# Task 1: Getting Started with R

This informal report contains a brief overview of the procedures and the results obtained in the two parts of this task.

## Part 1: Car distance prediction:

This task was pretty simple and all steps were well described. Therefore, I will move on directly to the results obtained. It’s only worth adding that in the linear regression, I removed the intercept term by subtracting by 1 because this increased the model’s performance. In the end, I obtained the following results:

|  |  |
| --- | --- |
| **Multiple R-Squared** | 0.9428 |
| **p-value** | <2.2e-16 |

The multiple R-squared is 0.9428 which is pretty close to 1, this models seems to be a good fit. The p-value is clearly smaller than 0.05 meaning that the relationship between these two variables is statistically significant.

The predictions obtained with the model mentioned above are:

|  |  |  |
| --- | --- | --- |
| **index** | **Predicted Distance** | **Real Distance** |
| 6 | 27.73984 | 16 |
| 9 | 30.82205 | 20 |
| 14 | 36.98645 | 26 |
| 16 | 40.06866 | 26 |
| 17 | 40.06866 | 28 |
| 19 | 40.06866 | 32 |
| 21 | 43.15086 | 32 |
| 22 | 43.15086 | 34 |
| 23 | 43.15086 | 34 |
| 29 | 52.39748 | 42 |
| 33 | 55.47968 | 50 |
| 35 | 55.47968 | 54 |
| 39 | 61.64409 | 60 |
| 42 | 61.64409 | 68 |
| 43 | 61.64409 | 70 |

It was predictable that this model would not be sensible to small variations, as can be seen above, due to the simplicity of the linear regression. Observing the plot of the car speed relative to the car distance, it is visible that the relationship between these two variables is not exactly linear. It seems to follow a slow exponential increase or at least polynomial. Additionally, because an ordinary least square method was used as an optimizer, the outliers negatively impact the final slope of the linear regression, thus limiting the quality of a linear regression for this problem. A different mathematical regression could be used to improve these results. For instance, I experimented implementing an exponential based regression and the Multiple R-Squared increased to 0.97 while the p-value remained stable.

On the other hand, the proportionality of the car speed to the car distance has been captured by this model and the predicted distance is never too far from the real distance.

In conclusion, this model is not perfect but it is quite a good fit.

## Part 2: Petal length prediction:

In this task I was supposed to firstly correct the mistakes found in the code shown and secondly to share the predictions results.

Starting with the mistakes found, many were just simple syntax errors that did not deserve much attention (such as plural word instead of singular, missing parentheses or quotation marks, etc.) others required further attention. Below follows a table with the suggested code and the corrections I have made (if required):

|  |  |
| --- | --- |
| **Original code** | **Corrections** |
| install.packages(readr) | Missing quotation marks. Correction:  *install.packages(“readr”)* |
| library("readr") | Quotations marks and unnecessary:  *library(readr)* |
| IrisDataset <- read.csv(iris.csv) | None |
| attributes(IrisDataset) | None |
| summary(risDataset) | Missing letter. Correction:  *summary(IrisDataset)* |
| str(IrisDatasets) | None |
| names(IrisDataset) | None |
| hist(IrisDataset$Species) | Unable to plot histogram of vector of strings in R. Either factorize:  *plot(as.factor(IrisDataset$Species))*  Or use other library such as ggplot2: *library("ggplot2")*  *qplot(IrisDataset$Species)*  Or it was pretended to create a histogram of the other variables:  *hist( IrisDataset$Sepal.Length)*  *hist( IrisDataset$Sepal.Width)*  *hist( IrisDataset$Petal.Length)*  *hist( IrisDataset$Petal.Width)* |
| plot(IrisDataset$Sepal.Length | Missing parenthesis:  *plot(IrisDataset$Sepal.Length)* |
| qqnorm(IrisDataset) | None |
| IrisDataset$Species<- as.numeric(IrisDataset$Species) | Cannot convert characters into numeric, convert into factor:  *IrisDataset$Species <- as.factor( IrisDataset$Species)*  If numeric is required then an additional step would be required:  *IrisDataset$Species <- as.numeric( as.factor(IrisDataset$Species))* |
| set.seed(123) | None |
| trainSize <- round(nrow(IrisDataset) \* 0.2) | Changed train size to 70% of original dataset:  *trainSize <- round(nrow(IrisDataset) \* 0.7)* |
| testSize <- nrow(IrisDataset) - trainSet | Changed trainset variable into trainSize:  *testSize <- nrow(IrisDataset) - trainSize* |
| trainSizes | Removed extra letter from “trainSizes”:  *trainSize* |
| testSize | None |
| trainSet <- IrisDataset[training\_indices, ] | Before this I had to select which indices were the training ones. I selected a random sample:  *training\_indices<-sample(seq\_len(nrow(IrisDataset)),size =trainSize*  This line itself received no correction. |
| testSet <- IrisDataset[-training\_indices, ] | None |
| set.seed(405)  trainSet <- IrisDataset[training\_indices, ]  testSet <- IrisDataset[-training\_indices, ] | This part was a repetition of the above. I removed it. |
| LinearModel<- lm(trainSet$Petal.Width ~ testingSet$Petal.Length) | The analysis goal was to predict a petal’s length using the petal’s width, not the other way around. Also it wouldn’t make sense using the testing set when fitting the model. Therefore I had to edit the formula and y parameters:  *LinearModel<-lm(Petal.Length~Petal.Width, trainSet)* |
| summary(LinearModel) | None |
| prediction<-predict(LinearModeltestSet) | Missing comma:  *prediction<-predict(LinearModel,testSet)* |
| predictions | Plural word instead of singular:  *prediction* |

