Informal Report on Installing and Using R

*Prediction Results & Performance Metrics*

*R Tutorial (Cars Dataset)*

* Variables
  + Independent Variable: Speed of Car
  + Dependent Variable: Distance Traveled by the Car
* Performance Metrics
  + Multiple R-Squared Value: 0.9554
    - Very close to 1; indicates a good fit to regression line
      * Linear relationship
  + P-Value: < 2.2E-16 ≈ 0
    - Relationship between Speed and Distance is Statistically Significant

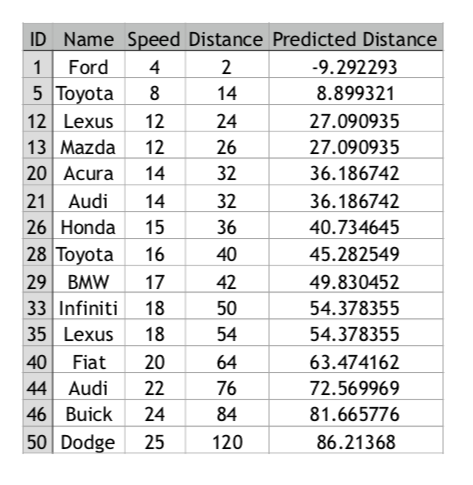


Figure : Predicted Traveled Distance based on Car Speed

*Find the Errors (Iris Dataset)*

* Variables
  + Independent Variable: Petal Width
  + Dependent Variable: Petal Length
* Performance Metrics
  + Multiple R-Squared Value: 0.9206
    - Very close to 1; indicates a good fit to regression line
      * Linear relationship
  + P-Value: < 2.2E-16 ≈ 0
    - Relationship between Petal Width and Petal Length is Statistically Significant

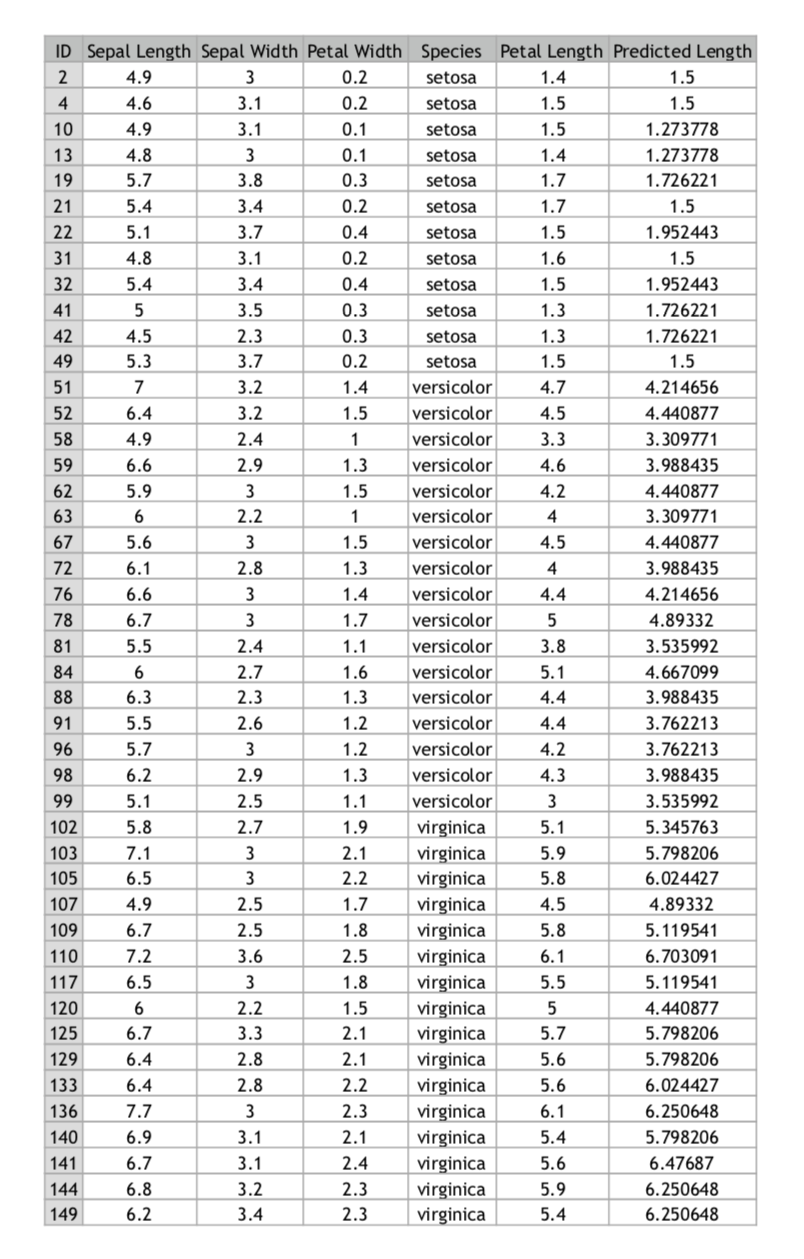


Figure : Predicted Petal Length based on Petal Width

*Errors & Warning Messages Encountered*

* Error: object ‘IrisDataset’ not found
  + First major obstacle was not being able to figure out how to load the Iris.csv data to the environment using the R Script. I solved it by adding **source(“Iris.R”)** to my script before calling the **read.csv** command.
* Second obstacle was not understanding why, when I ran the script, it would only execute the last line. I learned that if you want R to output multiple lines, you have to select them. If you do not select them, it will only output the line your cursor is on.
* Error in hist.default(IrisDataset$Species) : ‘x’ must be numeric
  + The way I fixed this was by instead of using the **hist()** graph, I used the **plot()** graph. This gave me the same graph that the histogram would have given me.
* Errors I noticed in given script file:
  + install.packages(readr) fixed to install.packages(“readr”)
  + library(“readr”) fixed to library(readr)
  + qqnorm(IrisDataset) does not specify a column
  + trainSize was set to 20% of data sets
    - fixed to 70%
  + missing specification of training\_indices
    - training\_indices<-sample(seq\_len(nrow(IrisDataset)),size=trainSize)
  + prediction<-predict(LinearModeltestSet)
    - prediction<-predict(LinearModel, testSet)

*Other Comments*

* It was really easy to start R Studio and R. All I had to do was choose the right installers to download, and the rest was done by the computer. Syntax was not really a big issue for me, since I have experience coding in Python; R and Python have similar syntax.
* The tutorial was somewhat useful in my opinion. It covered many of the big steps needed to complete the “Get Started with R” task, but I had to do a lot of outside Googling to see why R sometimes behaved in ways I did not expect it to.
  + I would recommend the tutorial to others. I got a really good grasp as to how I can use R and RStudio to predict data using a linear regression model.
* I think learning a lot of the commands was one of the most important lesson learned from this task. Now I know how to install and load packages in RStudio. I also learned basic syntax to code in R. For example, I wasn’t familiar with the commonly used **<-** to assign values, tables and other variables.
* I would recommend being really careful with syntax, spelling, capitalization and many other details in your code, since R takes what you type literally. For R, testset is not the same as testSet. Also, given that R is open source, people all around the world are using it and improving it, and a question, error or warning you might have, is probably already answered somewhere online. Therefore, I recommend constantly looking up documentation, forums or anything that can help you when you find yourself stuck on a specific line of code.