ECO 5720 **Name:**
Exam #1

There are 6 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:

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

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 four parts, suppose the residual for the above regression is denoted by $\hat{u}$.

c. Define $\hat{u}$ in terms of $score$ and $\hat{score}$.

Answer: We have $\hat{u} = score - \hat{score}$.

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

Answer: Zero.

e. What is the value of the correlation between $\hat{u}$ and *size*?

Answer: Zero.

f. What is the value of the correlation between $\hat{u}$ and $\hat{score}$?

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:

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

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. Suppose we use cross section data on about 400 observations across schooling districts and obtain the following estimated regression model:

$$\hat{tscore} = 607.300 + 3.850income - 0.042income^2$$

where,
tscore: average (fifth grade) test score in a district
income: average annual per capita income in a district in thousands of 1998 dollars
$income^2$: *income* squared.

a. After what value of *income* does additional district income begin to lower predicted *tscore*. Calculate the numerical value.

Answer: The turning point is given by |3.850/2(-0.042)|, i.e., about 45.83 in thousands of dollars.

b. State the formula for obtaining the effect of an additional unit of *income* (i.e., in thousands of dollars) on *tscore*. Calculate its numerical value for *income* = 20.

Answer: The effect is given by 3.850 – 2 x 0.042 x *income*. For *income* = 20, this is 2.17 points.

4. Consider a model allowing the height of a shrub to depend upon the amount of bacteria in the soil and the extent of sunlight received:

$$height = \beta_0 + \beta_1 bacteria + \beta_2 sun + \beta_3 bacteria \cdot sun + u$$

where,
$height$: height of a shrub in cm.
$bacteria$: measure of bacteria in soil (1000 per ml.)
sun: hours of daily sunlight.

a. State the formula for obtaining the effect of one additional unit of *sun* on *height*.

Answer: $\beta_2 + \beta_3 bacteria.$

For the above model, suppose we use data across about 700 observations and estimate the following regression:

$$\hat{height} = 35.7 + 4.2bacteria + 8.5sun + 1.8bacteria \cdot sun.$$

b. Calculate the effect of *sun* on *height* for *bacteria* = 10.

Answer: $8.5 + 1.8(10) = 26.5$.

5. Suppose we are interested in estimating a regression model corresponding to manufacturing firms. Our variables of interest are:
productivity: measure of firm-level productivity
exports: value of firm-level exports
age: a firm's age in years
skill: share of skilled labor in a firm:

The multiple regression of *productivity* on the explanatory variables discussed above is given by:

$$productivity = \beta_0 + \beta_1 exports + \beta_2 age + \beta_3 skill + u.$$

a. Suppose *u* includes factors such as managerial quality which renders *exports* endogenous. Would the ordinary least squares (OLS) estimator of β_1 be unbiased?

Answer: No.

b. Suppose *age* and *skill* are correlated with *exports* but not directly with *u*. Would the OLS estimators of β_2, and β_3 be unbiased?

Answer: No.

6. Consider a country-level cross-sectional data set on the Summer Olympics. The variables are
medals: total number of medals won
pop: population
pcy: GDP per capita
temp: average annual temperature.

Suppose the multiple regression of *medals* on *pop*, *pcy*, and *temp* is given by:

$$medals = \beta_0 + \beta_1 pop + \beta_2 pcy + \beta_3 temp + u.$$

a. What factors may *u* contain?

Answer: Factors related to a country's geography, age, and institutional support.

b. Using some examples from part (a), what does the assumption of homoskedasticity imply in the above model?

Answer: Variance of institutional support does not depend on population or per capita income.

c. Do we need homoskedasticity to obtain unbiased estimates of the coefficients?

Answer: No.

d. Do we need *u* to be normally distributed to obtain unbiased estimates of the coefficients?

Answer: No.

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

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

Here, $SST_{pop} = \sum_{i=1}^{n} (pop_i - \overline{pop})^2$, the total variation in *pop*

$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-3-1}}$, the standard error of the regression.

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

Answer: R-squared from the regression of *pop* on *pcy* and *temp*.

f. In terms of the above standard error formula in part (d), a high value of which term would correspond to the issue of multicollinearity?

Answer: R^2_{pop}.

g. Do we need the absence of multicollinearity to obtain unbiased estimates of the coefficients?

Answer: No.