Introduction

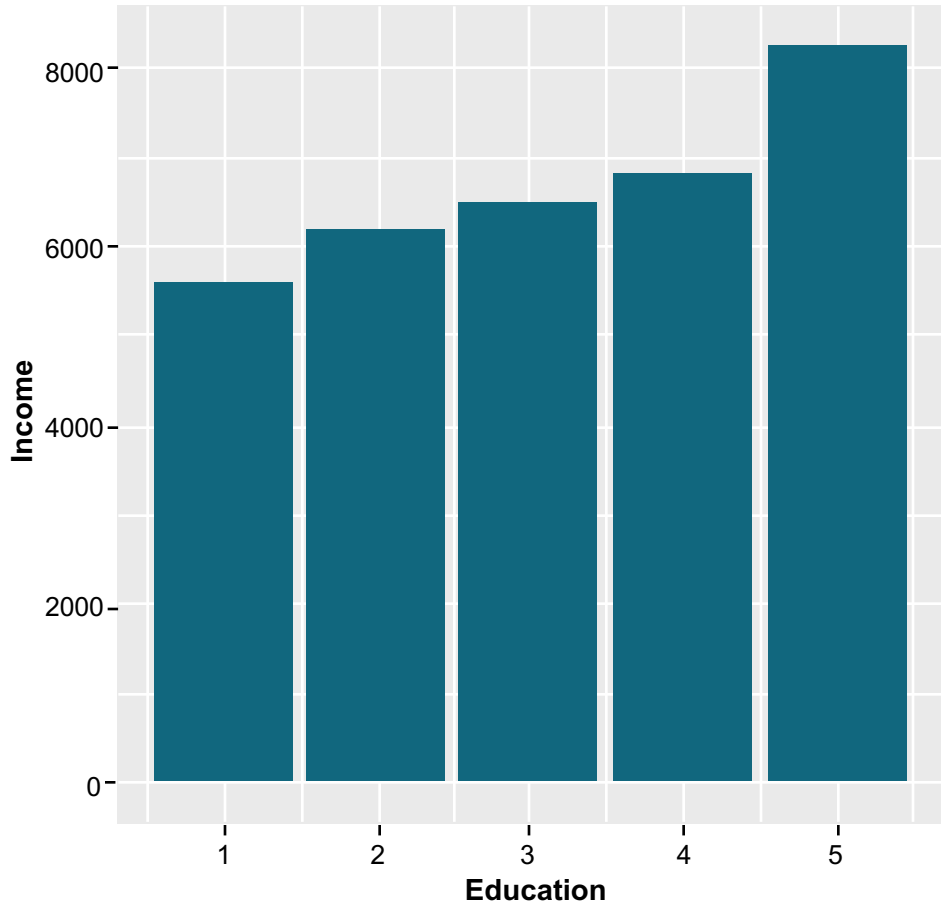
The Mighty Conditional Mean

Will Doyle

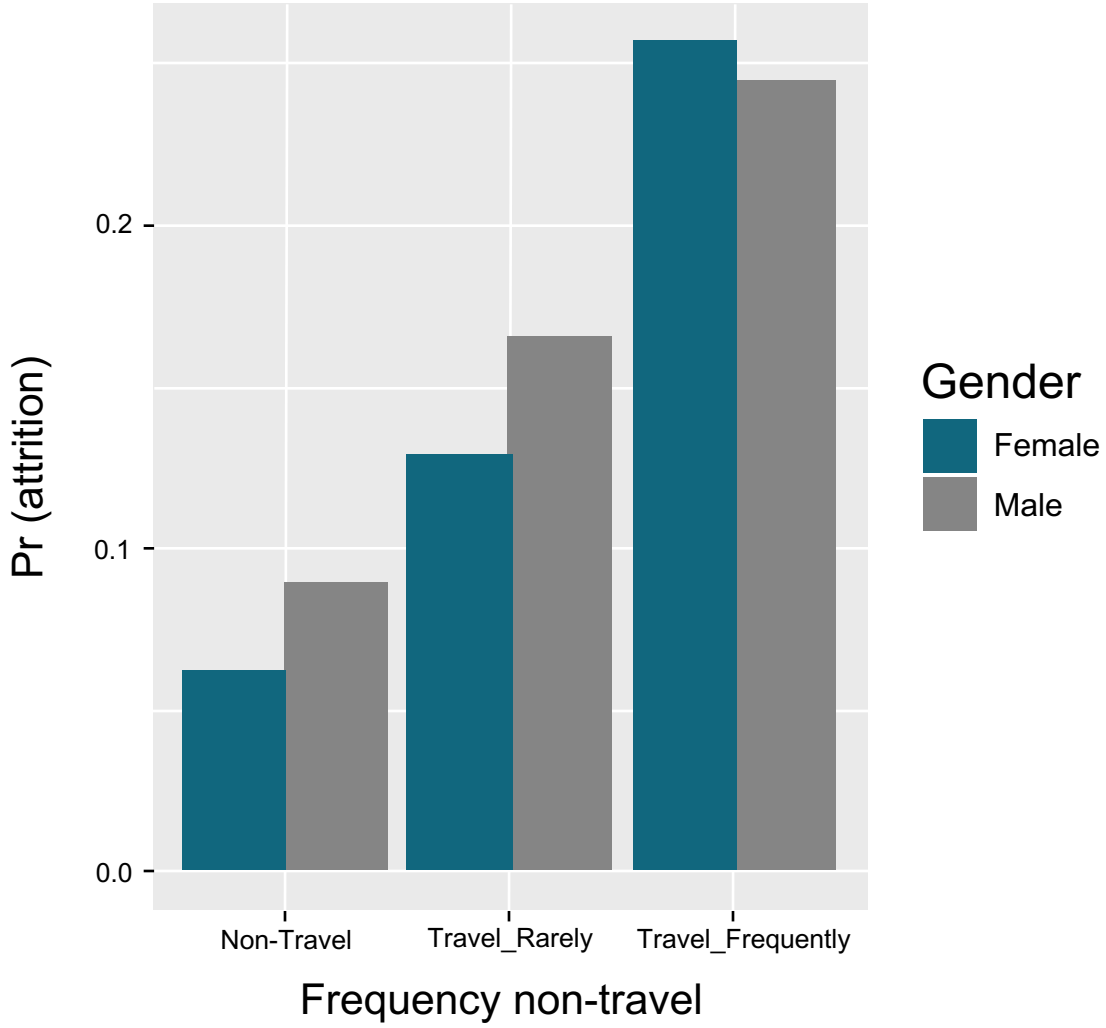
Applications of the Conditional Mean

- Course attendance by day of the week
- School enrollment by number of young people
- Teacher attrition by age
- Test scores by GPA
- Wages by education level

Example: Wages by Education Level

