Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Assumptions for Linear Regressions

Assumptions

- In the least squares model, beyond some technical assumptions, the following assumptions are sufficient for ordinary least squares (OLS) to work:
  - Data is random, independent draws from the population of X and Y
  - X is uncorrelated with ε (this actually is like assuming the model is "right", at least in the important ways)
- What does "work" mean?
  - It means your estimate of β will be the right one.
  - And hence, your predictions of $\hat{y}$ will also be good on average

Thinking About Bias

- What happens when your model is wrong (aka every time)?
- There's a simple relationship in a simple linear regression.
- Consider the case of schooling:

$$wage = \alpha + \beta\, educ + \varepsilon$$

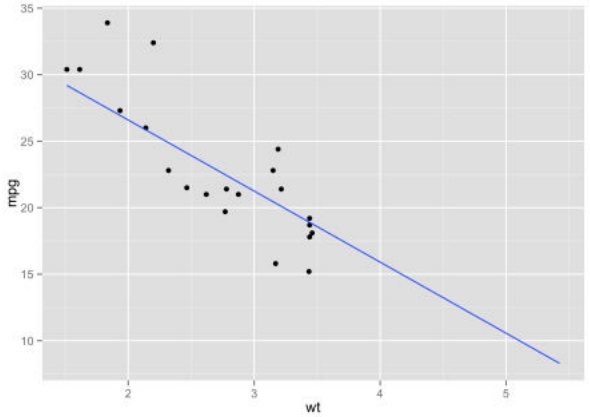
  - But what if smart people go to school more and get more education?
  - Because we don't measure IQ, highly educated people will also have high IQ.
  - Their wages will be extra high.
  - Thus, the line of best fit will have a bigger slope, and it will look like educ is more strongly related to wages than is in fact true.
- Simple models allow you to think through bias more clearly, but for more complicated models, nobody really knows what's going on.

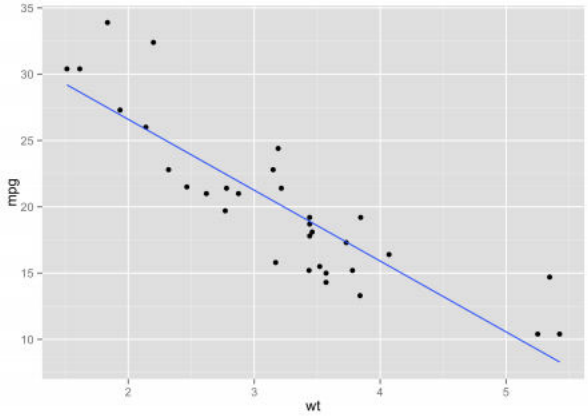
Georgia Tech

Extrapolation

- Extrapolation means using predictions outside the range of data where you fit your model.
- When would this work?
  - If your model is still “right” for new data
  - Regress wages on GPA for GA Tech grads in 2012 and 2013: should be OK
    - Regress wages on GPA for GA Tech grads in 2012 and grads in France in 2013: problems
  - Typically we warn against extrapolation

Extrapolation to new X





- Can be OK if the model is right

Georgia Tech

Extrapolation

But if your model isn't right, you will have trouble

Your model can be wrong for one of two reasons

1. The relationship you found in the data doesn't hold for different values of X
2. Unobserved factors which you can't control for are different in a new sample

Georgia Tech

Assumption: Data Drawn from Population

- Sometimes, this is a hard assumption to believe
- Example: Tweets
  - You want to look at the relationship between Tweet mood and the stock market
  - You create an average daily twitter mood and a stock market return variable
- Financial crisis vignette
  - A big problem was that mortgage backed securities were split up and rated as AAA (very high quality)
  - To rate these securities, agencies looked at the historical performance of comparable securities
  - Intuitively, a problem with this is that you have to assume that these sliced up securities were drawn from the same distribution as whether e.g. states repay loans

Causality and Correlation

- “Correlation does not imply causality”
- β is a measure of the correlation between X and Y
- When X is high, how high might we expect Y to be?
- What does this mean in the context of education/earnings?
  - People who go to school longer tend to make more money
  - But we don’t know whether that’s because they actually are more productive (assuming that productivity is related to pay)
  - Or whether there’s another trait which predicts schooling and income (Ambition? Family background?)

Causality

- How should we think about causality?
  - Imagine assigning a different value of X to an individual
    - Instead of me getting 21 years of education, maybe I got 16 instead
  - What would be the difference in my wages if I drew a different education level?
- Why do we care about this?
  - If we know the true relationship between X and Y , then we can safely extrapolate if we think that relationship should still hold

Lesson Summary

- Bias exists in models and is harder to think through if the model is more complex
- Extrapolation is when predictions are made outside of the range of data the model was fitted in
    - Extrapolation is typically warned against
- “Correlation does not imply causation”

Georgia Tech