Maximum Likelihood and Classical Non-linear Models Review

Advanced Microeconometrics
ITAM

Instructor: Cristián Sánchez

Spring 2022

Maximum Likelihood

- The principle of ML provides a means of choosing an asymptotically efficient estimator.
- ML provides consistent, efficient and asymptotically normal estimators.
- Drawback: Strong distributional assumptions and ML does not necessarily provides unbiased estimators.

The Basic Idea

- Consider a **random** sample of 10 observations from a Poisson distribution: $x = 5, 0, 1, 1, 0, 3, 2, 3, 4, 1$
- The associated density function is:

$$f(x_i; \theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

- The joint density (the likelihood) is:

$$f(x_1, \ldots, x_{10}; \theta) = \prod_{i=1}^{10} f(x_i; \theta) = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{10} x_i}}{\prod_{i=1}^{10} x_i!}$$

- This is the probability of observing this particular sample given the Poisson distribution.

The Basic Idea

- Key question: What value of θ would make this sample most probable?
- Let's apply the monotonic increasing transformation: $\ln L(\theta) = \ln(f(x_1, \ldots, x_{10}; \theta))$.
- We maximize the *log likelihood*. FOC:

$$[\theta] : \frac{\mathrm{d} \ln(L(\theta))}{\mathrm{d}\theta} = -10 + \frac{20}{\hat{\theta}} = 0 \Rightarrow \hat{\theta} = 2$$

- Notice that:

$$\frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} = -\frac{20}{\hat{\theta}^2} < 0$$

  So this is a maximum.

The ML Principle

- Assume the variables Y_i, $i = 1, 2, \ldots, N$ are independent and identically distributed with density $f(Y_i; \theta)$.
- The likelihood function, defined as a function of the unknown parameter vector, θ:

$$f(Y_1, \ldots, Y_N; \theta) = \prod_{i=1}^{N} f(Y_i; \theta) = L(\theta)$$

or alternatively

$$\ln L(\theta) = \sum_{i=1}^{N} \ln f(Y_i; \theta)$$

which is called the log likelihood function.

The ML Principle

- We denote by $\hat{\theta}$ the values of the parameters that maximize this function.
- The necessary condition is:

$$\frac{\partial \ln L(\hat{\theta})}{\partial \theta}$$

- Sufficient condition:

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta^2} < 0$$

or, in the vector case

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}$$

is negative definite.

Properties of ML Estimator

Under regularity conditions, the MLE ($\hat{\theta}$) has the following properties:

- Consistency: plim $\hat{\theta} = \theta$
- Asymptotic Normality:

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow N(0, I_1(\theta)^{-1})$$

where $I_1(\theta) = -E\left[\frac{\partial^2 \ln f(Y_1; \theta)}{\partial \theta \partial \theta'}\right]$

- Asymptotic Efficiency: $\hat{\theta}$ achieves the Cramér-Rao lower bound for consistent estimator.
- Invariance: The MLE of $\gamma = c(\theta)$ is $c(\hat{\theta})$

Properties of ML Estimator

Thus, an approximation to the sampling distribution of the ML estimator $\hat{\theta}$ for large N is:

$$\hat{\theta} \approx N\left(\theta, \frac{1}{N} I_N(\theta)^{-1}\right)$$

where $I_N(\theta)^{-1} = (N \times I_1(\theta))^{-1}$ is the Fisher information matrix for N observations (also the Cramér-Rao lower bound)

A consistent estimator of $I_1(\theta)$ is:

$$\hat{I}_1(\theta) = -\frac{1}{N} \frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}$$

Properties of ML Estimator

Additionally, under regularity conditions, if g is a continuously differentiable function, then:

$$\sqrt{N}(g(\hat{\theta}) - g(\theta)) \to N\left(0, \frac{dg(\theta)}{d\theta'} I_1(\theta)^{-1} \frac{dg(\theta)'}{d\theta'}\right)$$

As a consequence of this property, we can get the asymptotic distribution of functions of θ.

