ECN 480 Assign 2, Winter 2022

Tuesday, February 8, 2022 6:12 PM

ECN 480/PUB 580
Assignment #2
Due: Thursday, February 17, 2022 by end of day

Directions: Answer each question either on your own paper or electronically in a MS Word file. Compile your answers into a single computer file, and then upload it to Canvas under "Assignment #2." Contact me if you have any questions.

The following is a data set for the poverty rate and the percentage of students who are proficient in math (as measured by a standardized test) for various school districts in Genesee County. I got this data from an issue of *The Flint Journal* when I used to get the newspaper a long-long time ago in a galaxy far-far away:

District	Poverty	Proficient
Atherton	12	28
Beecher	36	9
Bendle	24	30
Bentley	10	34
Carman-Ainsworth	14	39
Clio	8	38
Davison	12	65
Fenton	6	59
Flint	35	14
Flushing	8	58
Genesee	19	30
Goodrich	6	66
Grand Blanc	7	71
Kearsley	12	47
Lake Fenton	4	49
Lakeville	7	37
Linden	8	66
Montrose	12	47
Mt. Morris	20	32
Swartz Creek	10	46
Westwood Heights	18	10

1. Open a blank data set in Stata using the `edit` command and enter in the above data set. Give your variables the same names as in the table. Use Stata commands to find the average poverty rate, the average proficiency rate, and the minimum and maximum values of the poverty and proficiency rates. **(3 points)**

```
. summarize Poverty Proficient

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     Poverty |         21    13.71429    8.894621         4         36
  Proficient |         21    41.66667    18.40471         9         71
```

---

2. Write down a simple regression model you would like to estimate where the poverty rate explains the proficiency rate. Write this in the form of the equation that follows the third bullet point under "Example 2.3 from the book" in Lecture #5. What sign do you think β_1 will take? Why? **(3 points)**

$$prof = \beta_0 + \beta_1 pov + u$$

β_1 should be negative; I expect poverty & proficiency to be inversely correlated...

---

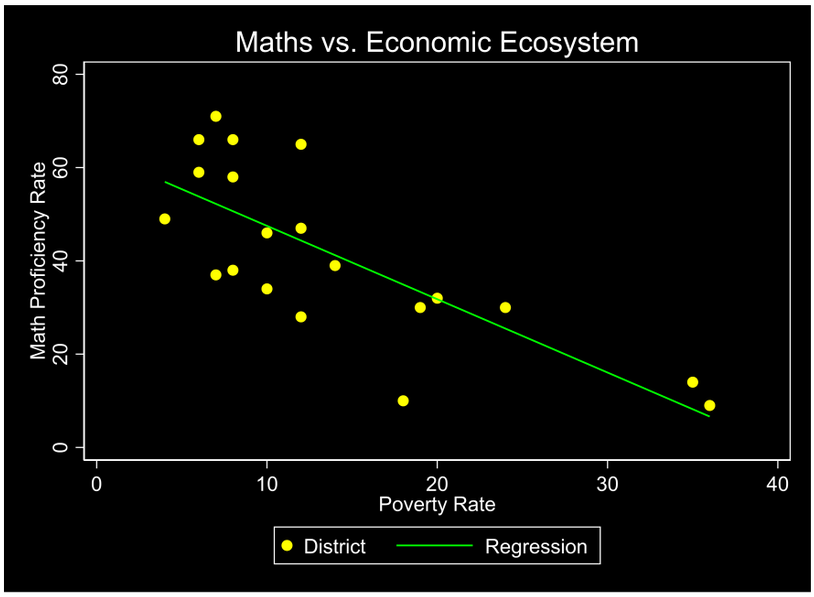
3. In Lecture #5, we said that simple regression is fitting a best-fit linear line through a scatterplot of observations for X and Y. In your simple regression, Y = proficiency in math while X = poverty rate. Stata can plot a best-fit linear line through a scatterplot with the following commend:

```
graph twoway scatter y-variable x-variable || lfit y-variable x-variable
```

The first part of the comment is the familiar scatter plot. The second part of the command tells Stata to plot a fitted line using linear regression (get it, *lfit*!!). The | | simply combines the two commands into one command. In fact, if you just typed `graph twoway lfit y-variable x-variable` you would get a graph with only the fitted regression line and not the scatter plot (but, we want both the regression line **and** the scatter plot for this question!). The `y-variable` is the name of the variable you want on the y-axis (e.g. Y), and `x-variable` is the name of the variable you want on the x-axis (e.g. X). Note that the | | is entered by holding down "shift" and then pressing the key directly above the "enter" key" twice.

