Forecasting & Predictive Analytics

Guillaume Chevillon
chevillon@essec.edu

ESSEC Business School

October-December 2017
Limited Dependent Variables

Limited Dependent Variables

- An LDV is broadly defined as a dependent variable whose range of values is substantively restricted:
  - A binary variable can only take two values: zero or one; percentages must be between 0 and 100; the number of times an individual is arrested in a give year is a nonnegative integer; GPA is between zero and 4.0 in most colleges
  - most variables are limited in some way, but not all of them need special treatment. Discrete variables can often be analyzed by linear models. But not always
  - Corner solutions to optimizing behaviors often arise (think of minimum wage, the firm may be wishing to pay the employee less).
  - Count variables
  - Duration

A-Dependent Binary variable

- We can imagine that tolerance y_i^* of an insect i to the insecticide is Normally distributed across insects, say

$$y_i^* \sim \mathsf{N}\left(\mu, \sigma_\varepsilon^2\right).$$

If an insect's tolerance is less than the dose x_i of the insecticide, the insect dies. The problem is that we cannot observe the tolerance of a particular insect: we only observe whether or not it dies. That is we observe y_i such that

$$y_i = \begin{cases} 1 & \text{if the insect dies} \\ 0 & \text{otherwise} \end{cases}$$

- Given this setup, the question of interest is: *what is the probability that insect i dies?* It is merely

$$P\left(y_i = 1\right) = P\left(y_i^* < x_i\right)$$

and hence the observations are generated by the following rule:

$$y_i = \begin{cases} 1 & \text{if } y_i^* < x_i \\ 0 & \text{otherwise} \end{cases}$$

In this example y^* is called a *latent or index variable*

Formulating a probability model

- Now our model is

$$y_i^* = X_i\beta + \varepsilon_i$$

  but we only observe y_i which takes values 0 or 1. (X_i is a row vector $X_i = [x_{i1}...x_{ik}]$)
- We would like to transform $X\beta$ into a *probability*. That is, we need a function F such that

$$P(y_i = 1) = F(X_i\beta)$$

  A natural choice for F is a distribution function or cumulative density as they lie between 0 and 1. The identity

$$P(y_i = 1) = X_i\beta$$

  does not yield the type of functions that we want.

- Choosing F to be the standard Normal leads to an attractive possibility; this is called the *Probit* model:

$$P\left(y_i = 1\right) = \Phi\left(X_i\beta\right) = \int_{-\infty}^{X_j\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

  The standard normal transformation constrains the probability to lie between zero and one.

- Choosing F to be the logistic distribution yields another attractive possibility, the *logit* model

$$P\left(y_i = 1\right) = \Lambda\left(X_i\beta\right) = \frac{\exp X_i\beta}{1 + \exp X_i\beta}.$$

  We are not constrained to restrict our choice to these two functions, but these are often used as they have **behavioral** interpretations. (we will denote $\partial\Phi = \phi$ and $\partial\Lambda = \lambda$)

A.1.-The Probit

Assume that we observe y that takes on one of two values 0 and 1. Define a latent variable y^* such that

$$y_i^* = X_i\beta + \epsilon_i$$

We do not observe y^* but rather y, which takes values 0 or 1 according to the following rule

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

We also assume that $\epsilon_i \sim \mathsf{N}\left(0, \sigma^2\right)$.

Hence y_i^* is distributed Normally, conditional on X, but y_i is not. It is straightforward to show that the previous rule generates a probit:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0) = P(X_i\beta + \epsilon_i > 0) \\ &= P(\epsilon_i > -X_i\beta) = P\left(\frac{\epsilon_i}{\sigma} > -\frac{X_i\beta}{\sigma}\right) \end{aligned}$$

and

$$\frac{\epsilon_i}{\sigma} \sim \mathsf{N}(0,1)$$

so by symmetry

$$\begin{aligned} P(y_i = 1) &= P\left(\frac{\epsilon_i}{\sigma} > -\frac{X_i\beta}{\sigma}\right) = P\left(\frac{\epsilon_i}{\sigma} < \frac{X_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{X_i\beta}{\sigma}\right) \end{aligned}$$

