Jennifer Roria

DA310

Prof. Charlene Cheng

Assignment Week 9

1. Read this csv file into R

```
getwd()
setwd("/Users/jennsmacbook/Documents/HW 9")
data <- read.csv('Website on Stress Level.csv')
data
```

2. Do a general analysis (boxplot, head, summary, etc) of the dataset (Exploratory Data Analysis)

   Note: Stress is the variable name for the dataset.

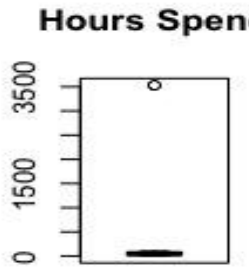
   - **Head and Summary**

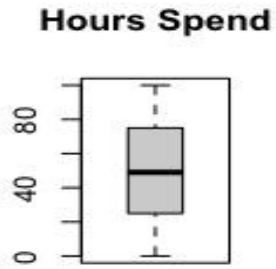
```
head(data$Hours.Spend.on.Website)
head(data$Stress.Test.Points..The.more.the.worse.)

summary(data$Hours.Spend.on.Website)
summary(data$Stress.Test.Points..The.more.the.worse.)
```

```
> head(data$Hours.Spend.on.Website)
[1] 24 50 15 38 87 36
> head(data$Stress.Test.Points..The.more.the.worse.)
[1] 21.54945 47.46446 17.21866 36.58640 87.28898 32.46387
> summary(data$Hours.Spend.on.Website)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   25.00   49.00   54.99   75.00 3530.16
> summary(data$Stress.Test.Points..The.more.the.worse.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  -3.84   24.93   48.97   49.94   74.93  108.87       1
```

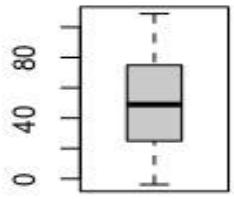
- **Boxplot**

```
13 boxplot(data$Hours.Spend.on.Website, main="Hours Spend", sub=paste("Outlier rows:", boxplot.stats(data$Hours.Spend.on.Website)$out))
14 stress <- data[-c(214),]
15 boxplot(data[-c(214),]$Hours.Spend.on.Website, main="Hours Spend", sub=paste("Outlier rows:", boxplot.stats(data$Hours.Spend.on.Website)$out))
16 boxplot(data$Stress.Test.Points..The.more.the.worse.)
```



