ECO 394M Homework 1

Steven Kim

Question 1

(a) The model does not differentiate the effect of being legal to purchase alcohol on alcohol consumption from the relationship between alcohol and age that would exist even without a drinking age law.

(b) i. $E(alcohol|age) = \beta_1 + \beta_3 overage$

ii. $E(alcohol|age) = \beta_1 + \beta_2 age + \beta_3 overage$

iii. when $\beta_2 = 0$, the model in (ii) becomes identical to the one in (i).

iv. The effect of the drinking age law is captured as β_3. If $\beta_3 = 0$, it would imply that the drinking law has no effect.

v. $E(alcohol|age = 25) - E(alcohol|age = 22) = \beta_2(25 - 22) + \beta_3(1 - 1) = 3\beta_2$
$E(alcohol|age = 22) - E(alcohol|age = 19) = \beta_2(22 - 19) + \beta_3(1 - 0) = 3\beta_2 + \beta_3$

(c) $E(alcohol|age) = \beta_1 + \beta_2 age + \beta_3 overage + \beta_4(overage * age)$
The effect of the drinking age law is captured in β_3 and β_4. If $\beta_3 = \beta_4 = 0$, it would imply that the drinking law has no effect.
$E(alcohol|age = 25) - E(alcohol|age = 22) = \beta_2(25 - 22) + \beta_3(1 - 1) + \beta_4(25 - 22) = 3\beta_2 + 3\beta_4$
$E(alcohol|age = 22) - E(alcohol|age = 19) = \beta_2(22 - 19) + \beta_3(1 - 0) + \beta_4(22 - 0) = 3\beta_2 + \beta_3 + 22\beta_4$

(d) The model doesn't give more information than the data. The comparison I did for (b) and (c) also loses the meaning, as they would not have the common parts. It would just be numbers, hard to interpret.

Question 2

(a)

$$S(b) = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})^2$$

first order conditions are

$$\frac{\partial S(b)}{\partial b_1} = \sum_{i=1}^{n} 2(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-1) = 0$$

$$\sum_{i=1}^{n} b_1 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_2 x_{i2} - \sum_{i=1}^{n} b_3 x_{i3}$$

$$\frac{\partial S(b)}{\partial b_2} = \sum_{i=1}^{n} 2(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i2}) = 0$$

$$\sum_{i=1}^{n} y_i x_{i2} - \sum_{i=1}^{n} b_1 x_{i2} - \sum_{i=1}^{n} b_2 x_{i2} = 0$$

$$\frac{\partial S(b)}{\partial b_3} = \sum_{i=1}^{n} 2(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i3}) = 0$$

$$\sum_{i=1}^{n} y_i x_{i3} - \sum_{i=1}^{n} b_1 x_{i3} - \sum_{i=1}^{n} b_3 x_{i3} = 0$$

$$\sum_{i=1}^{n} b_2 x_{i2} = \sum_{i=1}^{n} y_i x_{i2} - \sum_{i=1}^{n} b_1 x_{i2}$$

$$\sum_{i=1}^{n} b_3 x_{i2} = \sum_{i=1}^{n} y_i x_{i3} - \sum_{i=1}^{n} b_1 x_{i3}$$

$$\sum_{i=1}^{n} b_1 = \sum_{i=1}^{n} y_i - (\sum_{i=1}^{n} y_i x_{i2} - \sum_{i=1}^{n} b_1 x_{i2}) - (\sum_{i=1}^{n} y_i x_{i3} - \sum_{i=1}^{n} b_1 x_{i3})$$

$$\sum_{i=1}^{n} (1 - x_{i2} - x_{i3}) b_1 = \sum_{i=1}^{n} y_i (1 - x_{i2} - x_{i3})$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i (1 - x_{i2} - x_{i3})}{\sum_{i=1}^{n} (1 - x_{i2} - x_{i3})} = \frac{\sum_{i \in C} y_i}{n_C}$$

$$b_2 = \frac{\sum_{i=1}^{n} y_i x_{i2} - \sum_{i=1}^{n} b_1 x_{i2}}{\sum_{i=1}^{n} x_{i2}} = \frac{\sum_{i \in A} y_i - \frac{\sum_{i \in C} y_i}{n_C} \sum_{i=1}^{n} x_{i2}}{n_A}$$

$$= \frac{\sum_{i \in A} y_i - \frac{\sum_{i \in C} y_i}{n_C} n_A}{n_A} = \frac{\sum_{y \in A}}{n_A} - \frac{\sum_{i \in C}}{n_C}$$

