

Robust instrumental variables and accelerated life regressions

by

Anand Acharya

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in

partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Economics

Carleton University

Ottawa, Ontario

©2016

Anand Acharya

Abstract

This thesis considers the econometric problem of endogeneity in an accelerated life regression model. The proposed instrumental variables inference, based on inverting a pivotal statistic, is exact regardless of instrument quality. A (i) least squares statistic and (ii) distribution-free linear rank statistic allowing censoring are provided. A simulation confirms that the quality of exogenous variation determines an instrument's informative content. An original prospectively collected observational data set provides an empirical illustration, in which, the trauma status of a pediatric critical care patient instruments a possibly confounded illness severity index in a length of stay regression for a specific pediatric intensive care population. Results suggest a clinically relevant bias correction for routinely collected patient risk indices that is meaningful for informing policy in the health care setting.

Acknowledgements

Very rarely does one have the opportunity to pursue two life dreams. This thesis marks the culmination of an incredible personal journey that has allowed me to interact and learn from some amazing people without whom it would never have materialized.

The first and foremost person that I must thank is my supervisor, Lynda Khalaf. From the first course she taught me to my thesis completion, she never waived in her encouragement, her dedication and her belief that I could actually do this. I am convinced that her attention to detail and support for her graduate students is second to none.

Next, I wish to thank my co-supervisor Marcel Voia and the rest of my thesis committee, Gennady Shaikhet and Tim Ramsay. They listened to my ideas, encouraged me through my workshops and provided insightful guidance in developing original work. It was in Marcel's course that I had the opportunity to begin analyzing the clinical data set which has ultimately led to this thesis.

Last, but definitely not least, I would like to thank my amazing wife, Kusum and our two daughters, Nina and Jasmine. It is with their loving support, good humour, and encouragement that a dream became reality.

Once again, I am truly grateful for having had the opportunity to explore the field of econometrics and even more grateful to the people who made it possible.

Contents

1	Introduction	6
2	The accelerated life regression	8
2.1	Preliminaries: Characterizing event times	10
2.1.1	Outcome data structure and description	10
2.1.2	Parametric forms	12
2.2	Framework: Regression methods with time to event data . . .	15
2.2.1	Relative Risk	16
2.2.2	Accelerated life	18
2.2.3	Transformation models	19
2.3	Complications: Frailty, endogeneity and controls	23
2.3.1	Controls	26
2.3.2	Endogenous censoring	28

2.4	Assumptions	29
2.4.1	Specification and data structure	29
2.4.2	Censoring	29
2.4.3	Distributions	30
3	Endogenous accelerated life regression	31
3.1	Weak instruments and identification robustness	32
3.1.1	Linear instrumental variables	32
3.1.2	Weak instruments	33
3.1.3	Valid confidence sets	35
3.1.4	Anderson-Rubin procedure	37
3.2	Least Squares Inference	38
3.2.1	Simulation methods for location-scale family distributions	42
3.3	Rank inference	44
3.3.1	Linear rank statistic	45
3.3.2	Right-censoring	49
3.4	Simulation	53
3.4.1	Study design	55
3.4.2	Results	57

3.4.3	Discussion	59
4	Clinical inference in pediatric acute care	60
4.1	Pediatric intensive care unit length of stay	63
4.2	Pediatric illness severity	66
4.3	Pediatric trauma	69
4.4	Methods	73
4.5	Analysis, results and discussion	74
4.6	Conclusion	83

Chapter 1

Introduction

The aim of this thesis is to present an econometric methodology with associated biostatistical techniques to address the problem of unmeasured confounding or endogeneity in an accelerated life regression model. Motivation for this problem originates in an empirically relevant clinical study in the pediatric intensive care unit.

While any method hinges on assumptions, I focus on maintaining assumptions that are plausible or compatible with the specific data and environment relevant to the clinical analysis at hand. So, although this is not intended to be a unifying theory, it has provided a meaningful solution to a statistical problem of considerable empirical relevance, where one previously was not

known to exist.

Demonstrating the use of instrumental variables to overcome the complications of confounded interventions or exposures, with time to event outcomes, is the primary innovation of this thesis. Indeed, there appear to be limited examples of the use of instrumental variables in the duration literature. Moreover the specific instrumental variables techniques presented in this thesis draw from an active area of econometric research on robust inferential methods, which to date has not considered models of duration outcomes, with its unique complications.

Chapter 2

The accelerated life regression

The special relevance of time to event outcomes is well recognized in the fields of econometrics, statistics and reliability. In this thesis specific attention is focused on two particular aspects of the time origin of the model that complicate inference, (i) the inherent non-normality of the distribution of event times, and (ii) censoring, namely right-censoring. Accounting for these aspects has resulted in model specification and related estimators that are sufficiently distinct to produce a unique and dedicated literature. Within this literature, to date, there has been relatively less attention to the complications induced by unmeasured confounders or endogenous predictors. Accordingly, the primary aim of this thesis is directed on addressing

this problem with appropriately robust econometric techniques.

Although an appreciable amount of empirical duration analysis employs fully non-parametric methods such as Kaplan-Meier estimation (Kaplan & Meier, 1958), regression techniques remain a central modelling tool, particularly in policy relevant analysis. Two prominent paradigms are the (i) relative risk and (ii) accelerated life specifications, with the former being predominant, in part due to the overwhelming popularity of the Cox proportional hazard model. Indeed, Cox (1972) is a profound contribution to the field of statistics and has, as Reid (1994) suggested, spawned its own “cottage industry” (Reid (1994) pg. 440). Nevertheless, there are numerous desirable attributes to the overlooked accelerated life specification (Wei, 1992), (Kalbfleisch & Prentice, 2002). The popularity of the Cox model may to some degree be the ease of implementation with censoring and perceived parsimony in terms of minimal assumptions, particularly the absence of any distributional stance. However, the novel methods presented in this thesis provide an equivalently or comparably parsimonious alternative with the accelerated life specification, and most importantly, address some of the arguably under appreciated limitations of the Cox model in the policy relevant context (Aalen, Cook & Røysland, 2015).

2.1 Preliminaries: Characterizing event times

2.1.1 Outcome data structure and description

The primary outcome data of interest present as a single-spell time to event on each of the $i = 1..n$ individuals under study, denoted by the random variable T with realizations t_i . All random variables are suitably defined over the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The continuous random variable T is possibly imperfectly observed due to right censoring such that $t^* = \min(\tau, t)$, where τ is censored time. The presence of the latent t is recorded by the censoring indicator, d . There are a number of schemes under which censoring may arise (Klein & Moeschberger, 2003). It may simply be, as is often the case in clinical studies, that the study period has ended while some individuals have yet to experience the study event. Alternatively, an adverse event, possibly as extreme as death may result in the patient's premature removal from the study or the patient may be lost to follow up. A crucial assumption is whether or not the data generating process for the censoring indicator d , is independently distributed over the $i = 1..n$ individuals under study. The methods presented in this thesis relax this assumption, permitting a form of endogenous censoring as defined in §2.3.1, which is particularly relevant to

the clinical application in §4.

For the particular features of time to event analysis, there are a number of useful, mathematically equivalent functions variously used in analyzing the distribution of the the continuous non-negative random variable T . For a detailed summary see Cameron & Trivedi (2005). The regression models considered here are conveniently derived from either the survivor function $S(t) = \mathbb{P}(T > t)$, or the hazard function,

$$h(t) = \lim_{\Delta \rightarrow 0} \frac{\mathbb{P}[t \leq T < t + \Delta \mid T \geq t]}{\Delta}. \quad (2.1)$$

The hazard function, also referred to as the intensity function, force of transition, or failure rate, is widely accepted as the object or construct of interest in much applied work in duration analysis. An important distinction is that the hazard emphasizes conditional probabilities in comparison to the survivor function. There are a number of reasons for considering the hazard including the development of the risk process, ease of sub group comparisons, and ability to handle censoring (Cox & Oakes, 1984). However, the concept of instantaneous risk is possibly elusive and arguably raises difficulties when employed in a regression framework, particularly when causal interpretation is of interest (Aalen, Cook & Røysland, 2015). For example, the observed shape of the hazard may arise from changes within individuals

or changes between individuals without the ability to discriminate between the two sources.

As detailed in Cameron & Trivedi (2005) there are some convenient relationships between the distribution $F(t) = \mathbb{P}(T \leq t)$, survivor $S(t)$, density $f(t)$, hazard $h(t)$, cumulative or integrated hazard $H(t) = \int_0^t h(s)ds$ and quantile $Q(t) = F^{-1}(t)$ functions for any underlying distributional form,

$$S(t) = 1 - F(t) \quad h(t) = f(t)/S(t) \quad H(t) = -\ln S(t).$$

These definitions and relationships permit the development of model specification, censoring and computational methods.

2.1.2 Parametric forms

There are various special, completely specified, parametric failure time distributions of the random variable T that are useful for analysis of an assumed homogenous population of individuals under study. In this thesis, we consider three particular distributions, (i) the log-normal, (ii) the log-logistic, (iii) the Weibull, which includes the exponential as a special case. These distributions are the labels for each of the parametric time to event regression models to be discussed in §2.2.

The log-normal model has survivor and hazard function,

$$S(t) = 1 - \Phi\left(\frac{\ln t - \delta}{\sigma}\right) \quad \text{and} \quad h(t) = \frac{\exp(-(\ln t - \delta)^2/2\sigma^2)}{t\sigma\sqrt{t\pi}[1 - \Phi(\frac{\ln t - \delta}{\sigma})]},$$

where $\delta \in \mathbb{R}^+$ and $\sigma \in \mathbb{R}^+$ are the respective location and scale parameters. Although this specification is rarely utilized, in part due to the unusual hazard function, from a pedagogic perspective the log-normal model provides a convenient basis for extending the many innovations of linear normal theory to time to event models. Alternative and historic interest in this specification originates in reliability analysis of wear time for maintainable systems.

The log-logistic model with survivor and hazard,

$$S(t) = \frac{1}{1 + (\delta t)^p}, \quad h(t) = \frac{p\delta^p t^{p-1}}{1 + (\delta t)^p},$$

in which $\delta \in \mathbb{R}^+$ and $p \in \mathbb{R}^+$ are the respective location and scale parameters, is similar to the log-normal distribution, but with heavier tails. The log-logistic model is well suited to some particular common clinical survival analysis, for example modelling length of survival after cancer diagnosis, where the event rate increases initially and then decreases after diagnosis. The model is convenient for statistical inference and has an explicit, reasonably simple form for the survivor and hazard functions (Cox & Oakes, 1984).

The Weibull distribution of T has functional form,

$$S(t) = \exp(-\delta t^p) \quad h(t) = \delta p t^{p-1},$$

where $\delta \in \mathbb{R}^+$ and $p \in \mathbb{R}^+$ are the respective location and shape parameters.

The Weibull model offers the greatest flexibility in terms of the shape of the hazard and accordingly adapts well to a large number of differing survival experiences. As a special case, the exponential model arises when the shape parameter $p = 1$,

$$S(t) = \exp(-\delta t) \quad h(t) = \delta.$$

Naturally from the hazard modelling perspective a constant hazard facilitates a conceptually simple comparison of sub-groups (Cox & Oakes, 1984). Moreover, interest in the exponential model is motivated by its central role in the theory of point process (Cox & Isham, 1980) and counting process, with related Poisson regression (Aalen, 1989). Moreover, historically, the exponential model is possibly the first application directed to time to event outcomes and is particularly well suited to describing various idealized models in reliability studies.

Although the univariate duration outcome is interesting to model, the goal of much applied work is to relate a treatment, intervention, or exposure

to the time to event. In the conceptually simplest of cases, for example an indicator of treated or not treated, the comparing of survival curves has become standard practice. Indeed, Kaplan & Meier (1958) remains the most cited paper in the overall field of statistics. There is a large literature extending this fully non-parametric approach to comparing the survival experience of multiple sub-groups under study. However, study design may dictate the use and availability of other observables. This, of course is particularly pertinent in cases where the intervention or exposure of interest is possibly confounded, including the randomized study design. Naturally, a first step in overcoming the complications of a heterogenous population is to employ regression methods, readily admitting all observable information available for the given study design.

2.2 Framework: Regression methods with time to event data

The n -vector of possibly right-censored time to event outcomes $t \in \mathbb{R}^+$ is often observed concurrently with $(n \times g)$ potentially confounded interventions or exposures Y , further $(n \times k_1)$ concomitant variables X_1 , and other $(n \times k_2)$

instrumental variables X_2 .

