Johns Hopkins University
Economics 633
Econometrics
Censored and Truncated Regressions

Adrian Pagan

April 15, 2004

1 The Problem

Some data used in economic analysis has the characteristic that we do not observe values above or below a certain magnitude, owing to the operation of a *censoring* or truncation mechanism. Thus it may be that the central bank intervenes to stop an exchange rate falling below or going above certain values (floors and ceilings); dividends paid by a company may remain zero until earnings pass some threshold value; commodity stabilization funds may set reserve prices for commodities. In all these situations the observed data consists of a combination of measurements of some underlying latent variable and observations that pertain when the censoring mechanism is applied. Suppose we think of e_t as the deviation of the exchange rate from its equilibrium value and that the floor and ceiling imposed on this are e^- and e^+ respectively. If there were no bounds e_i would be e_i^*, and we would therefore observe $e_i = e_i^*$ if $e^- < e_i^* < e^+$ and $e_i = e^-$ if $e_i^* \leq e^-$, $e_i = e^+$ if $e_i^* \geq e^+$. Whilst the e_i^* might be thought of as a continuous random variable, the *observed data* e_i cannot be because it is equal to a number of values such as e^+ and e^- and so is *censored* at these points. Cesnoring can be *right censoring* (at an upper limit) or *left censoring* (at a lower limit). The complete sample of observations is available to us but the latent variables corresponding to the e^+ and e^- are not. Nevertheless we typically do have data at those points on what is happening in the economy and so on what affects the exchange rate. If for some reason though no data was published when these bounds were attained then we would be working with a *truncated* sample. Truncated samples occur a lot in micro work e.g. if one chooses to work with observations on families below the poverty line when estimating a relationship thought to hold for all household units. Since the analysis for both problems is

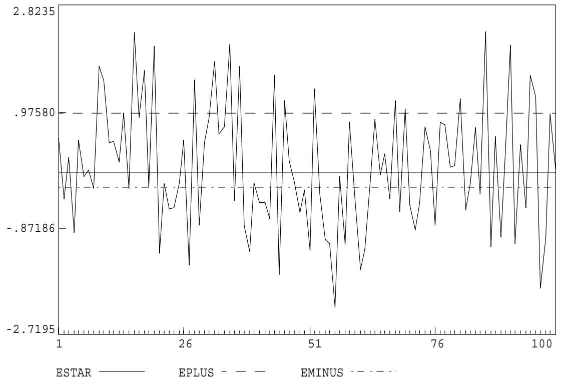

Figure 1:

very similar we will work through the censored case, pointing out the differences to the truncated case. Greene has fuller details.

The fact that e_i is not continuous creates problems when we come to analyze the determinants of exchange rates. It is likely that we can set up a model connecting e_i^* with some variables x_i, i.e., $e_i^* = x_i'\beta + u_i^*$, where u_i^* is a continuous random variable with $E(u_i^*) = 0$, but we do not observe e_i^*, only e_i. We are tempted therefore to regress e_i against x_i, but the error term in such a regression will generally have a non-zero mean, since it is $u_i = e_i - x_i'\beta = e_i - e_i^* + e_i^* - x_i'\beta = (e_i - e_i^*) + u_i^*$ and $E(u_i) = E[(e_i - e_i^*) + u_i^*] = E(e_i - e_i^*)$. Let's take a simple case where $x_i = 1$. Then $\beta = E(e_i^*)$. Suppose that $E(e_i^*) = 0$, *i.e.*, on average, without intervention, the exchange rate change is zero (a pretty good description of actual exchange rate movements). Then the graph below shows the situation where we have generated 100 observations on e_i^* (assuming the density of e_i^* is $\mathcal{N}(0,1)$) and have set $e^+ = 1$ and $e^- = -.25$.

