STA 310: Homework 1

Amaris

Instructions

- Write all narrative using full sentences. Write all interpretations and conclusions in the context of the data.
- Be sure all analysis code is displayed in the rendered pdf.
- If you are fitting a model, display the model output in a neatly formatted table. (The `tidy` and `kable` functions can help!)
- If you are creating a plot, use clear and informative labels and titles.
- Render and back up your work reguarly, such as using Github.
- When you're done, we should be able to render the final version of the Rmd document to fully reproduce your pdf.
- Upload your pdf to Gradescope. Upload your Rmd, pdf (and any data) to Canvas.

Exercises

Exercises 1 - 4 are adapted from exercises in Section 1.8 of @roback2021beyond.

Exercise 1

Consider the following scenario:

> Researchers record the number of cricket chirps per minute and temperature during that time. They use linear regression to investigate whether the number of chirps varies with temperature.

a. **Identify the response and predictor variable:** The response variable is the number of cricket chirps per minute, as this is the outcome being measured and analyzed. The predictor variable is temperature, since the researchers are interested in whether changes in temperature are associated with changes in the number of cricket chirps.

b. **Write the complete specification of the statistical model:**

   Let Y_i denote the number of cricket chirps per minute observed during minute i, and let X_i denote the temperature during that same minute.

   The relationship between temperature and chirping rate is modeled using simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

   The error terms are assumed to be independent and identically distributed with a normal distribution:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

c. **Write the assumptions for linear regression in the context of the problem.**

- **Linearity:** The expected number of cricket chirps per minute is a linear function of temperature. This means that, on average, changes in temperature are associated with proportional changes in the chirping rate, and this relationship can be adequately described by a straight line.
- **Independence:** The observations are independent of one another. In this context, this implies that the number of chirps recorded during one minute does not influence the number of chirps recorded during any other minute.
- **Normality:** For a fixed temperature, the deviations of the observed number of chirps per minute from the regression line are normally distributed. That is, the error terms in the model follow a normal distribution centered at zero.
- **Equal Variance:** The variability in the number of cricket chirps per minute is constant across all temperatures. This means that the spread of chirp counts around the regression line does not systematically increase or decrease as temperature increase or decrease.

Exercise 2

Consider the following scenario:

> A randomized clinical trial investigated postnatal depression and the use of an estrogen patch. Patients were randomly assigned to either use the patch or not. Depression scores were recorded on 6 different visits.

a. Identify the response and predictor variables.

b. Identify which model assumption(s) are violated. Briefly explain your choice.

Exercise 3

Use the Kentucky Derby case study in Chapter 1 of *Beyond Multiple Linear Regression.*

a. Consider Equation (1.3) in Section 1.6.3. Show why we have to be sure to say "holding year constant", "after adjusting for year", or an equivalent statement, when interpreting β_2.
b. Briefly explain why there is no error (random variation) term ϵ_i in Equation (1.4) in Section 1.6.6?

Exercise 4

The data set `kingCountyHouses.csv` in the `data` folder contains data on over 20,000 houses sold in King County, Washington (@kingcounty).

We will use the following variables:

- `price` = selling price of the house
- `sqft` = interior square footage

See Section 1.8 of Beyond Multiple Linear Regression for the full list of variables.

a. Fit a linear regression model with `price` as the response variable and `sqft` as the predictor variable (Model 1). Interpret the slope coefficient in terms of the expected change in price when `sqft` increases by 100.

b. Fit Model 2, where `logprice` (the natural log of price) is now the response variable and `sqft` is still the predictor variable. How is the `logprice` expected to change when `sqft` increases by 100?

c. Recall that $log(a) - log(b) = log(\frac{a}{b})$. Use this to derive how the `price` is expected to change when `sqft` increases by 100 based on Model 2.

d. Fit Model 3, where `price` and `logsqft` (the natural log of sqft) are the response and predictor variables, respectively. How does the price expected to change when sqft increases by 10%? *As a hint, this is the same as multiplying sqft by 1.10.*

Click here for notes on interpreting model effects for log-transformed response and/or predictor variables.

Exercise 5

The goal of this analysis is to use characteristics of 593 colleges and universities in the United States to understand variability in the early career pay, defined as the median salary for alumni with 0 - 5 years of experience. The data was obtained from TidyTuesday College tuition, diversity, and pay, and was originaly collected from the PayScale College Salary Report.

The data set is located in `college-data.csv` in the `data` folder. We will focus on the following variables:

variable	class	description
name	character	Name of school
state_name	character	state name
type	character	Public or private
early_career_pay	double	Median salary for alumni with 0 - 5 years experience (in US dollars)
stem_percent	double	Percent of degrees awarded in science, technology, engineering, or math subjects
out_of_state_total	double	Total cost for in-state residents in USD (sum of room & board + out of state tuition)

a. Visualize the distribution of the response variable `early_career_pay`. Write 1 - 2 observations from the plot.
b. Visualize the relationship between (i) `early_career_pay` and `type` and (ii) `early_career_pay` and `stem_percent`. Write an observation from each plot.
c. Below is the specification of the statistical model for this analysis. Fit the model and neatly display the results using 3 digits. Display the 95% confidence interval for the coefficients.

$$\begin{aligned} early\_career\_pay_i = \beta_0 + \beta_1\ out\_of\_state\_total_i + \beta_2\ type \quad (1) \\ + \beta_3\ stem\_percent_i + \beta_4\ type * stem\_percent_i \quad (2) \\ + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \quad (3) \end{aligned}$$

d. How many degrees of freedom are there in the estimate of the regression standard error σ?
e. What is the 95% confidence interval for the amount in which the intercept for public institutions differs from private institutions?

Exercise 6

Use the analysis from the previous exercise to write a paragraph (~ 4 - 5 sentences) describing the differences in early career pay based on the institution characteristics. *The summary should be consistent with the results from the previous exercise, comprehensive, answers the primary analysis question, and tells a cohesive story (e.g., a list of interpretations will not receive full credit).*

Grading

Total	**50**
Ex 1	8
Ex 2	4
Ex 3	7

Total	**50**
Ex 4	12
Ex 5	12
Ex 6	4
Workflow & formatting	3

The "Workflow & formatting" grade is to based on the organization of the assignment write up along with the reproducible workflow. This includes having an organized write up with neat and readable headers, code, and narrative, including properly rendered mathematical notation. It also includes having a reproducible Rmd/Quarto document that can be rendered to reproduce the submitted PDF.