# 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.

   A linear probability model is the same type of regression that we have been running, with the only exception that the dependent variable is now binary instead of a continuous variable.

   The model looks like the following:

$$\begin{aligned}\text{inlf}_i = \beta_0 + \beta_1\text{city}_i + \beta_2\text{faminc}_i + \beta_3\text{unem}_i + \beta_4\text{kidslt6}_i + \beta_5\text{kidsge6}_i \\ + \beta_6\text{age}_i + \beta_7\text{educ}_i + u_i\end{aligned}$$

   To run this in Stata, use the following command after loading the dataset:

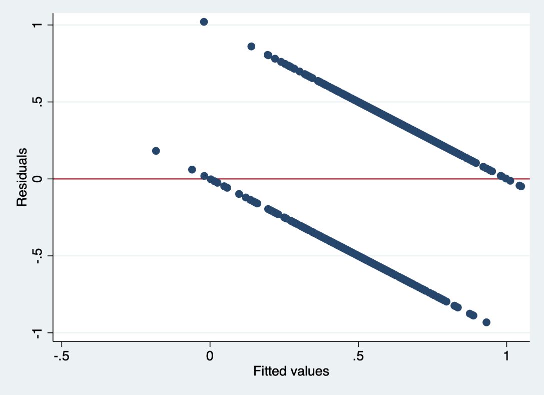
```
reg inlf i.city faminc unem kidslt6 kidsge6 age educ
```

   (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.

   Use the following command that we learned from Dis 6 to plot residual vs predicted $y$ plot:

```
rvfplot, yline(0)
```

   The graph looks like the following:



This pattern of residual suggests that the error is heteroskedastic instead of homoskedatic (refer to Dis 6 if you don't remember why this is the conclusion we draw), so **when running linear probability models, we always add the robust option to adjust for heteroskedastic error**.

The command in Stata is thus

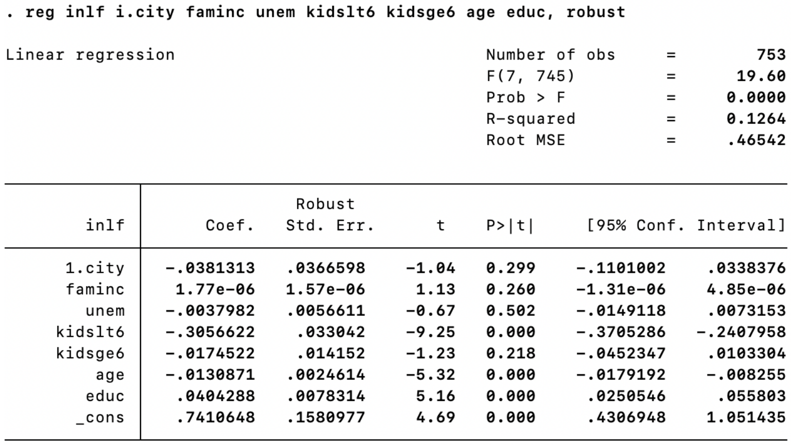
```
reg inlf i.city faminc unem kidslt6 kidsge6 age educ, robust
```

(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?

A nice feature about the linear probability model is that the coefficients estimated can be directly interpreted as the marginal effects. This means that

$$\frac{\partial \text{inlf}}{\partial \text{kidslt6}} = \beta_4$$

After accounting for robust standard error, our linear probability model's estimation becomes



```
. reg inlf i.city faminc unem kidslt6 kidsge6 age educ, robust

Linear regression                               Number of obs     =        753
                                                F(7, 745)         =      19.60
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1264
                                                Root MSE          =     .46542

------------------------------------------------------------------------------
             |               Robust
        inlf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.city |  -.0381313   .0366598    -1.04   0.299    -.1101002    .0338376
      faminc |   1.77e-06   1.57e-06     1.13   0.260    -1.31e-06    4.85e-06
        unem |  -.0037982   .0056611    -0.67   0.502    -.0149118    .0073153
     kidslt6 |  -.3056622    .033042    -9.25   0.000    -.3705286   -.2407958
     kidsge6 |  -.0174522    .014152    -1.23   0.218    -.0452347    .0103304
         age |  -.0130871   .0024614    -5.32   0.000    -.0179192    -.008255
        educ |   .0404288   .0078314     5.16   0.000     .0250546     .055803
       _cons |   .7410648   .1580977     4.69   0.000     .4306948    1.051435
------------------------------------------------------------------------------
```

The coefficient on `kidslt6`, $\beta_4$, is estimated to be -0.306. Thus, having one more kid less 6 years old decreases the probability of the wife joining the labor force by 0.306.