The observed data points, the e_i, are those within the bounds. Obviously e_i's density will have a mean that is not zero since there are more positive observations on e_i^* than negative ones, since the censoring point e^+ is larger in absolute value than $e^-.i.e.$ if we use the points that lie between the bounds (all that we have) to compute the sample mean it will always lie well away from the true value of zero that would be found (in large samples) if observations on the latent variable were available. When e_i^* is symmetrically distributed the sample mean of the e_i will not be zero unless the censoring is symmetric, i.e., e^+ and e^- are the same distance from the origin. More generally, a regression of e_i^* against x_i will generally mean that OLS is an inconsistent estimator of β.

We wish to do a more formal analysis of the impact of censoring of data upon the OLS estimator and how one can still estimate the parameters consistently. Historically, the problem was first investigated by Tobin, who noticed that expenditure on certain consumption items in household budgets could be zero and therefore proposed what has been referred to as the **TOBIT** model.

$$y_i^* = x_i'\beta + u_i^* \tag{1}$$

where y_i^* is a latent variable and $u_i^* \sim n.i.d.(0, \sigma^2)$. The observations are $y_i = 1(y_i^* > 0)y_i^*$ where $1(\cdot)$ is the indicator function taking the value unity if $y_i^* > 0$ and zero otherwise. Thus $y_i = y_i^*$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$.

2 Analysis of the OLS Estimator

Let us assume that there is a set of observations available on y_i, $i = 1, ..., N$, and these are either zero or positive. Let the N_0 zero values have indices $i \in I_0$ and the N_p positive ones $i \in I_p$, where I_0 and I_p are sets of indices. For example $y_i = \{0, 33.4, 0, 2.5, 5.6\}$ would give $I_0 = \{1, 3\}$, $I_p = \{2, 4, 5\}$.

Consider what would happen if we regressed y_i against $x_i (i \in I_p)$. This would correspond to a *truncated* sample since it is assumed that we only have data if y_i turns out to be positive. The model that is estimated would then be

$$y_i = x_i'\beta + v_i \quad i \in I_p \tag{2}$$

and we want to know what the density of the errors v_i associated with positive observations y_i is. To answer this question we see that $v_i = u_i^*$ only for $i \in I_p$ and $i \in I_p$ means $y_i^* > 0$ or $x_i'\beta + u_i^* > 0$ or $u_i^* > -x_i'\beta$. Hence the range of values of v_i we observe is from $-x_i'\beta$ to ∞ only and the density of v_i will be that of u_i^* over the range of values $u_i^* > -x_i'\beta$. Now if v_i is to have a proper density it must integrate to unity

$$\therefore \int_{-x_i'\beta}^{\infty} f_v(\lambda) d\lambda = 1$$

whereas

$$\begin{aligned} \int_{-x_i'\beta}^{\infty} f_{u^*}(\lambda) d\lambda &= F_i \\ &= \int_{-\infty}^{x_i'\beta} f_{u^*}(\lambda) d\lambda \end{aligned}$$

provided the density of u_i^* *is symmetric.*[1] In the Tobit model f_{u^*} is the $N(0, \sigma^2)$ density and so the above would be

[1] With the Tobit model the density is normal and so is symmetric. One normally sees F_i written as Φ_i, the conventional symbol for the cumulative normal distribution, but we keep the more general notation to emphasise that one could use other densities. If other non-symmetric densities are used F_i must be defined as the integral from $-x_i'\beta$ to ∞.

$$= \int_{-\infty}^{x_i'\beta} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}(\lambda/\sigma)^2\right) d\lambda.$$

Therefore we find the density of v_i, $i \in I_p$ as

$$pdf(v_i) \quad = \quad F_i^{-1}(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}(v_i/\sigma)^2\right) = h(v)$$

and the $pdf(v_i)$ is different to that of u_i^*. In particular

$$\begin{aligned} E(v_i) &= \int_{-x_i'\beta}^{\infty} \lambda \;\; h(\lambda) d\lambda \\ &= F_i^{-1} \int_{-x_i'\beta}^{\infty} \lambda f(\lambda) d\lambda \end{aligned}$$

where $f_{u^*}(u^*)$ is re-named as $f()$ for convenience. But $\int \lambda f(\lambda) d\lambda = -\sigma^2 f(\lambda)$ (reverse differentiation to check)

