Introduction to Multiple Linear Regression

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

Today's lecture

- Multiple Linear Regression
  - Assumptions
  - Interpretation

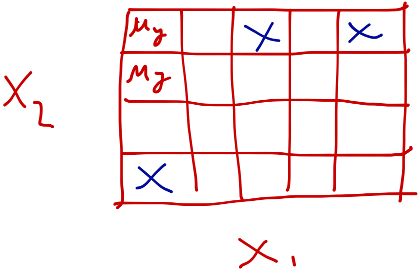
Motivation

Most applications involve more that one covariate – if more than one thing can influence an outcome, you need multiple linear regression.

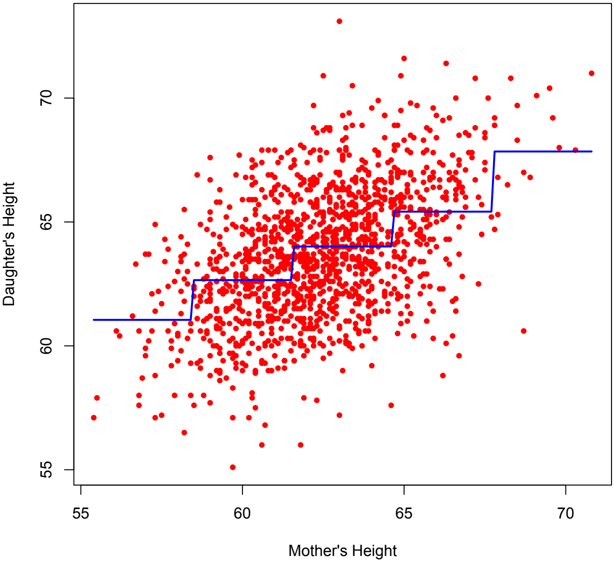
- Improved description of $y|\mathbf{x}$ bold face
- More accurate estimates and predictions
- Allow testing of multiple effects
- Includes multiple predictor types

Why not bin all predictors?

- Divide x_i into k_i bins
- Stratify data based on inclusion in bins across x's
- Find mean of the y_i in each category
- Possibly a reasonable non-parametric model



Why not bin all predictors?



Why not bin all predictors?

- More predictors = more bins
- If each x has 5 bins, you have 5^p overall categories
- May not have enough data to estimate distribution in each category
- Curse of dimensionality is a problem in a lot of non-parametric statistics

Multiple linear regression model

- Observe data $(y_i, x_{i1}, \ldots, x_{ip})$ for subjects $1, \ldots, n$. Want to estimate $\beta_0, \beta_1, \ldots, \beta_p$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \epsilon_i; \ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- Assumptions (residuals have mean zero, constant variance, are independent) are as in SLR
- Impose linearity which (as in the SLR) is a big assumption
- Our primary interest will be $E(y|\mathbf{x})$
- Eventually estimate model parameters using least squares

Predictor types

- Continuous
- Categorical
- Ordinal

Interpretation of coefficients

$\beta_0 = E(y|x_1 = 0, \ldots, x = 0)$

- Centering some of the x's may make this more interpretable

$$E(y_i | \vec{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 X_{2i} + \ldots + \beta_p x_{pi}$$

Interpretation of β_1

$$E(y_i \mid \vec{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

β_1 = expected change in Y for a 1 unit change in X_1, holding all other Xs constant

$$E(y \mid x_1 = 4, x_2 = k) =$$

$$E(y \mid x_1 = 5, x_2 = k) =$$

Example with two predictors

Suppose we want to regress weight on height and sex.

- Model is $y_i = \beta_0 + \beta_1 x_{i,age} + \beta_2 x_{i,sex} + \epsilon_i$
- Age is continuous starting with age 0; sex is binary, coded so that $x_{i,sex} = 0$ for men and $x_{i,sex} = 1$ for women

Example with two predictors

$\beta_1 =$

$\beta_2 =$

Coming up next...

Multiple linear regression models

- notation
- estimation
- inference