ApEc 8213: Econometric Analysis III -- Lecture #5

Censoring and Selection Models
Hansen, Chapter 27

A **censored regression** is situation where the **dependent variable has a "mass point" at a particular value**. For example, purchases of meat products could have many observations equal to 0 due to people who are vegetarians.

Selection occurs when **being in the sample is endogenous**. For example, a sample of employed people does not include people who are not working.

I. Censoring (27.2, 27.3 and 27.4)

Censoring occurs when the dependent variable has a large number of observations with the value of that variable at a "boundary" of the range of that dependent variable. **Often the boundary is a lower bound of 0, but other situations are possible**, **including upper bounds**. We will focus on a lower bound of 0, but the methods we discuss can be easily adapted to a lower bound of a different value, an upper bound, or even both lower and upper bounds.

Hansen gives an **example** of households in the Philippines that receive **remittances** from (former) members who now live in other countries. Most (80%) do not receive such remittances, so the value of this variable for them is 0. But 20% receive remittances, so the value is > 0.

Suppose you want to estimate the determinants of households' receipts of remittances. What should you do? Should you drop the 80% that do not receive remittances? We should consider this more carefully.

Tobin (1958) provided the following strategy. The model, now called a **Tobit**, is the following:

$$Y^* = X'\beta + e \qquad (27.1)$$

$$e \mid X \sim \mathrm{N}(0, \sigma^2)$$

$$Y = \max(Y^*, 0)$$

where $e \mid X$ means that e is assumed to be independent of all X variables. Here, ***Y** exists for everyone** but it is "latent" (not always observed). ***Y* is observed**. If $Y^* < 0$, then it is not observed, and instead we observe that $Y = 0$.

Notice that the error term ***e* is assumed to be normally distributed**. This is very **useful for maximum likelihood estimation**, but if this assumption is incorrect the maximum likelihood estimate of β will be biased and inconsistent, a point that we will return to below.

Interpreting this model can be a little difficult. If Y^* is "optimal" purchases of a good, one could argue that people who want to buy negative values (want to sell that amount?) cannot do that, so the optimal choice for them is to purchase 0 of that good. But it is a little strange to say that their optimal demand is a negative number.

With the above setup, there is Y^* and Y. **A third possible dependent variable**, which Hansen denotes as $\boldsymbol{Y^{\#}}$, is the situation where observations of $Y = 0$ are dropped:

$$\begin{aligned} Y^{\#} &= Y && \text{if } Y > 0 \\ &= \text{missing} && \text{if } Y = 0 \end{aligned}$$

Going back to the model in equation (27.1), a useful property is the **conditional probability of censoring** (conditional on X). If $e \mid X \sim \mathrm{N}(0, \sigma^2)$, then:

$$\mathrm{Prob}[Y^* < 0 \mid X] = \mathrm{Prob}[e < -X'\beta \mid X] = \Phi\left(-\frac{X'\beta}{\sigma}\right)$$

This expression can be used to calculate the expected value of the three different Y variables conditional on X:

$$m^*(X) = \mathrm{E}[Y^* \mid X] = X'\beta$$

$$m(X) = \mathrm{E}[Y \mid X] = X'\beta\Phi\left(\frac{X'\beta}{\sigma}\right) + \sigma\phi\left(\frac{X'\beta}{\sigma}\right) \qquad (27.2)$$

$$m^{\#}(X) = \mathrm{E}[Y^{\#} \mid X] = X'\beta + \sigma\lambda\left(\frac{X'\beta}{\sigma}\right) \qquad (27.3)$$

where $\lambda(\ )$ is defined as $\phi(\ )/\Phi(\ )$, which is called the **inverse Mills ratio.** Derivation of (27.2) and (27.3) can be a homework problem.

