Maximum Likelihood Estimation

Additional Topics

Raphael Cunha

Program in Statistics and Methodology – PRISM
Department of Political Science
The Ohio State University

cunha.6@osu.edu

March 28, 2014

Overview

We will cover two topics that are not usually covered in introductory maximum likelihood estimation (MLE) courses:

- Interpretation of multiplicative interaction terms in nonlinear models
- Models for truncation and sample selection (**Tobit** and the **Heckman selection model**)

Interaction terms in nonlinear models

Interaction terms in linear models (recap)

- We often have conditional hypotheses that can be captured by multiplicative interaction models. E.g.:

  H_1: An increase in X is associated with an increase in Y when condition Z is met, but not when condition Z is absent.

- Suppose that Y and X are continuous and Z is dichotomous. We can write a simple linear interactive model:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

- The marginal effect of X on Y (the effect of a one-unit change in X on Y) is given by:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z$$

- When $Z = 0$, $\frac{\partial Y}{\partial X} = \beta_1$. When $Z = 1$, $\frac{\partial Y}{\partial X} = \beta_1 + \beta_3$.

Interaction terms in linear models

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Hypothesis H_1: An increase in X is associated with an increase in Y when condition Z is met, but not when condition Z is absent.

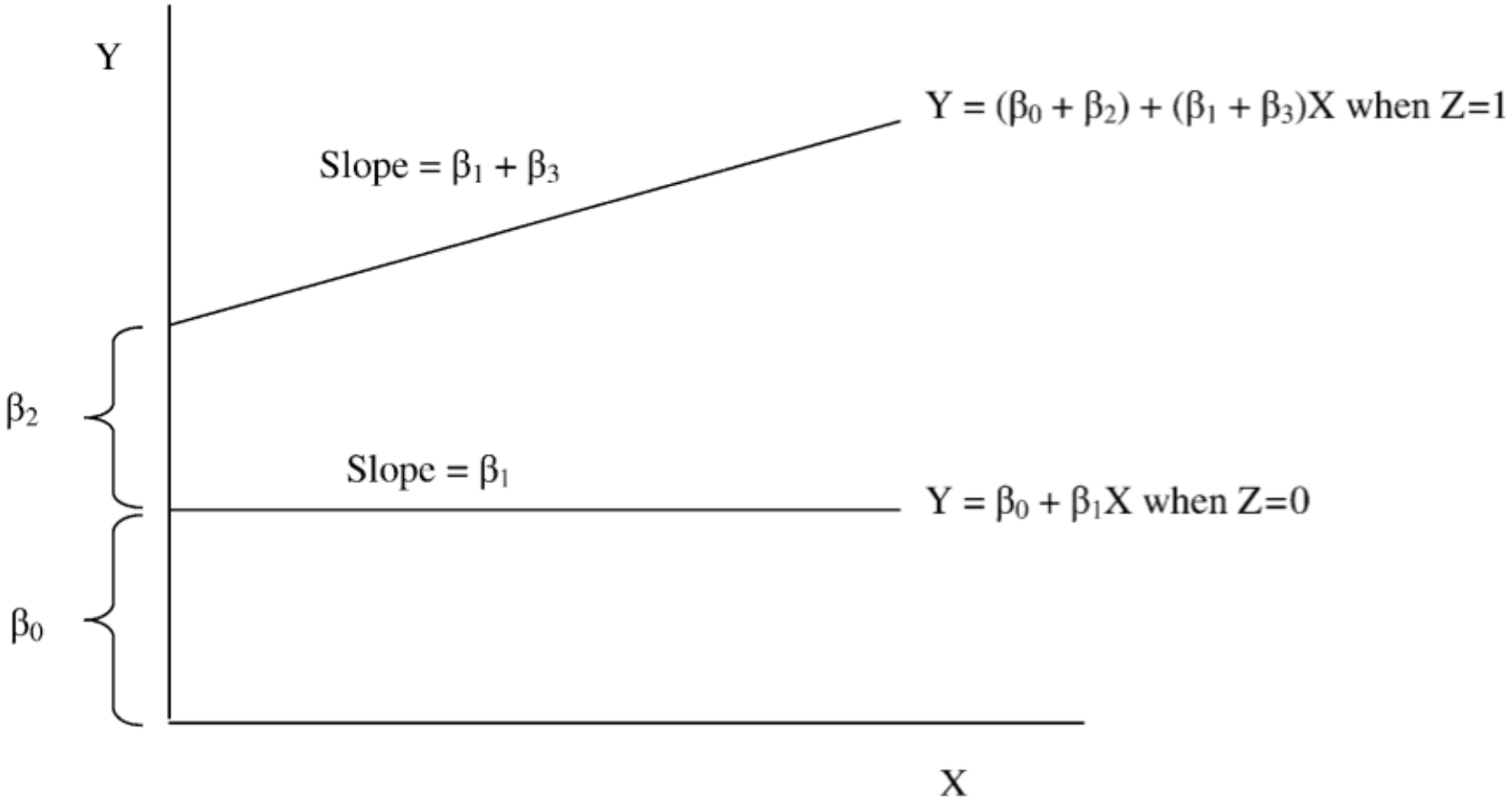

Figure: Brambor et al. 2006

Interaction terms in linear models

- We're usually not interested in the statistical significance or insignificance of the model parameters themselves. We care about $\frac{\partial Y}{\partial X}$, so we want to know the standard error of this quantity, which is given by:

$$\hat{\sigma}_{\frac{\partial Y}{\partial X}} = \sqrt{var(\hat{\beta}_1) + Z^2 var(\hat{\beta}_3) + 2Zcov(\hat{\beta}_1 \hat{\beta}_3)}$$

- If Z is dichotomous, we need only compute s.e.'s for $\frac{\partial Y}{\partial X}$ for when $Z = 0$ and $Z = 1$.

