PLSC 504

Sample Selection Models, I

November 14, 2017

Sample Selection In Theory

- Challenge: Inference to a Population from a Non-Random Sample
- Widespread Problem...
  - Heckman's wage equations...
  - Self-selection (e.g., into groups)
  - Surveys: "Screening" questions (*sometimes*...)
- Parallels in Missing Data, Causal/Counterfactual Inference

Sample Selection Basics

Observe:

$$
\begin{aligned}
Y_{1i}^* &= \mathbf{X}_i\boldsymbol{\beta} + u_{1i} \\
Y_{2i}^* &= \mathbf{Z}_i\gamma + u_{2i}
\end{aligned}
$$

$$
Y_{1i} = \begin{cases} Y_{1i}^* \text{ if } Y_{2i}^* > 0 \\ \text{missing if } Y_{2i}^* \leq 0 \end{cases}
$$

- Y_{2i}^* unobserved (except for sign);
- $\mathbf{X}_i$ observed iff Y_{1i} is observed;
- $\mathbf{Z}_i$ observed in every case.

Sample Selection Basics

$$
\begin{aligned}
\Pr(Y_{2i}^* \leq 0|\mathbf{X},\mathbf{Z}) &= \Pr(u_{2i} \leq -\mathbf{Z}_i\gamma) \\
&= 1 - \Pr(u_{2i} \geq -\mathbf{Z}_i\gamma) \\
&= 1 - \Pr(-u_{2i} \leq \mathbf{Z}_i\gamma) \\
&= 1 - \int_{-\infty}^{\mathbf{Z}_i\gamma} f(u_2)du_2 \\
&= 1 - F_{u_2}(\mathbf{Z}_i\gamma)
\end{aligned}
$$

Sample Selection Basics

Define:

$$D_i = \begin{cases} 1 & \text{if } Y_{1i} \text{ is observed.} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\Pr(D_i = 1) = F_{u_2}(\mathbf{Z}_i\gamma).$$

An Assumption

$$\{u_1, u_2\} \sim \mathcal{BVN}(0, 0, \sigma_1^2, 1, \sigma_{12})$$

Means

$$\Pr(D_i = 1 | \mathbf{Z}_i, \mathbf{X}_i) = \Phi(\mathbf{Z}_i \gamma).$$

Define:

$$\rho = \text{corr}(u_1, u_2).$$

Selection *Bias*

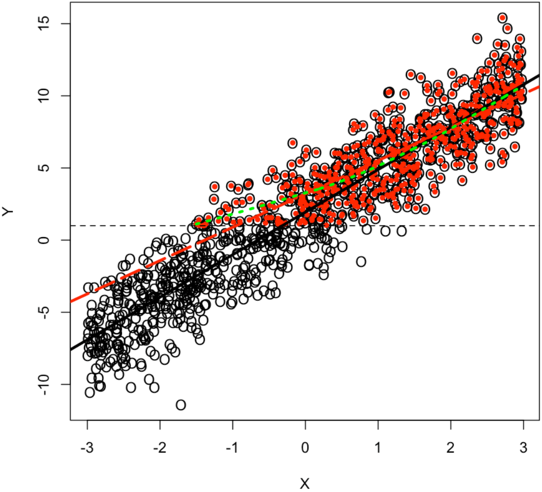
What we get:

$$\mathrm{E}(Y_{1i}|\mathbf{X}_i, \mathbf{Z}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right]$$

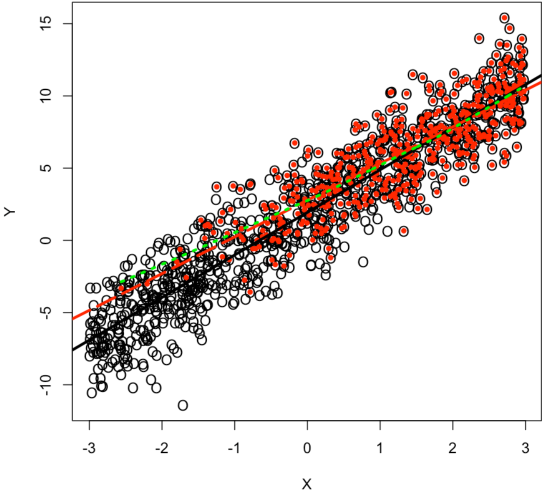
Without conditioning on $\mathbf{Z}$:

$$\mathrm{E}(Y_{1i}|\mathbf{X}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{E}\left\{\rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right]\middle|\mathbf{X}_i\right\}$$

