```
--------------------------------------------------------------------------------
      name:  PK
       log:
/Users/priyakoirala/Desktop/school/econometrics/projects/project3/koirala_project3.log
  log type:  text
 opened on:  27 Mar 2023, 21:08:56


. /*=====================================================================================
> The purpose of this exercise is to show how college tuition is related to
> adult men's highest education level. The hypothesis is that men pursue more
> education when tuition costs are lower.
>
> Open the HTV.dta data set. It includes data on a random sample of men in 1991.
> ===================================================================================*/
.
. use "/Users/priyakoirala/Desktop/school/econometrics/projects/project3/HTV.dta"
```

```
.
. /*==================================================================================
> (Q1): Use Stata's "sum, detail" command to show more detailed summary statistics
> for the tuit18 variable. Tuit18 is the average annual tuition (measured in $1000s)
> at nearby colleges when the men are 18 years old.
>
> What is the 75th percentile of college tuition in the sample? What does it mean?
> ==================================================================================*/
.

 sum tuit18, detail

                   college tuition, age 18
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%     .4102407              0
10%     2.444079              0       Obs               1,193
25%     6.057975              0       Sum of wgt.       1,193

50%     8.826549                      Mean           8.557239
                        Largest       Std. dev.      4.042644
75%     11.15503       18.17392
90%     14.16312       18.17392       Variance       16.34297
95%     15.00826       18.17392       Skewness      -.2158796
99%     18.17392       18.17392       Kurtosis       2.711644
```

**The 75th percentile of college tuition in the sample is $11,155.03. It means that 75% of the men who were sampled have college tuition expenses that are either equal or lower than this value.**

```
.
. /*==============================================================================
> (Q2): Estimate a multivariable regression relating men's level of education
> (Y=educ) to nearby college tuition at age 18 (X1=tuit18), their mother's
> education (X2=motheduc), their father's education (X3=fatheduc), a binary
> variable that equals 1 if they lived in the Northeastern US at age 18 (X4=ne18),
> a binary variable that equals 1 if they lived in the North-central US at age 18
> (X5=nc18), and a binary variable that equals 1 if they lived in the Southern US
> at age 18 (X6=south18). Note that all men lived either in the Northeast US, the
> North-central US, the Western US, or the Southern US.
>
> Use heteroskedasticity-robust standard errors.
>
> Interpret beta1hat in a sentence.
> ==============================================================================*/
.
. reg educ tuit18 motheduc fatheduc ne18 nc18 south18, robust

Linear regression                                 Number of obs     =      1,193
                                                  F(6, 1186)        =      63.10
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.2629
                                                  Root MSE          =     2.0194

------------------------------------------------------------------------------
             |               Robust
        educ | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      tuit18 |  -.0099148   .0245582    -0.40   0.686    -.0580972    .0382677
    motheduc |   .3198548    .038944     8.21   0.000      .243448    .3962615
    fatheduc |   .1771686   .0262213     6.76   0.000     .1257233    .2286139
        ne18 |     .77052   .2867912     2.69   0.007     .2078453    1.333195
        nc18 |   .5408396   .2531378     2.14   0.033     .0441917    1.037487
     south18 |   .2062903   .2145603     0.96   0.337    -.2146698    .6272503
       _cons |   6.577948   .4133719    15.91   0.000     5.766927     7.38897
------------------------------------------------------------------------------


 di e(r2_a)
.25920214
```

