Chapter 7: Linear models with more than one predictor

1. This exercise is based on data from an experiment investigating how caffeine affects performance on a verbal reasoning test [1]. In the study, university students were given either a low (2 mg/kg) or high (5 mg/kg) dose of caffeine before completing a test similar to the GRE (Graduate Record Examination – a standardized test commonly used for graduate school admissions in the United States). Students were also classified as either *introverted* or *extroverted*, based on a personality measure.

   You can load the dataset using:

```
caffeine_data <- read.csv("caffeine_data.csv")
```

   Make sure the file is saved in your working directory.

   The dataset includes the following variables:

   - `score`: the participant's GRE-style test score
   - `caffeine`: caffeine dose, either `"low"` or `"high"`
   - `personality`: participant personality type, either `"introvert"` or `"extrovert"`.

   Answer the following:

   - Fit a linear model that predicts test scores from caffeine intake alone. What does this model tell you about the effect of caffeine on performance?
   - Now fit a second model that includes both caffeine intake and personality as predictors. What does this model suggest about the combined effect of caffeine and personality on test performance? How does this differ from the simpler model in part (i)?
   - Provide a suitable visualisation to illustrate the effects of caffeine and personality on test scores. How well does your model fit the data?
   - Refit the model including an *interaction* between caffeine and personality. Which model (with or without the interaction) fits best? Is the interaction term necessary? Justify your answer based on model comparison and your interpretation of the interaction.
   - Select the model you consider most appropriate for describing the data. Report the model results in APA style.

2. In this exercise, you will analyse historical data from mid-19th century British counties, based on a study by prison chaplain John Clay[2]. The original paper explored how societal features –such as the presence of pubs, schools, and churches –might relate to crime rates. The dataset includes information about

[1] Gilliland, K. (1980). 'The interactive effect of introversion-extraversion with caffeine-induced arousal on verbal performance.' *Journal of Research in Personality,* 14, 482–492.

[2] John Clay (1857). 'On the Relation Between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-House.' *Journal of the Statistical Society of London*, 20 (1), 22–32.

the number of beerhouses, school attendance, church attendance, and recorded criminals across counties.

You can load the dataset using `read.csv("beerhall.csv")`. The relevant variables are:

- `county`: name of the county
- `region`: region name
- `criminals_per100k`: number of recorded criminals per 100,000 population
- `beerhouses_per100k`: number of ale/beer houses (pubs) per 100,000 population
- `attSchool_per10k`: school attendance per 10,000 population
- `attChurch_per2k`: church attendance per 2,000 population (contains missing values).

In the original article, Clay made several strong claims. You will evaluate two of them using multiple regression, focusing on the relationship between beerhouses, school attendance, and crime:

Claim 1: 'It is manifest that the amount of crime in a county mainly depends on the number of low-drinking houses which are suffered to infest it.' *Claim 2:* 'Our present system of popular education is of little or no efficacy in saving the industrial classes from the moral dangers created by those drinking houses.'

For this analysis, use `criminals_per100k` as the outcome variable, and test whether it is predicted by `beerhouses_per100k` and `attSchool_per10k`.

- Fit a multiple regression model predicting `criminals_per100k` from `beerhouses_per100k` and `attSchool_per10k`.
- Examine the model output.
  - Is there evidence that counties with more beerhouses tend to have more crime?
  - Is school attendance associated with lower crime, on average?

  What do the coefficients for each predictor suggest?
- Create two scatter plots that show the relationships between beerhouses and crime; and school attendance and crime.
- Now fit another model including also an interaction term between `beerhouses_per100k` and `attSchool_per10k`. It is recommended that you centre the variables, so as to make the coefficient easier to interpret. Is the *interaction* between beerhouses and school attendance statistically significant? What does it tell you about how the relationship between beerhouses and crime *changes depending on school attendance*?
- Now fit a second model including an interaction term between `beerhouses_per100k` and `attSchool_per10k`. To make the coefficients easier to interpret, it's helpful to centre the predictors (e.g., subtract the mean of each variable). Examine the model output.

  - Is the interaction statistically significant? Interpret what this means: how does the relationship between beerhouses and crime vary depending on the level of school attendance?
  - Translate the interaction into practical terms. For example: In counties with low school attendance, does a higher number of beerhouses predict a stronger increase in crime than in counties with high school attendance?
  - Visualise the interaction using a plot. One approach is to split school attendance into three thirds (e.g., low, medium, high attendance) and plot separate regression lines for each. (Refer to the approach used in the chapter for the AI Faces dataset.)
- Based on your results, do the data support Clay’s second claim – that education does little to mitigate the dangers posed by beerhouses?