From both an econometric and statistical perspective, a desirable model would be a specification using all available information in computing $S(t|Y, X_1) = \mathbb{P}(T > t|Y, X_1)$ for all possible Y, X_1 . If indeed other informative variables X_1 are observed, along with an intervention or exposure of interest Y , a regression model is commonly derived from the conditional survivor function $S(t|Y, X_1)$, typically following either the (i) relative risk or (ii) accelerated life paradigms. We define $\phi(Y, X_1; \beta, \delta)$ as the unspecified regression function. For a given paradigm, interest is focused on the appropriate estimators and subsequent interpretation for the regression coefficients $\beta, \delta \in \mathbb{R}$ associated, respectively with Y, X_1 .

2.2.1 Relative Risk

The most well know paradigm is the relative risk regression, where the fundamental underlying assumption is that the effect of the covariates X_1, Y is multiplicative (proportional) with respect to the unconditional baseline hazard function, $h_o(t)$,

$$h(t|Y, X_1; \beta, \delta) = h_o(t)\phi(Y, X_1; \beta, \delta), \quad (2.2)$$

which has the equivalent expression, in terms of the conditional survivor

function,

$$S(t|Y, X_1; \beta, \delta) = S_o(t)^{\phi(Y, X_1; \beta, \delta)}, \quad (2.3)$$

where $S_o(t)$ is the unconditional baseline survivor function.

Under the relative risk formulation, the interpretation arising from the resulting estimators for the parameters β and δ are in terms of the hazard function, or more specifically, the hazard ratio which is the *de facto* basis of the regression model. This is in contrast with a classical regression approach, where the conditional expectation function is the primary modelling construct.

The Cox regression leaves the distributional form of the hazard unspecified by utilizing the hazard ratio as the basis of the regression model. It provides a summary of the exposure or treatment effect in terms of the hazard ratio and is also viewed as a form of relative risk. The Cox regression model is the predominant specification in biostatistics and econometrics for time to event analysis. However, despite its widespread popularity, there are some possibly under appreciated limitations of the Cox model. In particular, proportionality to the hazard is not maintained when either covariates are deleted, omitted or measured with error. In the case of randomized study design, Aalen, Cook & Røysland (2015) prove that it is mathematically not

permissible to maintain randomization between survivors beyond period one. Accordingly it is “...unclear what the hazard ratio computed for a randomized survival study really means.” (Aalen, Cook & Røysland (2015) pg 582).

2.2.2 Accelerated life

An alternative is the accelerated life regression, which assumes the effect of covariates is multiplicative (proportional) with respect to survival time.

$$S(t|Y, X_1; \beta, \delta) = S_o(t * \phi(Y, X_1; \beta, \delta)), \quad (2.4)$$

with an equivalent expression in terms of the hazard function,

$$h(t|Y, X_1; \beta, \delta) = h_o(t * \phi(Y, X_1; \beta, \delta))\phi(Y, X_1; \beta, \delta). \quad (2.5)$$

The function $\phi(Y, X_1; \beta, \delta)$ is referred to as an acceleration factor, with the intuitive appeal of rescaling time directly. The *de facto* basis of the regression model is now the time ratio. A commonly cited example (Kleinbaum & Klein, 2012) is comparing the lifetime of differing species, for example a canine’s life span is accelerated by a constant factor of seven at every point of the survival curve relative to a human’s, perhaps readily seen by observing a dog at two years of age displaying comparable maturity to a fourteen year old teenager.

It is also convenient or useful to consider a hypothetical unobserved random variable, U corresponding to an individual's time in state if all of their observed covariate values were equal to zero. This, of course implicitly assumes all individuals share the same uncertain (baseline) failure times, U and otherwise only differ in the observed covariate patterns. Indeed, the actual or observed survival time, T given the observed covariates would be generated by $T = \exp(Y\beta + X_1\delta) * U$, if $\phi(Y, X_1; \beta, \delta) = \exp(Y\beta + X_1\delta)$. In which case the parameters β and δ have a direct interpretation of a marginal increase or decrease in survival time as a function of the covariate values. In this special, convenient and arguably plausible case (Cox, 1972), the time ratio basis of the regression model is analogous to the conditional expectation function in the equivalent log-linear representation as described in what follows.

2.2.3 Transformation models

Both regression paradigms have a convenient linear representation in the special case $\phi(Y, X_1; \beta, \delta) = \exp(Y\beta + X_1\delta)$. A general expression of the transformation model specifies the triple, a transformation function $\Lambda(t)$, regression function $\phi(Y, X_1; \beta, \delta)$, and distribution with density $f(\epsilon)$,

$$\Lambda(t) = \delta_i t_n + Y\beta + X_1\delta + \epsilon. \quad (2.6)$$

Various forms of this model have been studied originally by Box & Cox (1964) and subsequently by Han (1987), Ridder (1990), Horowitz (1996), Chen (2002) and recently Chiappori, Komunjer & Kristensen (2015). The proportional hazard specification restricts the distribution of the error term, ϵ to be type 1 extreme value (double exponential) (Gumbel) and the transformation function to be $\ln(H_o(t))$, where the latter is the log integrated baseline hazard function. In contrast, the accelerated life specification restricts the transformation to be the observable function $\ln(t)$, but does not restrict the error distribution to a single distribution. In principal, both linear representations are instructive for extending innovations from linear normal theory (Ridder, 1990). However, an important consideration is the proportional hazard linear representation introduces a new unobserved dependent quantity, the integrated or cumulative hazard. Only in the special case of the Weibull model, to be discussed, is this quantity observable.

Different distributional assumptions give rise to well known parametric duration models detailed previously. A convenient feature of the transformation model is not only to provide an additively linear representation of the underlying nonlinear specifications, but to also result in tractable distributions on ϵ . In particular the log-normal, log-logistic, and Weibull models give

rise to ϵ following a normal, logistic, or Gumbel distribution respectively.

$$t = \exp(Y\beta + X_1\delta) * U, \quad (2.7)$$

$$\ln(t) = \delta_i \ln + Y\beta + X_1\delta + \sigma\epsilon, \quad (2.8)$$

- $U \sim \text{Lognormal}(\exp(\delta_i), \sigma^2) \rightarrow \epsilon \sim \text{Normal}(0, 1)$,
- $U \sim \text{Loglogistic}(\exp(\delta_i), \sigma) \rightarrow \epsilon \sim \text{Logistic}(0, 1)$,
- $U \sim \text{Weibull}(\exp(\delta_i), \frac{1}{\sigma}) \rightarrow \epsilon \sim \text{Gumbel}(0, 1)$,

where the log-normal location, log-logistic location, and Weibull scale parameters are respectively captured in the transformed regression intercept, δ_i . This class of model is also known as the location-scale specification.

The accelerated life specification in each of the above cases also have equivalent alternative interpretations of wide interest in other types of applications. The log-normal model, although not commonly utilized in duration analysis, given the unusual hazard, is of course, in the absence of censoring the standard linear normal theory model.

The log-logistic model has equivalent representation as the proportional survival odds, in which survival odds ratio is assumed to be constant, where the survival odds is,

$$\frac{S(t)}{1 - S(t)} = \frac{\mathbb{P}(T > t)}{\mathbb{P}(T \leq t)}.$$

The exponential model is closely related to the Poisson count data model. Of course, both distributions have an assumed constant rate. Although the models have different data structure, in which the Poisson model considers the number of events, whereas, the exponential model considers the time to event, in certain cases they will yield equivalent parameter estimates.

The Weibull model has the attractive, well know theoretical property of being both an accelerated life and proportional hazard model.

Some justification for making otherwise seemingly strong parametric assumptions are discussed succinctly in Cox & Oakes (1984) pg. 24. Although particular attention is drawn to the numerous useful methods of specification testing, it is worthwhile pointing out at this point, that techniques to be discussed here are shown to be robust to parametric misspecification.

Both specifications have a semi-parametric form. In the proportional hazard model, as discussed, an unspecified baseline hazard gives the Cox proportional hazard model. The accelerated life analogue leaves the error distribution completely unspecified (Wei, 1992).

From an empirical perspective, there may not be consensus on the practitioner's ease of interpretation of resulting parameter estimates from the two regression paradigms. For example Hosmer, Lemeshow & May (2008)

suggest that “research-orientated clinicians” are most comfortable with the hazard ratio parameter interpretation, hence its overwhelming popularity. Whereas, Kalbfleisch & Prentice (2002) and Wei (1992) suggest that the direct interpretation of parameter estimates arising from the time ratio are natural and easily understood, a view also attributed to D.R. Cox (Reid, 1994). Irrespective, the semi-parametric accelerated life specification offers a viable alternative to the predominant Cox regression.

2.3 Complications: Frailty, endogeneity and controls

A primary difficulty with the Cox regression arises from the fact that observable covariates may not fully account for the heterogeneous population under study. It is well known that the baseline hazard is confounded when individual differences are unaccounted or omitted. Accordingly there is a large and extensive literature addressing the identification and inference of unobserved heterogeneity, or frailty within the duration framework (Elbers, 1982), (Heckman, 1984). Much of this literature considers directly modelling a proportional individual frailty term in the proportional hazard specification,

known as the mixed proportional hazard model (Lancaster, 1979), (Vaupel, Manton & Stallard, 1979), (Hougaard, 1984),

$$h(t|Y, X_1) = \nu h_o(t)\phi(Y, X_1, \beta, \delta), \quad \text{or,} \quad S(t|Y, X_1) = S_o(t)^{\phi(Y, X_1, \beta, \delta)\nu}$$

where the hazard rate for each individual is now a product of an unobserved individual specific quantity ν_i , the assumed baseline hazard, and other observed characteristics. In the linear representation (2.8), the n -column vector ν enters as an additional additively separable unobserved term, resulting in a compound additive error. Interestingly, the presence of this additional unobserved variable does not similarly complicate either the accelerated life specification or the linear transformation of the relative risk specification to the same degree. This was originally noted by Aalen (1989) and Struthers & Kalbfleisch (1986). Based on this work, Keiding, Anderson & Klein (1997) suggest “Accelerated failure time modelling seems to avoid these difficulties and also yield easily interpretable results. We propose that it would be advantageous to upgrade the accelerated failure time approach alongside the hazard modelling approach to survival analysis.” (Keiding, Anderson & Klein (1997) pg. 215).

The multiplicative frailty term adds an additional separable source of randomness to model (2.8), which if not explicitly specified, would naturally

at the very least, result in the misspecification of the distributional assumption on the random variable ϵ . The view taken here is that this is almost always the case and in addition this unobserved individual specific effect may plausibly be correlated with the intervention or exposure of interest. With this in mind, the methods presented in §3, while not explicitly specifying or assuming the heterogeneity, are robust to the distributional misspecification thus provoked. This is in contrast to a direct application of likelihood based methods, as is well documented in the frailty literature.

The frailty term may also be viewed as an omitted or missing covariate. For example, a patient's individual frailty, perhaps weak, strong or anywhere on the continuum in-between is typically not observable. In duration models, as described above, this complicates inference, irrespective of any further correlation with the intervention, treatment, or exposure of interest. In a somewhat benign case, this complication may solely manifest as an efficiency loss. However, the more serious presence of any correlation between a hypothesized frailty term or any omitted covariate, and the exposure Y , defines the statistical complication of unmeasured confounding, or similar econometric complication of endogeneity, in this case due to an

omitted variable. Expressed in terms of a conditional expectation,

$$\mathbb{E}(\nu | Y) \neq 0. \tag{2.9}$$

This condition may consequently be expressed alternatively as a failure of a moment condition or lack of orthogonality.

The primary focus of the robust methods presented in this thesis is to overcome the complication of endogeneity provoked by frailty or an omitted variable. In this sense, the use of the instrumental variables, X_2 is both complementary to the techniques of direct frailty modelling and moreover, innovative in providing a solution to the additionally plausible complication of endogeneity or unmeasured confounding in a duration model.

2.3.1 Controls

Often, the role of the concomitant variables X_1 is to provide useful information on discerning between the unobservable continuum of individual types or other unobserved variation that would preclude meaningful inference on the intervention or exposure. “Controlling” for the otherwise uninteresting but nevertheless intervening individual characteristics is a logical and natural first step in study design, including randomized trials. Of course, the presence of control variables is most poignant in the observational study design.