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

The main drawback of the linear probability model is that it doesn't restrict the predicted $y$ (i.e. the predicted probability) to be within 0 and 1.

For very small and very large values of the explanatory variables, this could cause the predicted probability to fall outside of the 0 to 1 range, resulting in interpretation problem.

(e) Run a logit model.

A logit model addresses the issue mentioned in (d) by strictly restricting the probability to be within 0 and 1. To do so, we use the logistic function $G(\cdot)$ to wrap around the explanatory variables, where $G(\cdot)$ has the nice property where its range always lie between 0 and 1:

$$\begin{aligned}\text{inlf}_i = G(&\beta_0 + \beta_1 \text{city}_i + \beta_2 \text{faminc}_i + \beta_3 \text{unem}_i + \beta_4 \text{kidslt6}_i + \beta_5 \text{kidsge6}_i \\ &+ \beta_6 \text{age}_i + \beta_7 \text{educ}_i) + u_i\end{aligned}$$

where

$$G(z) = \frac{\exp(z)}{1+\exp(z)}$$

Running this in Stata:

```
logit inlf i.city faminc unem kidslt6 kidsge6 age educ
```

(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.

The complicated bit about logit model is that now the coefficients are no longer the marginal effects, since now the explanatory variables are wrapped around the $G(\cdot)$ function. So one needs to use the chain rule when differentiating out the marginal effects.

Expression:

Denote $K \equiv \beta_0 + \beta_1 \text{city}_i + \beta_2 \text{faminc}_i + \beta_3 \text{unem}_i + \beta_4 \text{kidslt6}_i + \beta_5 \text{kidsge6}_i + \beta_6 \text{age}_i + \beta_7 \text{educ}_i$.

$$\frac{\partial \text{inlf}}{\partial \text{kidslt6}} = \frac{\partial G(K)}{\partial K} \times \frac{\partial K}{\partial \text{kidslt6}}\Bigg|_{\text{evaluated at mean level of variables}}$$

$$= \left[\exp(K)(1+\exp(K))^{-1} - \exp(K)^2(1+\exp(K))^{-2}\right] \times \beta_4 \Bigg|_{\text{evaluated at mean level of variables}}$$

$$= \frac{\exp(\overline{K})(1+\exp(\overline{K}) - \exp(\overline{K}))}{(1+\exp(\overline{K}))^2} \times \beta_4 = \frac{\exp(\overline{K})}{(1+\exp(\overline{K}))^2} \times \beta_4$$

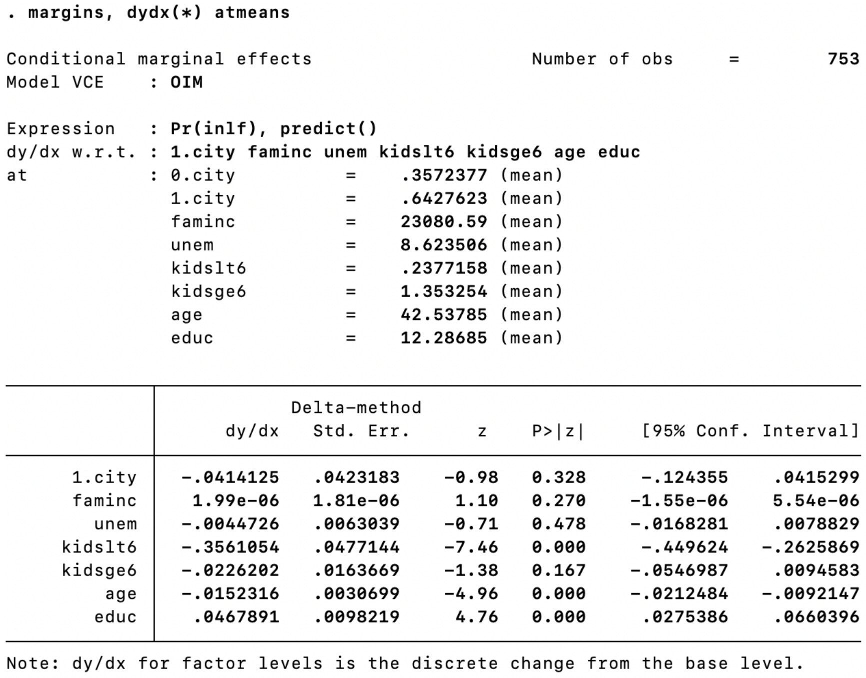
Here, $\overline{K} = \beta_0 + \beta_1 \overline{\text{city}} + \beta_2 \overline{\text{faminc}} + \beta_3 \overline{\text{unem}} + \beta_4 \overline{\text{kidslt6}} + \beta_5 \overline{\text{kidsge6}} + \beta_6 \overline{\text{age}} + \beta_7 \overline{\text{educ}}$. In estimation, replace all $\beta$ with $\hat{\beta}$.

