EDF 5401 Homework Assignment 1

(total points: 90)

Please read all questions carefully. If you are working with a group to complete this assignment, all members of your group must sign the statement below:

All members of this homework team contributed equitably to the assignment.

Member 1:

Member 2:

Member 3:

```r
library(DescTools)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

This assignment is based on the data file “jobsat.csv” (or “jobsat.sav” for SPSS users). This file contains 9 variables from 200 employees from a certain type of company, plus a variable “ID”. The meaning of each variable is as follow:

- Gender: 1=female; 2=male
- Age: the age of employee

- Physical environment: a scale score indicating how employees like the current physical environment in their companies; scale scores range from 1 to 30 with higher scores indicating positive feeling about the environment.
- Performance: a rating on employee's previous year's job performance using a scale of 1 to 50 with higher rating suggesting better performance.
- Previous year salary: previous year's salary in thousands of dollars
- Current salary: current annual salary in thousands of dollars
- Stress: employee's perceived stress for their job, higher score suggesting higher level of stress.
- Job satisfaction: satisfaction for the current job; scale scores range from 1 to 50 with higher scores indicating higher level of satisfaction.
- Rating (rating): supervisor's rating on this year's performance as measured by 0 (unsatisfactory) or 1 (satisfactory)

```r
jobsat <- read_csv("jobsat.csv")
```

```
Rows: 200 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): gender, rating
dbl (8): ID, age, environment, performance, preyearsalary, currentsalary, st...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

In this homework assignment, you only need to consider two variables: job satisfaction and current salary (plus the ID number). You try to describe and understand the relationship between job satisfaction and current salary. Commonly speaking, employee tends to feel satisfied with their current job if they perceived they are well-paid. You man use R, SPSS, or any other stat package you are familiar with.

1. (5 pts.) Suppose you want to predict one of the two variables from the other. Based on the description above and your understanding of the relationship between job satisfaction and current salary, which variable do you think is more appropriate to be conceptualized as outcome variable, and which as predictor?

2. (5 pts.) Use a statistics package to compute descriptive statistics (e.g., mean, median, standard deviation, skewness and kurtosis). Make two histograms or Q-Q plots, one for job satisfaction and the other for current salary. Do these two variables seem normally distributed?

3. (5 pts.) Use a statistics package to make a scatterplot of job satisfaction versus current salary. How would you characterize the relationship between job satisfaction and current salary? Be sure to describe the relationship in terms of three aspects: direction, strength, and shape.

4. (5 pts.) Check the scatterplot. Do any point(s) appear to be outlier(s)? What evidence supports your choice(s)? (Hint: making boxplots might be useful).

5. (5 pts.) Obtain the bivariate correlation between job satisfaction and current salary. Does the correlation you have obtained support your description in Question 3?

6. You have defined outcome variable and predictor in the Question 1. Conduct simple regression that predicts the outcome variable from the predictor using this sample of data. Save residuals from the model, as well as dfbetas. Then answer the following questions:

6.1 (5 pts.) Write equations for this regression model at the population level (two equations: one for observed scores and the other for the regression line). Define each element in the equation.

> In Rmarkdown, you can make an equation by enclosing it in a pair of `$`s. Displayed equations have double dollar signs. Within an equation, `_` is a subscript and `^` is a superscript. Greek letters are made by using a backslash (\) followed by the name of the letter. For example, `\epsilon` becomes ϵ.

Observed Scores:

$$Y_i =$$

Regression Line:

$$Y =$$

(6.2) (5 pts.) Check "Coefficients" table. What are the values of b_0 and b_1 the unstandardized regression equations for this model (two equations: one for observed scores and the other for the regression line).

(6.3) (5 pts.) Interpret the value of b_0 and b_1.

(6.4) (5 pts.) Obtain the value of MSE and SEE. What do these two values tell you? Comparing these two values to the variance and standard deviation of the outcome variable, does the predictor seem to help in predicting the outcome variable?

(6.5) (5 pts.) Check residuals and dfbetas, which case(s) seem to be outlier(s)? Support your answer with appropriate indices. Explain in words the meaning of the largest dfbeta value for b_1.

7. Exclude the outlier(s) you identified in the Question 6.5 and re-run the simple regression. Save residuals and dfbetas from the model, and make a plot with standardized residuals vs. predicted Y values. Then answer the following questions:

(7.1) (5 pts.) What are the value of slope (b_1) and its standard error (s_{b_1})? Use these two values to conduct the test of H0: $\beta_1 = 0$ at $\alpha = .05$. What is your t value? p value? What is your conclusion about the effect of the predictor on the outcome variable?

(7.2) (5 pts.) What's the null hypothesis of the F test? Using values from the ANOVA table, compute the F value. What is your p value? Suppose $\alpha = .05$, what is your statistical conclusion for this hypothesis testing?

(7.3) (5 pts.) Square the t value for the slope from Question 7.1, and compare it to the F value from Question 7.2. Explain your findings. Do the $F%-test and t*-test lead to the same statistical conclusion?

(7.4) (5 pts.) Compute the 95% confidence interval (C.I.) around the estimated slope (b_1). What does this interval tell you about the population value of slope? What does this interval suggest about the relationship between job satisfaction and current salary? Compare this information to that given by the t test in Question 7.1, does this C.I. lead to the same statistical conclusion as the t-test?

(7.5) (5 pts.) Describe the relationship between R^2 and the quantities in the ANOVA table. Interpret the meaning of this value.

(7.6) (5 pts.) Do you think this predictor (current salary) is important in predicting the outcome variable (job satisfaction)? Why do you think it is important (or not important)?

(7.7) (5 pts.) Check the scatterplot between standardized residuals and predicted Y values (SPSS default) or the scatterplot between square root of the absolute value of the residuals (R default). Does the assumption of homogeneity of error variance (homoscedasticity) seem to be met for this regression analysis?

(7.8) (5pts.) Does the fitted model excluding the outliers seem to be similar to or different from the model based on all 200 cases? Comment on any differences you see between these two models.