Probit/Logit must be estimated by maximum likelihood. Deriving the likelihood function is easy since

$$P\left(y_i = 0\right) = 1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)$$

and if we denote $i = 1, ..., m$ the observations such that $y_i = 0$ and $y_i = 1$ for $i = m+1, ..., n$, then

$$\begin{aligned} L &= P\left(y_1 = 0\right) \cdot P\left(y_2 = 0\right) \cdot ... P\left(y_m = 0\right) \\ &\quad \cdot P\left(y_{m+1} = 1\right) \cdot ... P\left(y_n = 1\right) \\ &= \prod_{i=1}^{m} \left[1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right] \prod_{i=m+1}^{n} \Phi\left(\frac{X_i\beta}{\sigma}\right) \\ &= \prod_{i=1}^{n} \left[\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{y_i} \left[1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-y_i} \end{aligned}$$

and the log-Likelihood is given by

$$\log L$$
$$= \sum_j \left\{ y_i \log \left[\Phi \left(\frac{X_i \beta}{\sigma} \right) \right] + (1 - y_i) \log \left[1 - \Phi \left(\frac{X_i \beta}{\sigma} \right) \right] \right\}$$

Another important aspect is that the coefficients β and σ always appear together: there are not separately identified: only their ratio matters. It is therefore convenient to normalize $\sigma = 1$.
The logL is globally concave but cannot be solved analytically, yet it is easy to maximize it using standard algorithms.

To test that $\beta = 0$, we use the likelihood ratio test

$$2\left[\log L\left(\alpha, \beta^*\right) - \log L\left(\alpha, 0\right)\right] \sim \chi^2\left(p-1\right)$$

where $\log L\left(\alpha, \beta^*\right)$ is the logL with a constant and the $p-1$ parameters β^* and $\log L\left(\alpha, 0\right)$ is the restricted model with only a constant.
How can me move from β to the impact of a variable on y?
Using

$$\frac{\partial \mathsf{E}\left[y\right]}{\partial x_j} = \frac{\partial \mathsf{P}\left[y = 1|X\right]}{\partial x_j} = \phi\left(X\beta\right)\beta_j$$

with the standard normal density

$$\phi\left(z\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Hence *the derivative of the probability with respect to x varies with the level of* X, i.e. *all the variables in the model.* Hence, the probit necessitates more information than just reporting the coefficients. It can be useful to compute the derivatives of the model at the mean values.

Partial effects

- If x_j is (roughly) continuous

$$\Delta \mathsf{P}\left[\widehat{y = 1}|X\right] \approx \phi\left(X\widehat{\beta}\right)\widehat{\beta}_j \Delta x_j$$

  but the scale factor $\phi\left(X\widehat{\beta}\right)$ depends on X.

- One option is to use $\phi\left(\overline{X}\widehat{\beta}\right)$, that of the "average" observation
- The former has no meaning if a regressor is binary (what is a person who is 47% female?), hence the use of the **average partial effect**:

$$n^{-1}\sum_{i=1}^{n}\phi\left(X_i\widehat{\beta}\right)$$

- For discrete regressors δ_i, such that $X\widehat{\beta} = X_*\widehat{\beta}_* + \delta\widehat{\beta}_k$ simply look at

$$\Phi\left(X_*\widehat{\beta}_* + (\delta+1)\,\widehat{\beta}_k\right) - \Phi\left(X_*\widehat{\beta}_* + \delta\widehat{\beta}_k\right)$$

Comparing estimates between Probit, Logit and Linear models

- $\phi(0) \approx .4$ and $\lambda(0) = .25$ hence to make magnitudes of probit and logit comparable, we can multiply probit coefficients by $.4/.25 = 1.6$, or multiply logit estimates by .625.
- In the linear model the equivalent value is 1, hence logit estimates should be divided by 4 and probit by 2.5 to be comparable to those of a linear regression.

Goodness of Fit

There exist various **pseudo R-squared** measures for binary response. McFadden (1974) suggests the measure

$$1 - \frac{\log L_{UR}}{\log L_O}$$

where $\log L_{UR}$ is logL in the estimated model and ℓ_O is the logL in the model with only an intercept.
Another option is to compute the squared correlation between y_i and $\widehat{y}_i$.