- If Z is continuous, useful to plot $\frac{\partial Y}{\partial X}$ against a substantively meaningful range of Z, using same formula above for the 95% confidence intervals

Interaction terms in linear models

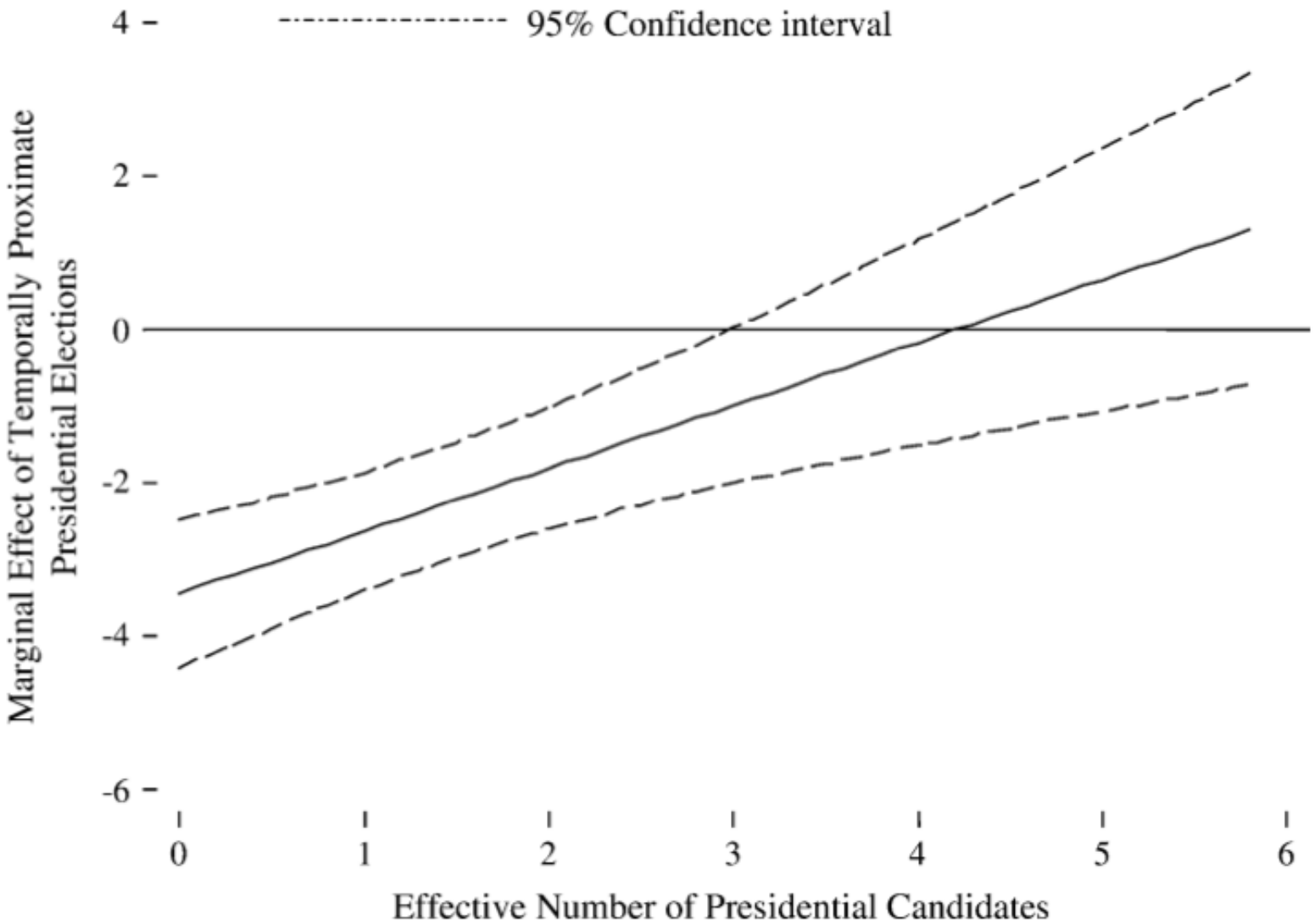

Figure: Marginal effect of temporally proximate presidential elections on the effective number of electoral parties. Example from Brambor et al. 2006.

Interaction terms in nonlinear models

- What about nonlinear models?
- Some have argued that it's not necessary to include an interaction term, because models like logit and probit force effect of all independent variables to depend on each other. E.g., consider the probit model:

$$E[Y] = \Phi(\beta_0 + \beta_1 X + \beta_2 Z) = \Phi(\cdot)$$

- Marginal effect of X is: $\frac{\partial \Phi(\cdot)}{\partial X} = \beta_1 \Phi'(\cdot)$
- Marginal effect of X depends on other independent variables whether the hypothesis being tested is conditional or not. It's just a function of how the model is parameterized. To substantively test a conditional hypothesis, one must include an interaction term, just like in the linear case (see Brambor et al. 2006)

Interaction terms in nonlinear models

- Ai & Norton 2003 show that interpreting interaction effects in nonlinear models is a lot more complicated than in linear ones
- For example, in our previous linear interactive model with a continuous Y, interaction effect of X and Z is cross-derivative of E[Y]:

$$\frac{\partial^2 E[Y|X,Z]}{\partial X \partial Z} = \beta_3$$

(But remember that not much can be learned from the statistical significance of β_3 alone)

