Regression: The Best Line

Ch. 5.2

Spring 2020

The Setup

So you've identified two numeric variables that seem to have a moderate to strong linear pattern. How can you find you find the best fitting line? "Eyeballing" it just isn't going to cut it. We need a more precise way.

- We can use the idea of **residuals** to quantify what "best line" means for modelling data.
- Slope and intercept for the best line can easily be estimated using simple summary statistics (or R).
- Predictions for new observations can then be made from our regression model.

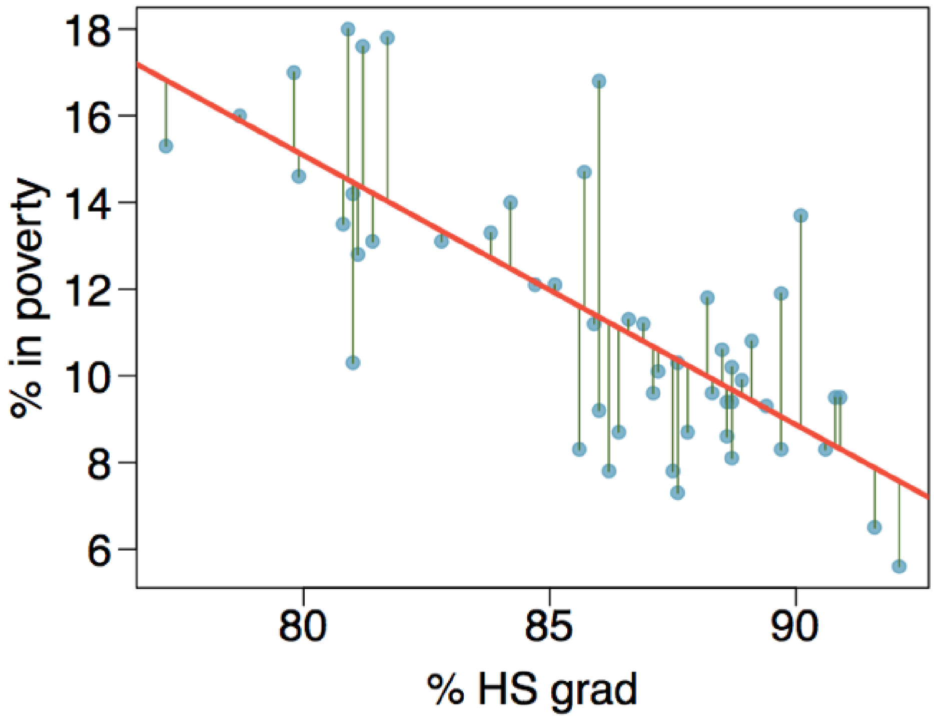
Most important

We will be able to interpret the slope coefficient for our best fitting linear regression equation. Why is this so important, because we don't want to lose sight of the original purpose of ANY statistical analysis, defineing patterns and relationships.

> The slope of a linear regression line is a way to represent the relationship between response and explanatory variables. It will be the most important aspect of the regression model and what we focus on for interpretation and inference.

Least Squares Regression

Recall from the last tutorial that **residuals** are the leftovers from a model, the error between an observed and predicted value.



It would make sense then, that the *best line* for any given scatterplot, is the one that minimizes all of the residuals. This is the idea behind **Least Squares Regression**, a method for finding that best line. We won't get into the technical details of how least squares regression works, but I want to cover just the basic concept:

- We consider many potential lines for the best line (each with different slope and intercept)
  - technology allows us to consider ALL potential lines at the same time
- We **square** all of the residuals for each line
  - residuals are squared so all become positive, and so we can give a bigger penalty to points that are really far from the model
- The line that gives us the **smallest sum of all squared residuals** is our best model for the data.
  - hence the name, least squares regression

I think there is a benefit to visualizing this process. Watch the video below of me walking you through different lines for a dataset and see the squared residuals change. Then you can follow the link provided to play around yourself.

Visualize least square regression