Multiple Linear Regression

ST552 Lecture 4

Charlotte Wickham

2019-01-14

Today

- Matrix warmup
- Multiple Linear Regression
- Matrix setup

Matrix warmup

See handout

Simple linear regression

Recall in simple linear regression:

Have n observations of a response y_i, and a single explanatory variable, x_i.

The response is related to the explanatory variable by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

where ϵ_i are independent and identically distributed with expected value 0, and variance σ^2.

Multiple linear regression

Now we have **more than one explanatory variable**.

Have n observations of a response, y_i and **a set of** explanatory variables, $(x_{i1}, x_{i2}, \ldots, x_{i(p-1)})$.

The response is related to the explanatory variables by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i(p-1)} + \epsilon_i \quad i = 1, \ldots, n$$

where ϵ_i are independent and identically distributed with expected value 0, and variance σ^2.

Example: Galápagos Islands

Faraway 2.6

Measurements on 30 Galápagos Islands are made.

First 5 islands:

	Species	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.3
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.1	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.8

Variable Descriptions

```
?gala
Species diversity on the Galapagos Islands

Format:

     The dataset contains the following variables

     'Species' the number of plant species found on the island

     'Endemics' the number of endemic species

     'Area' the area of the island (km$^2$)

     'Elevation' the highest elevation of the island (m)

     'Nearest' the distance from the nearest island (km)

     'Scruz' the distance from Santa Cruz island (km)

     'Adjacent' the area of the adjacent island (square km)
```

A possible model

$$\text{Species}_i = \beta_0 + \beta_1 \text{Area}_i + \beta_2 \text{Elevation}_i + \beta_3 \text{Nearest}_i + \beta_4 \text{Scruz}_i + \beta_5 \text{Adjacent}_i + \epsilon_i \quad i = 1, \ldots, n$$

E.g. $i = 1$, *Baltra*:

$$58 = \beta_0 + \beta_1 25.09 + \beta_2 346 + \beta_3 0.6 + \beta_4 0.4 + \beta_5 1.84 + \epsilon_1$$

Your turn:

- What does i index?
- What is the value of n?
- What is the value of p?

General matrix form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$y = X\beta + \epsilon$$

where

$$y_{n\times 1} = (y_1, y_2, \dots, y_n)^T$$

$$\epsilon_{n\times 1} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$$

$$\beta_{p\times 1} = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$$

$$X_{n\times p} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{pmatrix}$$

Galápagos: Matrix form

$$y_{30\times 1} = \begin{pmatrix} 58 \\ 31 \\ 3 \\ 25 \\ 2 \\ \vdots \end{pmatrix}, X_{30\times 6} = \begin{pmatrix} 1 & 25.09 & 346 & 0.6 & 0.6 & 1.84 \\ 1 & 1.24 & 109 & 0.6 & 26.3 & 572.33 \\ 1 & 0.21 & 114 & 2.8 & 58.7 & 0.78 \\ 1 & 0.1 & 46 & 1.9 & 47.4 & 0.18 \\ 1 & 0.05 & 77 & 1.9 & 1.9 & 903.82 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$\beta_{6\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}, \quad \epsilon_{30\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \end{pmatrix}$$

Your Turn

Write out the design matrix, X, for the following models, using the data for the first five islands:

$$\text{Species}_i = \beta_0 + \beta_1 \text{Area}_i + \beta_2 \text{Nearest}_i + \epsilon_i$$
$$\text{Species}_i = \beta_1 \text{Area}_i + \beta_2 \text{Area}_i^2 + \epsilon_i$$
$$\text{Species}_i = \beta_0 + \beta_1 1_{\{\text{Area}_i > 1\}} + \epsilon_i$$

where $1_{\{.\}}$ is an indicator variable that takes the value 1, when the condition in the argument is true, and 0 otherwise.

Fitted values and residuals

If we had an estimate for the β vector,

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}\right)^T$$

Then we can define fitted value and residual vectors:

$$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)^T = X\hat{\beta}$$
$$e = \hat{\epsilon} = (e_1, \ldots, e_n)^T = y - X\hat{\beta}$$

Questions to answer this week:

- How will we find $\hat{\beta}$?
- What properties do the estimates have?