Calculate using Stata:

The complicated math expression above can be easily calculated using Stata:

```
logit inlf i.city faminc unem kidslt6 kidsge6 age educ
margins, dydx(*) atmeans
```

The marginal effects are the following:



```
. margins, dydx(*) atmeans

Conditional marginal effects                    Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : 1.city faminc unem kidslt6 kidsge6 age educ
at           : 0.city          =    .3572377 (mean)
               1.city          =    .6427623 (mean)
               faminc          =    23080.59 (mean)
               unem            =    8.623506 (mean)
               kidslt6         =    .2377158 (mean)
               kidsge6         =    1.353254 (mean)
               age             =    42.53785 (mean)
               educ            =    12.28685 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.city |  -.0414125   .0423183    -0.98   0.328     -.124355    .0415299
      faminc |   1.99e-06   1.81e-06     1.10   0.270    -1.55e-06    5.54e-06
        unem |  -.0044726   .0063039    -0.71   0.478    -.0168281    .0078829
     kidslt6 |  -.3561054   .0477144    -7.46   0.000     -.449624   -.2625869
     kidsge6 |  -.0226202   .0163669    -1.38   0.167    -.0546987    .0094583
         age |  -.0152316   .0030699    -4.96   0.000    -.0212484   -.0092147
        educ |   .0467891   .0098219     4.76   0.000     .0275386    .0660396
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

The marginal effect estimate on `kidslt6` is -0.356. Thus, having one more kid less 6 years old decreases the probability of the wife joining the labor force by 0.356.

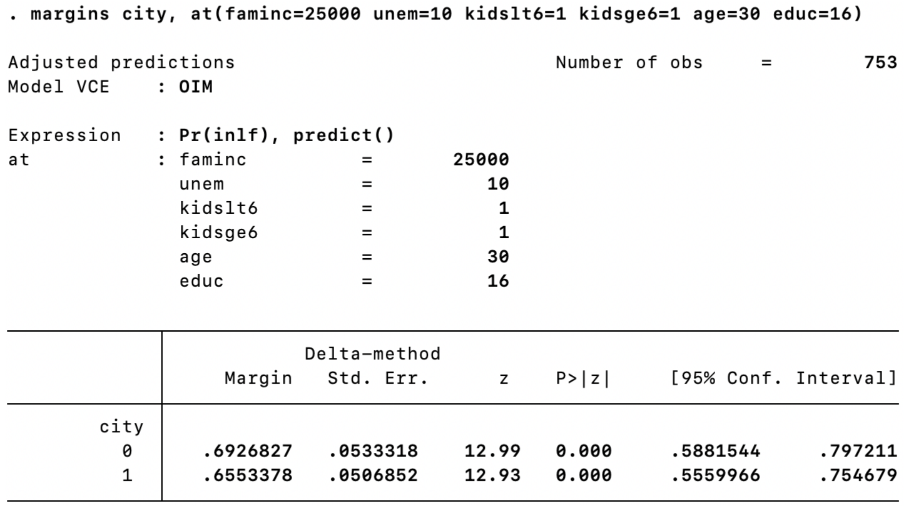
(g) What's the predicted probability of living in a city vs. not living in a city on participating in labor force for a respondent with the following characteristics?

- 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

When we want the predicted probability of the two levels in `city`, run the following command in Stata:

```
margins city, at(faminc=25000 unem=10 kidslt6=1 kidsge6=1 age=30 educ=16)
```

The output is the following:



```
. margins city, at(faminc=25000 unem=10 kidslt6=1 kidsge6=1 age=30 educ=16)

