-0.6205	0.1739
CMG_210	0.0286	0.2394	-0.4405	0.4978
CMG_211	-0.3559	0.223	-0.7931	0.0812
CMG_212	-0.116	0.2319	-0.5706	0.3386
CMG_213	-0.0728	0.2086	-0.4817	0.3361
HID_51406	-0.1007	0.1072	-0.3107	0.1093
HID_51982	-0.1242	0.0791	-0.2793	0.0308

Continued on next page

Table D.8 – continued Insignificant HER Gamma

Covariate	Estimate	SE	95% LB	95% UB
HID_51983	0.0041	0.0854	-0.1632	0.1714
HID_52003	0.1068	0.0863	-0.0624	0.2761
HID_52038	-0.1545	0.0832	-0.3176	0.0086
HID_52046	-0.0868	0.0822	-0.2479	0.0742
HID_54048	0.1858	0.1015	-0.0132	0.3847
HID_80015	0.0136	0.1581	-0.2963	0.3236
HID_80033	0.2767	0.1899	-0.0954	0.6489
HID_80044	-0.0695	0.0971	-0.2598	0.1208
HID_80052	0.0049	0.145	-0.2793	0.2891
HID_80070	-0.0336	0.1289	-0.2861	0.219
HID_80120	-0.2144	0.112	-0.4339	0.0052
HID_80122	0.1387	0.1722	-0.1989	0.4762
HID_80150	0.1502	0.1953	-0.2325	0.5329
HID_90109	-0.1177	0.1002	-0.3142	0.0787
HID_90130	-0.0111	0.1026	-0.2121	0.19
HID_90136	-0.2011	0.1112	-0.4189	0.0168
FIcnt2_1	0.0729	0.0561	-0.037	0.1828
FIcnt3_1	0.0514	0.1093	-0.1628	0.2655
FsYr_2008	0.1581	0.1139	-0.0651	0.3813

D.3 HEREM Model Results

This section will detail the significant and insignificant mean and variance of main effects of chapter 5, section 5.6

Table D.9: Significant HEREM Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	8.9042	0.0407	8.8245	8.984
IE_2	0.4102	0.039	0.3338	0.4866
IE_3	0.2688	0.0365	0.1973	0.3403
CmL_1	0.4156	0.0239	0.3688	0.4623
CmL_2	0.1645	0.0254	0.1147	0.2143
CmL_3	0.1645	0.0268	0.1119	0.217
CmL_4	0.2919	0.0287	0.2357	0.3481
CMG_160	1.327	0.1072	1.1168	1.5371
CMG_161	1.0751	0.0344	1.0078	1.1425
CMG_162	0.947	0.0324	0.8835	1.0104
CMG_163	0.6315	0.0372	0.5587	0.7044
CMG_164	0.1373	0.0543	0.0309	0.2437
CMG_165	0.6228	0.0445	0.5356	0.71
CMG_166	1.1259	0.0609	1.0066	1.2452
CMG_167	0.8505	0.1301	0.5956	1.1055
CMG_168	0.9958	0.0465	0.9047	1.0869
CMG_169	0.8758	0.0717	0.7354	1.0163
CMG_170	0.8047	0.0746	0.6584	0.951
CMG_171	0.7569	0.1596	0.444	1.0697
CMG_172	0.6655	0.0316	0.6035	0.7275
CMG_173	0.1977	0.0518	0.0961	0.2993
CMG_174	0.262	0.0319	0.1994	0.3245
CMG_175	0.1246	0.0306	0.0645	0.1846
CMG_176	-0.1992	0.0306	-0.2592	-0.1392
CMG_178	-0.1028	0.0384	-0.178	-0.0276
CMG_179	-0.5171	0.0324	-0.5806	-0.4535
CMG_180	0.4148	0.0483	0.3201	0.5095
CMG_181	0.7396	0.0321	0.6766	0.8026