Similarly,

$$b_3 = \frac{\sum_{y \in B}}{n_B} - \frac{\sum_{i \in C}}{n_C}$$

Therefore, the estimates we're looking for are

$$\widehat{\beta}_1 = \frac{\sum_{i \in C} y_i}{n_C}$$

$$\widehat{\beta}_2 = \frac{\sum_{y \in A}}{n_A} - \frac{\sum_{i \in C}}{n_C}$$

$$\widehat{\beta}_3 = \frac{\sum_{y \in B}}{n_B} - \frac{\sum_{i \in C}}{n_C}$$

(b) Yes. $\widehat{\beta}_2 = \frac{\sum_{y \in A}}{n_A} - \frac{\sum_{i \in C}}{n_C}$, which means the difference of sample mean when $c = A$ (which means $x_2 = 1$) and when $c = C$ (which means $x_4 = 1$).

(c) No. We get the results that consist of sample means because we only have indicator variables in the model. if there were other variables like age, the form would likely be more complicated.

(d) Yes. We can expect it to expand in the similar way. For example, if we had one more categories, the estimates would look like this:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i \in D} y_i}{n_D} \\
\widehat{\beta}_2 &= \frac{\sum_{y \in A}}{n_A} - \frac{\sum_{i \in D}}{n_D} \\
\widehat{\beta}_3 &= \frac{\sum_{y \in B}}{n_B} - \frac{\sum_{i \in D}}{n_D} \\
\widehat{\beta}_4 &= \frac{\sum_{y \in C}}{n_C} - \frac{\sum_{i \in D}}{n_D}
\end{aligned}$$

Question 3

(a) Let $\mathbf{X_2} = \mathbf{X_1 a}$. Then, by the partition regression method, $\widehat{\beta}_2 = ((M_{X_1}X_2)'(M_{X_1}X_2))^{-1}(M_{X_1}X_2)'(M_{X_1}y)$.
However, since $M_{X_1}X_2 = M_{X_1}X_1a = 0$, the inverse does not exist.
Therefore, we would not be able to obtain the estimate.

(b)

$$\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{u}$$

$$\mathbf{X_1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \mathbf{X_2} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \beta_1 \text{ and } \beta_2 \text{ are scalars.}$$

From the partitioned regression method, we know that $\widehat{\beta}_2 = ((M_{X_1}X_2)'((M_{X_1}X_2))^{-1}(M_{X_1}X_2)'(M_{X_1}y)$.

$$P_{X_1} = X_1(X_1'X)^{-1}X_1' = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} (\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix})^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} n^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

$$M_{X_1} = I - P_{X_1}$$

$$M_{X_1}X_2 = I_nX_2 - P_{X_1}X_2 = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \sum_{i=1}^n x_i \\ \vdots \\ \sum_{i=1}^n x_i \end{bmatrix} = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$$

$$M_{X_1}y = Iy - P_{X_1}y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \sum_{i=1}^n y_i \\ \vdots \\ \sum_{i=1}^n y_i \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

$$\begin{aligned}
\widehat{\beta}_2 &= ((M_{X_1}X_2)'((M_{X_1}X_2))^{-1}(M_{X_1}X_2)'(M_{X_1}y) \\
&= (\begin{bmatrix} (x_1 - \bar{x}) & \cdots & (x_n - \bar{x}) \end{bmatrix} \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix})^{-1} \begin{bmatrix} (x_1 - \bar{x}) & \cdots & (x_n - \bar{x}) \end{bmatrix} \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{y_i,x_i}}{s_{x_i}^2}
\end{aligned}$$

(c) Since X_1 and X_2 are not correlated at all, there would be nothing to be filtered out and therefore $\widehat{\beta}_2$ would stay the same. More specifically,

$$\begin{aligned}
M_{X_1} &= I - P_{X_1} \\
M_{X_1}X_2 &= (I - P_{X_1})X_2 \\
&= X_2 - P_{X_1}X_2 \\
&= X_2 \text{ (as } X_1 \text{ and } X_2 \text{ have zero correlation)}
\end{aligned}$$

$$\begin{aligned}
\widehat{\beta}_2 &= ((M_{X_1}X_2)'((M_{X_1}X_2))^{-1}(M_{X_1}X_2)'(M_{X_1}y) \\
&= (X_2'X_2)^{-1}X_2'y
\end{aligned}$$

This is the slope estimate in the simple regression model $E(y|x) = \beta_1 + \beta_2 X_2$.

