HW 04: Alumni jobs

Insert Name

> **!** Important
>
> Due: Friday, September 27, 11:59pm

In this homework, you'll use simple linear regression to analyze the relationship between the median early career salary and percent of alumni who perceive their job as making the world a better place for colleges and universities in the United States.

Learning goals

By the end of the homework you will be able to...

- assess model conditions for simple linear regression.
- calculate and interpret predicted values from the model.
- assess model fit using R^2 and RMSE

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 4 and upload the `hw-04.qmd` and `alumni-salaries.csv` file into the folder your just made.

Packages

We'll use the following packages in this homework.

```
library(tidyverse)
library(broom)
library(yardstick)
library(ggformula)
library(knitr)
library(patchwork)
# add more packages as needed
```

Data: Alumni jobs

Today's data set is part of the TidyTuesday College tuition, diversity, and pay.

The information in this data set was collected from the PayScale College Salary Report.

variable	class	description
rank	double	Potential salary rank within state
name	character	Name of school
state_name	character	state name
early_career_pay	double	Median salary for alumni with 0 - 5 years experience (in US dollars)
mid_career_pay	double	Median salary for alumni with 0 - 5 years experience (in US dollars)
make_world_better_percent	double	Percent of alumni who think they are making the world a better place
stem_percent	double	Percent of degrees awarded in science, technology, engineering, or math subjects

```
alumni <- read_csv("alumni-salaries.csv")
```

Exercises

Is there a relationship the typical early career pay for alumni and the percent of alumni who receive a STEM degree? To answer this question, we will use regression to predict the early career pay using the percent of STEM degrees.

Exercise 1

Fit a linear model that can be used to predict the typical (median) early career pay based on the percent of STEM degrees. Display the model output using 3 digits for numerical values.

Exercise 2

Calculate the predicted values and residuals from your model and save these results in a data frame. Print the first five rows of the new data frame.

Exercise 3

Before using the model for prediction, let's check the model conditions. To figure out how to do so, we type the following prompt into ChatGPT ("an AI powered language model") *"How do you check the model conditions for simple linear regression?"*

The response from ChatGPT shown below would receive partial credit. For each condition, state whether it is correct or incorrect/incomplete. If it incorrect/incomplete, state what is wrong or missing, then write the correct response. Please put all answers after the ChatGPT text.

> *In simple linear regression, you're trying to establish a linear relationship between two variables: a dependent variable (Y) and an independent variable (X). The model conditions or assumptions for simple linear regression are important to check to ensure that your model is valid and reliable. Here are the key assumptions and how to check them:*
>
> *1. Linearity Assumption:*
>
> - *Assumption: The relationship between the independent variable (X) and the dependent variable (Y) is linear.*
> - *How to Check: Create a scatterplot of Y against X. If the points roughly form a straight line, the linearity assumption is likely met.*
>
> *2. Independence Assumption:*
>
> - *Assumption: The residuals (the differences between observed Y and predicted Y) are independent of each other.*
> - *How to Check: Plot the residuals against the independent variable (X) or the predicted values. Look for patterns or trends in the residuals. There should be no clear structure or autocorrelation in the residuals.*
>
> *3. Homoscedasticity [Constant variance] Assumption:*

- *Assumption: The variance of the residuals is constant across all levels of the independent variable.*
- *How to Check: Plot the residuals against the predicted values or the independent variable (X). The spread of the residuals should remain roughly constant across the range of X.*

4. Normality Assumption:

- *Assumption: The residuals follow a normal distribution.*
- *How to Check: Create a histogram or a normal probability plot (Q-Q plot) of the residuals. If the residuals are approximately normally distributed, they should follow a bell-shaped curve in the histogram and lie along a straight line in the Q-Q plot.*

It's important to note that these assumptions may not always hold perfectly in real-world data, and minor violations may not necessarily invalidate your model. However, if you detect significant violations of these assumptions, you may need to consider alternative modeling techniques or transformations of your data to address the issues.

Exercise 4

Determine whether the model conditions hold.

> **!** Important
>
> Regardless of your previous answer, you can proceed assuming all the model conditions are satisfied.

Exercise 5

- 31 percent of the alumni at Johns Hopkins University received a STEM degree. Calculate the predicted early career pay at this university and the appropriate 95% interval. Interpret the interval in the context of the data.
- The actual early career pay at this university is $67,200. Is this what you expected based on your analysis? Briefly explain your reasoning.

Exercise 6

Next, let's consider how well the model fits the relationship between the early career pay and percent with a STEM degree.

- Calculate R^2 and interpret it in the context of the data.
- Calculate RMSE and interpret it in the context of the data.

Exercise 7

Do you think the model is useful for understanding and predicting the typical early career pay for alumni at a university? Briefly explain your reasoning.

Exercise 8

Fit another model which predicts the typical early career pay using only the percentage of alumni that believe their career will make the world a better place.

- Neatly display this model with 3 digits.
- Which model do you think is better - this model or the model fit in Exercise 1? Explain your response showing any analysis used to make the decision.

Submission

> **Warning**
>
> Before you wrap up the assignment, make sure you render your PDF and it appears how you want it to.

Upload the qmd and PDF files to Canvas.

Grading (50pts)

Component	Points
Ex 1	3
Ex 2	4
Ex 3	8
Ex 4	8
Ex 5	7
Ex 6	6
Ex 7	4
Ex 8	6

Component	Points
Workflow & formatting	4[1]

[1]The "Workflow & formatting" grade is to assess the reproducible workflowamd overall readability.