Simple Linear Regression

Ryan Miller

Introduction

- The purpose of this presentation is to introduce the *formal details* of linear regression
  - This will focus on the framing, logic, and notation used by statisticians
  - I'm assuming you're already familiar with basic concept of a straight-line model

Population-level Models

- Statisticians choose models that they presume will be useful *at the population level*
  - This choice might be informed by data exploration, but the goal typically is to generalize beyond the observed data

Population-level Models

- Statisticians choose models that they presume will be useful *at the population level*
  - This choice might be informed by data exploration, but the goal typically is to generalize beyond the observed data
- In **simple linear regression**, a straight-line is determine the *expected value* of an outcome variable at a given value of the explanatory variable

Model	Expectation
$y = \beta_0 + \beta_1 x + \epsilon$	$E(y) = \beta_0 + \beta_1 x$

- β_0 and β_1 are assumed to be fixed, but unknown *population parameters*

Errors

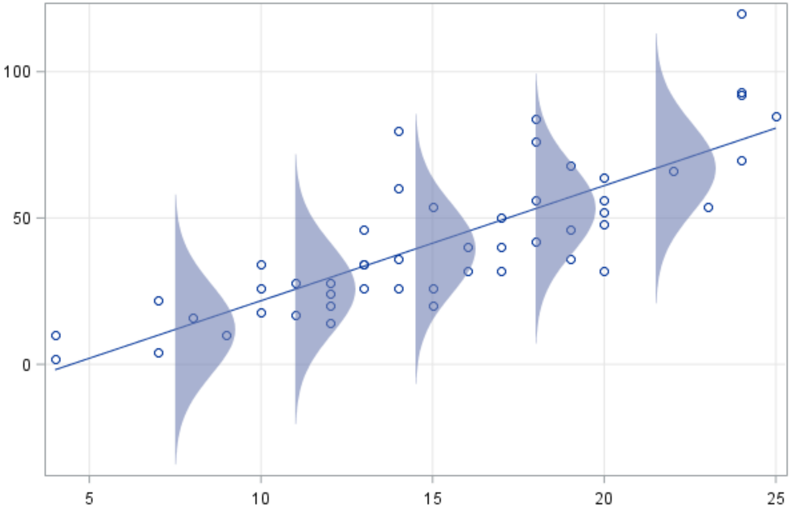
- The error component, ϵ, allows the model to be mathematically true without needing to pass through every data-point
  - These errors are assumed to follow a Normal distribution, $\epsilon \sim N(0, \sigma^2)$

Errors

- The error component, ϵ, allows the model to be mathematically true without needing to pass through every data-point
  - These errors are assumed to follow a Normal distribution, $\epsilon \sim N(0, \sigma^2)$
- While x and y are both observed, only y is a random variable
  - $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$, by virtue of ϵ

Model	Expectation
$y = \beta_0 + \beta_1 x + \epsilon$	$E(y) = \beta_0 + \beta_1 x$

Errors



Least Squares Estimation

- The slope and intercept are estimated from the observed data by solving for the line that minimizes the *residual sum of squares*

$$RSS = \sum_i r_i^2 \text{ where } r_i = y_i - E(y_i)$$

- These estimates can be found using calculus (something we'll gloss over):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})}$$

Maximum Likelihood

- The original motivation behind least squares regression was that $\sum_i r_i^2$ is differentiable, while $\sum_i |r_i|$ is not

Maximum Likelihood

- The original motivation behind least squares regression was that $\sum_i r_i^2$ is differentiable, while $\sum_i |r_i|$ is not
- It was later discovered that if y is Normally distributed, $\hat{\beta}_0$ and $\hat{\beta}_1$ are also the **maximum likelihood estimates** of β_0 and β_1
  - We won't go too far into likelihood theory in this course, but MLEs have some nice theoretical properties (which are shared by least squares estimates in the case of linear regression)

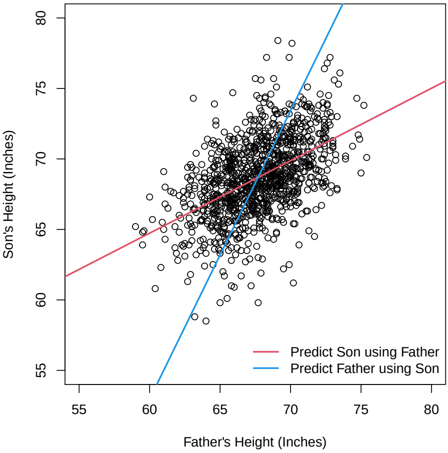
Two Regression Lines

- Because least squares optimizes *vertical deviations* (between y and $E(y)$), the explanatory and response variables are *not interchangeable*

$$\frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})} \neq \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})}$$

- This means that there are two possible regression lines for any pair of numeric variables

Two Regression Lines



Regression vs. Correlation

- The regression line is mathematically related to the correlation coefficient

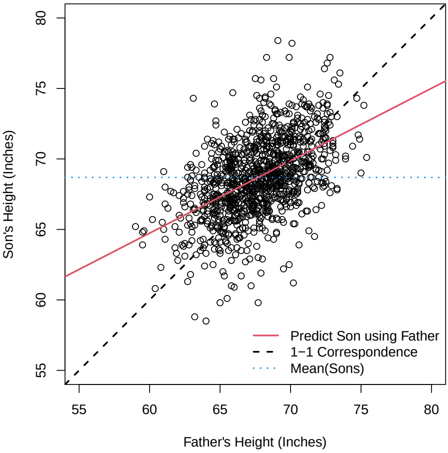
$$\hat{\beta}_1 = r * \frac{s_y}{s_x}$$

- When two variables are perfectly correlated ($r = 1$), the slope is just the ratio of standard deviations
  - Each 1 SD increase in x predicts a 1 SD increase in y

Regression vs. Correlation

- The regression line is mathematically related to the correlation coefficient
$$\hat{\beta}_1 = r * \frac{s_y}{s_x}$$
- When two variables are perfectly correlated ($r = 1$), the slope is just the ratio of standard deviations
  - Each 1 SD increase in x predicts a 1 SD increase in y
- When the correlation is imperfect, each 1 SD increase in x predicts an $r < 1$ SD increase in y
  - This "regression towards the mean" is how the method got it's name

Regression Towards Mediocrity



Closing Remarks

- Regression is a very general and widely-used modeling framework
  - It is statistical, so we can use our fitted models to make statistical inferences about a population
  - It is interpretable, so we can clearly describe the relationships suggested by the model
- In the coming weeks, we'll further generalize this method to incorperate *multiple explanatory variables*, a scenario in which the method really shines