Econ 613 - Assignment 4

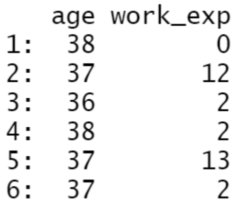
Zixuan Qi

April 21, 2022

Exercise 1

Question 1

The additional variables for the age and the total work experience measured in years are as follows.

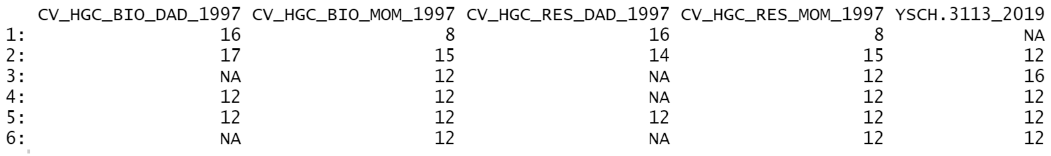


```
    age work_exp
1:   38        0
2:   37       12
3:   36        2
4:   38        2
5:   37       13
6:   37        2
```

Figure 1: Additional Variables

Question 2

The education variables indicating total years of schooling from all variables related to education are as follows.



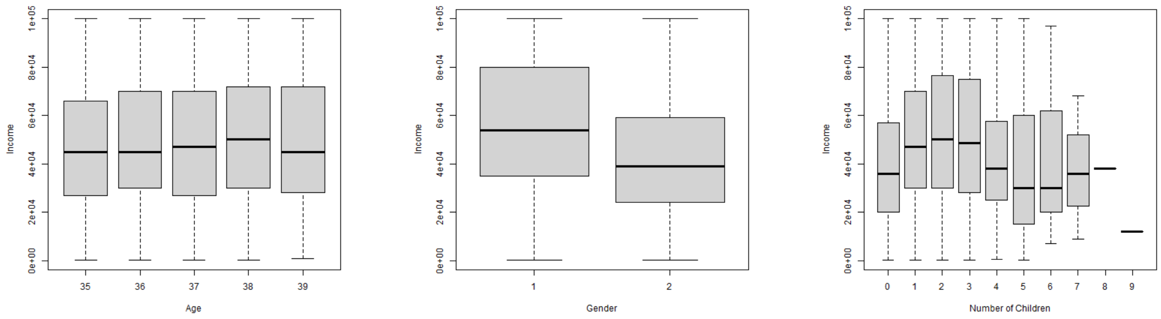
```
   CV_HGC_BIO_DAD_1997 CV_HGC_BIO_MOM_1997 CV_HGC_RES_DAD_1997 CV_HGC_RES_MOM_1997 YSCH.3113_2019
1:                  16                   8                  16                   8             NA
2:                  17                  15                  14                  15             12
3:                  NA                  12                  NA                  12             16
4:                  12                  12                  NA                  12             12
5:                  12                  12                  12                  12             12
6:                  NA                  12                  NA                  12             12
```

Figure 2: Education Variables

Question 3

Question 3.1

The Plots of the income data in different groups are as follows.