Question 4

(a)

```
. clear

. use "/Users/steven_unique/My Drive/FALL 2021/ECO 394M ECONOMETRICS/Problem Set
> s/PS1/HTV.DTA"

. regr educ motheduc fatheduc

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(2, 1227)      =    203.68
       Model |   1697.9676         2   848.9838   Prob > F        =    0.0000
    Residual |  5114.31207     1,227  4.1681435   R-squared       =    0.2493
-------------+----------------------------------   Adj R-squared   =    0.2480
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.0416

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .3041971   .0319266     9.53   0.000     .2415603     .366834
    fatheduc |   .1902858   .0222839     8.54   0.000     .1465669    .2340046
       _cons |   6.964355   .3198205    21.78   0.000     6.336899     7.59181
------------------------------------------------------------------------------
```

i) The slope estimate of **motheduc** indicates that holding **fatheduc** constant, when **motheduc** increases by 1, **educ** would increase by 0.3041971. Similarly, the slope estimate of **fatheduc** indicates that holding **motheduc** constant, when **fatheduc** increases by 1, **educ** would increase by 0..1902858.

ii) $0.3041971 + 0.1902858 = 0.4944829$. *educ* is expected to increase by 0.4944829 years.

iii) Yes. It means that with 0 education of both mother and father, a person would have at least 6.964355 years of education.

iv) The R-squared is 0.2493, so the correlation between y and $\widehat{y}$ is $\sqrt{0.2493} = 0.49929951$. This serves as an upper bound on the magnitude of the correlation between **y** and **motheduc** beceause if all the correlation comes from motheduc which means all the variation in the model is captured by motheduc only, the correlation between y and motheduc would be the same as the one we get from the model.

```
. gen parenteduc = motheduc + fatheduc

. regr educ parenteduc

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(1, 1228)      =    400.43
       Model |  1675.13029         1  1675.13029   Prob > F        =    0.0000
    Residual |  5137.14938     1,228  4.1833464   R-squared       =    0.2459
-------------+----------------------------------   Adj R-squared   =    0.2453
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.0453

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  parenteduc |   .2346699   .0117272    20.01   0.000     .2116623    .2576775
       _cons |   7.258604   .2946149    24.64   0.000       6.6806    7.836608
------------------------------------------------------------------------------
```

v) The new R-squared is $0.2459 < 0.2493$. The R-squared has declined because it captures less variation of the y variable, as two covariates have been collapsed. The magnitude of the parenteduc slope estimate, which is .2346699, seems sensible because it lies between the magnitude of matheduc, which is .3041971, and that of fatheduc, which is .1902858. An implicit restriction this model is facing is that, by only using the sum, it assumes that mother's and father's length of education have an equal effect on a child's years of education when Model 1 shows that **motheduc** has more effect than **fatheduc**.

(b)

```
. gen educinteraction = motheduc*fatheduc

. regr educ motheduc fatheduc educinteraction

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =    138.39
       Model |  1723.29513         3  574.431709   Prob > F        =    0.0000
    Residual |  5088.98455     1,226  4.15088462   R-squared       =    0.2530
-------------+----------------------------------   Adj R-squared   =    0.2511
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.0374

-------------------------------------------------------------------------------
           educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+--------------------------------------------------------------
       motheduc |   .1408771   .0733931     1.92   0.055    -.0031129    .2848672
       fatheduc |   .0285246   .0691587     0.41   0.680    -.1071579    .1642071
educinterac~n |    .013201   .0053442     2.47   0.014     .0027163    .0236858
          _cons |   8.906952   .8487192    10.49   0.000     7.241849    10.57205
-------------------------------------------------------------------------------
```

i) It might be expected that the partial effect of on variable changes as another variable changes.

ii) The partial effect of **motheduc** grows by 0.013201 when **fatheduc** grows by 1 and vice versa.

iii) It is 0.1408771 + 0.013201***fatheduc**. It depends on the value of fatheduc.

iv) $0.1408771 + 0.013201*\textbf{fatheduc} + 0.0285246 + 0.013201*\textbf{motheduc} = 0.1694017 + 0.013201(\textbf{fatheduc} + \textbf{motheduc})$

v) The R-squared is $0.2530 > 0.2459$. It was expected because we added an additional variable.

