UNIVERSITY OF MANNHEIM
School of Social Sciences

Advanced Quantitative Methods in Political Science: Selection Bias and Multi-Equation Models

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova

Week 11 - 11 May 2022

Quiz

Quiz

As dedicated experts of voting behavior in *Absurdistan* with its stable three-party system you analyze a data set of $N = 1000$ voters, i.e. all surveyed non-voters are already dropped. As expert of *Absurdistan* politics you know that *Age*, *Education* and *IssueDistance* are the key to predicting vote-choice. Before running a conditional logit model you take a look at the data and observe the following: There are 50 respondents with missing information on *Age* and 50 different respondents with missing information on *Education*. Finally, there are another different 50 respondents who place all parties at the same value on the issue scale where they locate themselves. Every respondent, though, reports her vote choice.

Which if the following statements is true?

1. There are 1000 cases (in long-format) in the data.
2. When estimating the model you use information of all cases in the data.
3. The information of 150 voters cannot be leveraged for estimation.
4. The number of cases (in long-format) used for estimating this model is 2700.

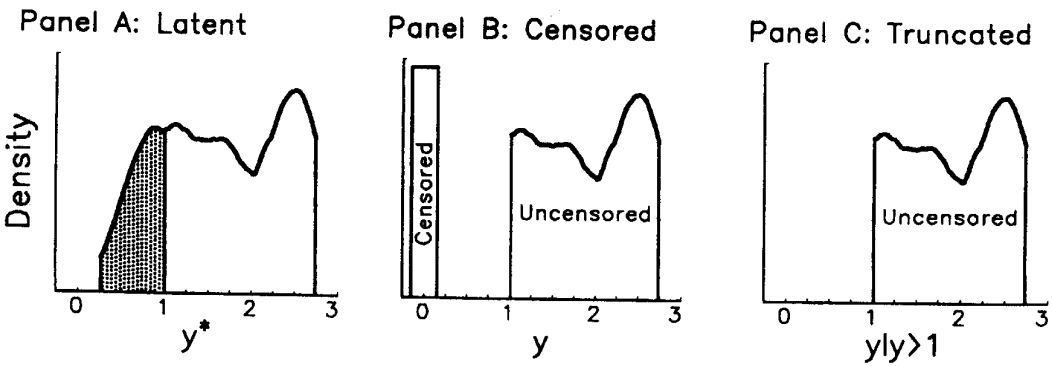
Intro

What should you take home from this class today?

- We talk about strategies how to deal with incompletely observed data and meet two types of non-random selection: *truncation* and *censoring*.
- We will learn how the likelihood theory of inference is dealing with those problems of non-random selection.
- We will get to know *Tobit* and *Heckman* models.
- We will meet *multi-equation models* (aka *Structural Equation Models*, SEM) and learn that we might gain efficiency when we model separate single-equation models simultaneously.
- We will learn how to estimate *Heckman models* as a particular multi-equation model in order to deal with non-random selected data.

Censored and Truncated Data

Censored and Truncated Data. What's the difference?



Taken from: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Newbury Park.: Sage, Figure 7.1.

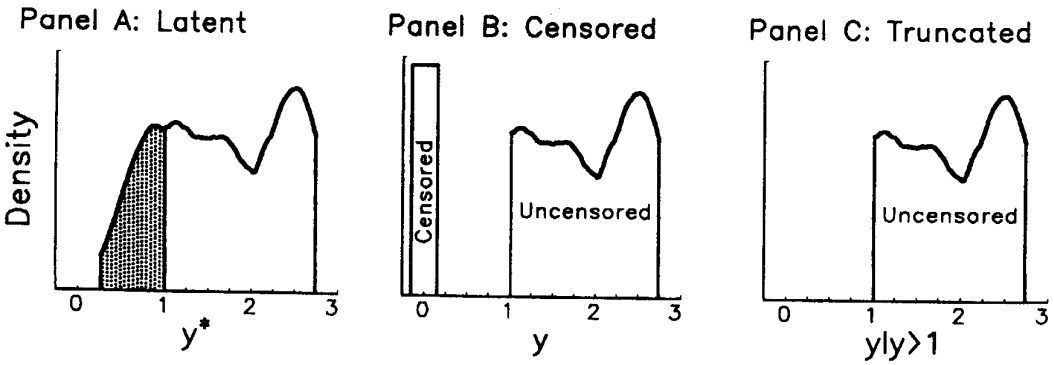
Incompletely Observed Data

There are two leading causes of incompletely observed data.

