Week8-Lab VII

Aamol Gote

10/18/2019

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

nrow(mlb11)

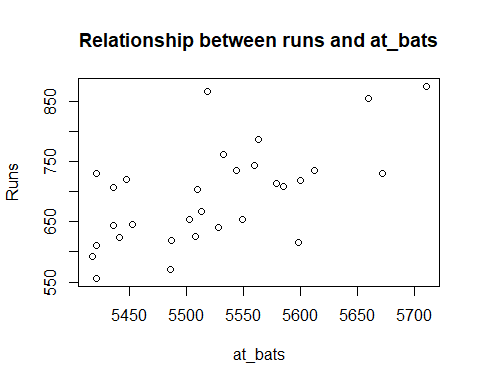
## [1] 30

**Excercise 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?

Scatterplot would be used to display the relationship between runs and one of the other Numerical variables

plot(mlb11$at\_bats, mlb11$runs, main = "Relationship between runs and at\_bats", xlab = "at\_bats", ylab = "Runs")



Relationship between runs and at\_bats is linear, but relationship is not strong relationship and it seems to be moderate one.

We can quantify the relationship between runs and at\_bats using correlation coefficient

cor(mlb11$at\_bats, mlb11$runs)

## [1] 0.610627

Team’s at\_bats can be used in a linear model to predict the number of runs given that we could see some linear trend in relationship. But, the relationship is not strong enough, so the prediction may not be of desired accuracy

**Excercise 2**

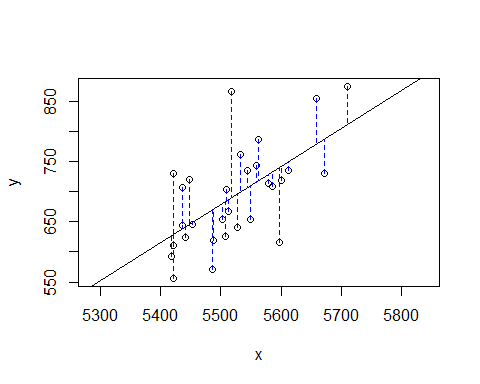
Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

1. Relationship between runs and at\_bats is positively correlated, as at\_bats increases runs also increases.
2. Relationship is not strong but moderatly stronger.
3. There are few outliars like at\_bats > 5650 as well runs near 850
4. There’s no curvature to the plot, so this is a linear relationship.

**Excercise 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?

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

Smallest Sum of squares: 123721.9

The neighboring value deviate from the smallest value by around 4000 - 5000.

==============================================

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

**Excercise 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?

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

##   
## 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

Equation of regression line

y = b0 + b1x

runs = 415.2389 + 1.8345 \* homeruns

In the context of the relationship between success of a team and its home runs slope tells that for every home run, number of total runs will also increase 1.8345.

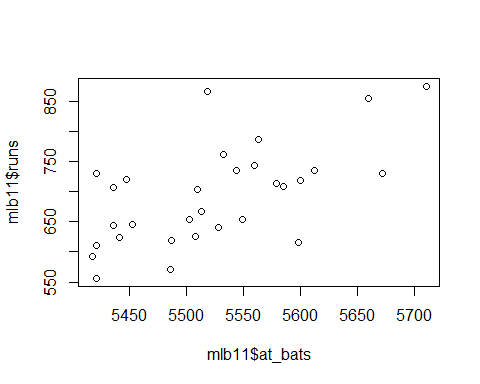
Slope = 1.8345

cor(mlb11$runs, mlb11$homeruns)

## [1] 0.7915577

Positive relationship with a correlation coefficient of 0.7915577, which is on stronger side (relatively stronger).

plot(mlb11$runs ~ mlb11$at\_bats)  
abline(m1)



**Excercise 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?

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

Least Square Regression line for runs vs at\_bats

y = b0 + b1x

b0 = -2789.2429 b1 = 0.6305;

x = 5578 (at\_bats)

y = -2789.2429 + 0.6305 \* (5578)

y =727.6861

Predicted runs (y) = 728

Estimated number of runs for 5578 at bats based on the linear regression formula above is 728.

Based on below data there are no teams with at\_bats = 5578, nearest match is at\_bats = 5579 for **Philadelphia Phillies** with runs **713**. So going by this there is an overestimate of 728 - 713 = 15 runs.

