Big Data and Security

Jeffrey Borowitz, PhD

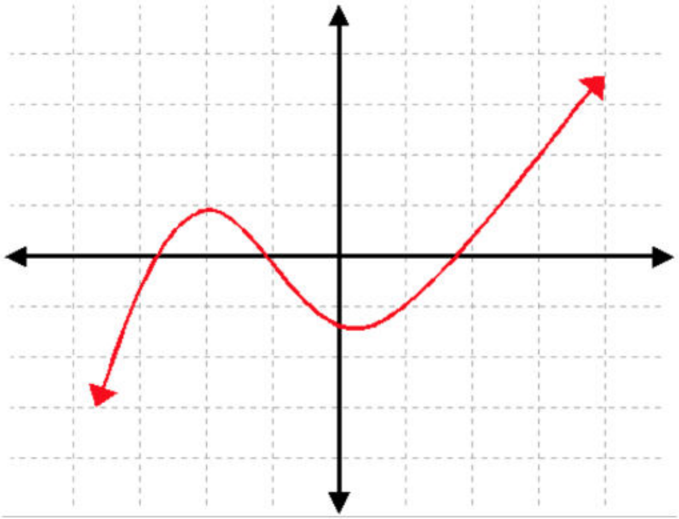
Lecturer

Sam Nunn School of International Affairs

Linear Regression

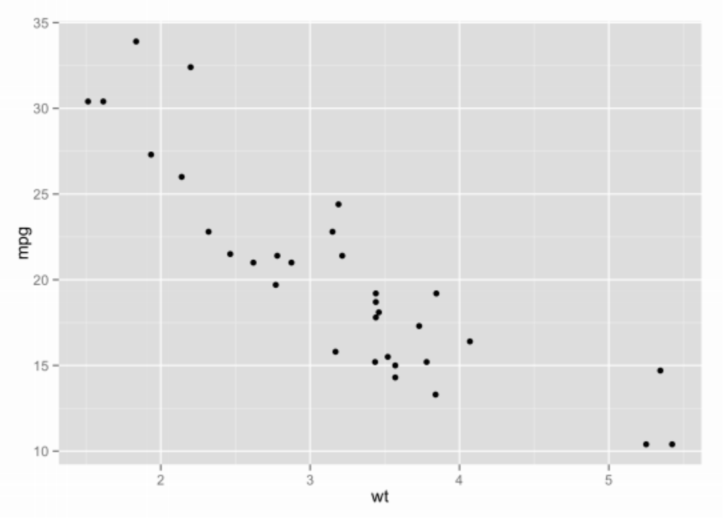
Regressions

- How do we determine the relationship between variables?
  - We will think about it like a function: X goes in, and Y comes out



