

Useful Commands in Stata

● Two-Stage Least Squares

- The structural form: $Y1 = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$
- The reduced form: $Y2 = \gamma_0 + \gamma_1 X1 + \gamma_2 X3 + \gamma_3 X4$

.reg Y1 Y2 X1 X2 X3 (X1 X3 X4)

- Check endogeneity: two ways

1) Hausman test

.reg Y1 Y2 X1 X2 X3 → obtain the coefficient(C1) and the s.e.(S1) of Y2

.reg Y1 Y2 X1 X2 X3 (X1 X3 X4) → obtain the coefficient(C2) and the s.e.(S2) of Y2

.di (C2 - C1)/sqrt(S2 - S1) → Based on this t-statistic, decide the endogeneity

2)

.reg Y2 X1 X3 X4

.predict v2hat, resid

.reg Y1 Y2 X1 X2 X3 v2hat

→ Based on the significance of the coefficient of v2hat, the endogeneity problem.

- Check overidentification problem → look at W1's class notes.

● Bivariate Probit

: biprobit estimates maximum-likelihood two-equation probit models -- either a bivariate probit or a seemingly unrelated probit (limited to two equations).

- First equation: $Y1 = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$

- Second equation: $Y2 = \gamma_0 + \gamma_1 X1 + \gamma_2 X2 + \gamma_3 X3$

.biprobit Y1 Y2 X1 X2 X3, robust

● Seemingly Unrelated Bivariate Probit Model

- First equation: $Y1 = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$

- Second equation: $Y2 = \gamma_0 + \gamma_1 X1 + \gamma_2 X2 + \gamma_3 X3$

.biprobit (Y1 = X1 X2 X3) (Y2 = X1 X3), robust

● Recursive Bivariate Probit model

- First equation: $Y1 = \beta_0 + \beta_1 X1 + \beta_2 X2$

- Second equation: $Y2 = \gamma_0 + \gamma_1 X1 + \gamma_2 X2 + \gamma_3 X3$

.biprobit (Y2 = X1 X2 X3) (Y1 = Y2 X1 X2), robust

- Obtain marginal effects

- a. Direct effect in the second equation (if X1 is dummy):

- mfx compute, at(mean Y2=1, X1=1) predict(pmarg1)*

- ➔ To save time, use “nose” command if you don’t need to get S.E.

- b. Direct/Indirect effects in the first equation: Since the way of obtaining these effects by using “mfx” commands is not clear, I recommend to read the Stata manual or write your own “do-file”. Also refer to Greene’s “*Economic Analysis*” and his article, “*Gender Economics Courses in Liberal Arts Colleges: Comment*”, from his web site.

- **Two-Stage Probit least Squares**

- : A simultaneous equation model in which one of the endogenous variables is continuous and the other is binary.

- First equation: $Y1 \text{ (binary)} = Y2 X1 X2 X3$

- Second equation: $Y2 \text{ (continuous)} = Y1 X1 X2 X4 X5$

- . cdsimeq (Y2 X1 X2 X3) (Y1 X1 X2 X4 X5)*

- **Logit or Probit estimation on grouped data**

- : blogit and bprobit produce maximum-likelihood logit and probit estimates on grouped (“blocked”) data; glogit and gprobit produce weighted least-squares estimates. In the syntax diagrams, pos_var and pop_var refer to variables containing the total number of positive responses and the total population.

- blogit deaths pop agecat exposed*

- blogit deaths pop agecat exposed, or*

- bprobit deaths pop agecat exposed*

- predict number*

- predict p if e(sample), p*

- glogit deaths pop agecat expose*

- glogit deaths pop agecat expose, or*

● Heckman Selection Model

: This method assumes that missing values of the dependent variable imply that the dependent variable is unobserved (not selected). Thus, it is a good way of predicting the value of the dependent variable that would be observed in the absence of selection (considering the missing values). If a data set specifies a binary variable that identifies the observations for which the dependent is observed/selected or not observed, it is much convenient to run this model. The heckman selection model assumes that there exists an underlying regression relationship

$$y_j = x_j\beta + u_{1j} \quad \text{--- regression equation ---}$$

The dependent variable, however, is not always observed. Rather, the dependent variable for observation j is observed if

$$z_j\gamma + u_{2j} > 0 \quad \text{--- selection equation ---}$$

, where

$$u_1 \sim N(0, \sigma)$$

$$u_2 \sim N(0, 1)$$

$$\text{corr}(u_1, u_2) = \rho$$

When $\rho \neq 0$, standard regression techniques applied to the first equation yield biased results. Heckman provides consistent, asymptotically efficient estimates for all the parameters in such models.

■ Example

If we assume that the hourly wage is a function of education and age whereas the likelihood of working (the likelihood of the wage being observed) is a function of marital status, number of children at home, and (implicitly) the wage.

Heckman wage educ age, select(married children educ age) cluster(county)

Heckman wage educ age, select(married children educ age) twostep

● Marginal Effects (partial change) in probit

: Probit magnitudes are hard to interpret. So use “dprobit” to get partial effects on response probabilities. “dprobit” also estimates maximum-likelihood probit models. Rather than reporting coefficients, dprobit reports the change in the probability for an infinitesimal change in each independent, continuous variable and, by default, the discrete change in the probability for dummy variables. probit may be typed without arguments after dprobit estimation to see the model in coefficient form.