(a) Age Groups

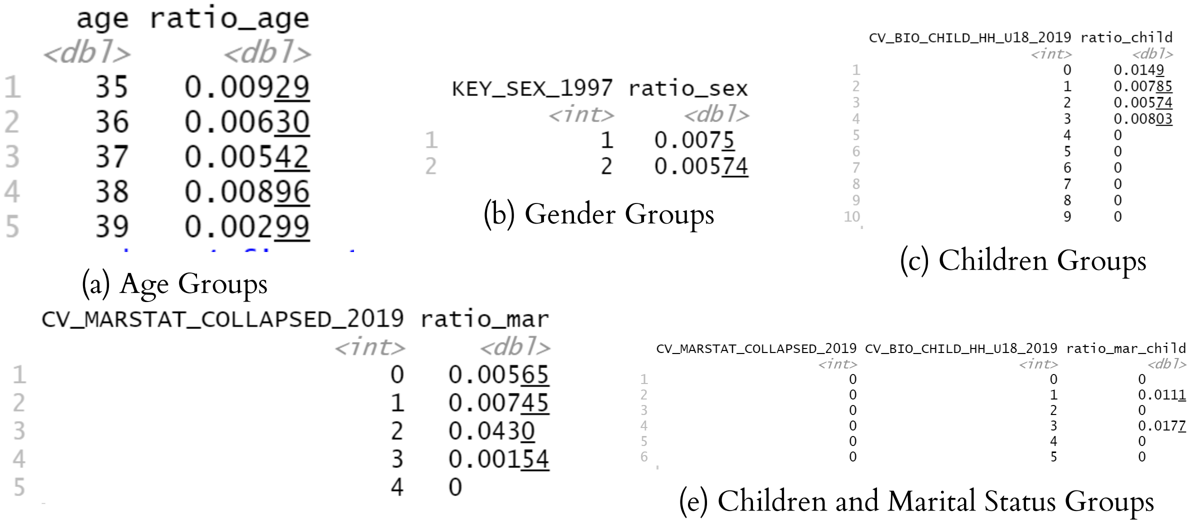
(b) Gender Groups

(c) Children Groups

Figure 3: Income Data

Question 3.2

The tables of share of ”0” in the income data in different groups are as follows.



```
    age ratio_age
  <dbl>     <dbl>
1    35   0.00929
2    36   0.00630
3    37   0.00542
4    38   0.00896
5    39   0.00299
```

(a) Age Groups

```
  KEY_SEX_1997 ratio_sex
         <int>     <dbl>
1            1    0.0075
2            2   0.00574
```

(b) Gender Groups

```
   CV_BIO_CHILD_HH_U18_2019 ratio_child
                      <int>       <dbl>
 1                        0      0.0149
 2                        1     0.00785
 3                        2     0.00574
 4                        3     0.00803
 5                        4     0
 6                        5     0
 7                        6     0
 8                        7     0
 9                        8     0
10                        9     0
```

(c) Children Groups

```
  CV_MARSTAT_COLLAPSED_2019 ratio_mar
                      <int>     <dbl>
1                         0   0.00565
2                         1   0.00745
3                         2    0.0430
4                         3   0.00154
5                         4   0
```

(d) Marital Status Groups

```
  CV_MARSTAT_COLLAPSED_2019 CV_BIO_CHILD_HH_U18_2019 ratio_mar_child
                      <int>                    <int>           <dbl>
1                         0                        0          0
2                         0                        1          0.0111
3                         0                        2          0
4                         0                        3          0.0177
5                         0                        4          0
6                         0                        5          0
```

(e) Children and Marital Status Groups

Figure 4: Share of 0

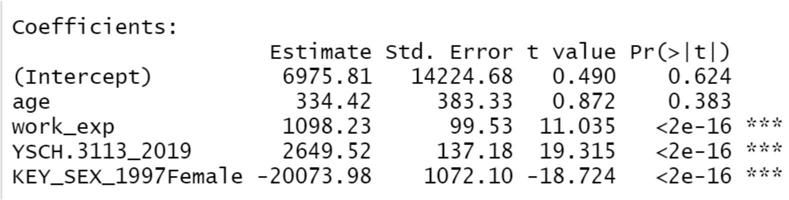
Question 3.3

1. In age groups, people whose age is around 38 have higher average wages.
2. In gender groups, there exists significant gaps in average wages between males and females.
3. In children groups, people who have one, two or three children earn more average wages than others.
4. For the share of 0 in the income data, it is clear that few people have zero wage. Female has less possibility to earn zero wage than male. People with a large number of children may not earn zero wage. It is more likely for people with separated marital status to earn zero wage.

Exercise 2

Question 1

I regard age, work experience, the total years of schooling of individual, and gender as regressors.



