Problem set 2

Due on Thursday September 22, 2022

Note: No credit will be given if you report only the final answers without showing formulas and calculations when appropriate. This applies to both theoretical and empirical questions. For the empirical questions, make sure to submit your R script and output files on Latte. No credit will be given if the R codes and output are missing.

Problem 1

In "An Economic Theory of Suicide," published in the Journal of Political Economy (1974), Hamermesh and Soss state "When unemployment rises individuals' expectations of future incomes (and utilities) are revised downward. Holding real income of the employed constant, an increased number of people will believe future prospects to have diminished and will commit suicide." As a crude test of their hypothesis, we can estimate the equation

$$suicide_i = \beta_0 + \beta_1 unemployment_i + u_i$$

where $suicide_i$ is the number of suicides per 100,000 population in year i, $unemployment_i$ is the unemployment rate in year i, and u_i is the error term. Data from 1968-1978 are given below:

year	suicide rate per 100,000	unemployment rate
1968	10.7	3.6
1969	11.1	3.5
1970	11.5	4.9
1971	11.7	5.9
1972	12.0	5.6
1973	12.0	4.9
1974	12.1	5.6
1975	12.7	8.5
1976	12.5	7.7
1977	13.3	7.0
1978	12.5	6.0

a. Compute the Ordinary Least Squares estimate of the parameters β_0 and β_1 (you can use an excel spreadsheet and use the formulas for intercept and slope that we discussed in class for the case of a univariate regression model)

b. How would these coefficients change if suicide rates were reported per 1000 population?

c. Calculate the predicted suicide rates for each year. Calculate the estimated error for each year.

d. What is the average estimated error? Why?

e. The unemployment rate was 5.8 in 1979. Based on your estimates, what suicide rate would you have expected that year?

Problem 2

At a recent county fair, you observed that at one stand people's weight was forecasted, and were surprised by the accuracy (within a range). Thinking about how the person at the booth could have predicted your weight fairly so accurately, you want to figure out how this could have been accomplished.

You remember from going to the doctor for your annual physical when you were younger that medical charts for children contain 5%, 25%, 50%, 75% and 95% lines for a weight/height relationship so you decide to conduct an experiment with 110 of your peers. You collect the data and calculate the following sums:

$$\sum_{i=1}^{n} Y_i = 17{,}375, \quad \sum_{i=1}^{n} X_i = 7{,}665.5,$$

$$\sum_{i=1}^{n} y_i^2 = 94{,}228.8, \quad \sum_{i=1}^{n} x_i^2 = 1{,}248.9, \quad \sum_{i=1}^{n} x_i y_i = 7{,}625.9$$

where X denotes height (measured in inches) and Y is the weight (in pounds). Small letters refer to deviations from means as in $z_i = Z_i - \overline{Z}$.

(a) Calculate the slope and intercept of the regression and interpret these.

(b) Find the regression R^2 and explain its meaning.

(c) Calculate the homoskedasticity-only standard errors and, using the resulting *t*-statistic, perform a test on the null hypothesis that there is no relationship between height and weight in the population of college students.

(d) What is the alternative hypothesis in the above test, and what level of significance did you choose?

(e) What other factors can you think of that might have an influence on the weight of an individual?

Problem 3

In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student the sum of hours in the four activities must be 168 (that's 24 hours over 7 days).

a. In the model

$$GPA_i = \beta_0 + \beta_1 Study_i + \beta_2 Sleep_i + \beta_3 Work_i + \beta_4 Leisure_i + u_i$$

does it make sense to hold Sleep, Work, and Leisure fixed, while changing Study?

b. Explain why this model violates OLS assumption #4 (regarding perfect multicollinearity).

c. How could you reformulate this model so that its parameters have a useful interpretation and it satisfies OLS assumption #4?

Problem 4

Suppose we want to estimate the effects of alcohol consumption (alcohol) on college grade point average (colGPA). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called attend). A standardized test score (say, SAT) and high school GPA (hsGPA) are also available.

a. Should we include attend along with alcohol as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$).

b. Should SAT and hsGPA be included as explanatory variables? Explain.

Problem 5 (empirical question)

Does a student's GPA rise or fall as the student moves from freshman year to senior year? In this question you will explore this issue – and other factors that are correlated with GPA – using data on college students collected by researchers at the Harvard School of Public Health. The researchers surveyed 9,890 undergraduate students at 119 four-year colleges in 2001.

a. Open the data set college.dta.[1] Use tidyverse to explore the data and answer the following questions: (i) What is the average GPA in the sample? (ii) How many men are in the sample?

[1] There are a few options available to import a STATA .dta file into R. A simple one is to rely on the `read_stata` command, which will import the.dta file into R as a data frame. See here for additional details: https://www.rdocumentation.org/packages/sjmisc/versions/1.0.0/topics/read_stata

How many women? (iii) What share of students are members of a fraternity or sorority? (iv) What share of students work during the school year? (v) What share of students report that they used marijuana in the past 30 days?

b. How many students did not report their GPA? Is this non-reporting likely to be random? Why or why not?

c. Estimate the regression of GPA on male and work. Interpret the regression coefficients (including the intercept).

d. Estimate the regression of GPA on freshman, sophomore, junior and senior for men only. Are all regression coefficients reported in the results? Explain what happened. What is the solution?

e. Estimate the regression of GPA on sophomore, junior and senior for men only. What is the interpretation of all coefficients in this regression (including the intercept)?

f. For the regression in (e), test the null hypothesis that there is no difference in the GPA of sophomores and juniors. What is the number of restrictions q for this test?

g. For the regression in (e), test the null hypothesis that the coefficients on sophomore, junior and senior are all zero, against the alternative that at least one coefficient is nonzero. State clearly the significance level of the test you are using and the critical value of the F statistic for this test.

h. Estimate the regression of GPA on age, sophomore, junior and senior for men only. What happens to the statistical significance of the coefficients in this regression? Explain why this is the case.

i. Estimate the same regression as in part (e), but now do this for all respondents (male and female). Include controls for male, work, marijuana, lightdrinker, moddrinker and heavy drinker. Calculate the predicted GPA for a male senior who works, is a moderate drinker and has not smoked marijuana in the past 30 days.

j. Using the regression in part (i), test the hypothesis (at the 5% significance level) that freshmen and sophomores have the same GPA on average, holding constant gender, work, type of drinking and marijuana use.

k. What is the adjusted R^2 for the regression in part (i)? What does this adjusted R^2 tell you about the fit of this regression? Does it indicate that omitted variable bias is likely to be a problem?