.dprobit Y X1 X2 X3, robust

→ It gives you the estimated effect of an independent variable at the mean values of the covariates and the discrete change of dummy variable from 0 to 1.

.matrix x= (2, 2, 3, 3, 1, 2)

.matrix colnames x = X1 X2 X3 X4 X5 X6

dprobit Y X1 X2 X3 X4 X5 X6, at(x) robust

→ If we vary X5 from 1 to 2,

matrix x= (2, 2, 3, 3, 2, 2)

matrix colnames x = X1 X2 X3 X4 X5 X6

dprobit Y X1 X2 X3 X4 X5 X6, at(x) robust

→ We can run the logit model and multinomial logit model too by substituting “dprobit” with “dlogit2” and “dmlogit2” commands.

● **Ordered Logit**

: When the dependent variable is ordinal, its categories can be ranked from low to high, but the distances between adjacent categories are unknown. If the intervals between adjacent categories are equal, the linear regression model can be used. In this model, we assume that the form of the error distribution is a logistic distribution with a mean of zero and a variance of $\pi^2/3$ (3.29).

ologit

Cf. In the ordered probit model, we assume the error term is distributed normally with mean zero and variance one.

● **Nested Logistic Estimation (nlogit)**

: nlogit estimates a nested logit model using full maximum-likelihood. The model may contain one or more levels. For a single-level model, nlogit estimates the same model as clogit.

■ **Examples**

Generate a new categorical variable named "type" that identifies the first-level set of alternatives based on the variable named "restaurant".

***. nlogitgen type = restaurant(fast: Freebirds | MamasPizza, family: CafeEccell |
LosNortenos | WingsNmore, fancy: Christophers | MadCows)***

The tree structure implied by these two variables can be examined using `nlogittree`.

. nlogittree restaurant type

Estimate the nested logit model.

*. nlogit chosen (restaurant = cost rating distance) (type = incFast incFancy
kidFast kidFancy), group(family_id)*

*. nlogit chosen (restaurant = cost rating distance) (type = incFast incFancy
kidFast kidFancy), group(family_id) ivc(fast=1, family=1, fancy=1)*

● **Conditional (fixed effects) Logistic Model (clogit)**

: `clogit` estimates what biostatisticians and epidemiologists call conditional logistic regression for matched case-control groups and what economists and other social scientists call fixed-effects logit for panel data. It also estimates McFadden's choice model. Computationally, these models are the same.

Example: matched case-control data

. clogit low lwt nonwhite smoke ptd, group(pairid)
. clogit low lwt nonwhite smoke ptd, group(groupid)
. clogit low lwt nonwhite smoke ptd, group(groupid) or

Example: Fixed-effects logit

. clogit union age grade not_smsa, group(idcode)

Example: McFadden's choice model

. clogit promoted tenureb fracatt female, group(poolid)

Example: predictions in 1:k matched data

. clogit y x1 x2 x3, group(id)
. predict phat if e(sample)
. predict idx, xb

● **“prvalue” and “prchange”**

: `prchange` computes discrete and marginal change for regression models for categorical and count variables. **Marginal change** is the partial derivative of the predicted probability or predicted rate with respect to the independent variables. **Discrete change** is the difference in the predicted value as one independent variable changes values while all others are held constant at specified values.

By default, the discrete and marginal change is calculated holding all other variables at their mean. Values for specific independent variables can be set using the `x()` option after `prchange`. For example, to compute predicted values when `educ` is 10 and `age` is 30, type `prchange, x(educ 10 age 30)`. Values for the unspecified independent variables can be set using the `rest()` option, e.g., `prchange, x(educ 10 age 30) rest(mean)`. The `if` and `in` conditions specify conditions for computation of means, min, etc., that are used with `rest()`.

The discrete change is computed when a variable changes from its minimum to its maximum (Min->Max), from 0 to 1 (0->1), from its specified value minus .5 units to its specified value plus .5 (-+1/2), and from its specified value minus .5 standard deviations to its value plus .5 standard deviations (-+sd/2).

`prchange` works with **`clolog`, `cnreg`, `gologit`, `intreg`, `logistic`, `logit`, `mlogit`, `nbreg`, `ologit`, `oprobit`, `poisson`, `probit`, `regress`, `tobit`, `zinb`, and `zip`**.

Examples

To compute discrete and marginal change for ordered probit at base values equal to the mean:

```
. oprobit warm yr89 male white age ed prst  
. prchange
```

To compute discrete and marginal change for a specific set of base values:

```
. prchange, x(male 1 white 1 age 30)  
. prchange, x(male 1 white 1 age 30) rest(median)  
. prchange, x(male 1 white max age min) rest(median)
```

To compute the discrete change for a set of base values while holding other variables constant to a group-specific median:

```
. prchange if white == 1 & male == 0, x(white 1 male 0) rest(median)
```

To list only the discrete change for specific variables:

```
. prchange, x(male 0 white 0) rest(median) list(age)
```