1. Truncation occurs when some observations on both the dependent variable as well as the regressors are lost.
   - We obtain inconsistent estimates if we run a regression when the *dependent variable is incompletely observed* because the sample is not representative of the population.
   - Say, you wanna predict political interest based on a student sample (non-students who might score low are excluded)
2. Censoring occurs when data on the dependent variable is lost (or limited) but not on the regressors. Think of it as a defect of the sample.
   - Again, we obtain inconsistent parameter estimates if we run a regression when the *dependent variable is incompletely observed* because the sample is not representative of the population.
   - Say, you wanna predict the amount of party-independent contributions a candidate receives when preparing her campaign. There is likely to be a cluster at 0 Euro across candidates (*a.k.a. censored from below* or *left-censored*) even if they got something.
   - Or suppose you wanna predict income but high-income people might be top-coded in a category, say, $\geq$ 100.000 Euro (*a.k.a. censored from above* or *right-censored*).

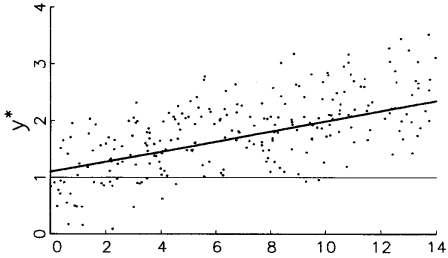
Say again. What's the difference?

- Intuitively, truncation entails greater information loss than censoring.
- If the mechanism by which the data is truncated or censored is independent of the dependent variable, run standard models.

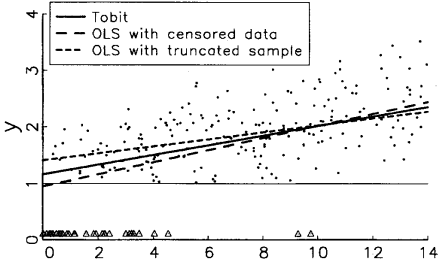


Taken from: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Newbury Park.: Sage, Figure 7.1.

Panel A: Regression without Censoring



Panel B: Regression with Censoring and Truncation



Taken from: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Newbury Park.: Sage, Figure 7.2.

Truncated Normal Model

Short Digression: Normal and Standard Normal Distributions

Suppose $y \sim N(\mu, \sigma^2 I)$. Thus, the pdf of y with mean μ and variance σ^2 is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

One can show that any linear transformation of the normal is itself normally distributed, i.e. $(a + by) \sim N(a + b\mu, b^2\sigma^2 I)$.

For a particular transformation with $a = -\frac{\mu}{\sigma}$ and $b = \frac{1}{\sigma}$, $z = \frac{y-\mu}{\sigma}$ has a standard normal distribution, i.e., $z \sim N(0, 1)$, with density

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Thus, a important characteristic is that one can rewrite every normal pdf $f(y)$ as a function of the standard normal as follows:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \frac{1}{\sigma} \phi(z)$$

Truncated Normal Distribution

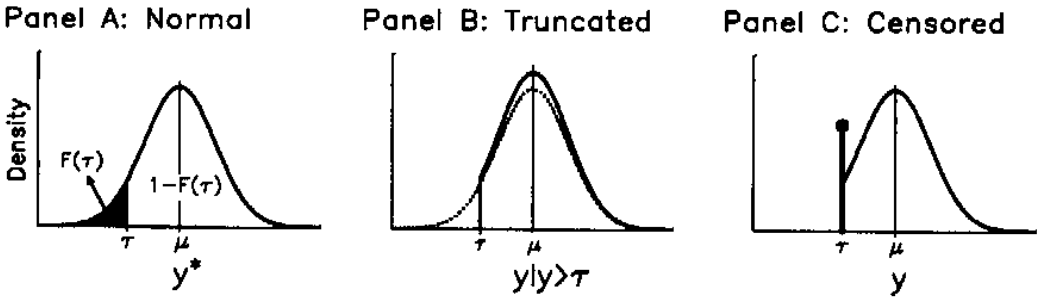


Figure 7.3. Normal Distribution With Truncation and Censoring

Taken from: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Newbury Park.: Sage, Figure 7.3.

