Homework 7

Due date: 11/14/2024

November 7, 2024

1 Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

2 Exercises

Exercise 1: Simple linear regression
Load the **mtcars** dataset inside the library **datasets**.

(a) Create a scatterplot with **hp** on the x-axis and **mpg** on the y-axis. Do you see any particular trend? **[5 pts]**

(b) Fit a linear regression with mpg as the response variable and hp as the predictor **[5 pts]**

(c) Test the hypothesis of no relationship between hp and mpg. Write down the null hypothesis and the alternative one. Do we reject the null hypothesis? Explain why. **[5 pts]**

(d) Write down the equation of the resulting line using the estimated coefficients. For a one unit increase of hp, how much is the mpg increasing? **[5 pts]**

(e) Create a 95% confidence interval for the coefficient of hp without using bootstrapping and/or the sandwich library. Which assumptions should we verify to rely on the just created confidence interval? **[5 pts]**

(f) Redo the plot in (a) adding the fitted regression line. Add the fitted values (the projection of the points over the line) and draw the residuals. Is this line the one minimizing the residual sum of squares? **[5 pts]**

(g) Write down the formula for the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE) (*Hint: you should have to study them by yourself*). Explain their meaning in words also using the plot created in (f). Compute the two values. **[5 pts]**

(h) Consider now the multiple R^2. Write down the formula and explain it in words in terms of residuals from the fitted line and residuals from the mean. (*Hint: for reference see the newly updated version of lecture 15 available on canvas*) **[5 pts]**

Exercise 2: Assumptions and confidence intervals
Consider the linear model studying weight (wt) as a function of age for the **Gestation** data from the **mosaicData** library.

(a) Generate a 90% and a 95% confidence interval for the coefficient of age. Which one is wider. Explain why. **[5 pts]**

(b) Do the same using the functions from the sandwich library. When should I rely on this library instead of using the summary? **[5 pts]**

(c) Generate 90% and a 95% confidence interval for the coefficient of age under the circumstances that no assumptions are verified (*Hint: you can't use summary() or the sandwich library*). **[10 pts]**

Exercise 3: Dealing with categorical variables
Let's read the data set Salaries from the **car** package and let's investigate relationship of the response variable **Salaries** with different regressors.

(a) Fit a linear model to study the relationship between the response variable salary and the regressor sex. Write the equation of the model (without using the estimated coefficients) and the resulting equation when the gender is female and when the gender is male (using the estimated coefficients). Show the results in a plot. **[5 pts]**

(b) Now repeat (a) studying the relationship between the response variable salary and the regressor discipline. Write the equation of the model (without using the estimated coefficients) and the resulting equation for each discipline (using the estimated coefficients). Show the results in a plot. **[5 pts]**

(c) Fit a parallel linear model to study the relationship between the response variable and the regressors sex and discipline. Write the equation of the model (without using the estimated coefficients) and the resulting equation for each discipline and sex alternatives (using the estimated coefficients). Show the results (lines) in a plot (*Hint: in the plot use colors to address one of the two categorical regressor and keep the other in the x-axis*). **[10 pts]**

3 Submission Instructions

Please submit your completed exercises by **November 14th** through **gradescope**. Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

4 Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

- **Lecture 3** - bootstrapping estimated sampling distributions and confidence intervals
- **Lecture 11** - least squares regression;
- **Lecture 12** - bias in least squares regression;
- **Lecture 15** - least squares regression;
- **Lab** - practicing all of the above