$$\begin{aligned} \therefore E(v_i) &= F_i^{-1}[-\sigma^2 f(\lambda)]_{-x_i'\beta}^{\infty} \qquad (3) \\ &= F_i^{-1}[\sigma^2 f(x_i'\beta)] \text{ as } f(\infty) = 0 \\ &= \sigma^2 F_i^{-1} f_i \end{aligned}$$

where $f_i = f(x_i'\beta)$. Now

$$= \sigma^2 F_i^{-1} f_i \neq 0 \text{ as } F_i > 0, f_i > 0.$$

and therefore, from (3), the expected value of v_i is non-zero and it also depends upon $x_i'\beta$. Consequently,

$$\hat{\beta} - \beta = \left(\sum x_i'x_i\right)^{-1} \sum x_i'v_i = \left(\sum x_i'x_i\right)^{-1} \sum x_i' \left(\sigma^2 F_i^{-1} f_i + \eta_i\right) \qquad (4)$$

where $\eta_i = v_i - E(v_i)$ has $E(\eta_i) = 0$. Therefore

$$\begin{aligned} E(\hat{\beta} - \beta | x_i'\beta) &= \left(\sum x_i'x_i\right)^{-1} \sum x_i' \left(\sigma^2 F_i^{-1} f_i + E(\eta_i | x_i'\beta)\right) \\ &= \left(\sum x_i'x_i\right)^{-1} \sum x_i'\sigma^2 F_i^{-1} f_i \end{aligned}$$

and so the OLS estimator would be biassed. This is simply because the conditional expectation is not linear in x_i and the non-linearity shows up as a biassed estimator. More formally, from (4) $\hat{\beta} - \beta \xrightarrow{p} 0$ only if $T^{-1}\sigma^2 \sum x_i' F_i^{-1} f_i \xrightarrow{p} 0$, and it is obvious that this will never happen. Hence the OLS estimator using only the positive observations is *inconsistent.*

The same conclusion holds if all observations are used i.e. the data is *censored* so that both the positive and zero values of y_i are available. To see this follow Greene and define

$$E[y_i|x_i'\beta] = 0 \times prob[y_i^* \leq 0|x_i'\beta] + E[y_i^*|y_i^* > 0, x_i'\beta] \times prob[y_i^* > 0|x_i'\beta].$$

The analysis above computed $E[y_i^*|y_i^* > 0, x_i'\beta]$ as $x_i'\beta + \sigma^2 F_i^{-1} f_i$ and $prob[y_i^* > 0|x_i'\beta] = F_i$ giving

$$E(y_i|x_i'\beta) = (x_i'\beta + \sigma^2 F_i^{-1} f_i)F_i.$$

For later reference we observe that

$$E(v_i^2|x_i'\beta) = \sigma^2 - \sigma^2 x_i'\beta(f_i|F_i) \quad \text{(Amemiya, \textit{Econometrica,} 1973)}$$

so the error term not only has a non-zero mean but is also heteroskedastic.

3 Weighted Least Squares and MLE

How can we get a consistent estimator of β? For a truncated sample one possibility is to write

$$y_i = x_i'\beta + \sigma^2 F_i^{-1} f_i + \eta_i \;\; i \in I_p$$

and regress y_i against x_i and $F_i^{-1} f_i$. The problem here is that we need a consistent estimator of β to form $F_i^{-1} f_i$, as that depends on β, and it is not clear where we get that from. Non-linear regression would be one possibility i.e. minimize $\sum_{i \in I_p} \eta_i^2$ w.r.t. β. Notice however that η_i must be heteroskedastic. Therefore, even if one got a consistent estimator of β by some means, one would need to allow for the heteroskedasticity if valid inferences were to be obtained or if a fully efficient estimator was to be found. Weighted non-linear least squares would be one possibility to handle that complication. Instead one might perform full maximum likelihood using the density function for $f(v_i)$ defined earlier. A similar choice obtains for the censored data case where

$$y_i = (x_i'\beta + \sigma^2 F_i^{-1} f_i)F_i + \eta_i$$

Again one has a problem about how to estimate F_i as β is unknown. In a later section where we discuss selectivity problems it emerges that in some instances where there is extra information we can estimate f_i and F_i without knowing β from auxiliary information, but that is not available in the pure Tobit model.

