Lin-Lab7

Bin Lin

2016-11-13

```
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

```
## The following object is masked from 'package:utils':
##
##     demo
```

```
startLab('Lab7')
```

```
## Warning: running command 'open 'C:/Users/blin261/Desktop/DATA606/Lab7/
## blin261-simple_regression.Rmd'' had status 34
```

```
## [1] "C:/Users/blin261/Desktop/DATA606/Lab7/blin261-simple_regression.Rmd"
```

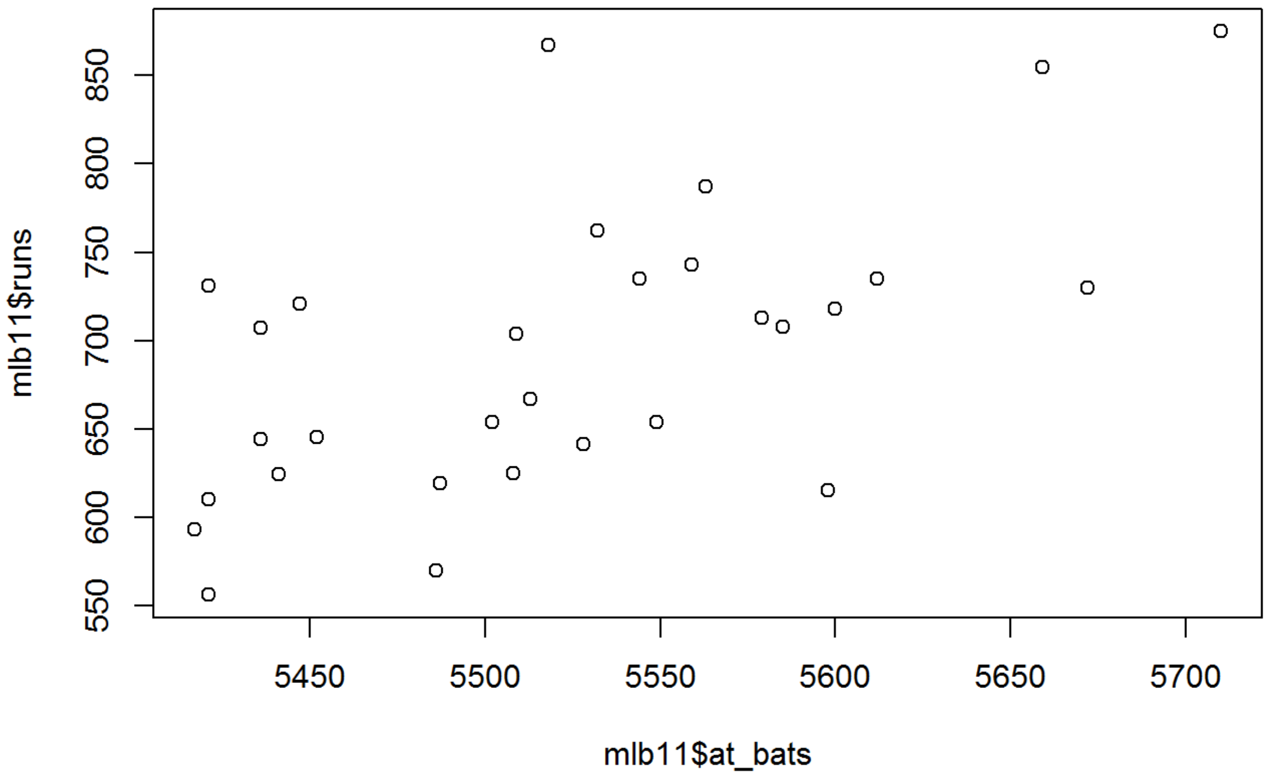
```
setwd('C:/Users/blin261/Documents/Lab7')
load("more/mlb11.RData")
mlb11
```

```
##                     team runs at_bats hits homeruns bat_avg strikeouts
## 1          Texas Rangers  855    5659 1599      210   0.283        930
## 2        Boston Red Sox  875    5710 1600      203   0.280       1108
## 3         Detroit Tigers  787    5563 1540      169   0.277       1143
## 4     Kansas City Royals  730    5672 1560      129   0.275       1006
## 5    St. Louis Cardinals  762    5532 1513      162   0.273        978
## 6          New York Mets  718    5600 1477      108   0.264       1085
## 7       New York Yankees  867    5518 1452      222   0.263       1138
## 8      Milwaukee Brewers  721    5447 1422      185   0.261       1083
## 9       Colorado Rockies  735    5544 1429      163   0.258       1201
## 10        Houston Astros  615    5598 1442       95   0.258       1164
## 11     Baltimore Orioles  708    5585 1434      191   0.257       1120
## 12   Los Angeles Dodgers  644    5436 1395      117   0.257       1087
## 13          Chicago Cubs  654    5549 1423      148   0.256       1202
## 14       Cincinnati Reds  735    5612 1438      183   0.256       1250
## 15    Los Angeles Angels  667    5513 1394      155   0.253       1086
## 16 Philadelphia Phillies  713    5579 1409      153   0.253       1024
## 17     Chicago White Sox  654    5502 1387      154   0.252        989
## 18     Cleveland Indians  704    5509 1380      154   0.250       1269
## 19  Arizona Diamondbacks  731    5421 1357      172   0.250       1249
## 20     Toronto Blue Jays  743    5559 1384      186   0.249       1184
## 21       Minnesota Twins  619    5487 1357      103   0.247       1048
## 22       Florida Marlins  625    5508 1358      149   0.247       1244
## 23    Pittsburgh Pirates  610    5421 1325      107   0.244       1308
## 24     Oakland Athletics  645    5452 1330      114   0.244       1094
## 25        Tampa Bay Rays  707    5436 1324      172   0.244       1193
## 26        Atlanta Braves  641    5528 1345      173   0.243       1260
## 27  Washington Nationals  624    5441 1319      154   0.242       1323
## 28  San Francisco Giants  570    5486 1327      121   0.242       1122
## 29      San Diego Padres  593    5417 1284       91   0.237       1320
## 30      Seattle Mariners  556    5421 1263      109   0.233       1280
##    stolen_bases wins new_onbase new_slug new_obs
## 1           143   96      0.340    0.460   0.800
## 2           102   90      0.349    0.461   0.810
## 3            49   95      0.340    0.434   0.773
## 4           153   71      0.329    0.415   0.744
## 5            57   90      0.341    0.425   0.766
## 6           130   77      0.335    0.391   0.725
## 7           147   97      0.343    0.444   0.788
## 8            94   96      0.325    0.425   0.750
## 9           118   73      0.329    0.410   0.739
## 10          118   56      0.311    0.374   0.684
## 11           81   69      0.316    0.413   0.729
## 12          126   82      0.322    0.375   0.697
## 13           69   71      0.314    0.401   0.715
## 14           97   79      0.326    0.408   0.734
## 15          135   86      0.313    0.402   0.714
## 16           96  102      0.323    0.395   0.717
## 17           81   79      0.319    0.388   0.706
## 18           89   80      0.317    0.396   0.714
## 19          133   94      0.322    0.413   0.736
## 20          131   81      0.317    0.413   0.730
## 21           92   63      0.306    0.360   0.666
## 22          95   72      0.318    0.388   0.706
## 23         108   72      0.309    0.368   0.676
## 24         117   74      0.311    0.369   0.680
## 25         155   91      0.322    0.402   0.724
## 26          77   89      0.308    0.387   0.695
## 27         106   80      0.309    0.383   0.691
## 28          85   86      0.303    0.368   0.671
## 29         170   71      0.305    0.349   0.653
## 30         125   67      0.292    0.348   0.640
```

Exercise 1: What type of plot would you use to display the relationship between runs and one of the other numerical variables? Plot this relationship using the variable at_bats as the predictor. Does the relationship look linear? If you knew a team's at_bats, would you be comfortable using a linear model to predict the number of runs?

I would fist use the scatter plot to get a very general idea about the relationship between two variable. As shown in the following runs and at_bats appear to be linearly related. Therefore, I am comfortable to use linear model to predict the number of runs.