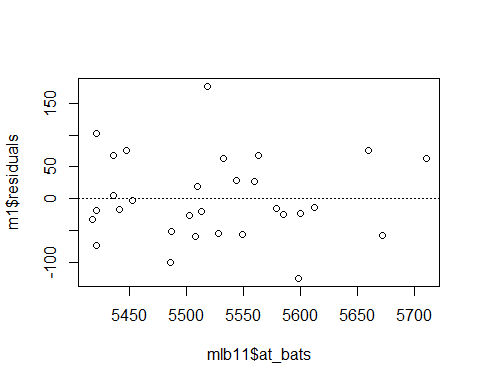
Residual of the prediction is 15

mlb11[order(mlb11$runs,mlb11$at\_bats),]

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

=================

plot(m1$residuals ~ mlb11$at\_bats)  
abline(h = 0, lty = 3)



**Excercise 6**

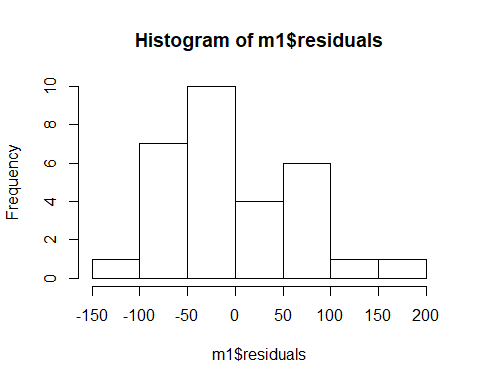
Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

There is no apparent pattern in the residual plot.

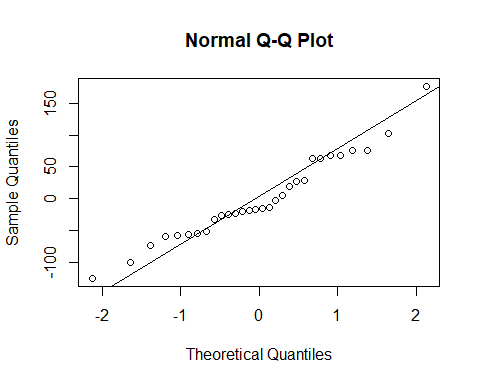
Distribution is scattered all over. It does indicate linear relationship between runs and at\_bats.

=============================================================

hist(m1$residuals)



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



**Excercise 7**

Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

Histogram is right skewed.

Nearly normal residual condition appears to be satisfied.

**Excercise 8**

Based on the plot in (1), does the constant variability condition appear to be met?

Constant variability condition appears to be satisfied barring some out-liars.

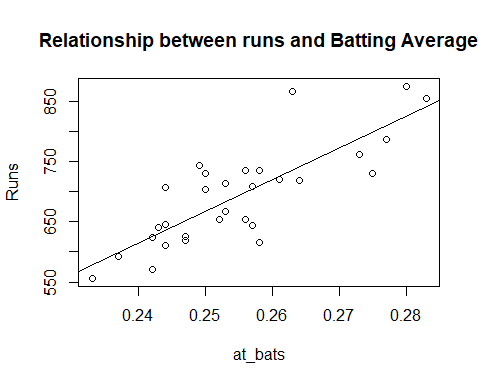
=============================================================

**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?

Another variable considered bat\_avg (Batting average), reason for choosing the same is higher the batting average higher number of runs.

plot(mlb11$bat\_avg, mlb11$runs, main = "Relationship between runs and Batting Average", xlab = "at\_bats", ylab = "Runs")  
m3 <- lm(runs ~ bat\_avg, data = mlb11)  
abline(m3)



cor(mlb11$runs, mlb11$bat\_avg)

## [1] 0.8099859

summary(m3)

##   
## 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

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

Least Square Regression line for runs vs bat\_avg

y = b0 + b1x

b0 = -642.8 b1 = 5242.2

y = -642.8 + (5242.2)\*x

Relationship between runs vs bat\_avg is positive, linear and relatively strong.

1. How does this relationship compare to the relationship between runs and at\_bats? Use the R2 values from the two model summaries to compare. Does your variable seem to predict runs better than at\_bats? How can you tell?

Relationship between runs and bat\_avg seems to be more stronger than runs and at\_bats.

R2 value for runs and bat\_avg = 0.6561

R2 value for runs and at\_bats = 0.3729

The variable “bat\_avg” is a much better predictor of “runs” than the variable “at\_bats” as is indcated by the R2 value.

1. 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).

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

##   
## 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

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

##   
## 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

bat\_avg <- lm(runs ~ bat\_avg, data = mlb11)  
summary(bat\_avg)

##   
## 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

strikeouts <- lm(runs ~ strikeouts, data = mlb11)  
summary(strikeouts)

##   
## 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

stolen\_bases <- lm(runs ~ stolen\_bases, data = mlb11)  
summary(stolen\_bases)

##   
## 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

wins <- lm(runs ~ wins, data = mlb11)  
summary(wins)

##   
## 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

cor(mlb11$runs, mlb11$hits)

## [1] 0.8012108

cor(mlb11$runs, mlb11$homeruns)

## [1] 0.7915577

cor(mlb11$runs, mlb11$bat\_avg)

## [1] 0.8099859

cor(mlb11$runs, mlb11$strikeouts)

