Discussion - Week11

David Simbandumwe

```
library(matlib)
library(pracma)
```

```
##
## Attaching package: 'pracma'
```

```
## The following objects are masked from 'package:matlib':
##
##     angle, inv
```

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Hotels Las Vegas Strip

Abstract: This dataset includes quantitative and categorical features from online reviews from 21 hotels located in Las Vegas Strip, extracted from TripAdvisor ([Web Link]).

Data Set Information: All the 504 reviews were collected between January and August of 2015.

Attribute Information: The dataset contains 504 records and 20 tuned features (as of â€œstatus = includedâ€, from Table 1 of the article mentioned below), 24 per hotel (two per each month, randomly selected), regarding the year of 2015. The CSV contains a header, with the names of the columns corresponding to the features marked as â€œstatus = includedâ€, from Table 1 of the aforementioned article.

- ***H0 - Hotel star rating does not predict the review score for a hotel in Las Vegas***
- ***H1 - Hotel star rating predicts the review score for a hotel in Las Vegas***

```
destfile <- tempfile()
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00397/LasVeg
df <- read_delim(destfile, delim=';' ,col_names = TRUE, col_types = list('Hotel stars
```

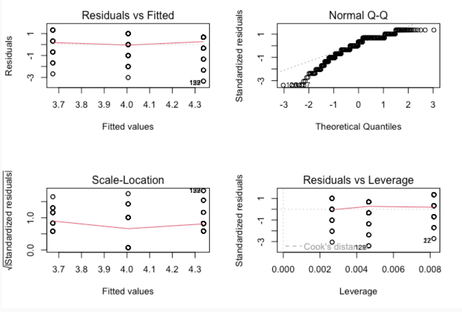
```
## Warning: One or more parsing issues, call `problems()` on your data frame for deta
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
mylm <- lm(df$Score ~ df$'Hotel stars')
summary(mylm)
```

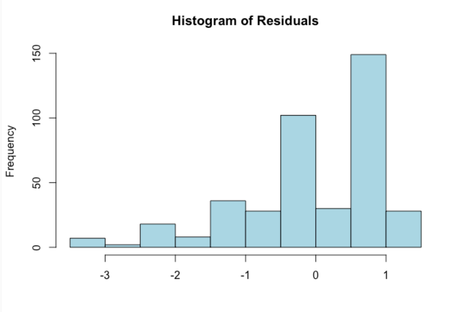
```
##
## Call:
## lm(formula = df$Score ~ df$"Hotel stars")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3369 -0.3369  0.3262  0.6631  1.3262
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.67908    0.26070  10.277  < 2e-16 ***
## df$"Hotel stars"   0.33156    0.06047   5.483 7.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9851 on 406 degrees of freedom
##   (96 observations deleted due to missingness)
## Multiple R-squared:  0.06895,    Adjusted R-squared:  0.06666
## F-statistic: 30.07 on 1 and 406 DF,  p-value: 7.351e-08
```

The linear regression model has p-value of less 0.05 so we reject the Null hypothesis in favour of the alternate hypothesis. However the model has a low predictive capability with an Adjusted R-squared of 0.068

```
par(mfrow=c(2,2))
plot(mylm)
```




```
hist(mylm$residuals, main = "Histogram of Residuals", xlab= "", col= "lightblue")
```



- ***Residuals vs Fitted - residuals are uniformly scattered around 0***
- ***Normal Q-Q - we see a stair step that deviates indicated line with obvious non linearity***
- ***Residual Historgram - the residial histogram is not normally distributed and skews right***