```
plot(mlb11$runs ~ mlb11$at_bats)
```

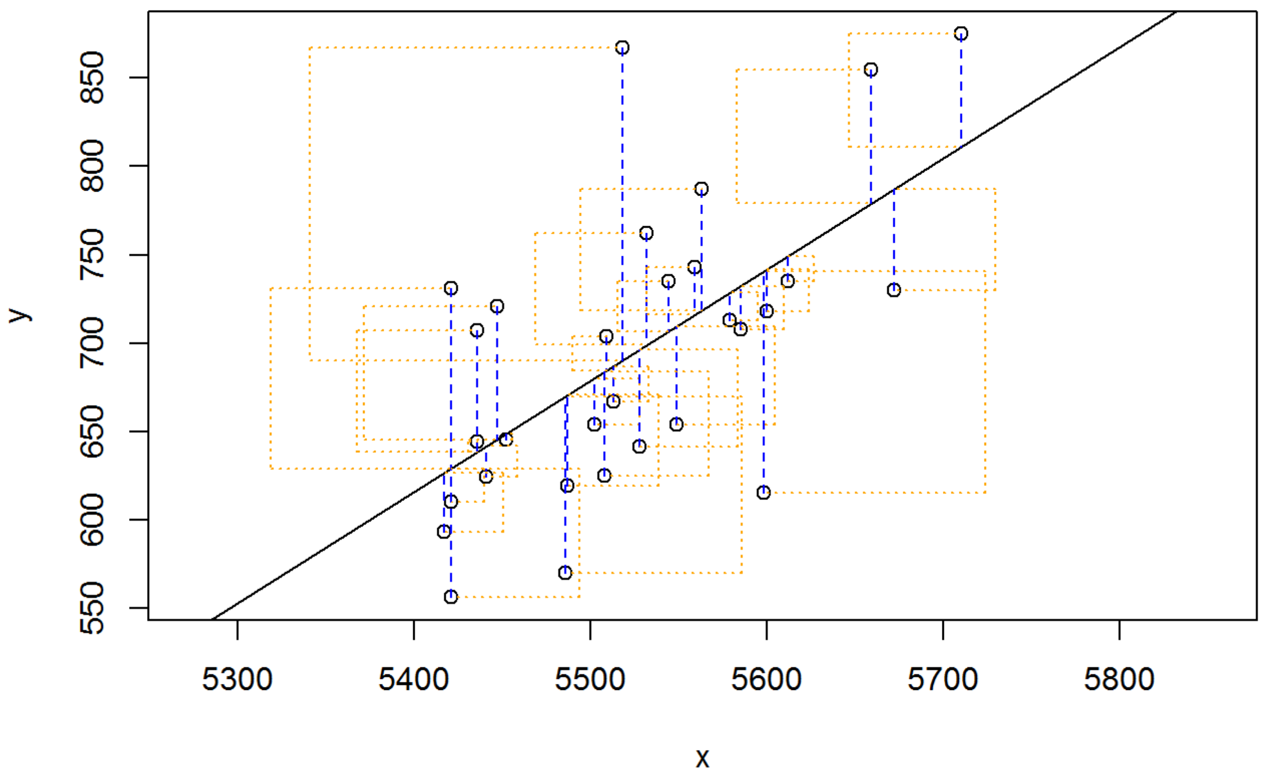


```
cor(mlb11$runs, mlb11$at_bats)
```

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

Exercise 2: Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations. It is a positive linear relationship. The strength of the relationship is moderately strong. There are some points look like outliers. As most of the team has number of at_bats less than 5600, those points that have at_bats greater than 5600 and at top right hand corner appear to be positive outliers.

```
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
```

Exercise 3: Using plot_ss, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors? The smallest sum of squares I got is 124567.5. I do not have a neighbor.

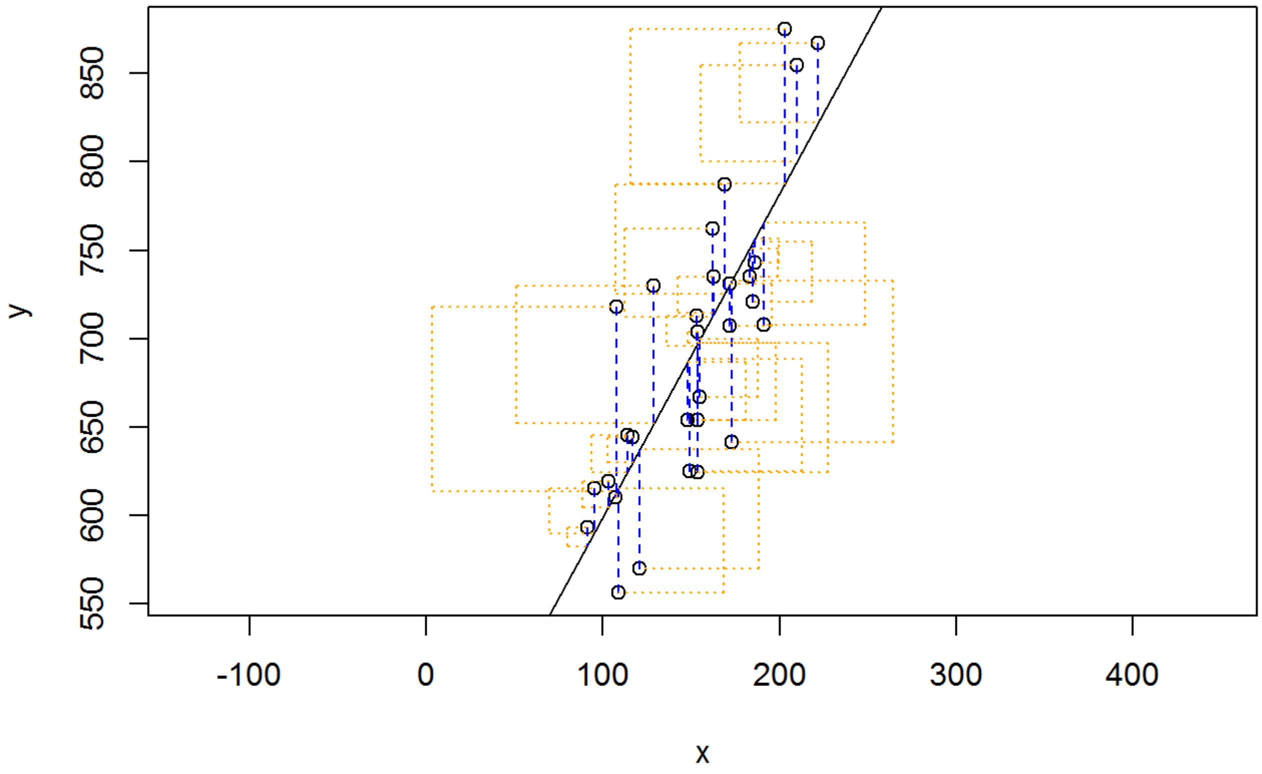
```
m1 <- lm(runs ~ at_bats, data = mlb11)
summary(m1)
```