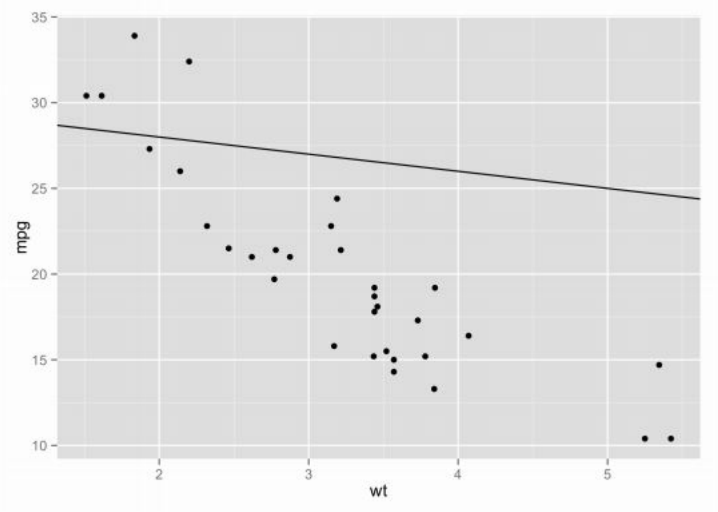
Regressions

- Instead of a nice functional relationship, we have messy clouds of data



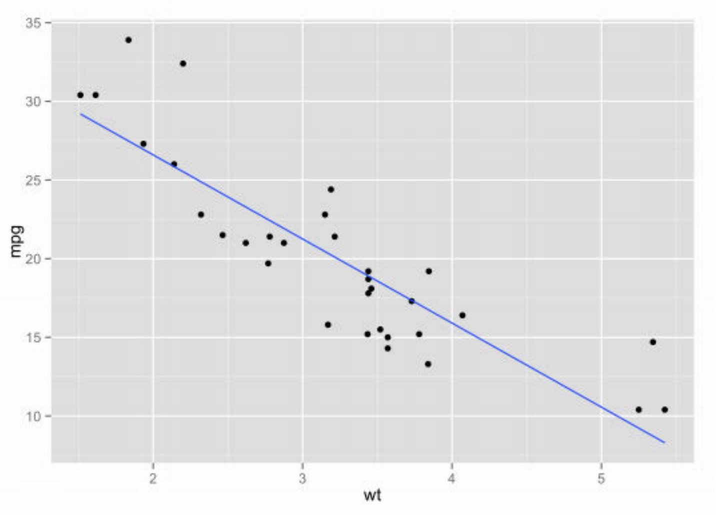
Regressions

- The goal is to pick a function that fits the data “well”



Regressions

- What does it mean for this line to fit “better” than the last?



Regressions

- The main model we use is called “linear regression”
  - We will take a straight line (hence linear)
  - We want it to be near the data.
  - The line which was “bad” was further away than seemed necessary
- We use an equation like:

$$\mathrm{Y} = \alpha + \beta X + \varepsilon$$

The Regression Model

- This is not a math course, but let’s look at what goes into this model
  - X, Y – random variables. We have draws of this data out of some population.
  - α, β – parameters. These parameters represent the relationship between X and Y
  - ε – This is the error term: the other things which are not in X but affect Y
- More intuitively
  - If Y is education and X is schooling
  - β is the amount that an extra unit of schooling is associated with higher wages
  - ε represents other factors which are not schooling that also affect wages

Predictions

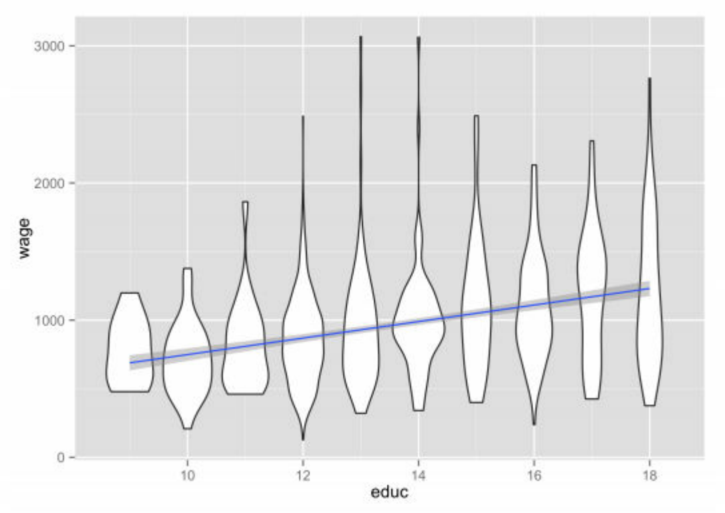
- How does this relate to prediction?
- We are trying to predict y_i with x_i
- How do we do this?
  - We know the formula for y_i from our model

    $$y_i = \alpha + \beta x_i + \varepsilon_i$$
  - But if we don't know y_i , we don't know ε_i either.
  - The best we can do is:

    $$\hat{y}^i = \alpha + \beta xi$$
- $\hat{y}^i$ is called the predicted value of y

Wages and Education

- X is years of schooling, Y is wages
- Prediction



What Things Can We Model With A Regression?

- What is Y ?
- What is X?

Examples

- Crime and policing at the state level
  - What is the sampling frame?
  - What is X, Y ?
- Crime as a function of income from survey data
  - What is the sampling frame?
  - What is X, Y ?

Where Next?

- First, we’ll talk about assumptions
- Linear regression was a very simple model
- But we can discuss a wide variety of extension now that we know about this model

Lesson Summary

- We try to predict an outcome random variable Y with an input random variable X
- To do this, we use a sample of pairs of (x_i , y_i) observation
- We choose the “best” parameters to fit the model
- We get a relationship between X and Y that can be used for prediction or analysis