Indeed, Stock (2010) summarizes the purpose and implications of the presence of the “controls” X_1 . Specifically he suggests the role of controls is to satisfy $E(\epsilon|Y, X_1) = E(\epsilon|X_1)$, which further implies that while $E(\epsilon|Y) = 0$, it would be the case that $E(\epsilon|X_1) \neq 0$. He points out that the control variable “proxies for deeper unmodeled effects remaining in the error term, so the control variable is correlated with the error term - indeed, that is why the control variable is included in the first place.” (Stock (2010) pg. 88). Accordingly, standard econometric practice has evolved to make a precise definition of the control variable, typically via the conditional mean independence assumption, which consequently precludes a causal interpretation on the coefficient of X_1 . However, if X_1 is sufficiently rich, such that $E(\epsilon|Y, X_1) = E(\epsilon|X_1)$ holds, the coefficient on the intervention Y has a causal interpretation. This has sharpened the focus on “measuring a single effect well instead of the vaguer goal of developing a model for the outcome.” (Stock (2010) pg. 88). The view taken here is, irrespective of whether or not the other observables X_1 are sufficiently rich to satisfy the role of controls, they will very likely be or at least following Stock’s heuristics, should assume to be correlated with the unspecified deeper unmodeled effects. Acknowledging and directly addressing this otherwise nuisance correlation, by providing suitably robust

inferential methods, is a purposeful contribution of this thesis.

2.3.2 Endogenous censoring

We have defined the dichotomous random variable d as the indicator for the right-censoring. When $d = 1$ the observed time t^* equals the censored time τ . We allow for censoring to be correlated with the unobserved type. In the notation used above, we define endogenous censoring as,

$$\mathbb{E}(\nu \mid d) \neq 0. \tag{2.10}$$

We view this as a form of selection bias, in that, for example, those on the “weaker” spectrum of types may have a systematically higher probability of prematurely exiting the study. This form of censoring is consistent with what appears in the empirical example of §4. This is in contrast with what would otherwise be assumed in the standard case where neither the censoring, controls, or intervention are believed to be correlated with the unobserved heterogeneity,

$$\mathbb{E}(\nu \mid d, X_1, Y) = 0. \tag{2.11}$$

2.4 Assumptions

Summarized are the key assumptions implied thus far by the preceding discussion. Some assumptions need not hold jointly or are, in fact mutually exclusive depending on the context, to be made explicit in the remainder as appropriate.

2.4.1 Specification and data structure

A 1 $\phi(Y, X_1, \beta, \delta) = \exp(Y\beta + X_1\delta)$

A 2 X_1, X_2 predetermined, or

A 3 X_2, ϵ pairwise stochastically independent.

A 4 (X_1, ϵ) independently distributed.

2.4.2 Censoring

C 1 d, ϵ independently distributed.

C 2 (d, ϵ) pairwise independently distributed (Endogenous censoring).

C 3 $t^* = \min(\tau, t)$ where τ is censored time and t need not be observed.

2.4.3 Distributions

D 1 ϵ *distribution unspecified.*

D 2 $\epsilon \stackrel{iid}{\sim} \text{Normal}(0, 1).$

D 3 $\epsilon \stackrel{iid}{\sim} \text{Logistic}(0, 1).$

D 4 $\epsilon \stackrel{iid}{\sim} \text{Gumbel}(0, 1).$

Chapter 3

Endogenous accelerated life regression

The primary objective of this chapter is to provide two methods to construct a valid confidence set on the parameter of interest β , which is the coefficient on the endogenous intervention or exposure Y , in the accelerated life regression (2.8). The two methods presented, the least squares based and the rank based, are innovative for a number of reasons. First, all the resulting confidence sets have a causal interpretation in the presence endogeneity or unmeasured confounding. Second, the inferential methods are robust to the identification status (i.e. valid in the presence of weak identification) of the

instrumental variables model. Third, both methods, particularly the rank based method are robust to distributional assumptions. Fourth, the rank based method admits (i) the precise definition of controls of §2.3.1, (ii) right censoring, and (iii) endogenous censoring as defined in §2.3.2.

The resulting confidence sets are informative on the quality of the instrument X_2 and inference in general. A weak instrument or invalid specification would result in a possibly disjoint, unbounded, or empty confidence set (Dufour, 1997), and accordingly may be viewed as a non-spurious signal to the inferential value of the data and model.

3.1 Weak instruments and identification robustness

3.1.1 Linear instrumental variables

The preceding linear representation of the accelerated failure time model, as derived from the assumed specification of the conditional survivor function, forms the basis for a structural model in a linear instrumental variables

system. Further defining $y \equiv \ln t$ and collecting $X_{1\iota} = [\iota, X_1]$,

$$y = Y\beta + X_{1\iota}\delta + \sigma\epsilon. \quad (3.1)$$

The standard practice is to assume an associated reduced form,

$$Y = X_{1\iota}\pi_1 + X_2\pi_2 + \mu. \quad (3.2)$$

It is emphasized that for the first stage regression neither the functional form or specification are assumed for the methods presented here. However, it is instructive to think of π_1 and particularly π_2 as intervening nuisance parameters, and accordingly not the focus of inference.

Inference based on the linear system is usually achieved via a number of standard econometric estimators, including two stage least squares, limited information maximum likelihood, and generalized method of moments following standard assumptions or associated regularity conditions. Attention is drawn to a particular regularity failure with considerable implications for econometric practice: the so called “weak instrument” problem.

3.1.2 Weak instruments

There is a substantial literature and on going research devoted to the non trivial complications arising from among others, the nuisance parameter π_2 ,

the coefficient representing the strength of the relationship between the instrument X_2 and the intervention or exposure of interest, Y . When this relationship is “weak”, the limiting distribution of the resulting instrumental variables estimator centres on the limiting distribution of the least squares estimator. Accordingly, historically asymptotically justified methods are not satisfactory for credible inference in the presence of identification failure, which may not be known. For a recent survey see Mikusheva (2013). Despite the large and extensive literature, there has, to date, been no direct focus on duration models. Naturally one of the goals of this thesis is to extend the weak instruments literature with an original contribution of the inclusion of the time origin model in the weak instrument methodology.

We set focus on providing a method robust to instrument quality in the context of a duration model. In general, it is well known that instrumental variable methods rely on the availability of appropriate instruments (Bound, Jaeger & Baker, 1995). An instrument is required to be (i) valid (i.e. not correlated with the structural disturbance ϵ) and presumed (ii) relevant (i.e. sufficiently correlated with, or sufficient information on the exposure Y such that $\pi_2 \neq 0$.), with an assumed first stage regression. Given these conditions, we define *identification robustness* as invariance to (ii). Our identification

strategy, maintaining (i), rests solely on the exclusion of the instrument in the model for the outcome of interest. We directly assess this requirement via an auxiliary regression, and accordingly collect all values of the regression coefficient β in model (2.8), that are compatible with this assumption.

The single identifying assumption placed on the instrument is assumption **A2** X_2 *predetermined*, or **A3** X_2 *independent of ϵ* . It is emphasized that there is no assumptions on the relationship nor functional form between the instrument and endogenous covariate.

3.1.3 Valid confidence sets

The motivation of reporting confidence sets for the parameter of interest is twofold. First, since the confidence set, by definition (Lehman, 1986), reports all values of the parameter accepted at a given level- α (type I rejection probability), the informative value of the underlying estimator and associated test is readily apparent. In other words, the sampling uncertainty in the data is revealed by the content of the confidence set. Second, a valid confidence set provides meaningful inference if the test from which it is derived maintains correct size irrespective of the identification status of the model. Accordingly and particularly in the weak instrument case the potentially insidious iden-

tification failure is revealed *ex post* by the form of the resulting confidence set. For example, an empty set implies model rejection. A set containing the entire real line implies a completely uninformative instrument. Whereas, a bounded confidence sets implies the instrument is informative, and the model specification is compatible with the data. Of course, the variability in the quality and balance of the instrument and the power of the underlying pivotal test, given the data at hand, translate to the width of the bounded confidence set, keeping in mind the above.

A requirement for the construction of “valid” confidence sets is the non-zero probability of an unbounded or empty set. “Accepting the possibility of an unbounded confidence set for a structural coefficient is simply a matter of logic and scientific rigor: the data may simply be uninformative about such coefficients.” (Dufour (1997) pg. 1383).

Naturally closely associated with confidence set construction is the underlying statistical test. The mechanics of inverting the associated statistical test hinges on the test being a pivotal quantity (Fisher, 1930), which of course, requires the test to be nuisance parameter free. A very early and still influential procedure for this purpose was introduced in the statistics literature by Anderson & Rubin (1949).

3.1.4 Anderson-Rubin procedure

Primary interest is focused on building a confidence set for β . For this purpose, under conditional independence of the instrument and within the linear Gaussian framework, Anderson & Rubin (1949) proposed inverting a least squares statistical test for the null hypothesis,

$$H_0 : \beta = \beta_0, \quad H_1 : \beta \neq \beta_0. \quad (3.3)$$

In the context of model (3.1-3.2) this translates to assessing the exclusion of the instrument in the auxiliary regression

$$y - Y\beta_o = X_{1i}\lambda + X_2\gamma + \omega, \quad (3.4)$$

which, rather than describing a statistical model *per se*, serves as a computational tool. The auxiliary regression is derived from the transformation,

$$y - Y\beta_o = Y\beta - Y\beta_o + X_{1i}\delta + \sigma\epsilon.$$

with reduced form,

$$\begin{aligned} y - Y\beta_o &= (X_{1i}\pi_1 + X_2\pi_2 + \mu)(\beta - \beta_o) + X_{1i}\delta + \sigma\epsilon \\ &= (X_{1i} \underbrace{(\pi_1(\beta - \beta_o) + \delta)}_{\lambda} + X_2 \underbrace{(\pi_2(\beta - \beta_o))}_{\gamma} + \underbrace{\mu(\beta - \beta_o) + \sigma\epsilon}_{\omega}) \\ &= X_{1i}\lambda + X_2\gamma + \omega, \end{aligned}$$

where ω is an $(n \times 1)$ vector of random disturbances and $\lambda, \gamma \in \mathbb{R}$. The auxiliary regression is the statistical and computational device of interest in the following regression, test and inferential methods.

3.2 Least Squares Inference

Assessing the exclusion of the instrument in the auxiliary regression implies a zero coefficient on X_2 which, moreover results by imposing $H_0 : \beta = \beta_0$ in model (3.1). Accordingly, in the context of the solely computational model (3.4), to test the hypothesis of the form $H_o : \gamma = 0$, the least squares based test statistic is,

$$T(\gamma_o) = \frac{\hat{\omega}'_c \hat{\omega}_c - \hat{\omega}'_u \hat{\omega}_u / k_2}{\hat{\omega}'_u \hat{\omega}_u / (n - k)}, \quad (3.5)$$

where $\hat{\omega}'_c \hat{\omega}_c$ is the residual sum of squares from the constrained regression and $\hat{\omega}'_u \hat{\omega}_u$ is the residual sum of squares from the unconstrained regression.

Using the fact that:

$$\hat{\omega}_c = M_1(y - Y\beta_o), \quad \hat{\omega}_u = M(y - Y\beta_o),$$

where $M = I - X(X'X)^{-1}X'$, in which $X = [X_{1\iota}, X_2]$ and $M_1 = I - X_{1\iota}(X'_{1\iota}X_{1\iota})^{-1}X'_{1\iota}$ are symmetric and idempotent, gives the generalized Anderson-

Rubin statistic:

$$GAR(\beta_o) = \frac{(y - Y\beta_o)'(M_1 - M)(y - Y\beta_o)/k_2}{(y - Y\beta_o)'M(y - Y\beta_o)/(n - k)}. \quad (3.6)$$

Theorem 3.2.1 *Under the null hypothesis imposing model (3.1) at the true parameter value of $\beta = \beta_o$, the distribution of $GAR(\beta_o)$ is completely determined by the distribution of $\overline{GAR}(\epsilon; X)$, where*

$$\overline{GAR}(\epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon'M\epsilon/(n - k)}. \quad (3.7)$$

The distribution of $\overline{GAR}(\epsilon; X)$ is completely determined by the distribution of the structural error, ϵ and remains exactly pivotal for any location-scale family in model (3.1). The distribution is invariant to β_0 , σ , and the data generating process linking Y to X_2 .