```
##
## 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
```

Exercise 4: Fit a new model that uses homeruns to predict runs. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between success of a team and its home runs?

The equation of the regression line is y = 415.2389 + 1.8345 * x. This equation indicates for each additional homerun, the total number of runs also increase by about 1.8345.

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



```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##     415.239        1.835
##
## Sum of Squares:  73671.99
```

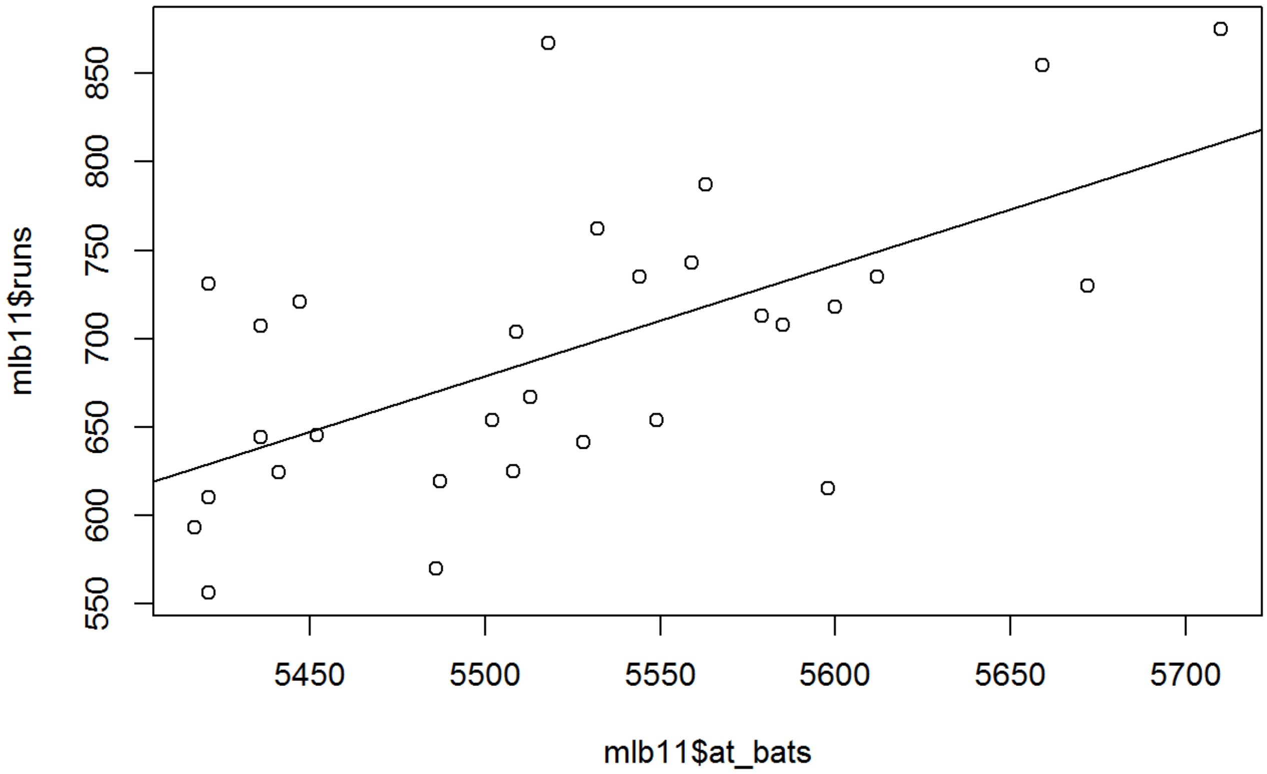
```
m2 <- lm(runs ~ homeruns, data = mlb11)
summary(m2)
```

```
## 
## 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
```

Exercise 5: If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,578 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

He or she will predict 728 runs for a team with 5578 at-bats. In acutal data, we can find a team with 5579 at-bats have 713 runs. Therefore, the model overestimate by around 14 to 15 runs.

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



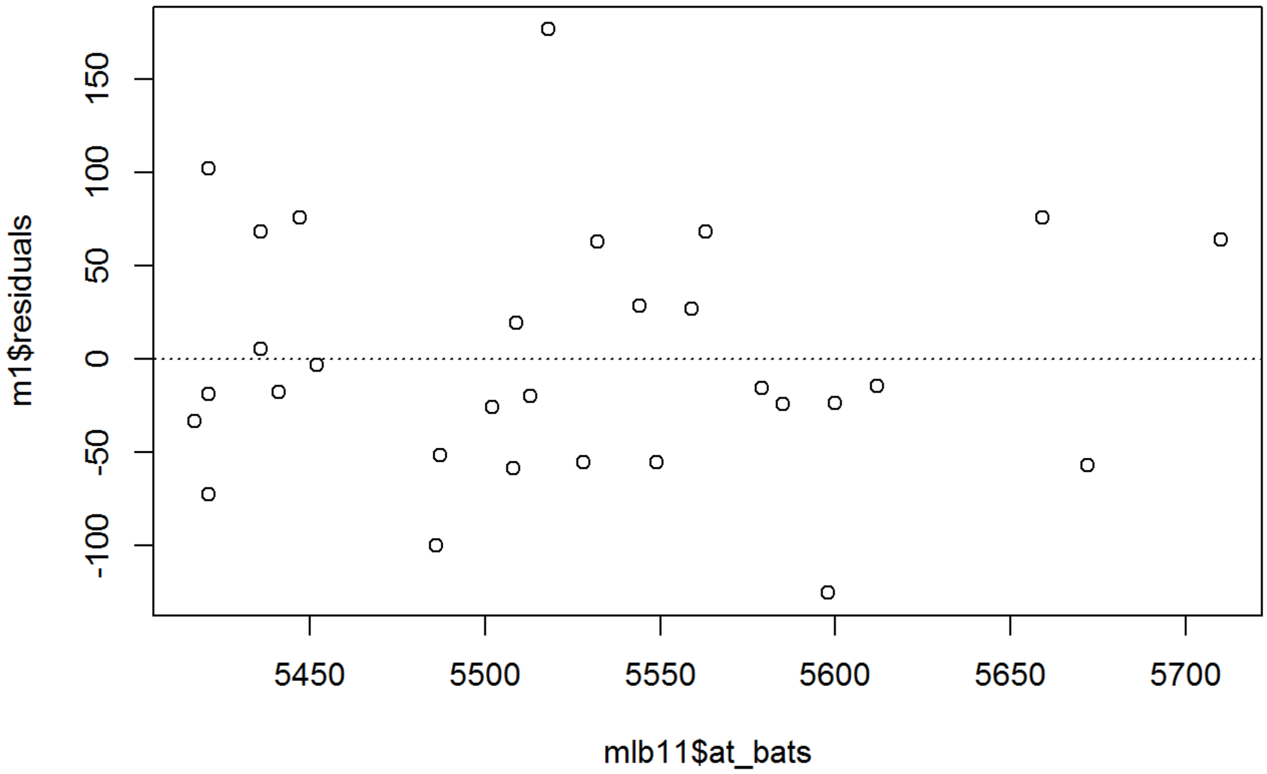
```
5578 * 0.6305 - 2789.2429
```