Although there is recent work using some *semi-parametric* methods to estimate β which treat f_{u^*} as unknown, the main way people estimate the Tobit model is by maximum likelihood. In this we think of y_i as a random variable that is a *mixture* of a discrete random variable (the zero values) and a continuous random variable (the positive ones). The likelihood for the Tobit model (a *limited dependent* variable model) requires us to determine the probability that the *r.v.* describes the observed data. Now the probability of getting a zero value is prob $(y_i^* \leq 0) = prob(x_i'\beta + u_i^*) \leq 0 = prob(u_i^* \leq -x_i'\beta) = \int_{-\infty}^{-x_i'\beta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(\lambda/\sigma)^2\right\} d\lambda = 1 - F_i$ as F_i is the integral of the normal density from $-x_i'\beta$ on to ∞ and the integral from $\int_{-\infty}^{\infty}$ must be unity. The probability of observing the non-zero data is just that from the normal density, so that the likelihood is

$$L^* = \prod_{i\in I_0}(1-F_i)\prod_{i\in I_p}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(y_i - x_i'\beta)^2}$$

making the log likelihood

$$L = \sum_{i\in I_0}\log(1-F_i) - \frac{1}{2\sigma^2}\sum_{i\in I_p}(y_i - x_i'\beta)^2 - \frac{N_p}{2}\log\ \sigma^2 - \frac{N_p}{2}\log\ 2\pi$$

Since the MLE has desirable properties of consistency and asymptotic normality in most cases, it is natural that one should estimate the unknown parameters β and σ^2 in the Tobit model above by maximizing L *wrt* β and σ^2. But this is not straightforward as β appears in F_i as well as in $(y_i - x_i'\beta)^2$ i.e. L is a non-linear function of β and an *iterative* procedure to maximize L *wrt* β, σ^2 is needed. Fortunately, one can show that there is a single maximum to L so that algorithms such as scoring or Newton-Raphson work well. The scores and Hessian (second derivative of the log likelihood) for the MLE are[2]

$$\begin{aligned}
L_\beta &= -\sum_{i\in I_0}(1-F_i)^{-1}f_i x_i + \sigma^{-2}\sum_{i\in I_p}x_i(y_i - x_i'\beta) \qquad (5)\\
L_{\sigma^2} &= \frac{1}{2\sigma^2}\sum_{i\in I_0}(1-F_i)^{-1}(x_i'\beta)f_i - \frac{1}{2\sigma^2}N_p + \frac{\sigma^{-4}}{2}\sum_{i\in I_p}(y_i - x_i'\beta)^2\\
L_{\beta\beta} &= -\sum_{i\in I_0}(1-F_i)^{-2}f_i[f_i - \sigma^{-2}(1-F_i)(x_i'\beta)]x_i x_i' - \sigma^{-2}\sum_{i\in I_p}x_i x_i'.
\end{aligned}$$

The Hessian is used to get estimated standard errors for the MLE of β.

4 Issues Arising in Testing and Interpreting Censored Regression

(i) It is a well known property of the MLE that an estimate of the covariance matrix of $\hat{\beta}$ and σ^2 can be obtained from the negative of the inverse of the second derivatives of L *wrt* β and σ^2.

[2] We have not given $L_{\beta\sigma^2}$ and $L_{\sigma^2\sigma^2}$.