Truncation Bias



Sample Selection Bias



Selection Bias: Substantive Effects

- *Specification Error* (unless $\rho = 0$)
- Indeterminate bias in $\hat{\boldsymbol{\beta}}$
- Including $\mathbf{Z}_i$ will not generally* remove the bias
- *Bias remains even if inference is limited to the "selected" group.* (This point is made nicely in Berk (1983)...)

*...unless sample selection is completely deterministic (i.e., determined by $\mathbf{X}$, $\mathbf{Z}$) (Heckman & Robb 1985).

E(Y) Under Selection

Conditional Density:

$$h(Y|\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\gamma,\sigma_1,\rho) = \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1\Phi(\mathbf{Z}_i\gamma)} \cdot \Phi\left[\frac{\frac{\rho(Y_{1i}-\mathbf{X}_i\boldsymbol{\beta})}{\sigma_1} + \mathbf{Z}_i\gamma}{\sqrt{1-\rho^2}}\right]$$

Note: $\rho = 0$ yields

$$\begin{aligned} h(Y|\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\gamma,\sigma_1,\rho=0) &= \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1\Phi(\mathbf{Z}_i\gamma)} \cdot \Phi\left[\frac{0+\mathbf{Z}_i\gamma}{1}\right] \\ &= \frac{\phi\left(\frac{Y_{1i}-\mathbf{X}_i\boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1}. \end{aligned}$$

Likelihood Under Selection

$$\begin{aligned}
\ln L(\boldsymbol{\beta}, \gamma, \sigma_1, \rho | Y_1) &= \sum_{i=1}^{N} (1 - D_i) \ln[1 - \Phi(\mathbf{Z}_i \gamma)] \\
&+ \sum_{i=1}^{N} D_i \ln[\Phi(\mathbf{Z}_i \gamma)] \\
&+ \sum_{i=1}^{N} D_i \ln \left\{ \frac{\phi\left(\frac{Y_{1i} - \mathbf{X}_i \boldsymbol{\beta}}{\sigma_1}\right)}{\sigma_1 \Phi(\mathbf{Z}_i \gamma)} \cdot \Phi\left[\frac{\frac{\rho(Y_{1i} - \mathbf{X}_i \boldsymbol{\beta})}{\sigma_1} + \mathbf{Z}_i \gamma}{\sqrt{1 - \rho^2}} \right] \right\}
\end{aligned}$$

Estimation

- MLE (above)
- Or, reconsider:

$$E(Y_{1i}|\mathbf{X}_i, \mathbf{Z}_i, D_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + \rho\sigma_1 \left[\frac{\phi(\mathbf{Z}_i\gamma)}{\Phi(\mathbf{Z}_i\gamma)}\right]$$

- Note that $\Phi(\mathbf{Z}_i\gamma) = \Pr(D_i = 1)$
- Suggests a *two-step* approach...

Heckman's Two-Step Estimator

1. Estimate $\hat{\gamma}$ from

$$\Pr(D_i = 1) = \Phi(\mathbf{Z}_i \gamma)$$

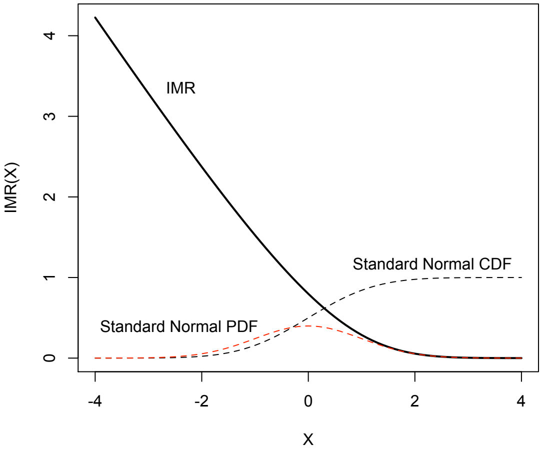
   and calculate the estimated *inverse Mills' ratio*:

$$\hat{\lambda}_i = \frac{\phi(\mathbf{Z}_i \hat{\gamma})}{\Phi(-\mathbf{Z}_i \hat{\gamma})}$$

2. Estimate $\boldsymbol{\beta}, \theta (\equiv \rho \sigma_1)$ as:

$$Y_{1i} = \mathbf{X}_i \boldsymbol{\beta} + \theta \hat{\lambda}_i + u_{1i}$$

What exactly is an "inverse Mills' ratio," anyway?