Adjusted predictions                            Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
at           : faminc          =       25000
               unem            =          10
               kidslt6         =           1
               kidsge6         =           1
               age             =          30
               educ            =          16

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        city |
          0  |   .6926827   .0533318    12.99   0.000     .5881544     .797211
          1  |   .6553378   .0506852    12.93   0.000     .5559966     .754679
------------------------------------------------------------------------------
```

Thus,

- For a married woman not living in a city (`city` = 0) who has the described characteristics (25000 family income, 10 county unemployment, 1 kid age less than 6 years old, 1 kid between 6 and 18, age 30, and have 16 years of education), the predicted probability of participating in labor force is 0.693.
- For a married woman living in a city (`city` = 1) who has the described characteristics, the predicted probability of participating in labor force is 0.655.

(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.

A likelihood ratio test tests the following hypothesis:

$$H_0 : \text{model parameters} = \text{parameters from (e)}$$
$$H_1 : \text{model parameters} = \text{parameters from (h)}$$

The test statistic is constructed by the likelihood ratio between the two models:

$$LR = \frac{\mathcal{L}_n(\text{parameters from (h)})}{\mathcal{L}_n(\text{parameters from (e)})}$$

which we can take the log version of it to obtain an equivalent log-likelihood test statistic:

$$\begin{aligned}&\log(\mathcal{L}_n(\text{parameters from (h)})) - \log(\mathcal{L}_n(\text{parameters from (e)}))\\ &= \ln(\text{parameters from (h)}) - \ln(\text{parameters from (e)})\end{aligned}$$

We transform it to log-likelihood numbers since log-likelihood is directly reported by Stata. (Notice that "ln" in here isn't the natural log, but rather the log-likelihood number!)
And often times, this statistic has a constant 2 multiplied in front to make this statistic follow a Chi-squared distribution. So the actual (and final) test statistic is

$$\text{Log-likelihood } LR = 2(\ln(\text{parameters from (h)}) - \ln(\text{parameters from (e)})) \quad \sim \quad \chi^2_q$$

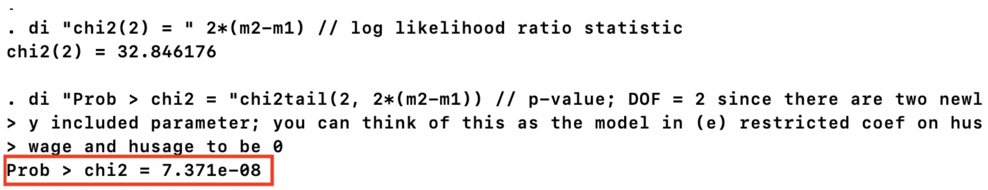
where $q$ is the number of restrictions. In this example, you can think about the restriction as model in (e) restricted coefficient on `huswage` and `husage` to be jointly 0, given that these two parameters are not included in (e). So $q = 2$ for this question.
To formally perform this test, we need the log-likelihood numbers from running logit command for model in (e) and model in (h). Then we can use Stata to construct p-value for interpretation. The Stata commands look like the following:

```
logit inlf i.city faminc unem kidslt6 kidsge6 age educ
scalar m1 = e(ll) // store log-likelihood level as m1

logit inlf i.city faminc unem kidslt6 kidsge6 age educ huswage husage
scalar m2 = e(ll) // store log-likelihood level of new model as m2

di "chi2(2) = " 2*(m2-m1) // log likelihood ratio statistic
di "Prob > chi2 = "chi2tail(2, 2*(m2-m1))
// p-value; the first argument in chi2tail(,) is q, the second is test statistic
```

The output in Stata is the following:



```
-
. di "chi2(2) = " 2*(m2-m1) // log likelihood ratio statistic
chi2(2) = 32.846176

