The Nature of Regression Analysis

Jamie Monogan

University of Georgia

Intermediate Political Methodology
Objectives

By the end of this meeting, participants should be able to:

- Explain the promise and expectations outlined in the syllabus.
- Load data in R & Stata, report descriptive statistics, and visualize the data.
- Define “regression” and other key terms.
- Distinguish between statistical and deterministic relationships, between regression and causation.
- Distinguish between the various types of data researchers use.
- Categorize variables based on their level of measurement.
The Motivation

- Model representation:
  \[ Y_i = \beta_1 + \beta_2 X_i + u_i \]
- Do taller fathers have taller sons? (\( H_0 : \beta_2 = 0, \quad H_A : \beta_2 > 0 \))
- Is there regression to the mean? (\( H_0 : \beta_2 = 1, \quad H_A : \beta_2 < 1 \))
- How tall would we expect Yao Ming’s (7’6” ) son to be?

Source: Gujarati & Porter 2009, 16
When people say, “regression to the mean,” this is a specific phenomenon.

The modern view of regression is much broader. From Gujarati & Porter (2009, 15):
Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

The terms dependent variable and explanatory variable have a variety of synonyms (many of which are posted on page 21).

Our focus this semester: linear regression, or linear models.
- Initially we will study simple, or two-variable regression analysis.
- Then, we will move to multiple regression analysis.
What Regression Can and Cannot Do

- We use regression models to understand *stochastic*, or *random* variables.
- There is always some variability that we cannot explain.
  - Some say unobserved variables create this.
  - Others say that there is an intrinsic variability in an outcome.
- The contrast: deterministic relationships.
  - The law of gravity.
  - Splitting water molecules: moles of hydrogen v. oxygen. (A perfect linear relationship.)
- Whereas *correlation* is a symmetric measure of association, *regression* coefficients are not.
- We create regression models to reflect our view of one variable’s dependence on others. Resist the temptation to think of this as a causal relationship.
- Clichéd but true: Correlation does not imply causation.
- Spring 2014: Ines Levin, Causal Inference
Types of Data

- Cross-section.
- Time series.
- Panel data.
  - Short panels (longitudinal analysis).
  - Long panels (time series cross-section analysis).
- Experimental v. field data.
Recall: Levels of Measurement

- **NOMINAL**: dichotomous, polychotomous, categorical
  - old/young, male/female, religions of the world, numbers on football uniforms, 501(c)3, telephone numbers

- **ORDINAL**: ranks, ordering
  - seniority in Congress, lower/middle/upper class, Likert scales

- **INTERVAL**: ordered with equal spacing
  - Fahrenheit/Celsius, education, age

- **RATIO**: has an meaningful absolute zero point
  - Kelvin, appropriations, vote percentage, crime rates
For Next Time

- Read Gujarati & Porter chapter 2 (Two-Variable Regression Analysis).
- From Gujarati & Porter, page 32, answer question 1.6. What kind of data would your study use?
- Download the file stateImmig0511.tab in the format of your choice from my Dataverse, http://hdl.handle.net/1902.1/16471
- Using either R or Stata, open this file. You also may want the file codebookStateImmig.txt to look-up variable descriptions.
- From these data please report the following:
  - A table for the variable of number of years with a Democratic governor from 2005-2011. (E.g., how many states fall into each category.)
  - The mean, standard deviation, variance, minimum, and maximum of public ideology (measured with the CCES) and the logged scale of immigrant policy from 2005-2011.
  - A scatterplot with immigrant policy on the vertical axis and public ideology on the horizontal axis.
  - With a pencil, draw what you think would be a “best fit” line on the scatterplot you turn in.