16: Introduction to Nonlinear Models
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

Note: Sections marked "(optional)" were not covered in lecture. You do not need to know this material for the exam, though the material represents applications of ideas and tools you do need to know.

The regression models we have considered to this point have all been linear in parameters. If the model's parameters are identified, linearity means that we can solve for the parameters in terms of moments of the DGP. Therefore, it also means that we can calculate estimates of these parameters directly.

Today we'll consider models that are nonlinear in parameters. In general, we can only find implicit solutions to nonlinear models. That is, the parameters can only be written as the solution to some optimization problem. As a result, our estimators of these parameters are generally calculated through numerical optimization rather than directly. Such estimators are called optimization estimators or M-estimators.

There are three primary estimation frameworks for the nonlinear case: nonlinear least squares (NLLS), maximum likelihood, and the method of moments. We'll discuss the first two today, and consider method of moments estimators next day.

Examples: Qualitative and Limited Dependent Variable Models

Now we'll consider some of the most commonly encountered nonlinear models. The first ones we'll discuss are typically applied to problems of binary choice, though extensions exist to cover more general discrete choice situations. The binary choice problem can be described as follows. An individual i is choosing between two alternatives, for example attending college or not. We write $y_i = 1$ if i chooses to attend college, and $y_i = 0$ if not.

The Linear Probability Model

The simplest starting point is to model the choice using choice probabilities that are linear in parameters. This gives rise to the **linear probability model**:

$$\begin{aligned} \Pr(y_i = 1|\mathbf{x}_i) &= \mathbf{x}_i'\beta \\ \Pr(y_i = 0|\mathbf{x}_i) &= 1 - \mathbf{x}_i'\beta \end{aligned}$$

which implies

$$\begin{aligned} E[y_i|\mathbf{x}_i] &= \mathbf{x}_i'\beta \\ y_i &= E[y_i|\mathbf{x}_i] + (y_i - E[y_i|\mathbf{x}_i]) \\ &= \mathbf{x}_i'\beta + \varepsilon_i. \end{aligned}$$

Notice that $E[y_i|\mathbf{x}_i] = \mathbf{x}_i\beta$ implies that

$$\varepsilon_i = \begin{cases} -\mathbf{x}_i'\beta & \text{if } y_i = 0 \\ 1 - \mathbf{x}_i'\beta & \text{if } y_i = 1 \end{cases}.$$

which means that:

$$var(\varepsilon_i|\mathbf{x}_i) =$$

In other words, OLS will provide consistent/unbiased estimators for the linear probability model (since $E(y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta$). But because the errors are necessarily heteroskedastic (unless $\beta = 0$) OLS is inefficient and the usual OLS standard errors will be wrong. Using GLS will fix both problems.

However there is a more fundamental problem with the linear probability model: it tends to generate probabilities greater than one or less than zero.

The Probit and Logit Models

A frequently encountered alternative to the linear probability model is the **probit model**. It assumes

$$\Pr(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i'\beta)$$

where Φ is the standard normal CDF. A common alternative is the **logit model**, or **logistic regression** which assumes instead that

$$\Pr(y_i = 1|\mathbf{x}_i) = \Lambda(\mathbf{x}_i'\beta) = \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}}.$$

The function $\Lambda(\mathbf{x}_i'\beta)$ is the CDF of the logistic distribution. The logistic distribution looks just like the normal, i.e., its PDF is symmetric, unimodal around zero, strictly positive everywhere, and "bell-shaped."

As CDFs, both the probit and logit transformations are nondecreasing (in fact, strictly increasing) functions that map real numbers to (the interior of) the $[0, 1]$ interval. We could use other transformations, though we rarely do.