A.2.-The Logit

The logit is similar to the probit, with now

$$P\left(y_i = 1\right) = \Lambda\left(X_i\beta\right) = \frac{\exp X_i\beta}{1 + \exp X_i\beta}$$

The latent variable interpretation is the same, but with ϵ_i which now follows what is called *a logistic distribution*. The main difference between the Normal and Logistic distributions is that the latter has more weight in the tails.
The derivative of the probability with respect to one element is now given by

$$\begin{aligned}\frac{\partial \mathsf{E}\left[y\right]}{\partial x_j} &= \frac{\exp X\beta}{\left(1 + \exp X\beta\right)^2}\beta_j \\ &= q\left(1 - q\right)\beta_j\end{aligned}$$

where $q = \frac{\exp X\beta}{1+\exp X\beta}$

Misspecification

What happens in these models if σ varies across individuals, i.e. the model is *heteroskedastic?* for instance if $\sigma_i = X_i\beta / X_i\gamma$? Then we get inconsistent estimates.
Fortunately, all is not so sombre. If we use the right specification and the model is linear, then the Maximum Likelihood Estimator is consistent even in the face of heteroscedasticity, nonnormal errors and serial correlation, provided that

$$\text{plim}\frac{1}{n}X'\epsilon = 0$$

i.e. that the regressors are orthogonal to the errors.
An estimate obtained under these circumstances is called a Quasi-Maximum Likelihood Estimator.
Unfortunately, here **any misspecification will result in inconsistent estimates**. But there are some robust estimators of the standard errors.

B-Multinomial models

- If a survey question allows for more than two answers then the qualitative variable is discrete but not binary
- Three main models:
  1. Ordered models: e.g. you want to know whether an individual will buy a product (depending on her revenue, household composition...), then the variable ”Purchase at price p” can take ”Yes”, ”Maybe” or ”No”.
  2. Sequential models: the outcome are seen at steps, e.g. level of studies: 0 for pre-high school, 1 for high school, 2 for college, 3 for postgraduate.
  3. Unordered models: this is the most important category which includes the multinomial logit.

we study them in turn.

B.1.-Ordered models

For instance assume that $y_i =$"Purchase at price p" can take values 1,2 and 3. We can them assume that there is a latent variable y_i^* such that

$$\begin{aligned} y_i &= 1 \text{ if } y_i^* - p \geq c_1 \\ y_i &= 2 \text{ if } c_1 > y_i^* - p \geq c_2 \\ y_i &= 3 \text{ if } c_2 > y_i^* - p \end{aligned}$$

where c_1 and c_2 must be estimated. Then again we assume that

$$y_i^* = X_i\beta + u_i$$

hence

$$
\begin{aligned}
P(y_i = 1) &= P(X_i\beta + u_i - p \geq c_1) = 1 - F(p + c_1 - X_i\beta) \\
P(y_i = 2) &= P(c_1 > X_i\beta + u_i - p \geq c_2) \\
&= F(p + c_1 - X_i\beta) - F(p + c_2 - X_i\beta) \\
P(y_i = 3) &= P(c_2 > X_i\beta + u_i - p) = F(p + c_2 - X_i\beta)
\end{aligned}
$$

and the Likelihood is ($y_{ij} = 1$ if $y_i = j$, 0 otherwise)

$$
\begin{aligned}
L = \prod_{i=1}^{n} & [1 - F(p + c_1 - X_i\beta)]^{y_{i1}} \\
& \times [F(p + c_1 - X_i\beta) - F(p + c_2 - X_i\beta)]^{y_{i2}} \\
& \times F(p + c_2 - X_i\beta)^{y_{i3}}
\end{aligned}
$$

note that different functions can be used for difference thresholds (β_1 and β_2) and can also be extended.

B.2.-Sequential models

Consider the model with the level of studies. This type of model considers the probability to stop at step j, which we define as $F_j\left(X_i, \beta\right)$, hence

$$P\left(Y_i = j\right) = F_j\left(X_i, \beta\right) \prod_{s=1}^{j-1} \left[1 - F_s\left(X_i, \beta\right)\right]$$

and we can construct the likelihood.