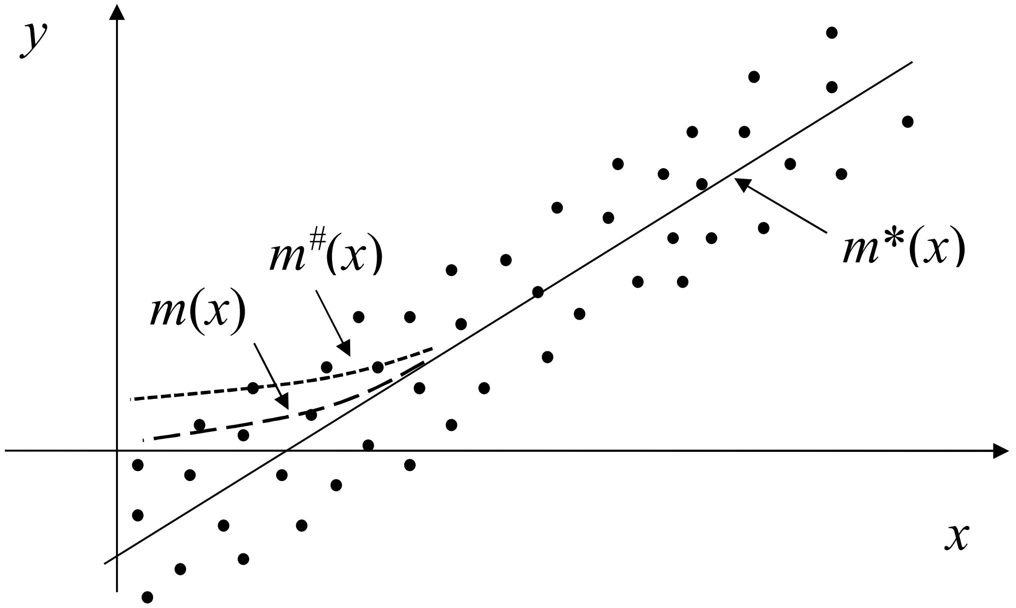
By the definition of Y^*, it is clear that $Y^* \leq Y$. This, and comparing equations (27.2) and (27.3), imply that:

$$m^*(X) \leq m(X) \leq m^{\#}(X)$$

Both inequalities are **strict inequalities if there is censoring** for a given value of X. Indeed, $\mathrm{E}[Y \mid X] < \mathrm{E}[Y^{\#} \mid X]$ follows from (27.2) and (27.3) whenever $\Phi(X'\beta/\sigma) < 1$, which implies that $\Phi(-X'\beta/\sigma) = \mathrm{Prob}[Y^* < 0 \mid X] > 0$ (whenever censoring is possible for a given value of X).

Clearly, if we observed Y^* we could use OLS to get an unbiased estimate of β. **When censoring is present** for at least some values of X, the fact that both $\mathrm{E}[Y \mid X]$ and $\mathrm{E}[Y^{\#} \mid X]$ are both $> \mathrm{E}[Y^* \mid X]$ implies that **OLS regression using Y or $Y^{\#}$ will lead to biased results**.

The figure below gives intuition for these biases, for the case of one X variable:



Bias of OLS Estimation

Consider an OLS regression of *Y* on the *X* variables under the assumptions of the Tobit model. Greene (1981) showed, **for the special case where all *X* variables are normally distributed**, that:

$$\mathrm{E}[\hat{\beta}_{\mathrm{OLS}}] = \beta(1 - \pi) \qquad (27.4)$$

where $\pi = \mathrm{Prob}[Y = 0]$. (Note that Hansen denotes $\hat{\beta}_{\mathrm{OLS}}$ by $\hat{\beta}_{\mathrm{BLP}}$ "best linear predictor".) Thus, for this special case, the **OLS estimates of *β* are all biased toward zero** (under-estimated in absolute value). While it is very unlikely that the *X* variables are normally distributed, this is a useful approximation of the bias, at least for continuous variables that have a distribution that is approximately symmetric.

At the end of Section 27.4, Hansen derives Greene's formula. He skips a few steps, so here is a more detailed derivation (though I do skip one step):

$$\mathrm{E}[\hat{\beta}_{\mathrm{OLS}}] = \mathrm{E}[XX']^{-1}\mathrm{E}[XY]$$

$$= \mathrm{E}[XX']^{-1}\{\mathrm{E}[XY|\, Y^* > 0]\times\mathrm{Prob}[Y^* > 0] + \mathrm{E}[XY|\, Y^* \leq 0]\times\mathrm{Prob}[Y^* \leq 0]\}$$

$$= \mathrm{E}[XX']^{-1}\mathrm{E}[XY^*|\, Y^* > 0](1 - \pi)$$

$$= \mathrm{E}[XX']^{-1}\mathrm{E}[XY^*](1 - \pi)$$

$$= \beta(1 - \pi)$$

Hansen shows (p.876) that = $\mathrm{E}[XY^*|\, Y^* > 0] = \mathrm{E}[XY^*]$.

