Project 2

Code

For Project 2 I will be testing a data set I created in a previous computer science class. I gathered this information from Pokemon website and extracted the information into a csv file on Excel. After cleaning and removing less relevant data and unecessary characters I was still left with a few NAs but I will be removing them here in R leaving plenty of data to use. My dataset is information about all Pokemon that have been released up to ~2020.

Hide

```
library(tidyr)
library(readr)
library(mosaic)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2

The 'mosaic' package masks several functions from core packages in order to add
additional features.  The original behavior of these functions should not be affected by this.

Attaching package: ‘mosaic’

The following objects are masked from ‘package:dplyr’:

    count, do, tally

The following object is masked from ‘package:Matrix’:

    mean

The following object is masked from ‘package:ggplot2’:

    stat

The following objects are masked from ‘package:stats’:

    binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test, quantile, sd, t.test, var

The following objects are masked from ‘package:base’:

    max, mean, min, prod, range, sample, sum
```

Hide

```
library(yarrr)
```

```
Loading required package: jpeg
Loading required package: BayesFactor
Loading required package: coda
************
Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).

Type BFManual() to open the manual.
************

Attaching package: ‘BayesFactor’

The following object is masked from ‘package:mosaic’:

    compare

Loading required package: circlize
========================================
circlize version 0.4.15
CRAN page: https://cran.r-project.org/package=circlize
Github page: https://github.com/jokergoo/circlize
Documentation: https://jokergoo.github.io/circlize_book/book/

If you use it in published research, please cite:
Gu, Z. circlize implements and enhances circular visualization
  in R. Bioinformatics 2014.

This message can be suppressed by:
  suppressPackageStartupMessages(library(circlize))
========================================

yarrr v0.1.5. Citation info at citation('yarrr'). Package guide at yarrr.guide()
Email me at Nathaniel.D.Phillips.is@gmail.com

Attaching package: ‘yarrr’

The following object is masked from ‘package:ggplot2’:

    diamonds
```

Hide

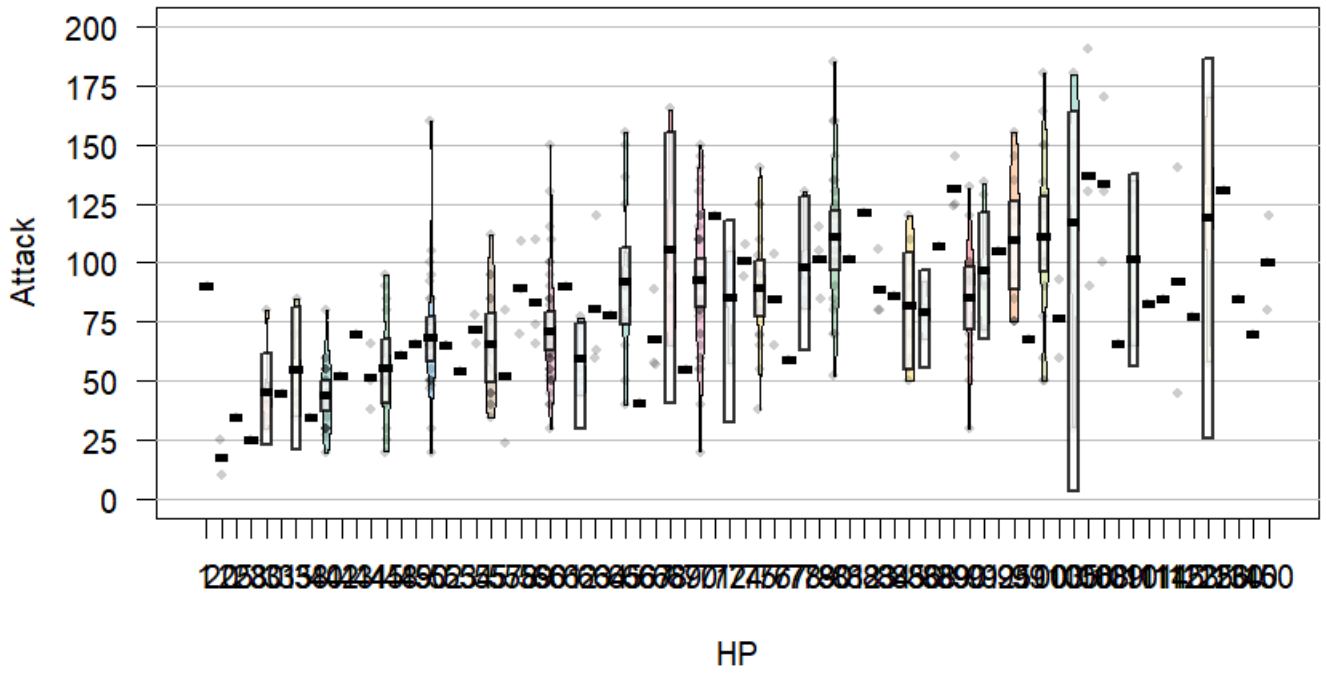
```
poke <- Pokemon %>% na.omit()
poke
```