B.3.-Multinomial Logit

- In the simple Logit model
  $P(y_i = 1) = \exp(X_i\beta) / [1 + \exp(X_i\beta)]$ hence since
  $P(y_i = 1) + P(y_i = 0) = 1$, a unique vector β is sufficient to describe the two alternatives. The probabilities thereof only vary according to the characteristics X_i.
- In case of several alternatives, we have to allow either for β or for X_i to vary according to the modalities.
- Notice that in the binary case, we can define β_0 and β_1 such that $\beta = \beta_1 - \beta_0$ and then

$$P(y_i = j) = \frac{\exp\left(X_i\left(\beta_j - \beta_0\right)\right)}{\sum_{k=0,1}\exp\left(X_i\left(\beta_k - \beta_0\right)\right)}$$

$$\frac{P(y_i = 1)}{P(y_i = 0)} = \frac{\exp(X_i\beta_1)}{\exp(X_i\beta_0)}$$

We can extend the previous to $m+1$ modalities with probabilities $(p_0, p_1, ..., p_m)$. Category 0 is taken as a reference, then

$$\frac{p_1}{p_1+p_0} = H\left(X_i\beta_1\right), ..., \frac{p_j}{p_j+p_0} = H\left(X_i\beta_j\right)$$

where H is a continuous and increasing function. We can solve

$$\frac{p_j}{p_0} = \frac{H\left(X_i\beta_j\right)}{1-H\left(X_i\beta_j\right)} = G\left(X_i\beta_j\right)$$

G is also increasing. Then we can solve as

$$p_j = \frac{G\left(X_i\beta_j\right)}{1+\sum_{k=1}^{m} G\left(X_i\beta_k\right)} = \frac{G\left(X_i\beta_j\right)}{\sum_{k=0}^{m} G\left(X_i\beta_k\right)}$$

if $G\left(X_i\beta_0\right) = 1$. We can let $G = \exp$ so $\beta_0 = 0$.

Estimation of the multinomial logit

- In this model, the parameters depend on the alternatives, but explanatory variables do not.
- Probabilities are expressed as ratios of exponentials $p_j/p_k = \exp\left(X_i\beta_j\right) / \exp\left(X_i\beta_k\right)$ hence:
  1. ratios are indifferent to a translation of parameters $\beta_j^* = \beta_j + \beta$
  2. these ratios change if the explanatory variables change
  3. parameters are seen as deviations from the reference β_0, the latter cannot be estimated.
-

$$\log L = \sum_{i=1}^{n} y_{i1} \log p_{i1} + ... + y_{im} \log p_{im} + y_{i0} \log p_{i0}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{m} y_{ik} X_i\beta_k - (m+1)\sum_{i=1}^{n} \log\left(1 + \sum_{k=1}^{m} \exp\left(X_i\beta_k\right)\right)$$

with partial effects $\partial p_{ij}/\partial X_i = p_{ij}\left[\beta_j - \sum_{k=0}^{m} p_{ik}\beta_k\right]$

B.4.-McFadden' conditional logit (Probabilistic choice)

Here the vector of parameters is constant across alternatives, but the variables vary:

$$P\left(y_i = j\right) = \frac{\exp\left(X_{ij}\beta\right)}{\sum_{k=0}^{m} \exp\left(X_{ik}\beta\right)} = \frac{\exp\left(X_{ij}\beta\right)}{1 + \sum_{k=1}^{m} \exp\left(X_{ik}\beta\right)}$$

$$\frac{P\left(y_i = j\right)}{P\left(y_i = k\right)} = \frac{\exp\left(X_{ij}\beta\right)}{\exp\left(X_{ik}\beta\right)} = \exp\left[\left(X_{ij} - X_{ik}\right)\beta\right]$$

where variables are seen as deviations from the reference point. This model is very useful when one wishes to predict the probability of a new (virtual) modality according to simulated explanatory variables.

Example

- Assume a municipality is considering creating a metro system. We want to model the probability that citizens will use it.
- Modalities are 1 (bus), 2 (car), 3 (bike), 0 (any other, such as walking, rollers, hitch-hiking...)
- Explanatory variables are average time from home to work using mode of transportation j (x_{1ij}), cost per mile of that mode (x_{2ij}).
- The probability that individual i uses mode j, say bus $j = 1$, is

$$P\left(i \text{ takes the bus}\right) = \frac{\exp\left(\beta_0 + \beta_1 x_{1i1} + \beta_2 x_{2i1}\right)}{1 + \sum_{k=1}^{3} \exp\left(\beta_0 + \beta_1 x_{1ik} + \beta_2 x_{2ik}\right)}$$

- Since mode 4 (metro) does not exist, x_{1i4} and x_{2i4} are unknown, but they can easily be estimated (simulated) and the probability that citizens will be using it estimated.