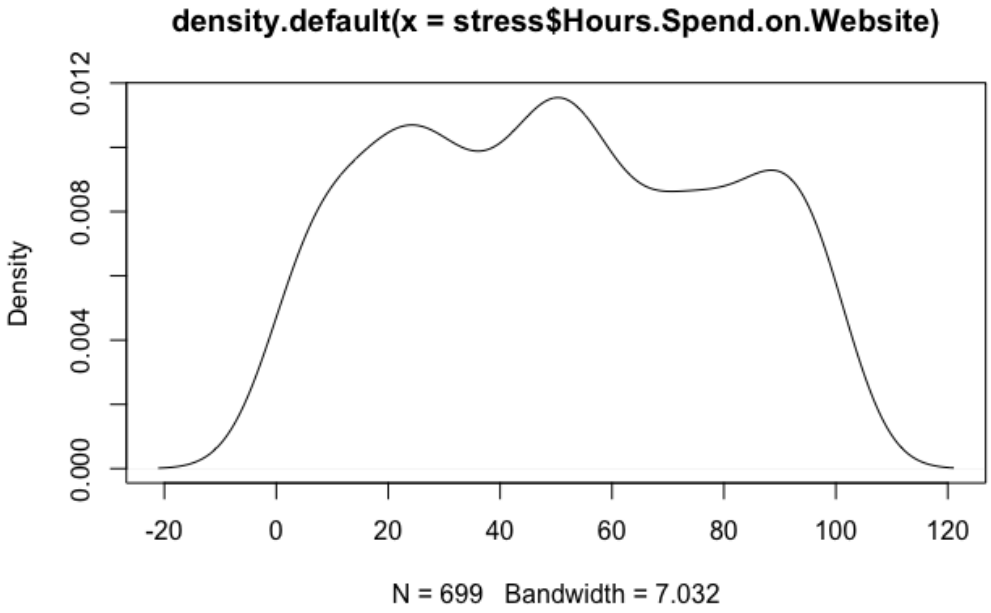


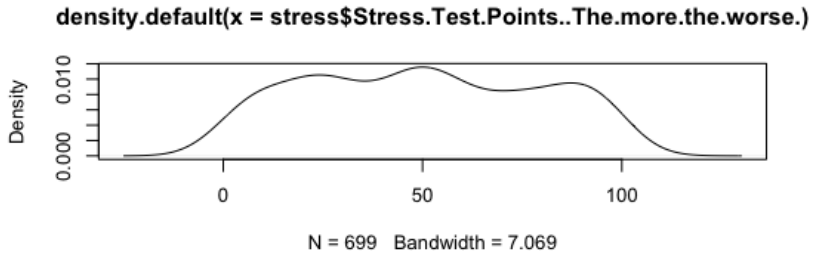
Outlier rows: 3530.15736! Outlier rows: 3530.15736!



- **Density plot**

```
stress <- data[-c(214),]
plot(density(stress$Hours.Spend.on.Website))
plot(density(stress$Stress.Test.Points..The.more.the.worse.))
```

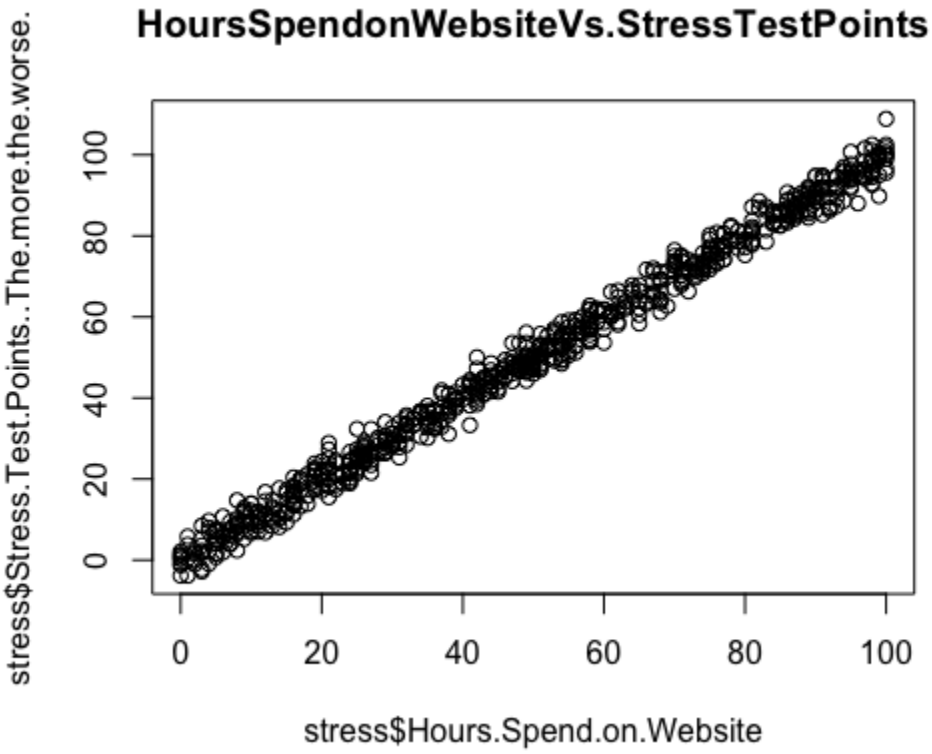




- **Scatterplot**

Note: Stress is the variable name for the dataset.

```
stress <- data[-c(214),]
scatter.smooth(x=stress$Hours.Spend.on.Website, y=stress$Stress.Test.Points..The.more.the.worse., main="HoursSpendonWebsiteVs.StressTestPoints")
```