# <dbl>	Name <chr>	Type 1 <chr>	Type 2 <chr>	Total <dbl>	HP <dbl>	Attack <dbl>	Defense <dbl>	Sp. Atk <dbl>	Sp. Def <dbl>
39	Jigglypuff	Normal	Fairy	270	115	45	20	45	25
40	Wigglytuff	Normal	Fairy	435	140	70	45	85	50
41	Zubat	Poison	Flying	245	40	45	35	30	40
42	Golbat	Poison	Flying	455	75	80	70	65	75
43	Oddish	Grass	Poison	320	45	50	55	75	65
44	Gloom	Grass	Poison	395	60	65	70	85	75
45	Vileplume	Grass	Poison	490	75	80	85	110	90
46	Paras	Bug	Grass	285	35	70	55	45	55
47	Parasect	Bug	Grass	405	60	95	80	60	80
48	Venonat	Bug	Poison	305	60	55	50	40	55

I would like to observe the two quantitative columns: HP and Attack. HP (response variable) describes the health points that each named Pokemon start at and Attack (explanatory variable) is there according attack power. I find it very interesting that from the looks of our plots there is somewhat of a linear relationship that may be happening here!

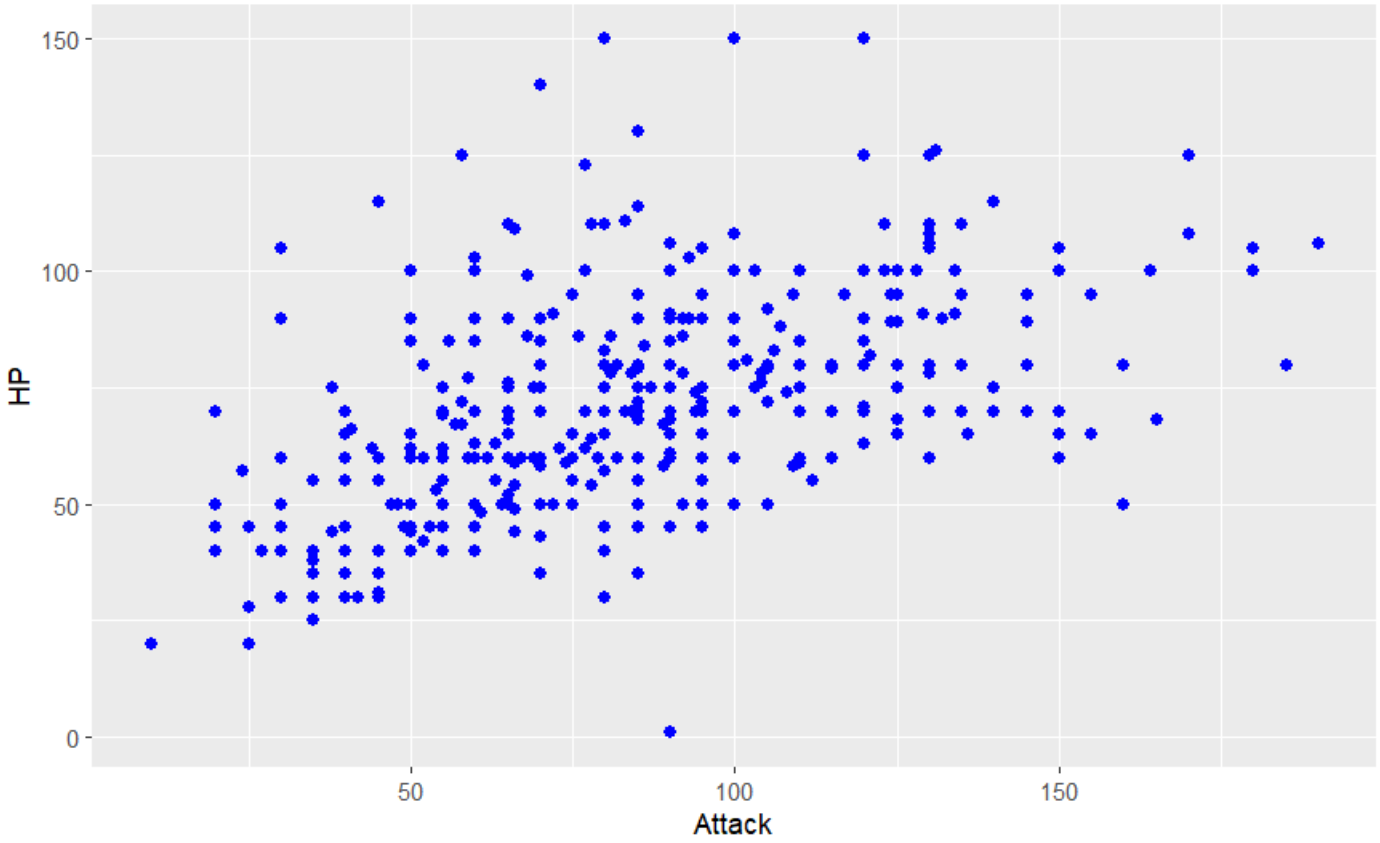
Hide

```
pirateplot(Attack ~ HP, data = poke)
```



