

PSTAT 126

Regression Analysis

Laura Baracaldo

Lecture 2

Simple Linear Regression Models

Quiz

- Roughly 4+ quizzes, on Canvas every week. (starting Next Monday (08/14))
- Notification will be sent out well in advance of the actual quiz time.
- You decide when to start the quiz (over a period of several hours).
- Once you start the quiz, **you have 20 minutes** to finish it.

Homework

- 4 Hw – First was posted today (08/08) and will be due in 1.5 weeks (08/18)
- You have to submit the .Rmd file and the generated PDF document.

Matrix Representation of the Regression Equation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Y = \underbrace{X\beta}_{\substack{\text{Systematic} \\ \text{Component}}} + \underbrace{\epsilon}_{\substack{\text{Random} \\ \text{Component}}}$$

Regression Equation:

- The column of ones incorporates the intercept term.
- The simplest example is the *null model*, where there is no predictor, just a constant mean: $Y = \mu + \epsilon$.

Estimating β

- We obtain estimation for β so that the systematic part explains as much of the response as possible, which is equivalent to "shrink" the random component as much as possible.
- For $\hat{Y} = X\hat{\beta}$ (predicted or fitted values) and $\hat{\epsilon} = Y - \hat{Y}$ (residuals); Our goal boils down to minimizing: $\|\hat{\epsilon}\| = \|Y - \hat{Y}\| = \|Y - X\hat{\beta}\|$.

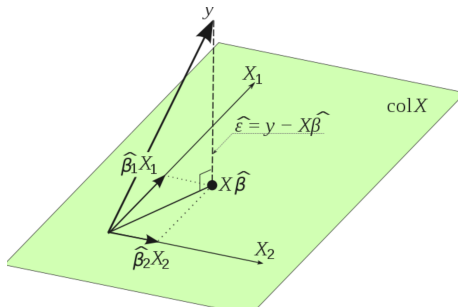
Geometric Interpretation

- The solution for $\hat{\beta}$ in this optimization problem corresponds to $\hat{\beta}$ s.t. \hat{Y} is the orthogonal projection of Y onto the column space of X :

$$\hat{Y} = X\hat{\beta} = Hy$$

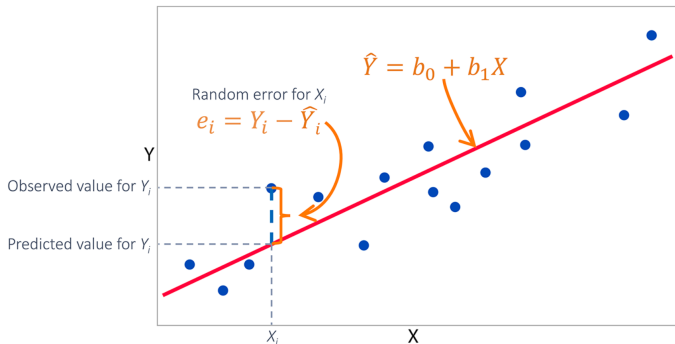
H is called the orthogonal projection matrix.

- For $p = 2$, the geometrical representation is:



Least-Squares (LS) Estimation

From a non geometric perspective: The best estimate of β , is the $\hat{\beta}$ that minimizes the sum of squared residuals (SSR):



$$\arg \min_{\beta} SSR = \arg \min_{\beta} \sum_{i=1}^n \hat{e}_i^2 = \arg \min_{\beta} \hat{e}^T \hat{e} = \arg \min_{\beta} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Least-Squares (LS) Estimation

We differentiate with respect to β :

$$\begin{aligned}\frac{\partial}{\partial \hat{\beta}}(y - X\hat{\beta})^T(y - X\hat{\beta}) &= \frac{\partial}{\partial \hat{\beta}}(y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}) \\ &= -y^T X - y^T X + 2\hat{\beta}^T X^T X \\ &= -2y^T X + 2\hat{\beta}^T X^T X \\ &= -2X^T y + 2X^T X\hat{\beta}\end{aligned}$$

This leads to the *Normal Equations*:

$$X^T X\hat{\beta} = X^T y$$

Least-Squares (LS) Estimation

provided $X^T X$ is invertible:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Which implies that the *fitted values* can be written as:

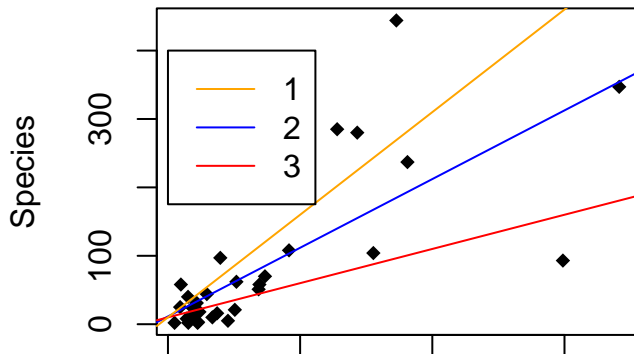
$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$$

Where H is called the *Hat Matrix*, and corresponds to the orthogonal projection of y onto the space spanned by the columns of X .

