Econometrics 1 - Problem Set 1 - Group 15

Felipe Montealegre[1], Paritosh Junare[2], and Ketki Balyan[3]

[1,2,3]Alma Mater Studiorum - Università di Bologna

October 2021

1 Question 1

1. For the following population model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

   (a) Following is the MATA output for the theoretical values of the coefficients:

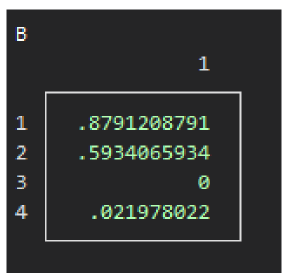

Figure 1: Population regression coefficients.

   (b) β_1 is the change in the conditional expectation of y given all regressors due to a marginal change in x_1 , holding all other regressors constant, that is, for every unit increase in x_1, there is an increase of 0.59 units in the conditional expectation of y given x, holding other x constant.

   (c) OLS estimator is defined as the minimizer of the expected squared difference between sample y and fitted y.

   (d) Total sum of squares (SST) is a measure of the total sample variation in y. It is defined as:

$$SST = \sum_{i=0}^{n} (y_i - \overline{y})^2.$$

Explained sum of squares (SSE) is a measure of total sample variation in $\hat{y}$.This is the part of variation in y explained by the regressors.

$$SSE = \sum_{i=0}^{n} (\hat{y}_i - \overline{y})^2.$$

Residual sum of squares (SSR) is a measure of total sample variation in $\hat{u}$. It is the unexplained part of the variation in y.

$$SSR = \sum_{i=0}^{n} (\hat{u}_i)^2.$$

(e) R-squared is the ratio of explained variation to the total variation.

$$R - squared = \frac{SSE}{SST} = \left(\frac{1 - SSR}{SST} \right).$$

Adjusted R squared is the modified version of R-squared, adjusted for the number of regressors. Unlike R-squared, adjusted R squared is relatively immune to the tendency of overfitting of model by adding irrelevant regressors.

$$AdjustedR - squared = 1 - \frac{SSR/(n-k)}{SST/n-1}.$$

```
sst
97.21797076

sse
44.52903988

ssr
52.68893088

r_squared
.4580330111

adjusted_r_squared
.4410965427
```

Figure 2: Sample variation estimates.

(f) OLS residuals are the differences between sample y and fitted values, $\hat{y}$. Fitted value, $\hat{y}$ is a function of sample y and obtained by multiplying the regressors with their respective OLS estimated coefficients.

(g) Sample average of OLS residuals, $\overline{\hat{u}}$ = -3.88471e-13, which is extremely close to zero. And figure 3 is the covariance vector of regressors and residuals. As the theoretical model suggests, both the

```
cov_xu_hat
                 1

1       -5.86484e-10
2       -5.85111e-10
3       -3.78675e-10
4       -3.84587e-11
```

Figure 3: Covariance between X and u.

sample average of residuals and their covariance with regressors are almost zero. This implies the OLS estimated coefficients are unbiased to their population values.

(h) Average fitted value, $\overline{\hat{y}}$ and average value of y are equal. $\overline{\hat{y}} = 9.8315$.

2. Following is the regression output obtained from STATA commands:

```
. reg y x*

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(3, 96)        =     27.04
       Model |  44.5290399         3  14.8430133   Prob > F        =    0.0000
    Residual |  52.6889309        96   .54884303   R-squared       =    0.4580
-------------+----------------------------------   Adj R-squared   =    0.4411
       Total |  97.2179708        99  .981999705   Root MSE        =    .74084

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          x1 |   .6141202   .0893371     6.87   0.000     .4367875    .7914529
          x2 |    .033728   .0763631     0.44   0.660    -.1178514    .1853075
          x3 |    .121349   .0782833     1.55   0.124    -.0340421      .27674
       _cons |  -.9562587   1.572035    -0.61   0.544    -4.076724    2.164206
------------------------------------------------------------------------------
```

Figure 4: Regression output using STATA commands.

