Lab 7

Priya Shaji

4/2/2019

## Introduction to linear regression

### Batter up

The movie Moneyball focuses on the “quest for the secret of success in baseball”. It follows a low-budget team, the Oakland Athletics, who believed that underused statistics, such as a player’s ability to get on base, betterpredict the ability to score runs than typical statistics like home runs, RBIs (runs batted in), and batting average. Obtaining players who excelled in these underused statistics turned out to be much more affordable for the team.

In this lab we’ll be looking at data from all 30 Major League Baseball teams and examining the linear relationship between runs scored in a season and a number of other player statistics. Our aim will be to summarize these relationships both graphically and numerically in order to find which variable, if any, helps us best predict a team’s runs scored in a season.

### The data

Let’s load up the data for the 2011 season.

#download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")  
load("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/Labs/Lab7/mlb11.RData")

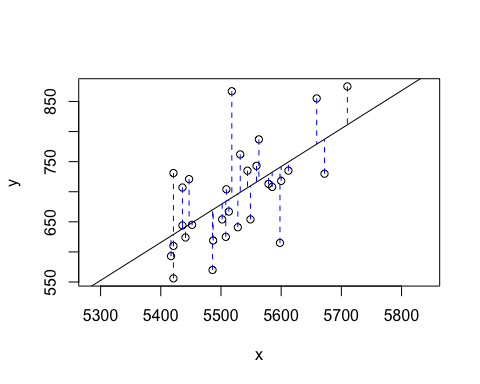
In addition to runs scored, there are seven traditionally used variables in the data set: at-bats, hits, home runs, batting average, strikeouts, stolen bases, and wins. There are also three newer variables: on-base percentage, slugging percentage, and on-base plus slugging. For the first portion of the analysis we’ll consider the seven traditional variables. At the end of the lab, you’ll work with the newer variables on your own.

### Sum of squared residuals

Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It’s also useful to be able to describe the relationship of two numerical variables, such as runs and at\_bats above.

Just as we used the mean and standard deviation to summarize a single variable, we can summarize the relationship between these two variables by finding the line that best follows their association. Use the following interactive function to select the line that you think does the best job of going through the cloud of points.

plot\_ss(x = mlb11$at\_bats, y = mlb11$runs)



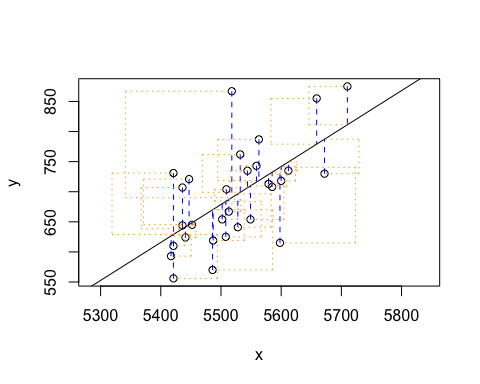
## Click two points to make a line.  
   
## Call:  
## lm(formula = y ~ x, data = pts)  
##   
## Coefficients:  
## (Intercept) x   
## -2789.2429 0.6305   
##   
## Sum of Squares: 123721.9

After running this command, you’ll be prompted to click two points on the plot to define a line. Once you’ve done that, the line you specified will be shown in black and the residuals in blue. Note that there are 30 residuals, one for each of the 30 observations. Recall that the residuals are the difference between the observed values and the values predicted by the line:

ei=yi−ŷ i

The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. To visualize the squared residuals, you can rerun the plot command and add the argument showSquares = TRUE.

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

### The linear model

It is rather cumbersome to try to get the correct least squares line, i.e. the line that minimizes the sum of squared residuals, through trial and error. Instead we can use the lm function in R to fit the linear model (a.k.a. regression line).

m1 <- lm(runs ~ at\_bats, data = mlb11)

The first argument in the function lm is a formula that takes the form y ~ x. Here it can be read that we want to make a linear model of runs as a function of at\_bats. The second argument specifies that R should look in the mlb11 data frame to find the runs and at\_bats variables.

