Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Anomaly Detection

Outlier Detection (Anomaly Detection)

- So far, we have focused on choosing models which will be useful to predict Y
- In many situations, this isn't really what we want to do with our data
  - Example: Enron project!
  - Our Y variable is really "fraud" which is unobserved
- The idea behind anomaly detection is to pick observations which seem unusual compared to the rest of the data

What Is Unusual?

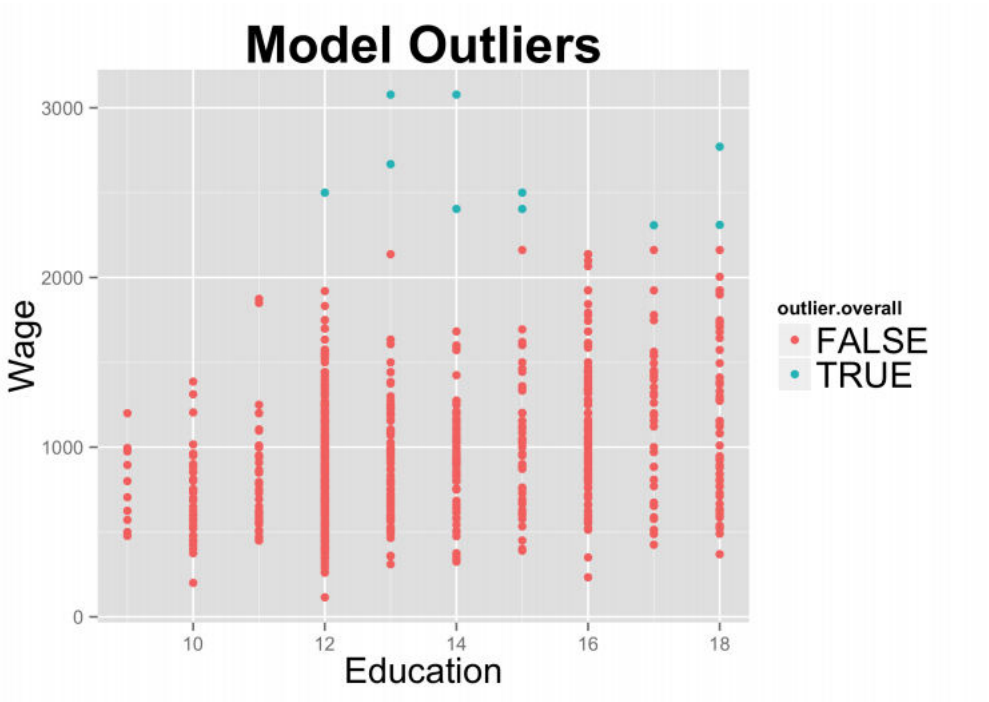
- We can determine what's "usual" in the data in the context of one of our models:

$$y = f(X) + \varepsilon$$

  - Look at the distribution of ε
  - Are there points very badly fit?
- Or we can do this without a model

Outlier Detection: No Model

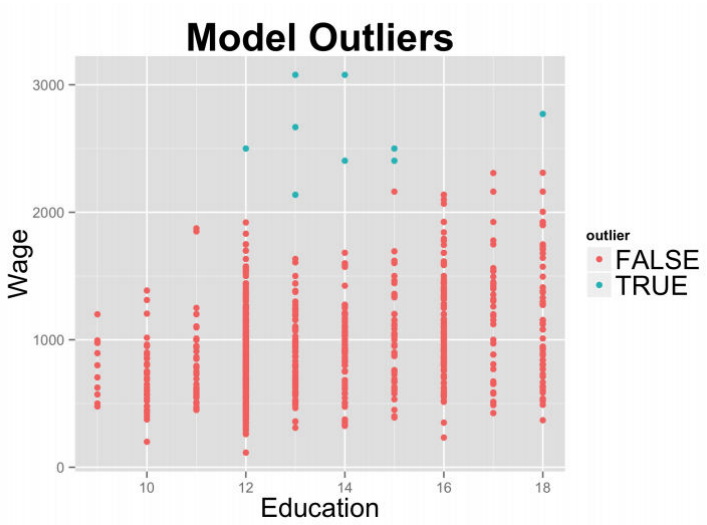
- Definition: An outlier is more than 3 standard deviations from mean