3. 1000 random samples obtained from the joint distribution of Y

(a) (see files dofile.do or log.pdf attached).

(b) An unbiased estimator is one whose expected value is equal to the value of the population parameter we are estimating. We calculated the difference between the mean of the estimated betas in the 1000 samples of $n = 100$ and the betas in the real model. Their difference is virtually zero.

Variable	Obs	Mean	Std. dev.	Min	Max
diff_b_0	1,000	-1.31e-08	1.838884	-5.481631	4.837348
Variable	Obs	Mean	Std. dev.	Min	Max
diff_b_1	1,000	3.81e-09	.0876936	-.4007812	.330343
Variable	Obs	Mean	Std. dev.	Min	Max
diff_b_2	1,000	-7.69e-10	.0819166	-.2453268	.247237
Variable	Obs	Mean	Std. dev.	Min	Max
diff_b_3	1,000	-1.45e-09	.0868876	-.284941	.3171248

Figure 5: Difference between the estimated betas for 1000 random samples of $n = 100$ and the true population parameters.

(c) In the graph below we can see the histogram plot of all four estimated betas. Their distribution resembles a normal distribution with their means (plotted in red) centered around the values of the population betas.

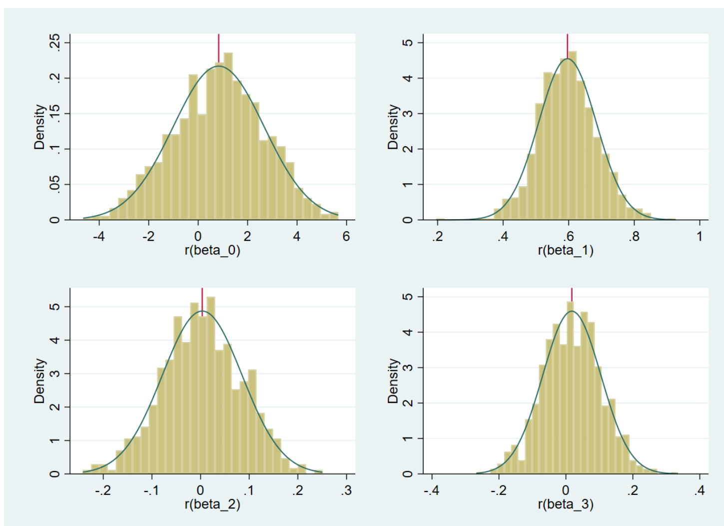

Figure 6: Distribution of estimated betas from 1000 random samples of $n = 100$.

2 Question 2

1. We have chosen the average hours worked per week (hourswm) as the measure of labor supply. Therefore, our model for equation 3 is given as:

$$hourwsm_i = \beta_0 + \beta_1 morekids_i + \beta_2 X_i + \epsilon_i$$

   where X_i includes $blackm, hispm, othracem, educm, agem1, agefstm$.

2. Following is the image of output obtained from STATA command:
   As observed, OLS method from both commands yield the same output.

```
. reg hourswm morekids educm agem1 agefstm blackm hispm othracem

      Source |       SS           df       MS      Number of obs   =   322,542
-------------+----------------------------------   F(7, 322534)    =   3626.21
       Model |   8418182.8         7  1202597.54   Prob > F        =    0.0000
    Residual |   106965359   322,534  331.640568   R-squared       =    0.0730
-------------+----------------------------------   Adj R-squared   =    0.0729
       Total |   115383542   322,541  357.732945   Root MSE        =    18.211

------------------------------------------------------------------------------
     hourswm | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    morekids |  -6.374662   .0684161   -93.17   0.000    -6.508756   -6.240568
       educm |   .7717739   .0150779    51.19   0.000     .7422216    .8013262
       agem1 |   .8708522   .0102514    84.95   0.000     .8507597    .8909447
     agefstm |  -1.605353   .0133654  -120.11   0.000    -1.631548   -1.579157
      blackm |    5.43159    .101549    53.49   0.000     5.232557    5.630623
       hispm |   2.491738   .1917857    12.99   0.000     2.115844    2.867633
    othracem |   4.365351   .1924748    22.68   0.000     3.988106    4.742597
       _cons |     17.262   .3179841    54.29   0.000     16.63876    17.88524
------------------------------------------------------------------------------
```

Figure 7: Regression output.

