

Problem Set 2Submission by Friday 6th October 2023, 8 amGeneral guideline for all problem sets:

Make sure to carefully explain your answers and interpret the results. Always state the null hypothesis and the alternative hypothesis you are testing, and the significance level you are considering.

Exercises:

1) You want to analyze the optimal use of your time investment as a graduate student. You distribute a survey among a random sample of recently graduated students, and ask their time allocation during graduate school. You have a question indicating average grades, the number of hours per week for study, the number of hours per week for sleep, and the number of hours per week for all other activities. You estimate the regression:

$$\log(\text{grades}) = b_0 + b_1 \text{study} + b_2 \text{sleep} + u$$

- Explain why it does not make sense to include the number of hours per week for other activities as an extra variable in the regression.
- You obtain an estimate for β_2 that is approximately 0.01. Give a careful interpretation.
- Indicate at least one variable that is likely included in u and that could cause a violation of the zero conditional mean assumption. Explain in which direction you expect the estimates of β_1 and β_2 to be biased.

2) For your job as a policy advisor, you need to come up with the best possible estimate of a parameter β by searching the literature. You find one study with an **unbiased** estimate of $\beta = 5.0$, with estimated variance of β of 8.0, from a regression with an R^2 of 0.86. Another study uses another dataset and finds an **unbiased** estimate of $\beta = 6.0$ with estimate variance of β of 4.0, from a regression with a R^2 of 0.43.

- On which of the two estimates will you rely? Why?
- Do you think you could come up with a better estimate if you had access to the two datasets? Explain.

3) Imagine you want to analyze whether people who were affected by hurricane Maria in Puerto Rico are more likely to favor the Democratic party rather than the Republican political party. You have data on 1000 households and know whether their property was on the path of the hurricane. Meteorologists have told you that the path of hurricane Maria was unpredictable, and also not related to the path of other hurricanes. You also know the amount of \$ people spent after the hurricane to fix their damaged houses, and whether they get any help from the government for doing so.

- would it be a good idea to include a variable for the \$ people spent to reconstruct their house?
- would it be a good idea to include a variable indicating the \$ amount received from the government as help?

Computer exercise

The dataset `nichh.dta` contains data from 582 households in Nicaragua. It is a subset from the 1998 Living Standard Measurement Survey (LSMS).¹ We want to analyze this data to understand the returns to different household assets (number of male and female adults, education and age of the household head, and total land owned) and the role of isolation (measured as the distance to school, and the distance to a health center). We will use total household consumption, expressed in Cordobas, as a measure of household welfare.²

- (1) Write a paragraph describing the dataset using the standard descriptive statistics (also called summary statistics, or “D-stats”). Add a table with the d-stats.
- (2) Specify a model that allows for increasing returns to household assets. Estimate the model and give a careful interpretation of each of the coefficients (sign, size and significance!).
- (3) What is the minimum significance level at which one can reject that hypothesis that age of the household head does not affect total household consumption?
- (4) One could hypothesize that household consumption would start declining after a household head has reached a certain threshold age. Can you find evidence in support of this hypothesis? If so, at which age does it start declining?
- (5) Estimate the model you specified in part (2) separately for households with household heads with at least primary school completed (i.e. 6 years of schooling or more) versus all others. Can you use these estimations to conclude that the returns to female labor are higher for households in which the household head is more educated?
- (6) Consider the alternatives of using either household consumption or $\log(\text{household consumption})$ as dependent variables. For which of the 2 alternatives is the normality assumption more likely to hold. Why?
- (7) Does this dataset allow you to understand the impact of distance to school and health services on households’ welfare? Why or why not?

¹ To be specific, this is a sample of households from the central urban region of Nicaragua.

² (Gold) Cordobas is the name of Nicaraguan’s currency; 1998 exchange rate: +/- 10 cordobas for one 1 US \$.