Continued on next page

Table D.9 – continued Significant HEREM Beta

Covariate	Estimate	SE	95% LB	95% UB
CMG_183	-0.3115	0.0788	-0.466	-0.1571
CMG_184	-1.1488	0.0683	-1.2826	-1.0149
CMG_185	-0.1773	0.0359	-0.2476	-0.107
CMG_193	-0.9061	0.0599	-1.0235	-0.7888
CMG_194	-1.2674	0.0601	-1.3851	-1.1497
CMG_195	-0.6092	0.0627	-0.732	-0.4863
CMG_196	-1.3025	0.0593	-1.4187	-1.1863
CMG_197	-1.3995	0.093	-1.5818	-1.2172
CMG_198	-1.5106	0.0766	-1.6607	-1.3605
CMG_199	-1.2042	0.0701	-1.3416	-1.0669
CMG_200	-1.4525	0.0614	-1.5728	-1.3323
CMG_201	-0.9457	0.0643	-1.0716	-0.8197
CMG_202	-1.7923	0.0597	-1.9093	-1.6753
CMG_203	-1.2352	0.0601	-1.3531	-1.1173
CMG_204	-1.7436	0.0606	-1.8623	-1.6249
CMG_205	-1.8498	0.0602	-1.9679	-1.7318
CMG_206	-1.8428	0.0653	-1.9708	-1.7148
CMG_207	-1.3232	0.0596	-1.44	-1.2063
CMG_208	-2.0491	0.0593	-2.1653	-1.9329
CMG_209	-1.5563	0.0602	-1.6744	-1.4383
CMG_210	-1.3665	0.0897	-1.5424	-1.1906
CMG_211	-1.4285	0.0678	-1.5615	-1.2956
CMG_212	-1.5449	0.0854	-1.7122	-1.3776
CMG_213	-1.5706	0.0666	-1.7011	-1.44
OOhCath_Y	-0.0642	0.014	-0.0916	-0.0368
OOhImplant_Y	-0.454	0.0327	-0.5181	-0.3898
OOhPtca_Y	-0.3802	0.0144	-0.4085	-0.3519
HID_51406	0.5013	0.0328	0.437	0.5655

Continued on next page

Table D.9 – continued Significant HEREM Beta

Covariate	Estimate	SE	95% LB	95% UB
HID_51423	-0.1819	0.0368	-0.254	-0.1098
HID_51597	-0.3209	0.0589	-0.4364	-0.2055
HID_51983	-0.282	0.0363	-0.3531	-0.2109
HID_51994	-0.4031	0.0413	-0.484	-0.3221
HID_52003	-0.2706	0.0353	-0.3397	-0.2015
HID_52038	-0.0733	0.0289	-0.1299	-0.0166
HID_52046	-0.1803	0.033	-0.2451	-0.1156
HID_53850	0.1935	0.0286	0.1373	0.2496
HID_53910	0.1466	0.0284	0.0909	0.2022
HID_53936	-0.084	0.0288	-0.1404	-0.0275
HID_53992	-0.3386	0.0432	-0.4232	-0.254
HID_53994	0.2065	0.0455	0.1173	0.2957
HID_54048	-0.11	0.0385	-0.1855	-0.0346
HID_80015	0.2329	0.0816	0.073	0.3929
HID_80016	0.286	0.0285	0.2301	0.3419
HID_80020	0.3792	0.0312	0.3181	0.4403
HID_80033	-0.474	0.1443	-0.7569	-0.191
HID_80041	0.3953	0.0357	0.3255	0.4652
HID_80042	0.3292	0.0315	0.2675	0.3909
HID_80043	0.2599	0.0311	0.1991	0.3208
HID_80044	0.301	0.0291	0.2441	0.358
HID_80052	-0.2788	0.0654	-0.407	-0.1506
HID_80070	-0.104	0.0527	-0.2073	-8.00E-04
HID_80120	0.1765	0.0338	0.1103	0.2428
HID_80122	-0.5109	0.0872	-0.6818	-0.3399
HID_80148	0.2661	0.0313	0.2048	0.3274
HID_90102	0.2495	0.0296	0.1915	0.3075
HID_90109	0.196	0.0295	0.1383	0.2538