```
## [1] 727.6861
```

Exercise 6: Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

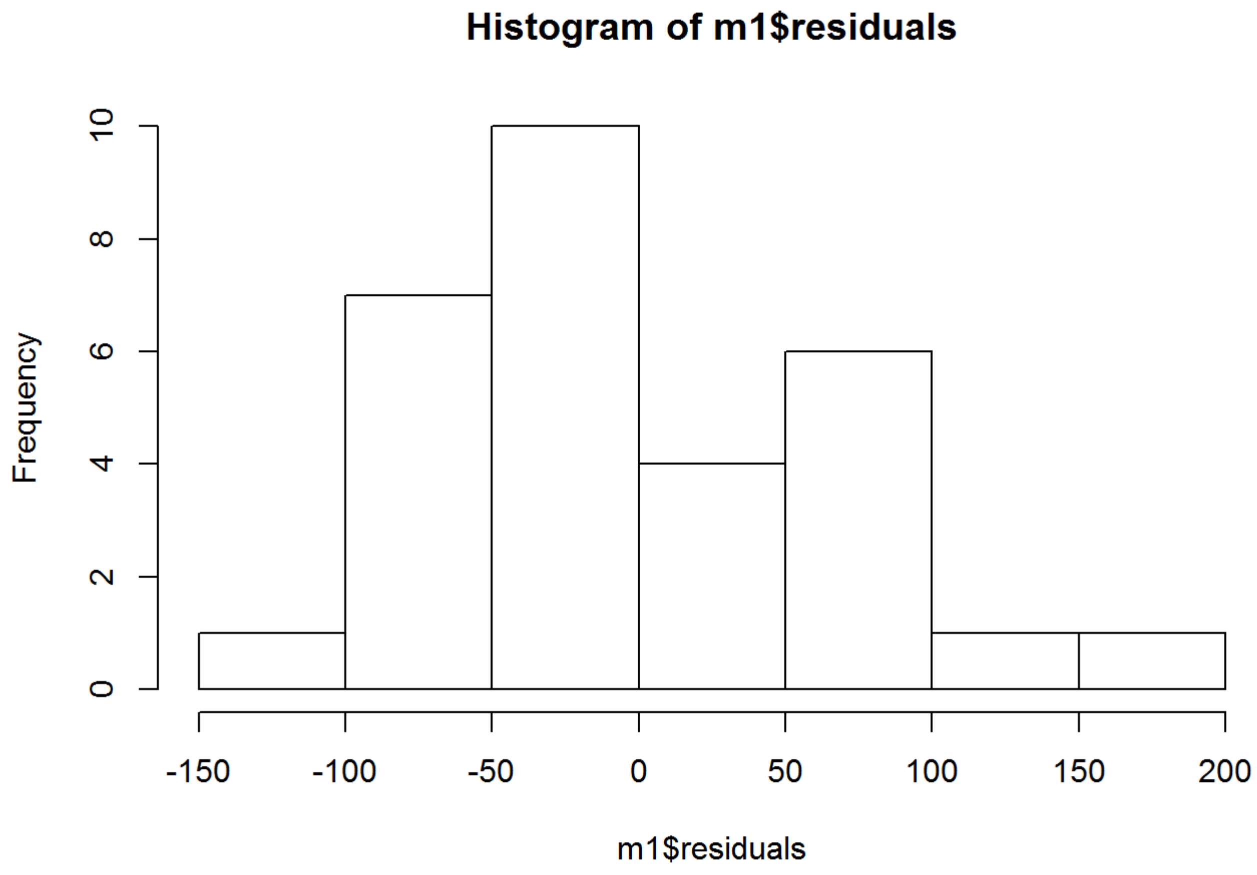
There is no pattern observed in the following figure. So we can conclude that the there exist linear relationship between residuals and at_bats.

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

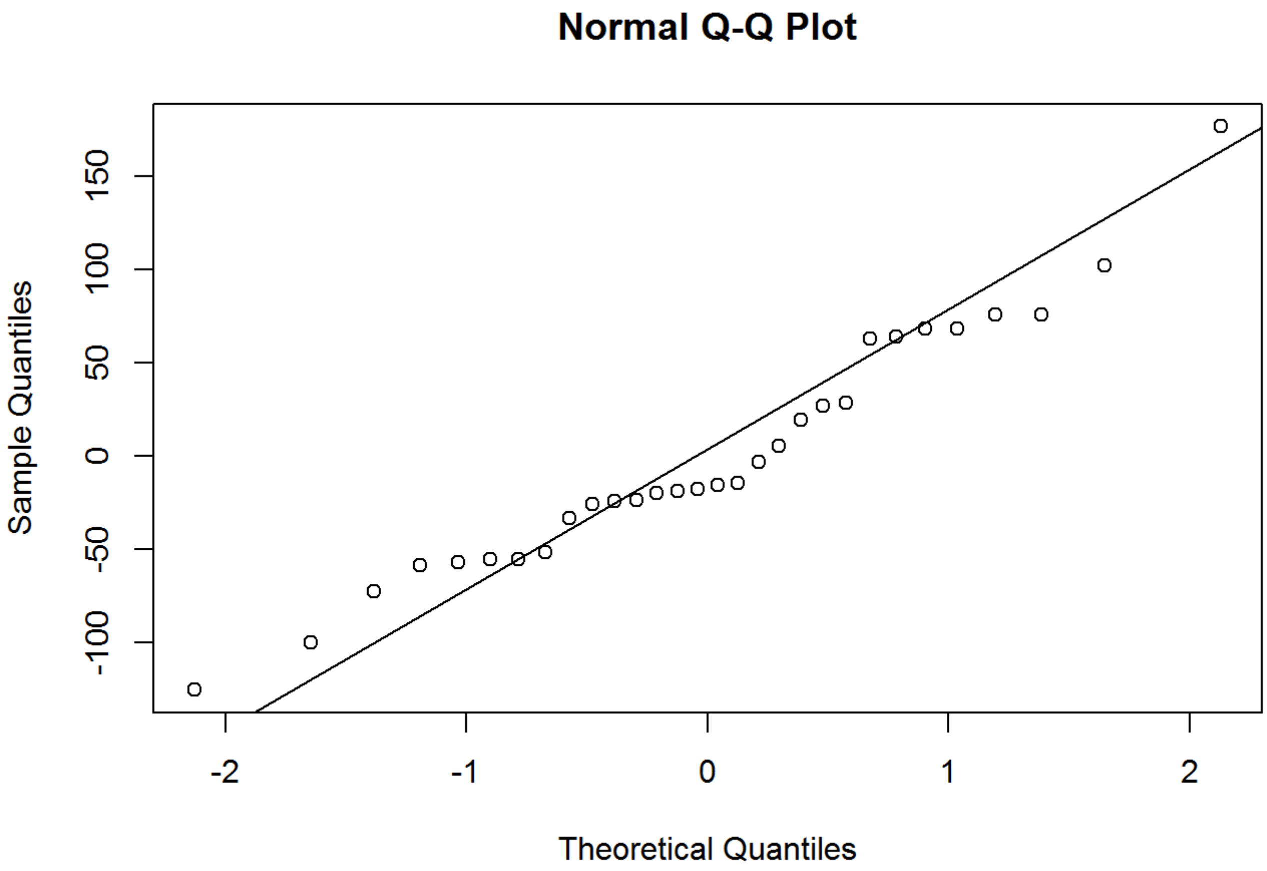


Exercise 7: Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met? The histogram is bell shaped and unimodel, the data does not deviate from qqline too much. And it is not bended. So the normal residual condition appear to be met.

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



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

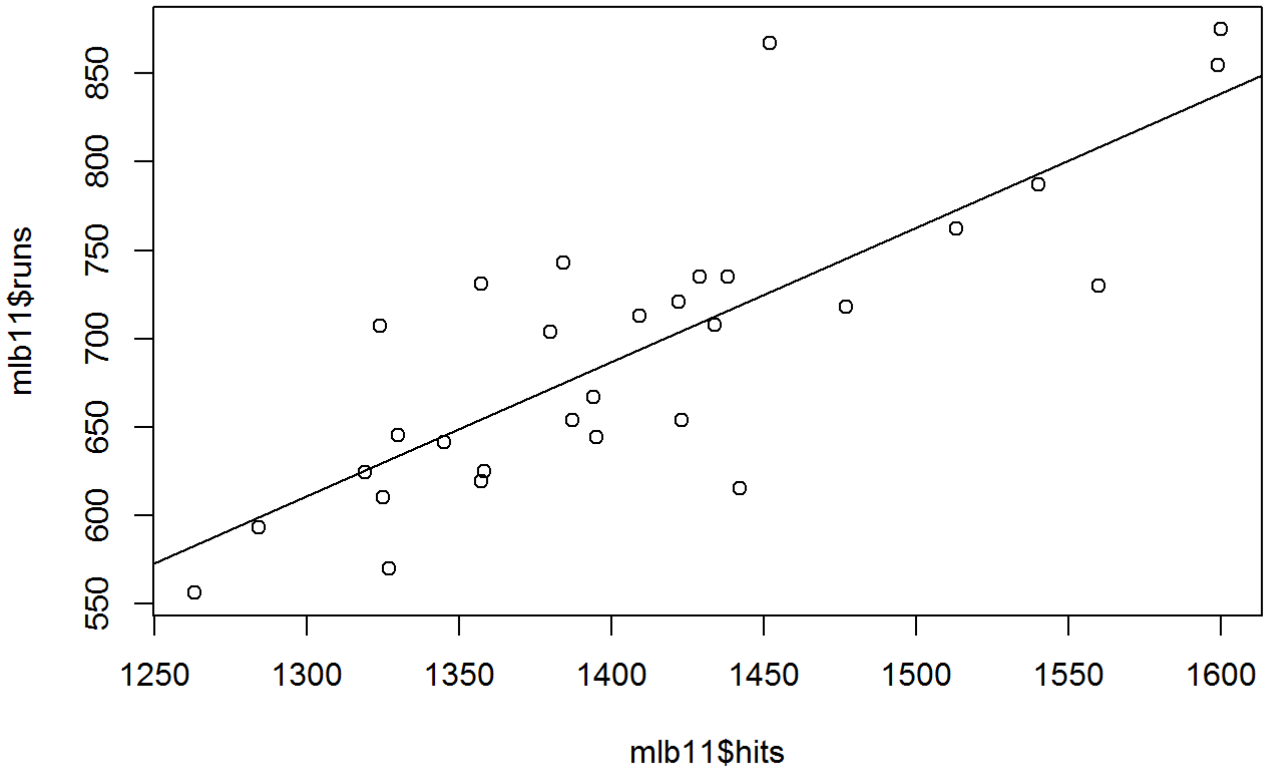


Exercise 8: Based on the plot in (1), does the constant variability condition appear to be met? Yes, because the variability of residuals around the 0 line appear to be roughly constant.

On Your Own 1. Choose another traditional variable from mlb11 that you think might be a good predictor of runs. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?

I choose number of hits, there exists the linear relationship between hits and runs variables.

