OpenIntro Chapter 8 Lab

Daniel Jackson

October 13th, 2023

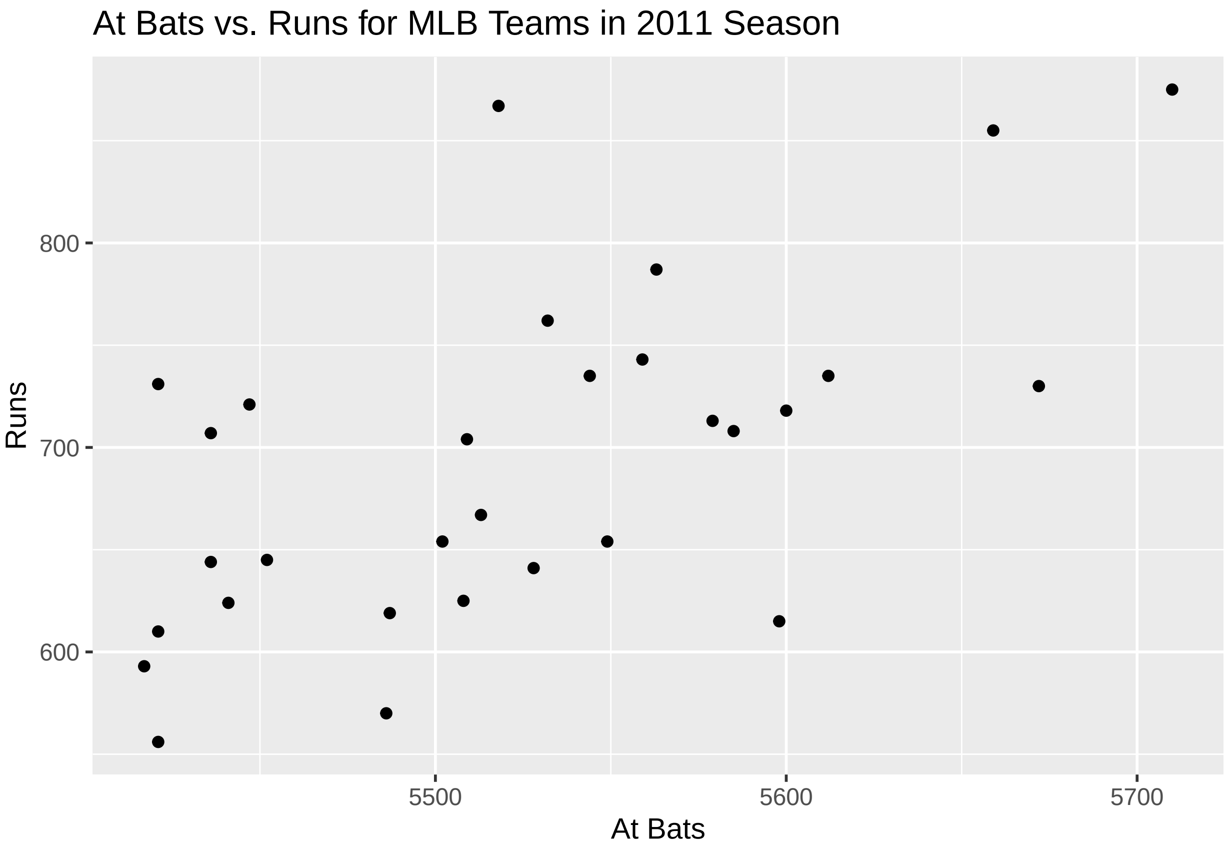
Read in data / install libraries

```r
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")

library(ggplot2)
```

Exercise 1

```r
ggplot(mlb11,
       mapping = aes(x = at_bats,
                     y = runs)) +
  geom_point() +
  labs(x = 'At Bats',
       y = 'Runs',
       title = "At Bats vs. Runs for MLB Teams in 2011 Season")
```



This plot has a relatively linear pattern. There seem to be a few outliers, but for the most part, you could say this relationship is linear. I would be comfortable using linear a model to predict runs based off knowing at bats. The more at-bats a team has in a season, the more they should be getting on base and the more runs they should be scoring.

```
cor(mlb11$runs, mlb11$at_bats)
```

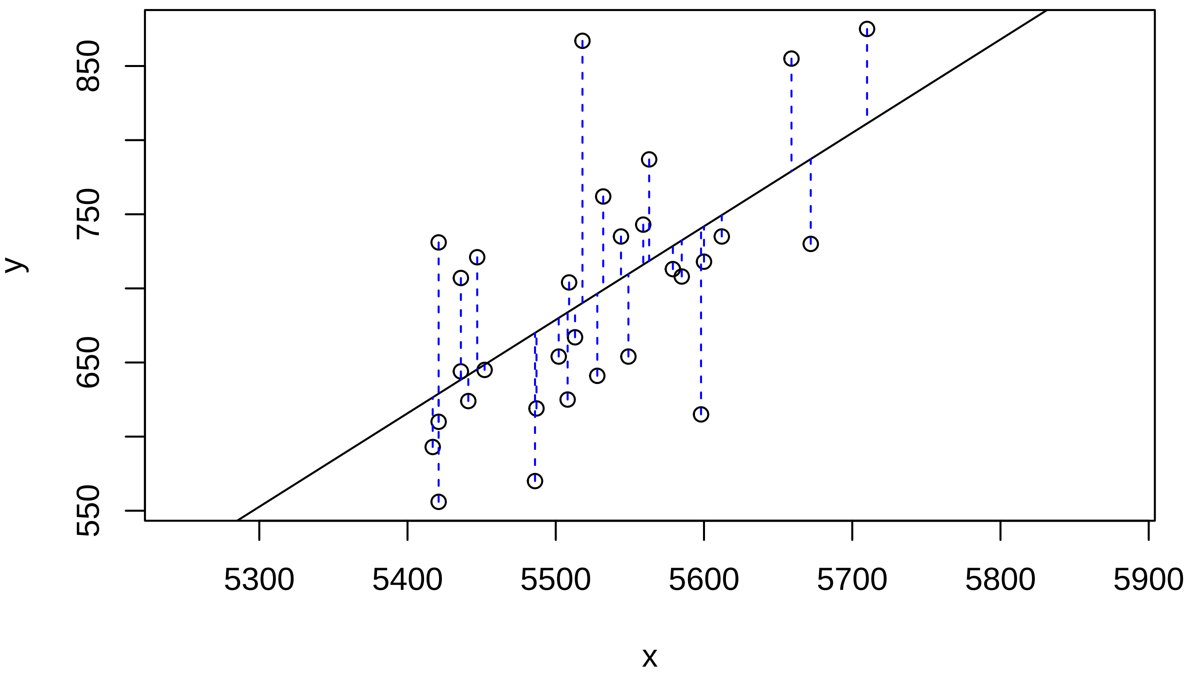
```
## [1] 0.610627
```

Returns 0.6106 = 61.6%. Not the most positive linear relationship, but enough to be considered linear.

Exercise 2

A majority of the data follows a positive linear realtionship. As there are more at-bats there seem to be more runs being scored by those teams. There are a few outliers in the data.I would not say that this plot has the strongest linear relationship, which can be seen by the correlation coefficient only being approximately 62%. There is not a significant constant variability in the data, but enought to try and create linear model.