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          6975.81   14224.68   0.490    0.624    
age                   334.42     383.33   0.872    0.383    
work_exp             1098.23      99.53  11.035   <2e-16 ***
YSCH.3113_2019       2649.52     137.18  19.315   <2e-16 ***
KEY_SEX_1997Female -20073.98    1072.10 -18.724   <2e-16 ***
```

Figure 5: OLS Results

1. The regression results show that the parameters of age is not significant. Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 2649.52 units. Holding other variables constant, for every one unit increase in the work experience, wages increase by an average of 1089.23 units.

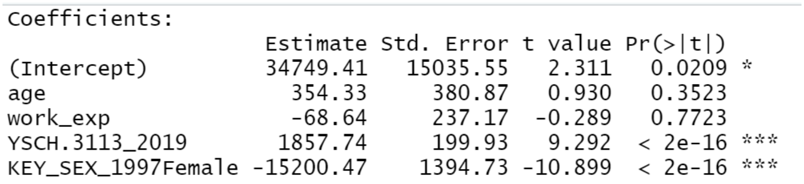
2. There exists selection problem because this regression only uses data of people with positive income. The selection of people who work is not random, thus using sub-population data to estimate the parameters leads to bias.

Question 2

The reason is that selection problem is a special case of endogeneity due to omitted variables. The Heckman selection model uses the probit model to represent the possibility of labor participation. Then, the samples of people who are not working are deleted, and the remaining sample points are shifted vertically downward according to their working probabilities. The smaller the working probability, the greater the downward displacement; the greater the working probability, the smaller the downward displacement. Specifically, adding inverse Mills Ratio into the OLS regression to eliminate the bias.

Question 3

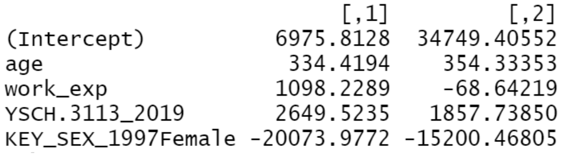
Please see the likelihood function in the code. I use other variables related to education and the number of children in the house as exogeneous variables in the probit model. The results of the Heckman selection model are as follows. It can be seen that the parameter of work experience is not significant.



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         34749.41   15035.55   2.311   0.0209 *
age                   354.33     380.87   0.930   0.3523
work_exp              -68.64     237.17  -0.289   0.7723
YSCH.3113_2019       1857.74     199.93   9.292  < 2e-16 ***
KEY_SEX_1997Female -15200.47    1394.73 -10.899  < 2e-16 ***
```

Figure 6: Heckman Results

The results of the OLS and the Heckman selection model are as follows. The first column is the results of the OLS, and the second column is the results of the heckman selection model.



```
                         [,1]         [,2]
(Intercept)         6975.8128  34749.40552
age                  334.4194    354.33353
work_exp            1098.2289    -68.64219
YSCH.3113_2019      2649.5235   1857.73850
KEY_SEX_1997Female -20073.9772 -15200.46805
```

Figure 7: OLS and Heckman Results

1. Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 1857.74 units. On average, female earns 15200.5 dollars less than male.

2. The parameters of age in both regressions are not significant. The effect work experience becomes insignificant in the Heckman Selection Model. The reason may be that people with high willingness of labor participation usually has more work experience and more years of schooling. In the biased OLS, when we only consider the people with more work experience, the effect of work experience may be amplified. The effect of years of schooling is also overestimated in the biased OLS, the reason may be same as above.

Exercise 3

Question 1

The histogram of the distribution of the income variable is as follows. The censored value is \$100000 because the distribution stop at the value \$100000