Statistical Inference

- Likelihood function is

$$\ln L(y;\theta) = \sum_{i=1}^{n} \ln f(y_i;\theta) \text{ with } \theta \in \mathbb{R}^{k}$$

- Let the test be

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

Wald Test

- Wald test statistic:

$$W = \sqrt{n}(\hat{\theta} - \theta_0)' I_{\hat{\theta}}^{-1} \sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} \chi_k^2$$

- Wald uses unrestricted estimate.
- It checks whether the null hypothesis and the relevant portion of the unrestricted estimate (which is the best choice of parameters under the alternative hypothesis) are very far apart.
- Intuitively, if the null hypothesis were true, then by the consistency of the ML estimator, the best choice under the alternative hypothesis should be getting close to the null hypothesis.

Rao's Score

- Score test statistic:

$$LM = \sqrt{n}\left(\frac{1}{n}\frac{\partial \ln L(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right)' I^{-1}_{\hat{\theta}=\theta_0} \sqrt{n}\left(\frac{1}{n}\frac{\partial \ln L(y;\theta)}{\partial \theta}\Big|_{\theta_0}\right) \overset{a}{\sim} \chi^2_k$$

- LM uses the restricted estimate: $\ln L(y;\theta) + \lambda(\theta_0 - \theta)$
- It checks whether the relevant portion of the score vector at the restricted estimate (which is the best choice of parameters under the null hypothesis) is close to zero.
- Intuitively, if the null hypothesis were true, then the gradient of the likelihood should be zero in the population at that parameter value and so restricting ourselves to that parameter value should produce a gradient close to zero.

Likelihood Ratio Test

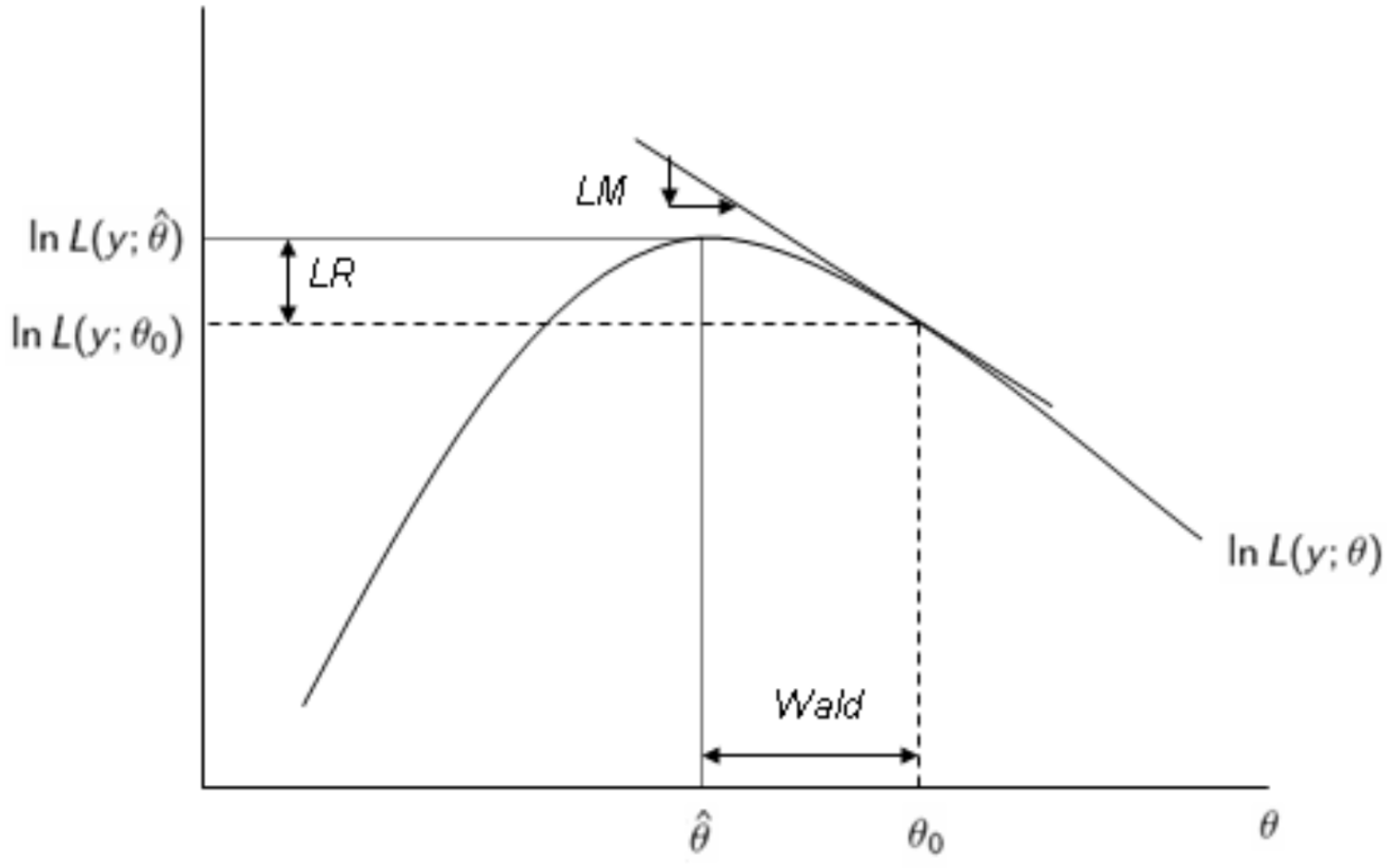
- Likelihood ratio test statistic:

