ECO 5720 **Name:**
Exam #1

There are 4 questions. Please try to attempt everything. All the best!

1. The effect of class size on pupil achievement is a question of policy relevance, and thereby analyzed by several studies. One such analysis is interested in exploring how class size affects test scores for fifth graders. It relies on a data set containing information across 1000 schools and estimates the following simple regression:

$$\widehat{score} = 75.32 - 0.031size.$$

Here, for any school:
score - average reading comprehension score (among fifth graders) varying between 0 and 100, and
size - average size of a class (for fifth graders).

a. Based on the simple regression above, what is the effect of increasing a school's average class size by 100 on the average reading score? Please explain the sign and magnitude of the impact.

Answer: As average class size increases by 100, the average reading score decreases by 3.1.

b. Given that students are not randomly allocated to classes of different size, please provide an example of an unobserved characteristic that could be correlated with both the dependent and independent variables.

Answer: Factors such as teacher quality may be correlated with both *score* and *size*.

For the next two parts, suppose the residual for the above regression is denoted by $\hat{u}$ such that for each observation, $\hat{u} = score - \widehat{score}$.

c. What is the average value of $\hat{u}$?

Answer: Zero.

d. What is the value of the correlation between $\hat{u}$ and the explanatory variable, *size*?

Answer: Zero.

2. A number of studies explore the determinants of crime rate across countries. Suppose one such analysis relies on data across countries to estimate:

$$\log(crime) = \beta_0 + \beta_{\log(GNP)}\log(GNP) + \beta_{educ}educ + \beta_{urban}urban + \beta_{\log(police)}\log(police) + u.$$

The estimated equation is

$$\widehat{\log(crime)} = \underset{(3.263)}{-7.67} + \underset{(0.319)}{1.061}\log(GNP) - \underset{(0.036)}{0.022}educ + \underset{(0.021)}{0.007}urban + \underset{(0.195)}{0.348}\log(police).$$

Here, the standard errors are in parentheses; the sample size $n = 65$ and $R^2 = 0.72$. Moreover,
crime: number of thefts as a percentage of population
GNP: per capita gross national product

educ: share of children enrolled in primary school
urban: percentage of population in urban areas, and
police: percentage of policemen in population

a. Suppose we have a two-tailed test with H_0: $\beta_{\log(GNP)} = 0$. For the 5% level of significance, what are the *t* distribution's critical values?

Answer: The *t* distribution's degrees of freedom is 65-4-1, or 60. The corresponding critical values are -2 and 2.

b. Given the estimated value of $\beta_{\log(GNP)}$, can we reject H_0 at the 5% level of significance? Explain by calculating the corresponding test statistic and comparing it against the *t* distribution's critical values.

Answer: The value of the test statistic is given by 1.061/0.319, i.e., 3.33. Accordingly, we reject H_0.

c. Given the information above, how much of the variation in log(*crime*) is explained by log(*GNP*), *educ*, *urban*, and log(*police*).

Answer: Given the R^2 value, it is 72%.

d. Suppose we are jointly testing whether the slope coefficients corresponding to log(*GNP*), *educ*, *urban*, and log(*police*) are zero, i.e., H_0: $\beta_{\log(GNP)} = 0$, $\beta_{educ} = 0$, $\beta_{urban} = 0$, and $\beta_{\log(police)} = 0$. Write the R-squared form of this *F* statistic and calculate its numerical value.

Answer: In this case, the unrestricted R^2 is the R^2 from the above regression. The restricted R^2 is zero. So, the *F* statistic is given by $[R^2/q] / [(1-R^2)/(n-k-1)]$, i.e., $[0.72/4] / [(1-0.72)/(65-4-1)] = 38.57$.

e. For H_0 in part (d) above and the 1% level of significance, what is the F distribution's critical value?

Answer: From Table G.3c, the critical value is about 3.65.

f. For H_0 in part (d) above, explain whether you reject H_0 at the 1% level of significance by comparing the F statistic value against the corresponding critical value.

Answer: We reject H_0.

g. State the null hypothesis that the elasticity of *crime* with respect to *GNP* is 1.

Answer: H_0: $\beta_{\log(GNP)} = 1$.

h. State the null hypothesis that the elasticity of *crime* with respect to *GNP* is the same as the elasticity of *crime* with respect to *police*.

Answer: H_0: $\beta_{\log(GNP)} = \beta_{\log(police)}$.

3. Consider a data set on home prices of homes. The variables are
price: house price, $1000s
bdrms: number of bdrms
sqrft: size of house in square feet
lotsize: size of lot in square feet.

Suppose the multiple regression of log(*price*) on *bdrms*, log(*sqrft*), and log(*lotsize*) is given by:

$$\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 \log(sqrft) + \beta_3 \log(lotsize) + \mathrm{u}.$$

a. What does the assumption of homoskedasticity imply in the above model?

Answer: Constant variance of u (the error term).

b. Do we need this assumption to obtain unbiased estimates of the coefficients?

Answer: No.

c. Do we need u (the error term) to be normally distributed to obtain unbiased estimates of the coefficients?

Answer: No.

Now, the standard error for $\hat{\beta}_1$ is given by

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SST_{bdrms}(1 - R^2_{bdrms})}}$$

Here, $SST_{bdrms} = \sum_{i=1}^{n} \left(bdrms_i - \overline{bdrms}\right)^2$, the total variation in *bdrms*
$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-3-1}}$, the standard error of the regression.

d. In this formula for standard error, what does R^2_{bdrms} indicate?

Answer: R-squared from the regression of *bdrms* on log(*sqrft*) and log(*lotsize*).

e. In terms of the above standard error formula in part (c), a high value of which term would correspond to the issue of multicollinearity among the explanatory variables?

Answer: R^2_{bdrms}.

f. Do we need the absence of multicollinearity to obtain unbiased estimates of the coefficients?
[Hint: Note that multicollinearity is not the same as perfect collinearity.]

Answer: No.

4. Consider a data set on NBA salaries around 1995. The variables are
wage: annual salary, $1000s
points: points per game
rebounds: rebounds per game.

The multiple regression of log(*wage*) on *points* and *rebounds* is estimated as:

$$\log\widehat{(wage)} = 5.909 + 0.075 points + 0.062 rebounds.$$

The simple regression of *rebounds* on *points* is estimated as:

$$\widetilde{rebounds} = 1.580 + 0.276 points.$$

a. From the multiple regression above, what is the (approximate) percentage increase in salary for one more point per game?

Answer: One more point per game increases salary by 100(0.075)% or 7.5%.

b. What is the numerical value of the slope estimate when log(*wage*) is regressed only on *points*?
[Hint: Think of the formula depicting how the coefficient estimate of a regressor changes upon omitting another regressor.]

Answer: Note that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$. Here, $\hat{\beta}_1 = 0.075$, $\hat{\beta}_2 = 0.062$, and $\tilde{\delta}_1 = 0.276$. So, $\tilde{\beta}_1 = 0.075 + (0.062 \times 0.276) = 0.092$.

c. Suppose that the multiple regression of log(*wage*) on *points* and *rebounds* satisfies the assumptions required for unbiasedness. However, the simple regression of log(*wage*) on *points* does not. In fact, the coefficient estimate of *points* likely suffers from an omitted variable bias in case of the simple regression. If *rebounds* has a positive effect on log(*wage*), *points* and *rebounds* are positively correlated, what is the likely sign of the omitted variable bias?

Answer: Positive or upward bias.