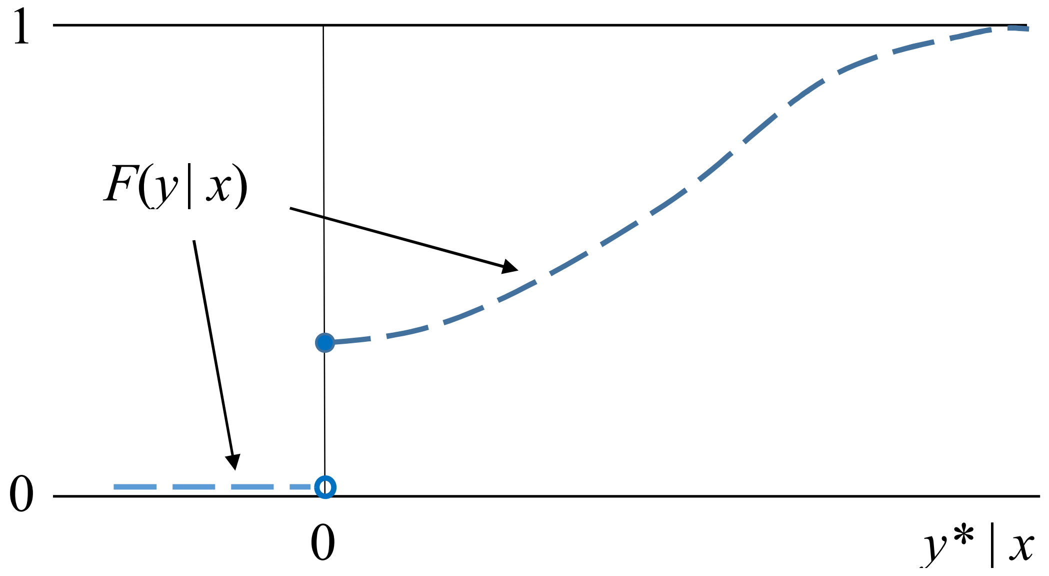
II. Tobit Model: Estimation and Identification (27.5 - 27.7)

The "classic" Tobit model is estimated by maximum likelihood. The observed value of Y combined a discrete part (when $Y = 0$) and a continuous part (when $Y > 0$):

$$F(y|\, x) = 0, \qquad \text{for } y < 0$$

$$= \Phi\left(\frac{e}{\sigma}\right) = \Phi\left(\frac{y - x'\beta}{\sigma}\right) \text{ for } y \geq 0$$

The following diagram should make this clearer:



The associated **density function** is:

$$f(y|\,x) = \Phi\left(\frac{-x'\beta}{\sigma}\right)^{1[y=0]}\left[\sigma^{-1}\phi\left(\frac{y-x'\beta}{\sigma}\right)\right]^{1[y>0]}$$

The **first part** is the **probability of censoring**, which is the probability that $y^* < 0$ (if $y^* = 0$ then $e = -x'\beta$). The second part is the density for a regression if e is normally distributed.

The **log likelihood** is the sum of the log density **over all of the observations**:

$$\ell_n(\beta, \sigma^2) = \textstyle\sum_{i=1}^{n} \log(f(Y_i|\,X_i))$$

$$= \textstyle\sum_{i=1}^{n}\{1[Y_i = 0]\log(f(Y_i|\,X_i)) + 1[Y_i > 0]\log[\sigma^{-1}\phi\left(\frac{Y_i - X_i'\beta}{\sigma}\right)]\}$$

$$= \textstyle\sum_{Y_i=0} \log(\Phi\left(\frac{-X_i'\beta}{\sigma}\right)) - \frac{1}{2}\sum_{Y_i>0}(\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(Y_i - X_i'\beta)^2)$$

The **first part** is the **same as** that for the **probit model for cases where *Y* = 0**, and the **second part** is the same as a standard **regression model** with a **normally distributed *e***.