$$LR = -2 \ln \left[\frac{L(y; \theta_0)}{L(y; \hat{\theta})} \right] \overset{a}{\sim} \chi^2_k$$

- Compares the highest value of the likelihood under the null hypothesis with the highest value of the likelihood under the alternative hypothesis.

- Intuitively, if the null hypothesis were true (and under our regularity conditions about the uniform convergence of the likelihood function), then the maximum of the likelihood under the null hypothesis and the maximum of the likelihood under the alternative hypothesis should be close.

Graphical Interpretation



All three are asymptotically distributed as χ^2_k

Motivation

- Suppose you are trying to estimate the following model:

$$D = 1 \text{ if and only if } \gamma Z + V \geq 0$$

  where $D = 1$ ($D = 0$) means that the behavior you are trying to understand is (not) observed.

- This is a binary choice model.
- The model is non-linear.
- We can use ML.

Here's the general approach:

$$\underbrace{\left[\begin{array}{c}\text{Economic model}\\ \text{(e.g. utility maximization)}\end{array}\right] \Rightarrow \left[\begin{array}{c}\text{Decision rule}\\ \text{(e.g. FOC)}\end{array}\right]}_{\text{Motivation: Index function and random utility models}}$$

$$\Rightarrow \underbrace{\left[\begin{array}{c}\textbf{Underlying regression}\\ \text{(e.g. solve the FOC for}\\ \text{a dependent variable)}\end{array}\right] \Rightarrow \left[\begin{array}{c}\textbf{Econometric model}\\ \text{(e.g. discrete variable}\\ \text{model)}\end{array}\right]}_{\text{Setup}}$$

$$\Rightarrow \underbrace{[\text{Estimation}]}_{\text{Estimation}} \Rightarrow \underbrace{[\text{Interpretation}]}_{\text{Marginal Effects}}$$

- We assume that we have an economic model and have derived implications of the model, e.g. FOCs, which we can test. Converting these conditions into an underlying regression usually involves little more than rearranging terms to isolate a dependent variable.

- Often this dependent variable is not directly observed, in a way that we'll make clear later. In such cases, we cannot simply estimate the underlying regression. Instead, we need to formulate an econometric model that allows us to estimate the parameters of interest in the decision rule/underlying regression using what little information we have on the dependent variable.

Suppose the marginal cost benefit calculation was slightly more complex. Let y_0 and y_1 be the net benefit or utility derived from taking actions 0 and 1, respectively. We can model this utility calculus as the unobserved variables y_0 and y_1 such that:

$$y_0 = \beta' x_0 + \varepsilon_0,$$
$$y_1 = \gamma' x_1 + \varepsilon_1.$$

Now assume that $(\varepsilon_1 - \varepsilon_0) \sim f(0, 1)$. Again, although we don't observe y_0 and y_1, we do observe y where:

$$y = 0 \text{ if } y_0 > y_1,$$
$$y = 1 \text{ if } y_0 \leq y_1.$$

In other words, if the utility from action 0 is greater than action 1, i.e., $y_0 > y_1$, then $y = 0$. $y = 1$ when the converse is true. Here the probability of observing action 1 is:

$$\begin{aligned}
\Pr\{y = 1\} &= \Pr\{y_0 \leq y_1\} = \Pr\left\{\beta' x_0 + \varepsilon_0 \leq \gamma' x_1 + \varepsilon_1\right\} \\
&= \Pr\left\{\varepsilon_1 - \varepsilon_0 \geq \beta' x_0 - \gamma' x_1\right\} \\
&= F(\gamma' x_1 - \beta' x_0).
\end{aligned}$$