We can estimate either model by maximum likelihood, which we know is asymptotically efficient. We usually write the likelihood for the probit model as

$$L(\beta) = \prod_{i=1}^{n} [\Phi(\mathbf{x}_i'\beta)]^{y_i} [1 - \Phi(\mathbf{x}_i'\beta)]^{(1-y_i)}$$

since this implies a convenient expression for the log-likelihood:

$$l(\beta) = \sum_{i=1}^{n} [y_i \ln \Phi(\mathbf{x}_i'\beta) + (1 - y_i) \ln(1 - \Phi(\mathbf{x}_i'\beta))].$$

The log-likelihood function of the logit is the same, but with Λ substituted for Φ.

Maximum Likelihood Estimation

There is no convenient closed form solution for the MLE of either the probit or logit model. Instead, we use iterative methods to find the maximum of the log-likelihood function. This is particularly easy to do for these models, since the probit and logit log-likelihoods are both strictly concave.

Marginal Effects

In linear models, the regression coefficients β measure the marginal effect of $\mathbf{x}_i$ on $E[y_i|\mathbf{x}_i]$. That is, $dE[y_i|\mathbf{x}_i]/d\mathbf{x}_i = \beta$. This is typically not the case for nonlinear models. In fact, for the probit model we see that

$$\frac{dE[y_i|\mathbf{x}_i]}{d\mathbf{x}_i} = \frac{d}{d\mathbf{x}_i}\Phi(\mathbf{x}_i'\beta) = \phi(\mathbf{x}_i'\beta)\beta$$

which depends on the data $\mathbf{x}_i$. Thus, there is no unique marginal effect of $\mathbf{x}_i$ that applies to every observation.

In practice, applied researchers will report "average marginal effects": either the marginal effect evaluated at the sample mean of the data[1] , $\phi\left(\bar{\mathbf{x}}'\hat{\beta}\right)\hat{\beta}$, or the sample mean of the marginal effects $\frac{1}{n}\sum_{i=1}^{n}\phi\left(\mathbf{x}_i'\hat{\beta}\right)\hat{\beta}$. An alternative is to report marginal effects for particular data values of interest.

In the logit model,

$$\frac{dE[y_i|\mathbf{x}_i]}{d\mathbf{x}_i} = \frac{d}{d\mathbf{x}_i}\Lambda(\mathbf{x}_i'\beta) = \frac{d}{d\mathbf{x}_i}\frac{e^{\mathbf{x}_i'\beta}}{1+e^{\mathbf{x}_i'\beta}} = \frac{e^{\mathbf{x}_i'\beta}}{\left(1+e^{\mathbf{x}_i'\beta}\right)^2} = \Lambda(\mathbf{x}_i'\beta)\left[1-\Lambda(\mathbf{x}_i'\beta)\right].$$

Inference

Since we estimate the probit and logit models by maximum likelihood, all our asymptotic distribution theory for MLEs applies. That is, the MLEs are consistent, asymptotically normal, invariant, and asymptotically efficient. As always, the asymptotic covariance of the MLE is obtained from the inverse information matrix.

We can obtain asymptotic standard errors of the marginal effects using the delta method. This is also true for predicted probabilities: $\Phi\left(\mathbf{x}_i'\hat{\beta}\right)$ in the probit model, and $\Lambda\left(\mathbf{x}_i'\hat{\beta}\right)$ in the logit. As with any MLE, we can test hypotheses using LRT, LM, or Wald tests.

The Probit and Logit as Latent Variable Models (optional)

Sometimes the probit and logit models are formulated as **latent variable** models. In this setting, the underlying choice depends on an unobserved (latent) variable y_i^*, such that

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$$

[1] This is the default of Stata's `dprobit` and `dlogit` commands, and so is the most common choice.

with $\varepsilon_i \sim N(0,1)$ in the probit case, or $\varepsilon_i \sim \Lambda(0, \pi^2/3)$ in the logit case. For example, y_i^* might represent the (net) utility of choosing some action. We don't observe utility, but only the choice that was made. That is, we observe y_i, where