Independence of Irrelevant Alternatives (IIA)

- This is a strong property of multinomial logit models: p_j/p_k does not depend on other alternatives.
- This is probably untrue when choice is between similar alternatives
- Assume that before metro is introduced bus is 30%, car 50%, bike 10%, other 10%, then the ratio of prob. bus over the others is 30/70=42.85%
- This ratio shouldn't change when metro is introduced for IIA to hold.
- It is likely that only former users of bus will now take the metro, assume that it's split evenly $P(bus) = P(metro)$, then IIA implies that $P(metro)/P(others) = .4285$, which when solving leads to $P(others) = .5384 < .70$ which isn't compatible with our assumption.

C-Corner Solutions: the Tobit model

- Assume you want to model monthly alcohol expense in the U.S.
- A large number of individual will spend zero
- How do you account for a nonzero probability that the variable is zero?
- If you model $y_i = X_i\beta + \varepsilon_i$, then
  1. the predicted $\widehat{y}_i$ could possibly be negative
  2. $y_i|X_i$ cannot be Normal (why?)
- The tobit model assumes a latent variable $y_i^* = X_i\beta + \varepsilon_i$, $\varepsilon_i|X_i \sim \mathsf{NID}\left(0, \sigma^2\right)$ that is observed only if it is positive

$$y_i = \max\left(y_i^*, 0\right)$$

What is the likelihood function for a Tobit model?

- The density of y_i is the same as that of y_i^* for positive values:

$$f_y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - X\beta)^2} = \frac{1}{\sigma}\phi\left(\frac{y - X\beta}{\sigma}\right)$$

  where ϕ is the density of a $\mathsf{N}(0,1)$.
- When y_i^* is negative, since $\frac{\varepsilon_i}{\sigma}|X_i \sim \mathsf{NID}(0,1)$

$$\begin{aligned} P(y_i = 0|X_i) &= P(y_i^* < 0|X_i) = P(\varepsilon_i < -X_i\beta|X_i) \\ &= P\left(\frac{\varepsilon_i}{\sigma} < -\frac{X_i\beta}{\sigma}|X_i\right) = \Phi\left(-\frac{X_i\beta}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{X_i\beta}{\sigma}\right) \end{aligned}$$

- Hence the log likelihood follows:

$$\log L(\beta,\sigma) = \sum_{i=1}^{n} 1_{y_i>0} \log\left[\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right] + 1_{y_i=0} \log\left[1 - \Phi\left(\frac{X
$$

Interpreting the estimates

What is the expected value of y?

$$\begin{aligned}\mathsf{E}[y|X] &= \mathsf{E}[y|y>0,X]\,P(y>0|X) + \mathsf{E}[y|y=0,X]\,P(y=0|X)\\ &= \mathsf{E}[y|y>0,X]\,P(y>0|X) = \Phi\left(\frac{X_i\beta}{\sigma}\right)\mathsf{E}[y|y>0,X]\end{aligned}$$

where

$$\mathsf{E}[y|y>0,X] = X\beta + \mathsf{E}[\varepsilon|\varepsilon > -X\beta, X]$$

and we use, for $z \sim \mathsf{N}(0,1)$

$$\mathsf{E}[z|z>c] = \frac{\phi(c)}{1-\Phi(c)}$$

hence, since ϕ is symmetric and $1-\Phi(-c) = \Phi(c)$

$$\mathsf{E}[y|y>0,X] = X\beta + \sigma\frac{\phi(X\beta/\sigma)}{\Phi(X\beta/\sigma)}$$

where $\frac{\phi(X\beta/\sigma)}{\Phi(X\beta/\sigma)} = \lambda(X\beta/\sigma)$ is called the *inverse Mills ratio.*