Interaction terms in nonlinear models

- However, intuition does not extend to nonlinear models. Consider a dichotomous y, two independent variables of interest, x and z, and a vector of additional independent variables $\mathbf{W}$. A simple interactive probit model:

$$E[y|x, z, \mathbf{W}] = \Phi(\beta_1 x + \beta_2 z + \beta_3 xz + \mathbf{W}\beta) = \Phi(\cdot)$$

- The interaction effect of x and z is the cross-derivative of the expected value of y:

$$\frac{\partial^2 \Phi(\cdot)}{\partial x \partial z} = \beta_3 \Phi'(\cdot) + (\beta_1 + \beta_3 z)(\beta_2 + \beta_3 x)\Phi''(\cdot)$$

- Very hard to interpret!

Interaction terms in nonlinear models

Implications for interactive effects in nonlinear models (Ai & Norton 2003):

- Interaction effect could be nonzero even if $\beta_3 = 0$
- Statistical significance of interaction effect cannot be tested with t-test on the coefficient of interaction term β_3
- Interaction effect is conditional on independent variables in non-trivial ways
- Interaction effect may have different signs for different values of covariates; sign of β_3 does not necessarily indicate sign of interaction effect

What to do?

Ai & Norton offer "formulas for the magnitude and standard errors of the estimated interaction effect in general nonlinear models", but Greene recommends visual interpretation.

Greene 2010:
"(. . .) the proposals made by Ai and Norton are likewise uninformative about interaction effects in the model. (. . .) the indicated relationships are inherently difficult to describe numerically by simple summary statistics, but graphical devices are much more informative."

Example

Data: U.S. House of Representatives vote on NAFTA (1993). 435 observations and 5 variables.
Dependent variable:

- `vote`: whether (=1) or not (=0) the House member in question voted for NAFTA.

Independent variables:

- `democrat`: whether the House member in question is a Democrat (=1) or a Republican (=0).
- `pcthispc`: the percentage of the House member's district who are of Latino/Hispanic origin.
- `cope93`: 1993 AFL-CIO (COPE) voting score of the member in question; ranges from 0 to 100, with higher values indicating more pro-labor positions.

Example

We have the following hypotheses:

- Higher COPE scores will correspond to lower probabilities of voting for NAFTA
- The effect of the former will be moderated by political party. In particular, the (negative) effect of COPE scores on pro-NAFTA voting will be greater for Democrats than for Republicans.

So we estimate an interactive logit model:

$$Pr(Y = 1|X) = \Lambda(\beta_0 + \beta_1 democrat + \beta_2 cope93 + \beta_3 democrat * cope93 + \beta_4 pcthispc),$$

where Λ is the logistic CDF.

Example

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.791640   0.275438   6.505 7.79e-11 ***
democrat         6.865556   1.547295   4.437 9.12e-06 ***
cope93          -0.036501   0.007598  -4.804 1.55e-06 ***
pcthispc         0.020911   0.007941   2.633  0.00846 **
democrat:cope93 -0.067054   0.018203  -3.684  0.00023 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure: Logit parameter estimates for the probability of voting for NAFTA.