The Setup

The random utility models provide the link between an underlying regression and an econometric model. Now we'll begin the process of flushing out the econometric model. First we'll consider different specifications for the distribution of ε and later examine how marginal effects are derived from our probability model. This will pave the way for our discussion of how to estimate the model.

Why $\Pr\{y = 1\}$?

- In the random utility models, the probability of observing $y = 1$ has the structure: $\Pr\{y = 1\} = F(\beta' x)$.

- Why are we so interested in the probability that $y = 1$? Because the expected value of y given x is just that probability:

$$E[y] = 0 \cdot (1 - F) + 1 \cdot F = F(\beta' x).$$

Common specifications for $F(\beta' x)$

How do we specify $F(\beta' x)$? There are three basic specifications that dominate the literature.

(a) Linear probability model (LPM):
$F(\beta' x) = \beta' x$

(b) Probit:

$$F(x) = \Phi(\beta' x) = \int_{-\infty}^{\beta' x} \phi(t)\, \mathrm{d}t = \int_{-\infty}^{\beta' x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, \mathrm{d}t$$

(c) Logit: $F(\beta' x) = \Lambda(\beta' x) = \dfrac{e^{\beta' x}}{1 + e^{\beta' x}}$

Deciding which specification to use

Each specification has its advantages and disadvantages.

(1) **LPM.** The linear probability model is popular because it is extremely simple to estimate. This simplicity, however, comes at a cost. To see what we mean:

$$\begin{aligned} y &= E[y|x] + (y - E[y|x]) \\ &= F(\beta' x) + \varepsilon \\ &= \beta' x + \varepsilon \end{aligned}$$

Deciding which specification to use

Notice that the error term:

$\varepsilon = 1 - \beta' x$ with probability $F = \beta' x$ and

$\quad - \beta' x$ with probability $1 - F = 1 - \beta' x$

This implies that:

$$\begin{aligned} var[\varepsilon|x] &= E[\varepsilon^2|x] - E^2[\varepsilon|x] = E[\varepsilon^2] \\ &= F \cdot (1 - \beta' x)^2 + (1 - F) \cdot (\beta' x)^2 \\ &= F - 2F\beta' x + F[\beta' x]^2 + [\beta' x]^2 - F[\beta' x]^2 \\ &= F - 2F\beta' x + [\beta' x]^2 \\ &= \beta' x - 2[\beta' x]^2 + [\beta' x]^2 = \beta' x(1 - \beta' x). \end{aligned}$$

So our first problem is that ε is heteroscedastic in a way that depends on β. Of course, absent any other problems, we could manage this with an FGLS estimator.

Deciding which specification to use

A second more serious problem, however, is that since $\beta' x$ is not confined to the $[0, 1]$ interval, the LPM leaves open the possibility of predicted probabilities that lie outside the $[0, 1]$ interval, which is nonsensical, and of negative variances:

$$\begin{aligned} \beta' x > 1 \Rightarrow E[y|x] = F = \beta' x > 1, \\ var[\varepsilon] = \beta' x(1 - \beta' x) < 0, \\ \beta' x < 0 \Rightarrow E[y|x] < 0, \\ var[\varepsilon] < 0. \end{aligned}$$

This is a problem that is harder to correct.

Deciding which specification to use

(2) **Probit vs. Logit.** The probit model, which uses the normal distribution, is sometimes (inappropriately) justified by appealing to a central limit theorem, while the logit model can be justified by the fact that it is similar to a normal distribution but has a much simpler form. The difference between the logit and normal distribution is that the logit has slightly heavier tails. The standard normal has mean zero and variance 1 while the logit has mean zero and variance equal to $\pi^2/3$.

