HW 01: Simple linear regression

Educaton & median income in US counties

> **!** Important
>
> This assignment is due on **Tuesday, January 28 at 11:59pm.** To be considered on time, the following must be done by the due date:
>
> - Final `.qmd` and `.pdf` files pushed to your GitHub repo
> - Final `.pdf` file submitted on Gradescope

Introduction

In this assignment, you will use simple linear regression to examine the association between the percent of adults with a bachelor's degree and the median household income for counties in the United States.

Learning goals

In this assignment, you will...

- Fit and interpret simple linear regression models.
- Construct and interpret bootstrap confidence intervals for the population slope, β_1.
- Create and interpret spatial data visualizations using R.
- Continue developing a workflow for reproducible data analysis.

Getting started

- Go to the sta210-sp25 organization on GitHub. Click on the repo with the prefix **hw-01**. It contains the starter documents you need to complete the lab.
- Clone the repo and start a new project in RStudio. See the Lab 00 for details on cloning a repo and starting a new project in R.

The following packages will be used in this assignment:

```
library(tidyverse)  # for data wrangling
library(tidymodels) # for modeling and inference
library(knitr)      # to neatly format tables
library(scales)     # to format visualizations

# load other packages as needed
```

Data: US Counties

The data are from the `county_2019` data frame in the usdata R package. These data were originally collected in the 2019 American Community Survey (ACS), an annual survey conducted by the United States Census Bureau that collects demographics and other information from a sample of households in the United States. The data frame `county_2019` contains county-level statistics from the ACS.

The data for this analysis are available in the file `us-counties-sample.csv` in the `data` folder of your repo. It contains a random sample of 600 counties in the United States. This is about 19% of the counties in the United States.

This analysis focuses on the following variables:

- `bachelors`: Percent of population 25 years old and older that earned a Bachelor's degree or higher
- `median_household_income`: Median household income in US dollars
- `household_has_computer`: Percent of households that have desktop or laptop computer

Click here for the full codebook for the `county_2019` dataset.

You will use two other data sets `county-map-sample.csv` and `county-map-all.csv` to create spatial visualizations of the ACS variables. Use the code below to load all of the data sets.

```
county_data_sample <- read_csv("data/us-counties-sample.csv")
map_data_sample <-  read_csv("data/county-map-sample.csv")
map_data_all <- read_csv("data/county-map-all.csv")
```

Exercises

There is a lot of public interest in understanding the impact of graduating college, i.e., obtaining a bachelor's degree, on one's future career and lifetime earnings. The common convention is that individuals who have earned a bachelor's degree (or higher) will earn more income over the course of a lifetime than an individual who does not have such a degree.

We will explore this at a county-level and examine the association between the percent of adults 25 years old + with a Bachelor's degree median household income. Specifically we'd like to answer questions such as, "do counties that have a higher percentage of college graduates have higher median household incomes, on average, compared to counties with a lower percentage of college graduates?".

Instructions

Type your responses to each question in your Quarto document. Write all narrative using complete sentences and include informative axis labels and titles on visualizations. Use a reproducible workflow by periodically rendering the Quarto document, writing an informative commit message, and pushing the updated `.qmd` and `.pdf` files to GitHub.

Part 1: Exploratory data analysis

Exercise 1

Create a histogram of the distribution of the predictor variable `bachelors`and calculate appropriate summary statistics. Use the visualization and summary statistics to describe the distribution. Include an informative title and axis labels on the plot.

Exercise 2

Let's view the data in another way. Use the code below to make a map of the United States with the color of the counties filled in based on the percent of residents 25 years old and older who have a Bachelor's degree. Fill in title and axis labels.

Then use the plot answer the following:

- What are 2 observations you have from the map?
- What is a feature that is apparent in the map that wasn't as easily apparent from the histogram in the previous exercise? What is a feature that is apparent in the histogram that is not as easily apparent from the map?

```
#county_map_data <- left_join(county_data_sample, map_data_sample)

county_map_data <- county_data_sample

ggplot(data = map_data_all) +
  geom_polygon(aes(x = long, y = lat, group = group),
    fill = "lightgray", color = "white"
    ) +
  geom_polygon(data = county_map_data, aes(x = long, y = lat, group = group,
    fill = bachelors)
    ) +
  labs(
    x = "Longitude",
    y = "Latitude",
    fill = "_____",
    title = "_____"
  ) +
  scale_fill_viridis_c(labels = label_percent(scale = 1)) +
  coord_quickmap()
```

Exercise 3

Create a visualization of the relationship between `bachelors` and `median_household_income` and calculate the correlation. Use the visualization and correlation to describe the relationship between the two variables.

This is a good place to render, commit, and push changes to your hw-01 repo on GitHub. Write an informative commit message (e.g. "Completed exercises 1 - 3"), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty.

Part 2: Modeling

Exercise 4

We will use a linear regression model to describe the relationship between `bachelors` and `median_household_income`.

