Endogeneity in parametric duration models with applications to clinical risk indices.

Anand Acharya[1] Lynda Khalaf [1]
Marcel Voia [1] David Wensley [2]
and The Canadian Critical Care Trials Group

[1]Department of Economics
Carleton University

[2]Department of Pediatrics
University of British Columbia

15th OxMetrics User Conference, London, U.K.
4 September 2014

Overview.

This paper provides an exact, identification robust method to correct for endogeneity in the location-scale family of parametric duration models using instrumental variables (IV) and the generalized Anderson Rubin (GAR) statistic, with applications to clinical risk indices.

- **Data** ⇒ Extreme valued.
- **Model** ⇒ Parametric location-scale family duration models; log-normal, log-logistic, Weibull.
- **Methods** ⇒ IV and the GAR statistic.
- **Empirical Application** ⇒ Length of stay (LoS) in the pediatric intensive care unit (ICU) and clinical risk index (*PIM2*).[1]

[1] Slater et. al. (2003)

Unique contributions.

This paper makes a unique contribution to the following 4 areas of study:

1. **Duration literature** $\Rightarrow$ corrects for endogeneity in one class of common duration models.
2. **Identification robust inference (IdR) literature** $\Rightarrow$ time to event outcomes.
3. **Simulation based inference literature** $\Rightarrow$ location-scale family distributions.
4. **Clinical health literature** $\Rightarrow$ provide the instrument: $Trauma_i$ for use in ICU studies.

Clinical risk indicies.

- Risk of mortality scores are routinely calculated at the time of admission.
- Typically derived from logistic regressions, and generally accepted as a proxy for illness severity.
- By construction, they omit individual specific effects, that are often unobservable.
- Nevertheless, they are extensively employed in risk-adjusting outcomes and stratifying patients.

Accelerated life models.[2]

Underlying assumption is covariates "accelerate" or "decelerate" observed $(n \times 1)$ time, t, by a constant factor, $exp(Y\beta + X_1\pi)$. Expressed as a transformation model:

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

where $y \equiv ln(t)$ is the $(n \times 1)$ vector of transformed durations, Y is the $(n \times 1)$ vector of observed risk scores, X_1 is the $(n \times k_1)$ matrix of observed covariates including intercept, and $\sigma\epsilon$ is the scaled $(n \times 1)$ vector of i.i.d. unobservables.

[2] Cox and Oakes (1984)

Distributional assumptions.

Different distributional assumptions on accelerated time lead to well known parametric duration models, in which the scaled unobservables follow their respective transformed distributions:

- $Lognormal(exp(\pi_1), \sigma^2) \rightarrow \epsilon \overset{iid}{\sim} Normal(0, 1)$,
- $Loglogistic(exp(\pi_1), \sigma) \rightarrow \epsilon \overset{iid}{\sim} Logistic(0, 1)$,
- $Weibull(exp(\pi_1), \frac{1}{\sigma}) \rightarrow \epsilon \overset{iid}{\sim} Gumbel(0, 1)$

where the *Lognormal* location, *Loglogistic* location, and *Weibull* scale parameters are respectively captured in the transformed regression intercept, π_1.

Structural model assumptions.

- Exogeneity
$$\mathbb{E}(\epsilon|X_1) = 0.$$
- Endogeneity
$$\mathbb{E}(\epsilon|Y) \neq 0.$$
- Availability of an $(n \times k_2)$ instrument, X_2 that satisfies:
$$\mathbb{E}(\epsilon|X_2) = 0.$$
- We explicitly make no assumptions on the data generating process that links Y and X_2.

Instrument X_2: Trauma status of patient.

Our contribution is to select the trauma status of a patient as an instrument.

- Trauma is an indicator variable for intensive care admission modality (n=644).
- Trauma etiologies:
  - Motor vehicle accidents.
  - Bicycle accidents.
  - Farm equipment accidents.
  - Near drownings.
  - Falls.
  - Child abuse.

Trauma: Intuition for random assignment.

- The selection of the instrument was based on the intuition that a patient that suffered a trauma was *as good as randomly assigned*.
- In other words, patients would otherwise not have a predisposition for trauma.
- The heterogenous types (i.e. frail or strong) would be equally as likely to suffer a trauma.