Continued on next page

Table D.9 – continued Significant HEREM Beta

Covariate	Estimate	SE	95% LB	95% UB
FIcnt2_1	-0.102	0.0142	-0.1299	-0.0742
FIcnt3_1	-0.2742	0.0302	-0.3334	-0.2151
FsYr_2008	0.0706	0.015	0.0413	0.1
fiBio_1	0.4695	0.0392	0.3927	0.5463
fiCel_1	0.1588	0.0394	0.0816	0.2359
fiCmo_1	0.4455	0.1157	0.2188	0.6723
fiCrd_1	0.2262	0.0342	0.1591	0.2933
fiDia_1	0.3432	0.0268	0.2907	0.3957
fiEnd_1	0.4543	0.0398	0.3763	0.5323
fiHrt_1	0.1282	0.0342	0.0613	0.1952
fiMeG96_1	0.6999	0.0388	0.6238	0.776
fiMeL96_1	0.4764	0.0323	0.4131	0.5398
fiNut_1	0.3117	0.053	0.2078	0.4156
fiPar_1	0.33	0.0664	0.1998	0.4601
fiPlr_1	0.3842	0.0278	0.3298	0.4386
fiRad_1	0.4234	0.1485	0.1324	0.7145
fiTra_1	0.5452	0.0616	0.4244	0.666
fiTub_1	0.5564	0.0495	0.4592	0.6535
fiVad_1	0.4684	0.0296	0.4104	0.5265

Table D.10: Insignificant HEREM Beta (Mean Estimate)

Covariate	Estimate	SE	95% LB	95% UB
CMG_177	-0.0781	0.0528	-0.1816	0.0254
CMG_182	0.0098	0.0332	-0.0552	0.0748
HID_51199	0.0163	0.0499	-0.0815	0.114
HID_51213	0.0043	0.0814	-0.1552	0.1638

Continued on next page

Table D.10 – continued Insignificant HEREM Beta

Covariate	Estimate	SE	95% LB	95% UB
HID_51444	-0.0207	0.0303	-0.0802	0.0387
HID_51748	-0.0366	0.0583	-0.1509	0.0777
HID_51754	-0.0073	0.0378	-0.0813	0.0667
HID_51982	-0.0417	0.0284	-0.0974	0.0141
HID_53917	0.052	0.0286	-0.0041	0.1081
HID_53932	0.0159	0.0321	-0.047	0.0789
HID_53988	0	0.0346	-0.0678	0.0677
HID_80150	0.0272	0.0777	-0.1251	0.1796
HID_90130	0.0238	0.0346	-0.044	0.0916
HID_90136	-0.0228	0.037	-0.0954	0.0498

Table D.11: Significant HEREM Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
Intercept	-0.4475	0.0789	-0.6021	-0.293
IE_3	0.2375	0.078	0.0847	0.3903
CmL_1	0.0441	0.0155	0.0137	0.0746
CMG_160	-1.2108	0.1537	-1.5121	-0.9095
CMG_161	-0.5635	0.068	-0.6967	-0.4303
CMG_162	-2.0173	0.0641	-2.143	-1.8916
CMG_163	-1.4516	0.0728	-1.5944	-1.3089
CMG_164	-0.4987	0.0974	-0.6896	-0.3079
CMG_165	-1.6616	0.0926	-1.8432	-1.4801
CMG_166	-1.7607	0.0863	-1.9299	-1.5915
CMG_167	-1.9867	0.1226	-2.227	-1.7465
CMG_168	-1.929	0.0861	-2.0978	-1.7601
CMG_169	-1.9707	0.1216	-2.209	-1.7324

