Homework Assignment #3

Due by hardcopy or Canvas as a *single file* at *beginning of class* on Tuesday September 19!

Written problems:

1. Wooldridge: Chapter 2, Problem 6 (Problem 2.6 in wooldridge_all.pdf)
2. Although we focused upon logarithms in class, one could use other non-linear transformations within a regression model. For the wage and education example that has been considered in lecture, suppose that we wanted to take the square root of education and relate wages to that. So the SLR model would be $wage = \beta_0 + \beta_1 educ + u$

a. Figure out the formula for the effect of *educ* on *wage* by taking the derivative *dE(wage|educ)/d educ*. Note that unlike the basic SLR model, this derivative depends upon the *x* variable (here, *educ*).

E(wage/education) = B(0) + B(1)sqrt(edu)

DE/D(edu) = b1/2sqrt(edu)

Wage = -4.464346+ 2.947042

Here is the Stata for the regression of this model (using *wage1.dta*):

b. Estimate the effect of *educ* on *wage* at both *educ=12* and *educ=16* (using your formula from the previous part). How do these effects compare to the estimated effects from the original SLR model? For your reference, the original regression (*wage* on *educ*) results were:

educ=12 -> wage = 5.74450695

educ=16 -> wage = 7.323822

Old model = -0.9048516 + 0.5413593

When edu = 12, wage = 5.59146

When edu = 16, wage = 7.7568972

c. Which regression (*wage* on *educ* or *wage* on *sqrteduc*) appears to give a better overall fit? Should overall fit (as measured by R-squared) be the only reason to choose one model specification over another? Explain.

The old model appears to give a better fit since by looking at the r-squared value since the old model has a higher value.Higher R-squared values generally indicate a better fit as they suggest that a larger proportion of the variance in the dependent variable is explained by the independent variable. However, this metric shouldn't necessarily be used always to choose the better overall fit since we need to look at outliers and other factors like it's practical significance.

Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

fitted log(*price*) = 9.40 + 0.312 log(*dist*) n = 135, R-squared = 0.162

(i) Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?

The coefficient 0.312 represents the estimated change in the log of housing price for a unit change in the log of the distance from the garbage incinerator while holding other factors constant.

The positive sign of the coefficient (0.312) shows that there is a positive relationship between the log of distance from the incinerator and the log of housing price. On average, houses farther away from the incinerator tend to have higher prices, which is what one might expect intuitively.

(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

The estimate provided in this simple regression doesn't provide an unbiased estimator of the ceteris paribus elasticity of price with respect to distance . Mainly because other factors affecting housing prices are not controlled for in this simple regression. In real life housing prices are influenced by numerous factors such as the size of the house, neighborhood quality, school district, number of bedrooms, etc. These uncontrolled factors can be confound variables.

(iii)What other factors about a house affect its price? Might these be correlated with distancefrom the incinerator?

- House size: Larger houses tend to have higher prices.
- Neighborhood characteristics: The quality of the neighborhood, crime rates, proximity to amenities (parks, shopping centers), and school quality can all influence housing prices.
- Number of bedrooms and bathrooms: More bedrooms and bathrooms typically lead to higher prices.
- Age and condition of the house: Newer and well-maintained houses would have higher prices.

Wooldridge Computer Exercise C2.6

Use the data MEAP93.DTA to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*). (Additional questions are given on the assignment sheet.) (CHECK APPENDIX)

```
. regress math10 expend
```

Source	SS	df	MS	Number of obs	=	408
				F(1, 406)	=	13.84
Model	1477.19666	1	1477.19666	Prob > F	=	0.0002
Residual	43339.9838	406	106.748729	R-squared	=	0.0330
				Adj R-squared	=	0.0306
Total	44817.1805	407	110.115923	Root MSE	=	10.332

math10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
expend	.0024557	.0006601	3.72	0.000	.001158	.0037534
_cons	13.35923	2.934111	4.55	0.000	7.591287	19.12718

(i) Do you think each dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate?

A diminishing effect seems to be more appropriate in this data.

(ii) In the population model *math10* = *beta0* + *beta1 log(expend)* + *u*, argue that (beta1 divided by 10) is the percentage point change in *math10* given a 10% increase in *expend*.

- When expend is 100 -> beta1 = beta1
- When expend is 110 -> 10% increase in log(expend) -> beta(1/10)

(iii)Use the data to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and R-squared.

N = 408

r^2 = 0.0297

math10^ = -69.3411 + 11.16439 ln (expend)

