San Francisco State University **Michael Bar**
ECON 312 **Fall 2018**

Problem set 6

Due Thursday, December 6, in class

Names of collaborators (type below):

Name 1 MINJIN PARK

Name 2 NISHANLANG KHONGLAH

Assignment Rules

1. Homework assignments must be typed. For instruction on how to type equations and math objects please see notes "Typing Math in MS Word".
2. Homework assignments must be prepared within this template. Save this file on your computer and type your answers following each question. Do not delete the questions.
3. Your assignments must be stapled.
4. No attachments are allowed. This means that all your work must be done within this word document and attaching graphs, questions or other material is prohibited.
5. Homework assignments must be submitted at the end of the lecture, in class, on the listed dates.
6. Late homework assignments will not be accepted under any circumstances, but the lowest homework score will be dropped.
7. The first homework assignment cannot be dropped.
8. You are encouraged to work on this homework assignment in groups of up to 3 people, and submit one assignment with up to 3 names typed on this page. Sharing the electronic version of your assignment with other teams is absolutely prohibited.
9. All the graphs should be fully labeled, i.e. with a title, labeled axis and labeled curves.
10. In all the questions that involve calculations, you are required to show all your work. That is, you need to write and explain the steps that you made in order to get to the solution.
11. This page must be part of the submitted homework.

Create an R script, which performs all the statistical analysis in this assignment. You can name the script `HW6`.R. **You can either print out the script and attach it as a separate page at the end of this assignment, or copy and paste its content at the end of your assignment. Make sure to add comments explaining which question you are solving, and every command in your script.**

Specification of Regression Models

1. (10 points). What are the likely adverse consequences of omitting relevant variables from regression models?

   Omitting relevant variables from regression models can result in biased and inconsistent estimators of coefficients on the included variables. In addition, omitting relevant variables can also result in biased standard errors of estimators, making all statistical tests invalid.

2. (20 points). List two reasons why researchers may omit relevant variables from the regression model, and propose solutions for the omitted variables problem. Present your answer in the following table:

Reason for omitted variables	Solution
1. Researchers may not be familiar enough with economic theory.	1. Review relevant economic theory.
2. Researchers may be unable to obtain data on the omitted variables.	2. Use proxy variables that are highly correlated with the omitted variables.

3. (5 points). What is adverse consequence of including irrelevant variable in your regression model?

   Including an irrelevant variable in the regression model makes the OLS coefficients inefficient, meaning that the coefficients do not have the smallest possible variance.

4. (5 points). How would you avoid inclusion of irrelevant variable in your regression model?

   To avoid the inclusion of irrelevant variables in the regression model, every variable in the model must be justified by economic or other theory.

5. (20 points). Suppose that Kevin estimated two models, and his fitted equations are:

$$\widehat{EARNINGS} = b_1 + 3S + EXP$$
$$EXP = d_1 - 0.5S$$

   Where S is schooling and EXP is experience. Dray is another researcher who estimated the following model:

$$\widehat{EARNINGS} = \tilde{b}_1 + \tilde{b}_2 S$$

   a. What would be the value of Dray's estimated coefficient on schooling, $\tilde{b}_2$?

      The value of Dray's estimated coefficient on schooling will be less than 3.

b. Explain intuitively why Dray's estimated coefficient on schooling is of the magnitude you answered in the previous section?

Dray's estimated coefficient on schooling is less than 3 because $\tilde{b}_2$ underestimates the net impact of schooling on earnings. That is, Dray's model does not capture the impact of experience on earnings, which is reflected in the underestimated value of $\tilde{b}_2$.

6. (20 points). For this question use the Educational Attainment and Earnings data: http://online.sfsu.edu/mbar/ECON312_files/wage21.csv.
   a. Estimate the following models,

$$[\text{model1}]: \log(EARNINGS) = \beta_1 + \beta_2 S + u$$
$$[\text{model2}]: \log(EARNINGS) = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

   Present both models in a summary table using `stargazer` package.

Earnings model regressed on Schooling and Earnings model regressed on Schooling and Experience

	Dependent variable:	
	log(EARNINGS)	
	(1)	(2)
S	0.1121***	0.1274***
	(0.0094)	(0.0092)
EXP		0.0381***
		(0.0051)
Constant	1.2657***	0.4122**
	(0.1307)	(0.1694)
Observations	540	540
R^2	0.2086	0.2824
Adjusted R^2	0.2071	0.2798
Residual Std. Error	0.5238 (df = 538)	0.4992 (df = 537)
F Statistic	141.8033*** (df = 1; 538)	105.6870*** (df = 2; 537)
Note:		*p<0.1; **p<0.05; ***p<0.01

b. Interpret the estimated slope coefficient on schooling in model1 (denote it $\tilde{b}_2$).

The estimated slope coefficient on schooling $\tilde{b}_2$ in model1 is 0.11. This means that every additional year of schooling adds 11 cents to hourly wages.

c. Interpret the estimates slope coefficient on schooling in model2 (denote is b_2).

The estimated slope coefficient on schooling b_2 in model2 is 0.13. This means that every additional year of schooling adds 13 cents to hourly wages, holding experience constant.

d. Explain why model1 estimated smaller impact of schooling on earnings that model2.

model1 includes only schooling as a regressor and fails to capture the effect of experience on earnings. By omitting a relevant variable, experience, the model underestimates the net impact of schooling on earnings.

Heteroscedasticity

7. (5 points). What are the consequences of Heteroscedasticity in regression model?

The consequences of heteroscedasticity in a regression model include:

i) OLS estimators are inefficient, meaning that they no longer have the lowest variance among all unbiased linear estimators.

ii) The estimated standard errors of the regression coefficients, s.e.(b), are biased, making the t-test and F-test invalid.

