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:

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

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 is 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:

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

The estimated equation is

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

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: $\beta_{\log(GNP)} = 0$. 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. The *t* distribution's degrees of freedom is 65-4-1, or 60. The corresponding critical value is 2. Accordingly, we reject H_0.

b. 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%.

c. 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., $[0.72/4] / [(1-0.72)/(65-4-1)] = 38.57$.

d. For H_0 in part (c) 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: From Table G.3c, the critical value is about 3.65. Hence, we reject H_0.

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

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

f. 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 1700 observations and obtain the following estimated regression model:

$$\widehat{bwght} = \underset{(77.561)}{3060.49} + \underset{(10.452)}{40.207}npvis - \underset{(0.341)}{0.850}npvis^2$$

where the standard errors are in parentheses, and
bwght: birth weight in grams
npvis: number of prenatal visits
npvis^2: *npvis* squared.

a. After what value of *npvis* does additional prenatal visits actually begin to lower predicted *bwght*. In other words, what is the value of *npvis* corresponding to the turning point of the quadratic relationship between *bwght* and *npvis*? Calculate the numerical value.

Answer: The turning point is given by |40.207/2(-0.850)|, i.e., about 23.65 visits.

b. State the formula for obtaining the effect of an additional unit of *npvis* on *bwght*. Calculate its numerical value for *npvis* = 10.

Answer: The effect is given by 40.207 – 2 x 0.85 x *npvis* grams. For *npvis*=10, this is 23.207 grams.

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: $\beta_2 + \beta_3 bacteria.$

b. State the formula for calculating the effect of an additional unit of *bacteria* on *height*.

Answer: $\beta_1 + \beta_3 sun.$

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

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

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

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

d. Calculate the effect of *bacteria* on *height* for *sun* = 2.

Answer: $4.2 + 1.8(2) = 7.8.$

5. Suppose we are interested in estimating a regression model corresponding to manufacturing firms. Our variables of interest are:
productivity: a measure of firm-level productivity
exports: value of a firm's exports
mqual: a measure of firm-level managerial quality.

If we have data on all three characteristics, the multiple regression of *productivity* on *exports* and *mqual* (given below) satisfies the assumptions required for unbiasedness:

$$productivity = \beta_0 + \beta_1 exports + \beta_2 mqual + u.$$

In other words, the ordinary least squares (OLS) estimator of β_1 from such a regression (i.e., $\overset{\wedge}{\beta_1}$) is unbiased.

However, suppose we do not have data on *mqual* and end up estimating a simple regression of *productivity* on *exports* as noted below:

$$productivity = \beta_0 + \beta_1 exports + u.$$

This simple regression does not satisfy the assumptions required for unbiasedness. In other words, the OLS estimator of β_1 from such a regression (i.e., $\widetilde{\beta_1}$) suffers from an omitted variable bias.

a. What is the likely sign (i.e., positive, negative, or zero) of the effect of *mqual* on *productivity*?

Answer: Positive.

b. What is the likely sign (i.e., positive, negative, or zero) of the correlation between *mqual* and *exports*?

Answer: Positive.

c. What is the likely sign (i.e., positive, negative, or zero) of $E(\widetilde{\beta_1}) - E(\overset{\wedge}{\beta_1})$?

Answer: Positive.

6. 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?

Answer: No.