A Few Things...

- Since $\sigma_1 > 0$, $\hat{\theta} = 0 \implies \rho = 0$
- Two-step approach:
  - Is "LIML"...
  - Consistent for $\hat{\boldsymbol{\beta}}$, *but*
  - Inconsistent estimating $\widehat{\mathbf{V}(\boldsymbol{\beta})}$; so
  - Standard errors require correction (e.g., bootstrap)
  - *Can* yield $\hat{\rho} \notin [-1, 1]$ (because $\hat{\rho} = \hat{\theta}/\hat{\sigma}_1$)
  - Sensitive to prediction of D_i (better prediction = better precision)

Identification, etc.

- If $\mathbf{X} = \mathbf{Z}$, then $\boldsymbol{\beta}, \gamma, \rho$ (formally) identified by nonlinearity of $\Phi(\cdot)$
- (Much) better: $\geq$ one covariate in $\mathbf{Z}$ not in $\mathbf{X}$
- But...
  - Factors causing Y_1 also (often) cause D
  - $\implies$ $\mathbf{X}, \mathbf{Z}$ highly correlated
  - ...just makes things worse (Stolzenberg and Relles 1997)

Some Practical Things

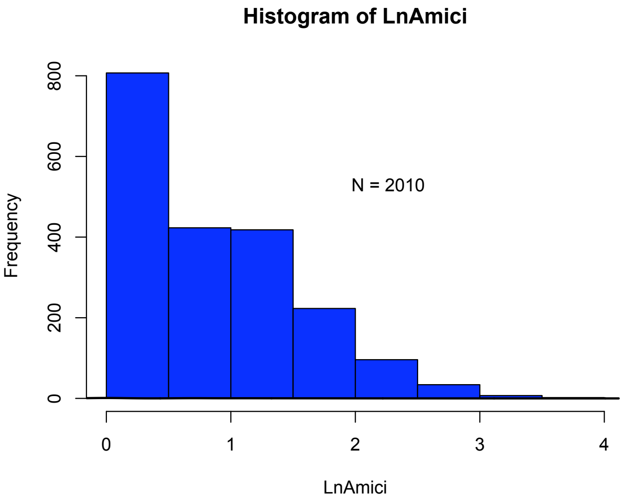
- In practice, few use two-step anymore,
- Sensitive to joint normality of $\{u_i, u_2\}$,
- *Very* sensitive to model specification...
- Key issue: *endogeneity* of selection...

Example: SCOTUS *Amicus* Briefs

- `LnAmici` $=$ ln($\#$ of briefs filed)
- For this to be defined, `Amici` > 0...
- Covariates:
  - `Year` -1900
  - `USPartic`: 1 if U.S. participated, 0 otherwise
  - `SCscore`: SCOTUS "Segal-Cover" liberalism score
  - `MultipleLegal`: 1 if multiple legal issues, 0 otherwise
  - `SGAmicus`: 1 if SG filed a brief, 0 otherwise