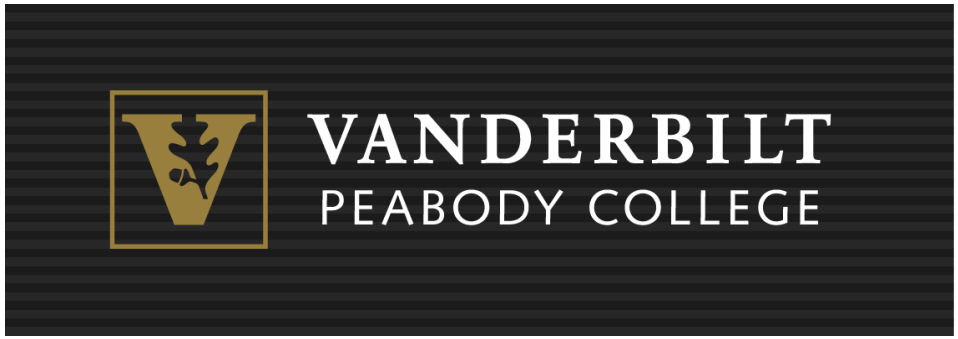


Example: Attrition by Travel and Gender



The Conditional Mean as a Prediction

- What do you predict the temperature will be where you live next July?
- You probably would look up average temperatures.
- What do you predict an engineer with 2 years of experience will make next year?
- You would probably look up average income by experience for engineers.