Marginal Effects

Unlike in linear models, the marginal effect of a change in x on $E[y]$ is not simply β. To see why, differentiate $E[y]$ by x:

$$\frac{\partial E[y|x]}{\partial x} = \frac{\partial F(\beta' x)}{\partial \beta' x} \frac{\partial \beta' x}{\partial x} = f(\beta' x)\beta.$$

These marginal effects look different in each of the three basic probability models.

1. **LPM.** Note that $f(\beta'x) = 1$, so $f(\beta'x)\beta = \beta$, which is the same as in the LR models, as expected.
2. **Probit.** Now, $f(\beta'x) = \phi(\beta'x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\beta'x)^2}{2}}$, so $f(\beta'x)\beta = \phi\beta$.
3. **Logit.** In this case:

$$\begin{aligned} f(\beta'x) = \frac{\partial \Lambda(\beta'x)}{\partial(\beta'x)} &= \frac{e^{\beta'x}}{1+e^{\beta'x}} - \frac{e^{\beta'x}}{(1+e^{\beta'x})^2} e^{\beta'x} \\ &= \frac{e^{\beta'x}}{1+e^{\beta'x}} \left(1 - \frac{e^{\beta'x}}{1+e^{\beta'x}}\right) \\ &= \Lambda(\beta'x)(1-\Lambda(\beta'x)) \end{aligned}$$

Giving us the marginal effect $f(\beta'x)\beta = \Lambda(1-\Lambda)\beta$.

Maximum Likelihood

Given our assumption that the ε are i.i.d., by the definition of independence, we can write the joint probability of observing $\{y_i\}_{i=1,\cdots,n}$ as

$$\Pr\{y_1, y_2, \cdots, y_n | X\} = \Pi_{y_i=0}[1 - F(\beta' x_i)] \times \Pi_{y_i=1}[F(\beta' x_i)].$$

Using the notational simplification

$F(\beta' x_i) = F_i, f(\beta' x_i) = f_i, f'(\beta' x_i) = f_i'$ we can write the likelihood function as:

$$L = \Pi_i (1 - F_i)^{1-y_i} (F_i)^{y_i}.$$

Since we are searching for a value of β that maximizes the probability of observing what we have, monotonically increasing transformations will not affect our maximization result. Hence we can take logs of the likelihood function; and since maximizing a sum is easier than maximizing a product, we take the log of the likelihood function:

$$\ln L = \sum_i (1 - y_i) \ln[1 - F_i] + y_i \ln F_i.$$

Now estimate $\hat{\beta}$ by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \ln L$$

Multinomial

- Suppose you are trying to understand the determinants of occupational choices among multiple sectors, or a travel model of urban commuters, or any model in which agents decide among multiple choices.
- We can generalize the binary choice model.

Multinomial Logit Model

- Assume the existence of $J+1$ choices.
- Each individual is characterized by x_i
- Consider:

$$\Pr(Y=j) = \frac{e^{\beta_j' x_i}}{1+\sum_{k=1}^{J} e^{\beta_k' x_i}} \text{ for } j = 1, \cdots, K$$

$$\Pr(Y=0) = \frac{1}{1+\sum_{k=1}^{J} e^{\beta_k' x_i}}$$

IIA

- Notice that

$$\ln\left[\frac{P_{ij}}{P_{i0}}\right] = \beta_j' x_i$$

  and likewise

$$\ln\left[\frac{P_{ij}}{P_{ik}}\right] = (\beta_j - \beta_k)' x_i$$

- Thus, the odds ratio does not depend on the other choices.

MLE

- The associated likelihood function is:

$$\ln L = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{ij} \Pr(Y_i = j)$$

  where $\Pr(Y_i = j)$ was defined before and d_{ij} is an indicator function that takes a value of 1 if alternative j is chosen.