. di "Prob > chi2 = "chi2tail(2, 2*(m2-m1)) // p-value; DOF = 2 since there are two newl
> y included parameter; you can think of this as the model in (e) restricted coef on hus
> wage and husage to be 0
Prob > chi2 = 7.371e-08
```

Since the p-value is less than 5% significance level, we conclude that we can reject the null hypothesis at 5% size, meaning that the alternative hypothesis (that the model parameters should be the one in (h)) is accepted. This suggests that the two added coefficients are jointly statistically significant.

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

A probit model is very similar to logit. It also restricts the probability to be within 0 and 1, just by using a different function. Probit uses the CDF function from a normal distribution, $\Phi(\cdot)$.

The probit model looks like the following:

$$\text{inlf}_i = \Phi(\beta_0 + \beta_1\text{city}_i + \beta_2\text{faminc}_i + \beta_3\text{unem}_i + \beta_4\text{kidslt6}_i + \beta_5\text{kidsge6}_i \\ + \beta_6\text{age}_i + \beta_7\text{educ}_i) + u_i$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Running this in Stata:

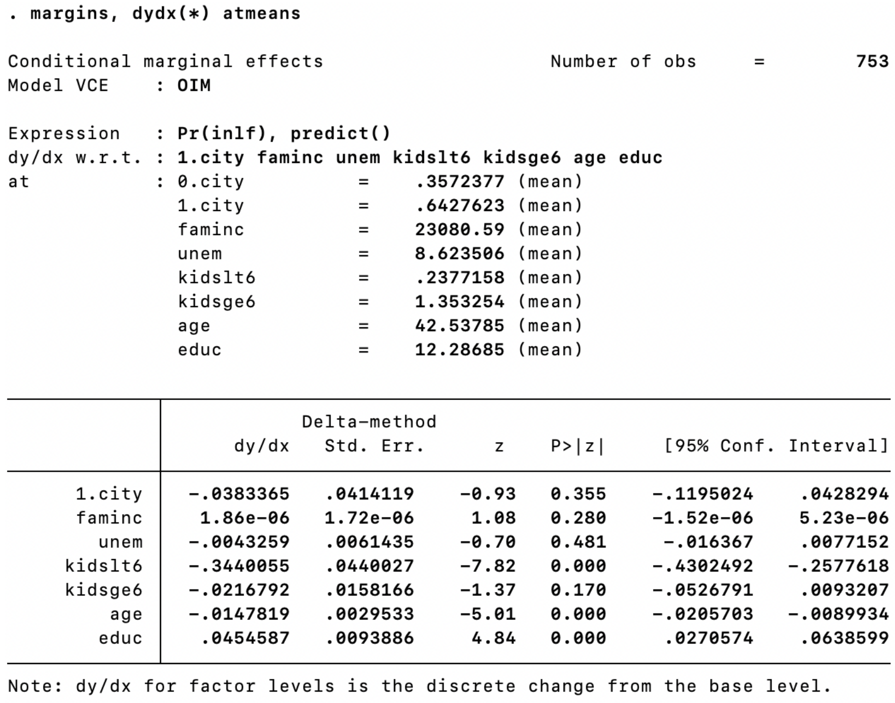
```
probit inlf i.city faminc unem kidslt6 kidsge6 age educ
```

(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.

Similar to what we mentioned in (f), since probit also wraps the explanatory variables by a function $\Phi(\cdot)$, the coefficient output is not directly the marginal effects. One can easily compute the marginal effects by Stata:

```
margins, dydx(*) atmeans
```

The resulting output looks like the following:



```
. margins, dydx(*) atmeans

Conditional marginal effects                    Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : 1.city faminc unem kidslt6 kidsge6 age educ
at           : 0.city          =    .3572377 (mean)
               1.city          =    .6427623 (mean)
               faminc          =    23080.59 (mean)
               unem            =    8.623506 (mean)
               kidslt6         =    .2377158 (mean)
               kidsge6         =    1.353254 (mean)
               age             =    42.53785 (mean)
               educ            =    12.28685 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.city |  -.0383365   .0414119    -0.93   0.355    -.1195024    .0428294
      faminc |   1.86e-06   1.72e-06     1.08   0.280    -1.52e-06    5.23e-06
        unem |  -.0043259   .0061435    -0.70   0.481     -.016367    .0077152
     kidslt6 |  -.3440055   .0440027    -7.82   0.000    -.4302492   -.2577618
     kidsge6 |  -.0216792   .0158166    -1.37   0.170    -.0526791    .0093207
         age |  -.0147819   .0029533    -5.01   0.000    -.0205703   -.0089934
        educ |   .0454587   .0093886     4.84   0.000     .0270574    .0638599
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

The marginal effect estimate on `kidslt6` is -0.344. Thus, having one more kid less 6 years old decreases the probability of the wife joining the labor force by 0.344.