## [1] -0.4115312

cor(mlb11$runs, mlb11$stolen\_bases)

## [1] 0.05398141

cor(mlb11$runs, mlb11$wins)

## [1] 0.6008088

Comparing R2 values and correlation coefficient for remaining variables

hits => Multiple R-squared: 0.6419

homeruns => Multiple R-squared: 0.6266

bat\_avg => Multiple R-squared: 0.6561

strikeouts => Multiple R-squared: 0.1694

stolen\_bases => Multiple R-squared: 0.002914

wins => Multiple R-squared: 0.361

hits => Correlation Coefficent: 0.8012108

homeruns => Correlation Coefficent: 0.7915577

bat\_avg => Correlation Coefficent: 0.8099859

strikeouts => Correlation Coefficent: -0.4115312

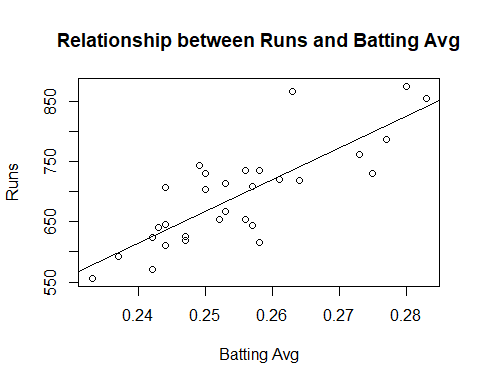
stolen\_bases => Correlation Coefficent: 0.05398141

wins => Correlation Coefficent: 0.6008088

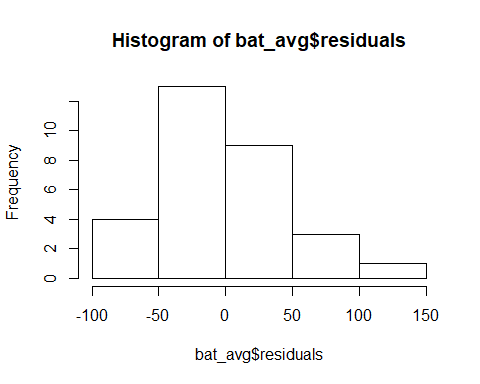
After running the summary statistics for all the variables, the variable which best predicts the runs based on R2 happened to be bat\_avg.

Correlation coeficient of bat\_avg seems to be much better than other variables.

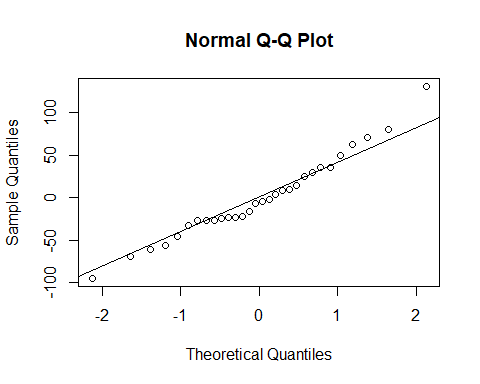
plot(mlb11$runs ~ mlb11$bat\_avg, main = "Relationship between Runs and Batting Avg", xlab = "Batting Avg", ylab = "Runs")  
bat\_avg <- lm(runs ~ bat\_avg, data = mlb11)  
abline(bat\_avg)



hist(bat\_avg$residuals)



qqnorm(bat\_avg$residuals)  
qqline(bat\_avg$residuals)



1. 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?

Three new variables new\_onbase new\_slug new\_obs

new\_onbase <- lm(runs ~ new\_onbase, data = mlb11)  
summary(new\_onbase)

##   
## 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

new\_slug <- lm(runs ~ new\_slug, data = mlb11)  
summary(new\_slug)

##   
## 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

new\_obs <- lm(runs ~ new\_obs, data = mlb11)  
summary(new\_obs)

##   
## 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

cor(mlb11$runs, mlb11$new\_onbase)

## [1] 0.9214691

cor(mlb11$runs, mlb11$new\_slug)

## [1] 0.9470324

cor(mlb11$runs, mlb11$new\_obs)

## [1] 0.9669163

new\_onbase => Multiple R-squared: 0.8491

new\_slug => Multiple R-squared: 0.8969

new\_obs => Multiple R-squared: 0.6561

new\_onbase => Correlation Coefficent: 0.9214691

new\_slug => Correlation Coefficent: 0.9470324

new\_obs => Correlation Coefficent: 0.9669163

Earlier Best prediction model was based on bats\_avg based on tradition variables

bat\_avg => Multiple R-squared: 0.6561

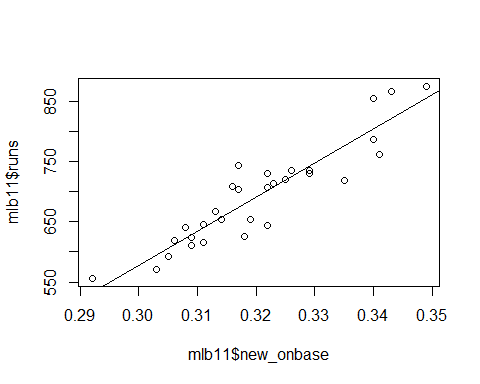
bat\_avg => Correlation Coefficent: 0.8099859

The newer variable seems to predict runs better than old variables. The R2 value for newer variables are higher than that of old variable. Given that the newer variables represent more advanced statistics of baseball it does make sense that they are better predictor of runs. So in summary result does make sense.