(c)

```
. regr educ motheduc fatheduc abil

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =    305.17
       Model |  2912.30705         3  970.769018   Prob > F        =    0.0000
    Residual |  3899.97262     1,226  3.18105434   R-squared       =    0.4275
-------------+----------------------------------   Adj R-squared   =    0.4261
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7836

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .1891314   .0285062     6.63   0.000     .1332051    .2450578
    fatheduc |   .1110854   .0198849     5.59   0.000     .0720733    .1500976
        abil |   .5024829    .025718    19.54   0.000     .4520268     .552939
       _cons |    8.44869   .2895407    29.18   0.000      7.88064     9.01674
------------------------------------------------------------------------------
```

i) When **abil** increases by one, holding other variables constant, **educ** is expected to increase by .5024829.

ii) The slope estimates of **motheduc** and **fatheduc** have declined from Model 1. This is expected because it is likely that longer education is correlated with higher **abil** and it is now included in the model to capture its effect.

iii)

```
. sum abil

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        abil |      1,230    1.796596    2.184406  -5.631463   6.263742

. gen abilstd = abil/2.184406

. regr educ motheduc fatheduc abilstd

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =    305.17
       Model |  2912.30706         3  970.769019   Prob > F        =    0.0000
    Residual |  3899.97262     1,226  3.18105434   R-squared       =    0.4275
-------------+----------------------------------   Adj R-squared   =    0.4261
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7836

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .1891314   .0285062     6.63   0.000     .1332051    .2450578
    fatheduc |   .1110854   .0198849     5.59   0.000     .0720733    .1500976
     abilstd |   1.097627   .0561785    19.54   0.000     .9874101    1.207843
       _cons |    8.44869   .2895407    29.18   0.000      7.88064     9.01674
------------------------------------------------------------------------------
```

The slope estimate of abilstd is larger than that of abil, because the regression is done on smaller values now. It is easier to interpret the slope estimate of abilstd than that of abil because we can see the relation between a standard deviation change of ability instead of the raw numbers of ability. None of the other estimates, including R-squared, change.

iv)

```
. replace abilstd = abilstd - 1.796596
(1,230 real changes made)

. regr educ motheduc fatheduc abilstd

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =    305.17
       Model |  2912.30706         3   970.76902   Prob > F        =    0.0000
    Residual |  3899.97262     1,226  3.18105434   R-squared       =    0.4275
-------------+----------------------------------   Adj R-squared   =    0.4261
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7836

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .1891314   .0285062     6.63   0.000     .1332051    .2450578
    fatheduc |   .1110854   .0198849     5.59   0.000     .0720733    .1500976
     abilstd |   1.097627   .0561785    19.54   0.000     .9874101    1.207843
       _cons |   10.42068   .3306905    31.51   0.000       9.7719    11.06946
------------------------------------------------------------------------------
```

The only change happens here is the intercept. It is because **abilstd** is now centered at an average ability, instead of an ability of zero.

v)

```
. regr abilstd motheduc fatheduc

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(2, 1227)      =    134.56
       Model |  221.069505         2  110.534753   Prob > F        =    0.0000
    Residual |  1007.93094     1,227  .821459609   R-squared       =    0.1799
-------------+----------------------------------   Adj R-squared   =    0.1785
```

```
       Total |  1229.00045     1,229  1.00000036   Root MSE        =    .90634

------------------------------------------------------------------------------
     abilstd | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .1048314   .0141734     7.40   0.000     .0770245    .1326382
    fatheduc |    .072156   .0098927     7.29   0.000     .0527475    .0915644
       _cons |   -3.14891   .1419803   -22.18   0.000    -3.427461   -2.870359
------------------------------------------------------------------------------

. predict uhat, residuals

. regr educ uhat

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(1, 1228)      =    266.39
       Model |  1214.33946         1  1214.33946   Prob > F        =    0.0000
    Residual |  5597.94022     1,228  4.55858324   R-squared       =    0.1783
-------------+----------------------------------   Adj R-squared   =    0.1776
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.1351

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        uhat |   1.097627   .0672511    16.32   0.000     .9656869    1.229567
       _cons |    13.0374   .0608783   214.16   0.000     12.91796    13.15684
------------------------------------------------------------------------------
```

It can be verified that the coefficient of the abilstd in this regression is the same as in the previous model. This shows that the partitioned regression approach works.