As usual, the ***maximum likelihood estimator*** is the values of β and σ that maximize this likelihood function:

$$\{\hat{\beta}_{\text{mle}}, \widehat{\sigma^2}_{\text{mle}}\} = \arg\max_{\beta,\sigma^2} \ \ell_n(\beta, \sigma^2)$$

To find the values of β and σ^2 that maximize the log-likelihood function, you need to search for it using numerical optimization methods.

Olsen (1978) provided a useful transformation for computation. Define $\gamma = \beta/\sigma$ and $v = 1/\sigma$. Then the likelihood function becomes:

$$\ell_n(\gamma, v) = \sum_{Y_i=0} \log(\Phi(-X_i'\gamma)) + n_1 \log(v/\sqrt{2\pi}) - \frac{1}{2}\sum_{Y_i>0}(Y_i v - X_i'\gamma)^2 \quad (27.5)$$

where n_1 is the number of observations that are > 0. **Each of these three terms is globally concave** in γ and v (as Hansen explains on p.877), **so the overall likelihood function is globally concave**, and so it generally converges quickly to the maximum.

Identification in Tobit Regressions

The **Tobit** model is based on **several assumptions** that **may not hold**. A **very general model** of the same process is:

$$Y^* = m(X) + e$$

$$\mathrm{E}[e] = 0$$

$$Y = \max(Y^*, 0)$$

where e is distributed as $F(e)$ and is independent of X.

There are two generalizations here. **First**, the $m(X)$ function is no longer linear in X, and it could even be "nonparametric" (this will be discussed next week). **Second**, e is no longer assumed to be normally distributed and could be heteroscedastic.

It turns out that **if we make no assumptions** about the $m(X)$ function and we make no functional form assumptions about e, then the censoring that occurs when $Y^* < 0$ makes it **impossible to estimate E[*Y*| *X*]**. The **intuition** is that censoring "snips the tails" of the distribution of e, and we need the whole distribution of e to estimate $\mathrm{E}[Y|\, X]$.

Yet it is **possible to estimate the *median* of *Y* conditional on *X*** (and, more generally, quantiles of Y conditional on X).

Censored Least Absolute Deviation (CLAD) Estimator

A median regression estimates the median of Y conditional on X: $\mathrm{Med}[Y|\, X] = X'\beta_{\mathrm{Med}}$. This is very insensitive to outlier data points, which can be useful for "noisy" data. Powell (1984) adapted this to censored data:

$$Y^* = X'\beta + e$$

$$\text{Med}[e|\, X] = 0$$

$$Y = \max(Y^*, 0)$$

The **most important point** is that we **no longer assume that *e* follows any particular distribution**.

This model can also be expressed as Med[$Y^* \mid X$] = $X'\beta$.

The **median has a very useful property** for this model:

$$\text{Med}[Y \mid X\,] = \max(X'\beta, 0)$$

[Draw a picture to show this.]

This can be estimated using least absolute deviations (LAD) methods (an application of m-estimation, which is explained in Chapter 22, of Hansen's book). The criterion to minimize is:

$$M_n(\beta) = (1/\text{n}) \sum_{i=1}^{n} |\; Y_i - \max(X_i'\beta, 0)\; |$$

Then $\hat{\beta}_{\text{CLAD}}$ can be defined as:

$$\hat{\beta}_{\text{CLAD}} = \underset{\beta}{\text{argmin}}\; M_n(\beta)$$

One potential problem is that $M_n(\beta)$ **may not be globally convex**, so you may end up "converging" to a local minimum rather than the global minimum.

At the end of Section 27.7, Hansen discusses extending the CLAD estimator to other quantiles (other than the median), which allows you to estimate "censored quantile regressions". I doubt that this is used very much, so this material is optional.

Derivatives for Tobit Models

Hansen does not present the derivatives of Y with respect to the X variables. The following is from Wooldridge (2010, pp.673-674). This is for continuous X variables. For binary X variables, see p.675 of Wooldridge.

There are two derivatives of interest: $\partial\mathrm{E}[Y\,|\,X\,]/\partial X_j$ and $\partial\mathrm{E}[Y\,|\,X,\,Y > 0]/\partial X_j$, for some variable X_j. For example, if Y is cigarettes smoked and X_j is the price of cigarettes, $\partial\mathrm{E}[Y\,|\,X\,]/\partial X_j$ is the change in consumption of cigarettes from a change in the price of cigarettes for the entire population, while $\partial\mathrm{E}[Y\,|\,X,\,Y > 0]/\partial X_j$ is the change of consumption for those who smoke cigarettes.