```
plot(mlb11$runs ~ mlb11$hits)
m3 <- lm(runs ~ hits, data = mlb11)
abline(m3)
```



2. How does this relationship compare to the relationship between runs and at_bats? Use the R22 values from the two model summaries to compare. Does your variable seem to predict runs better than at_bats? How can you tell? I think the relationship between runs and hits are stronger as it has higher correlation coefficient. The R^2 value for the relationship between runs and hits (0.6419) is higher than that between runs and at_bats (0.3729). Therefore, hits is a better than at_bats.

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

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

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

```
## [1] 0.8012108
```

```
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
```

```
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
```

3. Now that you can summarize the linear relationship between two variables, investigate the relationships between runs and each of the other five traditional variables. Which variable best predicts runs? Support your conclusion using the graphical and numerical methods we've discussed (for the sake of conciseness, only include output for the best variable, not all five).

It looks like bat_avg best predicts runs. It has R^2 value of 0.6561

```
m4 <- lm(runs ~ bat_avg, data = mlb11)
cor(mlb11$runs, mlb11$bat_avg)
```

```
## [1] 0.8099859
```

```
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
```

```
m5 <- lm(runs ~ strikeouts, data = mlb11)
cor(mlb11$runs, mlb11$strikeouts)
```

```
## [1] -0.4115312
```

```
summary(m5)
```

```
## 
## Call:
## lm(formula = runs ~ strikeouts, data = mlb11)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -132.27  -46.95  -11.92   55.14  169.76 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1054.7342   151.7890   6.949 1.49e-07 ***
## strikeouts    -0.3141     0.1315  -2.389   0.0239 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 76.5 on 28 degrees of freedom
## Multiple R-squared:  0.1694, Adjusted R-squared:  0.1397 
## F-statistic: 5.709 on 1 and 28 DF,  p-value: 0.02386
```

```
m6 <- lm(runs ~ stolen_bases, data = mlb11)
cor(mlb11$runs, mlb11$stolen_bases)
```