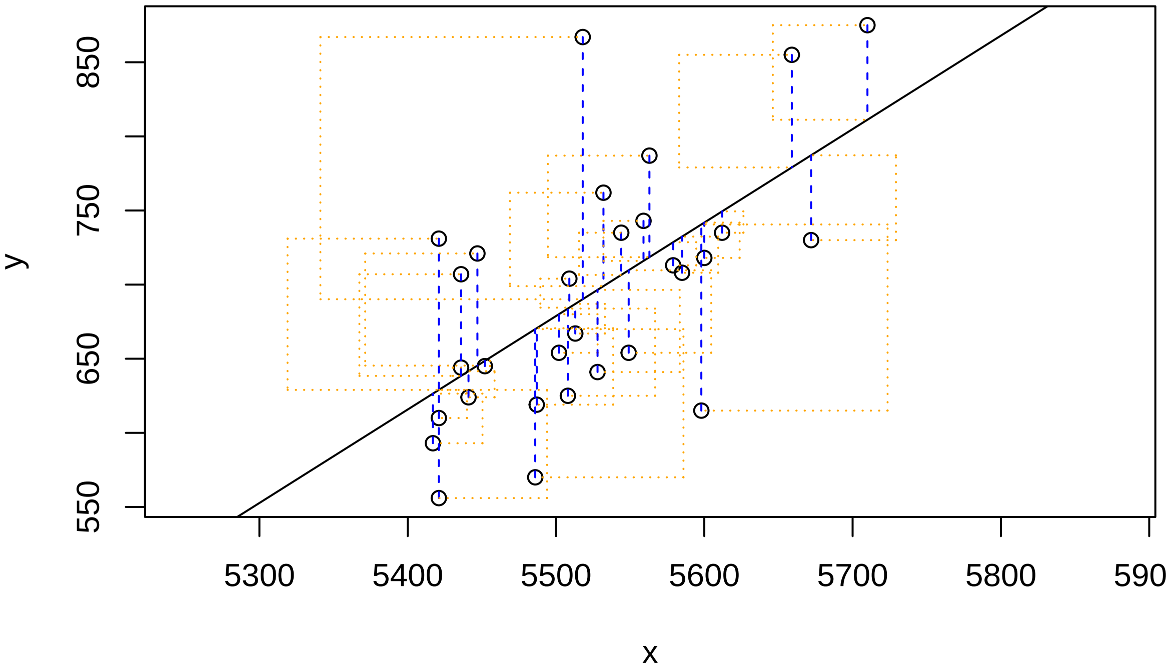
```
plot_ss(x = mlb11$at_bats, y = mlb11$runs)
```



```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##  -2789.2429       0.6305
##
## Sum of Squares:  123721.9
```

Exercise 3

```
plot_ss(x = mlb11$at_bats, y = mlb11$runs, showSquares = TRUE)
```



```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##  -2789.2429       0.6305
##
## Sum of Squares:  123721.9
```

Sum of squares: Test 1: 182980.4 Test 2: 163291.9 Test 3: 127231.1 Test 4: 124658.3 Test 5: 129166.0

Test 4 had least sum of squares.

Exercise 4

```
m1 <- lm(runs ~ at_bats, data = mlb11)
summary(m1)
```

```
##
## Call:
## lm(formula = runs ~ at_bats, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2789.2429   853.6957  -3.267 0.002871 **
## at_bats         0.6305     0.1545   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

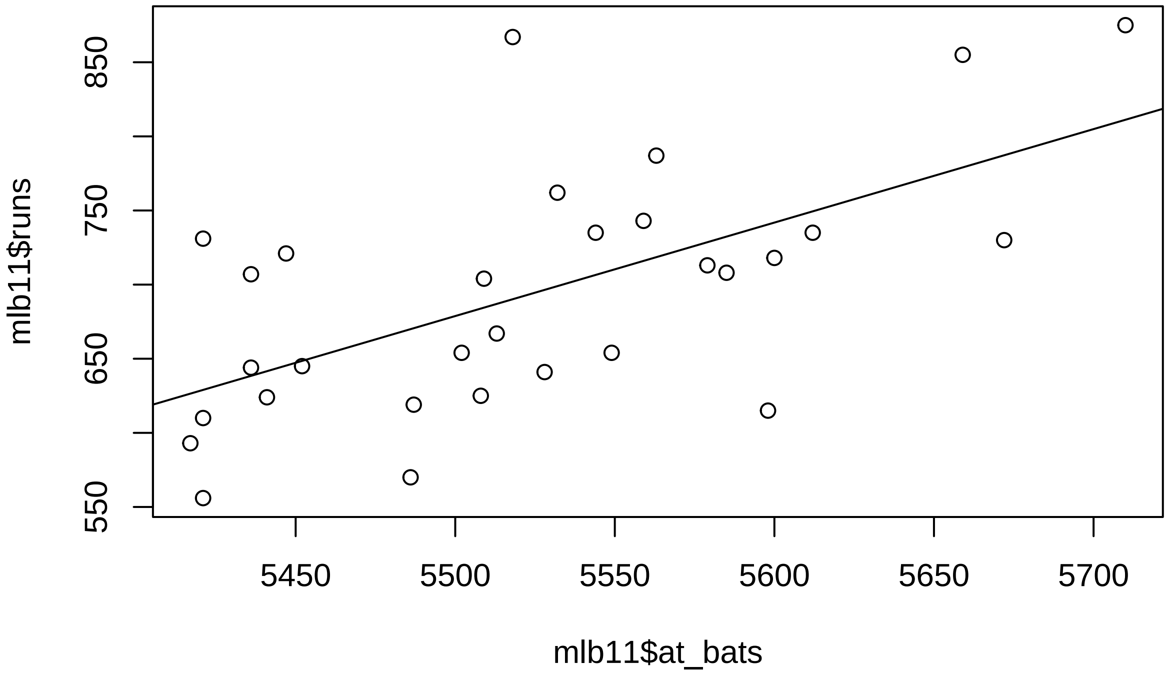
```r
m2 <- lm(runs ~ homeruns, data = mlb11)
summary(m2)
```

```
##
## Call:
## lm(formula = runs ~ homeruns, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.615 -33.410   3.231  24.292 104.631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 415.2389    41.6779   9.963 1.04e-10 ***
## homeruns      1.8345     0.2677   6.854 1.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.29 on 28 degrees of freedom
## Multiple R-squared:  0.6266, Adjusted R-squared:  0.6132
## F-statistic: 46.98 on 1 and 28 DF,  p-value: 1.9e-07
```

Linear model line: runs-hat = 415.2389 + 1.8345 X homeruns The positive slope shows that there is a positive relationship between homeruns and runs. The more homeruns you hit, the more runs you will score. For every homerun hit, on average you will have 1.8345 more runs.

Exercise 5