Truncated Normal Distribution

- Let y denote the observed realizations of a continuous random variable Y. Unlike the normal regression case, y is the incompletely observed value of a latent dependent variable y^*.
- With truncation (at, say, τ) we only observe $y = y^*$ for values above τ and lose the remaining observations. Thus our *truncated* sample is a subset of a larger population.
- Given that we lose observations, the area under the assumed normal does not integrate to 1. Thus, we have to re-scale (normalize) the distribution by using the information we observe to get a probability density distribution.
- Thus, given the normal pdf $f(y)$, we get the pdf of the trancated normal distribution as

$$f(y|y > \tau) = \frac{f(y)}{Pr(y > \tau)} = \frac{f(y)}{1 - Pr(y \leq \tau)} = \frac{\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})}{1 - \Phi(\frac{\tau-\mu}{\sigma})}$$

Deriving the Likelihood Function of a Truncated Normal

Let $f(y)$ be the pdf of a normal random variable (truncated from below),

$$f(y|y > \tau) = \frac{f(y)}{Pr(y > \tau)} = \frac{\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})}{1 - \Phi(\frac{\tau-\mu}{\sigma})}$$

Thus, the log-likelihood contribution L_i of observation i is

$$lnL_i = ln(\frac{1}{\sigma}\phi(\frac{y_i - \mu_i}{\sigma})) - ln(1 - \Phi(\frac{\tau - \mu_i}{\sigma}))$$

Then summing-up all N individual contributions assuming independent realizations and $\mu_i = X_i\beta$ gives us the log-likelihood.

$$\begin{aligned} lnL(\beta, \sigma) &= \sum_{i \in N}(ln(\frac{1}{\sigma}\phi(\frac{y_i - \mu_i}{\sigma})) - ln(1 - \Phi(\frac{\tau - \mu_i}{\sigma}))) \\ &= \sum_{i \in N}(ln(\frac{1}{\sigma}\phi(\frac{y_i - X_i\beta}{\sigma})) - ln(1 - \Phi(\frac{\tau - X_i\beta}{\sigma}))) \\ &= -\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - X_i\beta)^2 - \sum_{i=1}^{N}(ln(1 - \Phi(\frac{\tau - X_i\beta}{\sigma}))) \end{aligned}$$

Tobit Model

Tobit Model

- Let Y^* be a continuous unobserved variable
- Define the Tobit model through its *stochastic* and *systematic* component

$$\begin{aligned} Y_i^* &\sim N(y_i^* | \mu_i, \sigma^2) \\ \mu_i &= X_i \beta \end{aligned}$$

with a *censoring mechanism* (a.k.a *sample selection rule*):

$$y_i = \begin{cases} y & y^* > \tau \\ \tau_y & y^* \leq \tau \end{cases}$$

Observations with values at or below τ are not observed directly. Instead they are set to τ_y

- Finally, lets assume independent realizations (i.e., censored and uncensored observations are independent from one another).

Deriving the Likelihood Function of a Tobit

When a distribution is censored on the left, observations with values at or below τ are set to τ_y. Thus, there are two types of observations

- Uncensored observations, i.e. when $y^* > \tau$. We take the product over those observations as in the OLS-case (parameterized as a function of ϕ)
- Censored observations, i.e. when $y^* \leq \tau$. All we know for those observations is $Pr(y^* \leq \tau) = \Phi(\frac{\tau - \mu_i}{\sigma}) = 1 - \Phi(\frac{\mu_i - \tau}{\sigma})$

Let d_i an indicator scoring 1 if i^{th} observation is uncensored. Thus, the likelihood function of the Tobit (censored normal) model is a mixture of censored and uncensored obserations

$$\begin{aligned} L &= \prod_{i \in N} (\frac{1}{\sigma}\phi(\frac{y_i - \mu_i}{\sigma}))^{d_i} \cdot (1 - \Phi(\frac{\mu_i - \tau}{\sigma}))^{1-d_i} \\ &= \prod_{i \in N} (\frac{1}{\sigma}\phi(\frac{y_i - X_i\beta}{\sigma}))^{d_i} \cdot (1 - \Phi(\frac{X_i\beta - \tau}{\sigma}))^{1-d_i} \end{aligned}$$

Thus, taking the log yields

$$lnL(\beta, \sigma) = \sum_{i=1}^{N} [d_i(-ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2) + (1 - d_i)ln(1 - \Phi(\frac{X_i\beta - \tau}{\sigma}))]$$

Interpretation and Estimation

As usual, take predicted probabilities, expected values or first differences. Your quantities of interest can be summarized as ...