SCOTUS Decisions, 1953-1985

```
> SCOTUS<-read.dta("Data/SampleSelectionExample.dta",
+                  convert.factors=FALSE)
> summary(SCOTUS)
       ID              Docket              Amici             LnAmici     
 Min.   : 920764   Length:7156        Min.   : 0.0000   Min.   :0.000  
 1st Qu.:3790359   Class :character   1st Qu.: 0.0000   1st Qu.:0.000  
 Median :4100519   Mode  :character   Median : 0.0000   Median :0.693  
 Mean   :4116116                      Mean   : 0.8425   Mean   :0.757  
 3rd Qu.:4460624                      3rd Qu.: 1.0000   3rd Qu.:1.386  
 Max.   :4781050                      Max.   :39.0000   Max.   :3.664  
                                                        NA's   :5146   
      Year           USPartic         FedPetit          FedResp     
 Min.   :53.00   Min.   :0.0000   Min.   :0.0000   Min.   :1.000  
 1st Qu.:65.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000  
 Median :73.00   Median :0.0000   Median :0.0000   Median :3.000  
 Mean   :71.93   Mean   :0.3707   Mean   :0.1722   Mean   :2.593  
 3rd Qu.:80.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:3.000  
 Max.   :86.00   Max.   :1.0000   Max.   :1.0000   Max.   :3.000  
                                                                   
    SGAmicus           SCscore          MultipleLegal        select      
 Min.   :0.00000   Min.   :-0.22444   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:-0.12444   1st Qu.:0.000   1st Qu.:0.0000  
 Median :0.00000   Median :-0.01778   Median :0.000   Median :0.0000  
 Mean   :0.07868   Mean   : 0.13250   Mean   :0.149   Mean   :0.2809  
 3rd Qu.:0.00000   3rd Qu.: 0.47667   3rd Qu.:0.000   3rd Qu.:1.0000  
 Max.   :1.00000   Max.   : 0.66222   Max.   :1.000   Max.   :1.0000  
```


Histogram of LnAmici
N = 2010
Frequency
LnAmici
0
200
400
600
800
0
1
2
3
4

Estimates: OLS

```
> OLS<-lm(LnAmici~Year+USPartic+MultipleLegal+SCscore,data=SCOTUS)
> summary(OLS)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2328 -0.5837 -0.1223  0.4614  3.0901 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.737133   0.314843  -2.341   0.0193 *  
Year           0.020168   0.004134   4.879 1.15e-06 ***
USPartic      -0.174420   0.034968  -4.988 6.62e-07 ***
MultipleLegal  0.199667   0.038331   5.209 2.09e-07 ***
SCscore       -0.159575   0.117648  -1.356   0.1751    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.7275 on 2005 degrees of freedom
  (5151 observations deleted due to missingness)
Multiple R-squared: 0.1003,Adjusted R-squared: 0.09854 
F-statistic:  55.9 on 4 and 2005 DF,  p-value: < 2.2e-16
```

Estimates: Probit (Selection)

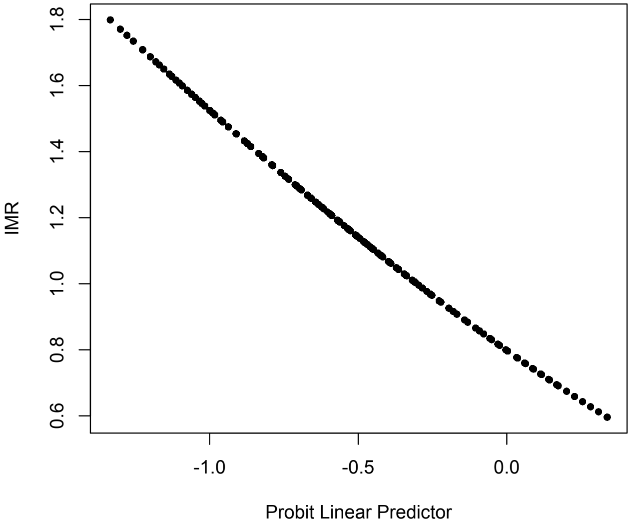
```
> SCOTUS$D<-SCOTUS$Amici>0
> probit<-glm(D~Year+USPartic+SCscore+MultipleLegal,data=SCOTUS,
  family=binomial(link="probit"))
> summary(probit)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.558970   0.273964  -9.341  < 2e-16 ***
Year           0.026875   0.003602   7.462 8.54e-14 ***
USPartic      -0.164948   0.034408  -4.794 1.64e-06 ***
SCscore       -0.089525   0.103323  -0.866    0.386
MultipleLegal  0.565585   0.043171  13.101  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8498.3  on 7155  degrees of freedom
Residual deviance: 8025.2  on 7151  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 8035.2
```

