Title

stata.com

Intro 4 — Endogenous sample-selection features

Description Remarks and examples Also see

Description

Endogenous sample-selection problems are handled by the `select()` option. ERMs provide probit and tobit selection. Probit selection is discussed below. Tobit selection is a variation on probit selection that uses censoring of a normal variable as an indicator of selection.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Is sample selection a concern in your research problem?

Say that you wish to fit the model

```
. eregress y x1 x2
```

We will tell you two stories about it. In the first, `y` is wage-and-salary income. In the second, `y` is a health outcome for people with a certain malady.

Both of these stories have issues of sample selection. Wages are observed only for people who work. Health outcomes are observed only for people with the malady who seek treatment. Do you care? You might not.

If you are an economist studying the effects of education, you might be perfectly satisfied measuring the return to schooling in terms of increased income of those who work. This would certainly be the situation if you were performing research to determine how schools could be improved.

If you are a medical researcher studying the effect of a treatment, you might be perfectly satisfied measuring the effect of the treatment on those who currently seek it. This would certainly be the situation if you were performing research to determine how the treatment could be improved.

Sample selection is of concern only when changing the selected population—those who work or those who are treated—is under consideration.

The problem and solution of endogenous sample selection

We wish to fit the model

```
. eregress y x1 x2
```

We observe `y` for some of but not all the sample. We observe `x1` and `x2` for the entire sample.

For instance, we might be doing a study of a walking program run by hospitals for patients after heart attacks. Doctors prescribe the program to patients who they believe will benefit. After six months in the program, recorded for each patient is

y	Meaning
1	I feel worse (tired)
2	I feel about like I did when I started the program
3	I feel better

The variable `y` will be missing for some of the observations in the data. Those observations correspond to the patients who were not prescribed the program. `y` could also be missing if patients were prescribed but dropped out of the program—were lost to follow-up—but we will ignore that right now. We will discuss lost to follow-up in [ERM] **Intro 5**.

Variable `y` is an ordinal variable, so rather than fitting the model by using `eregress`, we will fit it by using `eoprobit`:

```
. eoprobit y x1 x2
```

Do not type that command yet. If you did, the model would be fit using only the observations on patients who were prescribed the program, because `y` is missing otherwise. We are about to discuss those other patients. In fact, let's create a variable indicating whether patients were selected for inclusion in the program—we will need it later.

```
. generate selected = !missing(y)
```

There are two types of sample selection: exogenous and endogenous. Hardly any issues are created by exogenous sample selection. The real problems are raised by endogenous selection, and to discuss those issues, we need to tell you more about the walking program.

Doctors prescribed the program to their patients based on each patient's `x1` and `x2` values. Those variables are believed to predict how much a patient would benefit from the program. Indeed, patients in especially poor health might actually be harmed by the program. Say that we are conducting research to evaluate how well `x1` and `x2` predict a benefit and to consider whether the criteria for being prescribed the program should be loosened or tightened. Would extending the program to more patients be beneficial? Or is the program already being used by too many?

That the sample was selected on `x1` and `x2` causes no statistical issues, although it can cause complications. Assume that doctors also based their decisions on `x3` but that was for administrative reasons. That sounds horrible, but it is not necessarily bad; for example, if a patient lives far from the hospital, the doctor might not prescribe the hospital's walking program as readily. In any case, `x3`, the distance a patient lives from the hospital, affected the decision but is not believed to affect how beneficial the program is for the patient. If we are certain about that, we can ignore `x3`. If we are uncertain, we should add `x3` to the model to verify that the effect really is 0.

The above situation is called exogenous sample selection. It is not a reasonable story, but perhaps you do not yet see why. Anyway, if the only problem is exogenous sample selection, we can ignore it, and the only issue we have is to decide whether to include `x3` in our model. We can fit the model by typing

```
. eoprobit y x1 x2
```

or

```
. eoprobit y x1 x2 x3
```

Typing those commands is equivalent to typing

```
. eoprobit y x1 x2 if selected
```

or

```
. eoprobit y x1 x2 x3 if selected
```

We mention this merely to emphasize that because `y` is missing in the group for which `selected` is 0, all observations for which `selected` is 0 are omitted from the estimation subsample.