- ...*point estimates*, by averaging the simulations.
- ...*confidence intervals*, by sorting the simulations and taking the 2.5th and 97.5th percentile values for a 95% confidence interval for instance.
- ...or *standard errors*, by taking the standard deviation of the simulations.

In order to communicate your results, your summarized simulations of the quantities of interest can be ...

- ...presented in the text as a number or as a table.
- ...displayed as histograms, density estimates or in other graphs (e.g., ternary plot) to summarize the full sampling (or posterior) density.

You can also estimate this model in R for instance using the `library(Zelig)`.

Heckman as Extension of Tobit

Assumptions and Extensions of the Tobit Model

- The basic Tobit model can be easily coded-up to account for right-censoring or even for left- and right-censoring (with, in this case, three types of observations that make-up the likelihood: left-censored, uncensored and right-censored observations)
- Given the stochastic component of the model, *homoskedasticity* is assumed. If the errors are heteroskedastic, though, you need to model them (through parameterizing the variance function, e.g., as $\sigma_i^2 = e^{Z_i\gamma}$) in order to get consistent ML estimates assuming correct specification of σ_i^2.
- Tobit-type models need not to be constructed based on the normal distribution. Other assumptions about the stochastic component are conceivable, e.g., Poisson to model (censored) count processes (see King 1989, chapter 9).
- For Tobit models we cannot distinguish the data generating processes that drives the censoring and the dependent variable. *Sample selection models* (Heckman), a generalization of the Tobit model, are built on the idea that those two process can be separately modeled.

From Tobit to Heckman

- For Heckman models we assume that we can distinguish, and consequently, separately model, the data generating processes that drives the censoring (*selection equation*) and the dependent variable (*outcome equation*).
- Selection bias effects occur if unobserved factors, that influence which cases get selected into the sample (represented as error term of the *selection equation*), *are correlated with* unobserved factors influencing outcomes (represented as error term of the *outcome equation*) in the selected sample (akin to *omitted variable bias*).
- However, there might be variables that affect whether an observation is censored (i.e., selected into the sample) *without* determining the outcome. Hence, no bias!
- Whether there is selection bias or not, more generally, think of this as a strategy of how two simultaneous processes can be modeled at once. This is called a *system-of-equations* (in the Heckman case we have a *bivariate* dependent variable $Y_i = (Y_{1i}, Y_{2i})'$ consisting of two equations that are estimated simultaneously).

Take the voter-turnout-example of Timpone (1998)

- In the US you need to register first before you are eligible to vote (there is variance across states!). Thus registered voters are potentially a self-selected population.
- *Outcome equation* (who turns out to vote?) specifies standard kitchen-sink variables to model turnout.
- *Selection equation* (who is registered to vote?) includes variables to operationalize administrative barriers (e.g. closing time) in addition to standard kitchen-sink variables previously thought to affect turnout.
- Given that two processes are modeled simultaneously and the DV of each process is dichotomous, the dependent variable $Y_i = (Y_{1i}, Y_{2i})'$ of this system-of-2-equations is bivariate. Hence, Timpone estimates a *bivariate probit* model that allows the error terms to be correlated.

Timpone (1998, table 1)