vi) To account for different effects of **abilstd** depending on **motheduc**, I would introduce an interaction variable of **motheduc** and **abilstd**. If the hypothesis was correct, I would expect the coefficeint of **motheduc*abilstd** would be positive.

```
. gen mothabil = motheduc*abilstd

. regr educ motheduc fatheduc abilstd mothabil

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(4, 1225)      =    240.96
       Model |  2999.70914         4  749.927285   Prob > F        =    0.0000
    Residual |  3812.57053     1,225  3.11230248   R-squared       =    0.4403
-------------+----------------------------------   Adj R-squared   =    0.4385
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7642

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    motheduc |   .2978231   .0348672     8.54   0.000     .2294171    .3662292
    fatheduc |   .1081562   .0196766     5.50   0.000     .0695527    .1467597
     abilstd |  -.1101569   .2345894    -0.47   0.639    -.5703984    .3500845
    mothabil |    .104223   .0196672     5.30   0.000     .0656378    .1428082
       _cons |   9.103226   .4108515    22.16   0.000     8.297175    9.909276
------------------------------------------------------------------------------
```

In the new regression, the estimated partial effect of ability is $-0.110157 + 0.104223 *$ **motheduc**. The estimated partial effect of changing **motheduc** by a year is $0.2978231 + 0.104223 *$ **abilstd**.

(d)

```
. gen abilstd_squared = abilstd^2

. regr educ motheduc fatheduc abilstd abilstd_squared

      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(4, 1225)      =    244.91
       Model |  3027.03707         4  756.759267   Prob > F        =    0.0000
    Residual |  3785.24261     1,225  3.08999397   R-squared       =    0.4444
-------------+----------------------------------   Adj R-squared   =    0.4425
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7578

------------------------------------------------------------------------------
           educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+----------------------------------------------------------------
       motheduc |   .1901261   .0280957     6.77   0.000     .1350051    .2452472
       fatheduc |   .1089387   .0196014     5.56   0.000     .0704827    .1473946
        abilstd |   1.744496   .1197306    14.57   0.000     1.509596    1.979396
abilstd_squ~d |   .2414397   .0396232     6.09   0.000      .163703    .3191765
          _cons |   10.59507   .3271771    32.38   0.000     9.953183    11.23696
------------------------------------------------------------------------------
```

i) The relationship between a covariate and the dependent variable might be expected to have a non-linear but possibly quadratic relationship.

ii) $1.744496 + 2 * 0.2414397$**ablistd**.

iii) The estimated partial effect of **abilstd** in Model 3 is 1.097627, no matter the value of **abilstd** is. On the other hand, the estimated partial effect of **abilstd** in Model 4 depends on the value of **abilstd**. It increases by 0.4828794 when **abilstd** increases by 1.

(e) i)

```
. regr west18 ne18 nc18 south18
```

```
      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =         .
       Model |  152.950407         3  50.9834688   Prob > F        =         .
    Residual |           0     1,226           0   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  152.950407     1,229  .124451104   Root MSE        =         0

------------------------------------------------------------------------------
      west18 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        ne18 |         -1          .        .       .            .           .
        nc18 |         -1          .        .       .            .           .
     south18 |         -1          .        .       .            .           .
       _cons |          1          .        .       .            .           .
------------------------------------------------------------------------------
```

Yes, I get the expected results. It is because every observation belongs to one of the regions. Because of this perfect colinearity, we get R-squared of 1.

ii)

```
. regr educ west18
      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(1, 1228)      =      1.26
       Model |   6.9886564         1   6.9886564   Prob > F        =    0.2617
    Residual |  6805.29102     1,228  5.54176793   R-squared       =    0.0010
-------------+----------------------------------   Adj R-squared   =    0.0002
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.3541

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      west18 |  -.2137576   .1903482    -1.12   0.262    -.5872013    .1596862
       _cons |   13.06851   .0726144   179.97   0.000     12.92604    13.21097
------------------------------------------------------------------------------
```

Compared to all the other regions, years of education was lower by 0.2137576 on average if the child was in the west when 18.

iii)

```
. regr educ west18 ne18 nc18
      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(3, 1226)      =      7.38
       Model |  120.814836         3  40.2716118   Prob > F        =    0.0001
    Residual |  6691.46484     1,226   5.4579648   R-squared       =    0.0177
-------------+----------------------------------   Adj R-squared   =    0.0153
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    2.3362

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      west18 |   .2729996   .2263717     1.21   0.228    -.1711192    .7171184
        ne18 |   .9129506    .200097     4.56   0.000     .5203801    1.305521
        nc18 |   .5014193   .1776529     2.82   0.005     .1528818    .8499567
       _cons |   12.58175    .144058    87.34   0.000     12.29912    12.86438
------------------------------------------------------------------------------
```

Compared to the south, years of education was higher by 0.2729996 on average if the child was in the west when 18.

