ECN 480 Assign 5, Winter 2022

Saturday, March 26, 2022 1:35 PM

Carson Konzer

ECN 480/PUB 580
Assignment #5
Due: Thursday, March 31, 2022 by end of day

Directions: Answer each question electronically in a MS Word or .pdf file. Compile your answers into a single computer file, and then upload it to Canvas under "Assignment #5." Contact me if you have any questions.

1. Download the data set entitled "MEAP93.dta" from Canvas that is posted along with assignment #5. This contains data on the percentage of students in a school district being proficient on the math portion of the MEAP test along with some additional variables regarding the school district. The MEAP test is a test given to all high school students to see if they are proficient in various subjects. Suppose you suspect that math proficiency is related to poverty in the district and how much is spent on the schools in the district. The variables of interest are:

- math10: the percent of students passing the math section of the MEAP test.
- lnchprg: the percentage of students receiving free or reduced school lunch, as lower income students receive free or reduced lunch.
- expend: spending per student in the school district
- staff: number of staff per 1,000 students — <lnstaff?>
- salary: average teacher salary in the school district — co-linear?
- benefits: average teacher benefits in the school district — co-linear?
- droprate: the dropout rate in the school district
- gradrate: the graduation rate in the school district

Suppose you think that math10 = f(lnchprg, expend, salary, benefits, staff droprate, gradrate)

a. Estimate the multiple regression in Stata for this. Are there signs of multicollinearity? Why or why not? Refer to Lecture #15 for the signs regarding multicollinearity. **(3 points)**

```
. reg math10 lnchprg expend salary benefits staff droprate gradrate

      Source |       SS           df       MS      Number of obs   =       408
-------------+----------------------------------   F(7, 400)       =     14.21
       Model |  8924.93347         7  1274.9905   Prob > F        =    0.0000
    Residual |   35892.247       400  89.7306175   R-squared       =    0.1991
-------------+----------------------------------   Adj R-squared   =    0.1851
       Total |  44817.1805       407  110.115923   Root MSE        =    9.4726

------------------------------------------------------------------------------
      math10 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     lnchprg |  -.2598572   .0399512    -6.50   0.000    -.3383978   -.1813165
      expend |   .0005327   .0030446     0.17   0.861    -.0054528    .0065182
      salary |   .0001953    .000373     0.52   0.601    -.0005379    .0009285
    benefits |  -.0001613   .0004863    -0.33   0.740    -.0011173    .0007947
       staff |   .0361442   .1374901     0.26   0.793    -.2341493    .3064377
    droprate |  -.0046839    .123658    -0.04   0.970    -.2477846    .2384169
    gradrate |   .1017121   .0518261     1.96   0.050    -.0001734    .2035976
       _cons |   11.03807   15.07189     0.73   0.464    -18.59194    40.66808
------------------------------------------------------------------------------
```

Yes, there are signs of Multicolinearity.

The T-tests on Expend, Salary, benefits, staff, & droprate fail to reject Ho.

→ OLS has a hard time when trying to distinguish what variables are the ones effecting math10.

b. Calculate the variance inflation factors to see if multicollinearity is present. You can do this by typing `estat vif` after you estimate the regression. Which variables have a problematic variance inflation factor? Refer to Lecture #10 on how to interpret the variance inflation factor. **(3 points)**

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      expend |     25.31    0.039517
      salary |     16.02    0.062435
       staff |     15.17    0.065937
    benefits |      2.27    0.439592
    gradrate |      2.18    0.459296
    droprate |      2.09    0.479223
     lnchprg |      1.34    0.745701
-------------+----------------------
```

General Guideline: If $VIF_i > 5$ then we should suspect multicollinearity with such variable i & the other variables in the regression.

→ Expend, Salary, & Staff have a problematic variance inflation factor

```
    Mean VIF |      9.20
```

c. Suppose you think a fix for the multicollinearity is to drop the staff and salary variables and then re-estimate the model. Does this fix the multicollinearity? **(2 points)**

NO, while we no longer have any VIF's > 5, we can see that the OLS still has benefits & droprate failing to reject H_0, thus the model still has a hard time distinguishing what X's are important for Y.

```
. reg math10 lnchprg expend benefits droprate gradrate

      Source |       SS           df       MS      Number of obs   =       408
-------------+----------------------------------   F(5, 402)       =     19.84
       Model |  8871.21273         5  1774.24255   Prob > F        =    0.0000
    Residual |  35945.9678       402  89.4178302   R-squared       =    0.1979
-------------+----------------------------------   Adj R-squared   =    0.1880
       Total |  44817.1805       407  110.115923   Root MSE        =    9.4561

------------------------------------------------------------------------------
      math10 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    lnchprg |  -.2709573   .0370034    -7.32   0.000    -.3437017    -.198213
      expend |   .0016091    .000694     2.32   0.021     .0002447    .0029735
    benefits |  -.0001172   .0003654    -0.32   0.749    -.0008356    .0006012
    droprate |  -.0103834   .1230509    -0.08   0.933    -.2522869    .2315201
    gradrate |   .1001315   .0516475     1.94   0.053    -.0014014    .2016645
       _cons |    16.3272     5.8982     2.77   0.006     4.732029    27.92237
------------------------------------------------------------------------------