Anderson Rubin statistic.[3]

To obtain a confidence set on β, we invert the AR statistic associated with imposing $\beta = \beta_o$ in model (1). This implies testing the exclusion of the instruments in an auxiliary regression, which rather than describing a statistical model per se, serves as a computational tool:

$$f(y, Y, \beta_o) = X_1\zeta + X_2\gamma + \omega, \tag{2}$$

where ω is an $(n \times 1)$ vector of i.i.d disturbances and we further define:

$$f(y, Y, \beta_o) \equiv \; y - Y\beta_o. \tag{3}$$

[3] Anderson and Rubin (1949), Dufour(1997), Staiger and Stock(1997)

Test statistic.

Accordingly, to test the hypothesis of the form $H_o : \gamma = 0$, the test statistic is:

$$GAR(data; \beta_o) = \frac{RSS(\beta_o)_c - RSS(\beta_o)_u / k_2}{RSS(\beta_o)_u / (n-k)}, \tag{4}$$

where $RSS(\beta_o)_c$ is the residual sum of squares from the constrained regression, $RSS(\beta_o)_u$ is the residual sum of squares from the unconstrained regression and $k = (k_1 + k_2)$.

GAR Statistic.

Which gives the generalized Anderson-Rubin statistic:

$$GAR(data;\beta_o,) = \frac{(y - Y\beta_o)'(M_1 - M)(y - Y\beta_o)/k_2}{(y - Y\beta_o)'M(y - Y\beta_o)/(n-k)}, \qquad (5)$$

where $M = I$ - $X(X'X)^{-1}X'$ and $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$, in which $X = [X_1, X_2]$.

Pivotality.

Under the null hypothesis, the *GAR* statistic may be written as a function of the standardized structural error, ϵ.

$$\overline{GAR}(\beta_o, \epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon' M\epsilon/(n-k)}, \tag{6}$$

for which, the following theorem holds:

Theorem 1.

Under the null hypothesis, imposing model (1) at the true parameter value of $\beta = \beta_o$, the distribution of the GAR statistic is completely determined by the null distribution of $\overline{GAR}$, conditioned on X.
We note that the *null distribution* is:

1. Completely determined by the standardized distribution of the structural error, ϵ.
2. Invariant to the data generating process linking Y and X_2.
3. Invariant to β_0.
4. Invariant to scale, σ.

Furthermore, the statistic is nuisance parameter free and accordingly, pivotal.

Simulating AFT models.

Different distributional assumptions on the structural error lead to the various AFT models:

- $\mathit{Lognormal} \Rightarrow \epsilon \overset{iid}{\sim} N(0,1) \Rightarrow \overline{GAR} \sim F_{(k_2,(n-k))}$.
- $\mathit{Weibull} \Rightarrow \epsilon \overset{iid}{\sim} \mathit{Gumbel}(0,1) \Rightarrow \epsilon_j = ln(ln(\mathbf{u}_j))$.
- $\mathit{Loglogistic} \Rightarrow \epsilon \overset{iid}{\sim} \mathit{Logistic}(0,1) \Rightarrow \epsilon_j = ln(\frac{\mathbf{u}_j}{1-\mathbf{u}_j})$

Simulating null distribution.

For each draw j, where the $(n \times 1)$ vector $\mathbf{u}_j$ is drawn from the uniform [0,1] distribution, we have the jth realization of the GAR statistic:

$$\overline{GAR}_j = \frac{(\epsilon_j)'(M_1 - M)(\epsilon_j)/k_2}{(\epsilon_j)'M(\epsilon_j)/(n-k)}. \tag{7}$$

- Repeat for j=*1..J*.
- Construct the simulated null distribution.
- Appropriate α-level cut off value, $gar_{calc}(\alpha)$ may subsequently be utilized in the confidence set construction.