Proof 3.2.2 *Under $H_0 : \beta = \beta_0$, where (3.1) is the true model,*

$$y - Y\beta_o = X_{1\iota}\delta + \sigma\epsilon.$$

Using the fact:

$$M_1X_{1\iota} = 0, \quad MX_{1\iota} = 0,$$

implying from (3.6),

$$GAR(\sigma\epsilon; X) = \frac{\sigma\epsilon'(M_1 - M)\sigma\epsilon/k_2}{\sigma\epsilon'M\sigma\epsilon/(n - k)}.$$

This gives, the pivotal statistic:

$$\overline{GAR}(\epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon'M\epsilon/(n - k)}.$$

■

Empirically, these results permit the use of simulation based procedures in the absence of normality. Analytical construction of exact confidence sets is achieved via simulating the exact null distribution. First, however, it is instructive to consider the case where ϵ is normally distributed (Dufour & Taamouti, 2005), corresponding to the log-normal accelerated life model, in which case the GAR statistic follows an F distribution. The confidence set is

$$C_\beta(\alpha) = \{\beta_o : GAR(\beta_o) < F_{k_2, n-k}(\alpha)\}, \quad (3.8)$$

where α is the specified significance level. Following the notation of Dufour & Taamouti (2005), in the context of the accelerated life regression, the left hand side of the auxiliary regression (3.4) is of the form:

$$f(y, Y, \beta) = h_1(y, Y)' \beta + h_2(y, Y), \quad (3.9)$$

where,

$$h_1(y, Y) = -Y, \quad h_2(y, Y) = y \equiv \ln(t).$$

Accordingly, using the above notation and in the general case where $\beta_o \in \mathcal{R}^g$, (3.8) is written as the *quadric confidence set*:

$$C_\beta(\alpha) = \{\beta_o : \beta_o' A \beta_o + b' \beta_o + c \leq 0\}, \quad (3.10)$$

where:

$$A = h_1' B h_1, \quad b = -2h_1' B h_2, \quad c = h_2' B h_2,$$

in which,

$$B = M_1 - [1 + r(\alpha)]M, \quad \text{where } r(\alpha) = \frac{k_2 F(\alpha)}{(n - k)},$$

with M_1 and M are as previously defined. In what follows, the A matrix (referred to as the α -concentration matrix) plays a central role in determining the boundedness of the confidence set. To construct a confidence set on the $(n \times g)$ parameter of interest β_o , given the data, compute A , b , and c as defined by (3.10). Focus further attention on the case of a positive definite α -concentration matrix A . In this case the confidence set is:

$$C_{\beta_o} = [\tilde{\beta} + \sqrt{dA^{-1}}, \tilde{\beta} - \sqrt{dA^{-1}}], \quad (3.11)$$

if and only if $d \geq 0$, where:

$$\tilde{\beta} = -\frac{1}{2}A^{-1}b, \quad d = \frac{1}{4}b'A^{-1}b - c.$$

In the case where the model contains a single endogenous covariate, the quadric reduces to a quadratic inequality. The general solution admits empty or unbounded confidence sets with non zero probability. This conforms to the requirements for statistical validity as described in Dufour (1997).

3.2.1 Simulation methods for location-scale family distributions

The appropriate statistical distribution and related cut-off values depend on the error distribution in the structural model. As described above, in the log-normal model, the error distribution is standard normal, and accordingly, the GAR statistic is F distributed. However, when the error distribution is not standard normal, but belongs to the location-scale family, a direct result of **Theorem 3.2.1** is to allow simulation methods to be used to obtain the null distribution of the GAR statistic. For example in the Weibull model, the error distribution follows the standard type I extreme value distribution. Accordingly, the related critical values are calculated by simulating the GAR statistic, drawing from the related quantile function. For each draw j ,

$$\epsilon_j = -\ln(-\ln(u_j)), \tag{3.12}$$

where the $(n \times 1)$ vector u_j is drawn from the uniform $[0,1]$ distribution, giving the j th realization of the GAR statistic:

$$\overline{GAR}_j = \frac{(\epsilon_j)'(M_1 - M)(\epsilon_j)/k_2}{(\epsilon_j)'M(\epsilon_j)/(n - k)}. \quad (3.13)$$

Repeating this procedure for $j = 1..J$, simulations builds the simulated null distribution. The appropriate α -level cut off value may subsequently be utilized in analytical construction of exact confidence sets as described above,

$$C_\beta(\alpha) = [\beta_o : GAR(\beta_o) < gar_{calc}(\alpha)], \quad (3.14)$$

where $gar_{calc}(\alpha)$, for an α significance level, is exactly simulated by:

$$\epsilon_j = \ln\left(\frac{u_j}{1 - u_j}\right), \quad \epsilon_j = -\ln(-\ln(u_j)),$$

for the log-logistic and Weibull case respectively, where the n -vector $u_l \sim \text{Uniform}[0, 1]$ for each draw $j = 1, \dots, J$ simulations of \overline{GAR}_j

With the above least squares based procedure, the parametric assumptions of the accelerated life specifications only intervene for the computation of the quantity $gar_{calc}(\alpha)$. The numeric value of the statistic $GAR(\beta_o)$ is a direct function of the observed data and the hypothesized β_o . In this sense, the procedure is semi-parametric and accordingly less sensitive to the misspecification of the distribution in contrast to likelihood based methods.

Although the uncensored parametric least squares based result is standard, to the best of our knowledge this is the first extension of identification-robust instrumental methods to duration data and serves as an analytically tractable reference or baseline check.

3.3 Rank inference

An original contribution of this thesis is to provide a rank based inferential method for the construction of a valid confidence set on β that is (i) exactly distribution free and (ii) accounts for right censoring including endogenous censoring, as defined in §2.3.2. Moreover, the proposed rank statistics, which are derived from replacing the observed variate values with either (i) the expected value or (ii) the quantile of the order statistic in sampling a presumed distribution, are in general robust to misspecification and extreme values.

The basis for linear rank statistic we consider was introduced to the identification robust econometric literature by Andrews & Marmer (2008). We generalize their statistic by considering the duration framework with right censoring. Within the duration literature, rank methods have been used specifically with the accelerated life model (Kalbfleisch & Prentice, 2002)

and generally with the transformation model (Han, 1987), (Chen, 2002).

3.3.1 Linear rank statistic

The rank method presented here draws from a class of aligned (Hodges & Lehmann, 1962) linear rank statistics viewed as score function tests based on the data generating process arising from model (3.1) and assumptions **A3** and **A4**. In a similar fashion to §3.1.4, the hypothesis of interest remains,

$$H_0 : \beta = \beta_0, \quad H_1 : \beta \neq \beta_0.$$

However, the computational device is now the aligned rank auxiliary regression, analogous to (3.4):

$$\text{rank}(y - Y\beta_o - x_1\hat{\delta}(\beta_o)) = x_2\gamma + \omega, \quad (3.15)$$

in which $\hat{\delta} = (x_1'x_1)^{-1}x_1'(y - Y\beta_o)$ is the null restricted least squares estimator of δ , with x_1, x_2 expressed as X_1, X_2 with columns standardized to add to zero. The left hand side, the rank n -vector $\text{rank}(y - Y\beta_o - x_1\hat{\delta}(\beta_o)) \equiv r$, is the “rank” or positional label $[(1) \cdot (n)]$ of the ordered variate values of the quantity $(y - Y\beta_o - x_1\hat{\delta}(\beta_o))$. Let $z = y - Y\beta_o - x_1\hat{\delta}(\beta_o)$ be a random

variable representing the aligned residual about the hypothesized β_o ,

$$\begin{aligned}
 z &= y - Y\beta_o - x_1\hat{\delta}(\beta_o) \\
 &= y - Y\beta_o - x_1(x_1'x_1)^{-1}x_1'(y - Y\beta_o) \\
 &= m_1(y - Y\beta_o),
 \end{aligned}$$

where $m_1 = I - x_1(x_1'x_1)^{-1}x_1'$. The alignment of the left hand side of the auxiliary regression may heuristically be viewed as an estimating function around the hypothesized β_o . This approach would, for example, admit a wider class of transformations of the outcome t , or other null restricted estimators, provided they were cast as a function of the observed data, t, Y, x_1 and the hypothesized β_o .

As a purely computational device, the auxiliary regression (3.15) has conveniently transferred all dependent quantities, including possibly X_1 to the left side. Viewed in this way, the right side is comprised of a completely randomized quantity and an unobserved disturbance. Consequently, maintaining assumptions **A3**, **A4**, under the null hypothesis, the rank vector is considered exchangeable. That is, the random variable r is distributionally invariant to any and all re-ordering of the rank labels (1)..(n).

We have defined z as the random variable representing the aligned residual

about the hypothesized β_o . The order statistic $z^{(1)} \leq \cdot \leq z^{(n)}$ is the ordered variate values of z , with associated concomitant variate values of x_2 , where (i) is the rank label of the aligned residual order statistic. The rank vector is given by the corresponding rank labels $r = [(1) \cdot (n)]$, with the associated rank vector probability, $p(r)$,

$$p(r) = \int_{z^{(1)} \leq} \dots \int_{\leq z^{(n)}} \prod_{i=1}^n f(z^{(i)} - x_2 \gamma) dz^{(i)}. \quad (3.16)$$

Unlike a fully parametric marginal likelihood, the rank vector probability $p(r)$, is invariant to the distributional complications of unmeasured confounding and independent of location and scale (Kalbfleisch & Prentice, 2002).

A test of $H_o : \gamma = 0$, as implied by $H_0 : \beta = \beta_0$, is based on the associated score statistic derived from the aligned rank vector probability (Cox & Hinkley, 1974),

$$\begin{aligned} \mathcal{S} &= \left. \frac{d \ln p(r)}{d\gamma} \right|_{\gamma=0} \\ &= \sum_{i=1}^n c^{(i)} a(i), \end{aligned}$$

which is general form for the linear rank statistic, where $c : [0, 1) \rightarrow \mathbb{R}$ is a rank preserving non stochastic score and $a(i)$ is referred to as a regression constant (Randles & Wolfe, 1979). The generalized Andrews-Marmer

statistic is a quadratic form in \mathcal{S} , expressed in matrix notation,

$$GAM(\beta_o) = c' (p_2) c, \quad (3.17)$$

where $p_2 = x_2(x_2'x_2)^{-1}x_2'$ and the n -column vector, $c^{(i)} : [0, 1) \rightarrow \mathbb{R}$ is a rank preserving non stochastic score.

The score vector satisfy a nondecreasing and nonconstant condition, $c^{(i)} \leq \dots \leq c^{(n)}$ and $c^{(i)} \neq c^{(n)}$, where again, (i) is the rank label of the associated aligned residual order statistic. In general the score is selected according to a presumed cumulative distribution F_o in model (1), however given the robustness of the rank scores (Chernoff & Savage, 1958), this choice is made on power considerations. Moreover, as suggested in §2.3, since any presumed distribution may indeed be misspecified as a consequence of unaccounted heterogeneity, it is attractive to have accordingly robust methods to this complication. To be clear, a score based on a misspecified distribution would have lower power, but would continue to preserve size.

Two related and asymptotically equivalent scores (Randles & Wolfe, 1979) are the quantile F_o scores and the expected value F_o scores:

$$c^{(i)} = F_o^{-1} \left(\frac{\binom{i}{(n+1)}}{\binom{n+1}{(n+1)}} \right), \quad c^{*(i)} = E_{F_o}[V^{(i)}],$$

where $V^{(i)}$ is the i th order statistic in a random sample of size n . For example,

in the log-normal accelerated life model, the quantile F_o and expected value F_o scores follow from van der Waerden (1953) and Fisher & Yates (1963), respectively. Other well known classical expected value scores are those of Wilcoxon (1945),

$$c^{(i)} = \frac{2(i)}{(n+1)} - 1,$$

which is the expected value score when the presumed distribution is logistic.

The expected value score of Savage (1956),

$$c^{(i)} = \frac{1}{n} + \frac{1}{(n-1)} + \dots + \frac{1}{(n-(i)+1)} - 1,$$

arises when the presumed distribution is exponential.

3.3.2 Right-censoring

A primary contribution of this thesis is to use the above rank inference framework combined with the method of Prentice (1978), to account for right-censoring. As discussed in §2.1.1 there are many schemes where different forms of censoring may occur. A detailed reference is Klein & Moeschberger (2003). A particularly special case of censoring uniquely and originally considered here is endogenous censoring as detailed in §2.3.2.

Maintaining **C3**, the observed time is now t^* , which is the lesser of cen-

sored time, τ or event time t . The occurrence of $t^* = \tau$ is denoted by the censoring indicator, d . Let each of the censored observations, within any adjacent non-censored pair, $i = 1..k$, be indexed by the subscript $j = 1..m_i$. Conceptually, all censored observations now contribute to the rank vector probability via the survivor function. As analogously detailed in Prentice (1978), the associated rank vector probability, $p(r)$, is now,

$$p(r) = \int_{z^{(1)} \leq} \dots \int_{\leq z^{(n)}} \prod_{i=1}^k [f(z^{(i)} - x_2^{(i)}\gamma) \prod_{j=1}^{m_i} S(z^{(i)} - x_{2_{ij}}\gamma) dz^{(i)}]. \quad (3.18)$$

Again, as detailed in Prentice (1978), the associated score test of $H_o : \gamma = 0$ results in a composite expected value score for the statistic (3.16), in the case of Wilcoxon scores,

$$c^{(i)} = 1 - 2 \prod_{j=1}^i \frac{n_j}{n_j + 1}, \quad c_{m_i}^{(i)} = 1 - \prod_{j=1}^i \frac{n_j}{n_j + 1},$$

and for the Savage scores,

$$c^{(i)} = \sum_{j=1}^i n_j^{-1} - 1, \quad c_{m_i}^{(i)} = \sum_{j=1}^i n_j^{-1},$$

where m_i indexes all the right-censored residuals in any uncensored adjacent ordered interval $[(i), (i + 1))$.

