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
  - Interpretation
  - Notation

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

Interpretation of β_1

Omitted variable bias

What happens if we ignore x_2 and fit the simple linear regression:

$$y_i = \beta_0^* + \beta_1^* x_{i,1} + \epsilon_i^*$$

Does $\beta_1^* = \beta_1$?

Omitted variable bias

When should you be concerned?

If both of the following conditions are met, then $\beta_1^* = \beta_1$:

- The omitted variable is unrelated to the outcome
- The omitted variable is uncorrelated with the retained variable

Extra credit for problem set 1: create a simulation where you show an example of omitted variable bias.

Matrix notation

- 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

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

- Notation is cumbersome. To fix this, let
  - $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{ip}]$
  - $\boldsymbol{\beta}^T = [\beta_0, \beta_1, \ldots, \boldsymbol{\beta}_p]$
  - Then $y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i$

Matrix notation

- Let

$$\mathbf{y} = \left[\begin{array}{c} y_1 \\ \vdots \\ y_n \end{array}\right], \quad \mathbf{X} = \left[\begin{array}{cccc} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{array}\right], \quad \boldsymbol{\beta} = \left[\begin{array}{c} \beta_0 \\ \vdots \\ \beta_p \end{array}\right], \quad \boldsymbol{\epsilon} = \left[\begin{array}{c} \epsilon_1 \\ \vdots \\ \epsilon_n \end{array}\right]$$

- Then we can write the model in a more compact form:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

- $\mathbf{X}$ is called the *design matrix*

Matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- ϵ is a random vector rather than a random variable
- $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 I$
- Note that *Var* is an abuse of notation; in the present context it really means the "variance-covariance matrix"

Today's big ideas

- Multiple linear regression models, interpretation, notation, biases