3. Restricted model is given as:

$$hourwsm_i = \beta_0 + \beta_1 morekids_i + \beta_2 educm + u_i$$

   (a) No, $\hat{\beta}_1$, the OLS estimator for β_1 in model 4 is not unbiased for β_1 since the restricted model has omitted several variables of the population model (Model 3), and these regressors' coefficients in the population model are not zero and the regressors are correlated with $morekids_i$ and $educm_i$. Since neither source of bias is zero, the OLS estimates in model 4 are not unbiased.

   (b) $\hat{\beta}_1$, the OLS estimator of model 3, and $\tilde{\beta}_1$, the OLS estimator of model 4 will be very similar if either of the following conditions are satisfied:

      i. The omitted regressors' OLS coefficients are very close to zero.
      ii. The correlation between the omitted regressors and $morekids_i$ is very close to zero.

   Output: $\hat{\beta}_1$ = -6.374662, $\tilde{\beta}_1$ = -3.998152

```
. corr hourswm morekids blackm hispm othracem educm agem1 agefstm
(obs=322,542)

             |  hourswm morekids   blackm    hispm othracem    educm    agem1  agefstm
-------------+------------------------------------------------------------------------
     hourswm |   1.0000
    morekids |  -0.1077   1.0000
      blackm |   0.1057   0.0741   1.0000
       hispm |  -0.0094   0.0517  -0.0650   1.0000
    othracem |   0.0184   0.0089  -0.0634  -0.0304   1.0000
       educm |   0.0423  -0.1539  -0.0402  -0.1919  -0.0134   1.0000
       agem1 |   0.0533   0.0965  -0.0976  -0.0547   0.0090   0.2054   1.0000
     agefstm |  -0.1280  -0.1943  -0.1939  -0.0456   0.0493   0.4222   0.4089   1.0000
```

Figure 8: Correlation output.

```
. reg hourswm morekids educm

      Source |       SS           df       MS      Number of obs   =   322,542
-------------+----------------------------------   F(2, 322539)    =   2005.33
       Model |  1417133.91         2  708566.954   Prob > F        =    0.0000
    Residual |   113966408   322,539  353.341481   R-squared       =    0.0123
-------------+----------------------------------   Adj R-squared   =    0.0123
       Total |   115383542   322,541  357.732945   Root MSE        =    18.797

------------------------------------------------------------------------------
     hourswm | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    morekids |  -3.998152   .0683112   -58.53   0.000     -4.13204   -3.864264
       educm |   .2076339   .0139406    14.89   0.000     .1803107    .2349571
       _cons |   17.89738   .1784834   100.27   0.000     17.54756     18.2472
------------------------------------------------------------------------------
```

Figure 9: Restricted regression output.

(c) To confirm partialled out interpretation of the OLS estimates, following process is followed to obtain the estimate $\tilde{\beta}_1$:

i. Regress *hourswm* on *educm* and predict residuals, $\tilde{e}_2$.
ii. Regress *morekids* on *educm* and predict residuals, $\tilde{x}_1$.
iii. Regress the residual $\tilde{e}_2$ on $\tilde{x}_1$ and the regression coefficient is the partialled out interpretation of $\tilde{\beta}_1$.

Same procedure, after interchanging *morekids* and *educm*, provides us $\tilde{\beta}_2$.