The problem with the above story is that doctors know more about their patients than we do. They know more than what is recorded in our database. Doctors meet with their patients and get to know them, and doctors factor everything they know into their decisions. Doctors prescribed the walking program to patients who they believed would benefit. They predicted the benefit on the basis of `x1`, `x2`, and `x3`, as well as on information they know about the patients that is not recorded in the data.

Think of the decision that doctors make as a probit model:

$$\begin{aligned} p &= \Pr(\text{prescribed}) \\ &= \Pr(\beta_0 + \beta_1 \texttt{x1}_i + \beta_2 \texttt{x2}_i + \beta_3 \texttt{x3}_i + e_i.\texttt{selected} > 0) \end{aligned}$$

The important part of this model is $e.\texttt{selected}$. The error includes everything doctors know about their patients that is not recorded in the data. Because doctors presumably are making decisions in the patients' best interest, $e.\texttt{selected}$ will be correlated positively with $e.\texttt{y}$, which is the error in the model's main equation fit by

```
. eoprobit y x1 x2
```

If we fit the model ignoring this correlation, we would obtain results suitable for predicting outcomes among those who participated in the program but not among those who did not participate.

It is the nonzero correlation of $e.\texttt{y}$ and $e.\texttt{selected}$ that makes the sample-selection endogenous. `eoprobit` will produce estimates accounting for the correlation if we specify the `select()` option:

```
. eoprobit y x1 x2, select(selected = x1 x2 x3)
```

`eoprobit` will report $\widehat{\rho}$—the estimate of the correlation between the two errors—and it will report the coefficients in the outcome and selection models. Because we have now accounted for the endogenous sample selection, we can interpret the results in terms of the full population, not just those who were prescribed the treatment.

Endogenous sample selection handles missing not at random

`select()` can handle cases in which data are missing not at random (MNAR), also known as nonignorable missing data. It can handle them as long as that missingness is modeled in the `select()` equation. It can solve the problem of missing on unobservables.

Endogenous sample selection can be used with other features of ERMs

You can use `select()` with other features of ERMs, that is, with endogenous covariates, with treatment effects, and with observations that are correlated within panels or within groups. We have not discussed treatment effects or within-panel correlation yet. We will get to those in [ERM] **Intro 5** and [ERM] **Intro 6**.

In the meantime, we will show you one way that `endogenous()` can be used with `select()`. Above, we fit the model

```
. eoprobit y x1 x2, select(selected = x1 x2 x3)
```

In the story we told, `x3` measured an administrative reason we think affected doctors' decisions to prescribe the walking program. Let's imagine that `x3` was endogenous for one reason or another. In the original story, `x3` was the distance a patient lived from the hospital. Perhaps its value is measured with error. Or perhaps `x3` represents some other administrative reason we think is correlated with `y`. Because it is endogenous, we will now refer to this variable as `w3` instead of `x3`. We can address the problem by using the `endogenous()` option:

```
. eoprobit y x1 x2, select(selected = x1 x2 w3) endogenous(w3 = z1 z2, nomain)
```

We included suboption `nomain` because we do not want `w3` to be added to the main equation. `w3` appears only in the selection equation in this model.

Be careful not to omit `nomain` when it is necessary. Endogenous covariates can appear in the main equation, the selection equation, or both. Consider another example in which `x3` is not endogenous but `x2` is. Let's call it `w2` instead of `x2`. We could fit that model by typing

```
. eoprobit y x1, select(selected = x1 w2 x3) endogenous(w2 = z3 z4)
```

`w2` will appear in the main equation because we did not also specify `nomain`. Some users always type `nomain` and explicitly specify all the covariates that appear in the main equation. You could fit the same model by typing

```
. eoprobit y x1 w2, select(selected = x1 w2 x3) endogenous(w2 = z3 z4, nomain)
```

Mechanical notes

When you specify

```
. eoprobit y ..., select(selected = ...)
```

you can specify variables in just the `y` equation, just the `selected` equation, or both. When the same variables are specified in both equations, it is called functional-form identification. Statistically speaking, the situation would be better if there were some covariates that appeared in the `selected` equation that did not appear in the main equation, but no one is suggesting that you add irrelevant covariates to your model. Still, you should think about whether you have any such variables. We found such a variable (`x3`) in the story above.

Video example

Extended regression models: Endogenous sample selection

Also see

[ERM] **Intro 9** — Conceptual introduction via worked example