```
Stat 145, Fri 17-Sep-2021 -- Fri 17-Sep-2021
Biostatistics
Spring 2021


-------------------------------
Friday, September 17th 2021
-------------------------------
Wk 3, Fr
Topic:: least-squares regression
Read::  Lock5 2.6
HW::    WW Descriptive2 due Tues.

Regression wrap-up
 - R specifics
    lm(responseVar ~ explnatoryVar, data = <dataFrameName>)
```

Try lm(waiting ~ eruptions, data = faithful)

learn $b_0 = 33.47$, $b_1 = 10.73$

lm is for "linear model"

```
    can store the output/results
      myLMResults <- lm(respVar ~ explVar)

    predicted/fitted values are available in your output
      myLMResults$fitted.values

    residuals are also available in that output
      myLMResults$residuals

 - import and use of regression line
    meaning of slope, intercept
    prediction of response from explanatory value
      extrapolation vs. interpolation
```

From faithful example above

$$\text{predicted } \widehat{\text{waiting}} = 33.47 + (10.73)(\text{eruptions})$$

> head(faithful)

shows 1st eruption 3.6 mins followed by waiting time 79 (observed y_1)

The predicted waiting time at 3.6 mins in

$$\hat{y}_1 = 33.47 + (10.73)(3.6) = 72.098$$

The residual for the 1st eruption:

$$\text{residual \#1} = y_1 - \hat{y}_1 = 79 - 72.098 = 6.902.$$

Least-squares regression activity

Connect to our Microsoft Teams course. A copy of this document appears in the **Files** tab of the "Class activity log" channel. Use the **Posts** tab of the same channel to ask questions, or to browse and comment in the threads already started by others.

Working in groups of 2–3, complete the following tasks and answer the questions. We will go over answers in class, so record your responses to discuss later. This can be within R Markdown, if you like, but it is not required.

1. Display the first few lines of the data frame called `cars`. This is a built-in data set; you will not need to import it.
2. Decide on a quantitative variable to take role of *explanatory variable*. – speed (resp.: dist)
3. Working with the `cars` data frame, determine,
   (a) the mean and standard deviation for each quantitative variable,
   (b) the correlation between quantitative variables

	mean	sd
dist	42.98 ($\bar{y}$)	25.77 (s_y)
speed	15.4 ($\bar{x}$)	5.29 (s_x)

4. Use the formulas
$$b_1 = r\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}$$
to calculate slope b_1 and intercept b_0. Verify that the `lm()` command produces these same numbers.
5. State a useful way to think about/interpret your slope. – dist to stop increases approx. 3.93 for an increase of 1mph in speed.
6. Produce a scatter plot of the data in `cars`, along with regression line
7. What are the values of the variables for the point with the largest `dist`? Find the residual for that point. Filter that point out of the data, and use `lm()` to recompute the slope b_1 and intercept b_0. Did these seem to change much with the point omitted?
8. Above you have calculated each of
   - mean and standard deviation for both variables,
   - correlation,
   - slope, intercept of regression line.

   Do any of these change when the variables exchange roles (the one you had considered your explanatory variable becomes the response, and vice versa)? Which ones?
9. Given the points
$$(31, 37),\ (22, 56),\ (15, 68),\ (25, 60),\ (35, 41),$$
make a scatterplot, and find the correlation and equation of the least-squares regression line. Is there a positive association between values? Negative association? Is the association reasonably linear? Is predicting the y-value at $x = 42$ an example of interpolation or extrapolation?

3(b)

> cor(dist ~ speed, data = cars)

0.807 = r

4. $b_1 = r \cdot \frac{s_y}{s_x} = (0.807)\frac{25.77}{5.29} \approx 3.93$

$b_0 = \bar{y} - b_1\bar{x} = 42.98 - (3.93)(15.4) = $ y-intcpt.

lm(dist ~ speed, data = cars)

6. gf_point(dist ~ speed, data = cars) %>% gf_lm(type = "lm")

7. lm(dist ~ speed, data = filter(cars, dist < 115))

residual for our filtered point

#49 is the case

its dist was 120 : y

its predicted dist. is: $3.93(24) - 17.579 = 76.74$ $(\hat{y})$

residual = 120 $= y - \hat{y} = 120 - 76.74 = 43.26$.

9.

x	y
31	37
22	56
15	68
25	60
35	41

gf_point(c(37,56,68,60,41) ~ c(31,22,15,25,35))

cor
lm