Problem Set 3

Richard Gong
3032755754
GSI: Murilo
DIS 206

Problem 1

b)

```
. reg y x1

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    374.88
       Model |  44095.0772         1  44095.0772   Prob > F        =    0.0000
    Residual |  117389.163       998  117.624412   R-squared       =    0.2731
-------------+----------------------------------   Adj R-squared   =    0.2723
       Total |   161484.24       999  161.645886   Root MSE        =    10.845

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -6.471674   .3342491   -19.36   0.000    -7.127586   -5.815762
       _cons |  -8.291288   .3431456   -24.16   0.000    -8.964658   -7.617919
------------------------------------------------------------------------------
```

```
. reg y x2

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   3904.09
       Model |  128608.217         1  128608.217   Prob > F        =    0.0000
    Residual |  32876.0232       998  32.9419071   R-squared       =    0.7964
-------------+----------------------------------   Adj R-squared   =    0.7962
       Total |   161484.24       999  161.645886   Root MSE        =    5.7395

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |  -4.653856   .0744823   -62.48   0.000    -4.800016   -4.507696
       _cons |  -.3337941    .223727    -1.49   0.136    -.7728235    .1052352
------------------------------------------------------------------------------
```

Both the coefficients differ substantially from their specified coefficients in the population model due to omitted variable bias. Omitting x2 in the first regression leads the x1 coefficient to be more negative than its population value. Omitting x1 in the second regression leads the x2 coefficient to be less negative than its population value.

c)

```
. reg y x1 x2

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =   2726.52
       Model |  136523.194         2  68261.5969   Prob > F        =    0.0000
    Residual |  24961.0463       997  25.0361548   R-squared       =    0.8454
-------------+----------------------------------   Adj R-squared   =    0.8451
       Total |   161484.24       999  161.645886   Root MSE        =    5.0036

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   4.143288   .2330259    17.78   0.000     3.686011    4.600566
          x2 |  -5.961827   .0981208   -60.76   0.000    -6.154374    -5.76928
       _cons |   1.825082   .2297475     7.94   0.000     1.374238    2.275926
------------------------------------------------------------------------------
```

This regression results in coefficients much closer to the true population values. Compared to b), x1 now has the appropriate positive sign, and x2 is more negative. The coefficient estimate in c) for x1 is closer to the true value and has a lower standard error than that in b). The coefficient estimate in c) for x2 is closer to the true value and has a higher standard error than that in b). R^2 is higher in c) compared to both regressions in b). The regression on just x2 in b) has a much higher R^2 than the regression on just x1 in b). Adjusted R^2 values follow the behavior of the R^2 values.

d)

```
. reg y x1 x2 x3

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(3, 996)       =   1816.42
       Model |  136529.767         3  45509.9224   Prob > F        =    0.0000
    Residual |  24954.4729       996  25.0546917   R-squared       =    0.8455
-------------+----------------------------------   Adj R-squared   =    0.8450
       Total |   161484.24       999  161.645886   Root MSE        =    5.0055

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   4.300229   .3849936    11.17   0.000     3.544737    5.055721
          x2 |  -6.121832   .3274377   -18.70   0.000    -6.764379   -5.479285
          x3 |  -.1614444    .315189    -0.51   0.609    -.7799552    .4570664
       _cons |   1.983575   .3854454     5.15   0.000     1.227197    2.739953
------------------------------------------------------------------------------
```

The x1 and x2 coefficients are slightly larger in magnitude, and have higher standard errors relative to those in c). As x3 is an imperfect linear combination of x1 and x2, we get exaggerated slope coefficients and larger standard errors. The R^2 value is very slightly higher as x3 improves prediction by chance, while the R^2 adjusted value is slightly lower due to the penalty for an additional covariate.

e)

```
. di x1_coef_avg
3.9970031

. di x2_coef_avg
-6.0014029

. di x1_cor_avg
-.52371615

. di x2_cor_avg
-.86075681
```

The average of 1000 repetitions leads to coefficient means and correlations that are very close to the true specified values. See .do file code for details.

