Economics 142
Problem Set #4

1. Suppose we have a sample of size N on two variables, x_i and y_i. Define the correlation coefficient as

$$\rho_{xy} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \times \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{1/2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x_i and y_i. Consider the regression of y_i on a constant and x_i: $y_i = \beta_0 + \beta_1 x_i + u_i$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the OLS estimates of the coefficients β_0 and β_1.

(a) Define the OLS residual for observation i as: $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Show that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. Hint: look at problem set 3 !

(b) Show that $y_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) + \hat{u}_i$.

(c) Using (b) and the FOC for the OLS estimator, show that:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})\hat{u}_i = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2$$

(d) Show that

$$\hat{\beta}_1 \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 - \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2$$

Hint: take $\hat{\beta}_1$ inside the summation, then use (b) and (c).

(e) Recall from lecture that in the case of a regression with a constant and one other x variable:

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Using your results from (d), plus the definition of R^2 for the regression of y_i on a constant and x_i, show that

$$\rho_{xy}^2 = R^2$$

2. Using the data set from problem set number 3, find the correlation coefficient ρ_{we} between log wages and education for females in the sample. Then run the OLS regression of log wages on education, and get the R^2 coefficient, and verify that $\rho_{we}^2 = R^2$. Verify that the R^2 for the regression of education on

log wages is the same (i.e., you can reverse the role of x and y and you get the same R^2 coefficient, when there is only a single x and a constant). Show how you can modify the answer to 1 (e) when x and y are reversed.

3. (a) Using the data set from problem set number 3, compute mean log wages for female non-immigrants (imm=0) and the standard error of the mean. Then compute mean log wages for female immigrants (imm=1) and the standard error of the mean. Construct a test statistic to test if the means are the same.

(b) Another way to test if mean log wages of immigrant and non-immigrant females are the same is to run a regression of log wages on a constant and immigrant status. Check that when you do this, the estimated coefficient on the "immigrant" variable is EXACTLY equal to the difference in means you constructed in part (a). From the regression model you can test that immigrants and natives have the same mean log wages using the estimate of the "immigrant" coefficient, and the estimate of the standard error of this coefficient. Is this the same test statistic that you get in part (a)? HINT: the regression model standard errors assume homoskedasticity. Is that true here?

(c) You can implement "heteroskedasticity-robust" (HC) standard errors for a regression model in stata using the "robust" option with a "reg" statement. To get these standard errors in R you need to get the "sandwich" package:

install.packages("sandwich")

Suppose you fit a linear model and store the results in an object called "ols". You can then use the vcovHC function to compute a heteroskedasticity consistent variance covariance matrix:

ols$robse = vcovHC(ols, type="HC1")

Once you have the HC variance covariance matrix, compute the standard errors. Hint: what is the relationship between the standard errors and the diagonal of the variance covariance matrix? Compare these HC standard errors to the "regular" standard errors you computed in part (b).

Note: When you use the "robust" option with a "reg" statement in Stata, the software package automatically implements a particular way of calculating the $\widehat{V}$ matrix from Lecture 7. However, there are other ways of computing this matrix, and R wants you to specify a choice. For this reason, you need to specify the "type" of HC correction you have in mind. The sandwich packages allows for multiple HC corrections ranging from "HC0" to "HC4". Stata applies the "HC1" correction by default.

You can find out more about the corrections that R allows at https://cran.r-project.org/web/packages/sandwich/vignettes/sandwich.pdf.