**beta1hat is -0.0099148**

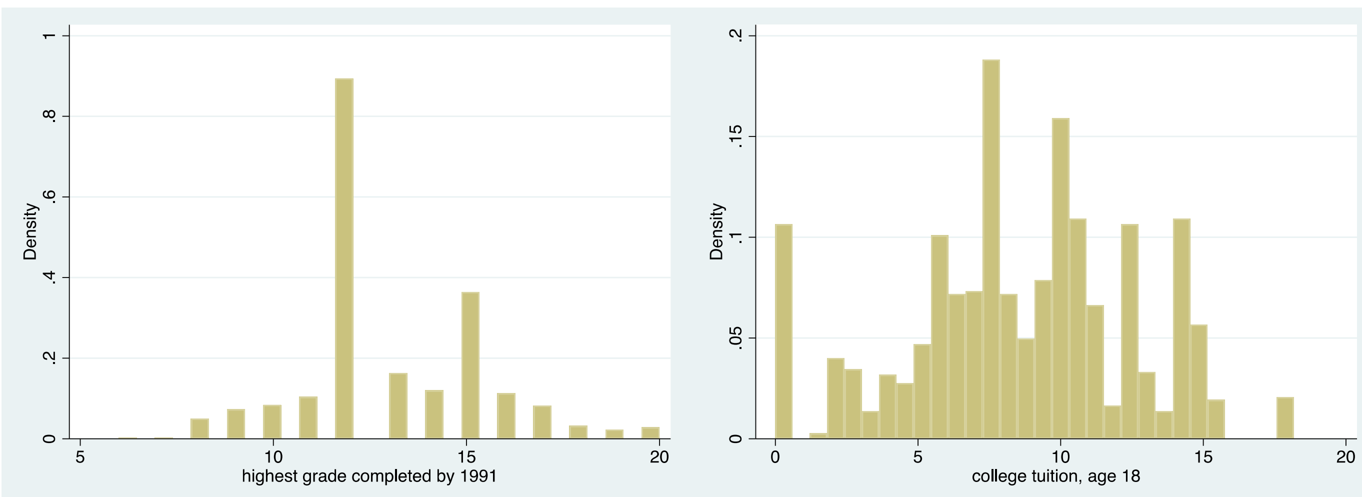
**Other variables held constant, on average, it is estimated that for every $1000 increase in tuition, the level of education for the men in the sample decreases by approximately 0.01 percentage points.**

```
.
. /*==========================================================================
> (Q3): Are the errors likely to be normally distributed in the model in (Q2)?
> Why would it matter?
> ==========================================================================*/
.
. histogram educ
(bin=30, start=6, width=.46666667)

. graph export "/Users/priyakoirala/Desktop/school/econometrics/projects/project3/Graph1_project3.pdf" file
/Users/priyakoirala/Desktop/school/econometrics/projects/project3/Graph1_project3.pdf saved as PDF format

. histogram tuit18
(bin=30, start=0, width=.6057972)

. graph export "/Users/priyakoirala/Desktop/school/econometrics/projects/project3/Graph2_project3.pdf" file
/Users/priyakoirala/Desktop/school/econometrics/projects/project3/Graph2_project3.pdf saved as PDF format
```




**The errors are unlikely to be normally distributed in the model in (Q2). When we make histogram of the educ variable, we can see that there is a high level of density at the 12-grade level. When we make a histogram of the tuit18, variable we can see that there is skew on the left side of the graph. The data does not follow a bell-shaped curve, which is associated with a normal distribution.**

**The assumption that errors are distributed normally matters because, although the assumption is not required for calculating unbiased estimates, checking the data distribution allows us a better understanding of our data. It allows us insight to any outliers we may have missed, or any other factors which could possibly have caused an extreme deviation in the data. If our data is way beyond the range of a "normal distribution", it could be difficult to draw a valid conclusion from our statistical inferences in the regression model.**

**However, the normality assumption is not necessary for the validity of our regression model as data can naturally vary. Presumably, most adult men have completed some form of high school therefore, there is higher density at that grade level. In addition, it would make sense that there is a skew towards the left side regarding tuition, as there is a smaller population of people who can afford extremely high tuition costs.**

```
.
. /*==============================================================================
> (Q4): Interpret beta6hat in a sentence.
> ==============================================================================*/
.
```

**beta6hat is 0.2062903.**

**This means that, all other variables held constant, it is estimated that for men aged 18 who lived in the southern region of the United States, on average, had 0.21 more years of education than men who lived in the western (reference) region of the United States.**

```
.
. /*=============================================================================
> (Q5): Why is the variable West excluded from the model in (Q2)?
> =============================================================================*/
.
```