VANDERBILT
PEABODY COLLEGE

The Logic of Prediction

Will Doyle

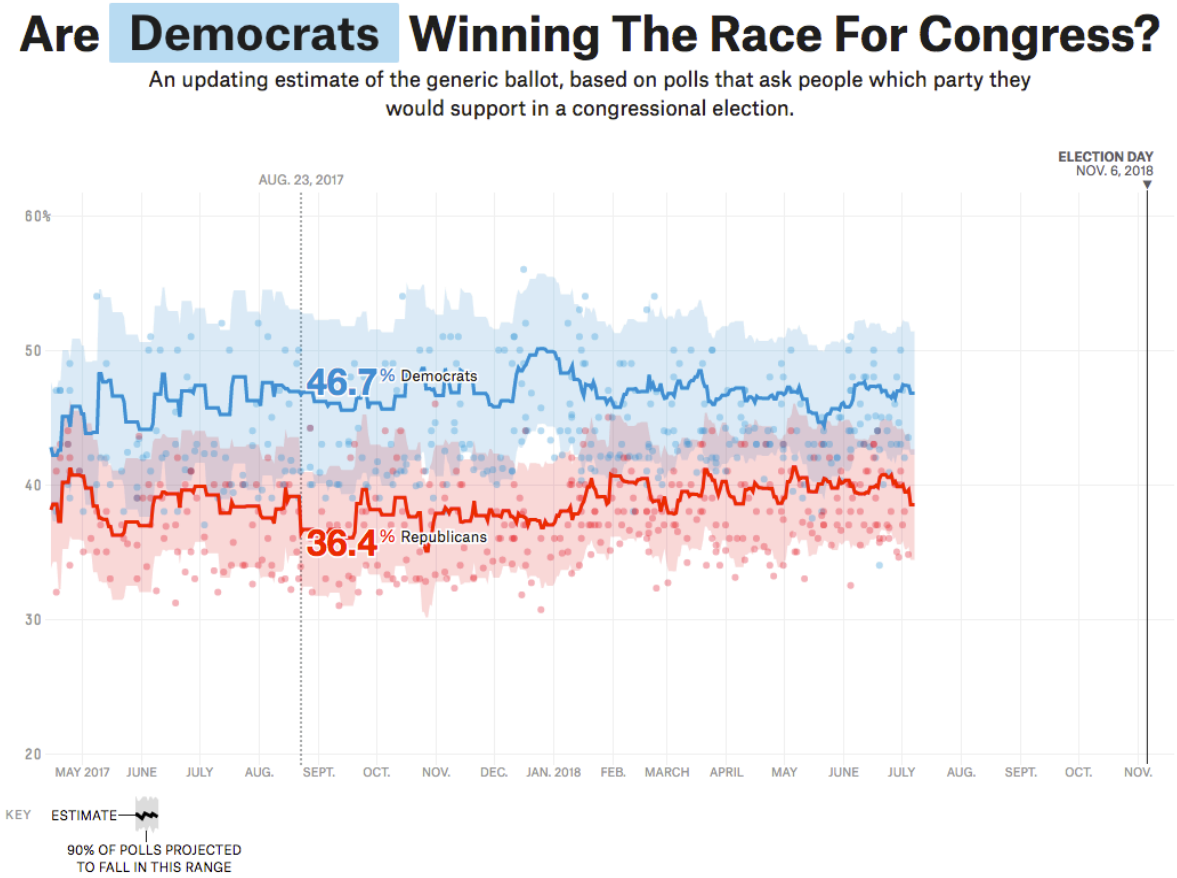
Predictions Using Data

- What do we need to predict?
  - How many students will enroll in my school?
  - How many customers will buy my product?
  - How many of my employees will leave next year?

