Dis 9: Binary Response; Panel Data; Instrumental Variable

1 Binary Response

1. Load the dataset from http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.dta into Stata (don't forget to first change your working directory).

   Dataset codebook is available at http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des

   In this dataset, `inlf` is a binary variable recording a married woman's decision on joining the labor force. Consider the following explanatory variables to enter the model:

   - Whether the respondent lives in a city (`city`)
   - Family's income level (`faminc`)
   - Unemployment rate in the county resided in (`unem`)
   - Number of kids less than 6 years old (`kidslt6`)
   - Number of kids age between 6 and 18 (`kidsge6`)
   - The respondent's age (`age`)
   - The respondent's years of education (`educ`)

   (a) Run a linear probability model.

   (b) Plot residual against predicted value of `inlf`. Does the error term look homoskedatic or heteroskedastic? Make appropriate adjustment in your estimation in (a) if needed.

   (c) What's the marginal effect of the number of kids less than 6 years old (`kidslt6`) on the decision of joining the labor force?

   (d) What's the drawback of linear probability model?

(e) Run a logit model.

(f) What's the marginal effect of the number of kids less than 6 years old (`kidslt6`) on the decision of joining the labor force when variables are at mean level? Write down the expression and then calculate using Stata.

(g) What's the predicted probability of living in a city vs. not living in a city on participating in labor force for a respondent with the following characteristics?

- Have $25,000 family income
- Have `unem`= 10
- Have 1 kid age less than 6
- Have 1 kid age between 6 and 18
- Age 30
- Have 16 years of education

(h) Add husband's hourly wage (`huswage`) and husband's age (`husage`) to your regression model. Use a likelihood ratio Chi-square test to determine whether the two added coefficients are jointly statistically significant.

(i) Run a probit model (without the added regressors mentioned in (h)).

(j) What's the marginal effect of the number of kids less than 6 years old (`kidslt6`) on the decision of joining the labor force when variables are at mean level? Calculate by Stata.

2 Panel Data

- Consider the following regression model:

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{it} + \varepsilon_{it} \\ &= \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it} \end{aligned}$$

  where

  - $i = 1, 2, \ldots, n$ is the index for each individual
  - $t = 1, 2, \ldots, T$ is the index for time
  - α_i is the time invariant error component
  - u_{it} is the time variant i.i.d. error component

- Model assumptions:

  1. **Zero conditional mean**: $E[u_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}, \alpha_i] = 0$
  2. **I.I.D. Draws**: $(X_{i1}, X_{i2}, \ldots, X_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT})$ for $i = 1, 2, \ldots, n$ are i.i.d. draws from their joint distribution
  3. **Large outliers are unlikely**: (X_{it}, u_{it}) have nonzero finite fourth moments (another way of saying that there's no data point that lives in a dramatically different region)
  4. **No perfect multicollinearity**: One of the regressors cannot be a perfect linear function of the other regressors.

- We'll discuss two scenarios for panel data:

  1. α_i is uncorrelated with X_{it}
  2. α_i is correlated with X_{it}

2.1 α_i uncorrelated with X_{it}: standard error adjustment

- From Dis 6, we learned that the structure of the error term matters for the standard error of $\hat{\beta}$s
- In our model $Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$, should we expect ε_{it} to be heteroskedastic?

  - Since $\varepsilon_{it} = \alpha_i + u_{it}$, for every individual i, there's a time-invariant component α_i that always exist for ε_{it}
  - This means that the error term is serially correlated
  - Solution when running OLS? Use **clustered standard error**

    This is the idea of having standard error cluster by group. For example, for individual i, the error terms ε_{it} across t won't be considered as i.i.d., so we allow for the possibility of serial correlation. $\Rightarrow$ Difficult to calculate by hand, but very easy to adjust for in Stata.

- Rule of thumb:

  - Clustered standard error is used when α_i is uncorrelated with X_{it}. OLS is still the regression technique; clustering standard errors simply correct the standard error calculations for $\hat{\beta}$ estimators.
  - In general (though not always true),

$$\text{homoskedatic standard error} < \text{robust standard error} < \text{clustered standard error}$$

2.2 α_i correlated with X_{it}: fixed effects and first differences

- One way to approach **omitted variable bias**, in the specific case that the omitted variable is invariant (either over time or over state)
- Two ways of estimating the parameter of interest β_1 in the model:

Method 1: Fixed Effects Regression (more commonly used in practice)

Take the average for individual i across time t:

$$\overline{Y_i} = \beta_0 + \beta_1 \overline{X_i} + \alpha_i + \overline{u_i}$$

and then subtract $\overline{Y_i}$ from Y_{it}:

$$\begin{aligned} Y_{it} - \overline{Y_i} &= (\beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}) - (\beta_0 + \beta_1 \overline{X_i} + \alpha_i + \overline{u_i}) \\ &= \beta_1 (X_{it} - \overline{X_i}) + (u_{it} - \overline{u_i}) \end{aligned}$$

Given that (X_{it}, u_{it}) for all i, t are i.i.d., we have that $(X_{it} - \overline{X_i}, u_{it} - \overline{u_i})$ are still i.i.d.