Continued on next page

Table D.11 – continued Significant HEREM Gamma

Covariate	Estimate	SE	95% LB	95% UB
CMG_170	-1.8506	0.0945	-2.0358	-1.6653
CMG_171	-2.0474	0.1717	-2.3839	-1.711
CMG_172	-2.1741	0.063	-2.2975	-2.0507
CMG_173	-0.7087	0.1068	-0.9181	-0.4993
CMG_174	-0.6462	0.0614	-0.7664	-0.5259
CMG_175	-1.2808	0.0588	-1.396	-1.1656
CMG_176	-1.1518	0.0588	-1.2671	-1.0365
CMG_177	0.3318	0.0918	0.1519	0.5118
CMG_178	-0.2395	0.0743	-0.3851	-0.0938
CMG_179	-0.4855	0.0629	-0.6087	-0.3623
CMG_180	-0.6886	0.0956	-0.8759	-0.5013
CMG_181	-1.1291	0.0643	-1.2551	-1.003
CMG_182	-0.7755	0.0667	-0.9062	-0.6448
CMG_184	-0.9541	0.2119	-1.3694	-0.5387
CMG_185	-0.2067	0.0666	-0.3373	-0.0761
CMG_193	-1.3359	0.0998	-1.5315	-1.1403
CMG_194	-0.4847	0.096	-0.6728	-0.2965
CMG_195	-1.1373	0.1094	-1.3517	-0.9228
CMG_196	-0.3047	0.0939	-0.4888	-0.1206
CMG_199	-0.2967	0.1079	-0.5083	-0.0852
CMG_200	-0.3956	0.0982	-0.5881	-0.2031
CMG_201	-0.9561	0.1137	-1.1789	-0.7334
CMG_203	-0.8152	0.098	-1.0072	-0.6232
CMG_204	-0.3079	0.0974	-0.4989	-0.1169
CMG_205	-0.3425	0.0969	-0.5325	-0.1525
CMG_207	-1.2024	0.0989	-1.3962	-1.0086
CMG_208	-0.3519	0.0949	-0.5379	-0.1659
CMG_211	-0.2522	0.1121	-0.4718	-0.0325

Continued on next page

Table D.11 – continued Significant HEREM Gamma

Covariate	Estimate	SE	95% LB	95% UB
OOhCath_Y	0.3506	0.0368	0.2785	0.4227
OOhImplant_Y	0.3015	0.075	0.1544	0.4486
OOhPtca_Y	0.3692	0.0402	0.2904	0.448
HID_51199	-1.1923	0.1511	-1.4884	-0.8961
HID_51213	-0.8048	0.2232	-1.2423	-0.3673
HID_51406	-0.2403	0.0687	-0.375	-0.1056
HID_51423	0.3831	0.0645	0.2568	0.5095
HID_51444	0.8523	0.0595	0.7356	0.969
HID_51597	-0.6235	0.1446	-0.9069	-0.3401
HID_51748	-0.8403	0.1572	-1.1484	-0.5322
HID_51754	-0.8155	0.0927	-0.9972	-0.6338
HID_51994	0.4379	0.0732	0.2946	0.5813
HID_52003	0.1881	0.0695	0.0519	0.3243
HID_53850	-0.2781	0.0613	-0.3983	-0.1579
HID_53910	-0.3953	0.0584	-0.5097	-0.2808
HID_53917	-0.4617	0.0588	-0.5769	-0.3465
HID_53932	-0.2545	0.063	-0.3779	-0.1311
HID_53936	-0.1496	0.0622	-0.2715	-0.0278
HID_53988	-0.6134	0.0773	-0.7648	-0.4619
HID_53992	-0.8182	0.1079	-1.0297	-0.6068
HID_53994	-0.4198	0.1014	-0.6186	-0.221
HID_80016	-0.5287	0.0589	-0.6442	-0.4132
HID_80020	-0.4497	0.062	-0.5713	-0.3282
HID_80041	-0.3223	0.0707	-0.4609	-0.1837
HID_80042	-0.404	0.0645	-0.5303	-0.2776
HID_80043	-0.2839	0.0633	-0.408	-0.1599
HID_80044	-0.1937	0.0607	-0.3126	-0.0747
HID_80120	-0.3905	0.0689	-0.5256	-0.2554