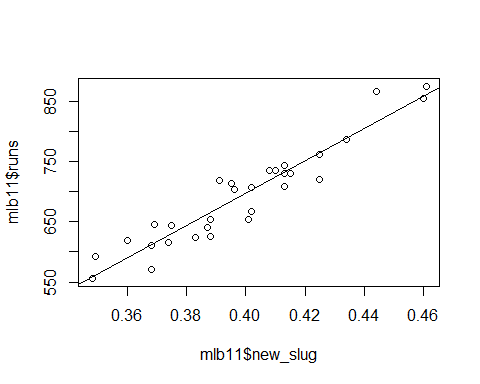
Of the new three variables the relationship between runs and new\_obs variable has the highest R2. It appears to be best predictor of runs.

Below are the plots for all three new varaibles

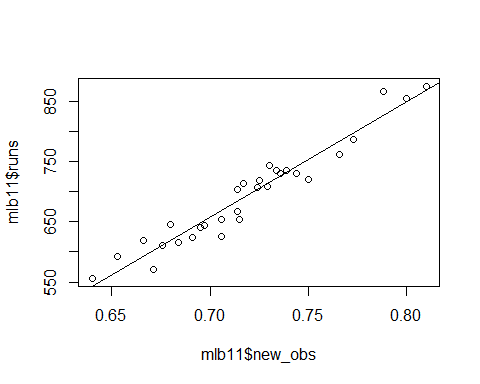
plot(mlb11$runs ~ mlb11$new\_onbase)  
abline(new\_onbase)



plot(mlb11$runs ~ mlb11$new\_slug)  
abline(new\_slug)



plot(mlb11$runs ~ mlb11$new\_obs)  
abline(new\_obs)

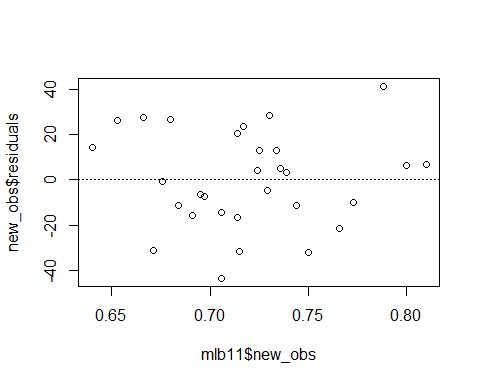


1. Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

new\_obs is the best predictor.

Linearity

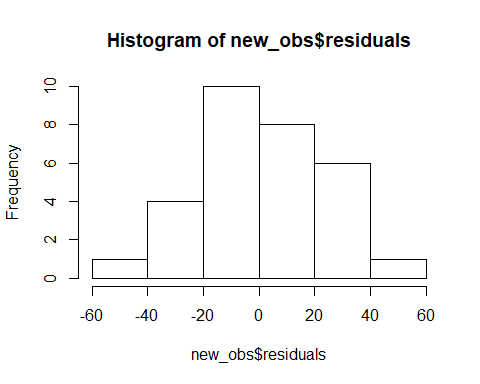
plot(new\_obs$residuals ~ mlb11$new\_obs)  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0



The relationship looks linear based on a residual plot as the variability of residuals is approximately constant across the distribution but does not indicate any curvatures or any indication of non-normality.

Normal residuals:

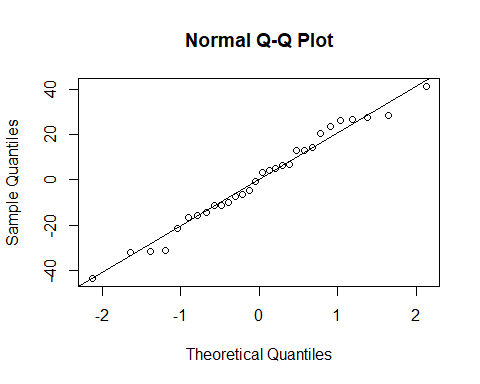
hist(new\_obs$residuals)



The residuals are approximately normaly distributed.

Constant variability

qqnorm(new\_obs$residuals)  
qqline(new\_obs$residuals)



The least squares line remains roughly constant so the condition constant variability has been met