The output of lm is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

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

Let’s consider this output piece by piece. First, the formula used to describe the model is shown at the top. After the formula you find the five-number summary of the residuals. The “Coefficients” table shown next is key; its first column displays the linear model’s y-intercept and the coefficient of at\_bats. With this table, we can write down the least squares regression line for the linear model:

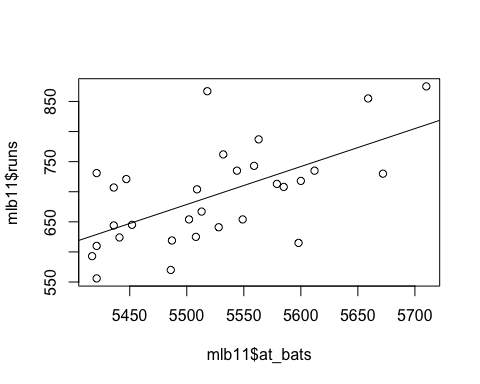
ŷ =−2789.2429+0.6305∗atbats

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, R2 . The R2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 37.3% of the variability in runs is explained by at-bats.

### Prediction and prediction errors

Let’s create a scatterplot with the least squares line laid on top.

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



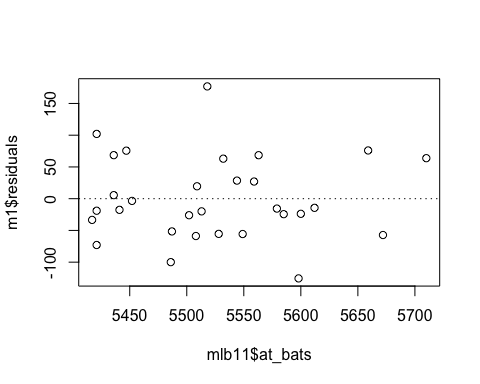
The function abline plots a line based on its slope and intercept. Here, we used a shortcut by providing the model m1, which contains both parameter estimates. This line can be used to predict y at any value of x . When predictions are made for values of x that are beyond the range of the observed data, it is referred to as extrapolation and is not usually recommended. However, predictions made within the range of the data are more reliable. They’re also used to compute the residuals.

### Model diagnostics

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

Linearity: You already checked if the relationship between runs and at-bats is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. at-bats. Recall that any code following a # is intended to be a comment that helps understand the code but is ignored by R.

plot(m1$residuals ~ mlb11$at\_bats)  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0



### Exercises

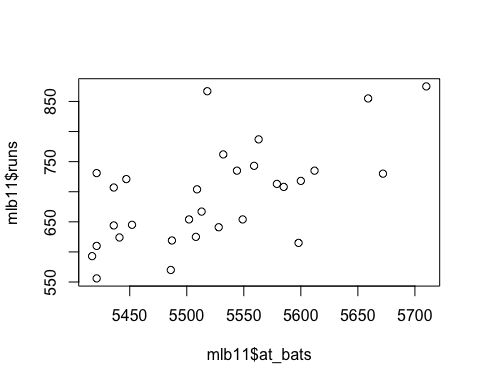
#### Exercise 1

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?

Answer 1)

I would use a scatter plot - they can be good for displaying numerical data like this. Per the plot below, I might be comfortable using the at\_bats - there appears to be an increase for in runs for an increase at\_bats.

plot(x = mlb11$at\_bats, y = mlb11$runs)



If the relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.

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.