(ii) Because the model is not a linear one, or one that just involves minimizing a quadratic form, there is no such thing as an R^2. We can however construct a pseudo$-R^2$ by *looking at the way in which the R^2 arises in a general linear model.* Let U designate an unrestricted regression model $y_i = \beta_0 + x_i'\beta_1$, and R the same model excluding all regressors bar the constant. The log-likelihoods for both models are

$$L_U = -\frac{T}{2}\log\hat{\sigma}_U^2 - \frac{1}{2\hat{\sigma}_U^2}\sum(y_i - \beta_{0,U} - x_i'\hat{\beta}_U)^2$$

$$L_R = -\frac{T}{2}\log\hat{\sigma}_R^2 - \frac{1}{2\hat{\sigma}_R^2}\sum(y_i - \hat{\beta}_{0R})^2$$

where $\hat{\beta}_{0,R}$ is the estimate of the constant term. Now $\hat{\sigma}_U^2 = T^{-1}\sum(y_i - \hat{\beta}_{0,U} - x_i'\hat{\beta}_U)^2$ and $\hat{\sigma}_R^2 = \frac{1}{T}\sum(y_i - \hat{\beta}_{0,R})^2$ so that

$$\begin{aligned} L_U &= -\frac{T}{2}\log\hat{\sigma}_U^2 - \frac{T}{2} \\ L_R &= -\frac{T}{2}\log\hat{\sigma}_R^2 - \frac{T}{2} \\ \therefore L_U - L_R &= \frac{T}{2}\log(\hat{\sigma}_R^2/\hat{\sigma}_U^2) = \frac{T}{2}\log(ESS_R/ESS_U) \\ &= -\frac{T}{2}\log(ESS_U/ESS_R) \end{aligned}$$

where ESS_R and ESS_U are the residual sums of squares in the restricted and unrestricted models. Now for a linear regression model the definition of R^2 is

$$1 - \frac{ESS_U}{\sum(y_i - \bar{y})^2} = 1 - \frac{ESS_U}{ESS_R}$$

since $\hat{\beta}_{0,R} = \bar{y} = mean$ of the y_i.

$$\begin{aligned} \therefore R^2 &= 1 - \frac{ESS_U}{ESS_R} \Rightarrow \frac{ESS_U}{ESS_R} = 1 - R^2 \\ \therefore L_U - L_R &= -\frac{T}{2}\log(1 - R^2) \end{aligned}$$

and hence we can solve for an R^2 from L_R, L_U and T. Thus an "R^2" for the Tobit model is available by computing the value of the likelihood with all x_i variables included (L_U) and the value with only a constant term (L_R) and then solving the above. There are other possibilities involving different ways of measuring ESS_U. Since in the regular regression model the residuals are just $y_i - \hat{E}(y_i)$ we might do the same here using the $E(y_i|x_i'\beta)$ for a Tobit model and then inserting the MLE of β and σ^2 in place of the unknown values.

(iii) Specification errors might be detected in the same way as in linear regression, i.e., by adding extra variables into the relation and seeing if they are significant. Thus a RESET type test would be available by adding $(x_i'\hat{\beta})^2$, $(x_i'\hat{\beta})^3$, etc to the model and the re-estimating with a Tobit estimator. The LM approach also yields specification tests which resemble those of the linear regression model in that they work with what are referred to as *generalized residuals.* These are defined by

$$-(1 - I(y_i > 0))[1 - \hat{F}_i]^{-1}\hat{f}_i + I(y_i > 0)\sigma^{-2}(y_i - x_i'\hat{\beta})$$

and one sees from (5) that the product of these with respect to x_i give the scores for β, just as would be true of the ordinary residuals in a regression model. Greene discusses this literature.

(iv) What is more of a problem with Tobit models is that the presence of heteroskedasticity causes the Tobit model estimators of β to be inconsistent. Hence one should test for the presence of heteroskedasticity, but so far packages have not allowed users to do this automatically. The Lagrange Multiplier test provides a way of testing for this. Greene has some discussion. It is generally relatively simple to write down the log likelihood for the case that the errors u_i^* are heteroskedastic. In the standard case $\sigma^{-1}u_i^*$ is taken to be $\mathcal{N}(0,1)$; now $\sigma_i^{-1}u_i^*$ has that density. The log likelihood then just follows. All that one then needs to do is to specify some form for σ_i^2.