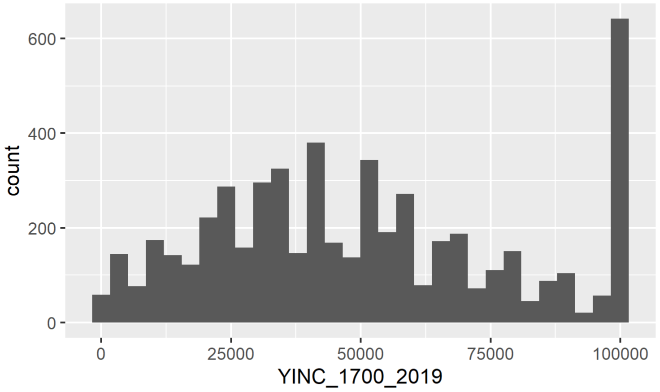
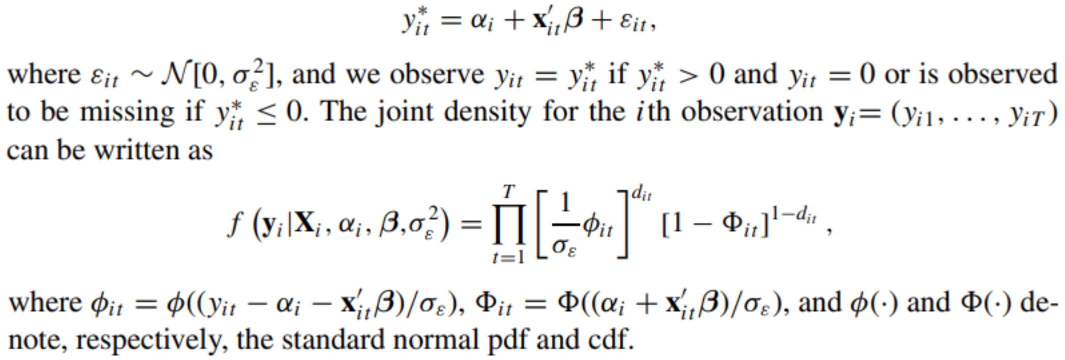


Figure 8: Distribution of Income

Question 2



$$y_{it}^* = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it},$$

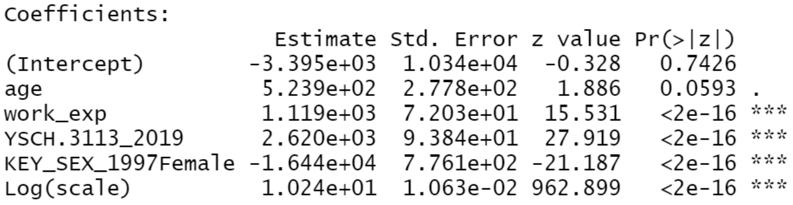
where $\varepsilon_{it} \sim \mathcal{N}[0, \sigma_\varepsilon^2]$, and we observe $y_{it} = y_{it}^*$ if $y_{it}^* > 0$ and $y_{it} = 0$ or is observed to be missing if $y_{it}^* \leq 0$. The joint density for the ith observation $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})$ can be written as

$$f\left(\mathbf{y}_i | \mathbf{X}_i, \alpha_i, \boldsymbol{\beta}, \sigma_\varepsilon^2\right) = \prod_{t=1}^{T} \left[\frac{1}{\sigma_\varepsilon}\phi_{it}\right]^{d_{it}} [1 - \Phi_{it}]^{1-d_{it}},$$

where $\phi_{it} = \phi((y_{it} - \alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon)$, $\Phi_{it} = \Phi((\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})/\sigma_\varepsilon)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the standard normal pdf and cdf.

Question 3

Please see the likelihood function in the code. The results of the tobit model are as follows. It can be seen that the parameter of work experience becomes significant.

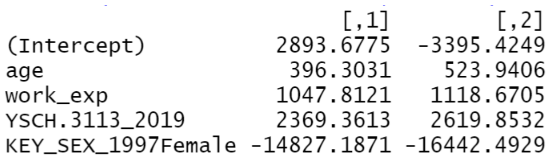


```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -3.395e+03  1.034e+04  -0.328   0.7426    
age                  5.239e+02  2.778e+02   1.886   0.0593 .  
work_exp             1.119e+03  7.203e+01  15.531   <2e-16 ***
YSCH.3113_2019       2.620e+03  9.384e+01  27.919   <2e-16 ***
KEY_SEX_1997Female -1.644e+04  7.761e+02 -21.187   <2e-16 ***
Log(scale)           1.024e+01  1.063e-02 962.899   <2e-16 ***
```