```r
plot(mlb11$runs ~ mlb11$at_bats)
abline(m1)
```



For 5578 at-bats:

runs-hat = -2789.2429 + 0.6305 X 5578

-2789.2429 + (0.6305 * 5578)

runs-hat = 717.6861, approx 718 runs.

Point estimate: (5578, 718)

In data set, Phillies had 5579 at-bats (closest to 5578 estimate) and 713 runs.

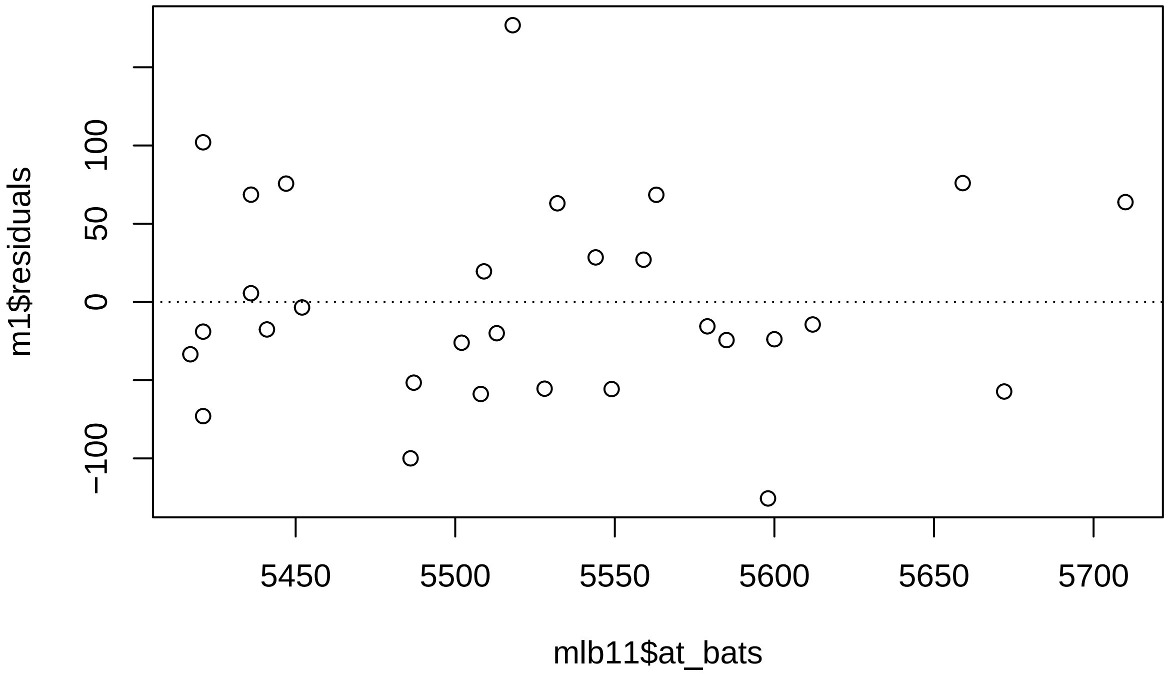
residual = observed runs - predicted runs

713 - 718

Residual = -5. Since residual of observed value is negative, our estimate was an over estimate based on the observed data.

Exercise 6

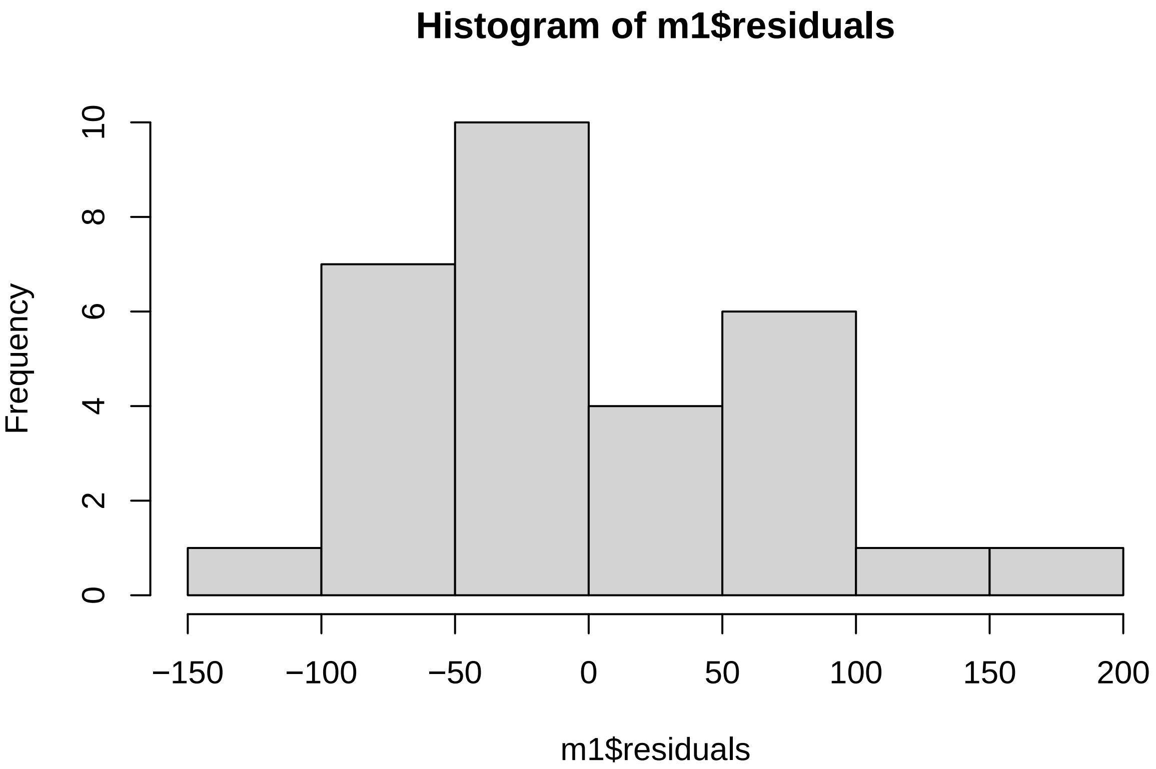
```
plot(m1$residuals ~ mlb11$at_bats)
abline(h = 0, lty = 3)
```



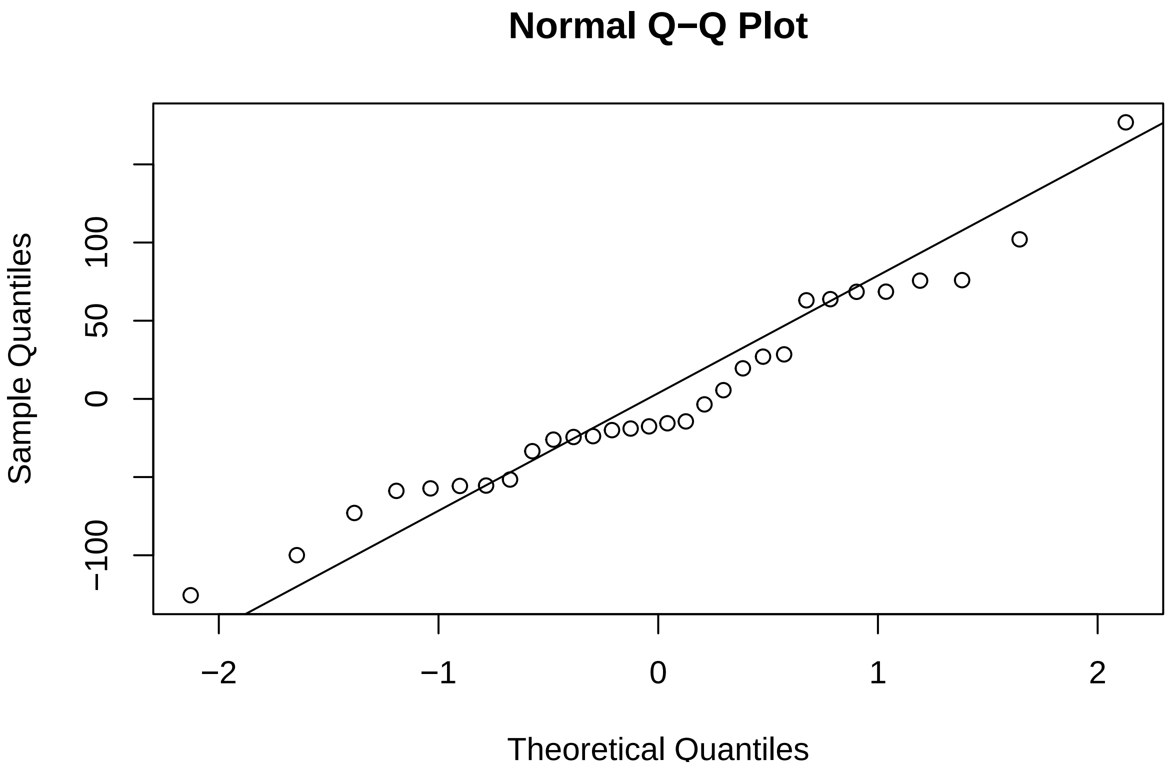
There is no major shape or curve in the data and all of the residuals seem to be hovering around the cloud around residual axis equaling 0. This tells us that the relationship between at-bats and runs is a linear one.

Exercise 7