Under exchangeability, the rank vector probability is $1/n!$ for any permutation of the ranks, $(1..n)$. An assumed independent censoring mechanism implies an equal individual probability of censoring, together giving:

Theorem 3.3.1 *Suppose the censoring mechanism is independent of the data generating process for the outcome y , as specified in model (1). Then, under the null hypothesis imposing model (1) at the true parameter value of $\beta = \beta_o$, the distribution of $GAM(\beta_o)$ is completely determined by the distribution of $\overline{GAM}(u; X_2)$, where*

$$\overline{GAM}(u; X_2) = c(\text{rank}(u))' (p_2) c(\text{rank}(u)), \quad (3.19)$$

in which the elements of u are independent draws from any assumed distribution.

The exchangeability of the aligned residuals permits a precise definition of the intended role of the controls, X_1 . As detailed in §2.3.1, Stock (2010) suggests the role of controls is to satisfy $E(\epsilon|Y, X_1) = E(\epsilon|X_1)$. Similarly, in Rubin (1990) the assignment mechanism for Y is *unconfounded* with y given X_1 if $Pr(Y|X_1, y) = Pr(Y|X_1)$. In both examples, this implies the correlation of X_1 with the unobservables. The aligned residuals, z , remain exchangeable under assumption **A4**, which allows for correlation between the controls and the unobserved types. In the context of right censoring, including the *ex post* indicator d in the vector of controls allows assumption **C2**, which permits the correlation of censoring and the unobserved types.

Corollary 3.3.2 *The distribution of $GAM(\beta_o)$ remains exactly pivotal in the presence of censoring that need not be independent of the data generating process for the outcome y .*

The distribution of $GAM(\beta_o)$ is exactly pivotal for any distributional assumption on model (1), invariant to β_0 , σ , and the data generating process linking Y and X_2 . As a result, confidence set construction is achieved via a search over β_o , satisfying:

$$C_\beta(\alpha) = [\beta_o : GAM(\beta_o, y, Y; X) < gam_{calc}(\alpha)], \quad (3.20)$$

where $gam_{calc}(\alpha)$, for an α significance level, is exactly simulated by, as an example, the n -vector $u_l \sim \text{Uniform}[0, 1]$ for each draw $l = 1, \dots, m$ simulations \overline{GAM}_l .

The following steps explicitly summarize the construction of a confidence set for β_o in practice. Step (1): calculate the critical value $gam_{calc}(\alpha)$ by first drawing the n -vector $u_l \sim \text{Uniform}[0, 1]$ and replacing the random variate values with the rank labels. Apply the desired score, $c^{(i)}$ to the rank labels and construct the statistic (3.19), saving the resulting value as a single element of an m -vector. Repeat this for $m = 1,000$ times and order the resulting m -vector. Select $gam_{calc}(\alpha)$ as the element corresponding to the

desired significance level from the ordered m -vector. Step (2) calculate value of statistic $GAM(\beta_o, y, Y; X)$ for a chosen β_o value by replacing the random variate values of z with the associated rank labels. Apply the desired score, $c^{(i)}$ to the rank labels and construct the statistic (3.17), saving the resulting value as a single element in the grid search over β_o . Step (3) retain all β_o not rejected, that is, satisfying $GAM(\beta_o, y, Y; X) < gam_{calc}(\alpha)$.

3.4 Simulation

The analytical objective of the methodology presented is to provide a valid confidence set on the parameter of policy interest, β . Accordingly, the statistical objective of the methodology presented is to provide an appropriate pivotal test that has sufficient properties for test inversion for a given α -level. Therefore, constructing a valid statistical test and hypothesis testing in general, is central to the proposed inferential procedure. With this in mind, a simulation study utilizing a Monte Carlo is conducted to confirm (i) the size-control and (ii) the power comparisons of the tests considered under various empirically relevant data structures.

A classical decision rule associated with a statistical test partitions the

data into an acceptance region and rejection region, where the former is the subset consistent with the null hypothesis, H_o . Naturally the inherent uncertainty of the data raise the possibility of an incorrect decision. The test may mistakenly reject H_o , when in fact, H_o is true, resulting in a type I error. Alternatively the test may correctly reject a false H_o . The probability of the latter being the power of the test. As detailed in Cameron & Trivedi (2005),

$$Size = \mathbb{P}(Reject\ H_o \mid H_o\ true), \quad Power = \mathbb{P}(Reject\ H_o \mid H_o\ false).$$

The fundamental strategy of the methodology presented, following the classical Neyman-Pearson approach, is to control the chance of incorrectly rejecting the null hypothesis, i.e. committing a type I error. Amongst such tests, naturally a test with greater power for all possible specifications is preferred. Accordingly, the purpose of the simulation is to confirm that all the tests considered show appropriate size control and sufficient power properties to provide meaningful inference via the related confidence sets for β .

In general, both the least squares and rank based tests do not rely critically on the parametric assumption. The least squares procedure is semi-parametric in the sense that the assumed distribution only intervenes in calculating the critical point or cut-off value. In the rank based test, the

assumed distribution would only guide the choice of the score function. Irrespective of the test, it is important to keep in mind that regardless of the assumed distribution, some degree of distributional misspecification is likely unavoidable given the presence of (i) frailty and (ii) censoring, both of which may intervene with unknown or arbitrarily specified distributional consequences. Indeed the robustness to this and possibly other sources of distributional misspecification is an attractive feature or property of all the tests considered.

The power of the test directly translates to the length of the confidence set. This is a useful, and as discussed perhaps critical, signal to the informative content of the instrument.

3.4.1 Study design

Following the notation of model (1), an empirically relevant simulation design adopts the data generating process:

$$y = Y\beta + X_1\delta + \epsilon, \quad Y = g(X_1\pi_1 + X_2\pi_2 + \sqrt{1 - \rho^2}\mu + \rho\epsilon),$$

in which various sampling schemes on ϵ , μ , and the sample balance of X_2 determine the different testable parametric models. The parameters π_2 , ρ , and β_o for an assumed β , reflect respectively, the instrument strength, degree

of confounding, and distance from the null. The function $g(\cdot)$ relaxes the first stage linearity, as non-linearity is empirically relevant in the clinical setting.

The simulation used a Monte Carlo technique programmed in STATA. An R-class program with an imbedded MATA program was written to permit use of the `simulate` command executed over a grid of β , π , X_2 , and n . The program was calibrated using parameter values obtained from the empirical data set. Following which, random variates were drawn from the normal, logistic, Gumbel, and exponential distributions, as appropriate. The instrument, censoring indicator, and covariates were generated to mimic the empirical data set. The resulting synthetic data and random variates were fed into the structural model describing the data generating process. The simulated data was passed to the MATA program which calculated the (i) numerical value of each statistic and (ii) the simulated null distribution and associated exact α -level cut off numerical value. A comparison of the two values provided an accept or reject. This was repeated 1,000 times and the total number of rejections recorded. The entire procedure was then repeated for a variety of specifications and possible parameter values.

3.4.2 Results

Included below are the power and size results for three distributions for the above data generating processes, the normal, logistic and Gumbel. The simulation used $\beta=.5$ as the true parameter value. Two instrument strengths, weak and moderate, were recorded. The indicator instrument X_2 was evenly divided in the sample, which we refer to as balanced. The sample size was 100.

β	π_2	GAR Norm	GAR Logistic	GAR Gumbel	GAM Norm	GAM Logistic	GAM Gumbel	GAM Exp	Wilcoxon	Savage	Wilcox Cen	Savage Cen
-0.25	0.4	0.145	0.141	0.147	0.138	0.135	0.143	0.134	0.126	0.13	0.125	0.118
0	0.4	0.099	0.101	0.099	0.101	0.105	0.1	0.105	0.096	0.099	0.096	0.094
0.25	0.4	0.062	0.065	0.067	0.06	0.062	0.067	0.056	0.061	0.057	0.062	0.055
0.5	0.4	0.05	0.05	0.049	0.053	0.053	0.059	0.056	0.052	0.054	0.052	0.048
0.75	0.4	0.085	0.086	0.085	0.084	0.086	0.089	0.089	0.081	0.085	0.08	0.083
1	0.4	0.256	0.255	0.256	0.261	0.258	0.257	0.244	0.249	0.236	0.248	0.233
1.25	0.4	0.595	0.595	0.595	0.59	0.591	0.576	0.534	0.571	0.524	0.565	0.517
-0.25	0.8	0.446	0.447	0.449	0.439	0.434	0.435	0.395	0.425	0.381	0.419	0.372
0	0.8	0.279	0.282	0.284	0.274	0.272	0.271	0.244	0.27	0.243	0.259	0.228
0.25	0.8	0.155	0.153	0.155	0.151	0.153	0.147	0.137	0.149	0.134	0.133	0.129
0.5	0.8	0.056	0.055	0.058	0.059	0.057	0.059	0.056	0.061	0.058	0.059	0.054
0.75	0.8	0.225	0.223	0.223	0.227	0.231	0.214	0.192	0.215	0.19	0.214	0.183
1	0.8	0.781	0.781	0.778	0.767	0.769	0.754	0.696	0.765	0.689	0.754	0.682
1.25	0.8	0.993	0.993	0.993	0.992	0.992	0.988	0.982	0.989	0.98	0.985	0.976

Normal DGP Sample size 100

β	π_2	GAR Norm	GAR Logistic	GAR Gumbel	GAM Norm	GAM Logistic	GAM Gumbel	GAM Exp	Wilcoxon	Savage	Wilcox Cen	Savage Cen
-0.25	0.4	0.078	0.079	0.078	0.081	0.083	0.08	0.075	0.083	0.076	0.078	0.078
0	0.4	0.064	0.065	0.063	0.065	0.065	0.068	0.066	0.069	0.065	0.068	0.068
0.25	0.4	0.057	0.059	0.058	0.06	0.054	0.057	0.062	0.066	0.061	0.064	0.057
0.5	0.4	0.046	0.046	0.046	0.048	0.047	0.047	0.046	0.049	0.046	0.052	0.049
0.75	0.4	0.059	0.059	0.059	0.055	0.056	0.061	0.051	0.056	0.055	0.057	0.049
1	0.4	0.141	0.143	0.146	0.149	0.149	0.134	0.129	0.161	0.124	0.146	0.11
1.25	0.4	0.347	0.347	0.35	0.353	0.344	0.341	0.308	0.367	0.298	0.343	0.287
-0.25	0.8	0.17	0.17	0.176	0.176	0.181	0.167	0.143	0.174	0.138	0.165	0.127
0	0.8	0.132	0.132	0.13	0.128	0.129	0.123	0.11	0.132	0.105	0.129	0.101
0.25	0.8	0.087	0.086	0.089	0.081	0.076	0.074	0.062	0.077	0.064	0.075	0.054
0.5	0.8	0.057	0.057	0.057	0.052	0.055	0.056	0.06	0.055	0.059	0.058	0.056
0.75	0.8	0.112	0.113	0.111	0.112	0.11	0.112	0.106	0.11	0.107	0.101	0.092
1	0.8	0.342	0.347	0.347	0.374	0.365	0.355	0.321	0.383	0.312	0.363	0.313
1.25	0.8	0.89	0.892	0.893	0.89	0.881	0.871	0.826	0.889	0.816	0.878	0.794

Logistic DGP Sample size 100

β	π_2	GAR Norm	GAR Logistic	GAR Gumbel	GAM Norm	GAM Logistic	GAM Gumbel	GAM Exp	Wilcoxon	Savage	Wilcox Cen	Savage Cen
-0.25	0.4	0.111	0.114	0.116	0.126	0.123	0.115	0.096	0.124	0.096	0.117	0.098
0	0.4	0.106	0.107	0.106	0.114	0.11	0.109	0.092	0.108	0.09	0.102	0.085
0.25	0.4	0.056	0.055	0.058	0.046	0.047	0.053	0.058	0.056	0.057	0.057	0.058
0.5	0.4	0.052	0.054	0.054	0.048	0.052	0.053	0.049	0.053	0.048	0.049	0.046
0.75	0.4	0.078	0.083	0.083	0.071	0.074	0.076	0.077	0.076	0.077	0.074	0.073
1	0.4	0.226	0.225	0.22	0.246	0.247	0.216	0.168	0.225	0.163	0.219	0.167
1.25	0.4	0.473	0.473	0.475	0.49	0.484	0.444	0.396	0.488	0.374	0.476	0.369
-0.25	0.8	0.301	0.302	0.3	0.35	0.346	0.292	0.227	0.335	0.221	0.337	0.225
0	0.8	0.215	0.213	0.213	0.231	0.237	0.205	0.173	0.226	0.173	0.209	0.162
0.25	0.8	0.125	0.124	0.129	0.135	0.138	0.123	0.107	0.132	0.108	0.13	0.114
0.5	0.8	0.054	0.054	0.055	0.047	0.053	0.054	0.055	0.053	0.055	0.053	0.056
0.75	0.8	0.156	0.155	0.156	0.177	0.179	0.151	0.118	0.178	0.117	0.173	0.12
1	0.8	0.602	0.605	0.608	0.653	0.648	0.591	0.485	0.652	0.468	0.621	0.472
1.25	0.8	0.974	0.973	0.975	0.97	0.971	0.954	0.919	0.971	0.913	0.965	0.899

Gumbel DGP Sample size 100

3.4.3 Discussion

In general with moderately strong instruments or better, power approaches one with sample sizes greater than 200 for both the least squares and rank statistic. Power increases in the effect size, however the power curve is not symmetric. Power increases markedly in instrument strength. Power is decreasing in instrument balance. Comparatively the rank statistics outperform with poor instrument quality and are not outperformed with good instruments, if the vector of controls satisfy the precise definition given in §2.3.1. Alternatively, if the vector X_1 is random, then there is a decreased power difference between the tests. The rank statistics generally outperform when nonlinearities in the first stage are introduced. However, although power in general increases in (i) overall sample size and (ii) the sample balance of the instrumented, in all cases the most notable power improvement is in instrument strength, which empirically would be reflected in the width of the resulting confidence set.