Georgia Tech

Outlier Detection: Using Regression Model

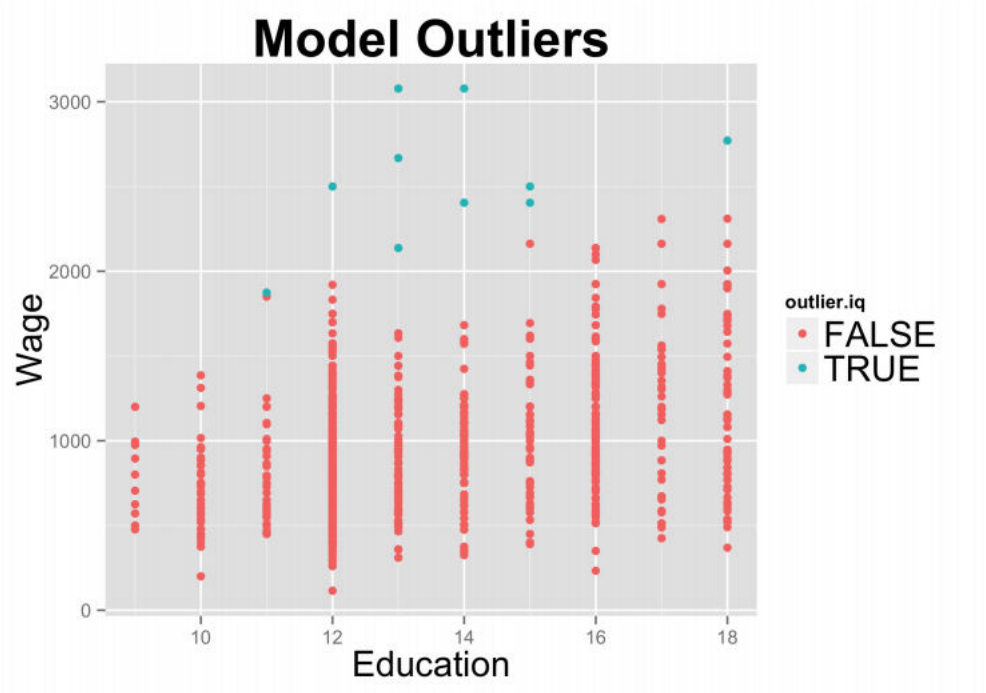
- Model: $wage = \alpha + \beta\, educ + \varepsilon$
- **Definition**: An outlier has $\hat{\varepsilon}$ more than 3 standard deviations from the mean



Georgia Tech

Outlier Detection Depends on Modeling

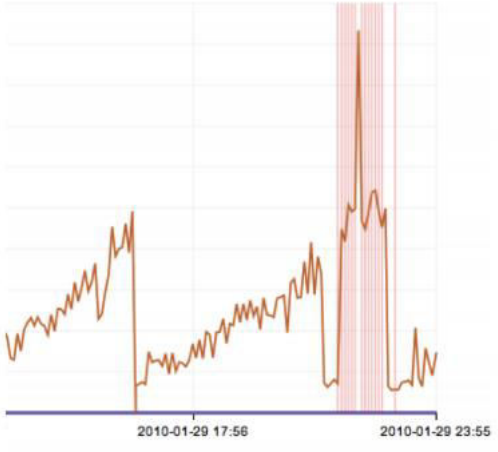
- Model: $wage = \alpha + \beta\, educ + \gamma IQ + \varepsilon$
- **Definition**: An outlier has $\hat{\varepsilon}$ more than 3 standard deviations from the mean



Georgia Tech

We Can Have Time Series Anomalies

- Some points aren't necessarily anomalies except when considered in their context
- This can apply to time series as well



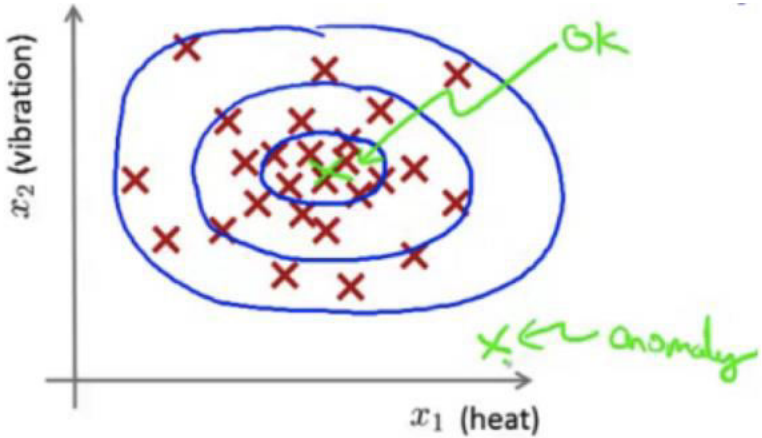
Georgia Tech

Key Pieces of Outlier Detection

1. Define a measure of distance
    - This can be based on a model (with residuals) or not
    - It can have multiple dimensions
2. Calculate this measure for each observation
3. Look at the points with the largest values

Anomaly in 2 Dimensions

- Once you have a measure of distance in 2 dimensions, you can calculate anomalies this way too
- Or you could think about using a model and looking at residuals



Georgia Tech

Lesson Summary

- Anomaly detection is to pick observations which seem unusual compared to the rest of the data
- It can be done with or without a model
- To find an anomaly, you pick a measure of distance, calculate it for each observation, then look at the points with the largest values
- Whether a point is an anomaly depends on its "context" or what model you're using to assess it.

Georgia Tech