iv)

```
. regr educ motheduc fatheduc abilstd abilstd_squared west18 ne18 nc18
      Source |       SS           df       MS      Number of obs   =     1,230
-------------+----------------------------------   F(7, 1222)      =    141.80
       Model |  3053.34975         7  436.192822   Prob > F        =    0.0000
    Residual |  3758.92992     1,222   3.0760474   R-squared       =    0.4482
-------------+----------------------------------   Adj R-squared   =    0.4451
       Total |  6812.27967     1,229  5.54294522   Root MSE        =    1.7539

--------------------------------------------------------------------------------
           educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
----------------+---------------------------------------------------------------
       motheduc |   .1955234   .0281776     6.94   0.000     .1402415    .2508053
       fatheduc |   .1053686   .0196234     5.37   0.000     .0668693    .1438679
        abilstd |   1.718812   .1198295    14.34   0.000     1.483718    1.953907
 abilstd_squ~d  |   .2362144    .039619     5.96   0.000     .1584855    .3139432
         west18 |  -.1507768   .1713097    -0.88   0.379    -.4868705    .1853169
           ne18 |   .3056072    .151779     2.01   0.044     .0078309    .6033835
           nc18 |   .1415781    .134219     1.05   0.292    -.1217471    .4049033
          _cons |   10.45244   .3389361    30.84   0.000      9.78748     11.1174
--------------------------------------------------------------------------------
```

Holding **motheduc**, **fatheduc** and **abilstd** constant, years of education was lower by 0.1507768 on average if the child was in the west when 18 compared to the south.

v) If the region did not matter, the slopes of **west18**, **ne18**, **nc18** would be all 0.

vi)

```
. regr educ motheduc fatheduc abilstd abilstd_squared south18 ne18 nc18
```

Source	SS	df	MS			
Model	3053.34975	7	436.192822	Number of obs	=	1,230
Residual	3758.92992	1,222	3.0760474	F(7, 1222)	=	141.80
				Prob > F	=	0.0000
Total	6812.27967	1,229	5.54294522	R-squared	=	0.4482
				Adj R-squared	=	0.4451
				Root MSE	=	1.7539

educ	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
motheduc	.1955234	.0281776	6.94	0.000	.1402415	.2508053
fatheduc	.1053686	.0196234	5.37	0.000	.0668693	.1438679
abilstd	1.718812	.1198295	14.34	0.000	1.483718	1.953907
abilstd_squ~d	.2362144	.039619	5.96	0.000	.1584855	.3139432
south18	.1507768	.1713097	0.88	0.379	-.1853169	.4868705
ne18	.456384	.1682742	2.71	0.007	.1262457	.7865223
nc18	.2923549	.1536704	1.90	0.057	-.0091322	.593842
_cons	10.30166	.3609697	28.54	0.000	9.593476	11.00985

Now, the west is captured in the intercept. Therefore, the slopes of **motheduc**, **fatheduc** and **abilstd** would stay the same while the coefficients of the indicator variables would change, now showing the difference between the corresponding region and the west. The results show that the slopes for **motheduc**, **fatheduc** and **abilstd** stay the same.

Question 5

```
. clear

. use "/Users/steven_unique/My Drive/FALL 2021/ECO 394M ECONOMETRICS/Problem Set
> s/PS1/FERTIL2.DTA"

. gen heducmissing = (heduc==.)

. replace heduc = 0 if heducmissing
(2,405 real changes made)
```

(a)

```
. gen age2 = age^2

. regr children age age2 educ evermarr heduc heducmissing

      Source |       SS           df       MS      Number of obs   =     4,361
-------------+----------------------------------   F(6, 4354)      =   1048.85
       Model |  12723.9061         6  2120.65102   Prob > F        =    0.0000
    Residual |  8803.27023     4,354  2.02188108   R-squared       =    0.5911
-------------+----------------------------------   Adj R-squared   =    0.5905
       Total |  21527.1763     4,360  4.93742577   Root MSE        =    1.4219

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .2786184   .0170281    16.36   0.000     .2452347    .3120021
        age2 |  -.0020174   .0002739    -7.36   0.000    -.0025544   -.0014803
        educ |  -.0612256   .0065688    -9.32   0.000    -.0741039   -.0483474
    evermarr |   .1754812   .1356173     1.29   0.196    -.0903978    .4413601
       heduc |  -.0646732    .007649    -8.46   0.000    -.0796692   -.0496772
heducmissing |  -.8460843   .1369495    -6.18   0.000    -1.114575   -.5775937
       _cons |  -2.809784   .2781238   -10.10   0.000    -3.355048    -2.26452
------------------------------------------------------------------------------
```