Continued on next page

Table D.11 – continued Significant HEREM Gamma

Covariate	Estimate	SE	95% LB	95% UB
HID_80148	-0.3097	0.0627	-0.4325	-0.1868
HID_90102	-0.3117	0.0691	-0.4472	-0.1762
HID_90109	-0.2088	0.062	-0.3303	-0.0873
HID_90136	-0.2739	0.0733	-0.4176	-0.1301
FIcnt2_1	0.2626	0.0402	0.1838	0.3414
FIcnt3_1	0.5495	0.0808	0.3912	0.7078
fiBio_1	-0.3422	0.054	-0.448	-0.2363
fiCrd_1	0.0969	0.0415	0.0156	0.1783
fiEnd_1	-0.2048	0.0582	-0.3189	-0.0907
fiMeG96_1	0.234	0.0646	0.1074	0.3605
fiMeL96_1	-0.069	0.0291	-0.1259	-0.012
fiTra_1	0.4534	0.1282	0.2022	0.7046
fiTub_1	-0.2704	0.125	-0.5155	-0.0253
fiVad_1	-0.0819	0.0384	-0.1571	-0.0066

Table D.12: Insignificant HEREM Gamma (Variance Estimate)

Covariate	Estimate	SE	95% LB	95% UB
IE_2	0.0274	0.0293	-0.03	0.0848
CmL_2	-0.0321	0.035	-0.1007	0.0364
CmL_3	-0.0197	0.0437	-0.1053	0.0659
CmL_4	-0.0239	0.0674	-0.1559	0.1082
CMG_183	-0.023	0.1459	-0.3089	0.263
CMG_197	-0.2886	0.1479	-0.5785	0.0013
CMG_198	-0.2341	0.1344	-0.4975	0.0293
CMG_202	-0.1853	0.0948	-0.3711	4.00E-04
CMG_206	-0.0934	0.1053	-0.2999	0.113

Continued on next page

Table D.12 – continued Insignificant HEREM Gamma

Covariate	Estimate	SE	95% LB	95% UB
CMG_209	-0.1083	0.0952	-0.2949	0.0783
CMG_210	0.1874	0.1379	-0.0829	0.4577
CMG_212	-0.0143	0.1348	-0.2785	0.2499
CMG_213	0.0518	0.1045	-0.153	0.2566
HID_51982	-0.055	0.061	-0.1746	0.0645
HID_51983	0.0813	0.0713	-0.0585	0.221
HID_52038	-0.081	0.064	-0.2064	0.0444
HID_52046	-0.0009	0.0675	-0.1331	0.1314
HID_54048	0.114	0.0746	-0.0322	0.2602
HID_80015	0.0155	0.1437	-0.2661	0.2971
HID_80033	0.2556	0.2392	-0.2133	0.7245
HID_80052	-0.0664	0.1216	-0.3048	0.172
HID_80070	-0.0988	0.0967	-0.2883	0.0907
HID_80122	0.0119	0.1565	-0.2948	0.3186
HID_80150	-0.0276	0.1405	-0.3031	0.2478
HID_90130	-0.0129	0.065	-0.1404	0.1145
FsYr_2008	0.0115	0.0112	-0.0105	0.0334
fiCel_1	-0.0235	0.0426	-0.1069	0.0599
fiCmo_1	0.3046	0.1946	-0.0768	0.6859
fiDia_1	-0.0286	0.0381	-0.1031	0.046
fiHrt_1	0.036	0.1057	-0.1712	0.2432
fiNut_1	0.0033	0.097	-0.1867	0.1934
fiPar_1	-0.028	0.1211	-0.2653	0.2094
fiPlr_1	0.0196	0.0475	-0.0734	0.1126
fiRad_1	0.3978	0.2475	-0.0872	0.8828