Example

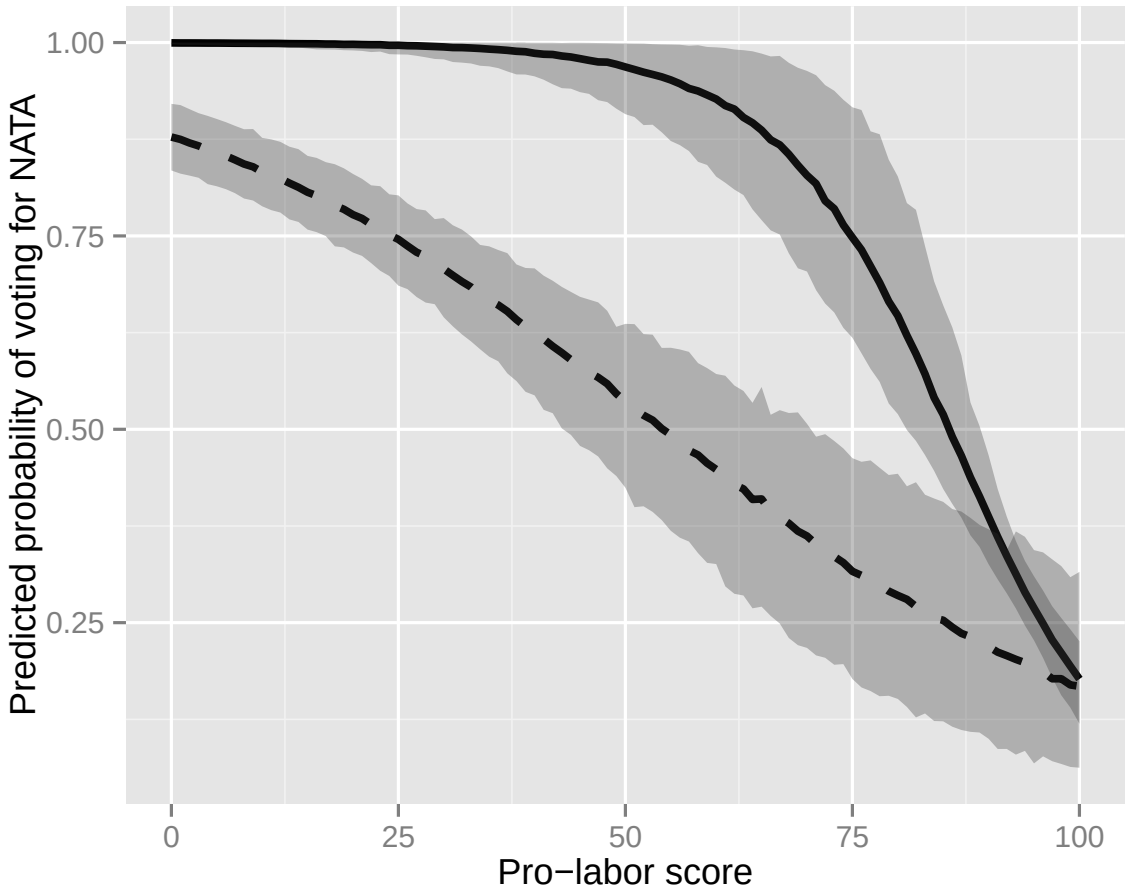

Figure: Predicted probability of voting for NAFTA as a function of pro-labor score (COPE93) for Democrats (continuous line) and Republicans (dashed line). Shaded areas are 95% bootstrapped CIs.

Example

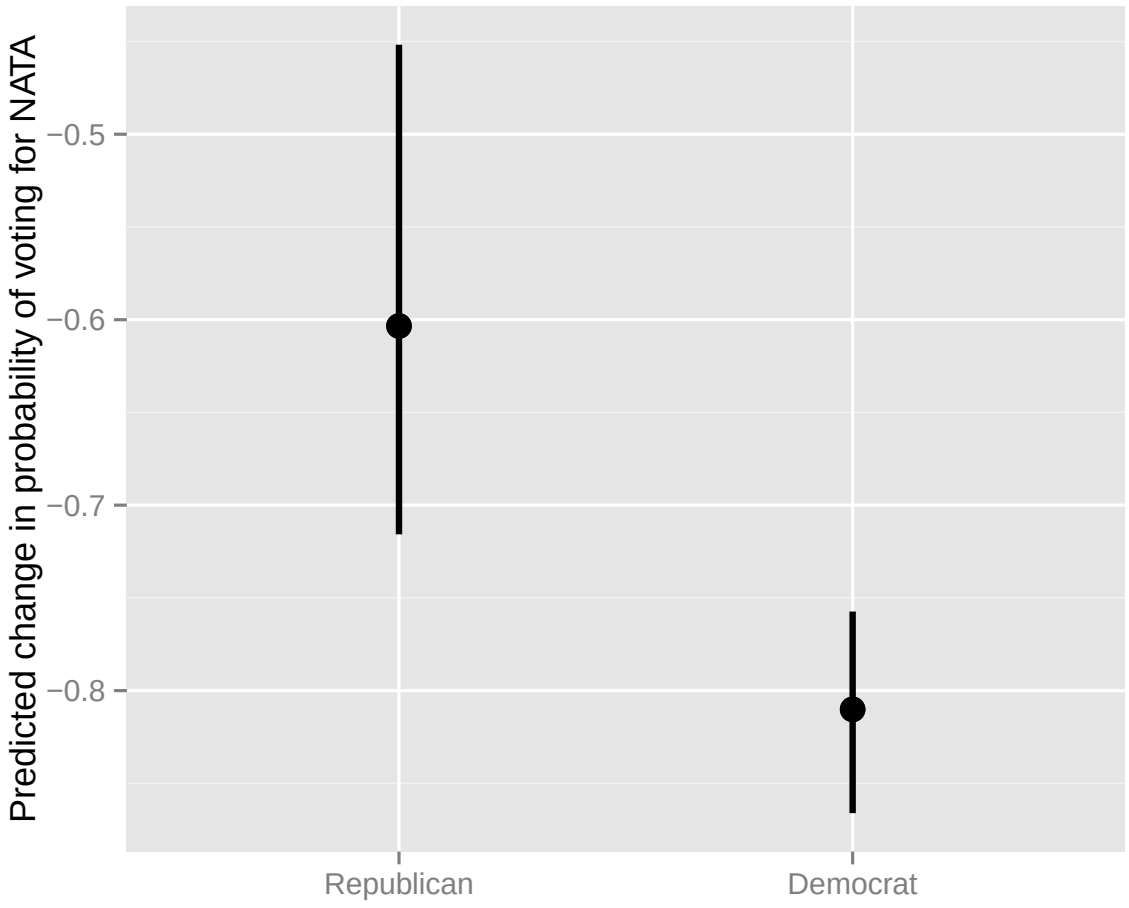

Figure: Discrete change in predicted probability of voting for NAFTA from a mean-centered, 2-standard-deviation change in COPE93 for Democrats and Republicans. Vertical bars are 95% bootstrapped CIs.

Example

Software implementation:

- Example R code available with the accompanying presentation materials
- Stata users: Be careful when using multiplicative interactions in Stata. The most common way of creating interaction terms is to generate a new variable equal to the product of the two interacting variables. If you do this, Stata will treat the interaction term as a third, distinct variable rather than two variables being interacted. When computing predicted probabilities, you might get wrong results. Make sure you know what the functions you are using are doing.

Interaction terms in nonlinear models

Recommended reading:

- Brambor, Thomas, William R. Clark & Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14: 63–82.
- Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58(Fall 2004): 807–820.
- Ai, Chunrong & Edward C. Norton. 2003. "Interaction terms in logit and probit models." *Economic Letters* 80: 123–129.
- Greene, William. 2010. "Testing hypotheses about interaction terms in nonlinear models." *Economic Letters* 107: 291–296.
- Berry, William D., Jacqueline DeMeritt & Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54(1): 248–266.

