Stat 145, Fri 19-Feb-2021 -- Fri 19-Feb-2021
Biostatistics
Spring 2021

Friday, February 19th 2021

Wk 3, Fr
Topic:: least-squares regression

Regression wrap-up

- sensitivity to outliers

  illustrate using "idea of regression" app

- R specifics

  ```
  lm(respVar ~ explVar)
  ```

  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
  ```

- extrapolation vs. interpolation

An activity on least-squares regression

Preparation: To get going, please

- I do not expect you to use a group-specific Etherpad for today's work. But, for the purpose of writing a note to the instructor when you are seeking help, open a browser tab and point it at

  `https://pad.disroot.org/p/s145-19feb2021`

  You will, instead, be building an R Markdown document to record your work and responses to questions. Choose a scribe/team member who will have the primary role of producing that R Markdown file in her/his account.
- Our Microsoft Teams class has channels marked Grp A, Grp B, . . . Grp J. Select one of these based on your group number: Group 01 should use the Grp A channel, Group 08 should use the Grp H channel. Enter that channel to meet with others from your group, turn on cameras, and have the scribe share his/her screen.
- Have the scribe open an R Markdown file, probably from a template as in earlier instances. Give it the title: "Group XX's regression", using your group number in place of XX. Insert the names of all group members as authors

Tasks: Complete the following tasks and answer the questions. Record your work and answers in the R Markdown file. If/when you seek help from the instructor, write a note in the Etherpad; have another group member, not the scribe, handle writing this note.

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*.
3. Working with the `cars` data frame, determine,
   (a) the mean and standard deviation for each quantitative variable,
   (b) the correlation between quantitative variables
4. Use the formulas

   $$b = r\frac{s_y}{s_x}, \qquad a = \overline{y} - b\overline{x}$$

   to calculate slope b and intercept a. Verify that the `lm()` command produces these same numbers.
5. State a useful way to think about/interpret your slope.
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 and intercept a. 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?

Submitting group work: Your group should complete these tasks/questions before Monday. The final submission will be the scribe's work, to be done by Monday at noon.

- Knit your group's R Markdown document to a .pdf file.
- Download that file to your local computer.
- Send an email to `scofield@calvin.edu` with *subject line* "group XX regression", attaching the .pdf file