```
hist(m1$residuals)
```



```
qqnorm(m1$residuals)
qqline(m1$residuals)
```



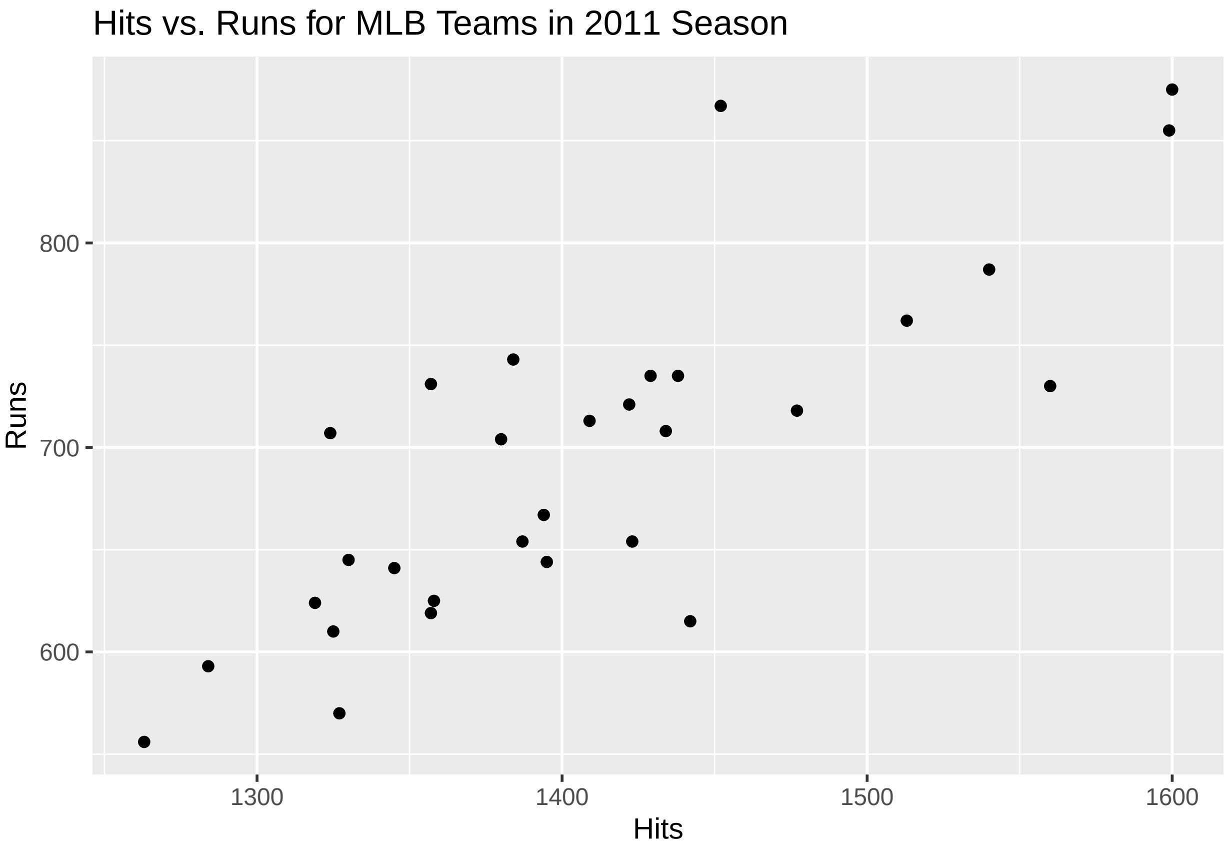
The residuals seem to follow a nearly normal distribution based on the bell shape of the residual histogram. The normal probability plot of the residuals also shows linear relationship, which also confirms nearly normal distribution of the residuals.

Exercise 8

Yes, there is enough constant variability in the plot to determine a linear relationship.

Own Your Own Question 1

```
ggplot(mlb11,
       mapping = aes(x = hits,
                     y = runs)) +
  geom_point() +
  labs(x = 'Hits',
       y = 'Runs',
       title = "Hits vs. Runs for MLB Teams in 2011 Season")
```



