Notes 4: Notation and Regression

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Data set notation, and other concepts

We will begin with a small review of some of the concepts you learned in the pre-requisite classes. You may do well to brush up on a few chapters of your pre-requisite book. In Freedman's book, I recommend chapters 4, 8 and 9 as needed. They are a good (and entertaining) review. They may be interesting even to those that know the pre-requisite material well, because this book's perspective is unusual, and quite useful for the way of thinking that I want you to develop.

- Example. Consider an observational data set to study the effect of smoking on birth weight.
  - i denotes one woman in the study.
  - For each woman, we observe their baby's birth weight y_i
  - For each woman, we observe the amount they smoked x_{1i}
  - For each woman, we observe a set of possible controls: age x_{2i}, education x_{3i}, and marital status x_{4i}.
  - Observation i is thus composed of a string of information on woman i, denoted $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$.
- Observational data set notation:
  - Observations are denoted by _____
  - For each observation, the outcome variable is denoted _____
  - For each observation, the treatment variable is usually denoted _____
  - For each observation, the controls are denoted _____

_____

- Observation i is thus composed of a string of numbers denoted _____

  _____

- Denote the total number of observations by _____

A data set is normally a table:

- Average outcome

- Empirical variance of outcome

- Empirical standard deviation of outcome

- Empirical covariance between treatment variable and outcome

- Correlation between treatment variable and outcome

  Some facts about the correlation:

  - It is between -1 and 1 (unit free)
  - It is a measure of the linear dependence between two variables.
  - It is a bad measure of other forms of dependence.

Notice: Correlation is not causation.

- Class attendance and grades may be correlated, it doesn't mean that class attendance causes any change in the grades. (Remember, the association could just be because of the confounders)
- Pregnancy smoking is correlated with birth weight. It doesn't mean that smoking in pregnancy causes any changes in the baby's birth weight. (Again, the association could just be because of the confounders)

2 Beginning Regression

Regression is a method to describe data. Later on we will see how this method can be useful to help us uncover causal effects. For now, we will think of regression as simply the next step in describing data (after mean, variance, correlation, etc.) Regression is a LINEAR description of data. We begin by thinking of two variables. Consider a scatter plot.

2.1 Scatter plot

It is a plot of the outcome variable as a function of the variable of interest (sometimes called "treatment variable," "causal variable" or "explanatory variable.")

2.2 Regression line

The regression line is the line that best predicts the average value of y for each value of x.

Examples:

- What is the average final grade for each number of attended classes.
- What is the average birth weight for each amount smoked

2.3 The Graph of Averages

It's a plot of the outcome averages for each level of the treatment variable.

The regression line is a smoothed version of the graph of averages.

ATTENTION: Never eyeball a regression line in the scatter plot. Do it in the graph of averages instead.

2.4 Regression method

It's a formula. A way of calculating the regression line. It turns out that this procedure has some great uses for understanding causal relationships, but we will only talk about them later in the course.

1. Find the point of averages $(\bar{x}_1, \bar{y})$.
2. Go one SD_{x_1} horizontally, and one $r_{y.x_1} \cdot SD_y$ vertically. This is the second point.
3. Trace the line that connects both points.