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

Then in the case of the probit model we have,

$$\begin{aligned} \Pr[y_i = 1|\mathbf{x}_i] &= \Pr[y_i^* > 0|\mathbf{x}_i] = \Pr[\mathbf{x}_i'\beta + \varepsilon_i > 0|\mathbf{x}_i] = \Pr[\varepsilon_i > -\mathbf{x}_i'\beta|\mathbf{x}_i] \\ &= \Pr[\varepsilon_i \leq \mathbf{x}_i'\beta|\mathbf{x}_i] \qquad \text{(symmetry)} \\ &= \Phi(\mathbf{x}_i'\beta) \end{aligned} \tag{1}$$

just as before. The derivation of the logit model is identical.

Nonlinear regression

The **nonlinear regression model** is

$$y_i = f(\mathbf{x}_i, \beta) + \varepsilon_i$$

where β and $\mathbf{x}_i$ are $k \times 1$, and $f : \mathbb{R}^k \to \mathbb{R}$. We will assume $E[\varepsilon_i|\mathbf{x}_i] = 0$, Thus $E[y_i|\mathbf{x}_i] = f(\mathbf{x}_i, \beta)$ is the regression function. The linear regression functions we have considered thus far are a special case. An example of a regression function nonlinear in parameters is

$$y_i = \beta_1 + \beta_2 e^{\beta_3 x_i} + \varepsilon_i$$

which might arise as the solution to a differential equation.

Another example is the probit or logit model. We could estimate this model using NLLS, since it implies $E[y_i|\mathbf{x}_i] = \Phi(\mathbf{x}_i'\beta)$ and hence

$$y_i = \Phi(\mathbf{x}_i'\beta) + \varepsilon_i.$$

Nonlinear Least Squares

The **nonlinear least squares** (NLLS) estimator minimizes

$$S(\beta) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \beta))^2 = (\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta))'(\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta)).$$

Define the matrix of first derivatives (the gradient matrix)

$$\mathbf{G}(\beta) = \begin{bmatrix} \frac{\partial f(\mathbf{x}_1,\beta)}{\partial \beta_1} & \frac{\partial f(\mathbf{x}_1,\beta)}{\partial \beta_2} & \cdots & \frac{\partial f(\mathbf{x}_1,\beta)}{\partial \beta_k} \\ \frac{\partial f(\mathbf{x}_2,\beta)}{\partial \beta_1} & \frac{\partial f(\mathbf{x}_2,\beta)}{\partial \beta_2} & \cdots & \frac{\partial f(\mathbf{x}_2,\beta)}{\partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{x}_n,\beta)}{\partial \beta_1} & \frac{\partial f(\mathbf{x}_n,\beta)}{\partial \beta_2} & \cdots & \frac{\partial f(\mathbf{x}_n,\beta)}{\partial \beta_k} \end{bmatrix} = \frac{\partial \mathbf{f}(\mathbf{X}, \beta)}{\partial \beta'}$$

which is $n \times k$.

The first derivative of the sum of squares function is

$$\frac{\partial S(\beta)}{\partial \beta'} = \frac{\partial}{\partial \beta'} (\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta))' (\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta)) = -2 (\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta))' \mathbf{G}(\beta).$$

Therefore the NLLS estimator $\hat{\beta}$ satisfies the FOC:

$$\mathbf{G}\left(\hat{\beta}\right)' \left(\mathbf{y} - \mathbf{f}\left(\mathbf{X}, \hat{\beta}\right)\right) = \mathbf{0}.$$

Notice the similarity between this and the least squares FOC $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

Computing the NLLS Estimator (optional)

Since the FOC that the NLLS estimator solves is not linear in β, we can't find a closed form solution. We could use Newton's method, which will be described below, but it is easier to use a first-order Taylor expansion of the regression function itself. This is called the **Gauss-Newton** method.