D.4 GLM Results

This section will detail the significant and insignificant mean of main effects of chapter 5, section 5.7

Table D.13: Significant GLM Beta

Covariate	Estimate	SE	t-value	p-value
Intercept	9.1387	0.0445	205.3137	0.0000
IE_2	0.5997	0.0780	7.6861	0.0000
IE_3	0.2090	0.0549	3.8091	0.0001
CmL_1	0.5299	0.0465	11.3955	0.0000
CmL_2	0.1443	0.0460	3.1359	0.0017
CmL_3	0.1631	0.0458	3.5622	0.0004
CmL_4	0.2793	0.0373	7.4892	0.0000
CMG_160	1.3851	0.1854	7.4719	0.0000
CMG_161	0.9965	0.0346	28.7682	0.0000
CMG_162	0.8636	0.0381	22.6675	0.0000
CMG_163	0.5689	0.0472	12.0503	0.0000
CMG_164	0.1695	0.0559	3.0296	0.0024
CMG_165	0.5710	0.0691	8.2578	0.0000
CMG_166	1.0956	0.1092	10.0338	0.0000
CMG_167	0.7946	0.3097	2.5659	0.0103
CMG_168	0.8764	0.0761	11.5102	0.0000
CMG_169	0.7878	0.1513	5.2051	0.0000
CMG_170	0.7480	0.1487	5.0297	0.0000
CMG_171	0.8432	0.4149	2.0323	0.0421
CMG_172	0.5699	0.0356	15.9950	0.0000
CMG_173	0.1546	0.0586	2.6403	0.0083
CMG_174	0.2110	0.0316	6.6748	0.0000
CMG_176	-0.2688	0.0302	-8.9139	0.0000

Continued on next page

Table D.13 – continued Significant GLM Beta

Covariate	Estimate	SE	t-value	p-value
CMG_177	0.1105	0.0414	2.6723	0.0075
CMG_179	-0.6334	0.0316	-20.0168	0.0000
CMG_180	0.4714	0.0536	8.7952	0.0000
CMG_181	0.6653	0.0333	19.9824	0.0000
CMG_183	-0.2257	0.0725	-3.1128	0.0019
CMG_184	-1.2248	0.0894	-13.7029	0.0000
CMG_185	-0.1221	0.0342	-3.5675	0.0004
CMG_193	-1.1053	0.0963	-11.4725	0.0000
CMG_194	-1.3629	0.0958	-14.2239	0.0000
CMG_195	-0.7958	0.0990	-8.0370	0.0000
CMG_196	-1.3632	0.0952	-14.3129	0.0000
CMG_197	-1.4245	0.1166	-12.2213	0.0000
CMG_198	-1.5559	0.1028	-15.1343	0.0000
CMG_199	-1.2159	0.0995	-12.2169	0.0000
CMG_200	-1.5194	0.0964	-15.7582	0.0000
CMG_201	-1.0897	0.0998	-10.9180	0.0000
CMG_202	-1.8420	0.0954	-19.3089	0.0000
CMG_203	-1.3959	0.0959	-14.5591	0.0000
CMG_204	-1.8238	0.0959	-19.0224	0.0000
CMG_205	-1.9390	0.0957	-20.2565	0.0000
CMG_206	-1.8554	0.0976	-19.0201	0.0000
CMG_207	-1.5161	0.0960	-15.7929	0.0000
CMG_208	-2.1496	0.0953	-22.5589	0.0000
CMG_209	-1.5802	0.0956	-16.5378	0.0000
CMG_210	-1.2428	0.1058	-11.7425	0.0000
CMG_211	-1.4626	0.0995	-14.7029	0.0000
CMG_212	-1.4659	0.1058	-13.8608	0.0000
CMG_213	-1.5162	0.0976	-15.5321	0.0000