Answer 2)

By looking at the above scatter plot, it is clear that the relationship is linear as the scatters are arranged some sort of a line, and the relationship between runs and at\_bats is a positive relationship as the slope is in upward directions(correlation is a positive value). However the strength of the relationship is not that strong as there are lot of variances of the scatters

#### 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?

Answer 3)

The smallest sums of squares got was 140241.6 compare to other four values (169310.1, 175082.9, 140264.5, 141743.7). But the minimum sums of squares value is 123722.0

#### 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?

Answer 4)

m2<-lm(formula = 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

According to the above summary, linear relationship between homeruns and runs is runs = 415.2389 + 1.8345\*homeruns

R- squared value for the model is 0.6266 which means 62.66% of data can be explained by this linear model

#### 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?

Answer 5)

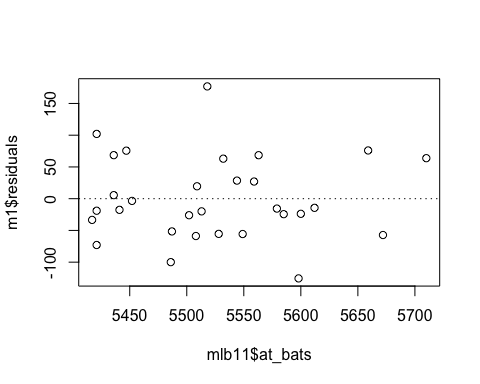
Relationship between runs and at\_bats is described by runs= -2789.2429 + 0.6305\* atbats . So for a at\_bats value of 5578, predicted runs should be 727.But for a at\_bats value of 5579, the number of runs scored is 713, so that the predicted value is an overestimate.

#### 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?

Answer 6)

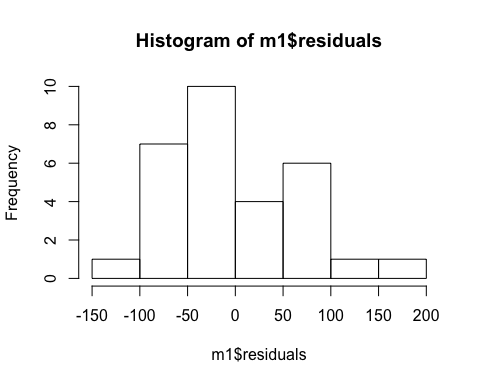
plot(m1$residuals ~ mlb11$at\_bats)  
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0



Points in the residual plot are randomly dispersed around the horizontal axis, which indicates that the linear model is appropriate.

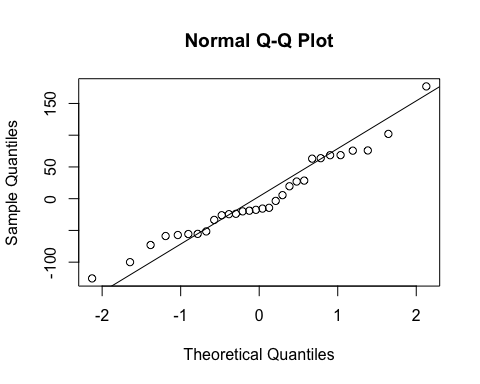
Nearly normal residuals: To check this condition, we can look at a histogram

hist(m1$residuals)



or a normal probability plot of the residuals.

qqnorm(m1$residuals)  
qqline(m1$residuals) # adds diagonal line to the normal prob plot



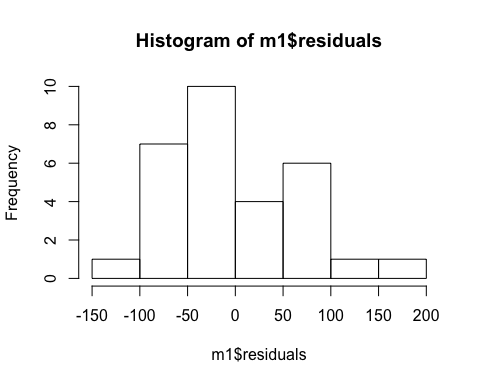
#### Exercise 7

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

Answer 7)