It turns out that $\partial\mathrm{E}[Y\,|\,X\,]/\partial X_j$ has a simple expression:

$$\partial\mathrm{E}[Y\,|\,X\,]/\partial X_j = \Phi\left(\frac{X'\beta}{\sigma}\right)\beta_j$$

The expression for $\partial \mathrm{E}[Y \mid X, Y > 0]/\partial X_j$ is more complicated:

$$\partial \mathrm{E}[Y \mid X, Y > 0]/\partial X_j = \beta_j\{1 - \lambda\left(\frac{X'\beta}{\sigma}\right)[\frac{X'\beta}{\sigma} + \lambda\left(\frac{X'\beta}{\sigma}\right)]\}$$

Endogenous variables in Tobit models

This is not discussed in Hansen's book, but it is in Wooldridge (2010, pp.681-685). Change (27.1) to:

$$Y_1{*} = X'\beta + \gamma Y_2 + e$$

$$\text{with } Y_2 = X'\delta_1 + Z'\delta_2 + v$$

where the vector Z is the excluded instruments for Y_2. If e and v are correlated, then Y_2 is endogenous and standard Tobit estimates for β and γ will be inconsistent.

Smith and Blundell (1986) developed a control function approach for this. Estimate the Y_2 equation by OLS and calculate the residuals as $\hat{v} = Y_2 - X'\hat{\delta}_1 - Z'\hat{\delta}_2$. Then estimate $Y_1{*} = X'\beta + \gamma Y_2 + \theta\hat{v} + e$ by OLS. This provides consistent estimates of β, γ and θ. The θ coefficient is a test for endogeneity. If it is insignificant, then the OLS standard errors for β and γ are correct. If θ is significant, then you need to adjust the standard errors, either using the Smith and Blundell formulas or by a 2-step bootstrap.

III. Sample Selection Bias (27.9, 27.10, and 27.11)

Sometimes your data are not from a random sample but **instead from a "selected" sample** of the population of interest. The one most familiar to economists is a sample of working individuals. You may be interested in estimating the impact of variables such as education and years of experience on the wages of the entire population, including those not currently working, but the selected sample includes only those who are working. Hanson gives some other examples in Section 27.9 of Chapter 27.

It is **useful to model this situation as a 2-stage process**. The **first stage is a random sample** of the Y and X variables **for the population of interest**, and the **second stage is the sample that you have**. We can **define the variable S** ("selection") as $S = 1$ for observations in your sample and $S = 0$ for the observations not in your sample.

Consider the **linear model $Y = X'\beta + e$**, with $\mathrm{E}[e| X] = 0$. The conditional mean for the observed (selected) sample is:

$$\mathrm{E}[Y| X, S = 1] = X'\beta + \mathrm{E}[e| X, S = 1]$$

Selection bias occurs when the second term is ≠ 0.

A standard approach is to **model the selection process**. Assume that there is some **latent variable S^*** $= X'\gamma + u$, such that $S^* > 0$ implies $S = 1$ and $S^* \leq 0$ implies $S = 0$. This can be expressed as:

$$S = 1[X'\gamma + u > 0]$$

Note: The X variables in the Y model can be different from those in the S model, by setting some β or γ terms equal to 0. Either way, ***X* is the same in both models**.

This expression for S implies that:

$$\mathrm{E}[Y|\, X, S = 1] = X'\beta + \mathrm{E}[e|\, u > -X'\gamma]$$

Let $\boldsymbol{e = \rho u + \varepsilon}$ be the equation for a **linear projection of *e* on *u***. Assume that all these errors (e, u and ε) are independent of X, and that u and ε are independent of each other.

Question: Is e independent of ε? of u?