(v) As was true of the discrete choice model the marginal effects of x_i upon the dependent variable need to be computed with some care. First, one has to define what the response variable is. In linear regression one focuses upon $\partial E(y)/\partial x$ but in Tobit models one could look at other measures e.g. $\partial E(y|y^* > 0)/\partial x$. In all instances the fact that $E(y)$ etc are non-linear functions of x_i means that the marginal response is not β. This is obvious once one sees $E[y_i|x_i'\beta] = [x_i'\beta + \sigma^2\frac{f_i}{F_i}]F_i$. A second problem arises if the x_i are dummy variables, (say) taking the value zero and one. Then the derivative is not relevant. One has to compute $E[y|x_i'\beta]$ with the dummy set to unity and then set to zero and subtract one from the other i.e. a finite difference is needed. In a linear model one doesn't need to do this since the response will be β regardless of whether we use a partial derivative or a finite difference. The difference in answers can be very large. One should note that the marginal effects given in programs such as STATA are partial derivatives so they need to be used cautiously with dummy variable among the $x's$.

5 Selected Samples

Selection problems arise when the sample presented to us has been selected by some non-random mechanism. It is similar to truncation where the sample available fails to be representative of the complete population due to observations being "deleted" if they fall below the truncation point. With selection, the truncation is more subtle; whether or not one sees a complete sample depends upon decisions made by individuals about some other choice. For example, suppose you wished to determine the income elasticity of demand for hotel accommodation and sampled tourists in Bermuda. This would be subject to potential selection problems since it is costly to travel to Bermuda and you therefore may be using a sample of people who have high incomes and these may be unrepresentative of the general population. You did not deliberately select a high income group but one has been presented to you by the choices made by the population at large.

How does one correct for a selection problem? When data is truncated the solution was to find the expected value of y_i recognizing the existence of a truncation mechanism, and that was then used to correct for the bias. To do this we needed to make some distributional assumptions. A similar situation occurs with selection. However, there is generally more information available now as it is typically assumed that there are two samples of information- one being the selected sample while the other describes what causes agents from the broader population to participate in the selected sample. Thus the first sample would be tourists from Bermuda whereas the second would be drawn from the population that these tourists come from and would consequently contain information about whether they choose to go to Bermuda or not. This second sample can therefore be used to predict which individuals will go to Bermuda and so correct for any biases that stem from the fact that work is only being done on a sample of Bermuda tourists.

From the description above we have two equations

$$\begin{aligned} y_{1i} &= x_{1i}'\beta_1 + u_{1i} \quad i = 1, ..., N & (6) \\ y_{2i}^* &= x_{2i}'\beta_2 + u_{2i}^* \quad i = 1, ..., n, & (7) \end{aligned}$$

where N is the number of observations in the selected sample and n is the number from the broader population. (6) is the equation to be estimated using the selected sample and (7) is to be used to predict whether an individual becomes part of the selected sample. The data available to estimate the second equation is x_{2i} and $y_{2i} = 1(y_{2i}^* > 0)$ i.e. a score of one is registered if the individual in that sample participates in the selected one i.e travels to Bermuda. The second equation will be used to measure the probability that (or *propensity* for) the individual to participate in the first sample. Since the first sample consists only of individuals for whom y_{2i} was unity, what we have are realizations from the conditional density $f(y_{1i}|y_{2i} = 1) = f(y_{i1}|y_{2i}^* > 0)$ and it is the characteristics of this conditional density that will be exploited in order to make a "selectivity bias correction".[3]

Although one might find $f(y_{i1}|y_{2i}^* > 0)$ and perform MLE, in practice researchers have often opted to work with regression solutions, largely because selection problems occur in combination with other difficulties such as endogenous x_i, and it becomes very difficult to find the likelihood in realistic cases. Consequently, the most important task is to find an expression for $E(y_{1i}|y_{2i}^* > 0)$ and this is

[3] The density is also conditional upon x_{1i} and x_{2i} but we will supress that dependence.

$$E(y_{1i}|y_{2i}^* > 0) = x_{1i}'\beta_1 + E(u_{1i}|y_{2i}^* > 0).$$

Now assume that $\begin{pmatrix} u_{1i} \\ u_{2i}^* \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right]$. From properties of the bivariate normal

$$u_{1i} = \rho u_{2i}^* + \eta_i,$$

where η_i is independent of u_{2i}^* and $\rho = \sigma_{12}/\sigma_{22}$.