```
## [1] 0.05398141
```

```
summary(m6)
```

```
##
## Call:
## lm(formula = runs ~ stolen_bases, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.94  -62.87   10.01   38.54  182.49
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  677.3074    58.9751  11.485 4.17e-12 ***
## stolen_bases   0.1491     0.5211   0.286    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.82 on 28 degrees of freedom
## Multiple R-squared:  0.002914,   Adjusted R-squared:  -0.0327
## F-statistic: 0.08183 on 1 and 28 DF,  p-value: 0.7769
```

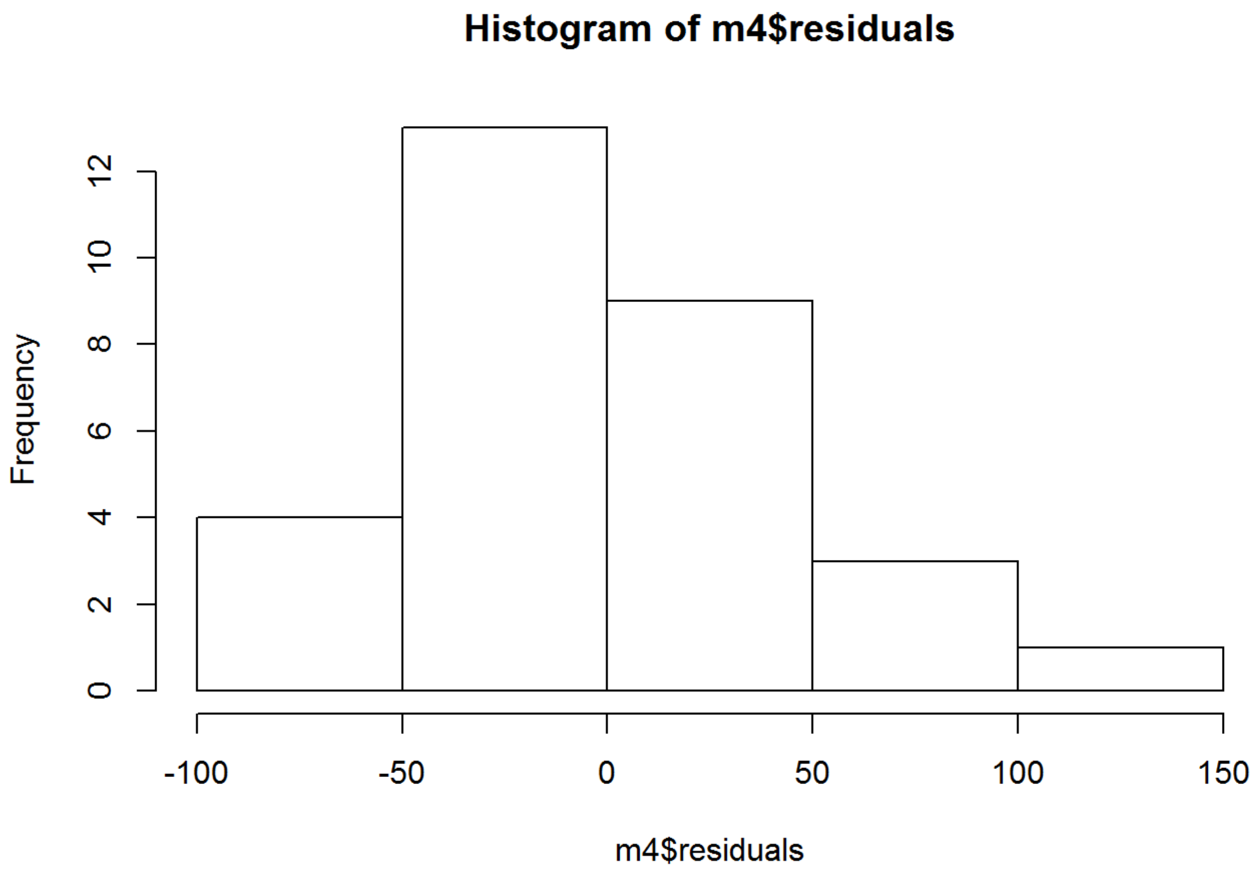
```
m7 <- lm(runs ~ wins, data = mlb11)
cor(mlb11$runs, mlb11$wins)
```

```
## [1] 0.6008088
```

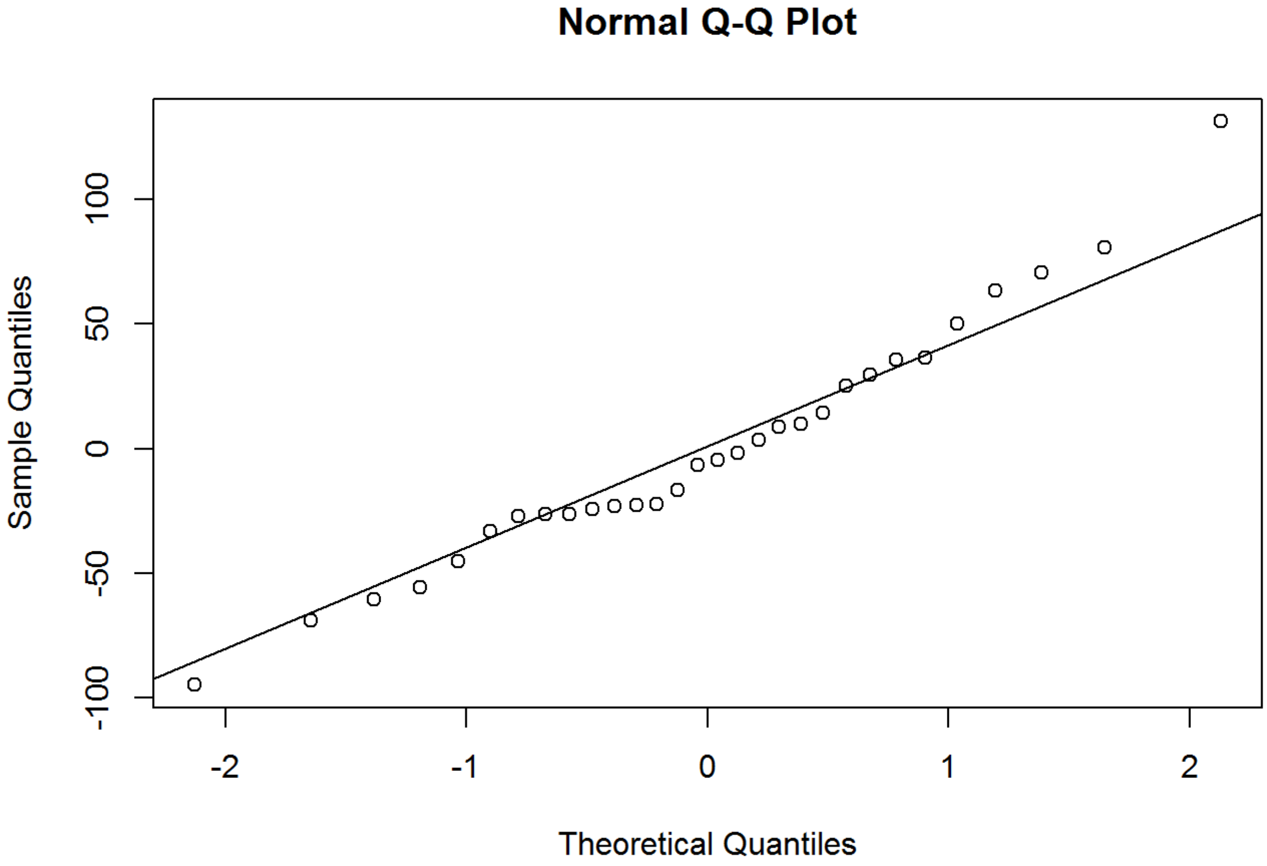
```
summary(m7)
```

```
##
## Call:
## lm(formula = runs ~ wins, data = mlb11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -145.450  -47.506   -7.482   47.346  142.186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  342.121     89.223   3.834 0.000654 ***
## wins           4.341      1.092   3.977 0.000447 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.1 on 28 degrees of freedom
## Multiple R-squared:  0.361,  Adjusted R-squared:  0.3381
## F-statistic: 15.82 on 1 and 28 DF,  p-value: 0.0004469
```

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



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



4. Now examine the three newer variables. These are the statistics used by the author of Moneyball to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we've analyzed, which seems to be the best predictor of runs? Using the limited (or not so limited) information you know about these baseball statistics, does your result make sense?

All three new variables are better than the old variables in terms of predicting a team's success. new_obs is the best predictor of runs. My result makes perfect sense.