This allows us to **write the above expression as**:

$$\mathrm{E}[Y|\, X, S = 1] = X'\beta + \rho\mathrm{E}[u|\, u > -X'\gamma] = X'\beta + \rho g(X'\gamma)$$

for some function $g(u)$. Note that if $u \sim \mathrm{N}(0, 1)$ then $g(u) = \lambda(u)$, the inverse Mills ratio. In this case:

$$\mathrm{E}[Y|\, X, S = 1] = X'\beta + \rho\lambda(X'\gamma) \qquad (27.6)$$

If $\rho = 0$, then $E[Y| X, S = 1] = X'\beta$ and we can use OLS to obtain unbiased estimates of β using only the selected sample. **So bias arises when unobserved components in the selection equation are correlated with the unobserved components in the equation for *Y*.** [Another case where there would be no selection bias would be if all elements of γ except the constant term were equal to 0, but this is highly unlikely.]

Thus, if $\rho \neq 0$ then OLS estimates on the selected sample are almost certainly biased. The best (only?) way to avoid this is to have some data on the selection process.

Heckman's Model ("Heckit")

Heckman (1979) provided a method to correct selection bias that can be used if one has data on the "non-selected" observations (not *Y*, but *X* and *S*). He explicitly **allows for the variables determining the selection process (*Z*) to be different from the *X* variables** determining *Y*:

$$Y^* = X'\beta + e$$

$$S^* = Z'\gamma + u$$

$$S = 1[S^* > 0]$$

$$\begin{aligned} Y &= Y^* && \text{if } S = 1 \\ &= \text{missing} && \text{if } S = 0 \end{aligned}$$

$$\begin{pmatrix} e \\ u \end{pmatrix} \sim \mathrm{N}\left(0,\ \begin{pmatrix} \sigma^2 & \sigma_{21} \\ \sigma_{21} & 1 \end{pmatrix}\right)$$

The **assumption that *u* is normally distributed** with a variance of 1 **means that a probit is used for the selection equation**. Note also that σ_{21} is the same as ρ in the projection equation above $e = \rho u + \varepsilon$.

While this model has been applied to many different economic and social phenomena, Heckman's original interest was estimating wage equations, so Y was wages and S was a dummy variable for being employed. The e term in the wage equation could represent unobserved "ability", which could be correlated with the unobserved factors u that determine employment.

The regression model is similar to one we had for eq. (27.6):

$$\mathrm{E}[Y|\,X, Z, S = 1] = X'\beta + \sigma_{21}\lambda(Z'\gamma) \qquad (27.7)$$

where $\lambda(\ \ )$ is the inverse Mills ratio.

Heckman proposed a **2-step method** to estimate β (and γ).

1. Use a probit model to estimate γ, call the estimate $\hat{\gamma}$.
2.
3. Construct $\hat{\lambda}_i = \lambda(Z_i'\hat{\gamma})$. Then run an OLS regression of Y_i on X_i and $\hat{\lambda}_i$ for the subsample with $S = 1$.

Heckman showed that the estimated β from the second step, denoted by $\widehat{\boldsymbol{\beta}}$, is **consistent and asymptotically normally distributed** under the assumptions of the model.

Note that the **second step** regression also **estimates $\boldsymbol{\sigma_{21}}$.** This can be used to **test for endogeneity** (correlation of *e* and *u*): if $\sigma_{21} = 0$ then there is no selection bias.

A final important point is that it is **generally a good practice** to "identify" the $\lambda(Z'\gamma)$ term by having **at least one *Z* variable that is not included in *X***. The idea is that both $X'\beta$ and $\lambda(Z'\gamma)$ are **functional form assumptions** that **could be mistaken**. If the "true" functional forms are almost identical, that is we have $\theta_X(X'\beta) \approx \theta_Z(Z'\gamma)$, and all the *Z* variables are included in *X*, then β is not identified.

At the end of Section 27.10, Hansen presents full simultaneous estimation of this model as an alternative to the 2-step method. In principle this is more efficient if the assumptions of the model are correct. But some economists think that this could exacerbate problems of bias if the model is not correct. Both methods can be implemented in Stata using the "heckman" command.

Nonparametric Estimation

Heckman's method could give inconsistent results if the functional form assumptions are incorrect. **A more general model** is the following:

$$Y^* = m(X) + e$$

$$S^* = g(Z) + u$$

$$S = 1[S^* > 0]$$

$$\begin{aligned} Y &= Y^* && \text{if } S = 1 \\ &= \text{missing} && \text{if } S = 0 \end{aligned}$$

Note that in addition to allowing for nonlinear functions for Y^* and S^*, this model **no longer assumes that the error terms *e* and *u* are normally distributed**.