Species Example

What is the best line that represents the relationship between y (species) and x (Elevation)?

$$\text{Species}_i = \beta_0 + \beta_1 \text{Elevation}_i + \epsilon_i \quad i = 1, \dots, 30$$



LS for Simple Linear Regression (One predictor)

Suppose we only have one predictor x that can be used to explain the response y . Based on a data set with subjects $i = 1, \dots, n$ the linear regression model is written as:

Simple Linear Regression Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$

Goal: To estimate β_0 and β_1 , by solving:

$$\arg \min_{\hat{\beta}_0, \hat{\beta}_1} SSR = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \quad (1)$$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \quad (2)$$

LS for Simple Linear Regression

$$(1) \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$
$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

LS for Simple Linear Regression

$$\begin{aligned}(1) \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= 0 \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} &= 0\end{aligned}$$

Where $\bar{y} = \sum_{i=1}^n y_i/n$ and $\bar{x} = \sum_{i=1}^n x_i/n$.

LS for Simple Linear Regression

$$\begin{aligned}(1) \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= 0 \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} &= 0 \\ \Rightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

Where $\bar{y} = \sum_{i=1}^n y_i / n$ and $\bar{x} = \sum_{i=1}^n x_i / n$.

LS for Simple Linear Regression

$$(2)0 = \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

LS for Simple Linear Regression

$$\begin{aligned}(2)0 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)\end{aligned}$$

LS for Simple Linear Regression

$$\begin{aligned}(2)0 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x})\end{aligned}$$

LS for Simple Linear Regression

$$\begin{aligned}(2)0 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) \\&= \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(x_i - \bar{x})\end{aligned}$$

LS for Simple Linear Regression

$$\begin{aligned}(2) 0 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) \\&= \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(x_i - \bar{x}) \\&= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \bar{x}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) + \bar{x}(x_i - \bar{x})\end{aligned}$$

LS for Simple Linear Regression

$$\begin{aligned}(2) 0 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\&= \sum_{i=1}^n x_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) \\&= \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n ((x_i - \bar{x}) + \bar{x})(x_i - \bar{x}) \\&= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \bar{x}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) + \bar{x}(x_i - \bar{x}) \\&\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

Species Example

```
data(gala, package = "faraway")
y <- gala$Species
x <- gala$Elevation
beta1 <- sum((y-mean(y))*(x-mean(x)))/sum((x-mean(x))^2)
beta1
```

```
## [1] 0.2007922
```

```
beta0 <- mean(y)-beta1*mean(x)
beta0
```

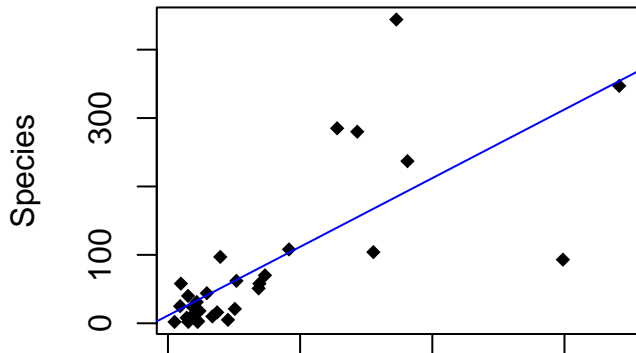
```
## [1] 11.33511
```

Species Example

```
data(gala, package = "faraway")  
fit<- lm( Species ~ Elevation, data=gala)  
fit
```

```
##  
## Call:  
## lm(formula = Species ~ Elevation, data = gala)  
##  
## Coefficients:  
## (Intercept)      Elevation  
##      11.3351         0.2008
```

Species Example



Species Example: SSR

```
fit<- lm( Species ~ Elevation, data=gala)
SSR <- sum((fit$residuals)^2); SSR
```

```
## [1] 173253.9
```

```
SSR2 <- sum((y- (10+0.3*x))^2); SSR2
```

```
## [1] 261110
```

```
SSR3 <- sum((y- (11+0.1*x))^2); SSR3
```

```
## [1] 267651.7
```

```
SSR4 <- sum((y- (11+0.2*x))^2); SSR4
```

```
## [1] 173268.9
```