Suppose we have a candidate solution β_s. If we expand $\mathbf{f}(\mathbf{X}, \beta)$ around β_s we get

$$\mathbf{f}(\mathbf{X}, \beta) \approx \mathbf{f}(\mathbf{X}, \beta_s) + \mathbf{G}(\beta_s)(\beta - \beta_s).$$

The Gauss-Newton method minimizes the linearized sum of squares function:

$$S_0(\beta) = [\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_s) - \mathbf{G}(\beta_s)(\beta - \beta_s)]' [\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_s) - \mathbf{G}(\beta_s)(\beta - \beta_s)]$$

by choice of β. Denoting the minimand by β_{s+1}, the first order condition is

$$-2\mathbf{G}(\beta_s)' [\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_s) - \mathbf{G}(\beta_s)(\beta_{s+1} - \beta_s)] = \mathbf{0}$$

which is linear in β_{s+1}. Solving the FOC for β_{s+1} gives

$$\beta_{s+1} = \beta_s + \left[\mathbf{G}(\beta_s)' \mathbf{G}(\beta_s)\right]^{-1} \mathbf{G}(\beta_s)' (\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_s)). \tag{2}$$

(Compare this to the linear case). This is a convenient updating equation that we can use to compute the NLLS estimator. Upon convergence, the FOC is satisfied and the solution is the NLLS estimator, $\hat{\beta}$. Unfortunately, there is no guarantee this method will actually converge. A variety of modifications to the Gauss-Newton method have been suggested to help in situations where it doesn't converge (see Greene).

Inference (optional)

Under some conditions, we can show the NLLS estimator to be consistent and asymptotically normal. Let β_0 denote the true parameter values. Then if

$$\operatorname{plim}\left(\frac{\mathbf{G}(\beta_0)' \mathbf{G}(\beta_0)}{n}\right) = \mathbf{Q} \text{ positive definite and finite} \tag{3}$$

$$\operatorname{plim}\left(\frac{\mathbf{G}(\beta_0)' \varepsilon}{n}\right) = \mathbf{0} \tag{4}$$

and if [1] the parameter space containing β is compact; [2] the sum of squares function S has a continuous and differentiable probability limit q; and [3] q has a unique minimum at β_0, then the NLLS estimator is consistent. There is a sketch of a proof in Greene; Amemiya (1985) has a formal proof. In fact, you can show

$$\hat{\beta} = \beta_0 + \left[\mathbf{G}\left(\beta_0\right)'\mathbf{G}\left(\beta_0\right)\right]^{-1}\mathbf{G}\left(\beta_0\right)'\varepsilon + o_p\left(n^{-1/2}\right). \tag{5}$$

[Recall that if $a_n = o_p\left(c_n\right)$, then $\left(a_n/c_n\right) \xrightarrow{p} 0$.] This is just like what we had in the linear model – a decomposition of the estimator into "truth" plus sampling error. Coupled with assumptions (3) and (4), equation (5) implies plim $\hat{\beta} = \beta_0$.

Rewrite (5) as:

$$\sqrt{n}\left(\hat{\beta} - \beta^0\right) = \left[\frac{\mathbf{G}\left(\beta_0\right)'\mathbf{G}\left(\beta^0\right)}{n}\right]^{-1}\frac{\mathbf{G}\left(\beta^0\right)'\varepsilon}{\sqrt{n}} + \sqrt{n}o_p\left(n^{-1/2}\right).$$

