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:

$$score = 75.32 - 0.031size.$$

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.

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.

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

c. What is the average value of u?

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

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

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

The estimated equation is

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

Here, the standard errors are in parentheses; the sample size n = 65 and $R^2 = 0.72$. Moreover,

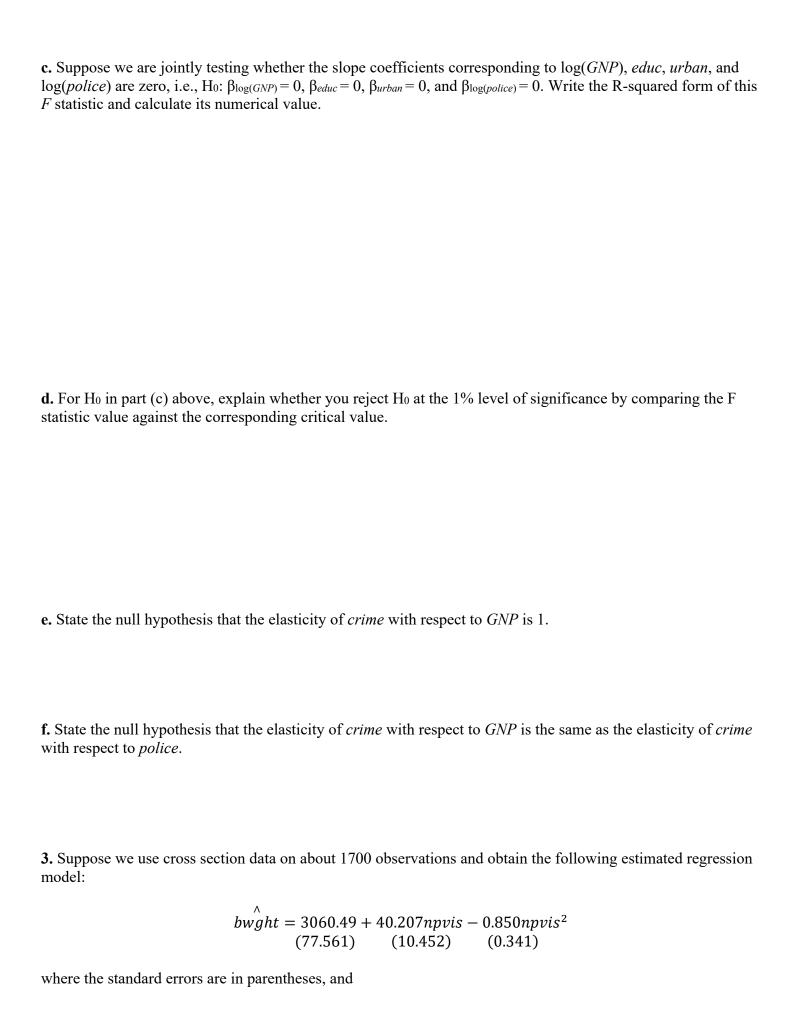
crime: number of thefts as a percentage of population

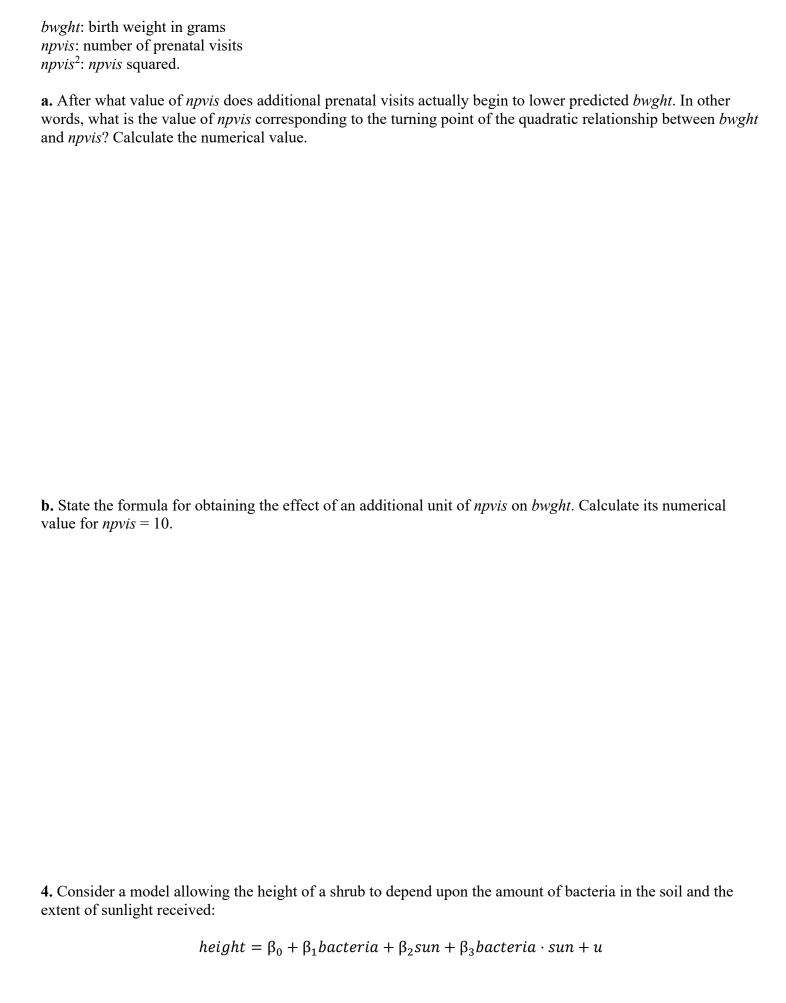
GNP: per capita gross national product

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 H₀: $\beta_{\log(GNP)} = 0$. Given the estimated value of $\beta_{\log(GNP)}$, can we reject H₀ at the 5% level of significance? Explain by calculating the corresponding test statistic and comparing it against the *t* distribution's critical values.

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





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

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

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

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

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

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

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?
- **b.** What is the likely sign (i.e., positive, negative, or zero) of the correlation between *mqual* and *exports*?
- **c.** What is the likely sign (i.e., positive, negative, or zero) of $E(\widetilde{\beta_1}) E(\widetilde{\beta_1})$?
- 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:

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

- a. What does the assumption of homoskedasticity imply in the above model?
- **b.** Do we need this assumption to obtain unbiased estimates of the coefficients?

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

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

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

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_{bdrms}^2 indicate?

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?

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