Problem 2

a)

```
. reg course_eval beauty, r

Linear regression                               Number of obs     =        463
                                                F(1, 461)         =      16.94
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0357
                                                Root MSE          =     .54545

------------------------------------------------------------------------------
             |               Robust
 course_eval |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      beauty |   .1330014   .0323189     4.12   0.000     .0694908    .1965121
       _cons |   3.998272   .0253493   157.73   0.000     3.948458    4.048087
------------------------------------------------------------------------------
```

The estimated slope of .1330014 is statistically significant at the 1% level.

b)

```
. reg course_eval beauty intro onecredit female minority nnenglish, r

Linear regression                               Number of obs     =        463
                                                F(6, 456)         =      17.03
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1546
                                                Root MSE          =     .51351

------------------------------------------------------------------------------
             |               Robust
 course_eval |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      beauty |     .16561   .0315686     5.25   0.000     .1035721    .2276478
       intro |    .011325   .0561741     0.20   0.840    -.0990673    .1217173
   onecredit |   .6345271   .1080864     5.87   0.000     .4221178    .8469364
      female |  -.1734774   .0494898    -3.51   0.001    -.2707337   -.0762212
    minority |  -.1666154   .0674115    -2.47   0.014    -.2990912   -.0341397
   nnenglish |  -.2441613   .0936345    -2.61   0.009      -.42817   -.0601526
       _cons |   4.068289   .0370092   109.93   0.000     3.995559    4.141019
------------------------------------------------------------------------------
```

The coefficient of beauty is .16561, so an increase by beauty of 1 raises course evaluation score by .16561 on average. The regression from a) suffers from OVB, the included control variables in b) theoretically influence course evaluations, and should be accounted for. Beauty is downwards biased in a). R^2 improves significantly as the additional controls improve the regression fit. The confidence intervals from a) and b) overlap so we cannot reject the null hypothesis that the effect of beauty is the same. The additional variables ‘female’, ‘minority’, ‘nnenglish’ have statistically significant negative effects on course evaluation score. The variable ‘onecredit’ has a statistically significant positive effect, and ‘intro’ appears to have no effect.

c)

```
. reg course_o beauty_o, r

Linear regression                               Number of obs     =        463
                                                F(1, 461)         =      27.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0599
                                                Root MSE          =     .51071

------------------------------------------------------------------------------
             |               Robust
    course_o |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    beauty_o |     .16561   .0313969     5.27   0.000     .1039112    .2273087
       _cons |   7.17e-09   .0237348     0.00   1.000    -.0466419    .0466419
------------------------------------------------------------------------------
```

Frisch-Waugh results in the exact same coefficient estimate for beauty.

d)

Professor Smith score = 4.068289 + .16561 * (0) + .011325 * (0) + .6345271 * (0) + -.1734774 * (0) + -.1666154 * (1) + -.2441613 * (0) = **3.9016736**

Problem 3

a)

The further the nearest four-year college, the fewer years of education completed. So negative sign for ‘dist’ coefficient. This is because a longer commute may be too expensive or time consuming to justify continued education.

b)

```
Linear regression                               Number of obs     =      3,796
                                                F(1, 3794)        =      29.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0074
                                                Root MSE          =     1.8074

------------------------------------------------------------------------------
             |               Robust
       yrsed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0733727   .0134334    -5.46   0.000    -.0997101   -.0470353
       _cons |   13.95586   .0378112   369.09   0.000     13.88172    14.02999
------------------------------------------------------------------------------
```

The estimated slope of -.0733727 is statistically significant at the 1% level. The distance to college does **not** explain a large fraction of the variance in educational attainment across individuals. The R^2 value of 0.0074 is small, indicating the linear relationship is a poor predictor.

c)

```
Linear regression                               Number of obs     =      3,796
                                                F(11, 3784)       =     183.54
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2829
                                                Root MSE          =     1.5383

------------------------------------------------------------------------------
             |               Robust
       yrsed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0308039   .0116178    -2.65   0.008    -.0535816   -.0080262
      bytest |   .0924474   .0030009    30.81   0.000     .0865638    .0983309
      female |   .1433777   .0502841     2.85   0.004     .0447912    .2419642
       black |   .3538083   .0674994     5.24   0.000     .2214695    .4861471
    hispanic |   .4023514   .0737302     5.46   0.000     .2577966    .5469063
    incomehi |   .3665952   .0622404     5.89   0.000     .2445672    .4886233
     ownhome |   .1456416   .0648174     2.25   0.025     .0185612    .2727221
     dadcoll |   .5699153   .0762509     7.47   0.000     .4204185    .7194121
     momcoll |   .3791836   .0835917     4.54   0.000     .2152945    .5430728
       cue80 |    .024418   .0092692     2.63   0.008     .0062449    .0425911
    stwmfg80 |  -.0502044   .0195902    -2.56   0.010    -.0886128    -.011796
       _cons |   8.861373   .2410771    36.76   0.000      8.38872    9.334027
------------------------------------------------------------------------------
```