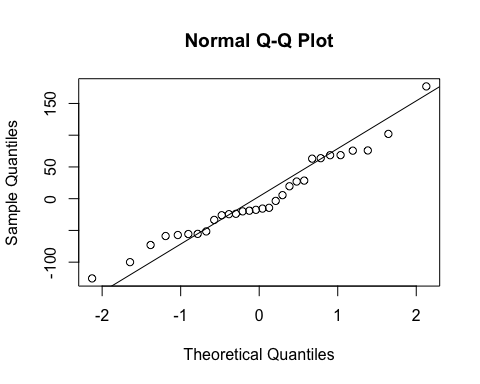
Nearly normal residuals: To check this condition, we can look at a histogram

hist(m1$residuals)



or a normal probability plot of the residuals.

qqnorm(m1$residuals)  
qqline(m1$residuals) # adds diagonal line to the normal prob plot



Yes the nearly normal residuals condition are met. The points in the qqline are all within the cloud and no obvious outliers can be found.

#### Exercise 8

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

Answer 8)

Points in the above residual plot are randomly dispersed and no patterns can be identified. Therefore the model satisfies the continuous variability condition

### On Your Own

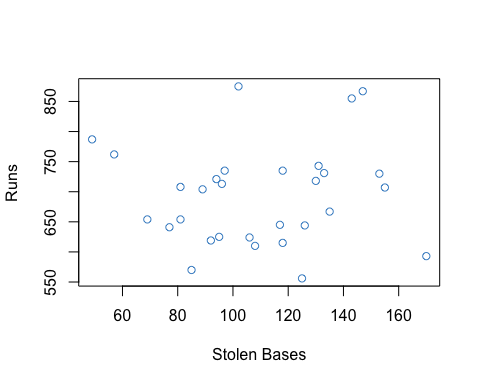
#### Ques\_a

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?

Answer a)

I have chosen Stolen Bases as a variable from mlb11

plot(mlb11$stolen\_bases,mlb11$runs, xlab = 'Stolen Bases', ylab = 'Runs' ,col = 'steelblue3')



There does seem to be a linear relationship.

#### Ques\_b

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?

Answer b)

m3 <- lm(mlb11$runs ~ mlb11$hits)  
m1<-lm(mlb11$runs ~ mlb11$at\_bats)  
summary(m1)

##   
## Call:  
## lm(formula = mlb11$runs ~ mlb11$at\_bats)  
##   
## 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 \*\*   
## mlb11$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 = mlb11$runs ~ mlb11$hits)  
##   
## 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 \*   
## mlb11$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

The R squared value for runs and at\_bats is 0.3729

The R squared value for runs and hits is 0.6419

With hits as the variable , we get a higher R square value, which represents the proportion of variability in the response variable is more when usinf hits as a varaible.

#### Ques\_c

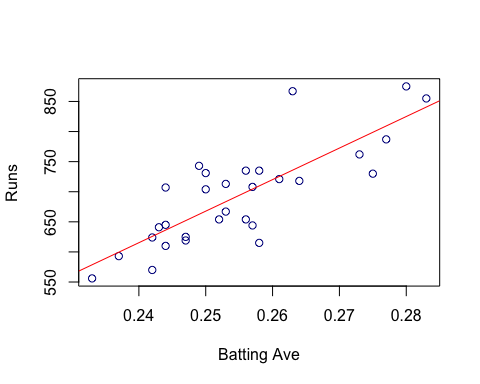
Now that you can summarize the linear relationship between two variables, investigate the relationships between runs and each of the other five traditional variables. Which variable best predicts runs? Support your conclusion using the graphical and numerical methods we’ve discussed (for the sake of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

Answer c)

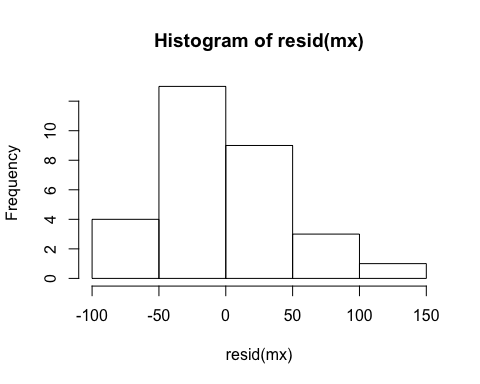
mx <- lm(mlb11$runs ~ mlb11$bat\_avg )  
summary(mx)