Label the axes of each graph appropriately, and give it an appropriate title. Copy and paste it into Microsoft Word. Then, in a sentence or two, comment on the relationship between the two variables based on the graph. Do we have a strong positive correlation, weak positive correlation, strong negative correlation, weak negative correlation? What do you think, and why? **(3 points)**

```
. graph twoway scatter Proficient Poverty || lfit Proficient Poverty
```



We seem to show a strong negative correlation. Impoverished areas may have less access to quality tutoring and education.

---

4. Estimate the simple regression model you wrote down in question #2 using the Stata command for a regression. Recall the command is:

```
reg y-variable x-variable
```

Where, like in question #3, `y-variable` is the name you are giving Y and `x-variable` is the name you are giving X. Copy-and-paste your results so that they look like they do in Stata (e.g. use Courier New, font size 8) **(3 points).**

```
. reg Proficient Poverty

      Source |       SS           df       MS      Number of obs   =        21
-------------+----------------------------------   F(1, 19)        =     25.97
       Model |  3912.15312         1  3912.15312   Prob > F        =    0.0001
    Residual |  2862.51354        19  150.658608   R-squared       =    0.5775
-------------+----------------------------------   Adj R-squared   =    0.5552
       Total |  6774.66667        20  338.733333   Root MSE        =    12.274

------------------------------------------------------------------------------
  Proficient | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     Poverty |  -1.572409   .3085706    -5.10   0.000    -2.218254   -.9265632
       _cons |   63.23113    5.00825    12.63   0.000     52.74874    73.71352
------------------------------------------------------------------------------
```

---

5. Interpret the $\hat{\beta}_1$ you obtained in question #4. How much does y-change if x-changes by 1 unit? Specify the units and remember the difference between percentage change and change in percentage points. We have a change in percentage points for X here. Refer to Lecture #5 to recall the difference between the two **(3 points)**.

A 1 percentage point increase in poverty predicts a 1.57 + + decrease + math proficiency.

---

6. Write down the null hypothesis that poverty does not affect math proficiency vs the alternative hypothesis that poverty lowers math proficiency. Use $\hat{\beta}_1$ to show the null and alternate hypotheses. Do we have a one-tailed or two-tailed test? Why? **(3 points)**

$H_0: \hat{\beta}_1 \neq 0 \ (\hat{\beta}_1 \geq 0)$ $H_A: \hat{\beta}_1 < 0$

We have a 1-tailed test as we only care about if poverty rate decreases maths proficiency if not + + + increases + + .

7. Conduct a t-test for your null vs. alternate hypothesis in question #6. You can use the t-test Stata gave you for $\hat{\beta}_1$. What is the critical value? Recall that you need to use the critical value from the t-table (Table G.2, p. 786) since you have a small sample size. Do you reject or fail to reject the null hypothesis? Why? **(3 points)**

Critical value @95% Confidence
One Tail
1.729

2 tailed t = -5.10
1 tailed t = -2.6

$|-2.6| > 1.729$

$\Downarrow$

We reject the null hypothesis as the test statistic is significant @ 95% confidence.

---

8. What is the R^2 for this regression? What does it tell you? **(2 points)**

0.5775. 57.75% of the variation in Proficiency is explained by Poverty.

Save your data set and then clear Stata using the `clear` command. If you don't save it, then you will have to re-enter it again if you want to return to these questions in the future.

For the remaining questions download the data set called "ATTEND.DTA" and load it into Stata. It is right below the .pdf file for Assignment #2 in the "Assignments" section of Canvas.

This is a data set that lists the number of classes attended in a current semester consisting of 32 classes (e.g. a 16 week semester) for various students as well as some other variables. We are interested in the variable "atndrte," which is the percentage of classes each student in the data set attended that semester. There are several additional variables as well. We are only concerned with two other ones for this assignment: priGPA: the student's GPA prior to the current semester in the data set and ACT: the student's ACT score.