```
m8 <- lm(runs ~ new_onbase, data = mlb11)
m9 <- lm(runs ~ new_slug, data = mlb11)
m10 <- lm(runs ~ new_obs, data = mlb11)
summary(m8)
```

```
##
## Call:
## lm(formula = runs ~ new_onbase, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.270 -18.335   3.249  19.520  69.002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1118.4      144.5  -7.741 1.97e-08 ***
## new_onbase    5654.3      450.5  12.552 5.12e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.61 on 28 degrees of freedom
## Multiple R-squared:  0.8491, Adjusted R-squared:  0.8437
## F-statistic: 157.6 on 1 and 28 DF,  p-value: 5.116e-13
```

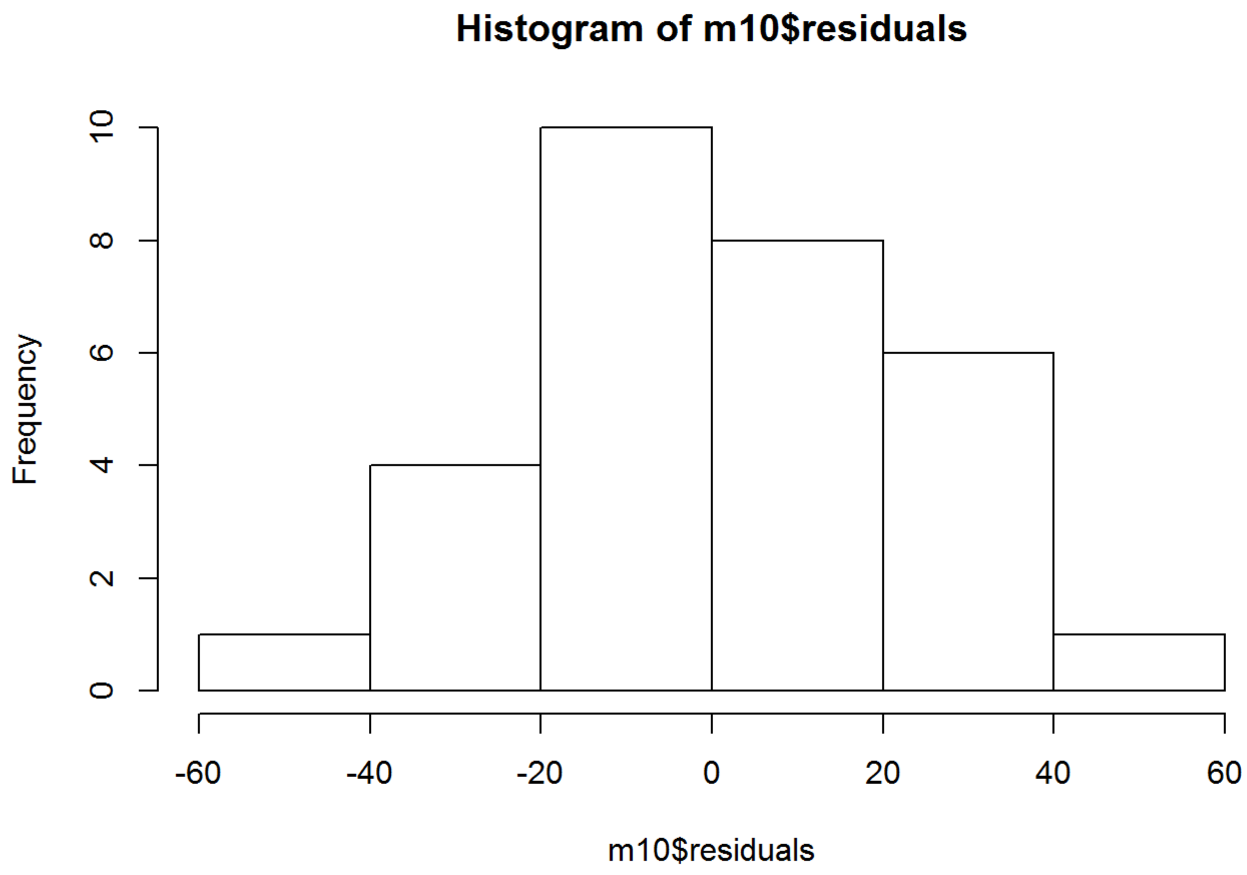
```
summary(m9)
```

```
##
## Call:
## lm(formula = runs ~ new_slug, data = mlb11)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.41 -18.66  -0.91  16.29  52.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -375.80      68.71   -5.47 7.70e-06 ***
## new_slug     2681.33     171.83   15.61 2.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.96 on 28 degrees of freedom
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8932
## F-statistic: 243.5 on 1 and 28 DF,  p-value: 2.42e-15
```

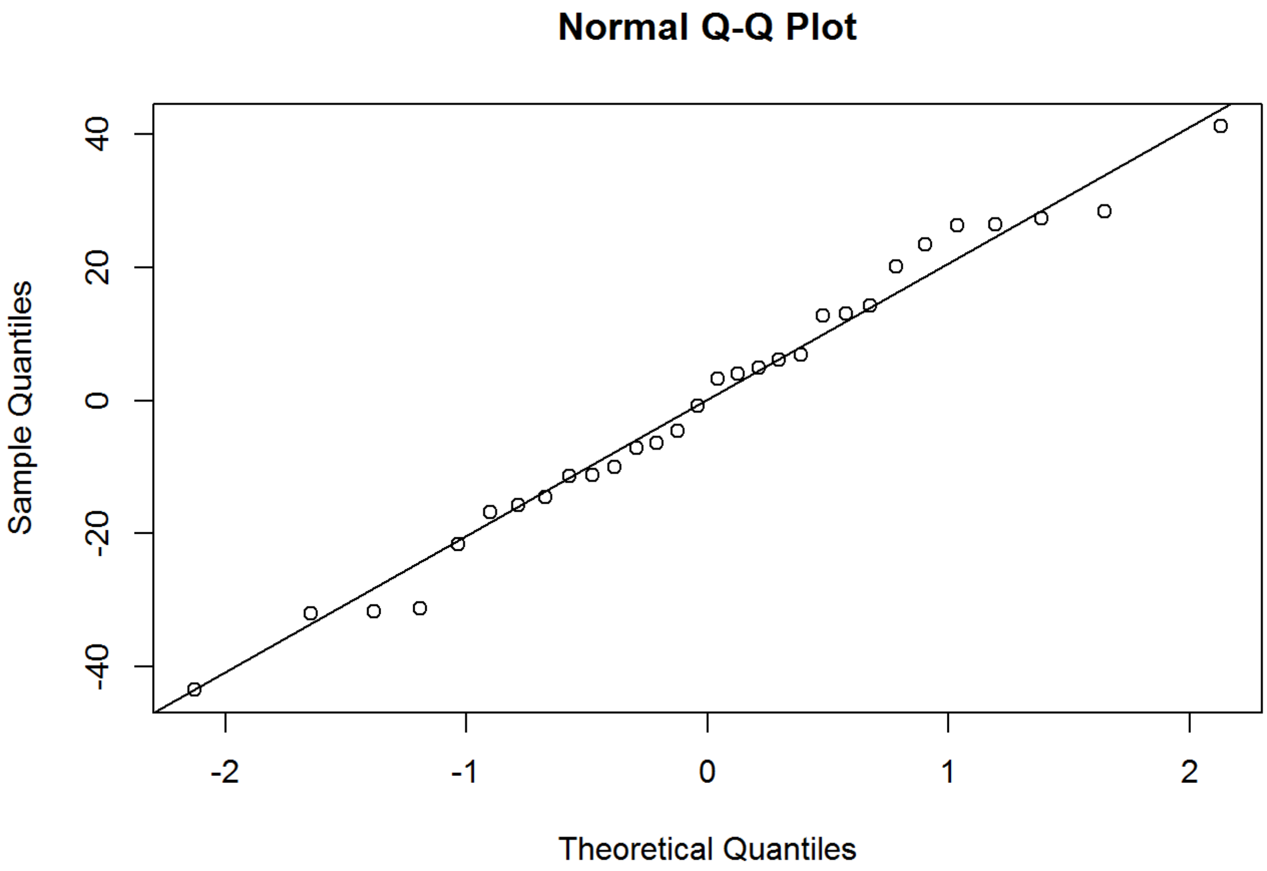
```
summary(m10)
```

