Endogeneity in parametric duration models with applications to clinical risk indices

A. Acharya[*], L.Khalaf[*], M. Voia[*] D. Wensley[†]

February 15, 2016

Abstract

We consider the problem of unmeasured confounding in an accelerated life regression model. The proposed inferential method, 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. We provide an empirical illustration with an original prospectively collected observational data set, 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.

1 Introduction

The analysis of many research studies is often complicated by the presence of unobserved factors that may affect both the exposure and the outcome. Such complications are usually dealt with by using a randomized controlled trial study design in which, given a large enough sample size, the confounders are presumably equally distributed between groups. However, in observational studies or in smaller randomized controlled trials, analyses rely on controlling

[*]Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Canada.

[†]Department of Pediatrics, University of British Columbia, British Columbia Children's Hospital, 4480 Oak Street, Vancouver, Canada.

for confounding using, among others, various regression techniques. These techniques only correct for known confounders and as such, researchers are still left with the possibility that an unknown or unmeasured confounder has led to the observed results. Duration outcomes are in particular susceptible. In the presence of unobserved heterogeneity, a relative risk duration model suffers from the confounding of the baseline hazard, irrespective of correlation with the exposure. Alternatively, a frailty accelerated life model does not suffer from unobserved heterogeneity with the same severity (Keiding, Anderson & Klein, 1997) however, continues to be complicated in the presence of unobserved factors affecting both exposure and outcome. We propose an instrumental variables approach to correct for unmeasured confounding in an accelerated life model and provide an illustration with an original prospective observational data set collected to analyze the relationship between illness severity and length of stay in a specific pediatric intensive care population.

We set focus on providing a method robust to instrument quality in the context of a duration model, which does not appear to have received attention in the literature. 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) and presumed (ii) relevant (i.e. sufficiently correlated with the exposure), 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 of, for example, illness severity that are compatible with this assumption. 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 informative content of the data (Imbens & Rosenbaum, 2005). Surprisingly, despite two decades of work on identification robust methods, the duration case was not analyzed, and particularly not from a finite sample perspective.

2 Exact semi-parametric inference

2.1 Model

An n-vector of possibly right-censored duration outcomes $t \in R^+$ is observed, along with $(n \times g)$ confounded interventions or markers Y, further $(n \times k_1)$ controls X_1 including intercept, and other $(n \times k_2)$ instrumental variables X_2. An assumed acceleration factor $exp(Y\beta + X_1\delta)$ gives the accelerated life regression, in which $y \equiv \log t$:

$$y = Y\beta + X_1\delta + \sigma\epsilon, \tag{1}$$

with random disturbance ϵ, where $\beta \in R^g$, $\delta \in R^{k_1}$, and $\sigma \in R$ are unknown parameters.

Assumption 2.1 (Exogeneity) $E(\epsilon \mid X_2) = 0$, holding *a fortiori* for a randomly assigned instrument.

Our interest is in constructing a confidence set for β. For this purpose, we invert an appropriate test statistic associated with the null hypothesis:

$$H_0 : \beta = \beta_0, \quad H_1 : \beta \neq \beta_0. \tag{2}$$

Under *Assumption 1* and within the linear Gaussian framework, Anderson & Rubin (1949) proposed inverting a least squares test that assesses the exclusion of the instruments in an auxiliary regression as formally defined below, which rather than describing a statistical model *per se*, serves as a computational tool. Andrews & Marmer (2008) introduce the rank analogue of this test. For the particular features of duration analysis, we generalize this inference strategy to (i) non-Gaussian errors in the least squares model, and (ii) aligned linear rank test statistics, as derived from the accelerated failure time model with possible right-censoring.

2.2 Least Squares Inference

To obtain a confidence set on β, we invert a generalized Anderson-Rubin test obtained from the uncensored auxiliary regression:

$$y - Y\beta_o = X_1\lambda + X_2\gamma + \omega, \tag{3}$$

where ω is an $(n \times 1)$ vector of random disturbances. If instrument exclusion holds, we would expect the coefficient on X_2 to be zero, which, moreover is implied by imposing $H_0 : \beta = \beta_0$

in model (1). Accordingly, in the context of the solely computational model (3), to test the hypothesis of the form $H_o : \gamma = 0$, the generalized Anderson-Rubin statistic is:

$$GAR(\beta_o) = \frac{(y - Y\beta_o)'(M_1 - M)(y - Y\beta_o)/k_2}{(y - Y\beta_o)'M(y - Y\beta_o)/(n - k)}, \tag{4}$$

where $M = I - X(X'X)^{-1}X'T$, in which $X = [X_1, X_2]$ and $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$.