Continued on next page

Table D.13 – continued Significant GLM Beta

Covariate	Estimate	SE	t-value	p-value
OOhCath_Y	-0.0329	0.0155	-2.1153	0.0344
OOhImplant_Y	-0.3617	0.0318	-11.3901	0.0000
OOhPtca_Y	-0.3624	0.0170	-21.2536	0.0000
HID_51199	-0.2201	0.0667	-3.3019	0.0010
HID_51213	-0.1908	0.0947	-2.0149	0.0439
HID_51406	0.3999	0.0405	9.8628	0.0000
HID_51423	-0.1189	0.0389	-3.0606	0.0022
HID_51597	-0.4891	0.0641	-7.6262	0.0000
HID_51748	-0.2026	0.0690	-2.9344	0.0033
HID_51754	-0.1915	0.0444	-4.3162	0.0000
HID_51982	-0.1880	0.0333	-5.6524	0.0000
HID_51983	-0.3427	0.0367	-9.3332	0.0000
HID_51994	-0.3429	0.0374	-9.1759	0.0000
HID_52003	-0.3118	0.0361	-8.6342	0.0000
HID_52038	-0.1499	0.0342	-4.3756	0.0000
HID_52046	-0.2477	0.0354	-6.9957	0.0000
HID_53936	-0.2020	0.0337	-5.9996	0.0000
HID_53988	-0.1337	0.0405	-3.2996	0.0010
HID_53992	-0.4890	0.0514	-9.5174	0.0000
HID_54048	-0.1596	0.0397	-4.0248	0.0001
HID_80015	0.1722	0.0836	2.0605	0.0394
HID_80016	0.1952	0.0353	5.5243	0.0000
HID_80020	0.2825	0.0372	7.6022	0.0000
HID_80033	-0.4146	0.0987	-4.1987	0.0000
HID_80041	0.2736	0.0432	6.3373	0.0000
HID_80042	0.2289	0.0394	5.8071	0.0000
HID_80043	0.1886	0.0380	4.9661	0.0000
HID_80044	0.2019	0.0363	5.5535	0.0000

Continued on next page

Table D.13 – continued Significant GLM Beta

Covariate	Estimate	SE	t-value	p-value
HID_80052	-0.3797	0.0697	-5.4449	0.0000
HID_80070	-0.1521	0.0580	-2.6215	0.0088
HID_80120	0.0994	0.0421	2.3613	0.0182
HID_80122	-0.5292	0.0810	-6.5323	0.0000
HID_80148	0.1763	0.0376	4.6910	0.0000
HID_90102	0.2146	0.0378	5.6719	0.0000
HID_90109	0.1623	0.0371	4.3722	0.0000
FIcnt2_1	-0.1694	0.0199	-8.4981	0.0000
FIcnt3_1	-0.4367	0.0404	-10.8107	0.0000
FsYr_2008	0.1084	0.0454	2.3883	0.0169
fiBio_1	0.4369	0.0519	8.4150	0.0000
fiCel_1	0.2445	0.0640	3.8229	0.0001
fiCmo_1	0.5636	0.0811	6.9504	0.0000
fiCrd_1	0.3115	0.0523	5.9559	0.0000
fiDia_1	0.3748	0.0335	11.1916	0.0000
fiEnd_1	0.5715	0.0498	11.4716	0.0000
fiHrt_1	0.1513	0.0454	3.3326	0.0009
fiMeG96_1	0.9063	0.0437	20.7521	0.0000
fiMeL96_1	0.6912	0.0432	15.9841	0.0000
fiNut_1	0.4409	0.0650	6.7884	0.0000
fiPar_1	0.4872	0.0736	6.6176	0.0000
fiPlr_1	0.4996	0.0382	13.0748	0.0000
fiRad_1	0.5083	0.1016	5.0023	0.0000
fiTra_1	0.5916	0.0662	8.9411	0.0000
fiTub_1	0.6170	0.0619	9.9757	0.0000
fiVad_1	0.6232	0.0414	15.0470	0.0000