The remainder term, $\sqrt{n}o_p\left(n^{-1/2}\right)$, converges in probability to zero. And the term in square brackets converges in probability to $\mathbf{Q}^{-1}$ by assumption. Assume

$$Var\left[\varepsilon|\mathbf{x}_i\right] = \sigma^2\mathbf{I}_n$$

As usual, we can invoke a CLT:

$$\frac{\mathbf{G}\left(\beta_0\right)'\varepsilon}{\sqrt{n}} \xrightarrow{d} N\left(\mathbf{0}, \sigma^2\mathbf{Q}\right)$$

and hence

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}\right) \tag{6}$$

We can use (6) as the basis for inference about β. Of course we need to estimate the elements of the variance, but there are obvious ways to do so. In particular,

$$s^2 = \frac{\left(\mathbf{y} - \mathbf{f}\left(\mathbf{X}, \hat{\beta}\right)\right)'\left(\mathbf{y} - \mathbf{f}\left(\mathbf{X}, \hat{\beta}\right)\right)}{n-k}, \quad \hat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{f}\left(\mathbf{X}, \hat{\beta}\right)\right)'\left(\mathbf{y} - \mathbf{f}\left(\mathbf{X}, \hat{\beta}\right)\right)}{n}$$

are both consistent estimators of σ^2 and

$$\hat{\mathbf{Q}} = \frac{\mathbf{G}(\hat{\beta})'\mathbf{G}(\hat{\beta})}{n}$$

is a consistent estimator of $\mathbf{Q}$. When interest centers on functions of β (e.g., elasticities), we can use the delta method for inference.

Numerical optimization

Suppose we wish to use a computer to find the maximum[2] of a function $f : \Theta \to R$, where $\Theta \subset R^k$:

$$\theta^* = \arg\ \max_{\theta} f(\theta)$$

How would we go about doing this? First, we need to impose some assumptions on the problem:

[2] If we instead want to minimize f. we can just maximize $-f$.

1. f has all the derivatives we might need.
2. Θ is closed and bounded. In combination with the existence of the first derivative for f this implies that $f(\theta)$ has at least one maximum.
3. θ^* is unique, and in the interior of Θ. This will allow us to impose the first order conditions.

When there is no closed form solution to the optimization problem, we must use an iterative search algorithm. That is, we construct a sequence $\theta_s, s = 0, 1, 2, \ldots, S$ that hopefully reaches an approximate solution to the problem. The typical algorithm has 3 parts:

1. A rule for constructing an initial guess θ_0.
2. An updating rule that gives θ_{s+1} as a function of $(\theta_0, \ldots, \theta_s)$.
3. A termination rule. Once the termination conditions have been met, we stop updating and pick the θ_i that came closest to solving the problem. Typical termination conditions include:
   - Parameter convergence: $\theta_{s+1} \approx \theta_s$.
   - Function convergence: $f'(\theta_{s+1}) \approx 0$.
   - Maximum iterations reached: Usually there is some hard limit on the number of iterations. This prevents your program from getting caught in an infinite loop (this can happen if you pick a bad starting value).

There is always a tradeoff between accuracy and computational cost. The cost of a given algorithm is roughly proportional to the number of times that f or its derivatives need to be calculated.

Finding a local optimum: Newton's method and its relatives

Newton's method is based on the idea of constructing a linear approximation to the first order conditions, and then solving that linear approximation. Suppose we have a θ_s. Then we can take a Taylor series approximation:

$$f'(\theta_{s+1}) \approx f'(\theta_s) + f''(\theta_s)(\theta_{s+1} - \theta_s)$$

where f'' is the Hessian of f. Since we would like for $f'(\theta_{s+1}) = 0$, we might try solving

$$f'(\theta_s) + f''(\theta_s)(\theta_{s+1} - \theta_s) = 0$$

for θ_{s+1}. Solving we get:

$$\theta_{s+1} = \theta_s - f''(\theta_s)^{-1} f'(\theta_s)$$

Notice that if we did find the actual maximum in step s, then $f'(\theta_s) = 0$ and $\theta_{s+1} = \theta_s$. More generally, if we are close to solving the FOC, then the steps will tend to be small. In addition we will tend to take smaller steps if the Hessian is big (i.e. the function has a lot of curvature).

