ECO 5720 **Name:**
Exam #1

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

1. Using data on fifth graders across schools, we estimate the effect of class size on test scores using a simple regression model:

$$score = \beta_0 + \beta_1 size + u.$$

For any school:
score - average reading comprehension score varying between 0 and 100,
size - average size of a class, and
u – unobserved characteristics affecting test scores.

The estimated equation is:

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

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 class size is not randomly determined across schools, please provide an example of an unobserved characteristic that could be correlated with both the dependent and independent variables.

Answer: A factors such as teacher quality may be correlated with both *score* and *size*. For example, urban districts may attract better quality teachers due to higher supplemental pay; these schools may also have larger class sizes than rural schools on average.

c. If the sample average value of *size* or $\overline{size}$ is 20, what is the sample average value of *score* or $\overline{score}$?

Answer: Note that the regression line passes through the sample average. So, the average value of score is $75.32 - 0.031 \times 20$, or 74.7.

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

d. What is $\hat{u}$ in terms of $score$ and $\hat{score}$?

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

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

Answer: Zero.

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

Answer: Zero.

g. 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:

$$\log(crime) = \beta_0 + \beta_1 \log(GNP) + \beta_2 educ + \beta_3 urban + \beta_4 \log(police) + u.$$

The estimated equation is

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

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: $\beta_1 = 0$. 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 critical values are -2 and 2.

b. Given the estimated value of β_1, 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: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$. 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: $\beta_1 = 1$.

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: $\beta_1 = \beta_4$.

3. Consider a data set on health indicators and other characteristics across school-age children. The variables are
health: indicator of health status
insec: self-reported measure of household food insecurity
income: household income
educ: parents' education in years.

Our multiple regression of interest is

$$health = \beta_0 + \beta_1 insec + \beta_2 income + \beta_3 educ + u.$$

Here, *u* represents unobserved factors affecting child health. It includes characteristics such as number of siblings, primary caregiver's health, and proximity to a medical facility.

a. Suppose *insec* and *income* are endogenous but *educ* is not correlated with *u*. However, *educ* is likely correlated with *insec* and *income*. Will ordinary least squares (OLS) yield biased estimates for none, some, or all of the β coefficients, i.e., β_1, β_2, and β_3?

Answer: Possibly for all of the β coefficients.

b. Suppose we disregard the information on endogeneity in (a). However, *insec* is measured with error for households with more children. Should we worry about bias in our OLS estimate of β_1?

Answer: Yes, due to measurement error.

c. If *u* does not follow a normal distribution, can we still assume our test statistics to follow distributions such as *t* or *z*?

Answer: Yes, provided we have a large sample.

4. 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 factors such as 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. Suppose we estimate a regression model from where we save the residuals $\hat{v}$; let us refer to this model as Model A. Next, we regress *medals* on $\hat{v}$ and obtain a coefficient estimate that is equal to $\hat{\beta}_1$, i.e., the coefficient estimate corresponding to *pop* when *medals* is regressed on *pop*, *pcy*, and *temp*. Can you identify the dependent and independent variables in Model A?

Answer: In Model A, the dependent variable is *pop* and the independent variables are *pcy* and *temp*.