Hide

```
ggplot() + geom_point( mapping = aes(x = Attack, y = HP ), data = poke, color = 'blue', size = 2 )
```

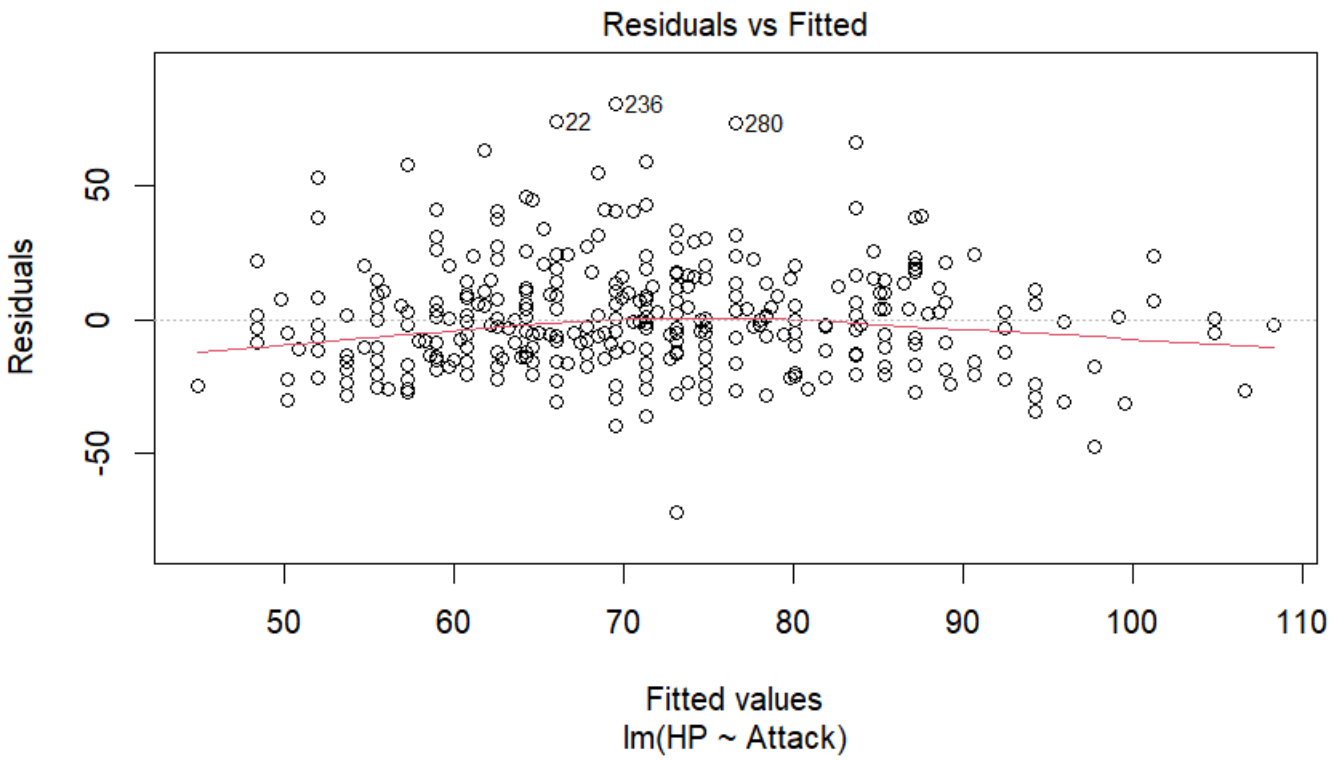


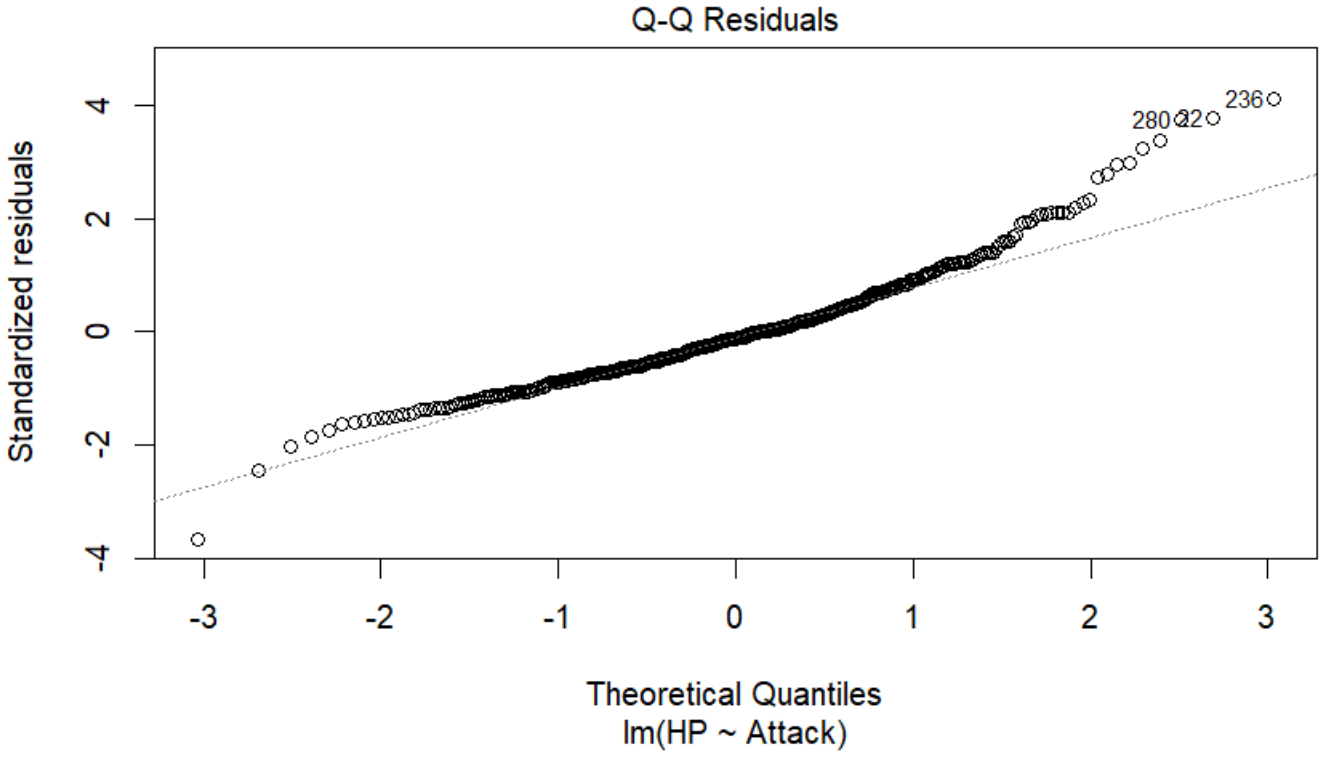
The four Assumptions for Linear Regression