A few additional considerations:

- Usually we will be able to find the first and second derivatives of f. In that case, each step requires us to evaluate the k functions in $f'(\theta)$ and the $(k^2+k)/2$ functions in $f''(\theta)$.
  - For sufficiently big k, the Hessian is much more costly to calculate than the gradient, and needs to be inverted (another costly action when k is big).
  - The Hessian will be negative definite and invertible at the maximum and sufficiently near the maximum. But further away, it may not be invertible (so the next Newton step doesn't exist), or it may not be negative definite (so the next Newton step might be in the direction of the minimum).
  - As a result there are a number of quasi-Newton methods that take the form:

    $$\theta_{s+1} = \theta_s - H_s f'(\theta_s)$$

    where H_s is a matrix constructed to approximate the inverse Hessian of f. The most common quasi-Newton method in econometrics is the BFGS (Broyden, Fletcher, Goldfarb, and Shanno) algorithm.
  - It is also common to combine quasi-Newton with line search.
- Sometimes the function f will be sufficiently complicated that it is difficult to solve analytically for its derivative and Hessian. In that case, we can construct an approximate derivative using the method of finite differences. Recall that:

  $$f'(\theta) = \lim_{\epsilon \to 0} \frac{f(\theta+\epsilon) - f(\theta)}{\epsilon}$$

  This suggests that if we pick a very small $e > 0$ that:

  $$f'(\theta) \approx \frac{f(\theta+e) - f(\theta)}{e}$$

  The finite difference approximation to the Hessian is similar. The tricky part is figuring out the right value of e. The approximation error is increasing in e, while the computer's rounding error is decreasing in e.

Quasi-Newton methods are a very fast way of (approximately) solving the first order conditions, and thus of finding a local optimum. However, you should note that they is not guaranteed to find the global optimum, unless the global optimum happens to be the unique local optimum as well.

Finding a global optimum

Usually we are interested in finding a global optimum, not a local optimum. So are there any algorithms that are designed to find the global optimum?

1. Brute force search: This approach simply tries out a very large number of candidate θ's, covering the whole range of Θ. The θ's can be generated randomly, or can be generated by dividing Θ into a fine grid and selecting one value in each grid. This is robust but very slow.

2. Newton with restarts: This approach combines brute force with quasi-Newton. Run a quasi-Newton search n times, each time with a different starting value (usually randomly drawn from Θ).

3. Simulated annealing. This is a global optimization method that works even when the objective function isn't differentiable. Each step involves selecting a new candidate θ^c at random in a neighborhood of θ_s, and accepting it (i.e., setting $\theta_{s+1} = \theta^c$) with some probability that is increasing in $f(\theta^c) - f(\theta_s)$ but is always nonzero. As the algorithm proceeds we gradually reduce a "temperature" (simulated annealing is actually based on physical models of annealing, which is a technique for strengthening metals and glass) parameter that controls the step size and the probability of accepting downward steps. Simiulated annealing is much slower than quasi-Newton, and so is only advised with very ill-behaved objective functions.

Appendix: The Tobit Model

I will skip the Tobit model in lecture, but I've left the discussion in the notes so you can see one more example of a nonlinear model.

The Tobit model is another example of a latent variable model. In this case, the latent variable y_i^* is **partially observed**, or **censored**. That is,

$$y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \sigma^2)$. Therefore $y_i^* \sim N(\mathbf{x}_i'\beta, \sigma^2)$ and if we observed y_i^* we could just estimate β and σ^2 by least squares. However, we do not observe y_i^*. Rather we observe y_i, where

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

A classic application of the Tobit model is to model labour supply (hours of work): it is either zero or positive. Expenditures on large items (automobiles, vacations, housing, etc.) are also frequently modeled using a Tobit. In each of these cases, a linear regression model is inappropriate because we observe "too many" outcomes of zero. Ignoring the zeros and estimating a linear regression model on the uncensored observations ($y_i^* > 0$) is also inappropriate, since $E[y_i^* | y_i^* > 0, \mathbf{x}_i] \neq \mathbf{x}_i'\beta$. In fact, we have the following result for the first moment of a truncated normal distribution.