(iv)How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

1.1%

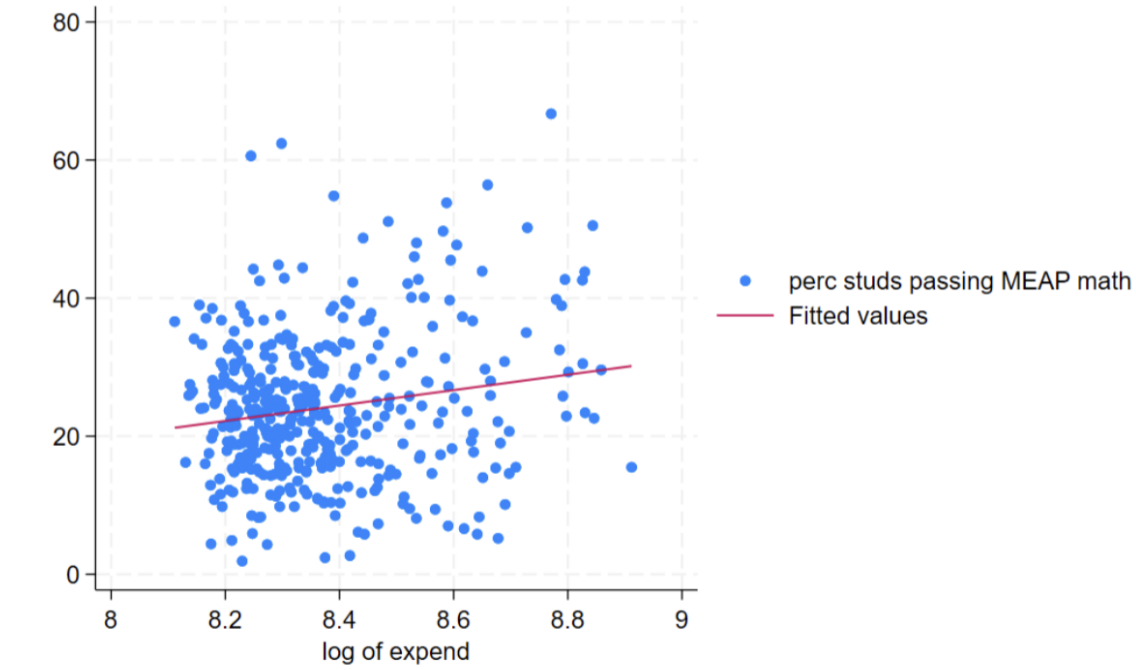
(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is that not much of a worry in this data set?

Because math10 is a maximum of 66.7, so we will never reach 100.

Computer problems (show any relevant Stata output):
Wooldridge: Chapter 2, Computer Exercise C6 (Problem C2.6 in wooldridge_all.pdf)

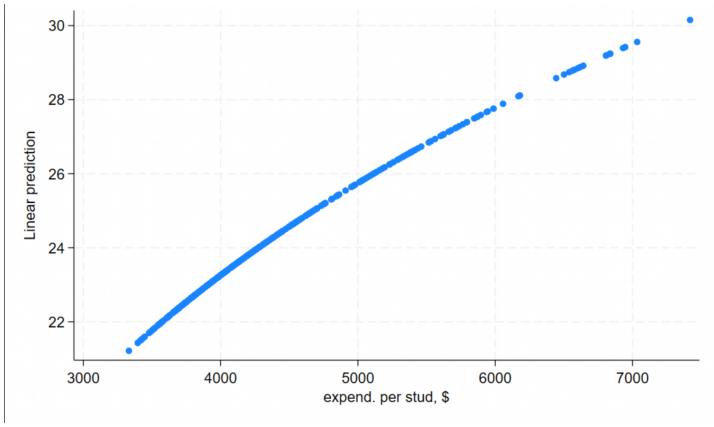
(Note that *log(expend)* is included in the data set as the variable named *lexpend*.) For the same dataset (MEAP93.DTA), also answer the following questions:

(vi) Plot *math10* versus *lexpend* with the fitted regression line shown.



(vii) To visualize the non-linearity described by this model, create the fitted values from the regression and then do a scatter plot (*without* the fitted line) of the fitted values versus *expend* (not *lexpend*). Explain how the effect of expenditures changes at higher values of expenditures.

We can see the diminishing nature of the slope which shows how the marginal value of expenditure starts descreaing as values soar.



(viii) Suppose that we wanted to measure *math10* as a fraction (a number between 0 and 1) rather than on a 0-to-100 scale. Specifically, we could do the following in Stata to re-scale the math10 variable:

. replace math10 = math10 / 100

(This replaces the original *math10* values with the values divided by 100.) If you re-ran the regression (now regressing the re-scaled *math10* upon *lexpend*), how would the following quantities change (compared to the original results)? Be specific, and try these on your own before checking your answers in Stata.

(a) R-squared: R-squared would not change

(b) SST: sst(new) = sst/100^2 = sst/10000

(c) the slope estimate:

New slope is b'

B' = b1/100

(d) the intercept estimate

b(0)' = b0/100

APPENDIX

```
. regress math10 expend
```

Source	SS	df	MS	Number of obs	=	408
				F(1, 406)	=	13.84
Model	1477.19666	1	1477.19666	Prob > F	=	0.0002
Residual	43339.9838	406	106.748729	R-squared	=	0.0330
				Adj R-squared	=	0.0306
Total	44817.1805	407	110.115923	Root MSE	=	10.332

math10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
expend	.0024557	.0006601	3.72	0.000	.001158	.0037534
_cons	13.35923	2.934111	4.55	0.000	7.591287	19.12718

```
. regress math10 ln_expend
```

Source	SS	df	MS	Number of obs	=	408
				F(1, 406)	=	12.41
Model	1329.42517	1	1329.42517	Prob > F	=	0.0005
Residual	43487.7553	406	107.112698	R-squared	=	0.0297
				Adj R-squared	=	0.0273
Total	44817.1805	407	110.115923	Root MSE	=	10.35

math10	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_expend	11.16439	3.169011	3.52	0.000	4.934677	17.39411
_cons	-69.3411	26.53013	-2.61	0.009	-121.4947	-17.18753

```
. sum math10
```

Variable	Obs	Mean	Std. Dev.	Min	Max
math10	408	24.10686	10.49361	1.9	66.7

```
. sum expend
```

Variable	Obs	Mean	Std. Dev.	Min	Max
expend	408	4376.578	775.7897	3332	7419

```
. predict math10_hat, xb

. sum math10_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
math10_hat	408	24.10686	1.807319	21.21697	30.15375