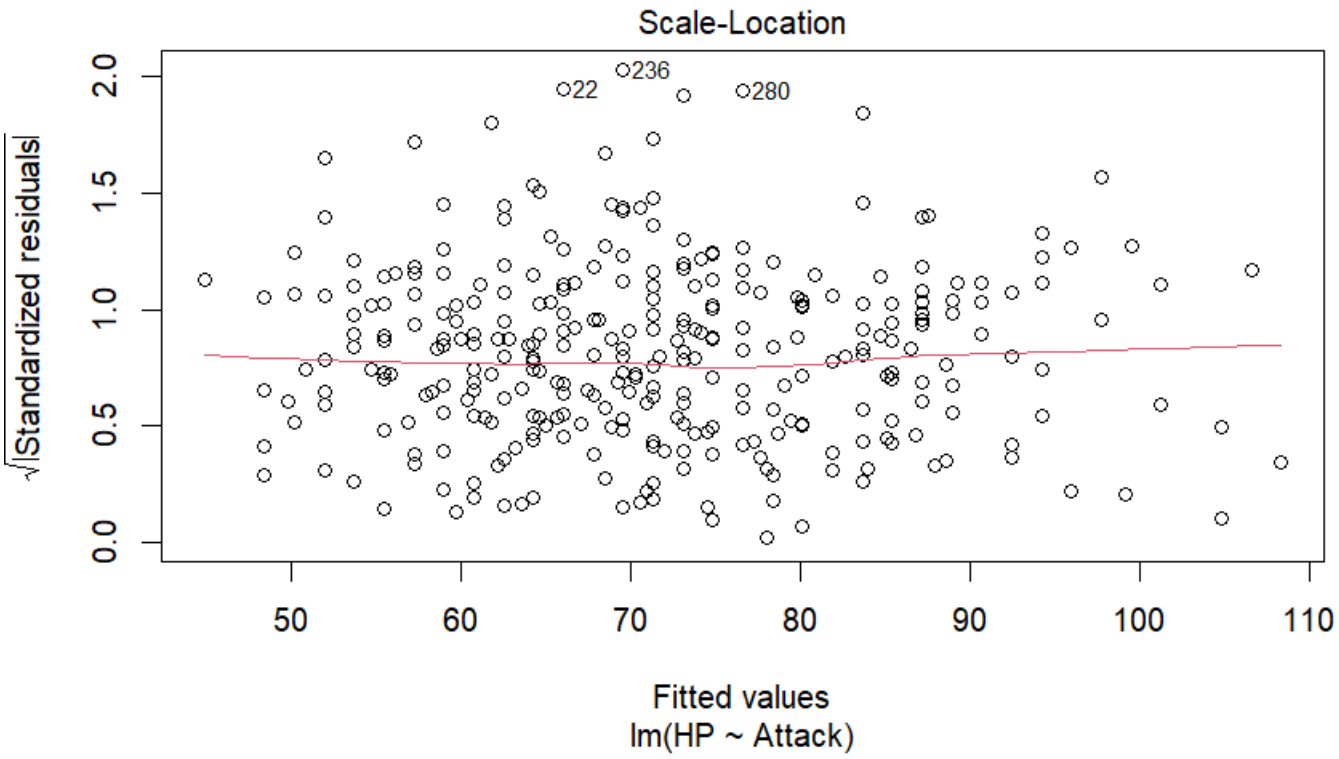
1. First we consider the Normality Assumption. Given the plot above we can see that there is a normal distribution of data points on the plot.
2. Next we can consider the Constant Variance Assumption. There are few outliers and the span has consistent variety throughout (as seen in pirate plot).
3. (The Independence Assumption) My y and x quantitative variables I am using are indeed