Chapter 4

Clinical inference in pediatric acute care

The preceding chapter of this thesis focused on proposing an econometric methodology with associated biostatistical techniques to address the general problem of unmeasured confounding or endogeneity in an accelerated life regression model. The motivation for this work originated from observations of length of stay outcomes derived from a clinical study in the pediatric intensive care unit. This chapter provides the background to this clinical study, the relevance of our econometric framework and the results obtained by applying our methods to the clinical setting.

Instrumental variable techniques have a long history in economics (Stock & Trebbi, 2003). Although historically far less employed in biostatistics and epidemiology, there is a growing amount of recent interest in clinic applications (Cawley, 2015). We contribute to this growing literature by (i) providing practitioners an example of how the choice and availability of an appropriate instrument may usefully be incorporated into study design and (ii) how our proposed robust instrumental variables method corrects for unmeasured confounding with length of stay outcomes. The specific clinical question we sought to answer using our generalized Anderson-Rubin statistic and generalized Andrews-Marmer statistic is: What is the relationship between a patient's length of stay in the pediatric intensive care unit and their illness severity score at the time of admission?

The goal of pediatric intensive care is to decrease the morbidity and mortality of critically ill children. This goal has been accomplished in developed countries by establishment of specialized services in pediatric intensive care units (PICU) (Pollack, Yeh, Ruttimann, Holbrook, & Field, 1984). These services delivered by a highly specialized team of health care providers in the PICU allow critically ill children to be seen quickly, given lifesaving therapies and be supported by heart lung machines if needed. Over the last twenty

years, these specialized services and form of acute care delivery have resulted in a dramatic drop in the mortality of critically ill children from 20% to 3-8% (Pollack, Patel & Ruttimann, 1996). However, these services come with a heavy cost in terms of health care resources and dollars. Specialized intensive care units accounted for 15.9% of inpatient direct expenses in Ontario acute care hospitals between 1999-2000 and 2003-2004, but only 8.1% of inpatient days (Leeb, Jokovic, Sandhu, & Ainck, 2006). The overall cost of providing intensive care also rose over this period in Ontario, from \$475 million to \$662 million (Leeb, Jokovic, Sandhu, & Ainck, 2006). As a result of these rising costs and intensity of resource use in the PICU, researchers, policy makers and economists have tried to determine the factors contributing to these costs and to develop appropriate models by which to analyze these factors.

We introduce an original prospectively collected observational data set to analyze illness severity and length of stay in the Canadian paediatric intensive care patient population. Observations ($n = 10,044$) were collected over a two year period at five tertiary care, level three trauma, pediatric intensive care units (British Columbia Children's Hospital, Stollery Children's Hospital, The Hospital for Sick Children, Children's Hospital of Eastern Ontario, McMaster Children's Hospital). Data collected included exact time of

admission and discharge providing precise length of stay along with physiologic, demographic, and therapeutic patient specific characteristics at the time of admission. The data set included 658 trauma patients and an overall mortality rate of 3.52% (354/10,044).

The physiologic, demographic, and therapeutic patient specific characteristics at the time of admission were used to determine an illness severity index (Slater, Shan & Pearson, 2003) (Pollack, Patel & Ruttimann, 1996) for each patient and although developed as a predictor of mortality, such scores are often used as a marker to assess quality and efficiency, organize health care delivery, allocate scarce resources, and stratify patients for research and therapy. However, regardless of how comprehensive, there remain unobserved risk factors that may affect both length of stay and illness severity.

4.1 Pediatric intensive care unit length of stay

PICU length of stay is an important contributor to PICU costs and is used by governments and health care systems to compare quality of care across units, allocate scarce resources and to determine per patient costs (Hsu, Lakhani & Brazelton , 2015). Therefore, the manner in which length of stay is analyzed

and reported is important as it has the potential to impact policy. However, there are several potential issues with reporting of length of stay data.

The first issue is that PICU length of stay follows a highly right-skewed distribution, even after a logarithmic transformation, leading to questions about how to deal with the very long stay patients. Some researchers and administrators have dealt with this issue by simply ignoring the long-stay patients and reporting the mean or median value of the remaining data. However, this approach results in loss of valuable information as although long-stay patients comprise only 2.1% to 8.1% of the PICU population they utilize more than 50% of the allotted resources (Pollack, Wilkinson & Glass, 1987), a result that remains consistent with our data set.

The second issue is that of right censoring. Patients with higher illness severity scores have longer PICU lengths of stay except for those patients who die and whose length of stay is therefore censored. Failure to correct for this censoring results in loss of useful information.

The third issue is that length of stay is a complex variable that is dependent on illness severity (Hsu, Lakhani & Brazelton , 2015). However, the relationship between PICU length of stay and illness severity may be affected by several confounding factors. In the clinical setting, a confounding factor

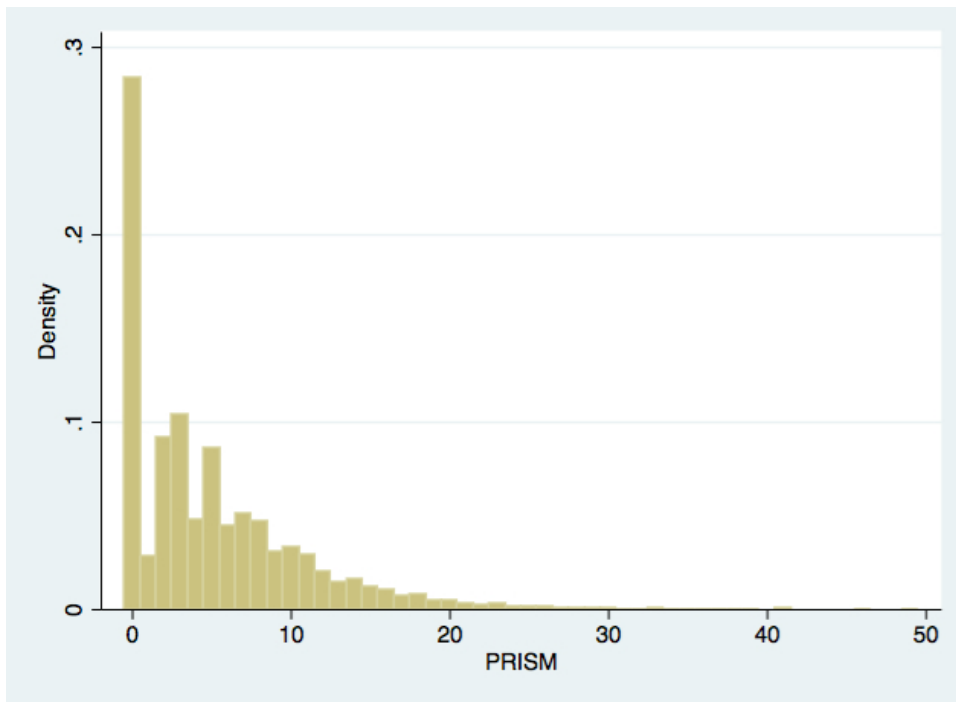
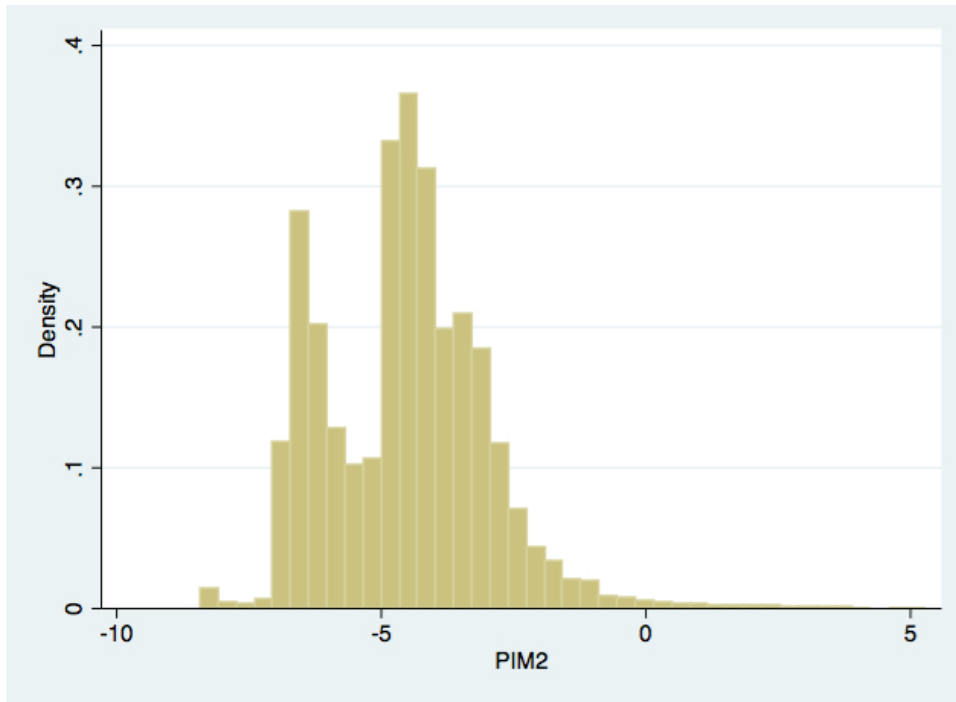
is defined as such if it meets three conditions. First, it is associated with both the risk factor of interest (illness severity) and the outcome (length of stay). Secondly, it is distributed unequally among the groups being compared (different PICUs) and thirdly, it is not an intermediary step in the causal pathway from the exposure of interest (illness severity) to the outcome of interest (length of stay). An example of a confounding factor of the relationship between PICU length of stay and illness severity is the need for mechanical ventilation. This confounder meets the three criteria outlined above as the need for mechanical ventilation has been shown to be related to both variables, may be unevenly distributed across different PICUs and in and of itself, does not result in worsening illness severity. Biostatistical methods have been developed to deal with known confounders including matching, stratification and multiple linear regression. However, less attention has been paid to analysis of relationships where there may be unknown confounders i.e. factors that affect both variables of interest but which cannot be measured. An example of such a confounder would be the so-called frailty factor whereby a “weaker” type of patient may be more likely to become sicker with a given illness than a “stronger” type may. In a randomized, controlled trial, one would assume that patients of different types would be evenly distributed

between the groups being studied, but this is not the case with observational data sets such as the current one analyzed.

4.2 Pediatric illness severity

Illness severity in the PICU population is most commonly measured by two scores: the Pediatric Risk of Mortality (PRISM III) score (Pollack, Patel & Ruttimann, 1996) and the Pediatric Index of Mortality score (PIM 2) (Slater, Shan & Pearson, 2003). The illness severity scores have been derived from large retrospective data sets of physiologic, therapeutic and demographic variables. The basis of the scores, in both cases, is a logistic regression for mortality. The resulting weights assigned to each retained predictor forms the weighted score or index for each patient. Although similar to a propensity score, the weights are derived from a training sample and subsequently prospectively validated using new and different patient observations. Both PRISM III and PIM 2 have been extensively used in the PICU literature and although derived from a patient's probability of mortality, they are primarily used as a measure of illness severity.