**The variable West is excluded from the model in (Q2) because it is used as a reference for the three binary region variables (ne18, nc18, south18) which are included in the model. If we were to include all four regional variables, the regression would result in perfect multicollinearity, which would not allow us to gain an accurate estimate of the regression coefficients. Though perfect multicollinearity itself does not generate bias in out model, it makes it difficult to interpret the regression coefficients as it inflates the standard errors of our estimates. It also makes it more difficult to identify the independent variables that are statistically significant.**

```
.
. /*===============================================================================
> (Q6) Test the joint statistical significance of beta4, beta5, and
> beta6. Use alpha=0.05. Write down the null and alternative hypotheses. How many
> restrictions are there? What do you conclude?
> ===============================================================================*/
.
. test ne18 nc18 south18

 ( 1)  ne18 = 0
 ( 2)  nc18 = 0
 ( 3)  south18 = 0

       F(  3,  1186) =    2.64
            Prob > F =    0.0483
```

```
>    H_0: beta4 = beta5 = beta6 = 0
>    H_1: beta5 != 0 &/or beta5 !=0 &/or beta6 = !0
>
>    This hypotheses test has three restictions.
>
>    P-value for the F statistic = 0.0483 < 0.05
```

**We reject the null hypothesis of no statistical significance at the 5% level and we accept the alternate hypothesis that beta4 beta5 and beta6 are jointly statistically significant. We conclude that the geographic locations Northeastern US, North-central US, and Southern US where men aged 18 attended college explains the variance in adult men's highest education level. We should keep the variables in the model.**

```
.
. /*============================================================================
> (Q7) Add the variable abil (a measure of cognitive ability) to the model from
> (Q2). Does ability help explain the variation in education, even after
> controlling for tuition, parents’ education, and geographic region? Use at
> least one test statistic to justify your answer.
> ============================================================================*/
.
. reg educ tuit18 motheduc fatheduc ne18 nc18 south18 abil, robust

Linear regression                               Number of obs     =      1,193
                                                F(7, 1185)        =     112.15
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4294
                                                Root MSE          =     1.7775

------------------------------------------------------------------------------
             |               Robust
        educ | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      tuit18 |  -.0189641   .0209396    -0.91   0.365    -.0600469    .0221187
    motheduc |   .1992246   .0346079     5.76   0.000      .131325    .2671241
    fatheduc |   .1049573   .0237609     4.42   0.000     .0583392    .1515755
        ne18 |   .6777296   .2460138     2.75   0.006     .1950583    1.160401
        nc18 |   .4665593    .218089     2.14   0.033     .0386757    .8944429
     south18 |   .2606951   .1869486     1.39   0.163    -.1060921    .6274823
        abil |   .4909632   .0281367    17.45   0.000     .4357599    .5461666
       _cons |   8.176346   .3870015    21.13   0.000     7.417062    8.935631
------------------------------------------------------------------------------


 di e(r2_a)
.42600577
```

**Yes, the variable abil explains the variation in education, even after controlling for tuition, parent's education, and geographic region. I.e., beta7 is statistically significant.**

**The adjusted R2 increases from 0.25920 in Q2 to 0.42601 when we add the variable abil. A higher adjusted R2 means that the model has improved.**

```
> H_0: beta7 = 0
> H_1: beta1 != 0
>
> t-statistic = (beta7hat-0)/std error of beta7hat
>                    = 0.490 / 0.028
>                    = |17.5| > 2.9
>                    = critical value for two sided alternative where alpha = 0.05
```

**So using t-statistic, we reject the null hypothesis of no statistical significance at the 5%. We conclude that a measure of cognitive ability is useful in determining adult men's highest education level. We should keep the variable abil in the model.**

```
.
. /*==========================================================================
> (Q8) Compare the coefficient estimates from the model in (Q2) to the estimates
> obtained from the model in (Q7). What do the differences tell us about omitted
> variable bias in the model in (Q2)?
> ========================================================================*/
.
```

**Omitting a variable can lead to either an overestimation or underestimation of the coefficient of our independent variable. The coefficients become unreliable, preventing the estimator from converging a probability to the true parameter value.**

**As we can see in the model from Q2, beta1hat has decreased from -0.00991 to -0.01896, beta2hat decreased from 0.31985 to 0.19922, beta3hat decreased from 0.17717 to 0.10496, beta4hat decreased from 0.77052 to 0.67772, beta5hat from 0.54084 to 0.46656, beta6hat increased from 0.020629 to 0.26070.**