Models for sample selection

Censoring and truncation

Two main causes of incompletely observed data:

- **Truncation**: some observations on both the dependent and independent variables are lost.
  - E.g.: Income may be the dependent variable and only low-income people are included in the sample.
- **Censoring**: information on the dependent variable is lost, but not on the regressors.
  - E.g.: People of all income levels may be included in the sample, but the income of high-income people may be top-coded and reported only as exceeding $100,000 per year.
  - E.g.: Level of delegation of authority to international treaties may be the dependent variable. All country-pairs may be included in the sample, but the delegation level is only recorded for those dyads that have actually signed a treaty. There is no information on what delegation level would be chosen by two dyads who have not signed a treaty.

Truncation entails greater information loss than censoring.

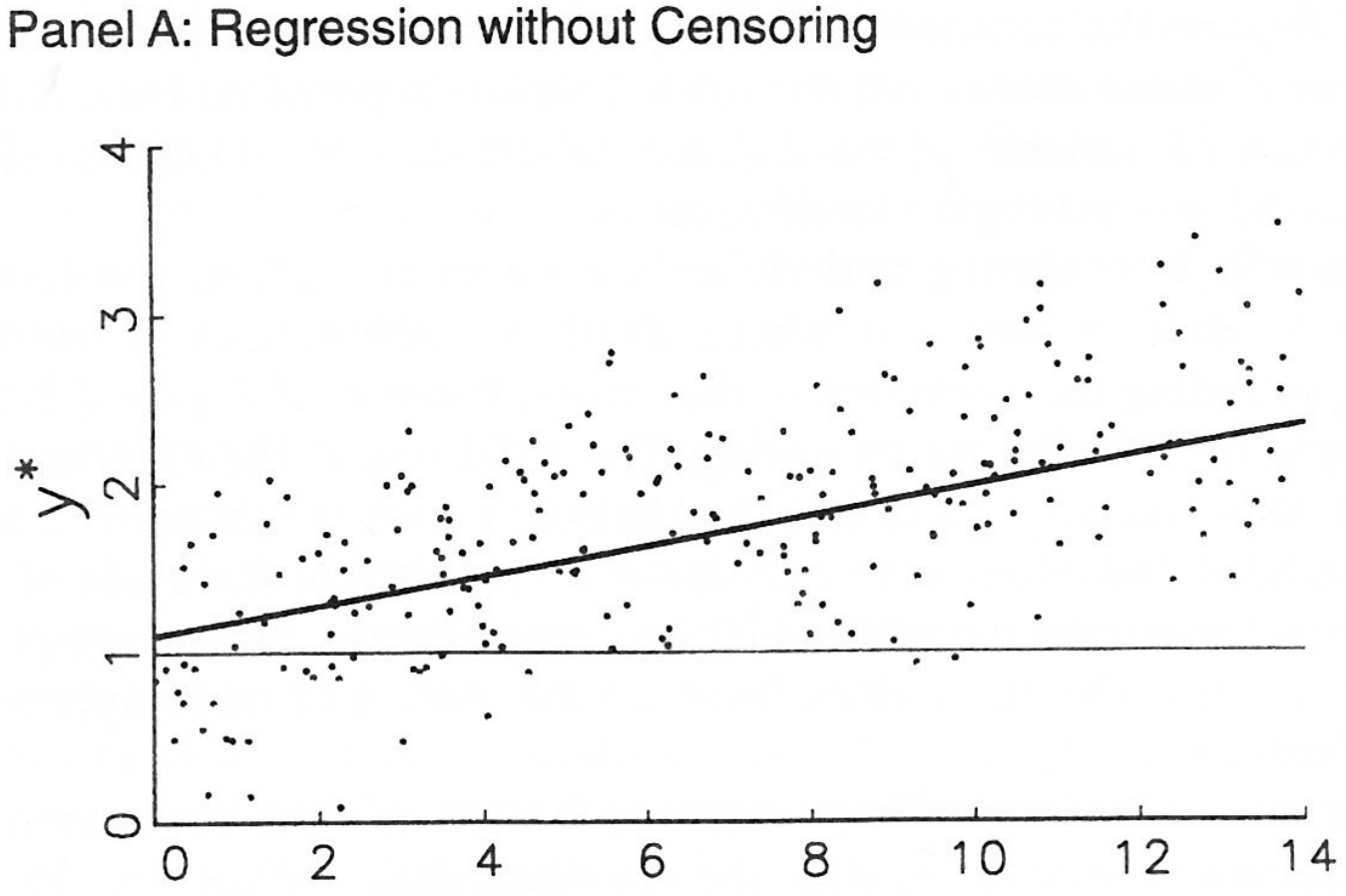

Figure: Long 1997.