Each patient in the study sample received a PIM 2 and PRISM III score



according to their observed factors at the time of admission. The distribution of the scores are presented in the histograms above.

The PIM 2 score weights were directly derived from a logistic regression for mortality in the original training sample, not the current sample. The score identifies nine risk factors and assigns weights derived from the logistic regression, to each of the risk factors. The risk factors include four physiologic measures (blood pressure, pupillary reaction, PaO_2 , base excess), one therapeutic (mechanical ventilation), three admission categories (elective, surgical, cardiac), and two diagnosis categories (high risk, low risk). The resulting scores from PIM 2 for the study sample are continuous and bimodal.

The PRISM III score used seventeen physiologic variables categorized into cardiovascular/neurologic, acid-base/blood gas, chemistry, and hematology. A number of other admission, therapeutic, and diagnosis were utilized. The score weights were again separately derived from two algorithms based on logistic regressions in the original training sample, not the current sample. The resulting scores from PRISM III for the study sample are categorical with a noticeable point mass at zero.

The scores are rarely used for mortality prediction but are extensively

used for quality assurance, research stratification, and risk adjustment. The scores are viewed as a signal of the intensity of treatment that patients require. In the high intensity environment of the PICU, the illness severity score is a useful summary measure for each patient. However, regardless of how comprehensive, the scores may not account for factors that are often difficult to observe. Indeed, we suggest that the scores are imperfect signals of a patient's type, when utilized in, for example, a length of stay analysis. Our intuition is that "frailer" patients would both score higher on the illness severity index for the same insult and experience a longer stay in the PICU irrespective of their score, as compared to "stronger" patients. In which case, the illness score is confounded or endogenous due to the missing observable types. Anecdotal evidence is the case of twins or siblings exposed to a flu virus, in which two otherwise observationally equivalent patients experiencing the same exposure have markedly different hospitalizations.

4.3 Pediatric trauma

The selection of trauma as an instrument hinges on its exclusion from explaining any portion of intensive care length of stay. This would be achieved,

if it were the case that trauma was as a purely chance event amongst the unobserved types otherwise not directly accounted for in the specified model for length of stay. A self selection to a trauma event, where the selection mechanism was in part determined by the unobserved type, would naturally preclude its use as an instrument.

Whether or not these conditions are satisfied requires careful consideration of the trauma aetiologies, that is, the source or cause of the trauma. An aetiology that was a completely chance event would likely not discriminate between the unobserved types, where an equal distribution of the unobserved were at risk of the particular aetiology.

For example suicide attempts amongst adolescents is a tragic event resulting in some of the PICU trauma admissions. Naturally, suicide falls into the choice side of the choice versus chance paradigm in models of selection. The fundamental question is whether or not the “types” self-selecting to suicide attempts coincide with the unobserved types that confound the length of stay-illness severity model. If it were the case that “strong” and “frail” are equally likely to self-select, this portion of trauma may still be useful as an instrument.

Alternatively, motor vehicle accidents, which comprise a large portion

of the trauma admissions, is in the context of paediatric patients, a chance event. Since children do not drive, there is no obvious self-selection to motor vehicle accidents.

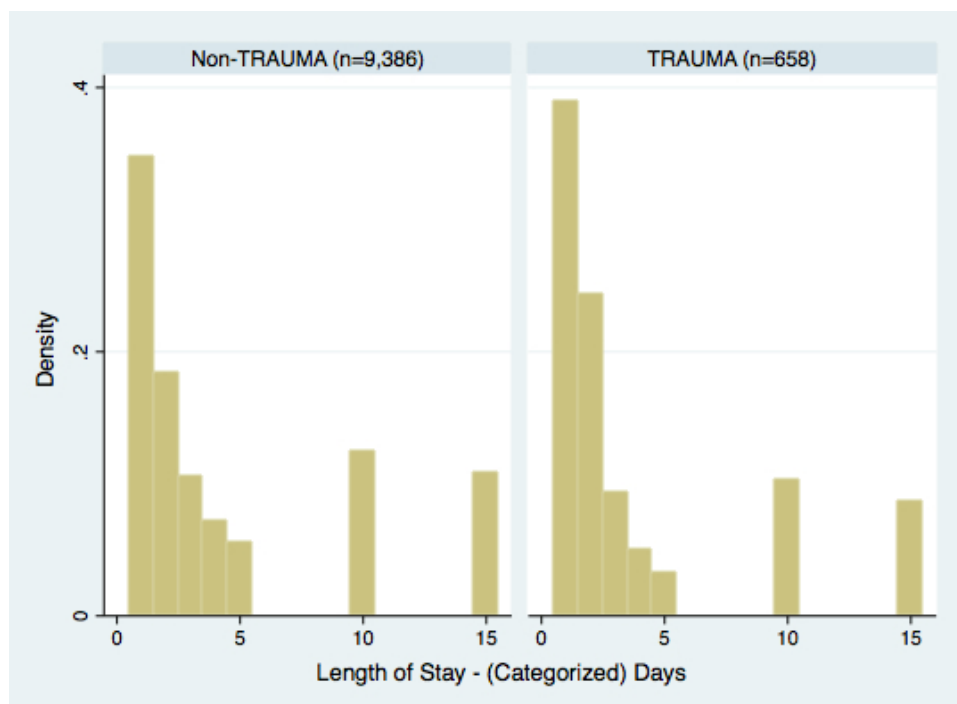
There are 658 trauma patients in the sample, comprising 6.55% of all the patients. A detailed breakdown of the number of patients in each of the eleven trauma aetiologies is,

(31) Bicycle	(5) Gun shot wound	(10) Self-inflicted: accidental
(21) Child abuse	(10) Knife wound	(11) Suicide attempt
(118) Falls	(23) Near drowning	(218) Motor vehicle accident
(3) Farm equipment	(185) Other	(658) Total

Our instrument, the trauma status of each patient, is coded as an indicator variable where the prevalent non-exclusive trauma etiologies are listed above. We assume that amongst children, in the context of a length of stay model, trauma is not correlated with the heterogeneous effect embodied in the structural error. The Canadian case provides a unique experimental-like setting, in that a homogeneous standard of care and observable patient characteristics across health delivery regions motivates the use of trauma as a proxy for randomization amongst otherwise unobservable heterogeneous

types.

If, indeed trauma randomizes over the otherwise unobserved types embedded in the structural error of, for example, model (3.1), we would expect trauma to have no observable relation to the outcome, in this case length of stay. The histogram for categorized lengths of stay suggest that in virtually



every portion of the distribution, there is no appreciable difference between the length of stay profile of trauma and non trauma patients. We view this as evidence that trauma is excluded from explaining length of stay.

4.4 Methods

The methods presented in the preceding sections of this thesis are utilized for the analysis of the length of stay and illness severity relationship. We use the accelerated life regression (2.8) as specified in §2.2.3, where for the specific analysis at hand, the outcome t = length of stay (Los_i) measured in hours, the marker of exposure of interest Y = the illness severity index ($PIM2_i$) or ($PRISMIII_i$), and the other covariates X_1 = patient's age category - infant, toddler, school age, and teenager ($Agecat_i$), an indicator for a chronic condition ($Chrndx_i$), and an indicator for previous admission to an intensive care unit ($Previcu_i$). The corresponding structural equation is,

$$\ln(LoS_i) = \delta_i + \beta PIM2_i + \delta_A Agecat_i + \delta_C Chrndx_i + \delta_P Previcu_i + \sigma \epsilon_i. \quad (4.1)$$

The other covariates, $Agecat_i$, $Chrndx_i$, $Previcu_i$, are known to be predictors of length of stay. Indeed previous pediatric studies have identified (i) respiratory therapy, (ii) pre-existing chronic condition, and (iii) previous admission to the PICU as risk factors for increased length of stay. The factors that were identified with decreased length of stay were, (i) elective procedures, (ii) post-operative status, (iii) increased age (Slater, Shan & Pearson, 2003), (Pollack, Patel & Ruttimann, 1996). Since PIM 2 and PRISM III

incorporate a number of these variables, we only include age, chronic conditions, and previous PICU admission as additional covariates. We do note, that model (4.1) is not a comprehensive approach, but rather a single component, to understanding the dynamics behind the movement of patients in and out of intensive care. For example a queueing model, for which model (4.1) provides one component, would better capture the effect of discharge practices and other hospital system influences.

4.5 Analysis, results and discussion

The respective statistics were programmed directly into the MATA matrix language as follows. For the GAR statistic, the critical values ($gar_{calc}(\alpha)$) were calculated as a function of the data (X_1, X_2) , and the random draws from the respective quantile function for the statistic (3.13) as detailed in §3.2.1. Next, the analytical construction of the confidence set for β was completed by solving the quadratic inequality using the data and calculated critical value. For comparative purposes, the resulting confidence sets are reported alongside (i) an uncorrected accelerated life analysis and (ii) a Gamma frailty analysis, both standard within the STATA package.

Model	Accelerated Life	Gamma Frailty	GAR (<i>size-controlled</i>)
Log-normal	(0.265, 0.293)	(0.248, 0.274)	(0.069, 0.194)
Log-logistic	(0.294, 0.321)	(0.287, 0.314)	(0.070, 0.193)
Weibull	(0.318, 0.352)	(0.275, 0.302)	(0.072, 0.191)

Table 4.1: 95% Confidence sets for PIM2 illness severity index.

For completeness we also report the regression results for both the accelerated life Weibull and log-logistic regression with and without the Gamma frailty correction for model (4.1). However, it is emphasized that following the arguments presented in §2.3.1, the other coefficient estimates are neither of interest, nor the focus of the analysis at hand. Indeed, as suggested by Stock (2010) we sharpen the focus on measuring a single effect well, rather than the vaguer agenda of developing a general model for length of stay. In this sense, the inclusion of the controls is simply a first attempt to proxy for the troublesome heterogeneity. It is also interesting to note that the point estimate for trauma in regression (4.1), supposing it were included, is .018 with a confidence interval $(-.082, .119)$. Again suggesting that trauma has no appreciable explanatory power, nor improves the fit, within the context of model (4.1). This further reinforces the arguments presented in §4.3.

	Weibull	Gamma frailty	Loglogistic	Gamma frailty
<i>Pim2</i>	.334 (.009)	.288 (.007)	.308 (.007)	.300 (.007)
<i>Agecat</i>	-.162 (.007)	-.149 (.006)	-.153 (.006)	-.152 (.006)
<i>Chrndx</i>	.347 (.024)	.190 (.019)	.204 (.020)	.192 (.019)
<i>Previcu</i>	.213 (.045)	.171 (.038)	.188 (.040)	.179 (.038)

Table 4.2: Coefficient estimates with standard errors for model (4.1).

The first analysis presented above was done with the PIM 2 illness severity score. The bounded robust GAR confidence sets are immediately informative on the quality of the instrument, in this analysis, all being in remarkably close agreement irrespective of model selection. An empty set implies model rejection, a set containing the entire real line implies a completely uninformative instrument, neither of which occur in our analysis. Although the bias correction of a gamma frailty modelling is in the same direction of the robust sets, the clinically relevant magnitude is notably different, giving relevant policy implications. The same analysis was then repeated for the other illness severity score PRISM III.

Model	Accelerated Life	Gamma Frailty	GAR (<i>size-controlled</i>)
Log-normal	(0.093, 0.101)	(0.093, 0.100)	(0.099, 0.316)
Log-logistic	(0.096, 0.104)	(0.095, 0.103)	(0.104, 0.305)
Weibull	(0.087, 0.097)	(0.092, 0.099)	(0.101, 0.313)

Table 4.3: 95% Confidence sets for PRISMIII illness severity index.

Again, the bounded confidence sets are informative on the quality of the instrument. However, interestingly with PRISM III the direction of the correction is opposite from the analysis with PIM 2.

The rank based analysis using the GAM statistic (3.16) was similarly programmed directly into the MATA matrix language. Again this consisted of calculating the critical values ($gam_{calc}(\alpha)$) for each of the eight GAM statistics. Since the rank statistic is distribution free, in principal any distribution could be used for the random draws in calculating the ranks. For ease and simplicity, we drew from the uniform (0,1) distribution. The random draws were then ranked and scored accordingly for each of the eight statistics. The four quantile scores simply scored using the respective quantile functions. The expected value scores used the respective formulas detailed in §3.3.1., again in a rather straightforward manor. The censored scores, however, re-

quired additional programming, as the risk set had to be adjusted on a per period basis. The procedure was repeated for one thousand simulations and ordered. The desired α -level critical value was selected from the appropriate position in the ordered simulated values.