independent and do not rely on each other. HP is a response variable to the Pokemon and Attack is a explanatory variable of the pokemon.
4. There are no points in the data that are more valuable or influential than the other points thus the model for the mean is correct. We are only looking at one group and that is pokemon.

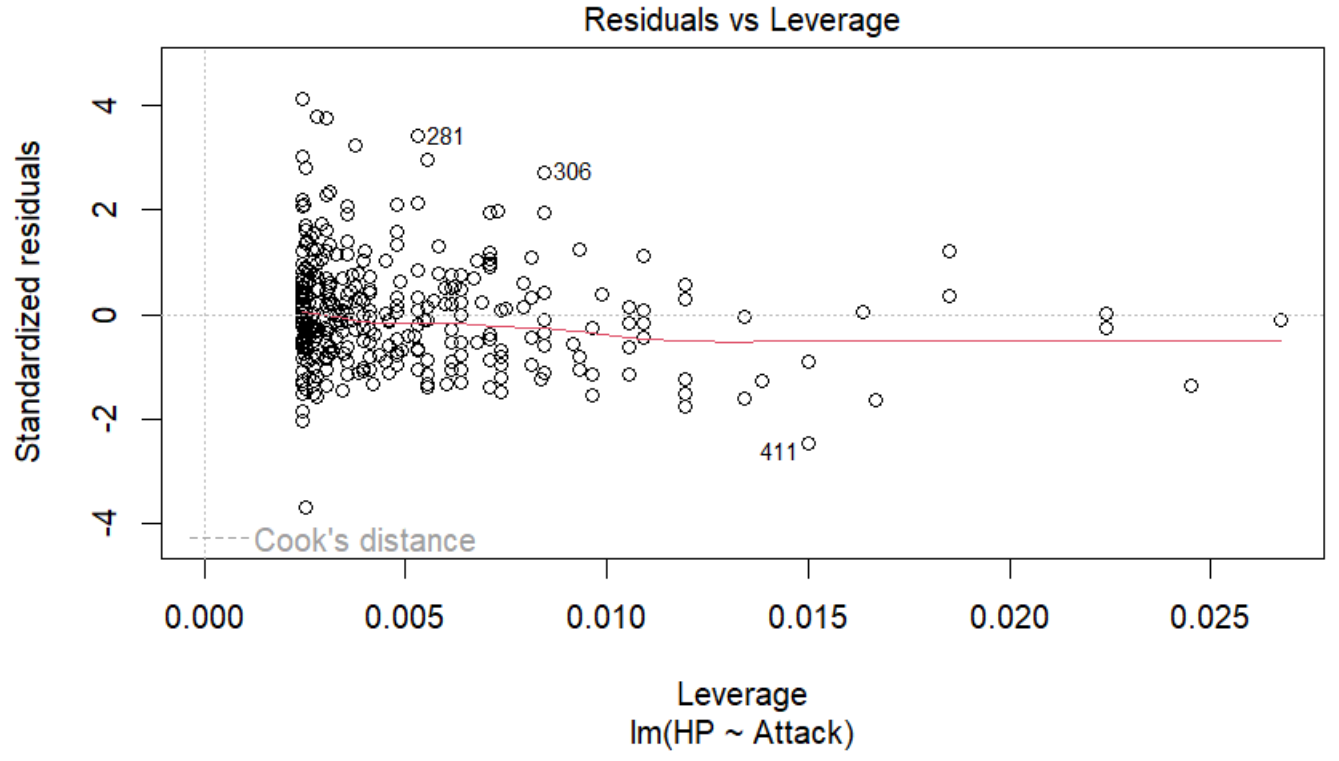
Hide

```
lm_poke <- lm( HP ~ Attack, data = poke)
plot(lm_poke)
```


