Week 1 Lab

(add your name)

4/1/2025

**Instruction to work on this sheet: In each code chunk, you can find some code being taken out purposely and left with three dots instead. Your task is to first uncomment the code line, then replace the three dots with the code you would write.**

## Topic: Set-up and Review: Multiple regression with continuous outcome variable

### Data

#loading libraries  
library(tidyverse)  
library(modelsummary)

#loading data and selecting key variables  
seda <- read.csv("https://daviddliebowitz.github.io/EDUC643\_23W/data/seda.csv") %>%  
 select(sedalea, math,  
 totenrl, perfl,   
 locale)

# data management: noticing the variable locale is coded as chr but should be changed to factor  
  
seda$locale <- factor(seda$locale)  
  
table(seda$locale)

##   
## Rural Suburb Town Urban   
## 5976 3360 2240 663

# Based on the table, decide which locale should be the reference category and change the code  
  
# seda$locale <- relevel(seda$locale, ref = "...")

The dataset we’ll be using in this lab is drawn from the Stanford Education Data Archive (SEDA) version 4.1. This is the same dataset you used for your final project. Take a look at this link here to remind yourself of the included variables

[here](https://daviddliebowitz.github.io/EDUC643_23W/assignments/assignment_6.html).

Brief variable descriptions we use here:

* sedalea, NCES district ID
* math, standardized math achievement score
* perfl, district-level proportion of students eligible for free- or reduced-price school lunch
* locale, Census-based location of schools attended by majority/plurality of students (Urban, Suburb, Town, Rural)
* totenrl - total district enrollment grades 3-8

**Table 1. Analytic sample summary statistics**

datasummary\_skim(data = seda[,3:4])

|  | Unique (#) | Missing (%) | Mean | SD | Min | Median | Max |
| --- | --- | --- | --- | --- | --- | --- | --- |
| totenrl | 9192 | 0 | 1813.6 | 7066.6 | 9.1 | 586.6 | 439767.3 |
| perfl | 12233 | 0 | 0.4 | 0.2 | 0.0 | 0.4 | 1.0 |

datasummary\_skim(data = seda, type = "categorical")

| locale | N | % |
| --- | --- | --- |
| Rural | 5976 | 48.8 |
| Suburb | 3360 | 27.5 |
| Town | 2240 | 18.3 |
| Urban | 663 | 5.4 |

1. Write a short paragraph describing the sample, key variables, and the summary statistics displayed in Table 1.

*Answer here*

### Research question

# RQ1: What is the relationship between percentage of FL students (perfl) and math achievement in Oregon districts?

*Outcome variable: ….*

*Predictor variable: ….*

**Figure 1. Visualization of the relationship between math achievement and percentage of FL students**

## Make a plot, remember to put the predictor on the x-axis and outcome variable on the y-axis  
# ggplot(seda,  
# aes(x = ..., y = ...)) +  
# geom\_point()+  
# geom\_smooth(method = 'lm') +  
# labs(x = "...",  
# y = "...") +  
# theme\_minimal(base\_size = 14)

1. Describe the strength of the relationship between percentage of FL students (perfl) and math achievement in Oregon districts? Use the terms strength, magnitude, direction, outliers, shape of curve (linear or othewise).

*Answer here*

# Use format: lm(outcome ~ predictor)  
#fit1 <- lm(... ~ ..., data = seda)  
#modelsummary(fit1)

1. Report the magnitude of the relationship from a simple regression model in one-two lines.

*Answer here*

### Multiple Regression

# RQ2: Does the relationship between math achievement and proportion of students receiving free/reduced lunch differ by locale?

1. We want to see if the relationship between math achievement test and percentage of students eligible for free lunch differs by locale.

The model is:

1. Which parameter (beta) is of direct interest for our research question?

*Answer here*

The null hypothsis is:

*Answer here*

1. Run the model. Did we reject the null?

fit2 <- lm(math ~ perfl + locale + perfl:locale, data = seda)  
anova(fit2)

## Analysis of Variance Table  
##   
## Response: math  
## Df Sum Sq Mean Sq F value Pr(>F)   
## perfl 1 325.92 325.92 14657.720 < 2.2e-16 \*\*\*  
## locale 3 8.33 2.78 124.843 < 2.2e-16 \*\*\*  
## perfl:locale 3 1.96 0.65 29.401 < 2.2e-16 \*\*\*  
## Residuals 12231 271.96 0.02   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Answer here*

1. Visualize the relationship and interpret.

**Figure 2. Visualization of the relationship between math achievement and percentage of FL students by locale**

#modelsummary(fit2)  
  
## Make a plot, add locale to the color aesthestics and give meaningful names  
## margins::margins(fit2,  
# at = list(locale = c(..., ..., ..., ...))) %>%   
# ggplot(fit,  
# aes(x = perfl, y = math, color = ...)) +  
# geom\_point()+  
# geom\_smooth(method = 'lm') +  
# labs(x = "...",  
# y = "...",  
# color = "...") +  
# theme\_minimal(base\_size = 14)

*Answer here*