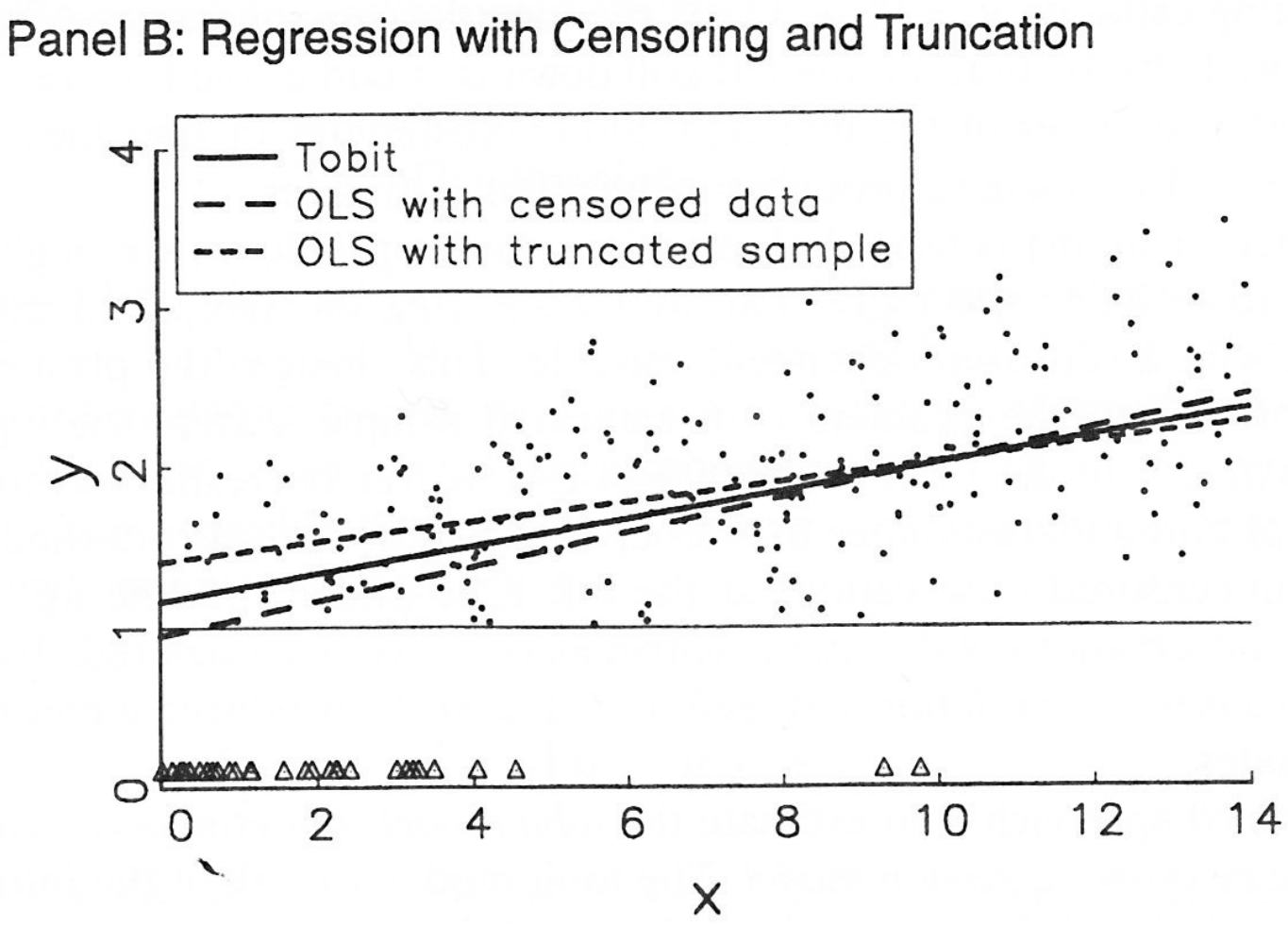

Figure: Long 1997.

Truncation and censoring cause inconsistency in OLS estimates of the slope parameter (increasing n doesn't solve the problem if the additional observations come from the same data-generating process).

- The typical censored normal regression model (a.k.a. Tobit, after James Tobin) starts with a latent (incompletely observed) variable y^*. For truncation from below, y^* is only observed if y^* exceeds a threshold. Thus,

$$y^* = \mathbf{x}'\beta + \varepsilon,$$

$$\varepsilon = \mathcal{N}[0, \sigma^2]$$

- The observed y is defined by:

$$y = \begin{cases} y^* & \text{if} \quad y^* > 0, \\ - & \text{if} \quad y^* \leq 0, \end{cases}$$

where – means y is observed to be missing. No particular value of y is necessarily observed, although it is common that we observe $y = 0$.

- Censoring and truncation change both the conditional mean and the conditional density.
- Consider MLE given censoring from below. For $y > L$, where L is the lower bound, the density of y is the same as that for y^*: $f(y|\mathbf{x}) = f^*(y|\mathbf{x})$.
- For $y = L$, the density is discrete with mass equal to the probability of observing $y^* \leq L$, or $F^*(L|\mathbf{x})$. Thus,

$$f(y|\mathbf{x}) = \begin{cases} f^*(y|\mathbf{x}) & \text{if} \quad y > L, \\ F^*(L|\mathbf{x}) & \text{if} \quad y = L. \end{cases}$$