```
m3 <- lm(runs ~ hits, data = mlb11)
summary(m3)
```

```
##
## Call:
## lm(formula = runs ~ hits, data = mlb11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.718  -27.179   -5.233   19.322  140.693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -375.5600   151.1806  -2.484   0.0192 *
## hits           0.7589     0.1071   7.085 1.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.23 on 28 degrees of freedom
## Multiple R-squared:  0.6419, Adjusted R-squared:  0.6292
## F-statistic:  50.2 on 1 and 28 DF,  p-value: 1.043e-07
```

Yes, at first glance, there seems to be a positive linear relationship between hits and runs. The more hits you have, the more runs you score. For every hit you have, on average you will have 0.7589 more runs.

Linear model: runs-hat = -375.56 + 0.7589 X hits

Question 2

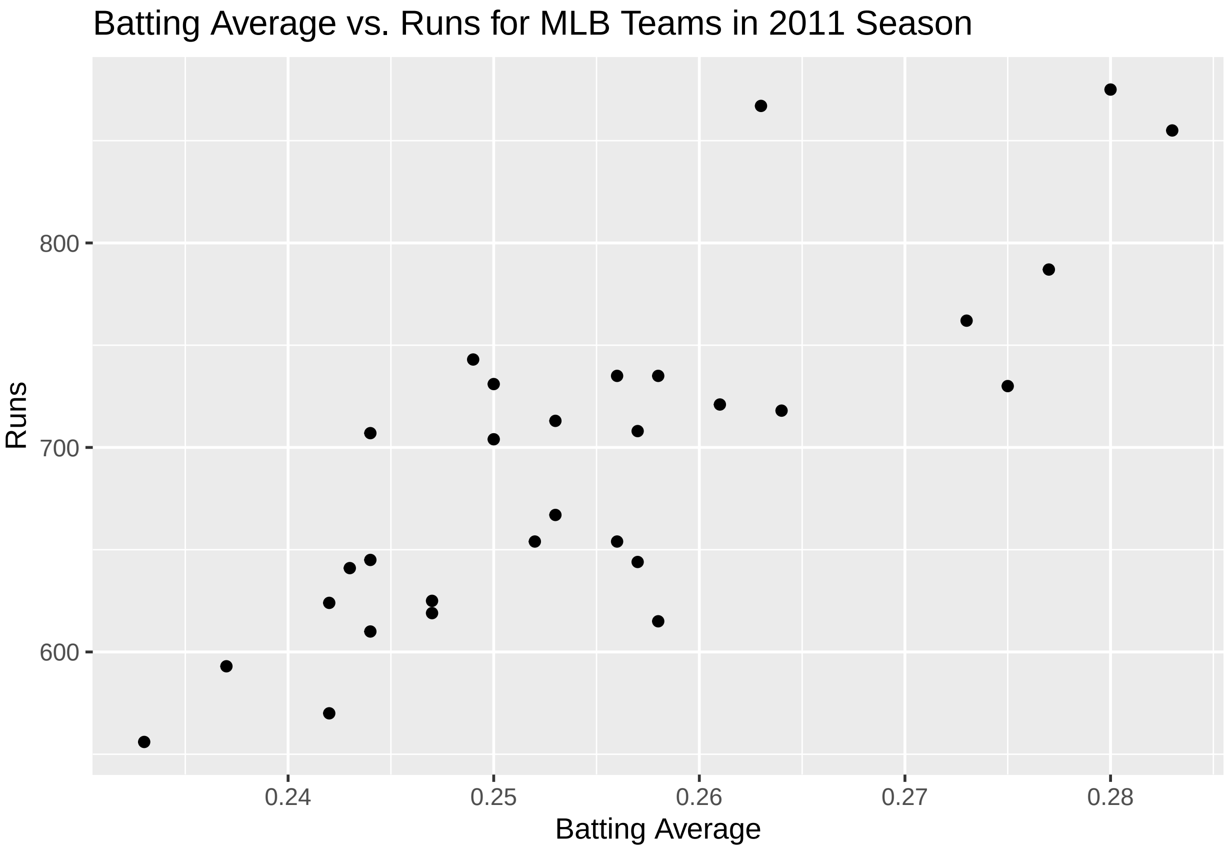
R-squared value for at-bats vs runs = 0.3729 R-squared value for hits vs runs = 0.6419 Since R-squared value for hits vs runs is greater than R-squared value for at-bats vs runs. Therefore, we can conclude that hits is a better predictor than at-bats to predict runs.

Question 3

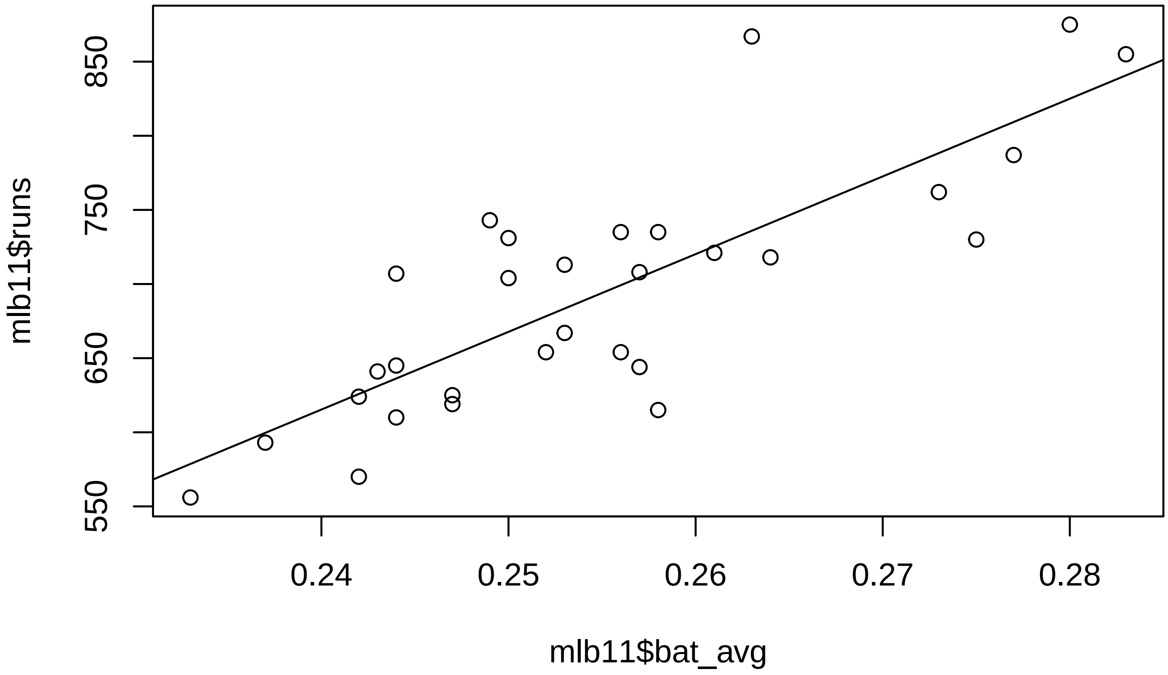
```r
m4 <- lm(runs ~ bat_avg, data = mlb11)
summary(m4)
```