We therefore see that

$$\mathsf{E}\left[y|X\right] = X\beta\Phi\left(X\beta/\sigma\right) + \sigma\phi\left(X\beta/\sigma\right)$$

where the expression to the right hand-side is always positive. It can also be shown that

$$\frac{\partial \mathsf{E}\left[y|X\right]}{\partial x_j} = \beta_j\Phi\left(X\beta/\sigma\right)$$

To compare OLS and Tobit estimates, we must therefore scale by $\Phi\left(X\beta/\sigma\right)$ which can be approximated with

$$n^{-1}\sum_{i=1}^{n}\Phi\left(X_i\widehat{\beta}/\widehat{\sigma}\right)$$

Notice that $\Phi\left(X\beta/\sigma\right)$ is always between zero and one!

Issues with Tobit

As usual with econometric models, we are restricted in our assumptions. A problem with Tobit models is that $\mathsf{E}[y|y > 0]$ is closely linked to $P(y > 0)$ and this can be limiting.
Example: assume that you want to relate people's life insurance coverage with age. Then the probability that $y > 0$ increases with age. Conditional on having life insurance, the value of the policy may be decreasing with age. This is not allowed in Tobit.
A way to test for Tobit is by estimating a Probit for $y = 0$ or $y > 0$ and checking whether the coefficients estimates are close to the $\widehat{\beta}/\widehat{\sigma}$ computed from a Tobit.

D-Censored and Truncated Regression

- Censored and truncated regression models arise where we cannot observe some data. This is very different from the Tobit where we observe whether somebody consumes alcohol or not, but where some chose a so-called 'corner solution'.
- Tobit as we have considered it is only for values being zero or positive. In fact this can be moved to a threshold. What we need to bear in mind is that there is *behavioral* reason why some data are all at a certain level.
- In practice it may be possible that we simply do not observe some values, but this is due to survey design (or legal constraints about data availability).
- Censoring occurs when y is missing, but we know that it is missing because it is above a threshold
- Truncation occurs when we exclude, on the basis of y, a subset of the population from our sampling scheme.
- These two methods mean that the *sample is not random*

D.1.-Censoring

- Example: Top coding. A variable is top coded when we only know its value up to a certain level (say that you report wage, and the top value is "wage is 100,000 or above" because you are not allowed to report wage when it is too high (why?)
- A Censored model is of the form:

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i|X_i, c_i \sim \text{NID}\left(0, \sigma^2\right)$$
$$w_i = \min\left(y_i, c_i\right)$$

- Here ε_i is independent of c_i and $P\left(w_i = c_i|X_i\right) = P\left(y_i \geq c_i|X_i\right) = P\left(u_i \geq c_i - X_i\beta|X_i\right)$
- log-Likelihood is

$$\ell\left(\beta, \sigma\right) = \sum_{i=1}^{n} 1_{w_i<c_i} \log\left[\frac{1}{\sigma}\phi\left(\frac{w_i - X_i\beta}{\sigma}\right)\right] + 1_{w_i=c_i} \log\left[1 - \Phi\left(\frac{c_i -}{}\right.\right.$$

- β is interpretable as in the linear model. (other example, see Wooldridge example 17.4) But standard OLS would give too small values

D.2.-Truncation

- In a truncated model, we do not observe a certain fragment of the population. Example: you only consider people who are 1.5 times above the poverty line (which depends on family size).
- The model starts normally

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i|X_i \sim \mathsf{NID}\left(0, \sigma^2\right)$$

but we only observe X_i if $y_i \leq c_i$ where c_i depends on exogenous variables, in particular X_i. Then, if $f\left(x; \mu, \sigma^2\right)$ is the pdf of a $\mathsf{N}\left(\mu, \sigma^2\right)$ and F the corresponding cdf, the density of y_i is

$$g\left(y_i|X_i, c_i\right) = \frac{f\left(y_i; X_i\beta, \sigma^2\right)}{P\left(y_i \leq c_i|X_i\right)} = \frac{f\left(y_i; X_i\beta, \sigma^2\right)}{F\left(c_i; X_i\beta, \sigma^2\right)}$$

D.3.-Sampling

- endogenous sample selection (e.g. Tobit): OLS is biased and inconsistent
- exogenous sample selection: observations are selected purely depending on X_i, then OLS is unbiased and consistent
- random sampling: OLS unbiased and consistent