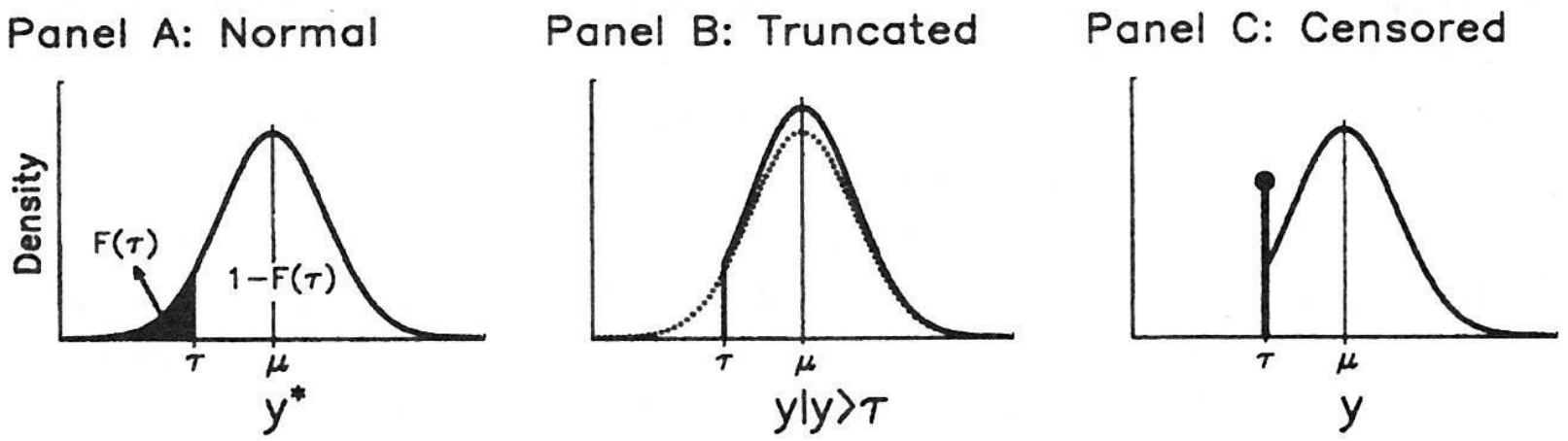

Figure 7.3. Normal Distribution With Truncation and Censoring

Figure: Long 1997.

- The density can then be written as:

$$f(y|\mathbf{x}) = f^*(y|\mathbf{x})^d F^*(L|\mathbf{x})^{1-d},$$

where $d = 1$ if $y > L$, and $d = 0$ if $y = L$.

- We can use this to compute the log-likelihood of the Tobit model. The censored density can be expressed as:

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y - \mathbf{x}'\beta)^2\right\}\right]^d \left[1 - \Phi\left(\frac{\mathbf{x}'\beta}{\sigma}\right)\right]^{1-d}.$$

- The density, therefore, is composed of two parts; uncensored observations ($d = 1$) contribute information to the first part, while censored observations ($d = 0$) contribute information to the second part (you'll notice that event history models deal with censoring in the same way).

- The Tobit MLE maximizes the censored log-likelihood function:

$$\ln L_N(\beta, \sigma^2) = \sum_i^N \Big\{ d_i \Big(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y_i - \mathbf{x_i'}\beta)^2 \Big) + (1 - d_i)\ln\Big(1 - \Phi\Big(\frac{\mathbf{x_i'}\beta}{\sigma}\Big)\Big)\Big\}$$

Limitations:

- A major weakness of the Tobit MLE is its heavy reliance on distributional assumptions
- Tobit assumes homoskedastic and normal errors. If either of those assumptions is violated, the Tobit MLE is inconsistent
- Consistent estimation with heteroskedastic errors can be done by specifying a model for the error variance: $\sigma_i^2 = \exp(\mathbf{z_i}'\gamma)$. Consistency then requires correct specification of the error variance model
- Tobit restricts the censoring mechanism to be from the same model as that generating the outcome variable. On the other hand, two-part models allow the censoring mechanism and the outcome to be modeled using separate processes (e.g., the Heckman selection model and zero-inflated count models)

Heckman selection

- In the standard Tobit model, the outcome variable y^* is observed if y^* exceeds some threshold (e.g., 0)
- We can make the model more general by letting the outcome variable be observed as a function of a second latent variable
- Let y_2^* denote the outcome of interest. We can introduce a different latent variable, y_1^*, and the outcome y_2^* is observed if $y_1^* > 0$ and not observed otherwise.
- E.g., y_1^* determines whether or not to work and y_2^* determines how much to work. Or y_1^* determines whether or not two countries sign a treaty and y_2^* determines how flexible the treaty is.

- A bivariate sample selection model then comprises a **selection equation**:

$$y_1 = \begin{cases} 1 & \text{if} \quad y_1^* > 0, \\ 0 & \text{if} \quad y_1^* \leq 0. \end{cases}$$

- And an **outcome equation**:

$$y_2 = \begin{cases} y_2^* & \text{if} \quad y_1^* > 0, \\ - & \text{if} \quad y_1^* \leq 0. \end{cases}$$

- The standard model specifies a linear model for the latent variables:

$$y_1^* = \mathbf{x_1'}\beta_\mathbf{1} + \varepsilon_1,$$
$$y_2^* = \mathbf{x_2'}\beta_\mathbf{2} + \varepsilon_2.$$

Where does selection bias come from?

E.g., suppose y_2^*, the outcome of interest, is wages and we want to estimate the effect of education on wages. y_1^* represents the utility of entering the labor market or the propensity to work. We might believe education also influences a person's decision to work.

Heckman selection

There are two selection effects at work:

- First, more educated people might be more likely to enter the labor force, and so we'll have a sample of educated people. This non-random aspect of the sample is what is commonly misunderstood to be the problem of 'selection bias' (see Sartori 2003).
- But this on its own does not bias the estimation of the outcome equation.

- The second selection effect is that some uneducated people will choose to enter the work force. This is because they decide that work is worthwhile because they have a high value on some unmeasured variable, which is captured by ε_1. For example, they may be smarter, but intelligence is not measured in our sample. That is, these people get into the sample not because they have high education, but because they have large error terms. The problem is that, whether or not education is correlated with the unmeasured intelligence in the overall population, these two variables will be correlated in the selected sample. If intelligence does lead to higher wages, then we will underestimate the effect of education on wages, because in the selected sample people with little education are unusually smart.