Logic of Prediction

- We predict future events from past occurrences.
- In the language of data science, we use data from the past to “train” our models, then “test” our predictions against future events.
- A good model will accurately predict future occurrences.
- A bad model will poorly predict future occurrences.

Examples: Good Predictions



Source: https://projects.fivethirtyeight.com/congress-generic-ballot-polls/?ex_cid=rrpromo

Examples: Bad Predictions

Data mining program designed to predict child abuse proves unreliable, DCFS says

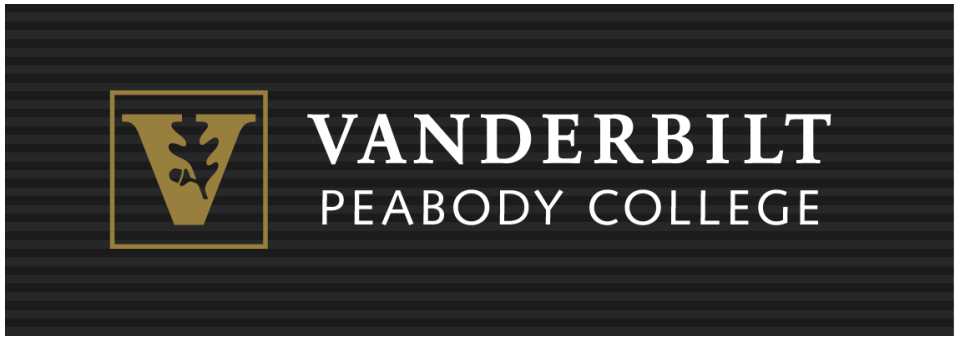
Beverly Walker, director of the Illinois Department of Children and Family Services, announced this week that the department is ending a high-profile program that used computer data mining to identify children at risk for serious injury or death. (Nancy Stone / Chicago Tribune)

By **David Jackson and Gary Marx · Contact Reporters**
Chicago Tribune

Source: http://www.chicagotribune.com/news/watchdog/ct-dcfs-eckerd-met-20171206-story.html

How We Check Our Models

- Make a prediction (e.g., predict conditional means for each group).
- Compare the prediction to the actual data.
- See how big the difference is.


VANDERBILT
PEABODY COLLEGE

How Wrong Are We?