Figure 9: Tobit Results

Question 4

The results of the OLS and the tobit model are as follows. The first column is the results of the OLS, and the second column is the results of the tobit model.



```
                          [,1]        [,2]
(Intercept)          2893.6775  -3395.4249
age                   396.3031    523.9406
work_exp             1047.8121   1118.6705
YSCH.3113_2019       2369.3613   2619.8532
KEY_SEX_1997Female -14827.1871 -16442.4929
```

Figure 10: OLS and Tobit Results

1. Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 2619.85 units; for every one unit increase in the work experience, wages increase by an average of 1118.67 units. On average, female earns 16442.49 dollars less than male.

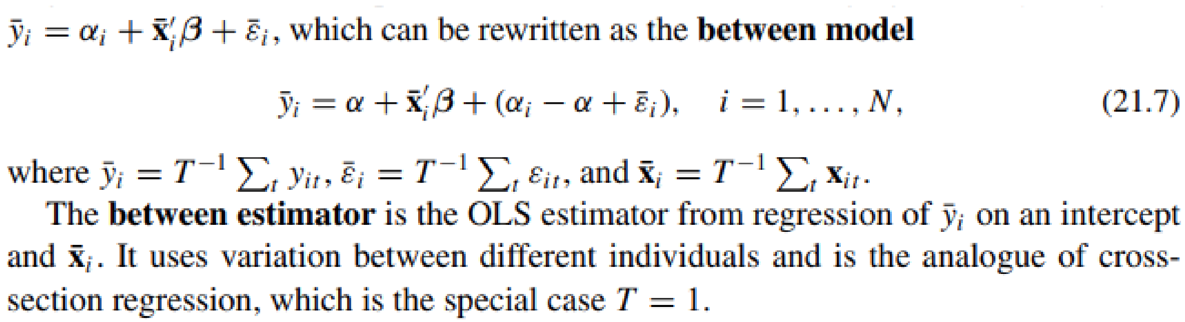
2. Compared with simple OLS results, the effects of these independent variables are greater because the wages are censored at $100000. Thus, in the simple OLS, the model ignores the effect of wages with higher than $100000.

Exercise 4

Question 1

There exists bias because some unobserved characteristics of individuals (abilities) affect wages. Thus, we need to add individual fixed effects in the model to avoid the endogeneity problem.

Question 2

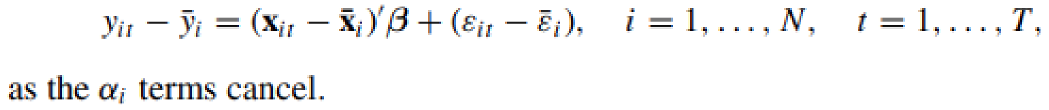


$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i$, which can be rewritten as the **between model**

$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + (\alpha_i - \alpha + \bar{\varepsilon}_i), \quad i = 1, \ldots, N, \qquad (21.7)$$

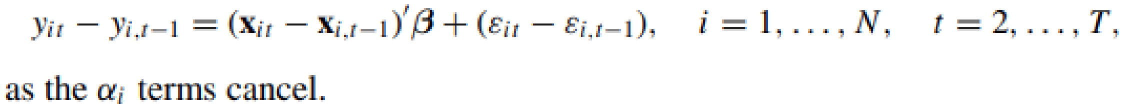
where $\bar{y}_i = T^{-1}\sum_t y_{it}$, $\bar{\varepsilon}_i = T^{-1}\sum_t \varepsilon_{it}$, and $\bar{\mathbf{x}}_i = T^{-1}\sum_t \mathbf{x}_{it}$.

The **between estimator** is the OLS estimator from regression of $\bar{y}_i$ on an intercept and $\bar{\mathbf{x}}_i$. It uses variation between different individuals and is the analogue of cross-section regression, which is the special case $T = 1$.