Now I will move on to the second part of this section, the results. Utilizing the linear regression as suggested, output the following results:

|  |  |
| --- | --- |
| **Multiple R-Squared** | 0.9271 |
| **p-value** | <2.2e-16 |

The multiple R-squared is 0.9271 which is pretty close to 1, this models seems to be a good fit. The p-value is clearly smaller than 0.05 meaning that the relationship between these two variables is statistically significant.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **index** | **Predicted Length** | **Real Length** | **index** | **Predicted Distance** | **Real Length** |
| **1** | 1.48243059262792 | 1.4 | **68** | 3.30330350345734 | 4.1 |
| **2** | 1.48243059262792 | 1.4 | **70** | 3.53091261731101 | 3.9 |
| **3** | 1.48243059262792 | 1.3 | **77** | 4.21373995887204 | 4.8 |
| **5** | 1.48243059262792 | 1.4 | **83** | 3.75852173116469 | 3.9 |
| **11** | 1.48243059262792 | 1.5 | **84** | 4.6689581865794 | 5.1 |
| **18** | 1.7100397064816 | 1.4 | **94** | 3.30330350345734 | 3.3 |
| **19** | 1.7100397064816 | 1.7 | **95** | 3.98613084501837 | 4.2 |
| **28** | 1.48243059262792 | 1.5 | **98** | 3.98613084501837 | 4.3 |
| **29** | 1.48243059262792 | 1.4 | **100** | 3.98613084501837 | 4.1 |
| **33** | 1.25482147877425 | 1.5 | **101** | 6.71744021126249 | 6.0 |
| **36** | 1.48243059262792 | 1.2 | **104** | 5.12417641428675 | 5.6 |
| **45** | 1.93764882033528 | 1.9 | **105** | 6.03461286970146 | 5.8 |
| **48** | 1.48243059262792 | 1.4 | **111** | 5.5793946419941 | 5.1 |
| **49** | 1.48243059262792 | 1.5 | **113** | 5.80700375584778 | 5.5 |
| **55** | 4.44134907272572 | 4.6 | **116** | 6.26222198355513 | 5.3 |
| **56** | 3.98613084501837 | 4.5 | **125** | 5.80700375584778 | 5.7 |
| **57** | 4.6689581865794 | 4.7 | **131** | 5.35178552814043 | 6.1 |
| **58** | 3.30330350345734 | 3.3 | **133** | 6.03461286970146 | 5.6 |
| **59** | 3.98613084501837 | 4.6 | **135** | 4.21373995887204 | 5.6 |
| **61** | 3.30330350345734 | 3.5 | **140** | 5.80700375584778 | 5.4 |
| **62** | 4.44134907272572 | 4.2 | **141** | 6.48983109740881 | 5.6 |
| **65** | 3.98613084501837 | 3.6 | **145** | 6.71744021126249 | 5.7 |
| **66** | 4.21373995887204 | 4.4 |

Once again, it was predictable that this model would not be sensible to small variations. As can be seen above, due to the simplicity of the linear regression some predicted values are slightly off.

Observing the plot of the petal length against the petal width, it seems that the relationship between these two variables is linear. There are quite a few outliers, but a linear regression seems to be a solid mathematical approach to modelling this relation.

Similarly to the first part the proportionality of the petal width to petal length has been captured by this model and the predicted length is never too far from the real value.

In conclusion, this model is quite a good fit.

## Conclusions

This task was an introduction to programming and modelling in R. Installing R and RStudio was no challenge whatsoever as every step of the installation is detailed in this tutorial as well as R’s website. This simple tutorial provided a solid framework to develop simple mathematical models in RStudio and a basic exploratory data analysis. I believe that the basics have been covered here. The next step will be to go into further detail and understand the specifics of this language (such as how to work with different data types, efficiency, etc.) and understand if this scales to datasets with thousands or millions of rows.