DS5110 Homework 5

Kylie Ariel Bemis

8 November 2022

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly three files:

- R Markdown (.Rmd)
- PDF report (.pdf)
- Miniposter (original)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–2 use the "miniposters" from Canvas Discussions.

Problem 1

Choose one of the "miniposters" created by your fellow classmates and posted on Canvas Discussions for Homework 2. Cite both the name of the student whose miniposter you choose and the original source of the dataset used in the miniposter. Include the *original* miniposter as one of your file uploads when submitting this assignment.

Download and import that dataset into R, put it into a tidy format (if necessary), and print the first several rows of the dataset.

Problem 2

To the best of your ability, reproduce the figures from the miniposter you chose in Problem 1. The data content and visual representation should be as similar as possible. Color schemes and themes do not have to be exactly the same.

You may contact the author of the original miniposter; if you do, cite and describe any information you receive from them.

(If you are contacted for information on reproducing figures from your own miniposter, you may provide it, but you are not obligated to respond or provide any help.)

Part B

Problems 3–5 use the `PimaIndiansDiabetes2` dataset from the `mlbench` package. You do not need to partition the dataset for any of the problems in Part B.

Problem 3

We would like to know if there is difference in blood pressure between people with diabetes and people without diabetes.

First remove missing values from the data using `na.omit()`. Then fit a model for blood pressue using diabetes as the only explanatory variable. Perform model diagnostics to check for any violations of model assumptions.

Visualize the relationship between blood pressure and diabetes. State the null and alternative hypotheses, choose an alpha value, and state the p-value and your conclusions.

Problem 4

We would like to consider `glucose`, `insulin`, `triceps`, `mass`, and `age` as possible covariates. Plot them each against `pressure` for consideration in the model as explanatory variables. Which variables would you consider including?

Use AIC to select the best model that also includes `diabetes` as a factor. Show your steps and reasoning, and then state the final model.

Hint: *AIC can be calculated using `AIC()` or `extractAIC()`; note that these two functions use different additive constants when calculating the likelihood, and so give differing values for AIC. However, they should lead to the same conclusions. You may find the `step()` function useful as well.*

Problem 5

Use your model from Problem 4 to test the same hypotheses as Problem 3.

State the null and alternative hypotheses, choose an alpha value, and state the p-value and your conclusions.

Are your results the same or different? How do you explain this?