```
## 
## 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
```

```
hist(m10$residuals)
```

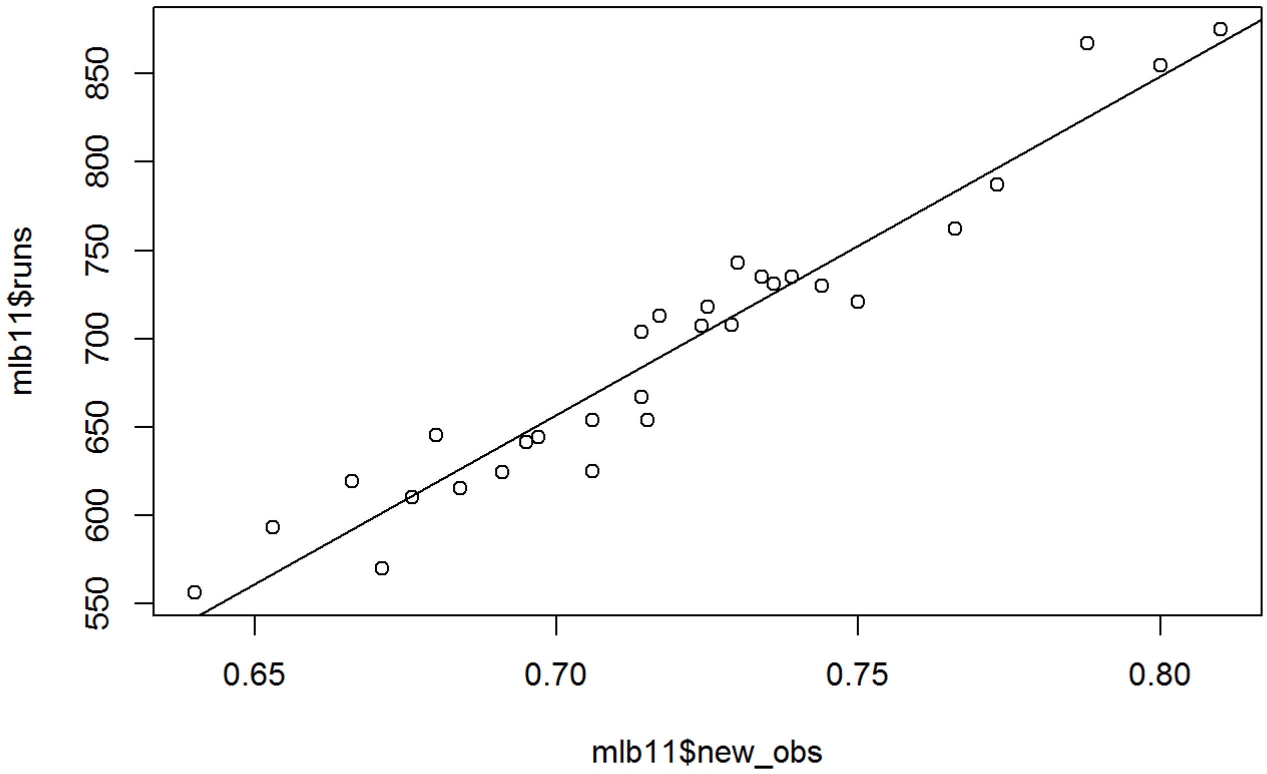


```
qqnorm(m10$residuals)
qqline(m10$residuals)
```

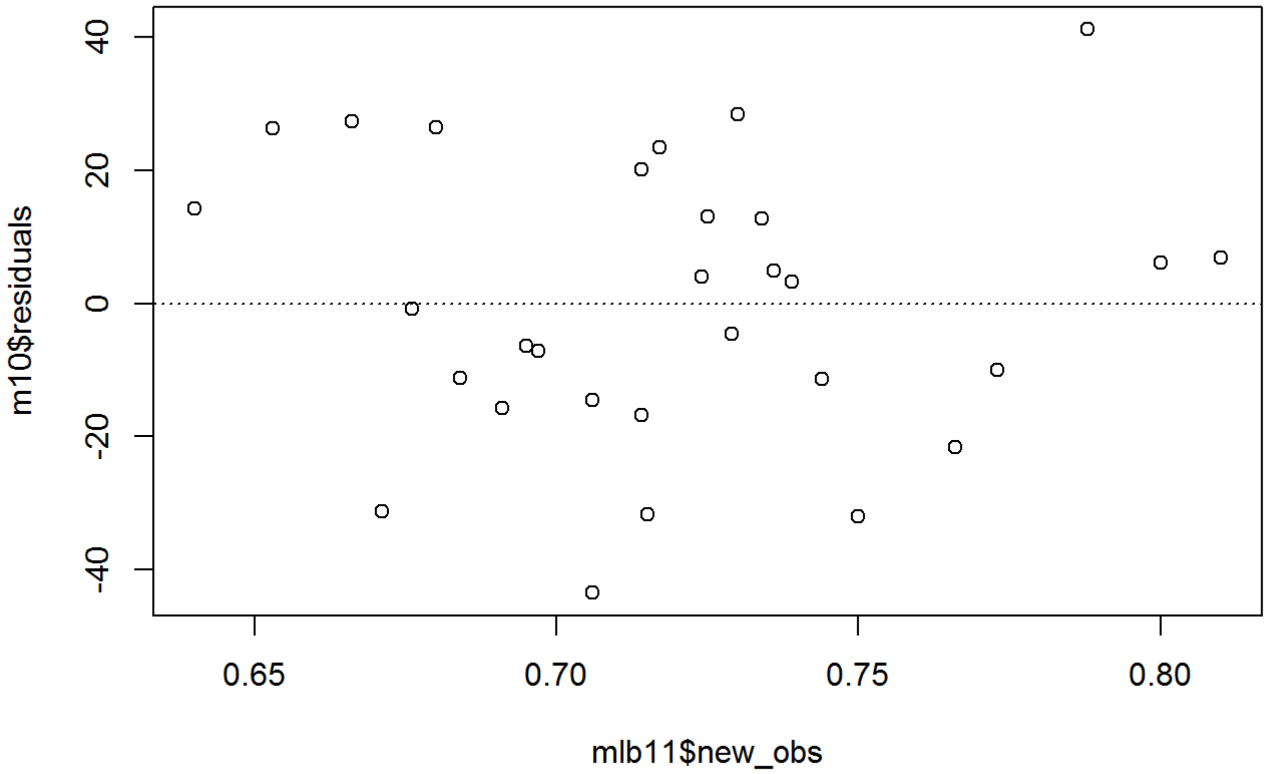


5. Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs. First of all, the scatterplot shows linear relationship. From the residual graphs above, the histogram is unimodel and bell shaped. All the data are very closed to the qqline and it is not bended. Furthermore, if we take a look at the residuals plot, the data points appear random without any patterns.

```
plot(mlb11$runs ~ mlb11$new_obs)
abline(m10)
```



```
plot(m10$residuals ~ mlb11$new_obs)
abline(h = 0, lty = 3)
```


40
20
0
-20
-40
m10$residuals
0.65
0.70
0.75
0.80
mlb11$new_obs