Stat 145, Thu 18-Feb-2021 -- Thu 18-Feb-2021
Biostatistics
Spring 2021

Thursday, February 18th 2021

Due::   PS03 due at 11 pm

Thursday, February 18th 2021

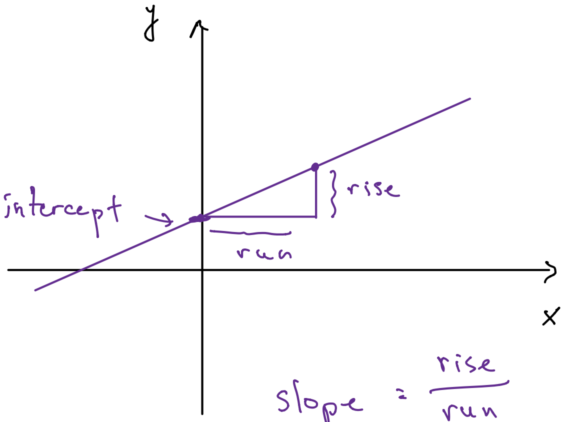
Wk 3, Th
Topic:: least-squares regression

More scatter plotting

- spruce data: Di.change ~ Ht.change
  - add color for Fertilizer
  - lm(Di.change ~ Ht.change, data = spruce)
- draft data: N69 ~ nday

Review features of a line y = intercept + slope * x

- intercept
- slope
  - meaning



slope = $\frac{\text{rise}}{\text{run}}$

gives how much rise corresponds to a change of 1 unit in x-coord.

least-squares regression: \hat y = a + bx

- identify slope as b, intercept as a
- offers a "prediction" to value of y for given x
- observed y vs. fitted/predicted \hat y-value
  - residual = observed - predicted
    - straight-line distance
    - positive if data point is above line, negative if below
- how data is used to choose a, b
  - want to make overall measure of residuals as small as possible
  - might add up residuals and try to make sum small
    - sum r_i does not prove to be effective
    - two alternatives:
      - sum |r_i|
      - sum r_i^2   better setup for calculus to take over and produce

Find a line $\hat{y} = a + bx$

x = expl. var.

a = intercept

b = slope

```
b = r s_y / s_x
a = ybar - b xbar
```

- use app, have groups make guesses

For spruce data frame, best-fit line

$$\overbrace{\text{Di.change}} = -0.5189 + 0.1459\left(\text{Ht.change}\right)$$

Purpose:

To predict values for Di.change at different levels of Ht.change.

In actual data set, one point

$$\left(\text{Ht.change}, \text{Di.change}\right) = \left(45, 5.4156\right) \quad \text{tree 1}$$

$$x_1 = 45 \qquad y_1 = 5.4156 \quad \left(\text{observed value at } x = 45\right)$$

↑ 1st tree

predicted value at $x=45$ →

$$\hat{y}_1 = -0.5189 + 0.1459\left(45\right)$$

$$= 6.0466 \quad \left(\text{what is predicted by model}\right)$$

For Tree #1, there is a positive residual

residuals = observed − predicted

$$r_i = y_i - \hat{y}_i$$

1st residual

$$r_1 = 5.4156 - 6.0466 = -0.631$$

How does software use data to decide the right values for slope (b) and intercept (a)?

Least-squares regression line: Chooses a, b so as to make

$$\sum r_i^2 \quad \text{(Sum of Squared Residuals)}$$

as small as possible. Good choice for use of calculus.

$$\boxed{\begin{aligned} \underline{b} &= r \cdot \frac{s_y}{s_x} \\ \underline{a} &= \bar{y} - b\bar{x} \end{aligned}}$$

For Spruce data

$\bar{x} = 30.93$ $\quad s_x = 15.05$

$\bar{y} = 4.0$ $\quad s_y = 1.788$

$r = 0.902$