Wednesday Week 14.1 April 3rd

Simple linear regression: Least squares and maximum likelihood

Example: Runoff (problem 3a Spring 2019)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, 25$$

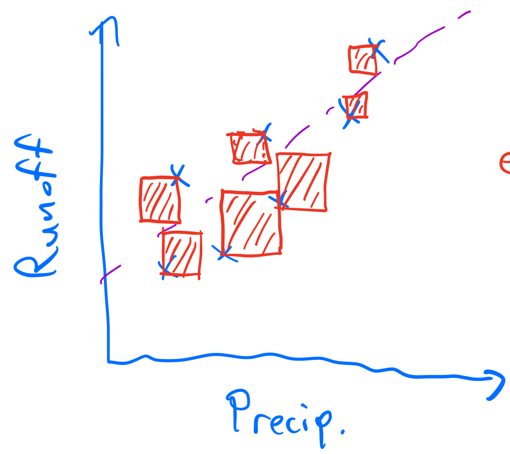
Y: Runoff (mm/year)

x: Precipitation (mm/year)

ε_i: independent and identically distributed (iid) noise, with $\varepsilon \sim N(0, \sigma^2)$

Qu 1: How can 'least squares' be used to find estimators for β_0 and β_1?

Answer: Geometrically:



Minimize sum of squared errors (SSE), i.e. the sum of the shaded areas to the left (as a function of β_0, β_1).

Mathematically:

$$SSE = \sum_{i=1}^{25} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

$$\frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^{25} \left(y_i - (b_0 + b_1 x_i)\right)$$

$$\frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^{25} \left(y_i - (b_0 + b_1 x_i)\right) x_i$$

We can therefore minimize SSEs by solving for b_0, b_1 such that:

$$\frac{\partial SSE}{\partial b_0} = \frac{\partial SSE}{\partial b_1} = 0.$$

As shown in the video lectures,

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{25} x_i y_i}{\sum_{i=1}^{25} x_i^2} = \frac{\sum_{i=1}^{25} (x_i - \bar{x}) y_i}{\sum_{i=1}^{25} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{25} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{25} (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

$\boxed{\text{Qu 2:}}$ What assumptions are necessary to use simple linear regression in this example?

$\boxed{\text{Answer:}}$

Assumptions

- $y_1, \dots, y_n$ realizations of random variables $Y_1, \dots, Y_n$
- $x_1, \dots, x_n$ fixed numbers
- $E[\varepsilon_i] = 0$
- x and Y have a linear relationship
- $\left(Var(\varepsilon_i) = \sigma^2 \right)$
- $\left(\varepsilon_i \text{ are iid} \right)$
- $\left(\varepsilon_i \sim N(0, \sigma^2) \right)$

Note:

- Not all of the above assumptions are always necessary. Throughout this block on simple linear regression, think about when these are used, and, perhaps more importantly, when they are <u>not</u> used.
- Are the assumptions different in least squares compared to maximum likelihood for simple linear regression? How/Which ones?

Example: Hot chocolate sales (problem 4b, Fall 2015)

Value of x_i	Ski conditions
1	bad
2	good
3	very good
4	excellent

Y_i: cups of hot chocolate sold

x_i: ski conditions

$i = 1, \ldots, 20$: day index

ε_i: error, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\boxed{\text{Qu:}}$ Do the modeling assumptions look reasonable from the plot in the slides?

Ans: No, not all of them. $Var(\varepsilon_i)$ increases with x_i (or else $E[Y_i]$ depends on x_i and also i itself). Y_i seems to increase for 'very good' and 'excellent' ski conditions as i increases.

Qu: Calculate $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$.

Ans:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{20} (x_i - \bar{x}) y_i}{\sum_{i=1}^{20} (x_i - \bar{x})^2} = \frac{237.15}{24.95} \approx \boxed{9.51}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 25.65 - \frac{237.15}{24.95} \cdot 2.45$$

$$\approx \boxed{2.36}$$

$$\hat{\sigma}^2 = \frac{1}{20} \sum_{i=1}^{20} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

$$= \frac{18}{20} \cdot \frac{1}{18} \sum_{i=1}^{20} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

$$= \frac{18}{20} \cdot 5.65^2$$

$$\approx \boxed{28.73}$$