```
##
## Call:
## lm(formula = runs ~ bat_avg, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.676 -26.303  -5.496  28.482 131.113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -642.8      183.1  -3.511  0.00153 **
## bat_avg       5242.2      717.3   7.308 5.88e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 49.23 on 28 degrees of freedom
## Multiple R-squared:  0.6561, Adjusted R-squared:  0.6438
## F-statistic: 53.41 on 1 and 28 DF,  p-value: 5.877e-08
```

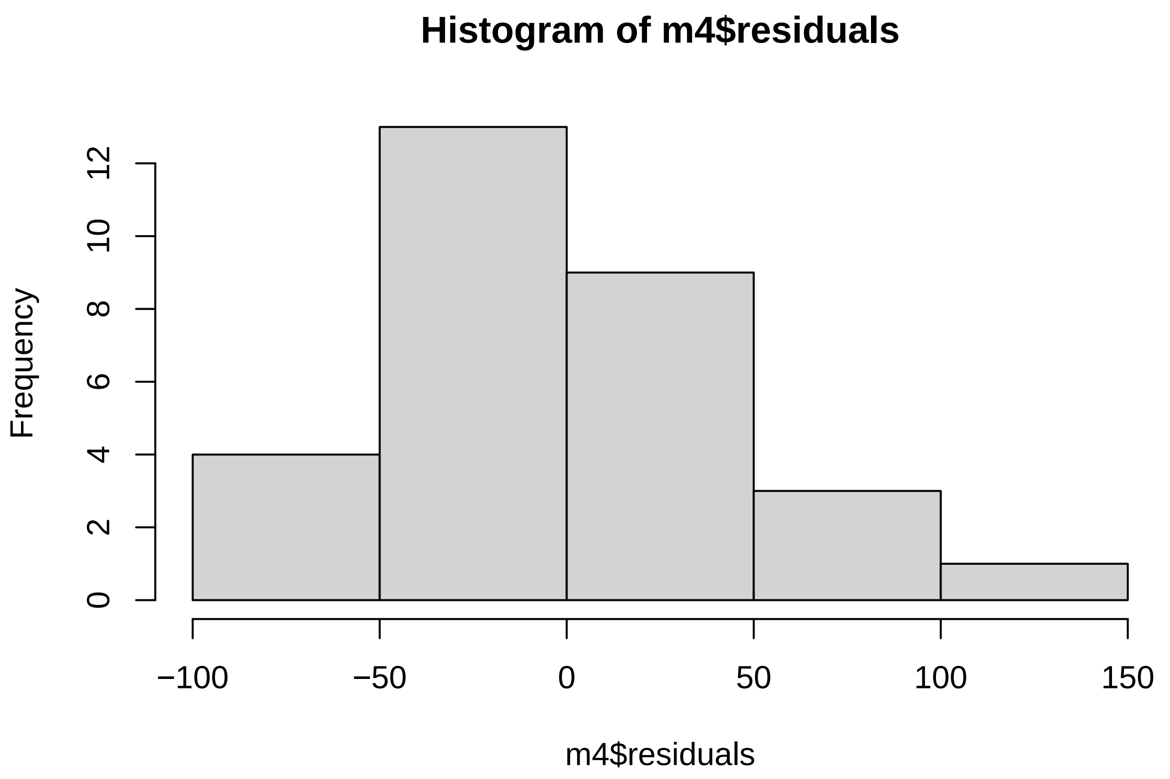
```r
ggplot(mlb11,
       mapping = aes(x = bat_avg,
                     y = runs)) +
  geom_point() +
  labs(x = 'Batting Average',
       y = 'Runs',
       title = "Batting Average vs. Runs for MLB Teams in 2011 Season")
```



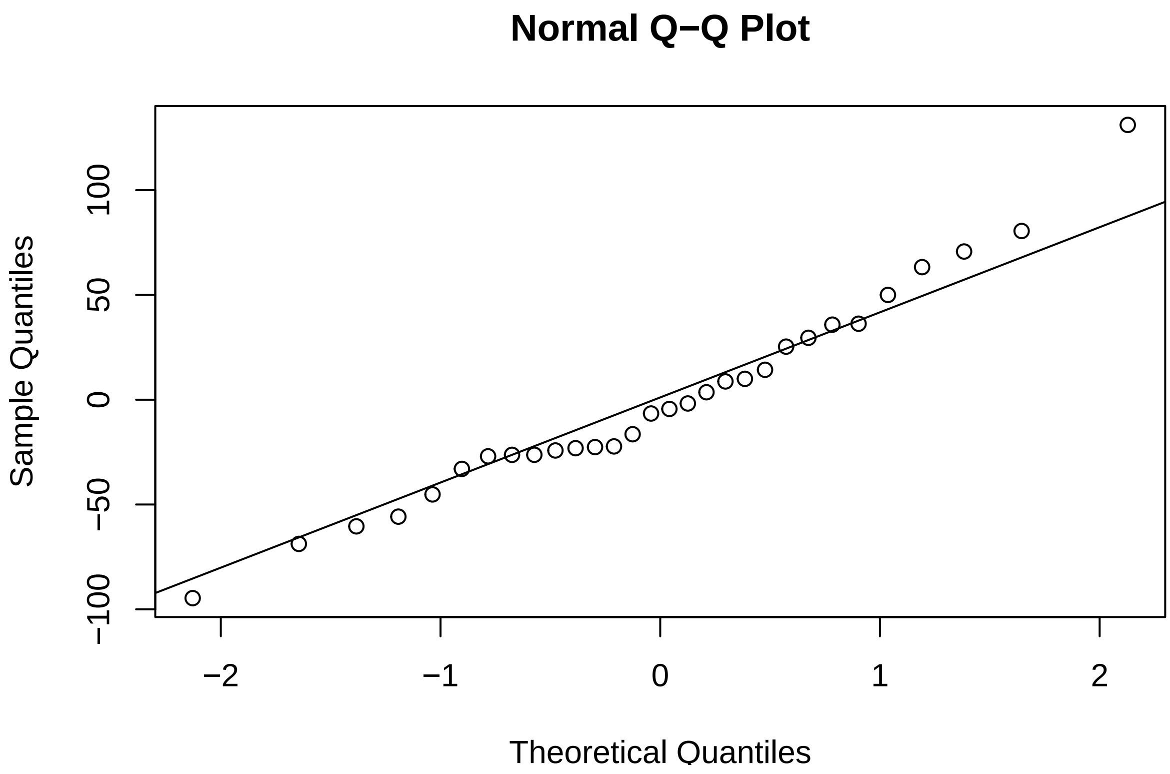
```r
plot(mlb11$runs ~ mlb11$bat_avg)
abline(m4)
```



```r
hist(m4$residuals)
```



```r
qqnorm(m4$residuals)
qqline(m4$residuals)
```



bat_avg best predicts runs as it has highest R-squared value (0.6561) among all of the predictors.

Question 4

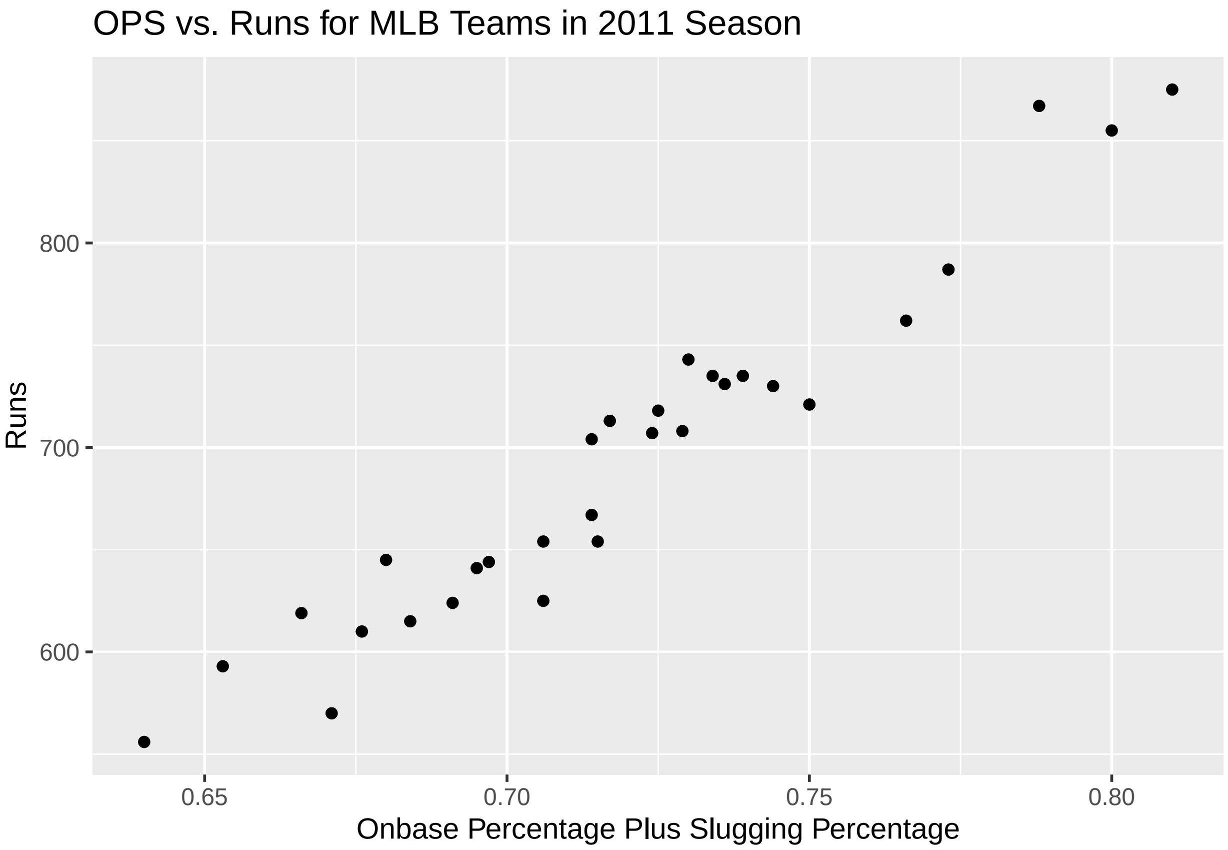
```
m5 <- lm(runs ~ new_obs, data = mlb11)
summary(m5)
```