- Thus, we need to solve:

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{N} (d_{ij} - \Pr(Y_i = j)) x_i \text{ for } j = 1, \cdots, K$$

MLE

- And we should interpret the results by studying:

$$\begin{aligned}\frac{\partial \Pr(Y=j)}{\partial x_i} &= \Pr(Y=j)[\beta_j - \sum_{k=0}^{J} \Pr(Y=k)\beta_k] \\ &= \Pr(Y=j)[\beta_j - \overline{\beta}]\end{aligned}$$

Multinomial Probit Model

- It is an appealing alternative to the previous models
- IIA does not hold
- Consider the following random utility model:

$$
\begin{aligned}
Y_1^* &= V_1 + \varepsilon_1 \\
Y_2^* &= V_2 + \varepsilon_2 \\
Y_3^* &= V_3 + \varepsilon_3
\end{aligned}
$$

where we assume that the residuals $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ have a trivariate normal distribution with mean zero and covariance matrix Σ:

$$
\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}
$$

Multinomial Probit Model

- Consider the probability that the first alternative will be chosen:

$$\Pr(Y_1^* > Y_2^*, Y_1^* > Y_3^*) = \Pr(\varepsilon_2 - \varepsilon_1 < V_1 - V_2, \varepsilon_3 - \varepsilon_1 < V_1 - V_3)$$

- Define $\eta_{21} = \varepsilon_2 - \varepsilon_1$, $\eta_{31} = \varepsilon_3 - \varepsilon_1$, $V_{12} = V_1 - V_2$, and $V_{13} = V_1 - V_3$, and

$$\begin{pmatrix} \eta_{21} \\ \eta_{31} \end{pmatrix} \sim N(0, \Sigma)$$

Multinomial Probit Model

- Thus, the probability that alternative 1 will be chosen is given by:

$$\Pr(Y_1^* > Y_2^*, Y_1^* > Y_3^*) = \int_{-\infty}^{V_{12}} \int_{-\infty}^{V_{13}} f(\eta_{21}, \eta_{31})\, \mathrm{d}\eta_{21} \mathrm{d}\eta_{31}$$

  where $f(\eta_{21}, \eta_{31})$ has a bivariate normal distribution.

- The other two probabilities can be similarly calculated.
- We can write down the likelihood and marginal effects using the previous cases.
- The presence of integrals represents a computational challenge.
- An alternative method used in the estimation of MNP is Monte Carlo methods (Lerman and Manski, 1982).

Limited Dependent

- The idea of this lecture is to show how by combining the linear regression equations and discrete choice analysis we can estimate general econometric models.
- These models can be linked to economic models.
- We use MLE to estimate these models.

Example

- Consider the following structure:

$$
\begin{aligned}
Y_0 &= \mu_0(X_0, \varepsilon_0) \\
Y_1 &= \mu_1(X_1, \varepsilon_1) \\
D &= 1[Y_1 - Y_0 - C(Z, \nu) \geq 0]
\end{aligned}
$$

- This is the Roy Model (Roy, 1956)
- This model can be used to characterize a variety of economic phenomena

The Tobit Model

- The Tobit model is a censored regression model:

$$Y_i = \beta' X_i + \varepsilon_i \text{ if } RHS > 0$$
$$Y_i = 0 \text{ otherwise}$$

  where ε_i are iid and $\varepsilon \sim N(0, \sigma^2)$

- We can estimate the model using the ML principle

Likelihood

- Assume the existence of N observations for (Y_i, X_i)
- Denote by N_1 the sample of individuals for which $Y_i > 0$, and N_0 the number for those with $Y_i = 0$
- Define

$$F_i = \int_{-\infty}^{\beta' X_i} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-t^2/2\sigma^2\right) \, \mathrm{d}t$$

$$f_i = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(\beta' X_i)^2/2\sigma^2\right)$$

- Thus, the likelihood function is:

$$L = \prod_{i=0} \Pr(Y_i = 0) \prod_{i=1} \Pr(Y_i > 0) f(Y_i | Y_i > 0)$$

- Notice that

$$\begin{aligned} \Pr(Y_i = 0) &= \Pr\left(\varepsilon_i > -\beta' X_i\right) \\ &= 1 - F_i \\ \Pr(Y_i > 0) f(Y_i | Y_i > 0) &= F_i \frac{f_\varepsilon(Y_i - \beta' X_i)}{F_i} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(Y_i - \beta' X_i)^2 / 2\sigma^2\right) \end{aligned}$$

- Thus,

$$L = \prod_{i=0}(1 - F_i) \prod_{i=1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-0.5\left(\frac{Y_i - \beta' X_i}{\sigma}\right)^2\right]$$

or

$$L = \sum_{i=0} \ln(1 - F_i) + \sum_{i=1} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}\left[0.5\left(\frac{Y_i - \beta' X_i}{\sigma}\right)^2\right]$$

- In this case, it is possible to show that the FOC are necessary and sufficient.
- No matter what the starting value, the solution will be the maximum.
- The asymptotic variance covariance matrix for the parameters can be estimated using the asymptotic properties of the MLE.