**These coefficients suggests that there was previously an upwards bias in our model in Q2, which has been corrected with the addition of the variable abil to the model in Q7.**

```
.
. /*=============================================================================
> (Q9) Test whether the relationship between father's education and adult son's
> education is the same as the relationship between mother's education and
> adult son's education. Use alpha=0.05. Write down the null and alternative
> hypotheses. How many restrictions are there? What do you conclude?
> ============================================================================*/
.
. test motheduc - fatheduc = 0

 ( 1)  motheduc - fatheduc = 0

       F(  1,  1185) =    3.26
            Prob > F =    0.0712
```

**>    H_0: beta2 - beta3 = 0**
**>    H_1: beta2 != 0 &/or beta3 != 0**
**>**
**>    This hypotheses test has one restriction.**
**>**
**>    P-value for the F-statistic: |0.0483| < 0.05**

**We reject the null hypothesis of no statistical significance at the 5% level. We conclude that the relationship between mother's education and adult son's education differs from the relationship between father's education and adult son's education. We should keep both variables in the estimate.**

```
.
. /*============================================================================
> (Q10) Consider the variable called tuit17, which is the average tuition of
> nearby colleges when the men are 17 years old. Should we add this variable to
> the model in (Q7)? Why or why not?
> ============================================================================*/
.
. reg educ tuit18 motheduc fatheduc ne18 nc18 south18 abil tuit17, r

Linear regression                               Number of obs     =      1,193
                                                F(8, 1184)        =      98.05
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4294
                                                Root MSE          =     1.7783

------------------------------------------------------------------------------
             |               Robust
        educ | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      tuit18 |  -.0204591   .0508123    -0.40   0.687    -.1201513     .079233
    motheduc |   .1992316   .0346283     5.75   0.000      .131292    .2671712
    fatheduc |   .1049772    .023796     4.41   0.000     .0582902    .1516641
        ne18 |   .6769675   .2497603     2.71   0.007     .1869453     1.16699
        nc18 |   .4659347   .2212261     2.11   0.035     .0318957    .8999736
     south18 |   .2604994   .1872257     1.39   0.164    -.1068319    .6278306
        abil |    .490936   .0281688    17.43   0.000     .4356697    .5462023
      tuit17 |   .0015459    .050493     0.03   0.976    -.0975199    .1006117
       _cons |   8.176131     .38755    21.10   0.000      7.41577    8.936492
------------------------------------------------------------------------------

di e(r2_a)
 .42552126
```

**No, we should not include the tuit17 variable into the model in (Q7), When we add variable tuit17 to our model, our adjusted r squared slightly decreases from 0.4260 to 0.4255 which suggests that the model is not a good fit.**

```
. corre tuit17 tuit18 motheduc fatheduc ne18 nc18 south18 abil
 (obs=1,193)

             |   tuit17   tuit18 motheduc fatheduc     ne18     nc18  south18     abil
-------------+------------------------------------------------------------------------
      tuit17 |   1.0000
      tuit18 |   0.9803   1.0000
    motheduc |  -0.0524  -0.0493   1.0000
    fatheduc |  -0.0056  -0.0000   0.5947   1.0000
        ne18 |   0.3641   0.3581   0.0529   0.0804   1.0000
        nc18 |   0.3738   0.3727  -0.0500  -0.0028  -0.4545   1.0000
     south18 |  -0.3906  -0.3881  -0.0670  -0.1057  -0.2841  -0.4371   1.0000
        abil |   0.0607   0.0556   0.3902   0.3805   0.0707   0.0261  -0.1004   1.0000
```

**Additionally, based on the data above, the correlation coefficient between tuit17 and tuit18 is 0.9803.**

**This means that there is an extremely high correlation between tuition for men aged 17 and tuition for men aged 18. By including both variables into the model, we risk having imperfect multicollinearity. Though imperfect multicollinearity itself does not generate bias in out model, it makes it difficult to interpret the regression**

**coefficients as it inflates the standard errors of our estimates. It also makes it more difficult to identify the independent variables that are statistically significant.**

**For example, in the regression model in (Q7) the standard error for tuit18 is 0.0209396. However, when we add tuit17 to the same regression model, we can see that the standard error for tuit18 nearly doubles to 0.0508123.**

**To avoid multicollinearity, the model should probably only contain one of the variables, tuti17 is too similar to the variable tuit18, which essentially measures the same thing (tuition for men).**

```
.
. /*================================================================================
> ================================================================================*/
.
. cap log close _all
```