We think that a student's attendance rate is determined by his/her GPA going into the semester as well as that student's ACT score. That is: atndrte = f(priGPA, ACT)

You want to estimate the following regression: $atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$

9. Estimate this in Stata using the `reg` command. Copy and paste you results in Courier New, font-size 8 **(3 points).**

```
. reg atndrte priGPA ACT

      Source |       SS           df       MS      Number of obs   =       680
-------------+----------------------------------   F(2, 677)       =    138.65
       Model |  57336.7612         2  28668.3806   Prob > F        =    0.0000
    Residual |  139980.564       677  206.765974   R-squared       =    0.2906
-------------+----------------------------------   Adj R-squared   =    0.2885
       Total |  197317.325       679   290.59989   Root MSE        =    14.379

------------------------------------------------------------------------------
     atndrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      priGPA |   17.26059   1.083103    15.94   0.000     15.13395    19.38724
         ACT |  -1.716553    .169012   -10.16   0.000    -2.048404   -1.384702
       _cons |    75.7004   3.884108    19.49   0.000     68.07406    83.32675
------------------------------------------------------------------------------
```

---

10. Explain $\hat{\beta}_1$ and $\hat{\beta}_2$. How does the attendance rate change if priGPA increases by 1 unit? If ACT increases by 1 unit? Specify the units for GPA and ACT. Note that we have a change in percentage points for the attendance rate, like in the previous regression. **(3 points).**

A 1 GPA point increase predicts a +17.26 percentage point change in attendance rate.
1 1 ACT 1 1 1 1 -1.72 1 1 1

---

11. Are any of the signs on $\hat{\beta}_1$ and $\hat{\beta}_2$ different than what you expected before running the regression? This is a very common issue in regression. Do you have any idea what might be causing the unexpected sign? There is no right or wrong answer to this. I am just curious what you think. **(2 points)**.

Yes, ACT's sign for $\hat{\beta}_2$. I would say that this is due to the confounding variables relating to a single Test Score. On the other hand, prior GPA is a variable that accounts for a student's performance over multiple years in multiple domains.

Possibly, students with higher ACT scores are better at learning independently, or grasp concepts quicker, and thus do not feel attending class is as important. ?

---

12. Suppose you have two students in class: Billy Madison and Eric Gordon. Billy Madison has a priGPA = 3.1 and ACT = 21. Eric Gordon is the perfect student with a priGPA=4.0 and ACT = 36. How many classes are each predicted to attend? **(3 points).**

$\hat{Y}_{Billy} = 75.7 + 17.26(3.1) - 1.72(21) = 75.7 + 53.51 - 36.12$

$= 93.09\%$ of classes.

$\hat{Y}_{Eric} = 75.7 + 17.26(4.0) - 1.72(36) = 75.7 + 69.04 - 61.92$

$= 82.82\%$ of classes.

---

13. Which $\hat{\beta}$ are statistically significant in the regression (e.g. statistically different from zero with a two-tailed test)? How do you know? **(3 points).**

2 tailed Critical Value @ 95% confidence: 1.96

Both ⇐ $|15.94| > 1.96$ & $|-10.16| > 1.96$

---

14. Rerun the regression, but remove the variable for ACT score. Does the $\bar{R}^2$ increase or decrease? What does that tell you about the relevance of this variable to the regression? Why? **(3 points).**

```
. reg atndrte priGPA

      Source |       SS           df       MS      Number of obs   =       680
-------------+----------------------------------   F(1, 678)       =    151.35
       Model |  36008.3571         1  36008.3571   Prob > F        =    0.0000
    Residual |  161308.968       678  237.918832   R-squared       =    0.1825
-------------+----------------------------------   Adj R-squared   =    0.1813
       Total |  197317.325       679   290.59989   Root MSE        =    15.425

------------------------------------------------------------------------------
     atndrte | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      priGPA |   13.36898   1.086703    12.30   0.000     11.23527    15.50268
       _cons |   47.12702   2.872615    16.41   0.000     41.48673    52.76732
------------------------------------------------------------------------------
```

$\bar{R}^2$ decreases. Thus ACT score does explain some of the variance in Attendance, and is important. Eg. Adding in ACT score increases $\bar{R}^2$ & R^2 as well. => our model is better with ACT information.