3. Check the four assumptions of linear regression (If one or more assumptions are not met, please explain the situation in the conclusion)

Independence of Observations: ✓From the model, the observations in the dataset are proven to be independent. Each observation of both stress test points and hours spent on website is independent of the others.
2. Normality: ✓Based on the residuals' normal Q-Q plot, it reveals that the points mainly follow a horizontal line which implies that the remainders have a normal distribution. The residuals histogram likewise displays a broadly bell-shaped distribution, which is consistent with the assumption of normality.
3. Linearity: ✓ There is a linear correlation between the stress test points and the number of hours spent on the website, which is further confirmed by the scatterplot model which demonstrates a positive linear association with an increase in stress test points of 0.990052 for each additional hour spent on the website.
Another evidence to support this is the high value of the multiple R-square. To conclude, itcan be seen from this that the linearity assumption is true.
4. Homoscedasticity: ✓ The dataset's scatter plot variance is constant over the scale of the dependent variables, as shown by the residual plot, which displays a random distribution or spread of dots around the horizontal line at 0. Therefore, The homoscedasticity assumption appears to be validated by this result.

4. Split the dataset into 4:1 ratio as training dataset and testing dataset

   Note: Stress is the variable name for the dataset.

```
install.packages("caTools")
library("caTools")
split=sample.split(stress$Stress.Test.Points..The.more.the.worse., SplitRatio=1/4)
TestData=subset(stress,split==TRUE)
TrainingData=subset(stress,split==FALSE)
TrainingData
```

5. Build up linear regression model on the training dataset

   Note: Stress is the variable name for the dataset.

```
linearMod <- lm (stress$Hours.Spend.on.Website ~ stress$Stress.Test.Points..The.more.the.worse., stress=train)
print(linearMod)
```

```
Call:
lm(formula = stress$Hours.Spend.on.Website ~ stress$Stress.Test.Points..The.more.the.worse.,
    stress = train)

Coefficients:
                                    (Intercept)  stress$Stress.Test.Points..The.more.the.worse.
                                         0.5713                                          0.9901

>
```

6. Check the P-value, R-squared value

   Note: Stress is the variable name for the dataset.

   summary(linearMod)

```
Call:
lm(formula = stress$Hours.Spend.on.Website ~ stress$Stress.Test.Points..The.more.the.worse.,
    stress = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3598 -1.8873  0.0081  2.0166  9.5167

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  0.571255   0.209970   2.721  0.00668 **
stress$Stress.Test.Points..The.more.the.worse. 0.990052   0.003633 272.510  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.794 on 697 degrees of freedom
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9907
F-statistic: 7.426e+04 on 1 and 697 DF,  p-value: < 2.2e-16
```

7. Run prediction on the testing dataset and check the performance

   predicted <- predict(linearMod, TestData)