Hansen claims that because u cannot be identified, it is OK to "normalize the distribution of u to a convenient functional form", and he imposes the assumption that it is normally distributed. I do not think there is a general consensus that this is OK, and Hansen does not always use this assumption.

If you do not assume that u is normally distributed, you can estimate γ in the selection equation using a semi-parametric estimation method, such as the Klein-Spady method (he discusses this in Section 25.11 in Chapter 25, but he prefers series estimation).

The functions ***m*(*X*) and *g*(*Z*)** can be **flexibly approximated** by $X'\beta$ and $Z'\gamma$, respectively, when X and Z can include power and interaction terms involving the original X and Z variables (see Chapter 20 in Hansen on series regression).

Whatever method you use to estimate $\boldsymbol{\gamma}$, there are **three possibilities for estimating $\boldsymbol{\beta}$**. The **first** is:

$$\mathrm{E}[Y|\, X, Z, S = 1] = X'\beta + h_1(Z'\gamma) \qquad (27.8)$$

where $h_1(Z'\gamma) = \mathrm{E}[e\,|\, u > - Z'\gamma]$. **We do not know the functional form of $\boldsymbol{h_1}$( ), but we can approximate it** with a simple polynomial in $Z'\gamma$. For example in (27.8) replace $h_1(Z'\gamma)$ with $\delta_1 Z'\hat{\gamma} + \delta_2(Z'\hat{\gamma})^2 + \delta_3(Z'\hat{\gamma})^3$.

Second, **assume that $\boldsymbol{u}$ is normally distributed** (based on Hansen's argument) but do not make any assumptions about the distribution if e. In this case, we have:

$$\mathrm{Prob}[S = 1|\, Z] = \mathrm{Prob}[u > - Z'\gamma \,|\, Z] = \Phi(Z'\gamma) = p(Z)$$

where $p(Z)$ is convenient notation, called the **propensity score**. Note that $Z'\gamma = \Phi^{-1}(p(Z))$, which means that we can express $h_1(Z'\gamma)$ as $h_1(\Phi^{-1}(p(Z)))$. Define $h_2(\;) = h_1(\Phi^{-1}(\;\;))$. Then we have $h_1(Z'\gamma) = h_2(p(Z))$, which implies:

$$\mathrm{E}[Y|\, X, Z, S = 1] = X'\beta + h_2(p(Z)) \qquad (27.9)$$

We do not know the functional form of $\boldsymbol{h_2}$( ), but we can approximate it with a simple polynomial in $p(Z)$, which equals $\Phi(Z'\gamma)$. For example in (27.9) replace $h_2(p(Z))$ with $\delta_1\Phi(Z'\hat{\gamma}) + \delta_2[\Phi(Z'\hat{\gamma})]^2 + \delta_3[\Phi(Z'\hat{\gamma})]^3$.

Third, again **assume that *u* is normally distributed** and do not assume anything about the distribution of *e*. Recall the inverse Mills ratio $\lambda(Z'\gamma) = \phi(Z'\gamma)/\Phi(Z'\gamma)$. It turns out that the inverse Mills ratio is invertible, so we can express $h_1(Z'\gamma)$ as $h_3(\lambda(Z'\gamma))$, where we define $h_3(\ ) = h_1(\lambda^{-1}(\ ))$, which implies:

$$\mathrm{E}[Y|\,X, Z, S = 1] = X'\beta + h_3(\lambda(Z'\gamma)) \qquad (27.10)$$

Again, we do not know the functional form of $h_3(\ )$, but we can approximate it with a simple polynomial in $\lambda(Z'\gamma)$. For example in (27.10) replace $h_3(Z'\gamma)$ with $\delta_1\lambda(Z'\hat{\gamma}) + \delta_2[\lambda(Z'\hat{\gamma})]^2 + \delta_3[\lambda(Z'\hat{\gamma})]^3$.

Hansen recommends (27.10) because, if it turns out that both *u* and *e* are normally distributed, then this is “first-order accurate”. I would recommend (27.8) because it does not make any assumptions about the distribution of *u* or *e*.

Whichever method you choose, you can calculate the appropriate variance-covariance matrix using GMM methods or bootstrap methods.

Also, you should have some variable in *Z* that is not one of the variables in *X*.