TABLE 1. Pooled Turnout Models for Full Electorate, 1980–88

Variable	Single Model	Selection Bias Model	
	Turnout	Registration	Turnout
Intercept	−2.1467** (.1838)	−1.9678** (.2118)	.1824 (.6056)
Administrative Barriers			
Closing date	−.0066* (.0029)	−.0078* (.0031)	—
Purge records	.0171* (.0086)	.0295** (.0090)	—
Demographics			
South	−.3788** (.0543)	−.3274** (.0567)	−.2302* (.0944)
Age	.0130** (.0019)	.0127** (.0022)	.0064 (.0033)
Age-squared	−.0002* (.0001)	−.0001 (.0001)	−.0002 (.0001)
Education	.0973** (.0108)	.1031** (.0126)	.0366 (.0202)
Race (black)	−.0843 (.0803)	.1041 (.0891)	−.2910** (.1122)
Gender (female)	−.0433 (.0496)	−.0641 (.0598)	.0059 (.0724)
Income	.0035* (.0014)	.0046** (.0017)	.0003 (.0021)
Time in home	.0129** (.0031)	.0136** (.0036)	.0070 (.0042)
Social Connectedness			
Church attendance	.5765** (.0678)	.4511** (.0735)	.4739** (.1204)
Group membership	.1477** (.0552)	.1512** (.0580)	.0628 (.0861)
Marital status	.1842** (.0535)	.0827 (.0601)	.2390** (.0725)
Time in home	.0010 (.0018)	.0039 (.0021)	−.0038 (.0023)
Home ownership	.2731** (.0591)	.3210** (.0595)	.0115 (.0973)
Political Attitudes: General			
External efficacy	.4884** (.0882)	.5139** (.0949)	.2075 (.1372)
Internal efficacy	.1721** (.0557)	.1473* (.0632)	.1231 (.0795)
Party differential	.1047* (.0496)	.1193* (.0503)	.0171 (.0693)
Strength of party identification	.1607** (.0263)	.1689** (.0268)	.0585 (.0469)
Trust in government	−.1081 (.1090)	−.1841 (.1216)	.0413 (.1503)
Political Attitudes: Election Specific			
Candidate differential	.0044** (.0011)	.0036** (.0012)	.0033* (.0016)
Candidate satisfaction	−.1032 (.0624)	−.0459 (.0663)	−.1361 (.0839)
RHO		−.3550 (.3937)	
n	3598	3598	
LLF initial	−2493.9	−4326.6	
LLF final	−1896.2	−2588.4	

Note: The dependent variables in these models are *Validated Registration* and *Validated Vote*. The full sample size of 3,598 is composed of 954 nonregistrants, 343 registered nonvoters, and 2,301 voters. The administrative barriers were not included in the second stage of the selection bias model. *$p < .05$, **$p < .01$. Standard errors are in parentheses (bootstrapped estimates for the selection bias models).

Multiple Equation Models

Multiple Equation Models: Why should we care?

- In order to understand Heckman models we need to first understand what multi-equation models are.
- We actually have seen already an example of a bivariate (i.e., 2-equation) model early in the semester. Remember?
  - Franklin, Charles H. 1991. "Eschewing Obfuscation? Campaigns and the Perception of Senate Incumbents". *American Political Science Review* 85(4): 1193-1214.
- We can gain some efficiency if we model simultaneous processes through a multi-equation set-up.
- If we wanna model recursive processes (e.g., Are certain institutions a cause or a consequence of a country's economic performance?) we need a multi-equation set-up as well (e.g., 2SLS, IV estimation, ect.).
- Modeling processes simultaneously is, of course, no free lunch. Specification errors in one equation bias estimates of other equations in the system as well.

Introduction to Multiple Equation Models

- Let Y_i be a $N \times 1$ vector for observation i $(= 1, \ldots, n)$
- Y_i is jointly (N-variate) distributed with a *stochastic* component

$$Y_i \sim f(y_i|\theta_i, \alpha)$$

- θ_i is a $N \times 1$ parameter vector, while α is typically a $N \times N$ matrix
- N *systematic components* are defined as (Example?):

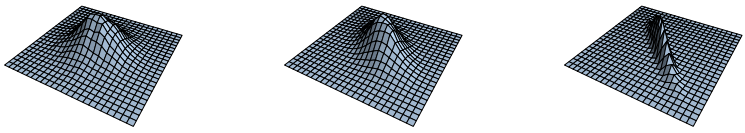
$$\begin{aligned} \theta_{1i} &= g_1(X_{1i}, \beta_1) \\ \theta_{2i} &= g_2(X_{2i}, \beta_2) \\ &\vdots \\ \theta_{Ni} &= g_N(X_{Ni}, \beta_N) \end{aligned}$$

- This model differs from N separate equation-by-equation models if the elements of Y_i are (conditional on X) correlated (i.e., stochastically dependent) or share parameters (e.g., a constraint such that $\beta_1 = \beta_2$)

Bivariate Normal Distribution, i.e. $N = 2$

Suppose you estimate two normal regressions simultaneously ...

$$Y_i = \begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \right]$$



(a) $\rho = 0.0$ (b) $\rho = 0.5$ (c) $\rho = 0.99$

ρ captures the degree to which both error terms are correlated. The uncertainty of a model's prediction (quantity-of-interest) depends also on the other model's error term (if $\rho > 0$). Separate estimation is inefficient if $\rho > 0$.