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
    gradrate |      2.17    0.460865
    droprate |      2.07    0.482276
      expend |      1.32    0.757871
    benefits |      1.29    0.775743
    lnchprg |      1.15    0.866213
-------------+----------------------
    Mean VIF |      1.60
```

Note: both R^2 & $\bar{R}^2$ have decreased, thus this model is worse at predicting math10 than the previous model.

d. Which variables are statistically significant for explaining the percentage of students who are proficient in math? Use a one or two-tailed test as you deem appropriate. **(3 points)**

From the above 2 questions I conclude that lnchprg, expend, & gradrate are statistically significant in explaining math10.

I ran regressions with the combinations including lnchprg, one of {salary, benefits, expend}, & one of {gradrate, droprate}. The regression with (lnchprg, expend, gradrate) explained the largest variance in math10. Additionally, with this regression, all of the p-values are below 0.05 & we reject H_0.

```
. reg math10 lnchprg expend gradrate

      Source |       SS           df       MS      Number of obs   =       408
-------------+----------------------------------   F(3, 404)       =     33.19
       Model |  8861.52812         3  2953.84271   Prob > F        =    0.0000
    Residual |  35955.6524       404  88.9991395   R-squared       =    0.1977
-------------+----------------------------------   Adj R-squared   =    0.1918
       Total |  44817.1805       407  110.115923   Root MSE        =    9.4339

------------------------------------------------------------------------------
      math10 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    lnchprg |  -.2705821   .0368813    -7.34   0.000    -.3430854   -.1980788
      expend |   .0015047    .000614     2.45   0.015     .0002976    .0027118
    gradrate |   .1032658   .0368944     2.80   0.005     .0307368    .1757947
       _cons |   15.70213   4.591338     3.42   0.001     6.676232    24.72802
------------------------------------------------------------------------------
```

This is also the combination from the above listed with the lowest p-values.

e. Suppose you have a school district that has 25% of students receiving free or reduced lunch, spends $5,550 per student, pays teachers a benefit package worth $6,500 dollars, has a 4% dropout rate and a 90% graduation rate. What percentage of students will be proficient in math? **(2 points)**

Using The Model Above;

$$M_{MATH10} = 15.70213 - 0.2705821(25) + 0.0015047(5550) + 0.1032658(90)$$

$$= \boxed{26.58\ \%}$$

= 26.58 %

Using The Model From Q: 1 (c) :

MATH10 = 16.3272 - 0.2709573(25) + 0.0016091(5,550) - 0.0001172(6,500)
- 0.0103834(4) + 0.1001315(90)

= 26.69 %

* possibly co-linearity of expend & benefits) & (gradrate & droprate) are causing the slight inflation in prediction.

---

2. Load in the data set entitled "smoke.dta" which is posted along with assignment #5. This is a data set containing the number of cigarettes a smoker smokes per day, the price of a pack of cigarettes in a state, along with some other variables. Suppose you think the number of cigarettes a smoker smokes a function of the smoker's education, age, square of age, the income of the smoker, and whether restaurants in the state restrict smoking. The variables of interest are:

- cigs: the number of cigarettes smoked per day
- educ: the education level of the smoker
- age: the age of the smoker
- agesq: the square of the age of the smoker
- lincome: the natural logarithm of the smoker's income
- restaurant: dummy variable = 1 if restaurants restrict smoking in that state.

You think that: cigs = f(educ, age, agesq, lincome, restaurant)

a. Estimate this regression in Stata. If you get an additional year of education, how many more (or less) cigarettes do you smoke per day as a result? **(3 points).**

```
. reg cigs educ age agesq lincome restaurant

      Source |       SS           df       MS      Number of obs   =       310
-------------+----------------------------------   F(5, 304)       =      6.93
       Model |  5535.81217         5  1107.16243   Prob > F        =    0.0000
    Residual |  48593.7362       304  159.847817   R-squared       =    0.1023
-------------+----------------------------------   Adj R-squared   =    0.0875
       Total |  54129.5484       309  175.176532   Root MSE        =    12.643

------------------------------------------------------------------------------
        cigs | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .5642898   .2928919     1.93   0.055    -.0120624    1.140642
         age |   1.089952   .2791078     3.91   0.000      .540724    1.63918
       agesq |  -.0113235   .0032092    -3.53   0.000    -.0176385   -.0050085
     lincome |   1.972289   1.113237     1.77   0.077    -.2183354    4.162914
  restaurant |   -2.72433    1.83095    -1.49   0.138     -6.32727    .8786107
       _cons |   -25.5103   10.80247    -2.36   0.019    -46.76739   -4.253219
------------------------------------------------------------------------------
```

Our regression predicts that an additional year of education will increase the number of cigarettes smoked per day by 0.5642898

---

b. Which $\hat{\beta}$ are statistically different from zero? Compare the test-statistic for each $\hat{\beta}$ to the two-tailed critical value for a 5% level of significance. **(3 points)**

For our regression, only age & agesq are statistically significant from Ø by a two-tailed t-test at a 5% level of significance.

* @ 10% significance, educ & lincome would also become statistically significant.

---

c. Test for heteroscedasticity using the Breusch-Pagan test using the `hettest, rhs iid` test we described in Lecture #17. Is there evidence of heteroscedasticity at a 5% level of significance? At a 10% level of significance? How do you know? **(3 points).**