Proposition 1 *Let $z \sim N(\mu, \sigma^2)$. Then*

$$E[z|z > a] = \mu + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$$

*where $\alpha = (a - \mu)/\sigma$. We usually call $\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$ the **inverse Mills ratio**.*

This result demonstrates that a simple linear regression of y_i on $\mathbf{x}_i$ in the uncensored sample won't consistently estimate β or σ^2, since

$$E[y_i | y_i > 0, \mathbf{x}_i] = E[y_i^* | y_i^* > 0, \mathbf{x}_i] = \mathbf{x}_i'\beta + \sigma \frac{\phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right)}.$$

That is, the conditional mean is a nonlinear function of β and σ^2. However, we could estimate this regression function by NLLS or maximum likelihood. This formulation is called the **truncated regression model**. It is appropriate if we have observations only on those i such that $y_i > 0$.

The Tobit model is a little different. Notice the inefficiency in the truncated regression approach: it is based only observations with $y_i > 0$ and ignores information contained in the remaining observations (i.e., the observations with $y_i = 0$). The Tobit model incorporates this additional information. Let $d_i = 1$ if $y_i > 0$ and $d_i = 0$ otherwise. Given $\varepsilon_i \sim N(0, \sigma^2)$, the likelihood function is

$$L\left(\beta, \sigma^2\right) = \prod_{i=1}^{n} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\beta}{\sigma}\right)\right]^{d_i} \left[\Phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right)\right]^{(1-d_i)}$$

Why? For the uncensored observations ($d_i = 1$), the density of y_i is the same as the density of y_i^* (since $y_i = y_i^*$), which is just $(1/\sigma)\,\phi\left(\left(y_i - \mathbf{x}_i'\beta\right)/\sigma\right)$ like in the linear regression under normality. For the censored observations ($d_i = 0$), the density of y_i is

$$\Pr\left[y_i = 0|\mathbf{x}_i\right] = \Pr\left[y_i^* < 0|\mathbf{x}_i\right] = \Pr\left[\mathbf{x}_i'\beta + \varepsilon_i < 0|\mathbf{x}_i\right] = \Pr\left[\varepsilon_i < -\mathbf{x}_i'\beta|\mathbf{x}_i\right] = \Phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right).$$

The conditional expectation of y_i in the Tobit model is

$$\begin{aligned} E\left[y_i|\mathbf{x}_i\right] &= E\left[y_i|y_i = 0, \mathbf{x}_i\right]\Pr\left[y_i = 0|\mathbf{x}_i\right] + E\left[y_i|y_i > 0, \mathbf{x}_i\right]\Pr\left[y_i > 0|\mathbf{x}_i\right] \\ &= \left(\mathbf{x}_i'\beta + \sigma\frac{\phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-\mathbf{x}_i'\beta}{\sigma}\right)}\right)\Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right) \end{aligned}$$

which remains nonlinear in β and σ^2. Estimation is almost always by maximum likelihood, though NLLS is possible. If estimated by maximum likelihood, our standard asymptotic distribution theory applies and we know how to do inference.

There are several marginal effects of interest in this model. One is measured by the slope coefficients, β :

$$\frac{d}{d\mathbf{x}_i}E\left[y_i^*|\mathbf{x}_i\right] = \beta.$$

This measures the marginal effect of $\mathbf{x}_i$ on the latent variable y_i^*. Since we don't observe y_i^*, this is not usually the object of interest. Typically, we are interested in

$$\frac{d}{d\mathbf{x}_i}E\left[y_i|\mathbf{x}_i\right] = \beta\Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)$$

although showing this equality requires a bit of work. Greene gives a proof.

The Tobit model depends strongly on the normality assumption and is sensitive to heteroscedasticity. Robust alternatives have been suggested – see Greene.