The Heckman Model

Heckman Model as a particular Multiple Equation Model

- Let Y_i vector for observation i $(= 1, \ldots, n)$
- $Y_i = (Y_{1i}^*, Y_{2i})'$ is bi-variate normal distributed with an
  1. *Selection equation*:
  $$y_{1i}^* = \mu_{1i} + u_i = X_{1i}\beta_1 + u_i,\ u_i \sim N(0,1)$$
  with an *stochastic censoring mechanism* (a.k.a *sample selection rule*)
  $$y_{1i} = \begin{cases} 1 & y_{1i}^* > 0 \\ 0 & y_{1i}^* \leq 0 \end{cases}$$
  2. *Outcome equation*: For all selected observations i, i.e., if $y_{1i}^* > 0$ one has
  $$y_{2i} = \mu_{2i} + \epsilon_i = X_{2i}\beta_2 + \epsilon_i,\ \epsilon_i \sim N(0, \sigma_2^2)$$
  whereby the error terms of both equations are correlated, i.e. $0 \neq \rho = corr(u_i, \epsilon_i)$. (Note that we get a Tobit as a special case if $y_{1i}^* = y_{2i}$)
- Thus, (check the dimensionality!)
$$Y_i = \begin{pmatrix} Y_{1i}^* \\ Y_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \right]$$

.... and the Likelihood

The likelihood function is a combination of the likelihood for censored and uncensored observations

- For $y_{1i}^* \leq 0$ all that is observed is that this event occurred. Thus, the density is the probability $Pr(y_{1i}^* \leq 0)$ that it occurred
- For $y_{1i}^* > 0$ we observe y_{2i} with a certain (conditional) probability. It is the probability of being selected, $Pr(y_{1i}^* > 0)$, multiplied by the bivariate density $f(y_{2i}|y_{1i}^* > 0)$.
- The likelihood function of a bivariate sample selection model is as follows:

$$L(\beta_1, \beta_2, \rho, \sigma_2^2) = \prod_{i=1}^{n} Pr(y_{1i}^* \leq 0)^{1-y_{1i}} \cdot \{f(y_{2i}|y_{1i}^* > 0) \cdot Pr(y_{1i}^* > 0)\}^{y_{1i}}$$

- One can show that the second term simplifies to a univariate normal distribution that can be easily handled computationally. Details can be found, for instance, in Amemiya's *Advanced Econometrics* (1985: 385-7) textbook.

Application: A Model to predict success in the Graduate Program

- Suppose we like to test whether GRE-scores predict success in terms of grades in our PhD-program.
- Obviously, in all the application files we have GRE-scores. Grades, however, are only available for the ones who join our program.
- Thus, the bottom of the distribution of the unobserved variable (Y^*_{1i}), *Admission Rating* of our PhD-program, is censored. The PhD selection committee only admits those candidates and monitor their performance who rate high on the latent *Admission Rating* variable.

Application: A Model to predict success in the Graduate Program

- Suppose we have the following *selection* and *outcome equation*:
  1. *Selection equation*:

$$AdmissionRating = \beta_{10} + \beta_{11}GRE + \beta_{12}TOEFL + u_i,\ u_i \sim N(0,1)$$

  with a *censoring mechanism* (a.k.a *sample selection rule*)

$$Admission = \begin{cases} 1 & AdmissionRating > 0 \\ 0 & AdmissionRating \leq 0 \end{cases}$$

  2. *Outcome equation*: For all enrolled (and former) students i, i.e., if $AdmissionRating > 0$ one has

$$Success = \beta_{20} + \beta_{21}GRE + \beta_{22}Math + \epsilon_i,\ \epsilon_i \sim N(0, \sigma_2^2)$$

- Admitted graduate students are not representative of applicants generally. Despite low GRE-scores applicants get admitted if they have high TOEFL-scores or because they have large error term – i.e., their applications have qualities that are uncorrelated with GRE or TOEFL scores (e.g., strong letter, University's reputation).