The Tobit Model: TS Procedure

- A two-stage procedure can also be used when estimating a Tobit model (Heckman, 1976).
- Because the likelihood function for the probit is well-behaved, we can define:

$$I_i = 1 \text{ if } Y_i > 0$$
$$I_i = 0 \text{ if otherwise}$$

  so we can estimate β/σ.

- With $\widehat{\beta/\sigma}$, we can construct $\phi(\widehat{\beta/\sigma}' X_i)$ and $\Phi(\widehat{\beta/\sigma}' X_i)$.
- On the other hand, we can write:

$$\begin{aligned} E(Y_i|Y_i > 0) &= \beta' X_i + E(\varepsilon_i|Y_i > 0) \\ &= \beta' X_i + E(\varepsilon_i|\varepsilon_i > -\beta' X_i) \end{aligned}$$

- Property of normal random variable (truncated mean):

$$E[\varepsilon|\varepsilon > c_i] = \sigma \frac{\phi\left(\frac{c_i}{\sigma}\right)}{1 - \Phi\left(\frac{c_i}{\sigma}\right)}$$

$$E[\varepsilon|\varepsilon \leq c_i] = -\sigma \frac{\phi\left(\frac{c_i}{\sigma}\right)}{\Phi\left(\frac{c_i}{\sigma}\right)}$$

so, in the context of our problem:

$$\begin{aligned} E(Y_i|Y_i > 0) &= \beta' X_i + E(\varepsilon|\varepsilon > -\beta' X_i) \\ &= \beta' X_i + \sigma \frac{\phi\left(\frac{-\beta' X_i}{\sigma}\right)}{1 - \Phi\left(\frac{-\beta' X_i}{\sigma}\right)} \end{aligned}$$

- We can use OLS to estimate β by using $\phi\left(\widehat{\beta/\sigma}'X_i\right)$ and $\Phi\left(\widehat{\beta/\sigma}'X_i\right)$ in:

$$Y_i = \beta' X_i + \sigma \frac{\phi\left(\widehat{\beta/\sigma}'X_i\right)}{1 - \Phi\left(\widehat{\beta/\sigma}'X_i\right)} + v_i$$

  where v_i is orthogonal to the controls in the regression.

- It is possible to show that (v_i) depends on i, so the error term in this regression is heteroskedastic.

- The two-limit tobit model can be analyzed in a similar way.

Truncated Regression Models

- There are many situation in which the appropriate model is a truncated regression model (e.g. earning regressions from data for the the negative-tax experiment).
- Consider the following model:

$$Y_i = \beta' X_i + \varepsilon_i$$

  where ε_i are iid and $\varepsilon \sim N(0, \sigma^2)$.
- Individuals are selected at random, but those with Y_i higher than L_i (truncation point) are eliminated from the sample.
- The estimation of the regression model by OLS from the truncated sample leads to biased estimates.

Likelihood

- Notice that the density function is the truncated normal density:

$$
\begin{aligned}
f(Y_i) &= \frac{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-(Y_i-\beta' X_i)^2/2\sigma^2\right)}{\Pr(Y_i \le L_i)} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-(Y_i-\beta' X_i)^2/2\sigma^2\right)}{\int_{-\infty}^{L_i-\beta' X_i}\frac{1}{\sigma\sqrt{2\pi}}\exp(-t^2/2\sigma^2)\,\mathrm{d}t} \text{ if } Y_i \le L_i \\
f(Y_i) &= 0 \text{ otherwise}
\end{aligned}
$$

- The log likelihood function is:

$$
\ln L = -n\ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2}\sum_i\left(\frac{Y_i-\beta' X_i}{\sigma}\right) - \sum_i \ln\Phi\left(\frac{L_i-\beta' X_i}{\sigma}\right)
$$

- The likelihood function in this model can also be shown to be globally concave.
- The two-stage method available for the censored regression model, is not feasible in the truncated regression model, because no observations are available for $Y_i > L_i$.