```
. hettest, rhs iid

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
```

We can test for evidence of heteroskedasticity

```
. hettest, rhs iid

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

    chi2(5) =    9.28
Prob > chi2 = 0.0985
```

We can test for evidence of heteroskedasticity via the p-value given w.r.t χ^2 distribution.

As p-value = 0.0985,

There is evidence of heteroskedasticity at a 5% level of significance,

| is not | | | | | 10% | | | .

---

d. Estimate the same regression as in part a., but have Stata calculate the heteroscedasticity-robust standard errors for you by adding `, robust` like we talked about in Lecture #16. Does anything change in terms of statistical significance? Compare these results to your results in part b. **(3 points)**

* We now reject H_0 for lincome & thus have one more statistically significant variable (:

```
. reg cigs educ age agesq lincome restaurant, robust

Linear regression                               Number of obs     =        310
                                                F(5, 304)         =       6.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1023
                                                Root MSE          =     12.643

------------------------------------------------------------------------------
             |               Robust
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5642898   .3309779↑    1.70   0.089↑   -.0870079    1.215588
         age |   1.089952   .2669926↓    4.08   0.000=    .5645643     1.61534
       agesq |  -.0113235   .0030769↓   -3.68   0.000=   -.0173783    -.0052687
     lincome |   1.972289   .9969255↓    1.98   0.049↓    .0105412    3.934038
  restaurant |   -2.72433   1.713223↓   -1.59   0.113↓   -6.095606    .6469467
       _cons |   -25.5103   9.930115↓   -2.57   0.011↓   -45.05077   -5.969843
------------------------------------------------------------------------------
```

For all $\hat{\beta}_i$ except that of educ, we see decreased standard error.

We additionally see that education is less significant now.

---

e. Suppose you think the heteroscedasticity is a function of the smoker's age, since there is such a large variance in the age of smokers in the data set. Estimate a weighted least squares regression by adding `[aweight = age]` to the end of the regression command like we talked about in Lecture #18. Do NOT include `, robust` Does anything change in terms of statistical significance? Compare your results to those in d. and in a. **(3 points)**

Compared to (a), we now reject H_0 for educ & thus have one more statistically significant variable.

```
. reg cigs educ age agesq lincome restaurant [aweight = age]
(sum of wgt is 12,125)

      Source |       SS           df       MS      Number of obs   =       310
-------------+----------------------------------   F(5, 304)       =      5.93
       Model |  4862.18104         5  972.436207   Prob > F        =    0.0000
    Residual |  49823.0163       304  163.891501   R-squared       =    0.0889↓↓
-------------+----------------------------------   Adj R-squared   =    0.0739
       Total |  54685.1974       309  176.974749   Root MSE        =    12.802

------------------------------------------------------------------------------
        cigs | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .7060197   .2923382↓↓    2.42   0.016↓↓   .1307572    1.281282
         age |   .9770046    .282955↑↑    3.45   0.001↑↑   .4202063    1.533803
       agesq |  -.0100339   .0030571↓↓   -3.28   0.001↑↑  -.0160498   -.0040181
     lincome |   1.520557   1.169527↑↑    1.30   0.195↑↑  -.7808361    3.821951
  restaurant |  -2.804753   1.868843↑↑   -1.50   0.134↓↑  -6.48226    .8727533
       _cons |  -20.61711   11.22402↑↑   -1.84   0.067↑↑  -42.70372    1.469505
------------------------------------------------------------------------------
```

Compared to (d), we now reject H_0 for educ, but we now also accept H_0 for lincome, thus we are trading the lincome variable for educ, by our assumption

In general, the assumption that heteroskedasticity is a function of age, seems to be a false one as we see different significant variables & decreased R^2.

---

f. We have estimated three different regressions to explain the number of cigarettes a smoker smokes per day: a regular regression, a regression with robust standard errors, and weighted least squares. Are there some common results across the three that suggest what determines the number of cigarettes a smoker smokes per day? **(2 points)**

smokes per day: a regular regression, a regression with robust standard errors, and weighted least squares. Are there some common results across the three that suggest what determines the number of cigarettes a smoker smokes per day? **(2 points)**

Across all 3 regressions, We can conclude that both age & age squared are statistically significant in determing th # of cigs smoked per day.

If we are to expand our level of significance to that of 10%, we see that educ is statistically significant across all regressions.

Last, whether or not restaurants let users smoke inside, within a given state is not statistically significant across all regressions.