Thus, we can regress $Y_{it} - \overline{Y_i}$ on $X_{it} - \overline{X_i}$ with no constant term to obtain the estimate of parameter of interest: $\hat{\beta}_1$.

Method 2: First Differences Regression (less common)

Use our model, write down how last period's relationship looks like:

$$Y_{i,t-1} = \beta_0 + \beta_1 X_{i,t-1} + \alpha_i + u_{i,t-1}$$

and then subtract $Y_{i,t-1}$ from Y_{it}:

$$\begin{aligned} Y_{it} - Y_{i,t-1} &= (\beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}) - (\beta_0 + \beta_1 X_{i,t-1} + \alpha_i + u_{i,t-1}) \\ &= \beta_1 (X_{it} - X_{i,t-1}) + (u_{it} - u_{i,t-1}) \end{aligned}$$

Given that (X_{it}, u_{it}) for all i, t are i.i.d., we have that $(X_{it} - X_{i,t-1}, u_{it} - u_{i,t-1})$ are still i.i.d.

Thus, we can regress $Y_{it} - Y_{i,t-1}$ on $X_{it} - X_{i,t-1}$ with no constant term to obtain the estimate of parameter of interest: $\hat{\beta}_1$.

2.3 Doing things in Stata

- When α_i is uncorrelated with X_{it}:

```
reg Y X, cluster(id_for_i)
// id_for_i is the name of the variable recording index for individual
```

- When α_i is correlated with X_{it}:

Method 1: Fixed Effects Regression

- The direct approach:

```
egen mean_Y = mean(Y), by(id_for_i)
// id_for_i is the name of the variable recording index for individual
egen mean_X = mean(X), by(id_for_i)
gen Y_diff = Y - mean_Y
gen X_diff = X - mean_X
reg Y_diff X_diff, noconstant
```

- The one-line approach:

```
xtreg Y X, i(id_for_i) fe
```

```
. xtreg pris polpc, i(state) fe

Fixed-effects (within) regression               Number of obs     =        714
Group variable: state                           Number of groups  =         51

R-sq:                                           Obs per group:
     within  = 0.3558                                         min =         14
     between = 0.5880                                         avg =       14.0
     overall = 0.5096                                         max =         14

                                                F(1,662)          =     365.66
corr(u_i, Xb)  = -0.5977                        Prob > F          =     0.0000

------------------------------------------------------------------------------
        pris |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       polpc |   1.730557   .0904999    19.12   0.000     1.552855    1.908258
       _cons |  -263.4766   24.33375   -10.83   0.000    -311.2572    -215.696
-------------+----------------------------------------------------------------
     sigma_u |  93.444658
     sigma_e |   58.04163
         rho |  .72160114   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(50, 662) = 23.32                    Prob > F = 0.0000
```

- Another one-line approach:

```
reg Y X i.id_for_i
```

Adding dummy variable for each individual accounts for the fixed component α_i.

Method 2: First Differences Regression

- The only approach (that I'm aware of):

```
xtset id_for_i id_for_t
// id_for_i is the name of the variable recording index for individual
// id_for_t is the name of the variable recording index for time
gen lag_Y = l.Y
// l. syntax here creates the lag variable (i.e. the t-1 value)
gen lag_X = l.X
gen Y_diff = Y - lag_Y
```

```
gen X_diff = X - lag_X
reg Y_diff X_diff, noconstant
```

*If you believe that u_{it} also correlates with X_{it} (instead of being independent), then clustered standard error can be used to adjust for that. Simply adding the `cluster(id_for_i)` option at the end of the fixed effects or first differences regression.

3 Instrumental Variable (IV)

3.1 IV overview

- Another method to approach **omitted variable bias** (and worries about **simultaneous causality**)
- What's the problem in the first place? $\rightarrow$ some variables might be **endogenous**!
  - **Exogenous variable**: Variable that is uncorrelated with the error.
  - **Endogenous variable**: Variable that correlates with the error term.

    Example. In a demand model where we regress quantity on price and product characteristics, price is considered as an endogenous variable, since price is correlated with unobserved demand shocks, and the unobserved demand shock is part of the error term that affects quantity sold.

    Example. In a return-to-schooling model where we regress average hourly earnings on years of education, years of education is endogenous, since years of education is correlated with unobserved ability, and the unobserved ability is part of the error term that affects hourly earnings.
- How IV addresses the endogeneity problem?
  - Choice of IV: two conditions for the IV Z on endogenous variable X
    1. **Relevance**: IV is correlated with X; that is, $Cov(Z, X) \neq 0$.
    2. **Exogeneity**: IV is uncorrelated with the error term in the model; that is, $Cov(Z, u) = 0$.
  - How IV affects β estimate?