Confidence set construction.[4]

To construct a confidence set on β_o, we invert equation (5) using the appropriate α-level cut off:

$$C_\beta(\alpha) = \{\beta_o : GAR(\beta_o, y, Y; X) < F_{k_2,n-k}(\alpha) \text{ or } gar_{calc}(\alpha)\}, \quad (8)$$

where α is the specified significance level, resulting in the *quadric confidence set*:

$$C_\beta(\alpha) = \{\beta_o : \beta_o' A\beta_o + b'\beta_o + c \leq 0\}, \quad (9)$$

[4] Dufour and Jasiak (2001), Dufour and Taamouti (2005)

in which,

$$A = Y'BY, \quad b = -2Y'By, \quad c = y'By,$$

where,

$$B = M_1 - [1 + r(\alpha)]M, \ \textit{and} \ r(\alpha) = \frac{k_2 gar_{calc}(\alpha)}{(n-k)}.$$

The analytical solution results from solving the quadratic inequality. We emphasize that the solution permits sets that are *closed, open, empty, or the union of two or more disjoint intervals.*[5]

[5] Dufour(1997)

Results: Size corrected confidence sets.

Model	*PIM2* 95% Confidence sets
AFT Lognormal (mle)	(.265 , .293)
AFT Loglogistic (mle)	(.294, .321)
AFT Weibull (mle)	(.318, .352)
IV	(.072 , .193)
GAR Lognormal	(.070 , .193)
GAR Loglogistic	(.067, .196)
GAR Weibull	(.069 , .194)

Table : *PIM2* 95% confidence sets. (n=10,044).

Discussion.

- Using *Trauma* as an instrument, we conclude from the IV regression that a conventional AFT regression provides biased estimates, in this case over-stating the effect of increased illness severity on length of stay.
- The size-corrected Anderson Rubin confidence sets, are strong evidence suggesting that *Trauma* does not suffer a weak instrument problem.
- Bridging the duration, identification robust, and clinical health literatures, the use of identification robust IV techniques provide a statistical procedure for overcoming endogeneity in the illness severity and length of stay relationship.

Conclusion and contributions.

1. A method to correct for endogeneity in a common class of duration models.
2. To the best of our knowledge, provided a first application of identification robust methods to time to event outcomes.
3. Proved the null distribution of the generalized Anderson Rubin statistic holds for the location-scale family distribution, giving exact simulation results.
4. Introduced *Trauma* as an instrument for future clinical studies in the critical care setting.

Proof of Theorem 1.

We need to show how:

$$\overline{GAR}(\beta_o, \epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon' M\epsilon/(n-k)}, \tag{10}$$

is derived from:

$$T(\gamma_o, f(y, Y, \beta_o)) = \frac{\hat{\omega}_c'\hat{\omega}_c - \hat{\omega}_u'\hat{\omega}_u/k_2}{\hat{\omega}_u'\hat{\omega}_u/(n-k)}, \tag{11}$$

which is the usual statistic for testing $H_o : \gamma = 0$ in model (2), 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.

Cont'd.

First we use the fact that:

$$\hat{\omega}_c = M_1 f(y, Y, \beta_o),$$

$$\hat{\omega}_u = Mf(y, Y, \beta_o),$$

where M_1 and M as defined in (5) are symmetric and idempotent, giving:

$$GAR(\beta_o, f(y, Y, \beta_o)) = \frac{f(y, Y, \beta_o)'(M_1 - M)f(y, Y, \beta_o)/k_2}{f(y, Y, \beta_o)'Mf(y, Y, \beta_o)/(n-k)}. \tag{12}$$

Cont'd.

Under the null hypothesis, where (1) is the true model, we have:

$$y - Y\beta_o = X_1\pi + \sigma\epsilon. \tag{13}$$

Using the fact:

$$M_1X_1 = 0,$$

$$MX_1 = 0,$$

Cont'd.

and as specified in the artificial regression:

$$f(y, Y, \beta_o) \equiv y - Y\beta_o, \tag{14}$$

we have:

$$GAR(\beta_o, \sigma\epsilon; X) = \frac{\sigma\epsilon'(M_1 - M)\sigma\epsilon/k_2}{\sigma\epsilon' M\sigma\epsilon/(n-k)}. \tag{15}$$

This gives, the pivotal statistic:

$$\overline{GAR}(\beta_o, \epsilon; X) = \frac{\epsilon'(M_1 - M)\epsilon/k_2}{\epsilon' M\epsilon/(n-k)}. \tag{16}$$