Residuals vs Fitted
Residuals
Fitted values
lm(HP ~ Attack)
22
236
280
50
0
-50
50
60
70
80
90
100
110


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(HP ~ Attack)
280
22
236
4
2
0
-2
-4
-3
-2
-1
0
1
2
3





Hide

```
summary(lm_poke)
```

```
Call:
lm(formula = HP ~ Attack, data = poke)

Residuals:
    Min      1Q  Median      3Q     Max
-72.056 -13.623  -2.391   9.725  80.469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.33617    2.56786   16.10   <2e-16 ***
Attack       0.35244    0.02862   12.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.6 on 412 degrees of freedom
Multiple R-squared:  0.2691,    Adjusted R-squared:  0.2673
F-statistic: 151.7 on 1 and 412 DF,  p-value: < 2.2e-16
```

- H0 (Null Hypothesis): The slope m is 0. There is no LINEAR effect of the level of Attack on describing the HP of any given pokemon.
- HA (Alternative Hypothesis): The slope m > 0. I.e. There is a non-constant linear relationship between level of Attack and HP of pokemon.

Our summary shows us that there is no linear effect happening because our R-Squared value is 0.2691 which is not close to one and that does not represent a strong linear model.

ANOVA I am going to conduct first a simple ANOVA (one way) test then a two way to gather more information about our data set. Assumptions:

- Independence comes from the experiment and in this case the important notation is that each pokemon is different and there are no multiples.
- Constant Variance works for this data set because all pokemon stats are relatively in the same ball park.

I need an explanatory variable with two or more values. I will make an interaction between type and speed of pokemon.

Hide

```
summary(poke$Speed)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   50.00   70.00   70.51   91.00  160.00
```

Hide

```
tally(poke$`Type 1`)
```

```
X
     Bug     Dark   Dragon Electric    Fairy Fighting     Fire   Flying    Ghost    Grass   Ground      Ice   Normal
      52       21       21       17        2        7       24        2       22       37       19       11       37
  Poison  Psychic     Rock    Steel    Water
      13       19       35       22       53
```

Hide

```
poke_test <- poke %>% mutate(TypesSpeed = interaction(poke$`Type 1`, Speed))
```