Estimates: Two-Step ("By-Hand")

```
> SCOTUS$IMR<-((1/sqrt(2*pi))*exp(-((probit$linear.predictors)^2/2))) /
  pnorm(probit$linear.predictors)
> OLS.2step<-lm(LnAmici~Year+USPartic+MultipleLegal+SCscore+IMR,data=SCOTUS)
> summary(OLS.2step)

Call:
lm(formula = LnAmici ~ Year + USPartic + MultipleLegal + SCscore +
    IMR, data = Day17)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.07914    3.58519  -2.253  0.02434 *
Year           0.07478    0.02688   2.782  0.00546 **
USPartic      -0.50500    0.16456  -3.069  0.00218 **
MultipleLegal  1.28738    0.53048   2.427  0.01532 *
SCscore       -0.33374    0.14490  -2.303  0.02137 *
IMR            2.75326    1.33926   2.056  0.03993 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7269 on 2004 degrees of freedom
  (5146 observations deleted due to missingness)
Multiple R-squared: 0.1022,Adjusted R-squared: 0.09999
F-statistic: 45.64 on 5 and 2004 DF,  p-value: < 2.2e-16
```

Estimates: Two-Step (Bad Specification)

```
> heckman2S<-heckit(D~Year+USPartic+SCscore+MultipleLegal, LnAmici~Year+USPartic
+SCscore+MultipleLegal,data=SCOTUS,method="2step")
> summary(heckman2S)
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
7156 observations (5146 censored and 2010 observed) and 13 free parameters (df = 7144)

Probit selection equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.558971   0.275385  -9.292  < 2e-16 ***
Year           0.026875   0.003622   7.420 1.31e-13 ***
USPartic      -0.164948   0.034366  -4.800 1.62e-06 ***
SCscore       -0.089524   0.103873  -0.862    0.389
MultipleLegal  0.565585   0.043298  13.063  < 2e-16 ***
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.07914    4.56334  -1.770   0.0767 .
Year           0.07478    0.03499   2.137   0.0326 *
USPartic      -0.50500    0.21993  -2.296   0.0217 *
SCscore       -0.33374    0.25058  -1.332   0.1829
MultipleLegal  1.28738    0.67647   1.903   0.0571 .

Multiple R-Squared:0.1022,Adjusted R-Squared:0.1
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio    2.753      1.668    1.65   0.0989 .
sigma            2.447         NA      NA       NA
rho              1.125         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
```


IMR
1.8
1.6
1.4
1.2
1.0
0.8
0.6
-1.0
-0.5
0.0
Probit Linear Predictor

Estimates: MLE (Bad Specification)

```
> heckmanML<-heckit(D~Year+USPartic+SCscore+MultipleLegal,
                    LnAmici~Year+USPartic+SCscore+MultipleLegal,
                    data=SCOTUS,method="ml")

> summary(heckmanML)

--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -6424.647
7156 observations (5146 censored and 2010 observed)
12 free parameters (df = 7144)

.
.
.
```

Estimates: MLE (Bad Specification)

```
Probit selection equation:
              Estimate Std. error t value  Pr(> t)
(Intercept)   -2.559549   0.331857  -7.713 1.23e-14 ***
Year           0.026862   0.004367   6.151 7.72e-10 ***
USPartic      -0.165173   0.043585  -3.790 0.000151 ***
SCscore       -0.090504   0.125536  -0.721 0.470946
MultipleLegal  0.566437   0.058852   9.625  < 2e-16 ***

Outcome equation:
              Estimate Std. error t value  Pr(> t)
(Intercept)   -8.06266    0.88402  -9.120  < 2e-16 ***
Year           0.08519    0.01182   7.205 5.80e-13 ***
USPartic      -0.49013    0.10103  -4.851 1.23e-06 ***
SCscore       -0.29510    0.34156  -0.864    0.388
MultipleLegal  1.26060    0.10607  11.885  < 2e-16 ***

Error terms:
      Estimate Std. error t value Pr(> t)
sigma  2.11218         NA      NA      NA
rho    0.99993    0.00742   134.8  <2e-16 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
Warning messages:
1: In sqrt(diag(vc)) : NaNs produced
2: In sqrt(diag(vc)) : NaNs produced
```

Estimates: MLE ("Better" Specification)

```
> betterML<-heckit(D~Year+USPartic+SCscore+MultipleLegal+SGAmicus,
           LnAmici~Year+USPartic+SCscore+MultipleLegal,
           data=SCOTUS,method="ml")

> summary(betterML)

--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 1: gradient close to zero
Log-Likelihood: -5689.492
7156 observations (5146 censored and 2010 observed)
13 free parameters (df = 7143)

.
.
.
```