8. (35 points). For this question use the Educational Attainment and Earnings data: http://online.sfsu.edu/mbar/ECON312_files/wage21.csv.

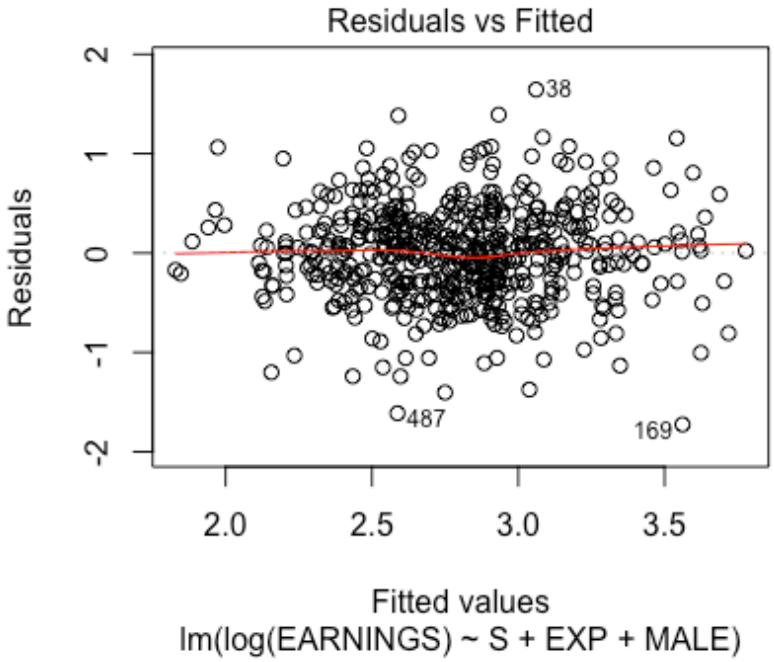
a. Regress earnings on schooling, experience and male dummy, and report the R output here.

Earnings model including dummy (MALE) variable

	Dependent variable:
	log(EARNINGS)
S	0.1215^{***}
	(0.0088)
EXP	0.0294^{***}
	(0.0050)
M	0.3054^{***}
	(0.0423)
Constant	0.4857^{***}
	(0.1622)
Observations	540
R^2	0.3460

Adjusted R^2	0.3424
Residual Std. Error	0.4770 (df = 536)
F Statistic	94.5403*** (df = 3; 536)
Note:	*p<0.1; **p<0.05; ***p<0.01

b. Perform model diagnostics by graphing the residuals against the fitted values. The R command: `plot(model1, which = 1)`. Present the graph here.



c. Does your graph from the last section support the presence of heteroscedasticity in your model? Explain briefly.

The graph does not confirm the presence of heteroscedasticity as the actual values are spread evenly around the fitted line, which is expected when the error terms are homoscedastic.

d. Perform the White test for heteroscedasticity. Describe the estimated model, report the regression output, and write the conclusion of the test.

The estimated model estimates the variance of the error term and the regression output is presented below.

White test summary

	Dependent variable:
	log(EARNINGS)
Y_hetero_pred	-0.2933
	(0.5367)
Y_hetero_pred2	0.0688
	(0.0948)
Constant	0.4999
	(0.7544)
Observations	540
R^2	0.0088
Adjusted R^2	0.0051
Residual Std. Error	0.3710 (df = 537)
F Statistic	2.3718* (df = 2; 537)
Note:	*p<0.1; **p<0.05; ***p<0.01

We fail to reject H_0 at significance level α as the reported p value is greater than α, and conclude that the original model does not suffer from heteroscedasticity.

e. Estimate the same model, but with the heteroscedasticity robust standard errors.

Uncorrected vs Robust Standard Errors

	Dependent variable:	
	log(EARNINGS)	
	uncorrected	robust
	(1)	(2)
Constant	0.4857***	0.4857***
	(0.1622)	(0.1639)
S	0.1215***	0.1215***
	(0.0088)	(0.0096)
EXP	0.0294***	0.0294***
	(0.0050)	(0.0051)
M	0.3054***	0.3054***
	(0.0423)	(0.0428)

Observations	540	540
R^2	0.3460	0.3460
Adjusted R^2	0.3424	0.3424
Residual Std. Error (df = 536)	0.4770	0.4770
F Statistic (df = 3; 536)	94.5403***	94.5403***
Note:		*p<0.1; **p<0.05; ***p<0.01

f. Suppose that you suspect that the variance of the error term is proportional to S^2 (schooling squared). Show how the model can be transformed in such a way that the error terms in the new model are homoscedastic.

The variance of the error term in the model is $var(u_i) = {\sigma_u}^2 {S_i}^2$.
We assume that the variance of the error is proportional to S^2. Now, we can transform the original variables and error term by dividing them *S*.

$$\frac{\ln Earnings}{S} = \beta_1 \frac{1}{S} + \beta_2 \frac{S}{S} + \beta_3 \frac{Exp}{S} + \beta_4 \frac{M}{S} + \frac{u}{S}$$

g. Obtain the Weighted Least Squares estimates of the model in this question, and report the R output here.

Weighted Least Squares Results

	Dependent variable:
	Y
	OLS
Constant	0.6415***
	(0.1526)
S	0.1076***
	(0.0092)
EXP	0.0315***
	(0.0048)
M	0.2974***
	(0.0418)
Observations	540
R^2	0.3089
Adjusted R^2	0.3051
Residual Std. Error	0.0358 (df = 536)

F Statistic	79.8701*** (df = 3; 536)
Note:	*p<0.1; **p<0.05; ***p<0.01