The Root Mean Squared Error

Will Doyle

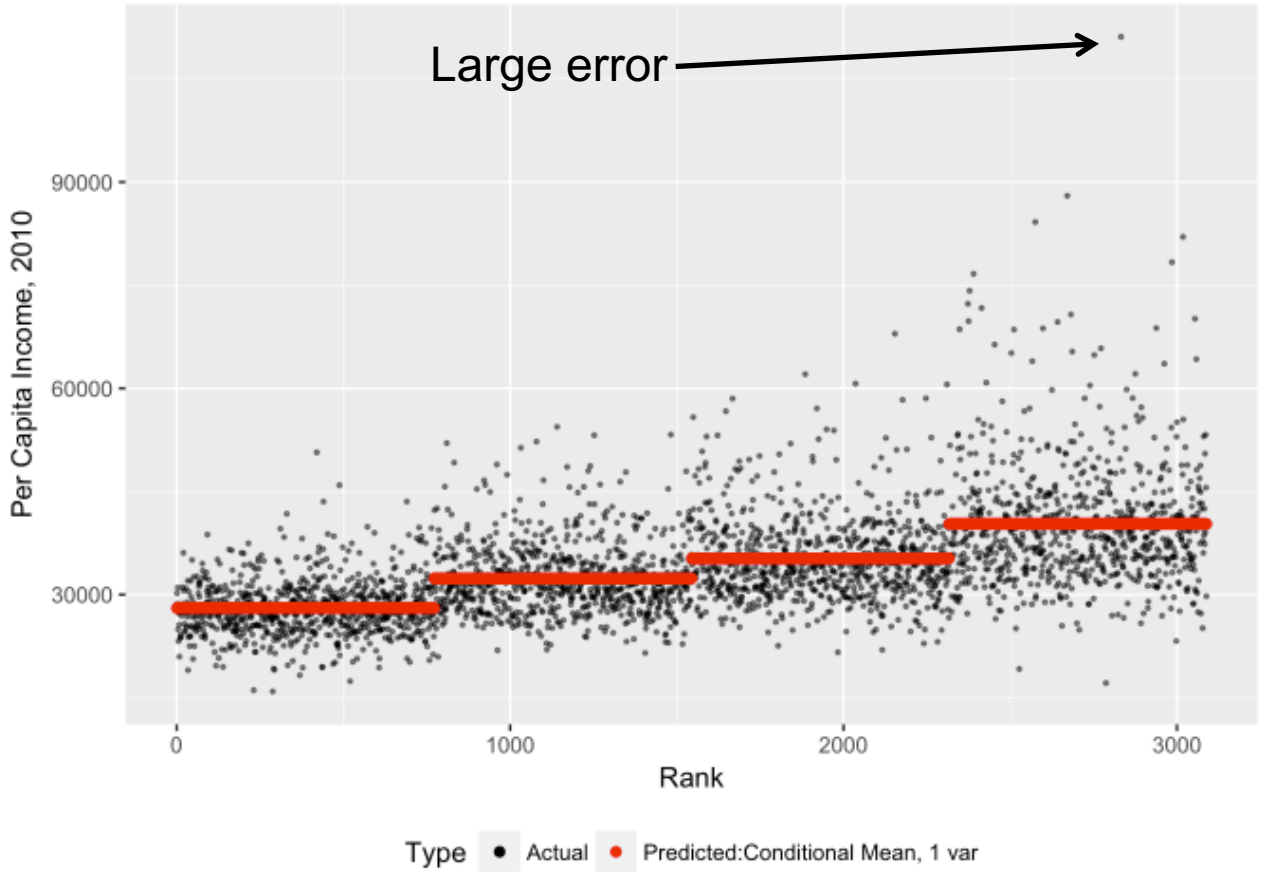
Quantifying Error

- Error = actual data – prediction
- Larger errors = less predictive accuracy
- We can quantify this using the MSE