Estimates: MLE ("Better" Specification)

```
Probit selection equation:
              Estimate Std. error t value  Pr(> t)
(Intercept)   -2.670268   0.289236  -9.232  < 2e-16 ***
Year           0.024971   0.003804   6.565 5.21e-11 ***
USPartic       0.080486   0.036022   2.234   0.0255 *
SCscore       -0.091135   0.109363  -0.833   0.4047
MultipleLegal  0.518324   0.045625  11.361  < 2e-16 ***
SGAmicus       2.167694   0.082758  26.193  < 2e-16 ***

Outcome equation:
              Estimate Std. error t value  Pr(> t)
(Intercept)   -0.177121   0.326280  -0.543 0.587233
Year           0.015413   0.004188   3.681 0.000233 ***
USPartic      -0.104100   0.036572  -2.846 0.004421 **
SCscore       -0.167759   0.117178  -1.432 0.152242
MultipleLegal  0.130377   0.039958   3.263 0.001103 **

Error terms:
      Estimate Std. error t value  Pr(> t)
sigma  0.73923    0.01270  58.199  < 2e-16 ***
rho   -0.29103    0.04419  -6.586 4.53e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
```

Extensions: "Probit-Probit"

- Selection + binary second stage ($Y_i \in \{0,1\}$) (a/k/a "Heckit").
- Assume errors are bivariate standard Normal [so, $\{u_1, u_2 \sim \mathcal{BVN}(0,0,1,1,\rho) \equiv \Phi_2(\cdot)]$
- Log-Likelihood:

$$
\begin{aligned}
\ln L(\boldsymbol{\beta}, \gamma, \sigma_1, \rho | Y_1) \quad = \quad & \sum_{Y_{1i}=1, D_i=1} \ln[\Phi_2(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\gamma, \rho)] \\
& + \sum_{Y_{1i}=0, D_i=1} \ln[\Phi_2(-\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\gamma, -\rho)] \\
& + \sum_{D_i=0} \ln \Phi(-\mathbf{Z}_i\gamma)
\end{aligned}
$$

More Extensions

- Different outcome stages:
  - Poisson (Greene 1995)
  - Durations (Boehmke et al.)
  - Count/binary/ordinal (Mirand and Rabe-Hesketh 2005)
- Selection stage is <u>ordered</u> (Chiburis & Lokshin 2007)
- Multiple-stage models (not much... finance?)

Sample Selection: Software

- R (`selection` and `heckit` in `sampleSelection` package)
  - Binary selection
  - Continuous/binary outcomes
  - Also tobit, etc. models

- Stata
  - `heckman` (binary-continuous model)
  - `heckprob` (binary-binary model)
  - `oheckman` (ordered-continuous)
  - `dursel` (binary-duration model)
  - `gllamm` (various multilevel models w/selection)

Further Readings: References

Articles by Heckman (1974, 1976, 1979).

Breen, Richard. 1996. *Regression Models for Censored, Sample Selected, or Truncated Data*. Thousand Oaks, CA: Sage.

Stolzenberg, Ross M. and Daniel A. Relles. 1997. "Tools for Intuition about Sample Selection Bias and Its Correction." *American Sociological Review* 62:494-507.

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." *The Journal of Human Resources* 33:127-169.

Winship, Christopher and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327-350.

Further Readings: Applications

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43:1209-1230.

Blanton, Shannon Lindsey. 2000. "Promoting Human Rights and Democracy in the Developing World: U.S. Rhetoric versus U.S. Arms Exports." *American Journal of Political Science* 44:123-131.

Hart, David M. 2001. "Why Do Some Firms Give? Why Do Some Give a Lot?: High-Tech PACs, 1977-1996." *The Journal of Politics* 63:1230-1249.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57:587-616.

Nooruddin, Irfan. 2002. "Modeling Selection Bias in Studies of Sanctions Efficacy." *International Interactions* 28: 57-74.

Timpone, Richard J. 1998. "Structure, Behavior and Voter Turnout in the United States." *American Political Science Review* 92: 145-158.

Von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *American Political Science Review* 99:611-622.