Big Data and Security

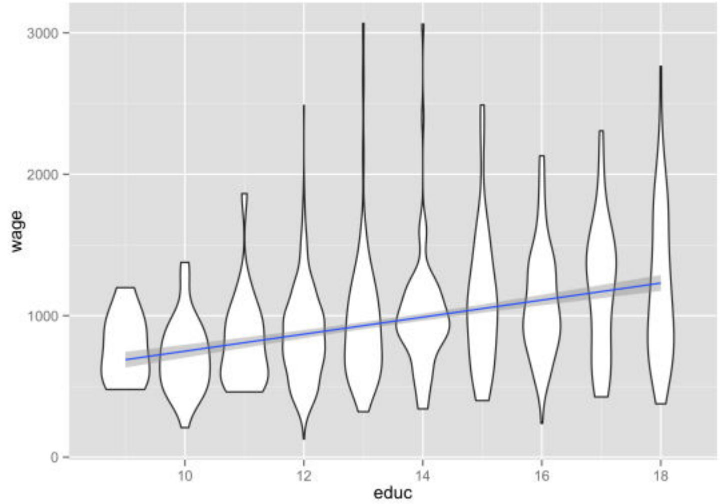
Jeffrey Borowitz, PhD
Lecturer
Sam Nunn School of International Affairs

Generalizing Linear Regression

Our Linear Regression Model

- X is years of schooling, Y is wages

$$wages = \alpha + \beta\, edu + \varepsilon$$



How Did We Find the Line of Best Fit?

- Think about what the residuals would look like if we moved the line around
  - They would grow!
- What we're doing is moving the line around (aka changing α and β) to minimize these squared residuals
- Another names for this linear regression model is "least squares"
- For each data point, you can write the equation:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- So if you pick a particular α and β, you can calculate your εi :

$$\varepsilon_i = y_i - \alpha - \beta x_i$$

Georgia Tech

Least Squares

So you have your expression for each residual ε_i

1. Square each one: ε_i^2
2. Add them all up: $\sum_i \varepsilon_i^2$
3. This is just a function that depends on your data, and your parameters of interest:

$$F(\alpha, \beta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

4. Find α, β to minimize $F(\alpha, \beta)$

These are your estimates

Law of Large Numbers Again

- How does the law of large numbers relate?

$$F(\alpha, \beta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

- Think about α, β as not random - just numbers like 3, or 7.2
- x_i and y_i are summed up, so they are like sample averages
- So you get a kind of combined average function

$$F(\alpha, \beta) \approx \sum_i \gamma_i(\alpha, \beta) = \mathrm{E}[\gamma(\alpha, \beta)]$$

Where $\gamma_i(\alpha, \beta) = (y_i - \alpha - \beta x_i)^2$

- If we have large samples, the α and β we choose are the ones that optimize this function

Georgia Tech

An Aside on Estimation

- We can estimate any function of this form by just taking the sum of the individual data components

$$F(\alpha, \beta) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

- But remember in polling data: the number of points you need to consider is tied to your desired accuracy
- Using math, something like the central limit theorem applies to F(α, β) as well
  - So we can get "accurate" estimates of α and β even with a sample of data points
- Estimating F with just a sample of the data points is called *Stochastic Gradient Descent* and is very widely used in all sorts of big data estimation

Georgia Tech

Least Squares Generalizes Lots of Ways

$$\min_{\alpha,\beta} F(\alpha, \beta) = \sum_i (y_i - \alpha - \beta x_i)^2$$

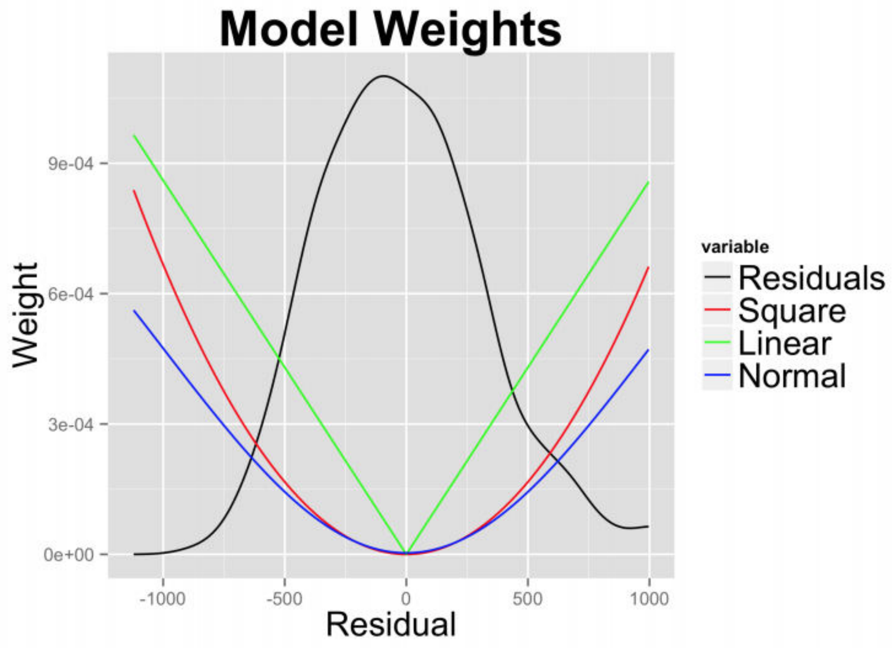
- You can have more variables than X
  - Control for more things: in addition to education, gender, age, experience, etc. might be useful
- You could control for categorical information with “dummy” variables which are 1 or 0 depending whether things are true or false
- Instead of $\alpha + \beta X$ you could have really complicated functions parameters and data

Georgia Tech

Least Squares vs. Maximum Likelihood

- In least squares, the loss function looks like $\sum \varepsilon^2$, so it increases quadratically
- But why quadratic?
  - If we did $\sum |\varepsilon|$, we would penalize big misses less heavily
  - If we did $\sum |\varepsilon^3|$, we would penalize big misses more heavily
  - Answer: squared turns out to be “best” under some assumptions
- Another option is to specify the probability distribution for each ε
  - This is like saying, given schooling, what is the distribution of wages?
  - If you specify a distribution for ε, then you can use this to specify the shape of the distribution
  - This is “maximum likelihood”

Georgia Tech


Model Weights
9e-04
6e-04
3e-04
0e+00
Weight
-1000
-500
0
500
1000
Residual
variable
Residuals
Square
Linear
Normal

Why Least Squares?

- It turns out, least squares has some really nice mathematical properties
  - It’s “unbiased” under a relatively wide variety of conditions
  - Under some relatively specific conditions, it’s the lowest variance estimator there is

Does Linear Make Sense?

- Least Squares seems like it’s restricting us to **linear** functions of X
- But actually this isn’t that much of a limitation
  - We can have two variables: X and Z on the right hand side
  - What if we made $Z = X^2$ - that’s OK
  - And math says that any function of X can be expressed as a sum of polynomials (like X, X^2 , X^3 , etc.)

Georgia Tech

Lesson Summary

- A linear regression model, or "least squares", is used to find the line of best fit
- Least Squares allows one to generalize by controlling for more variables and by controlling for categorical information with “dummy variables”
- Least Squares is “unbiased” under a wide variety of conditions