i) $.2786184 + -2 * 0.0020174\mathbf{age}$

ii) 0.1754812

(b)

```
. regr children age age2 educ heduc heducmissing if evermarr == 1

      Source |       SS           df       MS      Number of obs   =     2,079
-------------+----------------------------------   F(5, 2073)      =    300.63
       Model |  4563.36644         5  912.673289   Prob > F        =    0.0000
    Residual |  6293.42144     2,073  3.03590036   R-squared       =    0.4203
-------------+----------------------------------   Adj R-squared   =    0.4189
       Total |  10856.7879     2,078  5.22463324   Root MSE        =    1.7424

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .5254392   .0390565    13.45   0.000     .4488451    .6020333
        age2 |  -.0054568    .000581    -9.39   0.000    -.0065961   -.0043175
        educ |  -.0759259   .0118672    -6.40   0.000    -.0991988    -.052653
       heduc |  -.0551317   .0106142    -5.19   0.000    -.0759472   -.0343162
heducmissing |  -.8397852   .1683276    -4.99   0.000    -1.169894   -.5096763
       _cons |  -6.777208   .6338617   -10.69   0.000     -8.02028   -5.534136
------------------------------------------------------------------------------

. regr children age age2 educ heduc heducmissing if evermarr == 0
note: heduc omitted because of collinearity.
note: heducmissing omitted because of collinearity.

      Source |       SS           df       MS      Number of obs   =     2,282
-------------+----------------------------------   F(3, 2278)      =    972.55
       Model |  2956.05077         3  985.350258   Prob > F        =    0.0000
    Residual |  2307.97026     2,278  1.01315639   R-squared       =    0.5616
-------------+----------------------------------   Adj R-squared   =    0.5610
       Total |  5264.02103     2,281  2.30776898   Root MSE        =    1.0066

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .2290825   .0176561    12.97   0.000     .1944587    .2637062
        age2 |  -.0014574   .0003161    -4.61   0.000    -.0020773   -.0008375
        educ |  -.0549112   .0064503    -8.51   0.000    -.0675604    -.042262
       heduc |          0  (omitted)
heducmissing |          0  (omitted)
       _cons |   -2.87956   .2326078   -12.38   0.000    -3.335706   -2.423415
------------------------------------------------------------------------------
```

The effects of age and education are different in the two regressions. They look larger in the first regression where **evermarr** is 1. Stata drops **heduc** and **heducmissing** because when **evermarr** is 0, **heduc** is 0 and **heducmissing** is 1, which means these three are perfectly correlated in which case the inverse of $X'X$ cannot be calculated and therefore an unique OLS estimator cannot be calculated.

(c)

```
. regr children age age2 educ electric
```

```
      Source |       SS           df       MS      Number of obs   =     4,358
-------------+----------------------------------   F(4, 4353)      =   1451.87
       Model |   12294.619         4  3073.65474   Prob > F        =    0.0000
    Residual |  9215.41316     4,353  2.11702577   R-squared       =    0.5716
-------------+----------------------------------   Adj R-squared   =    0.5712
       Total |  21510.0321     4,357  4.93689055   Root MSE        =     1.455

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .3370397   .0165166    20.41   0.000     .3046588    .3694206
        age2 |  -.0026696   .0002718    -9.82   0.000    -.0032025   -.0021367
        educ |  -.0788944   .0062513   -12.62   0.000    -.0911502   -.0666387
    electric |  -.3777901   .0673176    -5.61   0.000    -.5097669   -.2458133
       _cons |  -4.247568   .2406003   -17.65   0.000    -4.719267   -3.775869
------------------------------------------------------------------------------
```

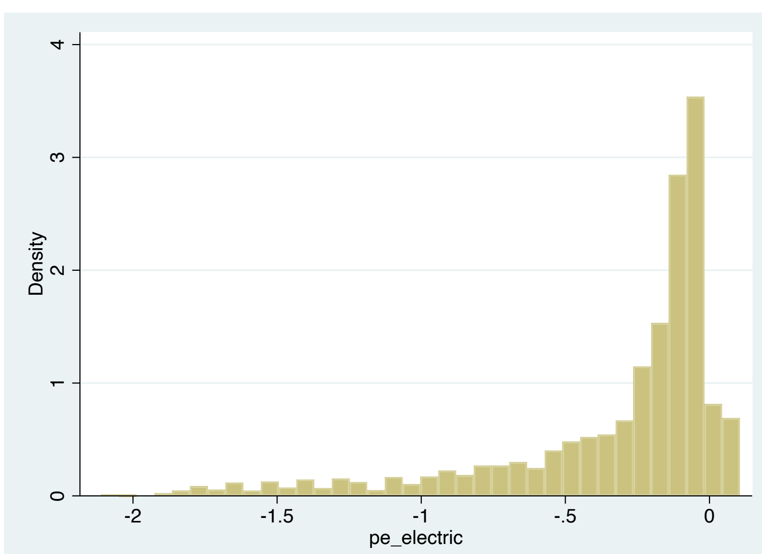
The estimated partial effect of **electric** is -0.3777901. Since it is not reasonable to assume that having electricity would negatively affect the number of children a woman would have, we should not think of the estimated partial effect as the true causal effect but rather think that there would be a cofounder.