Hence

$$E(u_{1i}|y_{2i}^* > 0) = \rho E(u_{2i}^*|y_{2i}^* > 0) + E(\eta_i|y_{2i}^* > 0).$$

Because η_i is independent of u_{2i}^* it must be independent of y_{2i}^* and so the last expectation is $E(\eta_i) = 0$, leaving us to evaluate the first. But this comes from results on moments of truncated normal random variables (see Greene), viz that for an $N(0, \sigma^2)$ random variable, u, that is truncated at a,

$$E(u|u > a) = \sigma\frac{\phi(a)}{\{1 - \Phi(a)\}}$$

In this case $y_{2i}^* > 0$ implies $u_{2i}^* > -x_{2i}'\delta_2$ so $a = -x_{2i}'\delta_2$ and

$$\begin{aligned} E(u_{2i}^*|u_{2i}^* &> -x_{2i}'\delta_2) = \sigma_{22}\frac{\phi(-x_{2i}'\delta_2)}{\{1 - \Phi(-x_{2i}'\delta_2)\}} \\ &= \sigma_{22}\frac{\phi(-x_{2i}'\delta_2)}{\Phi(x_{2i}'\delta_2)} = \sigma_{22}\phi_i/\Phi_i \end{aligned}$$

where

$$\phi_i = (2\pi)^{-1/2}\exp\left\{-(1/2)\left(x_{2i}'\delta_2\right)^2\right\}$$

and

$$F_i = \int_{-\infty}^{x_{2i}'\delta_2} f(\lambda)d\lambda.$$

Consider defining $x_{2i}'\beta_2/\sigma_{22}^{1/2}$ as $x_{2i}'\alpha$ so that

$$f_i = \sigma_{22}^{-1/2}\left[(2\pi)^{-1/2}\exp(-1/2)(x_{2i}'\alpha)^2\right] = \sigma_{22}^{-1/2}\phi_i$$

$$\begin{aligned} F_i &= \int_{-\infty}^{x_{2i}'\alpha} (2\pi)^{-1/2}\exp(-(1/2)\psi^2)d\psi \text{ using } \psi \\ &= \frac{\lambda}{\sigma_{22}^{1/2}}(\text{note } d\psi = (d\lambda)\sigma_{22}^{-1/2}) = \int_{-\infty}^{x_{2i}'\alpha} \phi(\psi)d\psi. \end{aligned}$$

Hence $\rho f_i/F_i = \rho\sigma_{22}^{-1/2}\phi_i/\Phi(x_{2i}'\alpha)$ and we can regress y_{11} against x_{1i} and $\phi_i/\Phi(x_{2i}'\alpha)$ to get a consistent estimator of β_1.

The second equation can be written as

$$\begin{aligned}\sigma_{22}^{-1/2}y_{2i}^* &= x_{2i}'\left(\frac{\beta_2}{\sigma_{22}^{-1/2}}\right)+\left(\frac{u_{2i}^*}{\sigma_{22}^{1/2}}\right)\\ &= x_{2i}'\alpha+\nu_i\end{aligned}$$

where ν_i is $\mathcal{N}(0,1)$. But we have observations on $y_{2i} = 1(y_{2i}^* > 0)$ and so this is a Probit model. Applying Probit we can estimate α, getting $\hat{\alpha}$, and then proceed to get $\hat{\beta}_{1i}$ by regression of y_{1i} against x_{1i} and $\hat{\phi}_i/\hat{\Phi}_i$, where $\hat{\alpha}$ replaces α in these terms. We have to be careful to get the correct covariance matrix for two reasons: Replacing α with $\hat{\alpha}$ has an effect on the distribution of $\hat{\beta}$ and the errors are heteroskedastic. Programs such as STATA do this correctly. This estimator is sometimes referred to as Heckman's two-step estimator.