Hide

```
summary(poke_test$TypesSpeed)
```

```
   Water.70     Water.60       Bug.40       Bug.65  Electric.86    Dragon.95     Steel.30     Water.30     Grass.40
          7            6            5            5            5            5            4            4            4
   Steel.50     Water.50     Grass.55      Rock.70     Grass.80     Fire.100   Dragon.110     Grass.30       Bug.36
          4            4            4            4            4            4            4            3            3
  Normal.50      Rock.50      Fire.55     Water.55     Grass.60    Normal.60       Bug.75    Dragon.80     Ghost.80
          3            3            3            3            3            3            3            3            3
 Psychic.80     Water.85   Normal.100  Psychic.100     Water.15    Normal.20       Bug.30      Rock.30      Rock.35
          3            3            3            3            2            2            2            2            2
   Water.35    Ground.40     Water.40       Bug.42       Bug.45  Electric.45      Rock.45     Grass.50       Ice.50
          2            2            2            2            2            2            2            2            2
     Bug.55    Ground.55      Rock.55     Ghost.56      Dark.58       Bug.60      Dark.60     Steel.60       Ice.65
          2            2            2            2            2            2            2            2            2
   Water.65     Water.67       Bug.70  Electric.70     Grass.70    Normal.70   Psychic.70     Steel.70    Normal.71
          2            2            2            2            2            2            2            2            2
    Rock.71    Normal.75    Normal.80     Water.81       Bug.85    Normal.85    Poison.85    Dragon.90      Fire.90
          2            2            2            2            2            2            2            2            2
   Ghost.90       Bug.95    Water.100 Electric.101      Bug.105    Water.108  Psychic.110     Rock.110     Dark.115
          2            2            2            2            2            2            2            2            2
      Bug.5     Grass.10     Grass.15    Normal.15       Bug.20      Dark.20      Fire.20     Ghost.20     Grass.20
          1            1            1            1            1            1            1            1            1
    Rock.20     Water.22      Rock.23     Steel.23       Bug.25    Ground.25       Ice.25     Steel.28      Fire.30
          1            1            1            1            1            1            1            1            1
    (Other)
        178
```

Hide

```
lm_1 <- lm( HP ~ TypesSpeed , data = poke_test) #  * means include the interaction term
anova(lm_1)
```

```
Analysis of Variance Table

Response: HP
            Df Sum Sq Mean Sq F value  Pr(>F)
TypesSpeed 276 156316  566.36  1.2889 0.04723 *
Residuals  137  60200  439.42
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- H0 (Null Hypothesis): The TypesSpeed interaction in our data set is significant to the HP of the pokemon.
- HA (Alternative Hypothesis): TypesSpeed interaction is not significant to the HP of the pokemon.

  Our one way ANOVA test shows that there is a p value of 0.04723, we want 95% accuracy,this is just enough to accept the null hypothesis. The Types and Speed of the pokemon are significant to the HP of the pokemon, however, we don’t know in which ways it is significant but we can see that it makes a significant difference.

Parametric Version:

Hide

```
Tobs_2 <- anova(lm_1)
Tobs_2
```

```
Analysis of Variance Table

Response: HP
            Df Sum Sq Mean Sq F value  Pr(>F)
TypesSpeed 276 156316  566.36  1.2889 0.04723 *
Residuals  137  60200  439.42
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
Tobs_2[1,4]
```

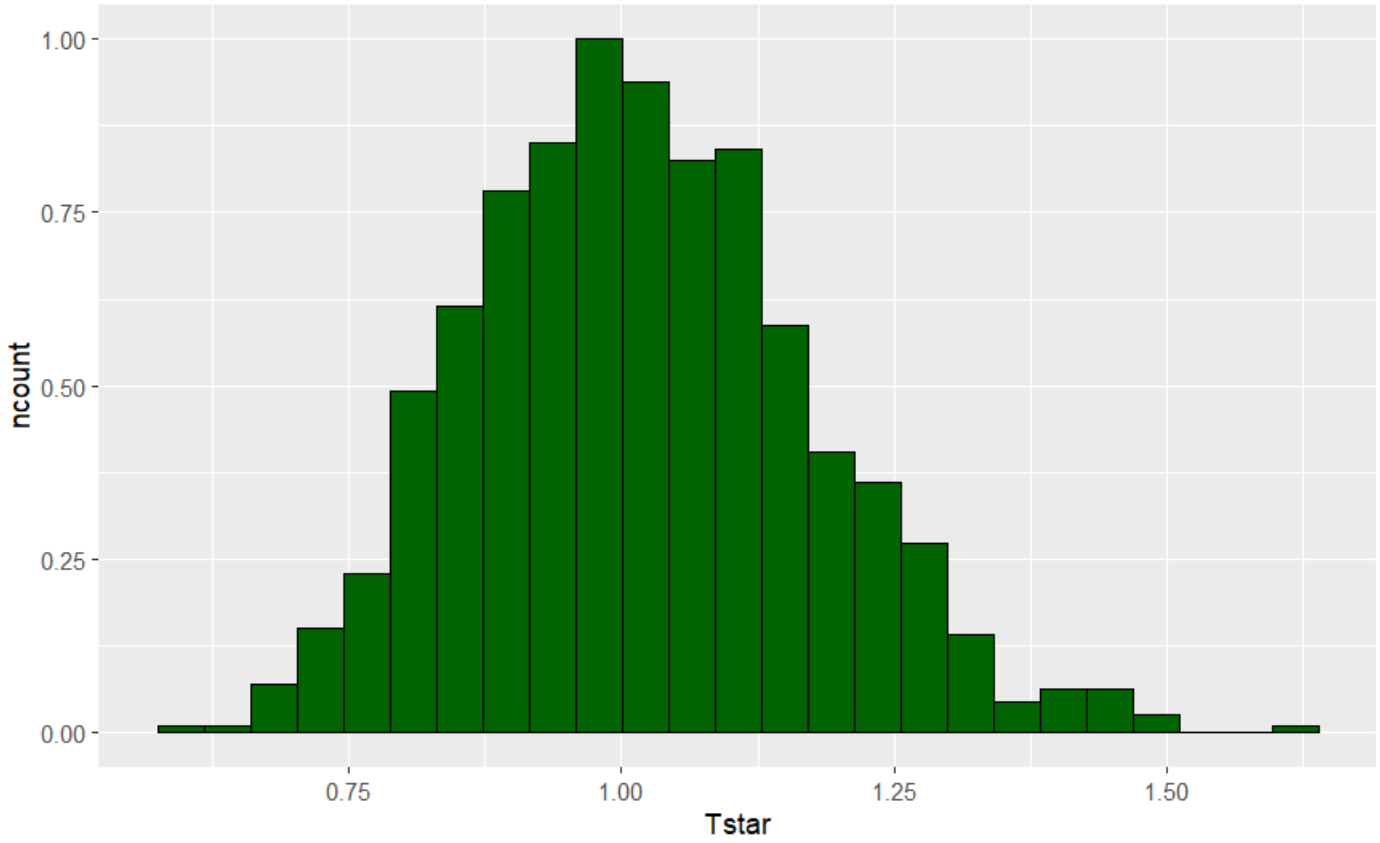
```
[1] 1.288894
```

Hide

