Homework 03: Coffee grades

Inference for simple linear regression using mathematical models

> **!** Important
>
> Due: Friday, September 20, 11:59pm

Introduction

In this weeks homework you will analyze data from over 1,000 different coffees to explore the relationship between a coffee's aroma and its flavor grade.

Learning goals

By the end of the homework you will...

- be able to use mathematical models to conduct inference for the slope
- be able to assess conditions for simple linear regression

Getting started

- Go to RStudio and login with your College of Idaho Email and Password.
- Make a subfolder in your hw directory to store this homework.
- Log into Canvas, navigate to Homework 3 and upload the `hw-03.qmd` and `coffee-grades.csv` file into the folder your just made.

Packages

The following packages are used in the homework.

```
library(tidyverse)
library(ggformula)
library(broom)
library(knitr)
```

Data: Coffee grades

The dataset for this homework comes from the Coffee Quality Database and was obtained from the #TidyTuesday GitHub repo. It includes information about the origin, producer, measures of various characteristics, and the quality measure for over 1,000 coffees. The coffees can be reasonably be treated as a random sample.

This homework will focus on the following variables:

- `aroma`: Aroma grade, 0 (worst aroma) - 10 (best aroma)
- `flavor`: Flavor grade, 0 (worst flavor) - 10 (best flavor)

Click here for the definitions of all variables in the data set. Click here for more details about how these measures are obtained.

```
coffee <- read_csv("coffee-grades.csv")
```

Exercises

! Important

Make sure to do the following as you complete the assignment:

- Write all code and narrative in your Quarto file. I should be able to read all your code in the rendered PDF.
- Write all narrative in complete sentences.
- Use informative axis titles and labels on all graphs.
- You should periodically **render** your Quarto document to produce the updated PDF to make sure your output is as you expect it to be.

Goal: The goal of this analysis is to use linear regression to understand variability in coffee flavor grades based on the aroma grade.

Exercise 1

Visualize the relationship between the aroma and flavor grades. Write two observations from the plot.

Exercise 2

Fit the linear model using aroma grade to understand variability in the flavor grade. Neatly display the model using three digits and include the **95%** confidence interval for the model coefficients in the output.

Exercise 3

- Interpret the slope in the context of the data.
- Assume you are a coffee drinker. Would you drink a coffee represented by the intercept? Why or why not?

Exercise 4

Input the model name in the code below to calculate the regression standard error, $\hat{\sigma}_\epsilon$. State the definition of this value in the context of the data.

```
glance(____)$sigma
```

Exercise 5

Do the data provide evidence of a statistically significant linear relationship between aroma and flavor grades? Conduct a hypothesis test using mathematical models to answer this question. In your response

- State the null and alternative hypotheses in words and in mathematical notation.
- What is the test statistic? State what the test statistic means in the context of this problem.
- What distribution was used to calculate the p-value? Be specific.
- State the conclusion in the context of the data using a threshold of $\alpha = 0.05$ to make your decision.

Exercise 6

- What is the critical value (i.e. t^*) used to calculate the 95% confidence interval displayed in Exercise 2 using the `qt` function? Show the code and output used to get your response.
- Is the confidence interval consistent with the conclusions from the hypothesis test? Briefly explain why or why not.

Exercise 7

- Calculate the 95% confidence interval for the mean flavor grade for coffees with aroma grade of 7.5. Interpret this value in the context of the data.
- One coffee produced by the Ethiopia Commodity Exchange has an aroma of 7.5. Calculate the the 95% prediction interval for the flavor grade for this coffee. Interpret this value in the context of the data.
- How do the predicted values compare? How do the intervals compare? If there are differences in the predictions and/or intervals, briefly explain why.

Exercise 8

We'd like to check the model conditions to assess the reliability of the inferential results. To do so, we will create a data frame called `coffee_aug` that includes the residuals and predicted values from the model. Input the name of your model in the code below.

```
#|eval: false
coffee_aug <- augment(_____)
```

Make a scatterplot of the residuals (`.resid`) vs. fitted values (`.fitted`). Use `gf_hline()` to add a horizontal dotted line at $residuals = 0$.

> **Note**
>
> The **linearity condition** is satisfied if there is random scatter of the residuals (no distinguishable pattern or structure) in the plot of residuals vs. fitted values.
>
> The **constant variance** condition is satisfied if the vertical spread of the residuals is relatively across the plot.

- Is the linearity condition satisfied? Briefly explain why or why not.
- Is the constant variance condition satisfied? Briefly explain why or why not.

Exercise 9

> **Note**
>
> The **normality** condition is satisfied if the distribution of the residuals is approximately normal. This condition can be relaxed if the sample size is sufficiently large ($n > 30$).

Make a histogram or density plot of the residuals (`.resid`). Is the normality condition satisfied? Briefly explain why or why not.

Exercise 10

> **Note**
>
> The **independence** condition means that knowing one residual will not provide information about another. We often check this by assessing whether the observations are independent based on what we know about the subject matter and how the data were collected.

Is the independence condition satisfied? Briefly explain why or why not.

Submission

> **Warning**
>
> Before you wrap up the assignment, make sure you have rendered your document and that the PDF appears as you want it to.

To submit your assignment, upload the .qmd and PDF files to Canvas.

Grading (50 pts)

Component	Points
Ex 1	4
Ex 2	3
Ex 3	4
Ex 4	3
Ex 5	8
Ex 6	5

Component	Points
Ex 7	8
Ex 8	6
Ex 9	4
Ex 10	2
Workflow & formatting	3[1]

[1]The "Workflow & formatting" grade is to assess the reproducible workflow and document format. This includes having at least 3 informative commit messages, a neatly organized document with readable code and your name and the date in the YAML.