Theorem 2.2 *Under the null hypothesis imposing model (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*

$$\overline{GAR}(\epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon' M\epsilon/(n - k)}. \tag{5}$$

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 (1). The distribution is invariant to β_0, σ, and the data generating process linking Y to X_2. Consequently, we extend Dufour & Taamouti (2005), with analytical construction of exact confidence sets achieved via simulating the exact null distribution:

$$C_\beta(\alpha) = [\beta_o : GAR(\beta_o) < gar_{calc}(\alpha)], \tag{6}$$

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

$$F_{k_2,n-k}(\alpha), \quad \epsilon_l = ln(\frac{u_l}{1 - u_l}), \quad \epsilon_l = -ln(-ln(u_l)),$$

in the log-normal, log-logistic, and Weibull case respectively, where the n-vector $u_l \sim$Uniform$[0, 1]$ for each draw $l = 1, \ldots, m$ simulations of $\overline{GAR}_l$. Approximate confidence sets are available via $\chi^2_{k_2}(\alpha)$, under standard least squares motivated regularity conditions.

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.

2.3 Aligned rank inference with right censoring.

We draw 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 (1). These test statistics, 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 distribution-free and moreover, robust to general misspecification and extreme values. The aligned rank auxiliary regression, analogous to (2):

$$rank(y - Y\beta_o - x_1\hat{\delta}(\beta_o)) = x_2\gamma + \omega, \quad \hat{\delta} = (x_1'x_1)^{-1}x_1'(y - Y\beta_o), \tag{7}$$

in which $\hat{\delta}$ is the null restricted estimator with x_1, x_2 expressed as X_1, X_2 with columns standardized to add to zero. In this form we have 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, the rank vector is considered exchangeable.

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), giving the rank analogue of the Anderson-Rubin statistic which we refer to as the generalized Andrews-Marmer statistic:

$$GAM(\beta_o) = c'\,(p_2)\,c, \tag{8}$$

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

The score vector satisfy a nondecreasing and nonconstant condition, $c^{(i)} \leq ... \leq c^{(n)}$ and $c^{(i)} \neq c^{(n)}$, where (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 and may indeed be misspecified in the presence of unmeasured confounding. Two related and asymptotically equivalent scores (Randles & Wolfe, 1979) are the quantile F_o scores and the expected value F_o scores:

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

where $V^{(i)}$ is the ith 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 know classical expected value scores are those of Wilcoxon (1945) and Savage (1956):

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

Utilizing the framework of Prentice (1978), we provide an analytically tractable right-censored generalization under the assumption of the independence of the censoring mechanism and the outcome, y. The composite expected value scores are, in the case of Wilcoxon scores:

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

For the Savage scores:

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

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 2.3 *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*

$$\overline{GAM}(u; X_2) = c(rank(u))'\,(p_2)\,c(rank(u)), \tag{9}$$

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 an example Stock (2010) suggests the role of controls is to satisfy $E(\epsilon|Y, X_1) = E(\epsilon|X_1)$. Similarily, 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. In the context of controlling for death, including an *ex post* indicator for mortality in the vector of controls effectively accounts for censoring that need not be independent of the data generating process for the outcome y.

Corollary 2.4 *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:

$$C_\beta(\alpha) = [\beta_o : GAM(\beta_o, y, Y; X) < gam_{calc}(\alpha)], \tag{10}$$

where $gam_{calc}(\alpha)$, for an α significance level, is exactly simulated by, as an example, the n-vector $u_l \sim$Uniform$[0,1]$ for each draw $l = 1, \ldots, m$ simulations of $\overline{GAM}_l$. Again, approximate confidence sets are available via $\chi^2_{k_2}(\alpha)$ (Hajek & Sidak, 1999).