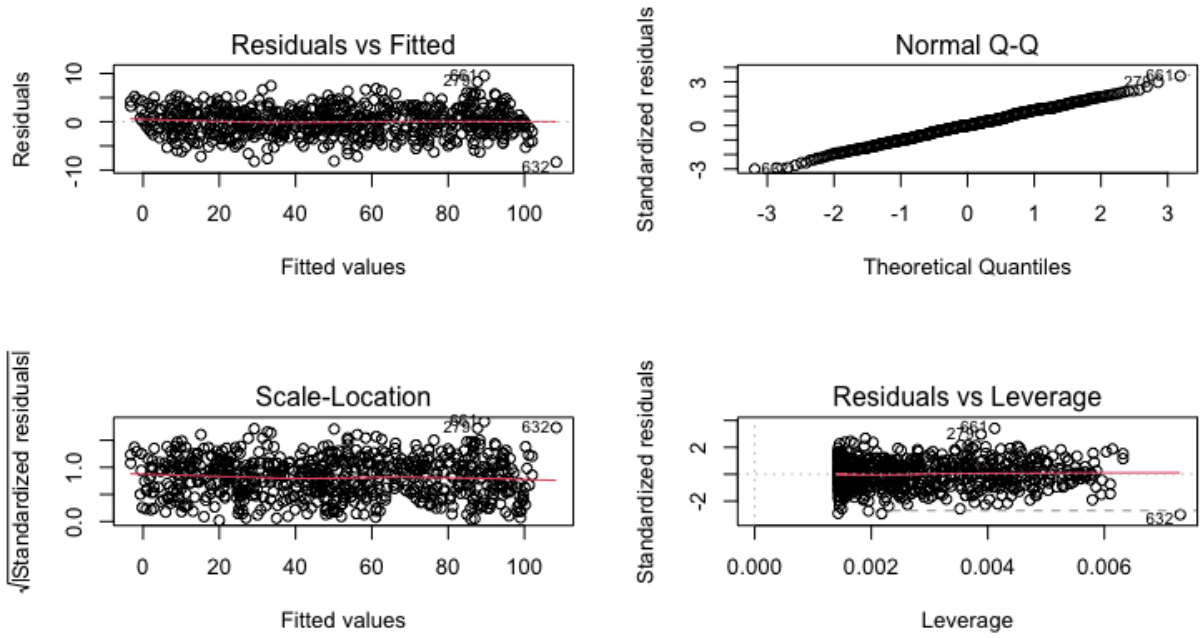
   print(predicted)

   print(TestData)

8. Run more diagnostics of the model

par(mfrow= c(2,2))

plot(linearMod)



9. Conclusion

The model demonstrates a significant positive correlation between stress test points and website runtime. The stress test points variable has a statistical significance p-value of less than 0.05, suggesting that it has a strong linear correlation with the dependent variables. The R-square value is quite large (0.9907), meaning that the predictor variable can account for 99.07% of the variation in the responder variable. There are no obvious patterns or outliers in the residuals, which shows that the linear regression's presumptions are satisfied. Overall, I believe that it that the dataset and this linear regression model are verifiable and satisfactory.

Do you believe you can use linear regression?

According to the data collected, the linear regression model appears to be a good preference for the available dataset, with a high R-squared value and a low p-value indicating that the model explains a significant percentage of the data's variability and the correlation between the determinant and independent variable is statistically significant. In addition, the scatter plot displays a positive linear connection between the data. Furthermore, the scatter plot displays a random distribution of dots surrounding a horizontal line, which means the homoscedasticity assumption is likely true. Lastly, the model's histogram displays a broadly bell-shaped distribution, which is consistent with the assumption of normality.

What is your model and linear relationship you find?

The model generated by the dataset model can be written as follows: Hours.Spend.on.Website = 0.990052 * Stress.Test.Points..The.more.the.worse. + 0.571255 or y = 0.990052x + 0.571255

Where y is the predicted value for hours spent on the website, x is the value of the stress test score, and the coefficients 0.990052 and 0.571255 are the slope and intercept of the linear model, respectively. The slope of the scatter plot demonstrates a positive linear association with an increase in stress test points of 0.990052 for each additional hour spent on the website.

How does this model behave (p-value, r-squared, performance)?

The model acts quite well based on the resulting inputs. The model is statistically significant because of the extremely low p-value ($p < 2.2e-16$), which shows that the discrepancy between the observed and predicted values is highly unlikely to be the accidental consequence. The model accounts for 99.07% of the variation in the dependent variable (Hours.Spent.on.Website) based on the independent variable, according to the high R-squared value (0.9907). (Stress.Test.Points..The.more.the.worse.). This shows that the two variables have a significant

association. Also, the residual standard error is rather low (2.794), which shows that the model fits the data well. Overall, the model's performance with this data is satisfactory.

How do you want to use your final model to do predictions?

The resulting model might be used to estimate a person's duration of time spent on a website based on the results of their stress test. For instance, if someone has a stress test score of 90, using this model we can predict that the person has spent approximately 85.1 hours on the website. Moreover, before employing the model for any real-world applications, it's crucial to test its effectiveness using relevant information/real data. It's crucial to take into account that the data used to train the model and any assumptions made during the modeling process can all affect how accurate the model's predictions are.