    Consider the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

    here, X_i is endogenous, and Z_i is a valid instrument (valid = relevant + exogenous).

    Comparing the $\hat{\beta}_1$ estimates:

$$\hat{\beta}_{1,OLS} = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$
$$\hat{\beta}_{1,IV} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, X_i)} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}$$

    Notice that

$$\hat{\beta}_{1,OLS} = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)} = \frac{\widehat{Cov}(X_i, \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i)}{\widehat{Var}(X_i)}$$

$$= \beta_1 + \frac{\widehat{Cov}(X_i, u_i)}{\widehat{Var}(X_i)} \quad \xrightarrow{p} \quad \beta_1 + \frac{\overbrace{Cov(X_i, u_i)}^{\neq 0}}{Var(X_i)} = \beta_1 + \text{bias}$$

$$\hat{\beta}_{1,IV} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, X_i)} = \frac{\widehat{Cov}(Z_i, \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i)}{\widehat{Cov}(Z_i, X_i)}$$

$$= \beta_1 + \frac{\widehat{Cov}(Z_i, u_i)}{\widehat{Cov}(Z_i, X_i)} \quad \xrightarrow{p} \quad \beta_1 + \frac{\overbrace{Cov(Z_i, u_i)}^{=0}}{Cov(Z_i, X_i)} = \beta_1$$

Thus, when X_i is endogenous, using Z_i as an instrument yields consistent estimate!

> Side note: Unfortunately, $\hat{\beta}_{1,IV}$ often isn't an unbiased estimator (that is, $E[\hat{\beta}_{1,IV}] \neq \beta_1$). But it is consistent (that is, when sample size n grows Large, $\hat{\beta}_{1,IV} \xrightarrow{p} \beta_1$).
> IV estimator is nonetheless better than plain OLS estimator (the plain OLS estimator doesn't account of the endogeneity issue at all).

- How to perform IV estimation in practice: **Two Stage Least Squares (2SLS / TSLS)**
  - Say that our model is the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Here, X_1 is an endogenous variable, and we have two instruments Z_1, Z_2 that are both valid. X_2 and X_3 are exogenous variables.
2SLS is carried out in the following way:

1. **First stage**: Regress the endogenous variable on instruments and exogenous variables. Obtain predicted value of the endogenous variable.
   In other words, run the following regression:

$$X_{1i} = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + \pi_3 X_{2i} + \pi_4 X_{3i} + v_i$$

   And then obtain the predicted value $\hat{X}_1$:

$$\hat{X}_{1i} = \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + \hat{\pi}_2 Z_{2i} + \hat{\pi}_3 X_{2i} + \hat{\pi}_4 X_{3i}$$

2. **Second stage**: Regress Y on the predicted endogenous variable from first stage, along with other exogenous variables.
   In other words, run the following regression:

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

  - Why does this two-stage approach work? (Equivalently, what does the first stage do?)
    * The first stage obtains predicted X_1 from our instruments and other exogenous variables
      $\Rightarrow$ Instruments and the other exogenous variables are all uncorrelated with the original model's error term u
      $\Rightarrow$ This predicted X_1 is constructed by things uncorrelated with u

$\Rightarrow$ We are eliminating the endogenous component of X_1 (getting rid of part of X_1 that correlates with the error u)

* Can we only include the IVs in the first stage but not the other exogenous variables?
$\Rightarrow$ NO! Since X_1 very much could be correlated with the other Xs, so if we didn't include the other exogenous Xs in the first stage, then the predicted X_1 obtained has less exogenous component than it would have had.

3.2 Doing things in Stata

- Performing 2SLS directly:

```
reg X1 Z1 Z2 X2 X3
predict X1_hat
reg Y X1_hat X2 X3
```

Though in this way, the standard error estimates cannot be trusted, since running the second stage separately does not account for the fact that `X1_hat` is estimated from the first stage.

- Performing 2SLS in one line:

```
ivregress 2sls Y (X1 = Z1 Z2) X2 X3
```

```
. ivregress 2sls Y (X1 = Z1 Z2) X2 X3

Instrumental variables (2SLS) regression          Number of obs   =        935
                                                  Wald chi2(3)    =     134.57
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0607
                                                  Root MSE        =     .40794
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
X1	.141278	.0141604	9.98	0.000	.113524	.1690319
X2	.0311679	.0046156	6.75	0.000	.0221216	.0402143
X3	.0110878	.0027553	4.02	0.000	.0056876	.0164881
_cons	4.435582	.2296195	19.32	0.000	3.985536	4.885628

```
Instrumented:  X1
Instruments:   X2 X3 Z1 Z2
```