$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

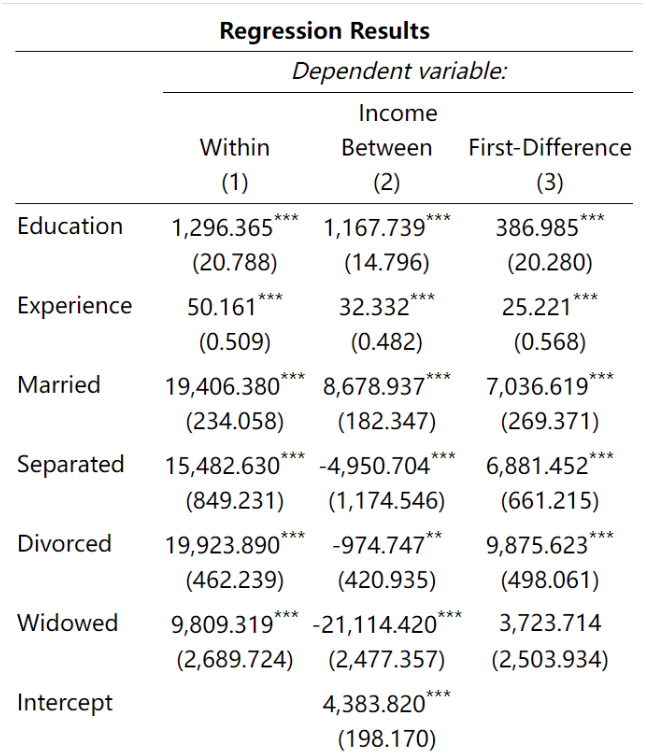
as the α_i terms cancel.



$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}), \quad i = 1, \ldots, N, \quad t = 2, \ldots, T,$$

as the α_i terms cancel.

The regression results are as follows.



Regression Results

	Dependent variable:		
	Income		
	Within	Between	First-Difference
	(1)	(2)	(3)
Education	1,296.365***	1,167.739***	386.985***
	(20.788)	(14.796)	(20.280)
Experience	50.161***	32.332***	25.221***
	(0.509)	(0.482)	(0.568)
Married	19,406.380***	8,678.937***	7,036.619***
	(234.058)	(182.347)	(269.371)
Separated	15,482.630***	-4,950.704***	6,881.452***
	(849.231)	(1,174.546)	(661.215)
Divorced	19,923.890***	-974.747**	9,875.623***
	(462.239)	(420.935)	(498.061)
Widowed	9,809.319***	-21,114.420***	3,723.714
	(2,689.724)	(2,477.357)	(2,503.934)
Intercept		4,383.820***	
		(198.170)	

Figure 12: Regression Results

Question 3

The interpretation of these results are as follows.

1. **Within Estimator.** Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 1296.37 units; for every one unit increase in the work experience, wages increase by an average of 50.161 units. On average, people with married status earn 19406.38 dollars more compared to the never-married individuals. Others variables related to marital status have the similar explanations as above.

2. **Between Estimator.** Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 1167.74 units; for every one unit increase in the work experience, wages increase by an average of 32.33

units. On average, people with married status earn 8678.94 dollars more compared to the never-married individuals. Others variables related to marital status have the similar explanations as above.

3. **First-Difference Estimator.** Holding other variables constant, for every one unit increase in the years of individual schooling, wages increase by an average of 386.99 units; for every one unit increase in the work experience, wages increase by an average of 25.22 units. On average, people with married status earn 7036.62 dollars more compared to the never-married individuals. Others variables related to marital status have the similar explanations as above.

The comparison of these results are as follows.

1. **Within Estimator.** The model measures the association between indiindividual-specific deviations of regressors from their time-averaged values. Thus, the fixed effect is added in this model, and the within estimators are consistent.

2. **Between Estimator.** This model uses variation between different individuals and is the analogue of cross section regression. However, if the regressors $\bar{x}_i$ are not independent of the error term $\alpha_i - \alpha + \bar{\epsilon}_i$. The estimators are biased.

3. **First-Difference Estimator.** This model measures the association between individual-specific one-period changes in regressors and individual-specific one-period changes in the dependent variable. Like the within estimator, this model also yields consistent estimated parameters.