The estimated effect of dist is now -.0308039. It is much lower in magnitude compared to the slope in b). This suggests negative OVB exists in b). We still need relevant controls to claim OVB, and such controls are appropriately included in c).

d)

'dadcoll' indicates if the father is a college graduate. The coefficient effect tells us that individuals whose dads are college graduates attain an additional .5699153 years of education on average.

e)

'ue80' and 'stwmfg80' proxy labor opportunities for low education individuals. The more local labor opportunities, the less individuals would be inclined to get more education. 'ue80' has a coefficient of .02441, which makes sense. Higher unemployment leads to fewer labor opportunities, so it may be better to stay in school. 'stwmfg80' has a coefficient of -.0502044, which seems to suggest higher wages for low-skilled workers encourages attending more schools. This is unexpected.

f)

Bob education = 8.861373 + -.0308039 (2) + .0924474 (58) + .1433777 (0) + .3538083 (1) + .4023514 (0) + .3665952 (1) + .1456416 (1) + .5699153 (0) + .3791836

(1) + .024418 (7.5) + -.0502044 (9.75) = **15.1005852**

Problem 4

a)

Using white men alone controls for race, which likely has a significant effect on log hourly wages.

b)

The coefficient on 9 of -0.277 is the mean difference in log hourly wages between those who have completed 9 years of education and those who have completed 12 years of education.

c)

The 14th year of college can correspond to the completion of a 2-year degree, so it has its own diploma effect. The 15th year has no diploma effect.

d)

Following college graduation, a senior would get the average diploma effect of a Bachelor's degree 0.245 (which is relative to high school). Completing the 16th year of education would give an effect of 0.178 (which is relative to 12 years). So the average difference following graduation is 0.423

e)

If your goal is to earn a higher wage, a professional degree is better than a doctoral degree. The marginal effect over a bacher's is 0.286 for a professional degree, which is higher than 0.067 for a doctoral degree.

f)

Perform an F-test

H_0: Diploma dummies have coefficients of 0
H_1: Some diploma dummies do not have coefficient 0

q = 8: There are 8 diploma effect variables.
k = 26: There are dummies for 0-18 years of education excluding 12. And the 8 diploma dummies

n = 8957: observations

$F = [(0.154 - 0.147) / 8] / [(1 - 0.154)/(8957 - 26 - 1)] = 9.23611111111$

The p-value is 0.0 using the $F_{8,8930}$ distribution. Therefore, we reject the null at the 5% level. Diploma effects exist.

```
* 1a
set obs 1000
set seed 0
generate u = rnormal(0,5)
gen x1 = rnormal()
gen x2 = exp(x1)
gen y = 2 + 4 * x1 - 6 * x2 + u
* 1b
reg y x1
reg y x2
* 1c
reg y x1 x2
* 1d
gen v = rnormal(0, 0.5)
gen x3 = 1 + x1 - x2 + v
reg y x1 x2 x3
* 1e
gen x1_coef_avg = 0
gen x2_coef_avg = 0
gen x1_cor_avg = 0
gen x2_cor_avg = 0
forvalues i = 1/1000 {
	quietly {
		replace u = rnormal(0,5)
		replace x1 = rnormal()
		replace x2 = exp(x1)
		replace y = 2 + 4 * x1 - 6 * x2 + u
		reg y x1 x2
		replace x1_coef_avg = x1_coef_avg + _b[x1]
		replace x2_coef_avg = x2_coef_avg + _b[x2]
		cor y x1
		replace x1_cor_avg = x1_cor_avg + r(rho)
		cor y x2
		replace x2_cor_avg = x2_cor_avg + r(rho)
	}
}
replace x1_coef_avg = x1_coef_avg / 1000
replace x2_coef_avg = x2_coef_avg / 1000
replace x1_cor_avg = x1_cor_avg / 1000
replace x2_cor_avg = x2_cor_avg / 1000
* 2a
use "TeachingRatings.dta", clear
reg course_eval beauty, r
* 2b
reg course_eval beauty intro onecredit female minority nnenglish, r
* 2c
reg course_eval intro onecredit female minority nnenglish, r
predict fit_c , xb
gen course_o = course_eval - fit_c
reg beauty intro onecredit female minority nnenglish, r
predict fit_b , xb
gen beauty_o = beauty - fit_b
reg course_o beauty_o, r
```

```
* 3b
reg yrsed dist, r
* 3c
reg yrsed dist bytest female black hispanic incomehi ownhome dadcoll momcoll cue80 stwmfg80, r
```