Week 1: Exercises

ETC3580

Contents

Exercise 1A

Using the `swiss` data, with `Fertility` as the response variable, complete the following tasks.

1. Make a tibble for the `swiss` data set.
2. Produce pairwise scatterplots and boxplots of all variables, looking for any outliers or unusual observations.
3. We are interested in what characteristics of a canton affect the fertility rate. Fit a model using all available predictors, and visualize the resulting fit using `visreg`.
4. Use F-tests to test the significance of each variable in the model.

Exercise 1B

We will use the GSS data set to look at a model of job prestige as a function of various demographic variables. Information about all variables is available in the GSS Codebook.

The following code will fit a model for job prestige with year, age, sex and race as predictors. Make sure you understand what each line is doing.

```
library(tidyverse)
load("gss.RData")

gss %>%
  select(prestige, year, age, sex, race) %>%
  mutate(prestige = as.numeric(substr(prestige,1,2)),
         age = as.numeric(substr(age, 1, 2))) ->
  mydata

summary(mydata)

GGally::ggpairs(mydata)

fit <- lm(prestige ~ year + age + sex + race, data=mydata)
summary(fit)
```

1. Why are there two clusters of observations?

2. Add other demographic variables that you think might be relevant.
3. Try including an interaction of age with sex, and age with race.
4. Include interactions where appropriate, and visualize the resulting models.
5. Find the best model you can (by minimizing the AIC)
6. Can you find a model with R^2 more than 10%?
7. Does it make any difference if you only use data after 2000? [Use the `filter()` function from the `dplyr` package (part of the tidyverse).]
8. What is the single most important predictor you can find? (based on AIC).