To calculate the confidence set (3.18) the statistic $GAM(\beta_o, y, Y; X)$ was calculated for a grid of possible values for β_o . The data, y, Y, X_1 were used with the candidate values of β_o to calculate aligned variate values that were then ranked. The ranks were subsequently scored, as above, for each of the eight variants. The scores were then used with X_2 in calculating the value of statistic (3.16). All the grid values of β_o that were not rejected, that is, less than $gam_{calc}(\alpha)$, as calculated above, were collected to form the α -level confidence set. The least rejected value is reported with the confidence sets and may reasonably be considered a point estimate in the rank inference framework (Hodges & Lehmann, 1963). This “point-value” may or may not be the mid-point of the confidence set, as the collected values not rejected need not be symmetric. As an indication of the computational burden, the grid search conducted in increments of .005 over the possible values of the coefficient β (with starting values derived from an uncorrected regression), keeping in mind the sample size of 10,044, ran for eight hours using a dual-

core i7 processor running Stata/MP 13.1 multi-core. As above, the first analysis used the PIM 2 illness severity index as the marker of interest.

Quantile Scores			
Normal	Logistic	Gumbel	Exponential
(0.060, 0.170)	(0.065, 0.175)	(0.080, 0.210)	(0.095, 0.250)
0.115	0.120	0.145	0.175

Expected Value Scores		Censored Scores	
Wilcoxon	Savage	Wilcoxon	Savage
(0.040, 0.160)	(0.095, 0.250)	(0.070, 0.240)	(0.100, 0.335)
0.100	0.175	0.155	0.215

Table 4.4: Rank inference 95% confidence sets with least rejected point values for PIM 2 illness severity index.

The bounded confidence sets are consistent with the results from the least squares analysis. The normal and logistic scores result in very similar sets, with the logistic shifting ever so slightly towards the upper tail, as would be expected. The Gumbel and exponential do capture more of the tail behaviour. With the expected value scores, the Savage and exponential scores are identical, which is consistent with the fact that the Savage score is

based on drawing from the exponential distribution. However, the Wilcoxon does not exactly coincide in the same manner with the logistic, possibly not capturing as much tail behaviour. The censored scores correctly reflect the fact that despite a thankfully small number of deaths (3.5%), not accounting for the very ill who are censored in this manner, does understate the illness severity length of stay relationship. However, given the very small number of censored, not surprisingly the width of the confidence set has increased, but nevertheless providing meaningful information.

Despite the very small numbers, we further control for death by including the censoring indicator d in the vector of controls X_1 . Consistent with **Corollary 3.3.2**, our analysis explicitly allows for the correlation between those who die and their unobserved type. Intuitively, if we believe, for example a “frail” type would have a higher illness score and length of stay irrespective of that score, attributable solely to their type, then it seems logical that death would also be correlated with their “frail” type.

Quantile Scores			
Normal	Logistic	Gumbel	Exponential
(0.035, 0.160)	(0.040, 0.165)	(0.060, 0.205)	(0.090, 0.250)
0.100	0.105	0.135	0.170
Expected Value Scores		Censored Scores	
Wilcoxon	Savage	Wilcoxon	Savage
(0.030, 0.150)	(0.090, 0.250)	(0.055, 0.240)	(0.095, 0.335)
0.085	0.170	0.150	0.215

Table 4.5: Rank inference 95% confidence sets with least rejected point values for PIM 2 illness severity index, controlling for death.

The results, in all cases, shift the confidence sets and least rejected values towards a reduced effect, which would be the direction of correction we would expect if indeed those who died were systematically of a “frail” or related type. We believe this form of controlling for “endogenous” censoring to be a unique contribution with useful clinical relevance.

The analysis was repeated for the PRISM III illness severity score.

Quantile Scores			
Normal	Logistic	Gumbel	Exponential
(0.145, 0.480)	(0.115, 0.415)	(0.165, 0.455)	(0.175, 0.044)
0.295	0.255	0.295	0.295
<hr/>			
Expected Value Scores		Censored Scores	
Wilcoxon	Savage	Wilcoxon	Savage
(0.230, 0.850)	(0.180, 0.440)	(0.360, 0.950)	(0.190, 0.750)
0.395	0.295	0.600	0.375
<hr/>			

Table 4.6: Rank inference 95% confidence sets with least rejected point values for PRISM III illness severity index.

With PRISM III, there is a notable difference in the width and location of the robust confidence sets. Nevertheless, the comparative performance of the scores and censoring are similar to the PIM 2 analysis and in the same direction of correction as the least squares analysis for PRISM III. However, the informative content is decreased, keeping in mind that (i) over 30% of patients score zero on the PRISM III scale, which may simply reflect the fact that PRISM III is not as informative when analyzing the illness severity length of stay relationship.

Although the precision of the estimates, as reflected by the wider confidence sets, are comparatively lower, the censoring corrections do indeed provide useful policy relevant information. For example, any comparative type adjustments, whether for payment mechanisms, quality assurance or patient safety would be understating the adjustment for illness severity if censoring were to be ignored. Admittedly, controlling for death does not have an appreciable empirical impact, possibly given the low incidence of mortality. However, theoretically this innovation remains attractive given the higher incidence of mortality in other clinical settings.

4.6 Conclusion

Our procedure of inverting a pivotal robust test statistic is clinically informative without relying on a qualitative intermediate interpretation of the first stage regression, common to conventional instrumental methods. Moreover, being empirically motivated to employ a procedure that was robust to extreme values, we have found that the procedure extracts useful information from the extremes, be they (i) trauma (6.58% of sample), (ii) mortality (3.54% of sample) or (iii) long stay (12.18% of sample), all of which would

in practice, otherwise be excluded from analysis.

Duration outcomes are central to many clinical studies. Dealing with unobserved factors in such models has produced a number of specialized methods. Our clinical analysis provides practitioners with a complementary, unique, and robust method to exploit other exogenous variation that may indeed be present, via the method of instrumental variables. Our clinical analysis has a practical example of instrument selection. A known limitation of the instrumental variables methodology is the availability of appropriate instruments. Accordingly, our prescription to practitioners is to employ robust econometric techniques combined with careful instrument selection. In such a manner, the method of instrumental variables offers a very useful tool in disentangling confounding effects in the clinical setting.

From a theoretical and methodological perspective, our proposed inference strategy has a number of promising directions for future research. The most prominent of which is extending the censored rank test to the larger class of general transformation models.

Bibliography

AALEN, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

AALEN, O. O., BORAN, Ø & GJESSING, H. K. (2008). *Survival and event history analysis: a process point of view* Springer, New York

AALEN, O. O., COOK, R. J. & RØYSLAND, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* **21**, 579–593.

ANDERSON, T. W. & RUBIN, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics* **20(1)**, 46–63.

ANDREWS, D. W. K. & MARMER, V. (2008). Exactly distribution free

- inference in instrumental variables regression with possibly weak instruments. *Journal of Econometrics* **142**, 183–200.
- BOUND, J., JAEGER, D. A. & BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **904**, 443–450.
- BOX, G. E. P & COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society* **26**, 211–252.
- CAMERON, C. & TRIVEDI, P. (2005). *Microeconometrics*, Cambridge University Press.
- CAWLEY, J. (2015). A selective review of the first twenty years of instrumental variables models in health services research and medicine. *Journal of Medical Economics* **18**, 721–734.
- CHEN, S. (2002). Rank estimation of transformation models. *Econometrica* **70**, 1683–1697.
- CHERNOFF, H. & SAVAGE, I. R. (1958). Asymptotic normality and effi-

- ciency of certain non-parametric test statistics. *The Annals of Mathematical Statistics* **29**, 972–994.
- CHIAPPORRI, P. A., KOMUNJER, A. & KRISTENSEN, D. (2015) Nonparametric identification and estimation of transformation models. *Journal of Econometrics* **188**, 22–39.
- COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* **34**, 187–220.
- COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- COX, D. R. & ISHAM, V. (1980). *Point processes*. London: CRC Press.
- COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.
- DUFOUR, J. M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* **65**, 1365–1387.
- DUFOUR, J. M. & TAAMOUTI, M. (2005). Projection-based statistical in-

- ference in linear structural models with possibly weak instruments. *Econometrica* **4**, 1351–1365.
- ELBERS, C. & RIDDER, G. (1990). True and spurious duration dependence: The identifiability of the proportional hazards model. *Review of Economic Studies* **49**, 402–411.
- FISHER, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
- HAN, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics* **35**, 303–316.
- HECKMAN, J. J. & SINGER, B. (1984). The identifiability of the proportional hazards model. *Review of Economic Studies* **60**, 231–243.
- HODGES, J. L. & LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics* **33(2)**, 482–497.
- HODGES, J. L. & LEHMANN, E. L. (1963). Estimates of location based on ranks. *The Annals of Mathematical Statistics* **34(2)**, 598–611.
- HOROWITZ, J. L. (1996). Semiparametric estimation of a regression model

- with an unknown transformation of the dependent variable. *Econometrica* **64**, 103–137.
- HOSMER, D. W., LEMESHOW, S. & MAY, (2008). *Applied survival analysis: regression modelling of time-to-event data*. John Wiley & Sons.
- HOUGAARD, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71**, 75–84.
- HSU, B. S., LAKHANI, S. & BRAZELTON, T. B. (2015). Relationship Between Severity of Illness and Length of Stay on Costs Incurred During a Pediatric Critical Care Hospitalization. *S D Med* **68**, 339–344.
- KALBFLEISCH, J. D. & PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.
- KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association* **53**, 457–481.
- KEIDING, N., ANDERSON, P. K. & KLEIN, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16**, 215–224.

- KLEIN, J. P. & MOESCHBERGER, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd ed.
- KLEINBAUM, D. G. & KLEIN, M. (2012). *Survival Analysis*. Springer.
- LANCASTER, T. (1979). Econometric methods for the duration of unemployment. *Econometrica* **47**, 939–956.
- LEEB, K., JOKOVIC, A., SANDHU, M. & AINCK, G.(2006). CIHI Survey: Intensive Care in Canada. *Healthcare Quarterly* **9**, 32–33.
- LEHMAN, E. L. (1986). *Testing Statistical Hypotheses*. Wiley.
- MIKUSHEVA, A. (2013). Survey on statistical inferences in weakly-identified instrumental variables models. *Applied Econometrics* **29**, 116–131.
- POLLACK, M. M., PATEL, K. & RUTTIMANN, U. E. (1996). PRISM III: An updated Pediatric Risk of Mortality score. *Critical Care Medicine* **24**, 743–752.
- POLLACK, M. M., WILKINSON, J. D. & GLASS N. L. (1987). Long-stay pediatric intensive care unit patients: outcome and resource utilization. *Pediatrics* **80**, 855–860.

- POLLACK, M. M., YEH T. S., RUTTIMANN, U. E. HOLBROOK P. R. & FIELDS A. I.(1984). Evaluation of pediatric intensive care. *Critical Care Medicine* **12**, 376–383.
- PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167–179.
- RANDLES, R. H. & WOLFE, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. John Wiley & Sons.
- REID, N. (1994). A Conversation with Sir David Cox. *Statistical Science* **9**, 439–455.
- RIDDER, G. (1990). The non-parametric identification of generalized accelerated failure-time model. *Review of Economic Studies* **57**, 162–182.
- RUBIN, D. B. (1990). Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference* **25**, 279–292.
- RUTTIMANN, U. E. & POLLACK, M. M. (1996). Variability in duration of stay in pediatric intensive care units: a multi-institutional study. *J Pediatrics* **128**, 35–44.

- SAVAGE, I. R. (1956). Contributions to the theory of rank order statistics – the two-sample case. *Ann. Math. Statist.* **27**, 590–615.
- SLATER, A., SHAN, F. & PEARSON, G. (2003). PIM2: A revised version of the paediatric index of mortality. *Intensive Care Medicine* **29**, 278–285.
- STOCK, J. H. (2010). The other transformation in econometric practice: Robust tools for inference. *Journal of Economic Perspectives* **17**, 177–194.
- STOCK, J. H. & TREBBI, F. (2003). Retrospectives: Who invented instrumental variables regression? *The Journal of Economic Perspectives* **17**, 177–194.
- STRUTHERS, C. A. & KALBFLEISCH, J. D. (1986). Misspecified proportional hazards models. *Biometrika* **74**, 363–369.
- VAN DER WAERDEN, B. L. (1953). Ein neuer Test für das Problem der zwei Stichproben. *Math. Annalen* **126**, 93–107.
- VAUPEL, J., MANTON, K. & STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.
- WEI, L. J. (1992). The accelerated failure time model: a useful alternative

to the Cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871–1879.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.