(d)

```
. gen ageelectric = age*electric
(3 missing values generated)

. gen age2electric = age2*electric
(3 missing values generated)

. gen educelectric = educ*electric
(3 missing values generated)

. regr children age age2 educ electric ageelectric age2electric educelectric

      Source |       SS           df       MS      Number of obs   =     4,358
-------------+----------------------------------   F(7, 4350)      =    844.60
       Model |  12392.2517         7  1770.32168   Prob > F        =    0.0000
    Residual |  9117.78038     4,350  2.09604147   R-squared       =    0.5761
-------------+----------------------------------   Adj R-squared   =    0.5754
       Total |  21510.0321     4,357  4.93689055   Root MSE        =    1.4478

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .3271644   .0177036    18.48   0.000     .2924564    .3618723
        age2 |  -.0023715   .0002912    -8.14   0.000    -.0029424   -.0018006
        educ |  -.0687357   .0069977    -9.82   0.000    -.0824548   -.0550165
    electric |   -.599813    .695946    -0.86   0.389    -1.964222    .7645958
 ageelectric |   .0754904   .0482619     1.56   0.118    -.0191275    .1701083
age2electric |  -.0020197   .0007896    -2.56   0.011    -.0035677   -.0004717
educelectric |  -.0212059   .0163257    -1.30   0.194    -.0532126    .0108008
       _cons |   -4.27699     .25999   -16.45   0.000    -4.786703   -3.767277
------------------------------------------------------------------------------
```

The estimated partial effect of electric is $-0.599813 + 0.0754904 * \mathbf{age} - 0.0020197 * \mathbf{age}^2 + -0.0212059\mathbf{educ}$.

```
. gen pe_electric = _b[electric] + _b[ageelectric]*age + _b[age2electric]*age2 +
>  _b[educelectric]*educ

. hist pe_electric
(bin=36, start=-2.1105685, width=.06155454)
```



(e)

```
. regr children age age2 educ if electric == 0
```

```
      Source |       SS           df       MS      Number of obs   =     3,747
-------------+----------------------------------   F(3, 3743)      =   1746.95
       Model |  11334.5009         3  3778.16697   Prob > F        =    0.0000
    Residual |  8095.04807     3,743  2.16271656   R-squared       =    0.5834
-------------+----------------------------------   Adj R-squared   =    0.5830
       Total |   19429.549     3,746  5.18674559   Root MSE        =    1.4706

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .3271644   .0179829    18.19   0.000     .2919071    .3624216
        age2 |  -.0023715   .0002958    -8.02   0.000    -.0029514   -.0017916
        educ |  -.0687357   .0071082    -9.67   0.000    -.0826719   -.0547994
       _cons |   -4.27699   .2640927   -16.20   0.000     -4.79477    -3.75921
------------------------------------------------------------------------------

. regr children age age2 educ if electric == 1
      Source |       SS           df       MS      Number of obs   =       611
-------------+----------------------------------   F(3, 607)       =    190.12
       Model |  960.976371         3  320.325457   Prob > F        =    0.0000
    Residual |   1022.7323       607  1.68489671   R-squared       =    0.4844
-------------+----------------------------------   Adj R-squared   =    0.4819
       Total |  1983.70867       610  3.25198143   Root MSE        =     1.298

------------------------------------------------------------------------------
    children | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .4026547   .0402541    10.00   0.000     .3236006    .4817089
        age2 |  -.0043912    .000658    -6.67   0.000    -.0056835   -.0030989
        educ |  -.0899416   .0132244    -6.80   0.000    -.1159128   -.0639704
       _cons |  -4.876803   .5787919    -8.43   0.000    -6.013481   -3.740125
------------------------------------------------------------------------------
```

The coefficient estimates for **age**, **age2**, and **educ** in the regression for the electric = 0 are the same as in the model with interactions in part (d).