```
##
## Call:
## lm(formula = runs ~ new_obs, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.456 -13.690   1.165  13.935  41.156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -686.61      68.93  -9.962 1.05e-10 ***
## new_obs      1919.36      95.70  20.057  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.41 on 28 degrees of freedom
## Multiple R-squared:  0.9349, Adjusted R-squared:  0.9326
## F-statistic: 402.3 on 1 and 28 DF,  p-value: < 2.2e-16
```

On base percentage, slugging percentage and OPS (on base percentage plus slugging percentage) are all better predictors compared to our other predictors. The R-squared values of these three predictors are all higher than the R-squared values of the other predictors.

OPS (onbase percentage plus slugging percentage) is the best predictor as it has the highest R-squared value of 0.9349

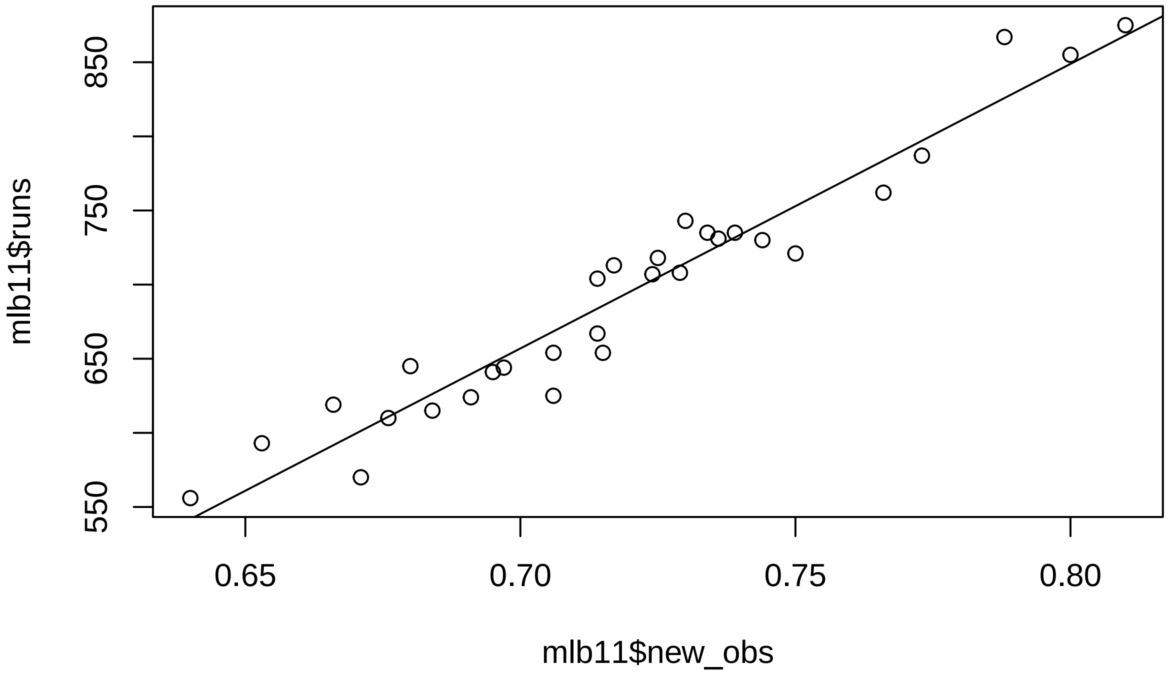
```
ggplot(mlb11,
       mapping = aes(x = new_obs,