The aligned rank method is also compatible with non-deterministic possibly discrete transformations of the observed time, provided they are cast as an estimating function around the hypothesized β_o. For example, the proportional hazard model has a linear representation in which the transformation of observed time is variously approximated by a step function of the data and β_o.

3 Monte Carlo

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

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

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 $h(\cdot)$ relaxes the first stage linearity, as non-linearity is not uncommon in the clinical setting.

In general with moderately strong instruments or better, power approaches one with sample sizes greater than 150 for both the least squares and rank statistic. 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 above. Alternatively, if the vector X_1 are strictly exogenous or random, then there is a decreased power difference between the tests. With very weak instruments the power of the least squares statistic is possibly non standard; in certain cases an increasing sample size may not translate to

	Least Squares Statistic		
	Normal	Logistic	Gumbel
Log-normal	09 19 55	09 19 55	12 27 62
Log-logistic	06 12 29	06 12 29	09 18 37
Weibull	14 28 56	14 28 56	19 35 64

	Rank Statistic					
	Quantile Scores				Expected Value Scores	
	Normal	Logistic	Gumbel	Exponential	Wilcoxon	Savage
Log-normal	17 38 79	17 38 79	16 38 78	16 35 71	16 37 79	16 35 70
Log-logistic	10 18 41	09 18 41	09 17 38	09 16 35	09 18 41	09 16 35
Weibull	17 34 67	17 34 67	15 30 59	12 25 48	17 32 65	13 24 45

Table 1: Percent power weak-moderate-strong instrument. Sample size 100.

increased power, in particular, as the distance from the null increases or in the presence of non-linearities. Therefore, 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.

4 Application

As an illustration 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 consisting of 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. These specific characteristics are used to determine an illness severity index (Slater, Shan & Pearson, 2003) 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.

Our instrument, the trauma status of each patient, is coded as an indicator variable where the prevalent non-exclusive trauma etiologies are; bicycle accidents, motor vehicle accidents, near drownings, falls, and farm equipment accidents. 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.

The bounded robust 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 (Abbring & Van Den Berg, 2007) is in the same direction of the robust sets, the clinically relevant magnitude is notably different, giving relevant policy implications. A further correction for censoring shows that despite the low mortality rate, the censored sets, although being slightly wider, are usefully informative in reflecting the joint increase in length of stay and illness severity induced by this group.

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.

			size-controlled
Model	Accelerated Life	Gamma Frailty	Generalized Anderson-Rubin
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 2: Confidence sets for Pediatric index of mortality (PIM2) illness severity index.

	Quantile Scores			
	Normal	Logistic	Gumbel	Exponential
95% Confidence sets	(0·060, 0·170)	(0·065, 0·175)	(0·080, 0·210)	(0·095, 0·250)
Least Rejected	0·115	0·120	0·145	0·175
	Expected Value Scores		Censored Scores	
	Wilcoxon	Savage	Wilcoxon	Savage
95% Confidence sets	(0·040, 0·160)	(0·095, 0·250)	(0·070, 0·240)	(0·100, 0·335)
Least Rejected	0·100	0·175	0·155	0·215

Table 3: Rank inference confidence sets for Pediatric index of mortality (PIM2) illness severity index.

References

Abbring, J. H. & Van Den Berg, G. J. (2007). The unobserved heterogeneity distribution in duration analysis. *Biometrika* **94(1)**, 87–99.

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.

Chernoff, H. & Savage, I. R. (1958). Asymptotic normality and efficiency of certain non-parametric test statistics. *The Annals of Mathematical Statistics* **29**, 972–994.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. 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 inference in linear structural models with possibly weak instruments. *Econometrica* **4**, 1351–1365.

Hajek, J. , Sidak, Z. & Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press.

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.

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.

Imbens, G. W. & Rosenbaum, P. R. (2005). Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education. *J. R. Statist. Soc. A* **168(1)**, 109–126.

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.

Rubin, D. B. (1990). Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference* **25**, 279–292.

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.

van der Waerden, B. L. (1953). Ein neuer Test für das Problem der zwei Stichproben. *Math. Annalen* **126**, 93–107.

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