What is Regression?

Jeffrey Arnold

April 12, 2016

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two or more variables are related
- Examples:
  - Does turnout vary by types of mailers received?
  - Is the quality of political institutions related to average incomes?
  - Does conflict mediation help reduce civil conflict?

Notation and conventions

- Y - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Log GDP per capita
  - Number of battle deaths
- X - the independent variable or explanatory variable or regressor or right-hand-side variable or treatment or predictor
  - Social pressure mailer versus Civic Duty Mailer
  - Average Expropriation Risk
  - Presence of conflict mediation
- Generally our goal is to understand how Y varies as a function of X:

$$Y = f(X) + \text{error}$$

Conditional expectation review

- How to describe relationship between X and Y?
- **Definition** The **conditional expectation function** (CEF) or the **regression function** of Y given X, denoted $E[Y|X=x]$ is the function that gives the mean of Y at various values of x.
- Note that this is a function of the *population* distributions.
- Regression at its most fundamental is about how the mean of Y changes as a function of X

Conditional with Discrete Variables

Define Y to be the prestige of a profession, $X = ("bc", "wc", "prof")$ (Duncan data):

$$\mu_{bc} = E[Y|X = bc]$$
$$\mu_{wc} = E[Y|X = wc]$$
$$\mu_{prof} = E[Y|X = prof]$$

Each μ for each group is a conditional expectation.

How to calculate Conditional Expectation Function with Discrete Variables

- How do we calculate this? We've already done this: it's just the usual sample mean among the men and then the usual sample mean among the women:

$$\widehat{E}[Y_i|X_i = wc] = \frac{1}{n_{wc}} \sum_{i:X_i=wc} Y_i$$

$$\widehat{E}[Y_i|X_i = bc] = \frac{1}{n_{bc}} \sum_{i:X_i=bc} Y_i$$

$$\widehat{E}[Y_i|X_i = prof] = \frac{1}{n_{prof}} \sum_{i:X_i=prof} Y_i$$

Continuous covariate (I): each unique value gets a mean

- What if X is continuous? Can we calculate a mean for every value of X?
- Not really, because remember the probability that two values will be the same in a continuous variable is 0.
- Thus, we'll end up with a very "jumpy" function, $\widehat{E}[Y_i|X_i = x]$, since n_x will be at most 1 for any value of x.
- You can imagine that this will jump around a lot from sample to sample. The estimates, $\widehat{E}[Y_i|X_i = x]$, will have high sampling variance.
- For some values of x we never observe anything

Continuous covariate (II): stratify and take means

- So, that seems like each value of X won't work, but maybe we can take the continuous variable and turn it into a discrete variable. We call this **stratification**.
- Once it's discrete, we can just calculate the means within each **strata**.

Continuous covariate (III): model relationship as a line

- The stratification approach was fairly crude: it assumed that means were constant within strata, but that seems wrong.
- Can we get a more global model for the regression function? Well, maybe we could assume that it is linear:

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 x$$

- Why might we do this? Parsimony, first and foremost: 2 numbers to predict any value.
- Some other nice properties we'll talk about in the coming weeks.
- Here is the linear regression function for the weight-active minutes relationships:
- We'll see soon how we estimate this line. It's a bit more complicated that the stratify and calculate means.

Parametric vs. nonparametric models

- The conditional mean approach for discrete independent variables are **nonparametric** because they make no assumptions about the functional form of $E[Y_i|X_i = x]$.
- We just estimate the mean among each value of x.
- With continuous independent variables, this approach breaks down because of the number of values.
- Need to make **parametric** assumptions about the functional form of $E[Y_i|X_i = x]$ in order to make progress
- These are parametric because they involve writing the functional form in terms of parameters, like the linear model.

Bias-variance tradeoff

- How we model the regression function, $E[Y_i|X_i = x]$, affects our the behavior of our estimates:
- Low bias (function "nails" every point)
- High variance (drastic changes from sample to sample)

Bias-variance tradeoff

- How we model the regression function, $E[Y_i|X_i = x]$, affects our the behavior of our estimates:
- Higher bias (misses "local" variation)
- Low variance (slope and intercept will only change slightly from sample to sample)

References

These slides derived almost without change from Matthew Blackwell, What is Regression.