What if we just run an OLS for all observations we have data for?

- Group of students that were admitted because of high GRE scores are representative of the group of applicants with high GRE scores.
- However, the group of admitted students with low GRE scores are *not* representative of the group of all applicants with this score. Assuming that the selection committee has done a good job, those admitted low-score students perform better than the non-admitted ones.
- Thus, running regression on the selected sample might wrongly show that GRE does not systematically predict success in graduate school.

Identification, Interpretation and Estimation

- Selection bias models, as the Timpone-Example shows, are not bounded to have a normally distributed stochastic component but can be in principle fit to any theory about the selection and outcome processes.
- *Identification* of those processes is an issue with selection bias models, though. You need at least one variable (and the more the better!) that only predicts *selection* but not the *outcome* (otherwise identification hinges solely on non-linearity of the selection equation, hence on distributional assumptions that cannot be checked rigorously).
- *Interpretation.* As usual, calculate expected values, predicted probabilities and first-differences using statistical simulations.
- You can estimate these models in R for instance using the `library(sampleSelection)`.

Case Selection and Selection Bias

Selection on the Dependent Variable

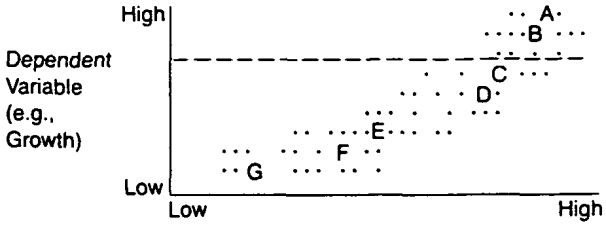


Fig. 1. Assumed relationship between factor *X* and the dependent variable

Taken from: Geddes, Barbara. 1997. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics" *Political Analysis* 2(1): 131-50; Figure 1.

Selection on the Dependent Variable

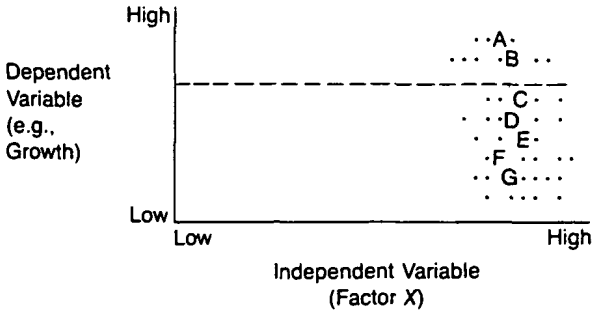


Fig. 2. An alternative possibility for the relationship between factor *X* and the dependent variable

Taken from: Geddes, Barbara. 1997. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics" *Political Analysis* 2(1): 131-50; Figure 2.

Selection on the DV - Endpoints in a Time Series

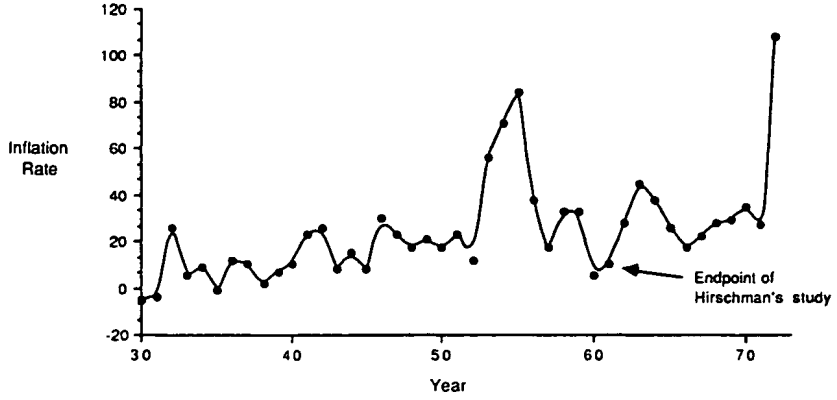


Fig. 11. Inflation in Chile, 1930–72. (Data for 1930–61 from Hirschman

Taken from: Geddes, Barbara. 1997."How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics" *Political Analysis* 2(1): 131-50; Figure 11.