Heckman selection

- The Heckman selection model is about **selection on unobservables**.
- The problem arises because the error term in the selection equation, ε_1, is correlated with the error term in the outcome equation, ε_2. Errors in the outcome equation will be correlated with explanatory variables.
- Thus, inconsistent estimates of β_2.
- Heckman (1976, 1979) showed that the selection problem can be treated as an **omitted variable problem**.

Heckman selection

- Assuming the correlated errors are joint normally distributed and homoskedastic:

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N} \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right]$$

- The bivariate sample selection model has likelihood function

$$\mathsf{L} = \prod_i^n \{Pr[y_{1i}^* \leq 0]\}^{1-y_{1i}} \{f(y_{2i}|y_{1i}^* > 0) \times Pr[y_{1i}^* > 0]\}^{y_{1i}}$$

- We can use ML to estimate the parameters.

- Heckman proposed a **two-step estimator** (LIML, limited-information maximum likelihood).
- For the subsample with a positive y_2^*, the conditional expectation of y_2^* is given by:

$$E(y_2^*|\mathbf{x_2}, y_1^* > 0) = \mathbf{x_2'}\beta_\mathbf{2} + E(\varepsilon_2|\varepsilon_1 > -\mathbf{x_1'}\beta_\mathbf{1})$$

- Given the assumption of joint normality and homoskedasticity of the errors, it can be shown that

$$E(\varepsilon_2|\varepsilon_1 > -\mathbf{x_1'}\beta_\mathbf{1}) = \frac{\sigma_{12}}{\sigma_1}\frac{\phi(\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1)}{\Phi(-(\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1))}$$

- Heckman's two-step proposal is to estimate the so-called inverse Mills ratio

$$\lambda(\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1) = \frac{\phi(\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1)}{\Phi(-(\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1))}$$

  by way of a Probit model, and then estimate the following equation using OLS:

$$y_2 = \mathbf{x_2'}\beta_\mathbf{2} + \sigma_{12}\lambda(\widehat{\mathbf{x_1'}\beta_\mathbf{1}/\sigma_1}) + \varepsilon_2$$

Heckman selection

- OLS standard errors will be wrong, though:
  - First, because of heteroskedasticity in ε_2. Could correct for heteroskedasticity by using robust s.e.'s, but...
  - Second, $\hat{\lambda}$ is an estimator of λ, so the inverse Mills Ratio is estimated with uncertainty. Need to take uncertainty into account.
- Heckman 1979 provides a correction

Heckman selection

- Sample selection problem as a special case of omitted variable problem (λ being the omitted variable).
- Heckman's estimator is consistent if ε_1 is normally distributed and ε_2 is independent of λ.
- Two-step procedure has efficiency loss compared to FIML
- But needs less restrictive distributional assumptions than FIML estimator (MLE requires joint normality of errors)
- Still, LIML estimates very sensitive to distributional assumptions
- Distributional assumptions of LIML can be weakened even further to permit semiparametric estimation

Identification issues:

- The bivariate sample selection model is theoretically identified without any restriction on the regressors. That is, $\mathbf{x_1}$ and $\mathbf{x_2}$ can be the same
- If $\mathbf{x_1} = \mathbf{x_2}$, then parameters are identified by nonlinearity of $\lambda(\cdot)$ (inverse Mills Ratio)
- However, $\lambda(\cdot)$ is approximately linear over a wide range of its argument. Thus, likely **collinear**
- Identification when $\mathbf{x_1} = \mathbf{x_2}$ is fragile. Results not robust
- If $\mathbf{x_2}$ and λ highly correlated, then Heckman model might do worse than OLS on the selected sample. Cure could be worse than the disease.

Heckman selection

- Problem is less severe the greater the variation in $\widehat{\mathbf{x}_1'\beta_1}$ across observations (i.e., the better a probit model can discriminate between participants and nonparticipants)
- In practice, one needs one or more variables in $\mathbf{x_1}$ that are good predictors of y_1^* and do not appear in $\mathbf{x_2}$ (exclusion restriction). But hard to find them!
- Don't just remove a variable from $\mathbf{x_2}$ or add any variable to $\mathbf{x_1}$. Theoretically unmotivated and leads to misspecification of either or both equations.

Software implementation:

- Tobit:
  - Stata: `tobit`
  - R: `vglm` function of the `VGAM` package
- Heckman selection model:
  - Stata: `heckman` (both FIML and two-step estimators)
  - R: `heckit` function of the `sampleSelection` package

Recommended reading:

- Long, Scott J. 1997. *Regression Models for Categorical and Limited Dependent Variables.* (Chapter 7)
- Cameron, A. Colin & Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* (Chapter 16)
- Heckman, J. J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic Social Measurement* 5(4): 475–492.
- Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 53–161.
- Puhani, Patrick A. 2000. "The Heckman Correction for Sample Selection and its Critique." *Journal of Economic Surveys* 14(1): 53–68.

Recommended reading (cont.):

- Winship, Christopher & Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18: 327–350.
- Stolzenberg, Ross M. & Daniel A. Relles. 1997. "Tools for Intuition about Sample Selection Bias and its Correction." *American Sociological Review* 62 (June 1997): 494–507.
- Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources* 33(1): 127–169.
- Sartori, Anne E. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11: 111–138.
- Nooruddin, Irfan. 2002. "Modeling Selection Bias in Studies of Sanctions Efficacy." *International Interactions* 28: 59–75.