STAT 420: Methods of Applied Statistics

Simple Linear Regression

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

University of Illinois at Urbana-Champaign
January 23, 2017

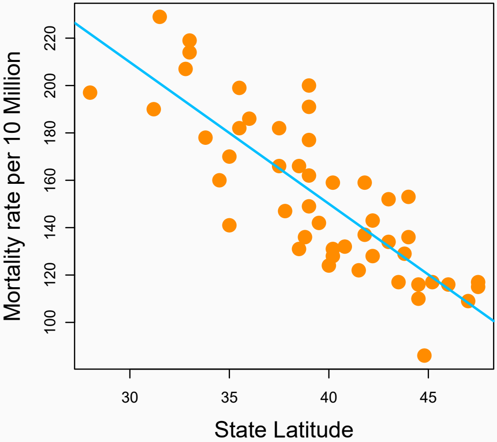
Simple Linear Regression

- Let's look at a simplified version of a linear model — with only one predictor.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- X is the predictor, Y is the outcome, and ϵ is a random error.
- β_0 and β_1 are unknown regression coefficients that we want to estimate.
- Suppose that a researcher can set the value of X, and perform experiments to observe Y. Repeatedly perform such experiments on different values of X will allow us to collect a set of data.

Simple Linear Regression



Skin cancer mortality rate per 10 million (1950s) by state latitude.

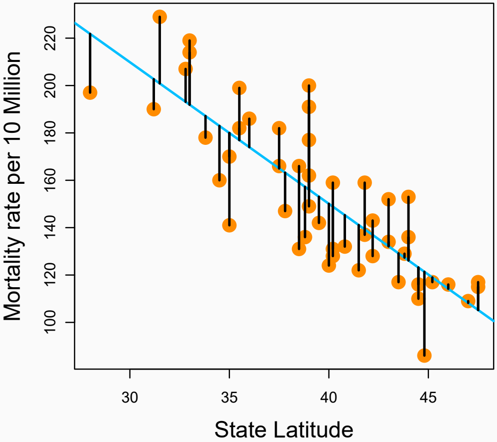
Simple Linear Regression

- What is the optimal line (with intercept $\widehat{\beta}_0$ and slope $\widehat{\beta}_1$) that describes this relationship based on the observed data?
- There are n observations, and for each $i \in 1, \ldots n$, we have
  - y_i the observed mortality rate for state i
  - x_i the latitude for state i
- Usually this is obtained by minimizing the sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

- Interpretation: $y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ measures the (vertical) distance between the observed point and the fitted line.

Simple Linear Regression



Skin cancer mortality rate per 10 million (1950s) by state latitude.

Minimizing the SSE

- How to minimize the sum of squared errors (SSE)?

$$
\begin{aligned}
(\widehat{\beta}_0, \widehat{\beta}_1) &= \arg\min_{\beta_0, \beta_1} \text{SSE} \\
&= \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}
$$

- It is usually believed that the technique was first discovered around 1805 by Adrien Marie Legendre (1752-1833).
- This is a quadratic function of both β_0 and β_1, hence is convex about its argument
- Take the derivative with respect to the parameters and set to zero:

$$
\frac{\partial \text{SSE}}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial \text{SSE}}{\partial \beta_1} = 0
$$

Legendre (left) and Fourier (right).

Minimizing the SSE

$$
\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
\frac{\partial \text{SSE}}{\partial \beta_0} &= 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0 \\
\frac{\partial \text{SSE}}{\partial \beta_1} &= 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \overset{\text{set}}{=} 0
\end{aligned}
$$

$$
\Longrightarrow \qquad
\begin{aligned}
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_y}{s_x}
\end{aligned}
$$

where r_{xy} is the sample correlation coefficient, and s_y and s_x are the sample standard error.

Example 1

Suppose we observe 8 sample points:

$$\mathbf{x} = (0.7, -0.1, 0.4, 0.3, -2.2, -2.5, -0.4, -1.3)^\mathsf{T},$$
$$\mathbf{y} = (2.2, -1.0, -0.5, 2.8, -2.8, -3.4, 0.1, -2.1)^\mathsf{T}.$$

Find the optimal simple linear regression line that describes the data.

Example 1

Suppose we observe 8 sample points:

$$\mathbf{x} = (0.7, -0.1, 0.4, 0.3, -2.2, -2.5, -0.4, -1.3)^\mathsf{T},$$
$$\mathbf{y} = (2.2, -1.0, -0.5, 2.8, -2.8, -3.4, 0.1, -2.1)^\mathsf{T}.$$

Find the optimal simple linear regression line that describes the data.

- First we calculate $\bar{x} = -0.6375$, $s_x = 1.221167$, $\bar{y} = -0.5875$, and $s_y = 2.235709$.
- Calculate $\widehat{\beta}_1 = r_{xy}\frac{s_y}{s_x} = 1.593462$
- Calculate $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} = 0.4283319$

Example 1

```r
# input the data
> x = c(0.7, -0.1, 0.4, 0.3, -2.2, -2.5, -0.4, -1.3)
> y = c(2.2, -1.0, -0.5, 2.8, -2.8, -3.4, 0.1, -2.1)

# calculate the means
> ybar = mean(y)
> xbar = mean(x)

# calculate β1
> sum((x - xbar)*(y - ybar))/sum( (x- xbar)^2)
[1] 1.593462

# calculate β0
> ybar - beta1*xbar
[1] 0.4283319

# another way to calculate β1 using the correlation coefficient
> cor(x, y)*sd(y)/sd(x)
[1] 1.593462
```

Example 1

```
# validate the result using the build-in function lm()
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.4283       1.5935
```

Example 2

Suppose a researcher wants to perform a linear regression on the samples he collected. However, the original data was lost, and he has only the access to some summary statistics

$$\bar{x} = 0.2875, \qquad \bar{y} = 0.0075,$$
$$\sum_{i=1}^{n} x_i y_i = 1.5941, \qquad s_x = 0.5460704.$$

Can you still figure out the regression line based on these available information?

Example 2

Suppose a researcher wants to perform a linear regression on the samples he collected. However, the original data was lost, and he has only the access to some summary statistics

$$\bar{x} = 0.2875, \qquad \bar{y} = 0.0075,$$
$$\sum_{i=1}^{n} x_i y_i = 1.5941, \qquad s_x = 0.5460704.$$

Can you still figure out the regression line based on these available information?

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x^2} = 0.7554315$$
$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = -0.2096866$$

Example 2

Suppose a researcher wants to perform a linear regression on the samples he collected. However, the original data was lost, and he has only the access to some summary statistics

$$\bar{x} = 0.2875, \qquad \bar{y} = 0.0075,$$
$$\sum_{i=1}^{n} x_i y_i = 1.5941, \qquad s_x = 0.5460704.$$

Can you still figure out the regression line based on these available information?

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x^2} = 0.7554315$$
$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = -0.2096866$$

To perform linear regressions, the original data is not necessary as long as some key statistics are calculated.