Homework 7: Multiple Linear Regression

Mario Giacomazzo

October 14, 2023

Instructions:

The purpose of this homework assignment is to practice doing multiple linear regression. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you what I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. Also, if you see the comment "#DO NOT CHANGE", then I don't want you to make any modifications to that code. **You should knit your RMD file to a PDF after you answer every question.**

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under HW7.

Questions

Q1 (2 Points)

In the **Ecdat** library, there is a dataset called **Clothing** that contains sales data of 400 men's fashion stores from 1990 in the Netherlands. The response variable for this assignment will be *tsales* which contains the total annual sales for each of these stores in 1990 measured in Dutch guilders (currency in Netherlands prior to Euro). The first thing I want you to do is to create a dataframe called **Clothing2** which contains all the data in **Clothing** but removes the other optional response variables, *sales* and *margin.* We only want the targeted response variable *tsales* and 10 other potential predictor variables (nown,nfull, npart, . . . ,start). After doing this, you should definitely read the documentation about the **Clothing** dataset so you understand each of the predictor variables.

Use the `str()` function on **Clothing2** to show a preview of this dataset. This should be the only output from this code chunk. Going forward we we only use **Clothing2**.

Q2 (3 Points)

Fit the full linear regression model to predict **tsales** using the data in **Clothing2**. This is the linear regression model where all 10 potential predictor variables are included. When creating your formula in the `lm()` function, you can use code in the form of `y~.`. The period symbol tells the `lm()` function to include all other (non-response) variables in the dataset as predictors.

In your output, I only want to see the following three diagnostic plots from this model:

1. Histogram of residuals
2. Normal quantile plot of residuals
3. Scatterplot of residuals versus fitted values

Q3 (3 Points)

Based on the diagnostic plots, there seems to be a decent deviation away from the Normality assumption on the residuals. Create a new variable in **Clothing2** called **tsales.cubic** which is the cubic root of **tsales**.

Fit the full linear regression model to predict **tsales.cubic** using the data in **Clothing2**. This is the linear regression model where all 10 potential predictor variables are included. You are ignoring the original response variable, **tsales**, in your set of predictor variables. When creating your formula in the `lm()` function, you can use code in the form of `y~. - x33`. The period symbol tells the `lm()` function to include all other (non-response) variables in the dataset as predictors. The `- x33` part tells the `lm()` function to remove the variable `x33` from the set of predictor variables.

In your output, I only want to see the following three diagnostic plots from this model:

1. Histogram of residuals
2. Normal quantile plot of residuals
3. Scatterplot of residuals versus fitted values

Q4 (5 Points)

Based on the diagnostic plots from the previous model, the transformation seemed to work. Therefore, going forward, the variable **tsales.cubic** will be the response variable in all future models.

I want you to fit the simple linear regression model for the relationship of **tsales.cubic** versus **tsales**. Use the code chunk to run any necessary code required to answer the questions that are below the code chunk. Write in complete sentence(s) and round your numbers to 2 decimal places.

What is the R-squared statistic for the full model in Question 3 where we ignored the original response variable?

REPLACE THIS SENTENCE WITH YOUR ANSWER

What is the R-squared statistic for simple linear regression model that was discussed in this question?

REPLACE THIS SENTENCE WITH YOUR ANSWER

Do you think it is surprising that the simple linear regression does a far better job at understanding the variation in our response variable than the full model in Question 3? Why or why not?

REPLACE THIS SENTENCE WITH YOUR ANSWER

Q5 (2 Points)

Next, I want you to fit the empty linear regression model to predict **tsales.cubic**. An empty linear regression model is the linear regression model where there are predictor variables (no coefficients). You are fitting a linear regression model where there is only a intercept. To do this, you would run the `lm()` function with a model in the form `y~1`. The *1* in the model expression tells the `lm()` function that the only parameter on the right side of the formula for predicting y is the y-intercept.

In your output, I only want to see the output from the `summary()` function and the output from the `anova455()` function applied to this model.

Q6 (4 Points)

Next, I want you to fit another linear regression model to predict **tsales.cubic** based off the output from the full model in Question 3. I want to use the `summary()` function on the full model from Question 3 to identify all of the predictor variables that had significant p-values for their individual t-tests. Then, I want you to fit another linear regression model predicting **tsales.cubic** using only those significant variables.

In your output, I only want to see the output from the `summary()` function applied to the full model in Question 3 and the `summary()` function applied to the smaller model based on the significant variables. Then, answer the question below the code chunk in complete sentences.

After removing the variables from the full model that were not proven to have coefficients significantly different from 0, which of the remaining predictor variables became statistically insignificant according to individual t-tests from the smaller model?

REPLACE THIS SENTENCE WITH YOUR ANSWER

Q7 (9 Points)

The `cor()` function in R can be used on a dataframe to show the correlation of every pair of variables in a dataframe. I want you to use the `cor()` function to print out the correlation matrix only for the potential predictor variables in **Clothing2**. We want to use this matrix to identify predictor variables that are highly correlated with each other. When you run this function, the correlations will print with too many decimal places. I want you to run the `round()` function on the correlation matrix to round all correlations to 2 decimal places.

After you do this, you will notice two variables that are highly correlated (>0.80). Both of these variables are in the smaller model in the previous Question 6. I want you to refit the smaller linear regression model from Question 6 twice where each of them has only 1 of these 2 highly correlated predictor variables. I want you to print out the ANOVA table using the `anova455()` function for each of these modified smaller models.

Your output should only contain the correlation matrix of all 10 predictor variables rounded to 2 decimal places and both ANOVA tables from the two models that are subsets of the model in Question 6.

After all of this, answer the question below the code chunk using complete sentences.

Based off the ANOVA output, which of the two highly correlated predictor variables would be better to remove and why?

REPLACE THIS SENTENCE WITH YOUR ANSWER