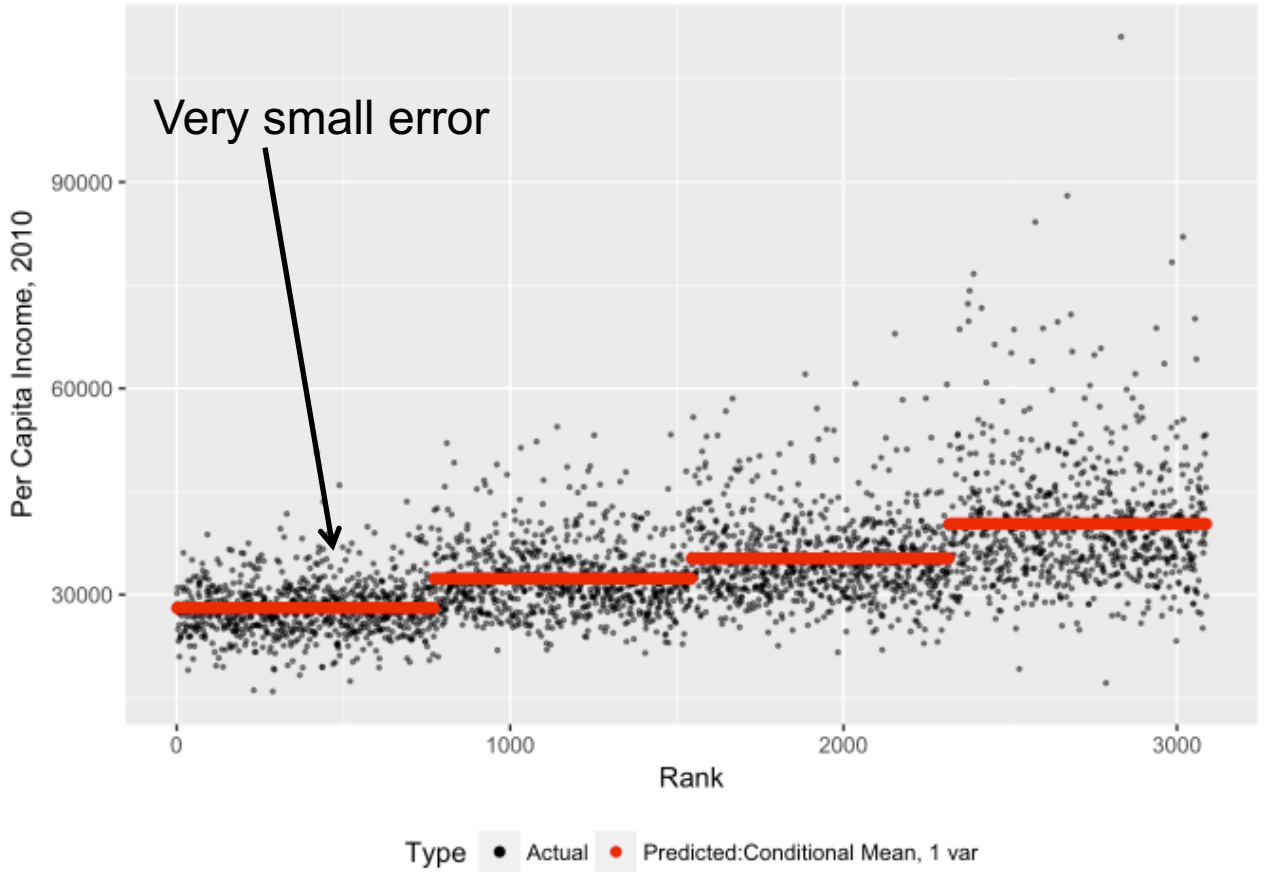
Root Mean Squared Error

$$RMSE(\hat{Y}) = \sqrt{1/n \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

Root Mean Squared Error

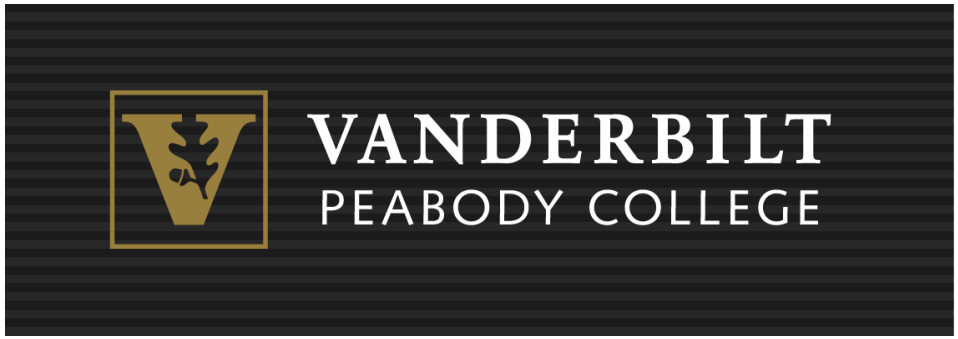


Root Mean Squared Error



Root Mean Squared Error

- This measure “averages” the size of all of the error terms.
- The square term is necessary, as errors average out to 0 for many predictors.
- It is expressed in units of the outcome variable.
- What’s “big” or “small” depends on the outcome of interest.


VANDERBILT
PEABODY COLLEGE