##   
## Call:  
## lm(formula = mlb11$runs ~ mlb11$bat\_avg)  
##   
## 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 \*\*   
## mlb11$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

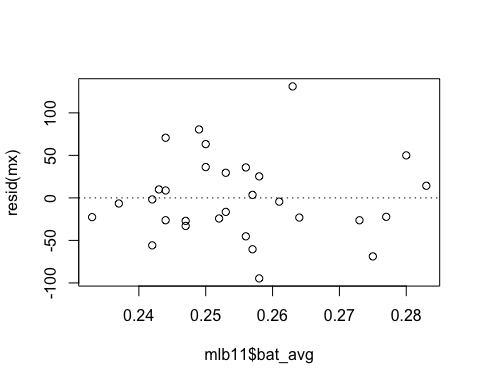
plot(mlb11$bat\_avg, mlb11$runs, xlab = 'Batting Ave', ylab= 'Runs', col = 'darkblue')  
abline(mx, col = 'red')



hist(resid(mx))



plot(mlb11$bat\_avg, resid(mx))  
abline(h = 0, lty = 3)



#### Ques\_d

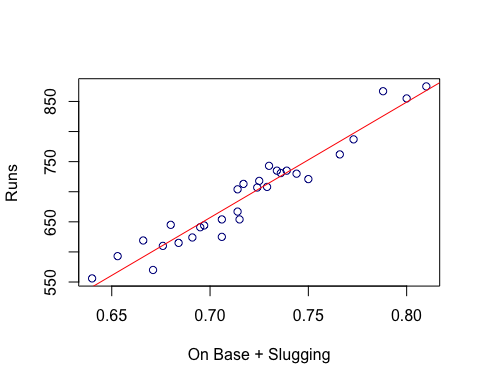
Now examine the three newer variables. These are the statistics used by the author of Moneyball to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we’ve analyzed, 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?

Answer d)

my <- lm(mlb11$runs ~ mlb11$new\_obs)  
summary(my)

##   
## Call:  
## lm(formula = mlb11$runs ~ mlb11$new\_obs)  
##   
## 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 \*\*\*  
## mlb11$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

plot(mlb11$new\_obs, mlb11$runs, xlab = 'On Base + Slugging', ylab= 'Runs', col = 'darkblue')  
abline(my, col = 'red')

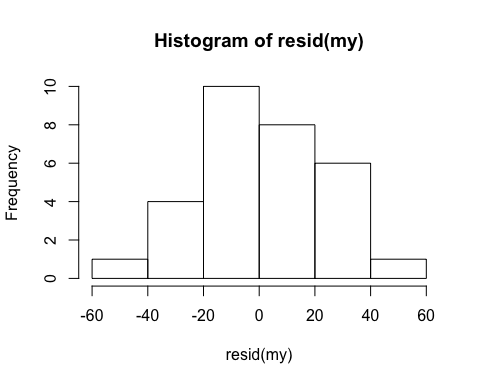


#### Ques\_e

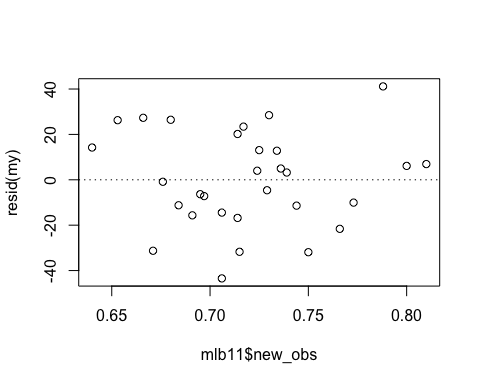
Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

Answer e)

hist(resid(my))



plot(mlb11$new\_obs, resid(my))  
abline(h = 0, lty = 3)



qqnorm(resid(my))  
qqline(resid(my))