```
N <- 1000
Tstar <- matrix(NA, nrow = N)
for (b in (1:N)){
  Tstar[b] <- anova(lm(HP ~ shuffle(TypesSpeed), data=poke_test))[1,4]
}
```

Hide

```
tibble(Tstar) %>% ggplot(aes(x = Tstar)) +
  geom_histogram(aes(y = ..ncount..), bins=25, col=1, fill='darkgreen')
```



Most of our results from this test get around 1.288894, as seen in our anova test. We also get a nice bell curve, indicating that there is significance.

Prediction and Confidence Intervals: (for HP and Attack):

Hide

```
N <- 1000
S <- 50
Rstar_poke <- matrix(NA, nrow = N)
for (b in 1:N){
  dfs <- sample(poke, size = S, replace = TRUE)
  lm_t <- lm(HP ~ Attack, data = dfs)
  Rstar_poke[b] = rsquared(lm_t)
}
```

Hide

```
tibble(Rstar_poke) %>% ggplot(aes(x=Rstar_poke)) + geom_histogram(aes( y = after_stat(ncount)), bins=20, col=1, fill='skyb
lue' ) +
   geom_vline(xintercept = quantile(Rstar_poke, c(0.05, 0.95)), col="red", lwd=2)
```

Hide

```
qdata(Rstar_poke, c(0.05, 0.95))
```

Prediction and Confidence Intervals: (for HP and Defense):

Hide

```
N <- 1000
S <- 50
Rstar_poke2 <- matrix(NA, nrow = N)
for (b in 1:N){
  dfs2 <- sample(poke, size = S, replace = TRUE)
  lm_t2 <- lm(HP ~ Defense, data = dfs2)
  Rstar_poke2[b] = rsquared(lm_t2)
}
```

Hide

```
tibble(Rstar_poke2) %>% ggplot(aes(x=Rstar_poke2)) + geom_histogram(aes( y = after_stat(ncount)), bins=20, col=1, fill='sk
yblue' ) +
   geom_vline(xintercept = quantile(Rstar_poke2, c(0.05, 0.95)), col="red", lwd=2)
```

Hide

```
qdata(Rstar_poke2, c(0.05, 0.95))
```

Prediction and Confidence Intervals: (for HP and Speed):

Hide

```
N <- 1000
S <- 50
Rstar_poke3 <- matrix(NA, nrow = N)
for (b in 1:N){
  dfs3 <- sample(poke, size = S, replace = TRUE)
  lm_t3 <- lm(HP ~ Speed, data = dfs3)
  Rstar_poke3[b] = rsquared(lm_t3)
}
```

Hide

```
tibble(Rstar_poke3) %>% ggplot(aes(x=Rstar_poke3)) + geom_histogram(aes( y = after_stat(ncount)), bins=20, col=1, fill='sk
yblue' ) +
   geom_vline(xintercept = quantile(Rstar_poke3, c(0.05, 0.95)), col="red", lwd=2)
```

Hide

```
qdata(Rstar_poke2, c(0.05, 0.95))
```