# 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
  - $\alpha_i$ 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. $\alpha_i$ is uncorrelated with $X_{it}$
  2. $\alpha_i$ is correlated with $X_{it}$

## 2.1 $\alpha_i$ 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 $\varepsilon_{it}$ to be heteroskedastic?
  - Since $\varepsilon_{it} = \alpha_i + u_{it}$, for every individual $i$, there's a time-invariant component $\alpha_i$ that always exist for $\varepsilon_{it}$
  - 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 $\varepsilon_{it}$ 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 $\alpha_i$ 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 $\alpha_i$ 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 $\beta_1$ 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 $\alpha_i$ 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 $\alpha_i$ is correlated with $X_{it}$:

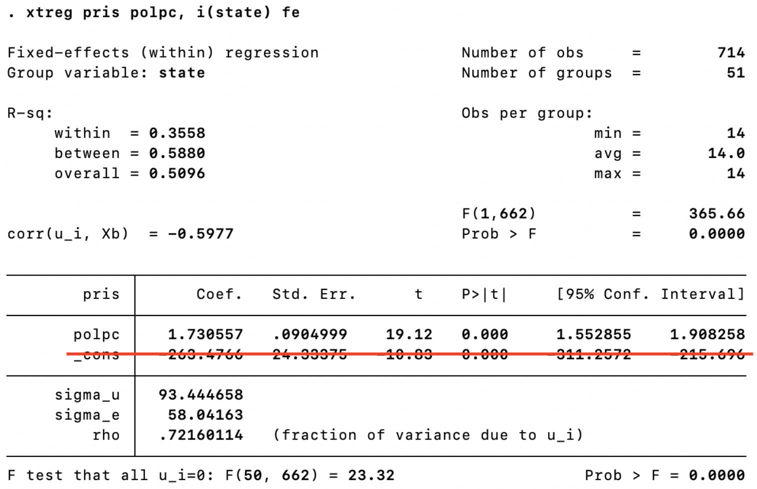
**Method 1: Fixed Effects Regression**

- 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 $\alpha_i$.

**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? → 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 $\beta$ 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
      ⇒ Instruments and the other exogenous variables are all uncorrelated with the original model's error term $u$
      ⇒ This predicted $X_1$ is constructed by things uncorrelated with $u$

⇒ 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?
  ⇒ NO! Since $X_1$ very much could be correlated with the other $X$s, so if we didn't include the other exogenous $X$s 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

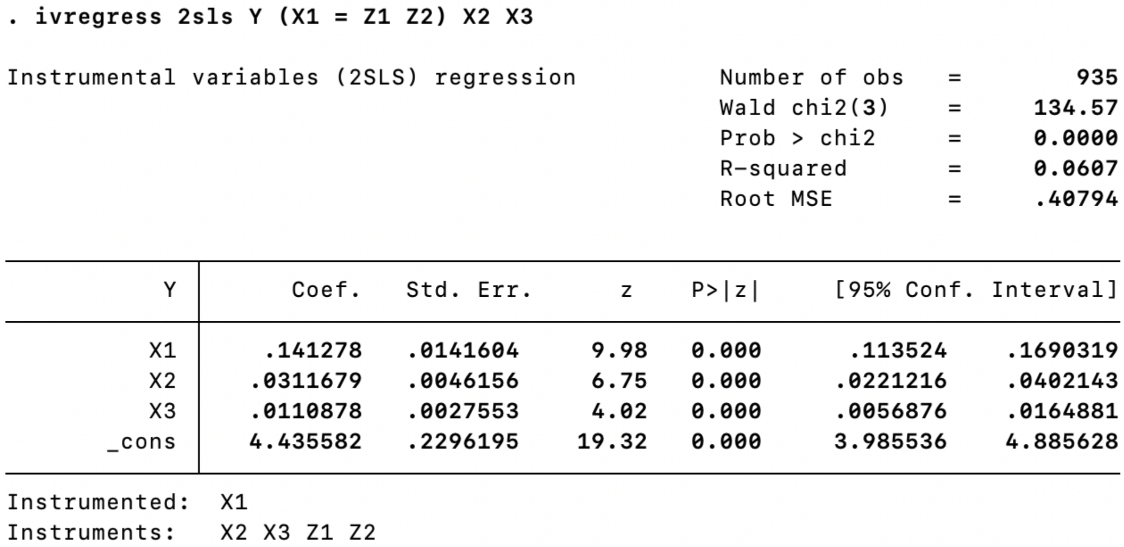
- Performing 2SLS directly:

```
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
```