Write the form of the statistical (theoretical) model we will use for this task using mathematical notation. Use variable names (`bachelors` and `median_household_income`) in the equation for your model[1].

$$\text{median\_household\_income}$$

Exercise 5

- Fit the regression line corresponding to the statistical model in the previous exercise. Neatly display the model output using 3 digits.
- Write the equation of the fitted model using mathematical notation. Use variable names (`bachelors` and `median_household_income`) in the equation.

Exercise 6

- Interpret the slope. The interpretation should be written in a way that is meaningful in the context of the data.
- Is it useful to interpret the intercept for this data? If so, write the interpretation in the context of the data. Otherwise, briefly explain why not.

Now is a good time to render your document again if you haven't done so recently and commit (with a meaningful commit message) and push all updates.

Part 3: Inference for the U.S.

We want to use the data from these 600 randomly selected counties to draw conclusions about the relationship between the percent of adults age 25 and older with a bachelor's degree and median household income for the over 3,000 counties in the United States.

Exercise 7

- What is the population of interest? What is the sample?
- Is it reasonable to treat the sample in this analysis as representative of the population? Briefly explain why or why not.

[1] Click here for a guide on writing mathematical symbols using LaTex.

Exercise 8

Conduct a hypothesis test for the slope to assess whether there is sufficient evidence of a linear relationship between the percent of adults age 25 and older with a bachelor's degree and the median household income in a county. Use a randomization (permutation) test. In your response:

- State the null and alternative hypotheses in words and mathematical notation
- Show all relevant code and output used to conduct the test. Use `set.seed(2023)` and 1000 iterations to construct the appropriate distribution.
- State the conclusion in the context of the data.

Exercise 9

Next, construct a 95% confidence interval for the slope using bootstrapping with `set.seed(2023)` and 1000 iterations.

- Show all relevant code and output used to calculate the interval.
- Interpret the confidence interval in the context of the data.
- Is the confidence interval consistent with the results of the test from the previous exercise? Briefly explain why or why not.

Now is a good time to render your document again if you haven't done so recently and commit (with a meaningful commit message) and push all updates.

Reproducibility

Exercise 10

You are asked to use a reproducible workflow for all of your work in the class, and the goal of this question to is better understand potential real-world implications of doing (or not) doing so. Below are some real-life examples in which having a non-reproducible workflow resulted in errors that impacted research and public records.

Table 1: Source: Ostblom and Timbers (2022)

Reproducibility error	Consequence	Source(s)
Limitations in Excel data formats	Loss of 16,000 COVID case records in the UK	(Kelion 2020)

Reproducibility error	Consequence	Source(s)
Automatic formatting in Excel	Important genes disregarded in scientific studies	(Ziemann, Eren, and El-Osta 2016)
Deletion of a cell caused rows to shift	Mix-up of which patient group received the treatment	(Wallensteen et al. 2018)
Using binary instead of explanatory labels	Mix-up of the intervention with the control group	(Aboumatar and Wise 2019)
Using the same notation for missing data and zero values	Paper retraction	(Whitehouse et al. 2021)
Incorrectly copying data in a spreadsheet	Delay in the opening of a hospital	(Picken 2020)

Choose one of the scenarios from the table and read the linked article discussing what went wrong. Then,

- Briefly describe what went wrong, i.e., what part of the process of was not reproducible and what error or impact that had.
- Then, describe how the researchers could make the process reproducible.

Now is a good time to render your document again if you haven't done so recently and commit (with a meaningful commit message) and push all updates.

Submission

> **Warning**
>
> Before you wrap up the assignment, make sure all documents are updated on your GitHub repo. We will be checking these to make sure you have been practicing how to commit and push changes.
> Remember – you must turn in a PDF file to the Gradescope page before the submission deadline for full credit.

To submit your assignment:

- Go to http://www.gradescope.com and click *Log in* in the top right corner.
- Click *School Credentials* → *Duke NetID* and log in using your NetID credentials.

- Click on your *STA 210* course.
- Click on the assignment, and you'll be prompted to submit it.
- Mark the pages associated with each exercise. All of the pages of your lab should be associated with at least one question (i.e., should be "checked").
- Select the first page of your PDF submission to be associated with the *"Workflow & formatting"* section.

Grading (50 points)

Component	Points
Ex 1	5
Ex 2	5
Ex 3	4
Ex 4	3
Ex 5	4
Ex 6	4
Ex 7	3
Ex 8	7
Ex 9	7
Ex 10	5
Workflow & formatting	3[2]

Ostblom, Joel, and Tiffany Timbers. 2022. "Opinionated Practices for Teaching Reproducibility: Motivation, Guided Instruction and Practice." *Journal of Statistics and Data Science Education* 30 (3): 241–50. https://doi.org/10.1080/26939169.2022.2074922.

[2]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.