                     y = runs)) +
  geom_point() +
  labs(x = 'Onbase Percentage Plus Slugging Percentage',
       y = 'Runs',
       title = "OPS vs. Runs for MLB Teams in 2011 Season")
```



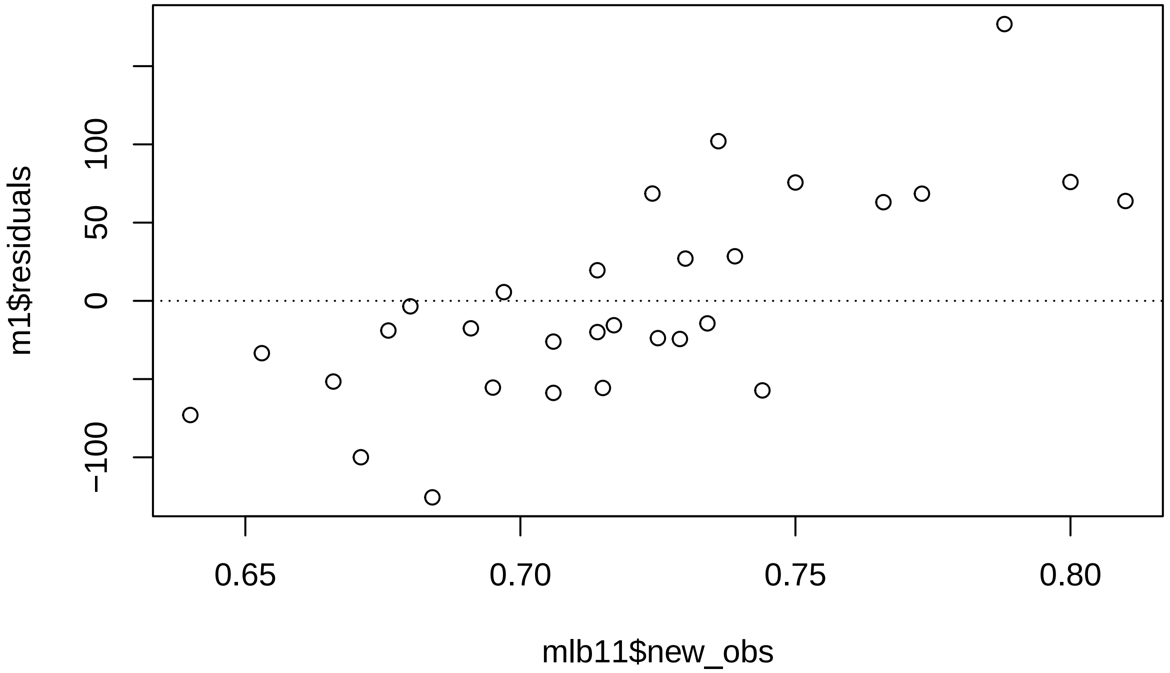
This result makes sense. On base percentage and slugging percentage (OPS) were the two best predictors in this dataset. OPS is on base percentage plus slugging percentage. Combining the two creates a better predictor of runs rather than looking at each individually. The best players and teams in Major League Baseball are judged off of how high their OPS is.

Question 5

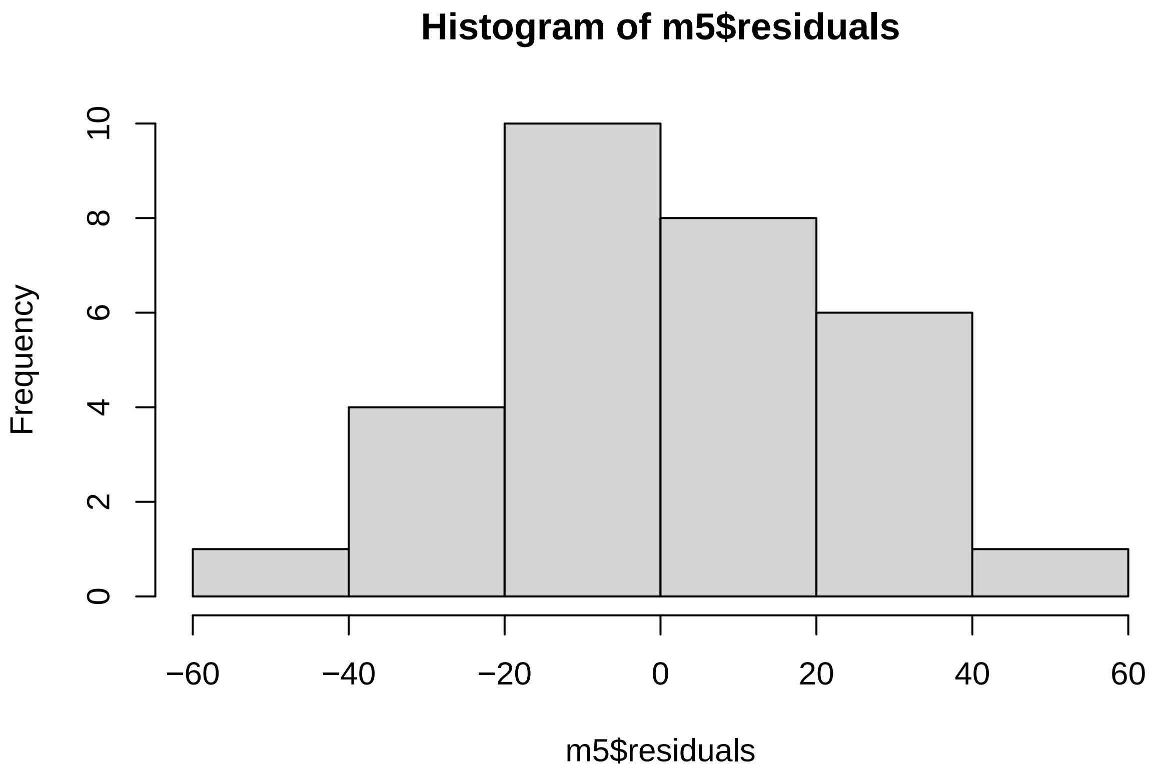
```
plot(mlb11$runs ~ mlb11$new_obs)
abline(m5)
```



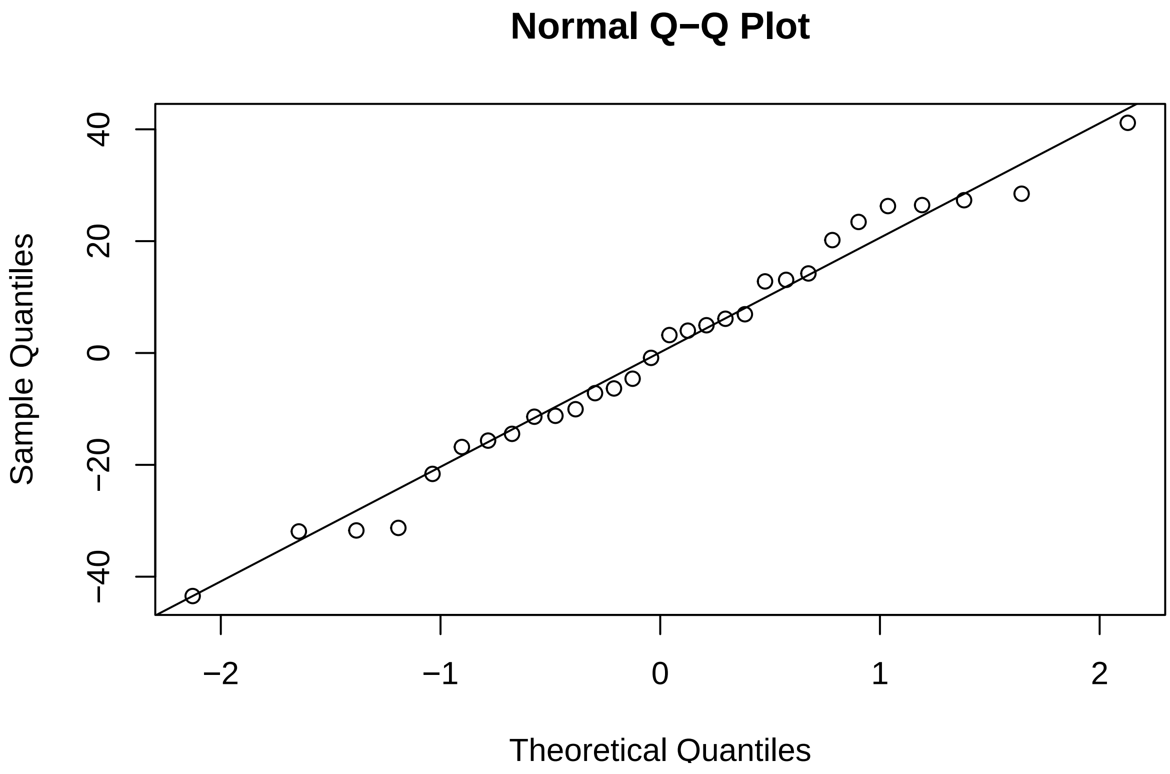
```
plot(m1$residuals ~ mlb11$new_obs)
abline(h = 0, lty = 3)
```



```
hist(m5$residuals)
```



```
qqnorm(m5$residuals)
qqline(m5$residuals)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles