Lab 7 Introduction to linear regression

anjal hussan

12/3/2017

Batter up

The movie Moneyball (http://en.wikipedia.org/wiki/Moneyball_(film)) 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("mlb11.RData")
summary(mlb11)
```

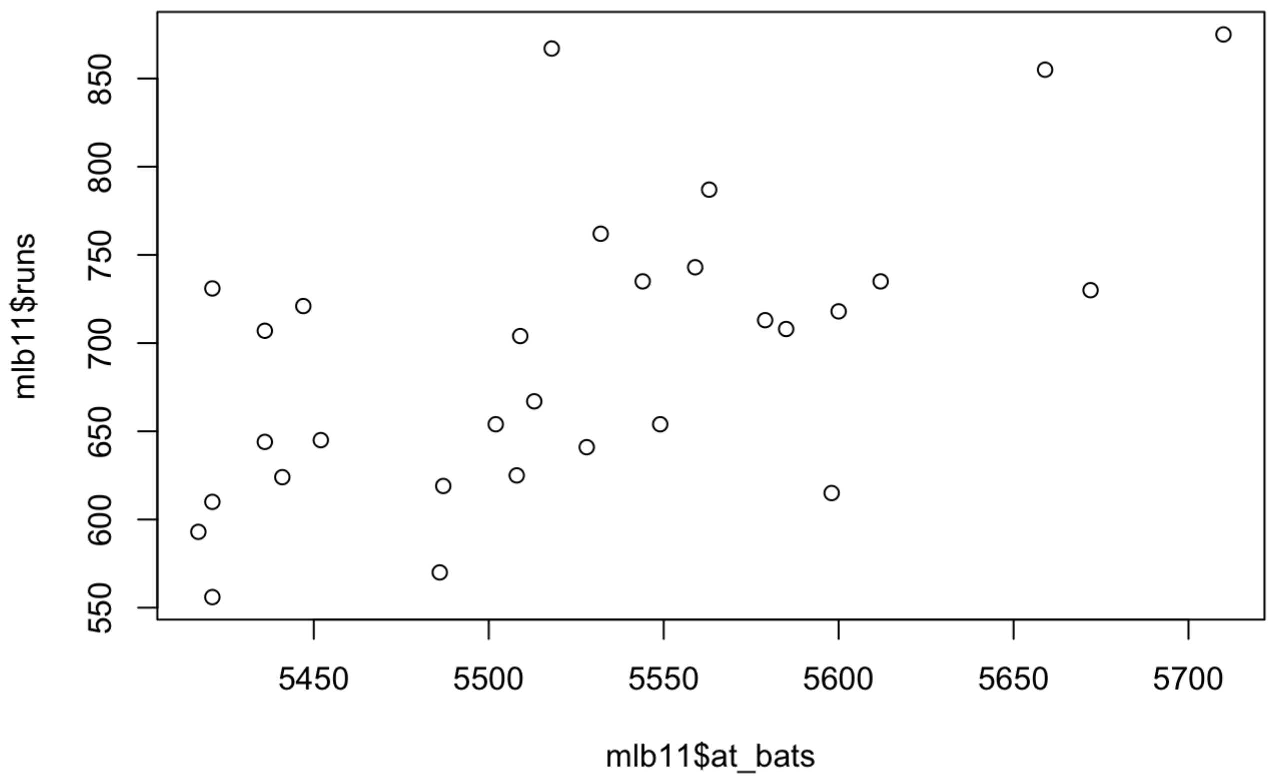
```
##                    team         runs          at_bats          hits     
##  Arizona Diamondbacks: 1   Min.   :556.0   Min.   :5417   Min.   :1263  
##  Atlanta Braves      : 1   1st Qu.:629.0   1st Qu.:5448   1st Qu.:1348  
##  Baltimore Orioles   : 1   Median :705.5   Median :5516   Median :1394  
##  Boston Red Sox      : 1   Mean   :693.6   Mean   :5524   Mean   :1409  
##  Chicago Cubs        : 1   3rd Qu.:734.0   3rd Qu.:5575   3rd Qu.:1441  
##  Chicago White Sox   : 1   Max.   :875.0   Max.   :5710   Max.   :1600  
##  (Other)             :24                                                
##     homeruns        bat_avg         strikeouts    stolen_bases   
##  Min.   : 91.0   Min.   :0.2330   Min.   : 930   Min.   : 49.00  
##  1st Qu.:118.0   1st Qu.:0.2447   1st Qu.:1085   1st Qu.: 89.75  
##  Median :154.0   Median :0.2530   Median :1140   Median :107.00  
##  Mean   :151.7   Mean   :0.2549   Mean   :1150   Mean   :109.30  
##  3rd Qu.:172.8   3rd Qu.:0.2602   3rd Qu.:1248   3rd Qu.:130.75  
##  Max.   :222.0   Max.   :0.2830   Max.   :1323   Max.   :170.00  
##                                                                  
##       wins          new_onbase        new_slug         new_obs      
##  Min.   : 56.00   Min.   :0.2920   Min.   :0.3480   Min.   :0.6400  
##  1st Qu.: 72.00   1st Qu.:0.3110   1st Qu.:0.3770   1st Qu.:0.6920  
##  Median : 80.00   Median :0.3185   Median :0.3985   Median :0.7160  
##  Mean   : 80.97   Mean   :0.3205   Mean   :0.3988   Mean   :0.7191  
##  3rd Qu.: 90.00   3rd Qu.:0.3282   3rd Qu.:0.4130   3rd Qu.:0.7382  
##  Max.   :102.00   Max.   :0.3490   Max.   :0.4610   Max.   :0.8100  
## 
```

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.

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?

A scatter plot is good starting point to visualize runs vs. at bats.

```
plot(mlb11$at_bats,mlb11$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
```

Sum of squared residuals

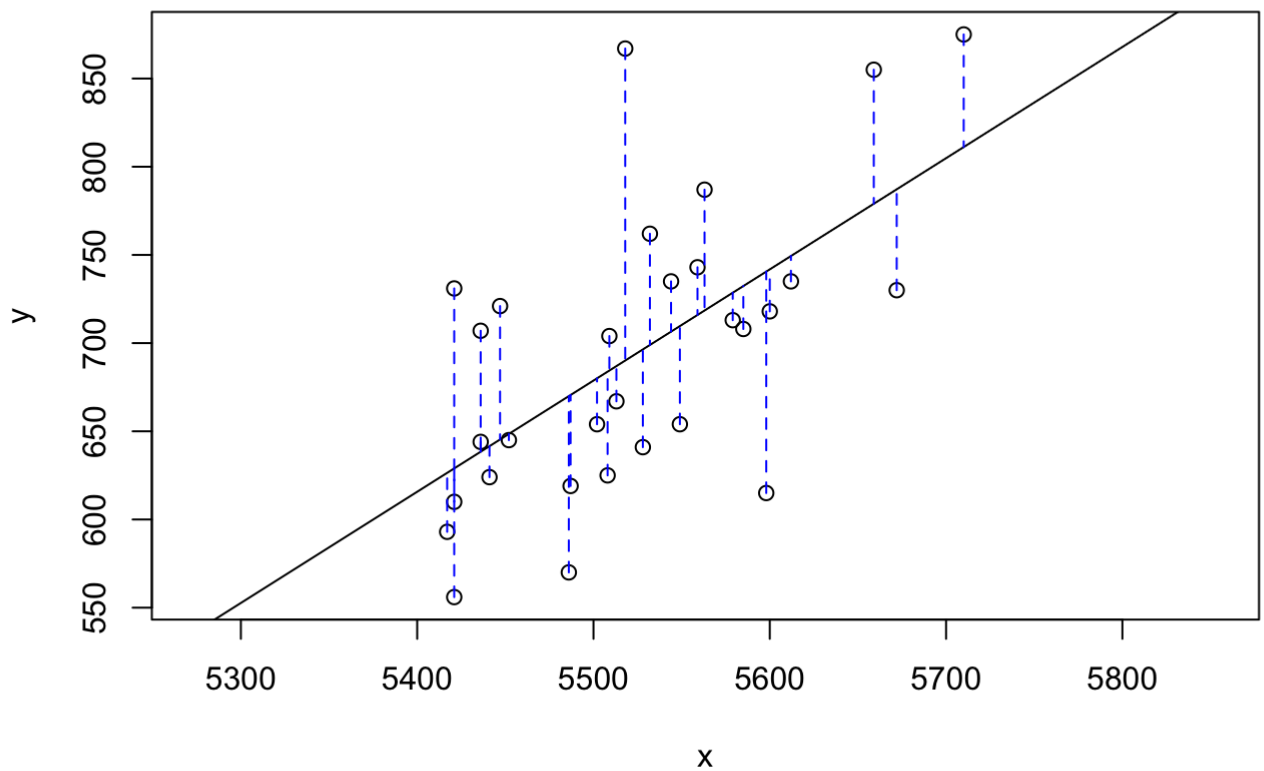
Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as `runs` and `at_bats` above.

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.

The relationships is linear; we can draw a straight line through the data with almost equal numbers of data points on the both side of the line. The relationship is strong there is scatter, but there is a definite slope to the trend line.

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)
```



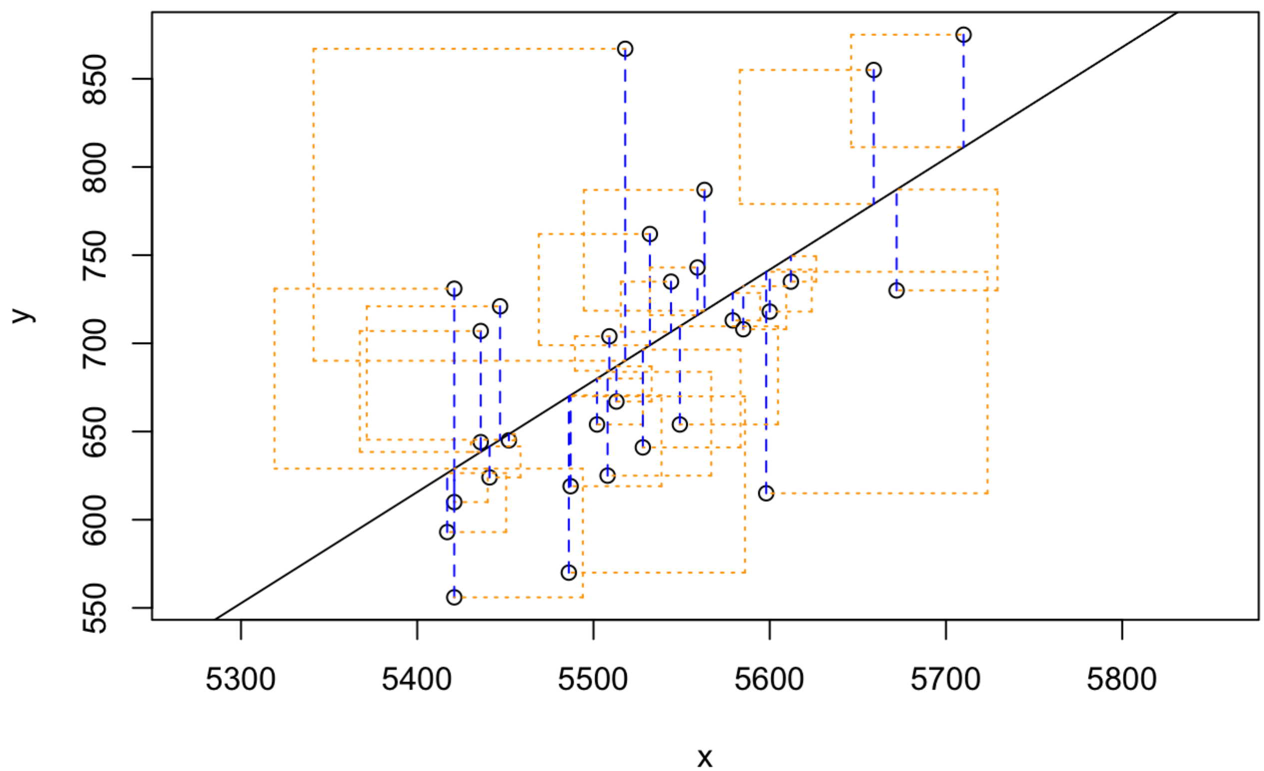

```
## Click two points to make a line.

## 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:

$$e_i = y_i - \hat{y}_i$$

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)
```

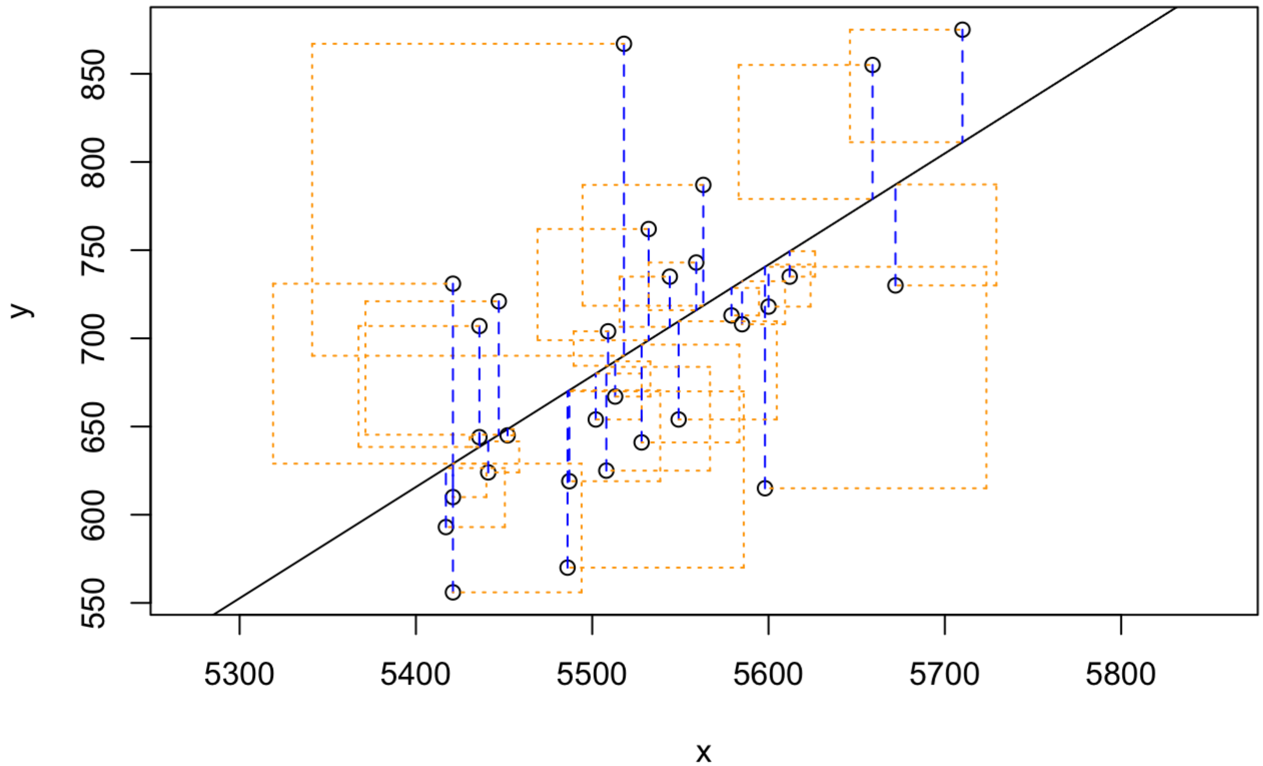



```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##  -2789.2429       0.6305
##
## Sum of Squares:  123721.9
```

Note that the output from the `plot_ss` function provides you with the slope and intercept of your line as well as the sum of squares.

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(mlb11$at_bats,mlb11$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:

$$\hat{y} = -2789.2429 + 0.6305 * atbats$$

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, R^2. The R^2 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.

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?

```
m2 <- lm(mlb11$runs ~ mlb11$homeruns)
summary(m2)
```

```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$homeruns)
##
## 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 ***
## mlb11$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
```

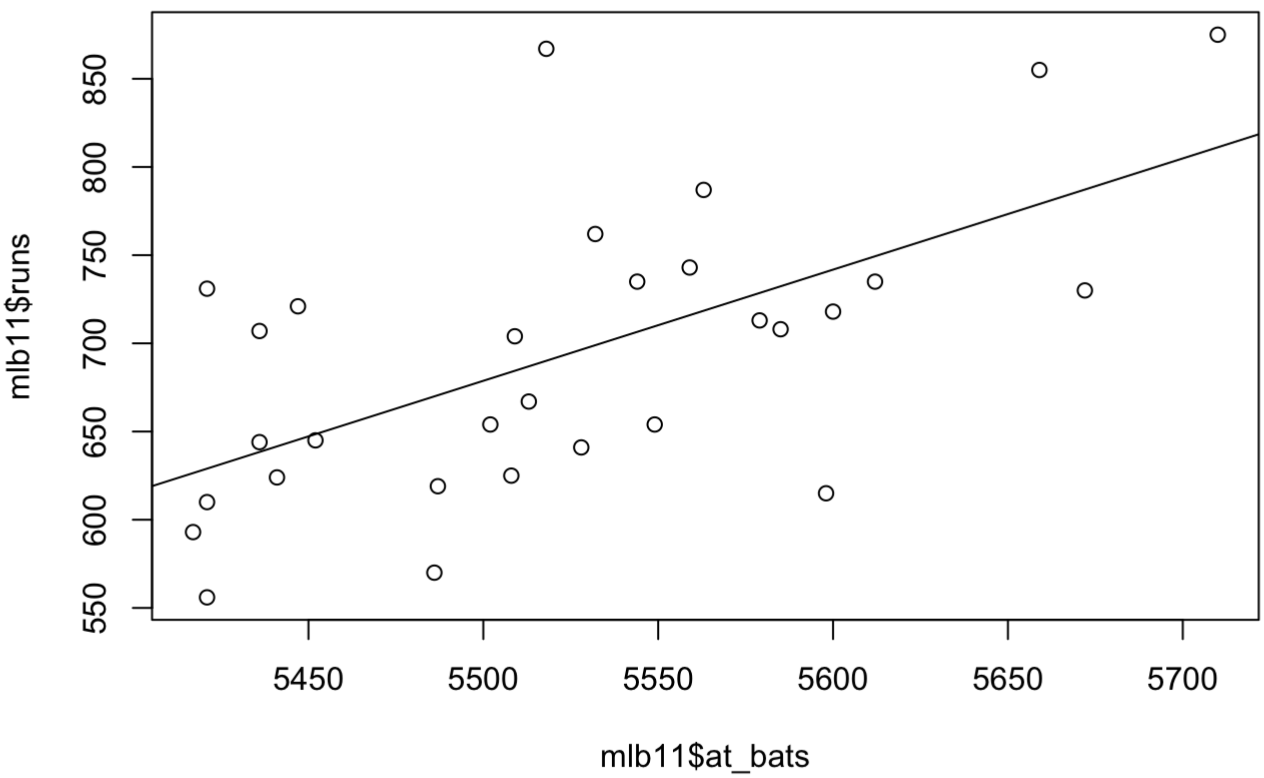
$y = 415.2389 + 1.8345 * homeruns$

The more home runs a team has the more successful the team

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)
```

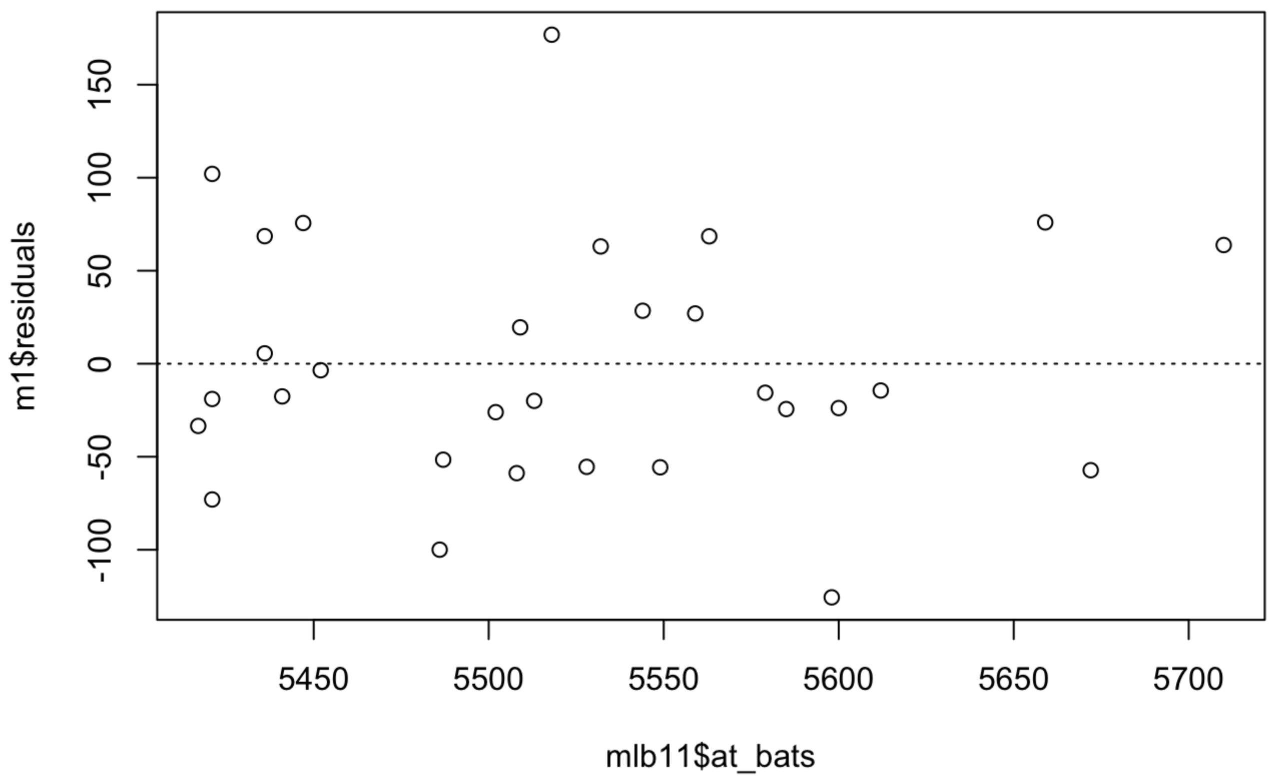


Model diagnostics

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
```

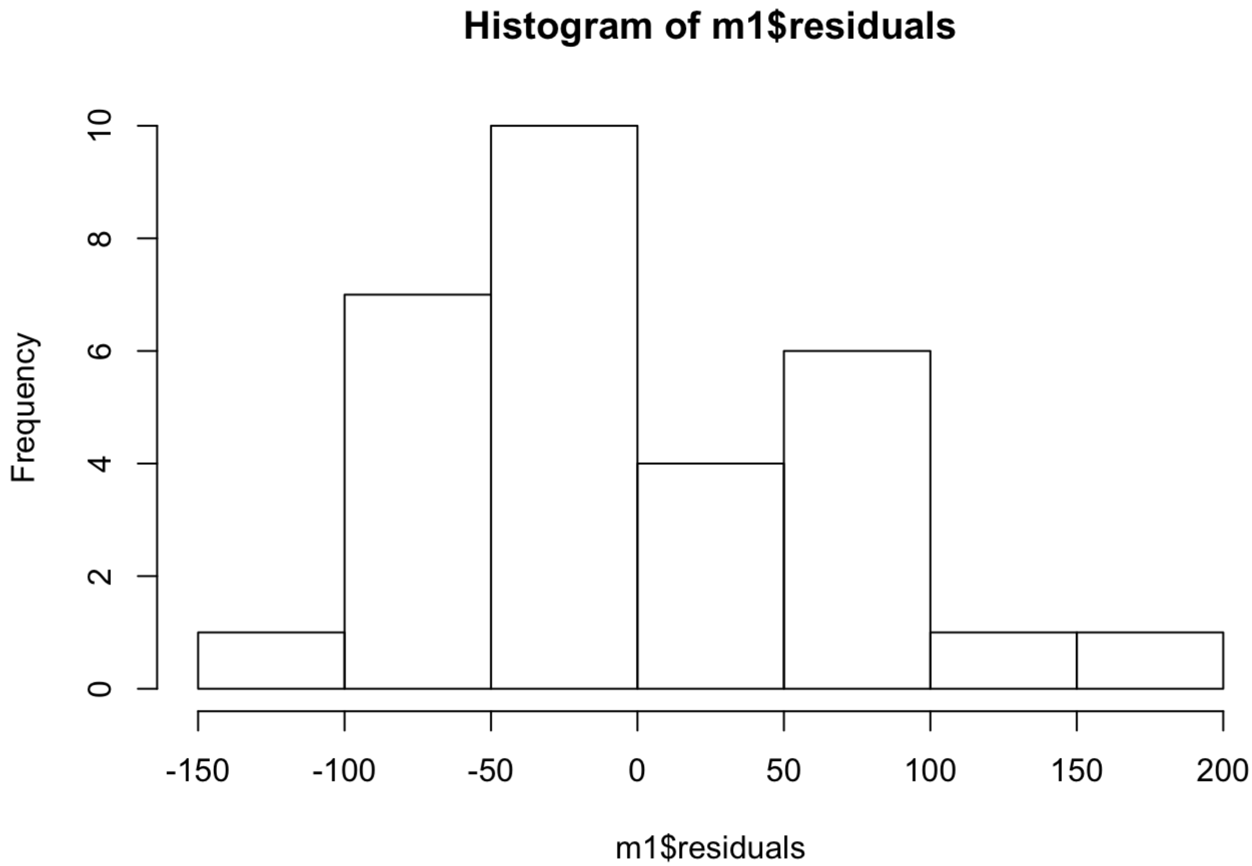


6. Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

   Ans: There is no pattern in the residuals. The data seems fairly evenly distributed above and below the trend line.

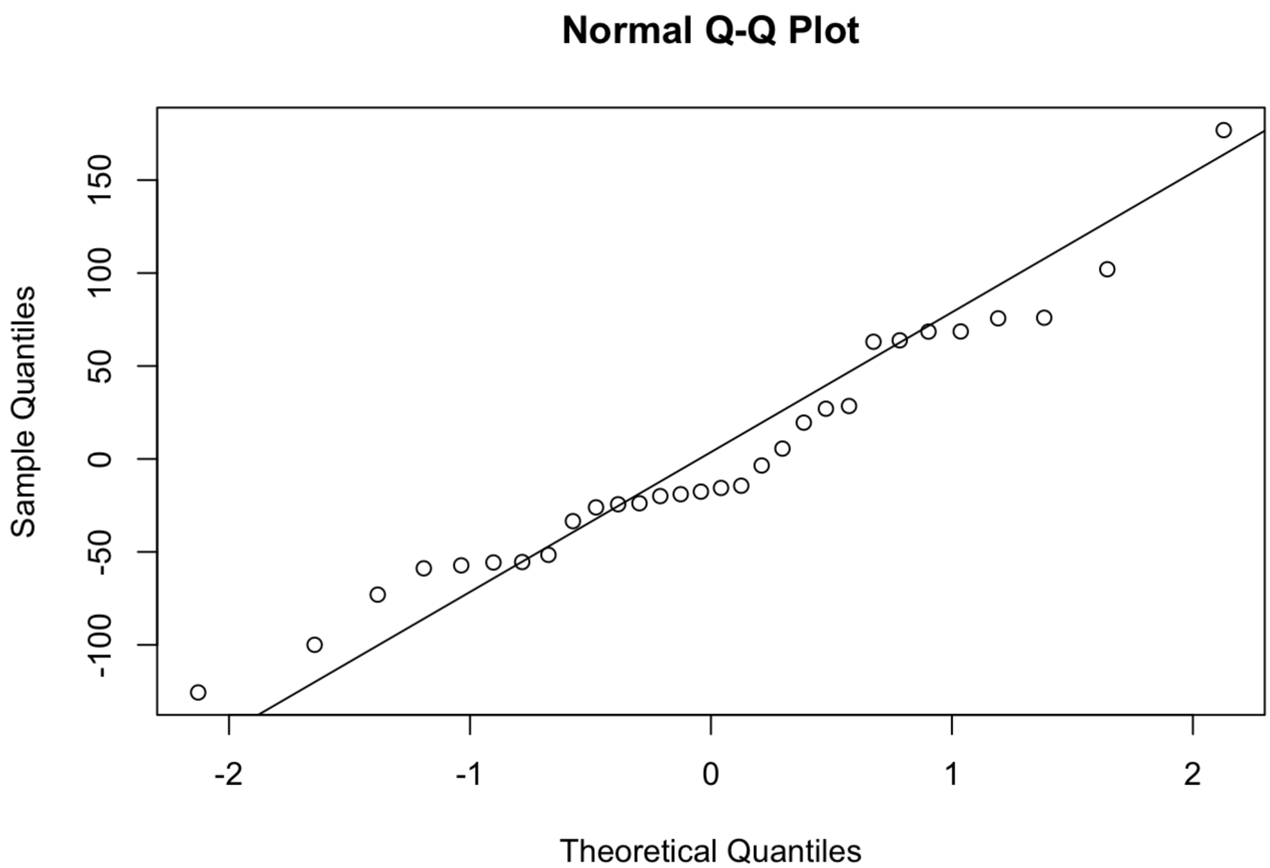
Nearly normal residuals: To check this condition, we can look at a histogram

```
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
```



7. Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met? Ans: The models are valid and the QQ Normal plot seems to follow the theoretical values. Hence, the normal residual condition have been met.

Constant variability:

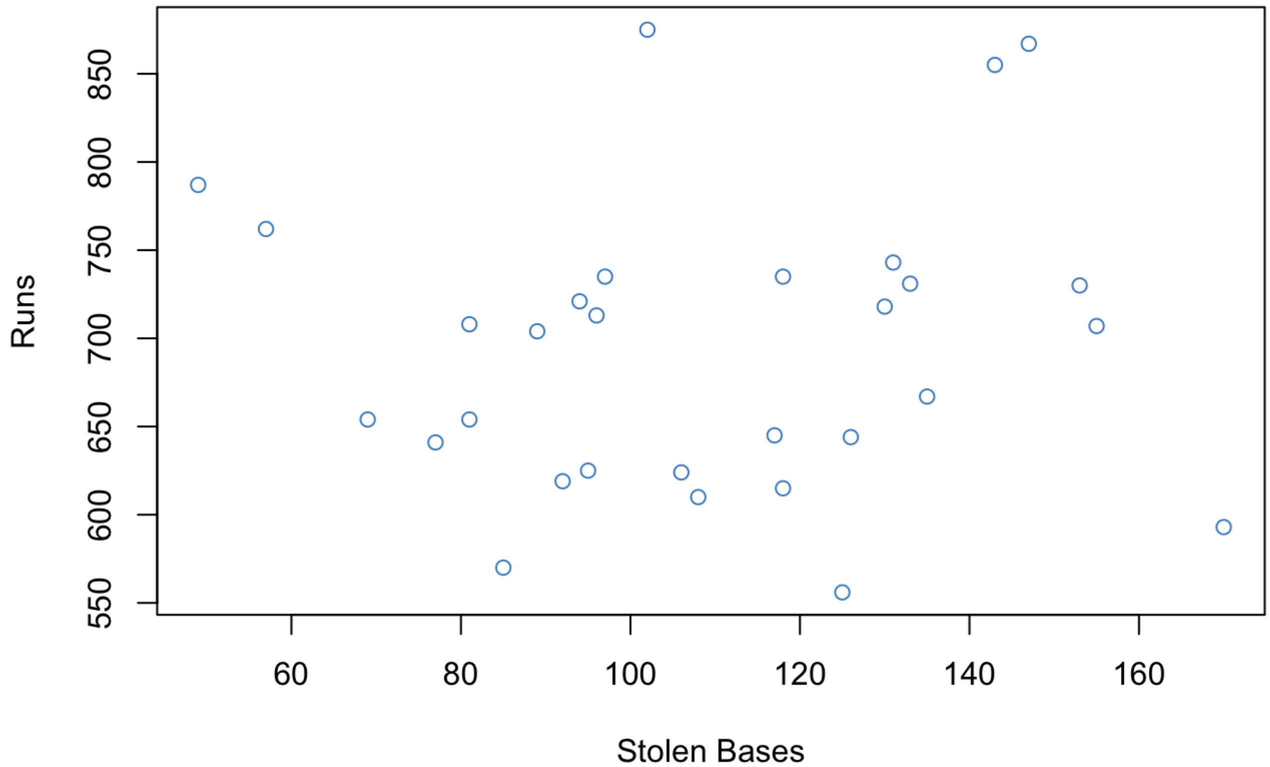
8. Based on the plot in (1), does the constant variability condition appear to be met? Ans: When the residuals are plotted they randomly vary form -100 to 100 with the exception of 1 point at 150. variability condition seems like fulfilled

---

On Your Own

- 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?

```
plot(mlb11$stolen_bases,mlb11$runs, xlab = 'Stolen Bases', ylab = 'Runs' ,col = 'steelblue3')
```



There does not seem to be a linear relationship.

- How does this relationship compare to the relationship between `runs` and `at_bats` ? Use the R^2 values from the two model summaries to compare. Does your variable seem to predict `runs` better than `at_bats` ? How can you tell?

```
m00 <- lm(mlb11$runs ~ mlb11$stolen_bases)
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
```

```
print("============================================================================")
```

```
## [1] "============================================================================"
```

```
summary(m00)
```

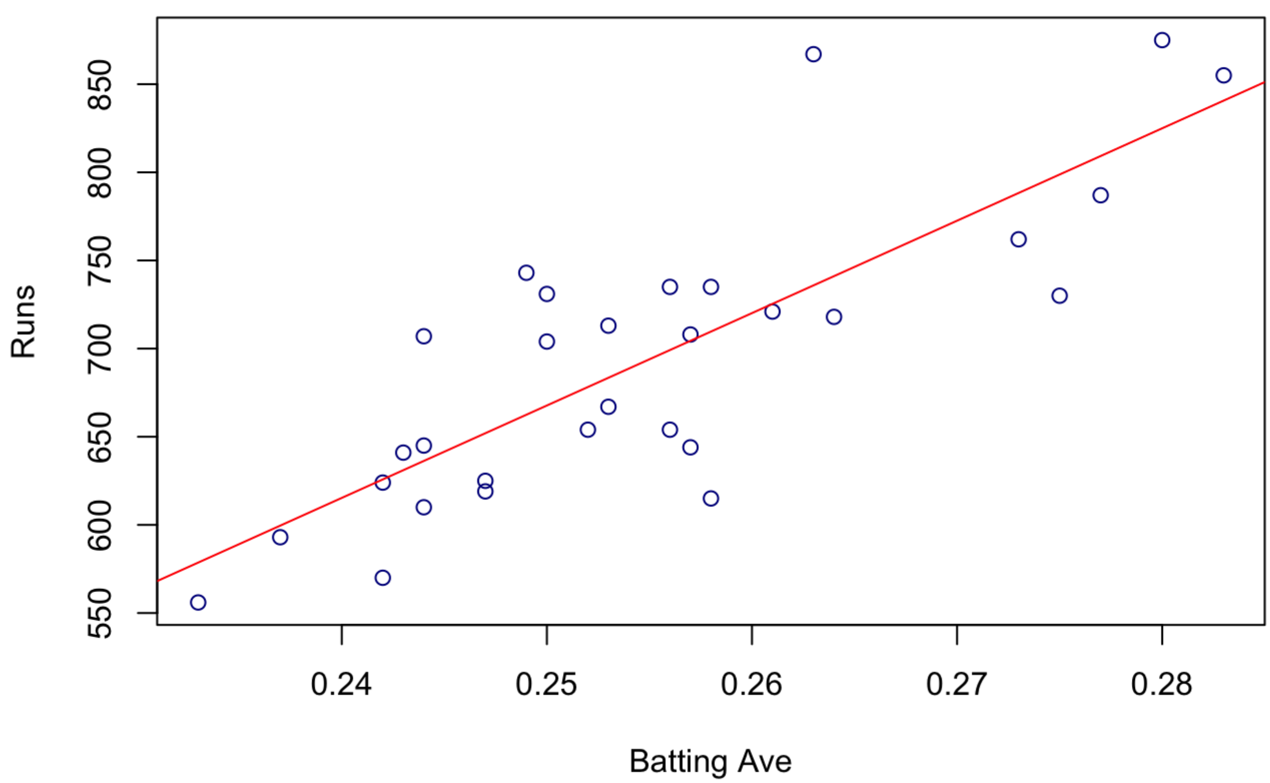
```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$stolen_bases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -139.94  -62.87   10.01   38.54  182.49
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        677.3074    58.9751  11.485 4.17e-12 ***
## mlb11$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
```

- 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).

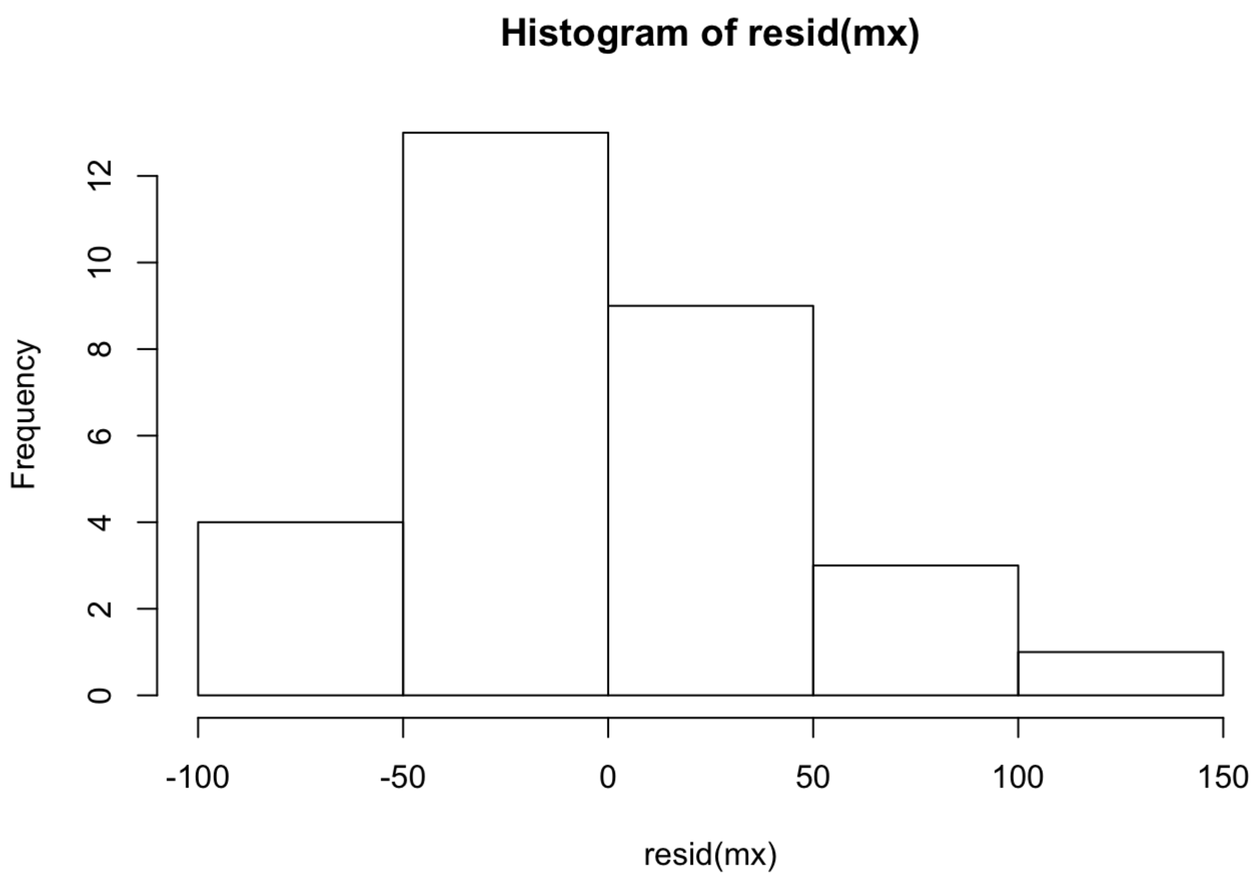
```
mx <- lm(mlb11$runs ~  mlb11$bat_avg )
summary(mx)
```

```
##
## 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
```

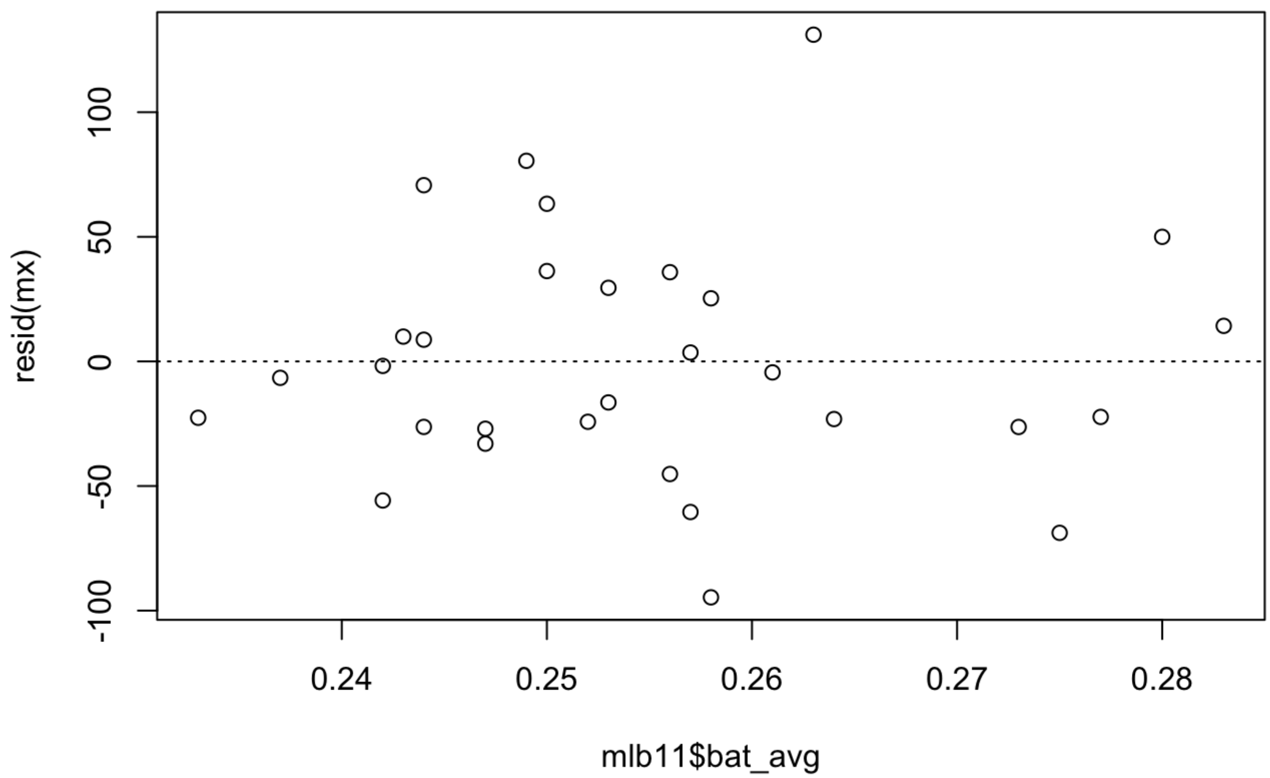
```
plot(mlb11$bat_avg, mlb11$runs, xlab = 'Batting Ave', ylab= 'Runs', col = 'darkblue')
abline(mx, col = 'red')
```



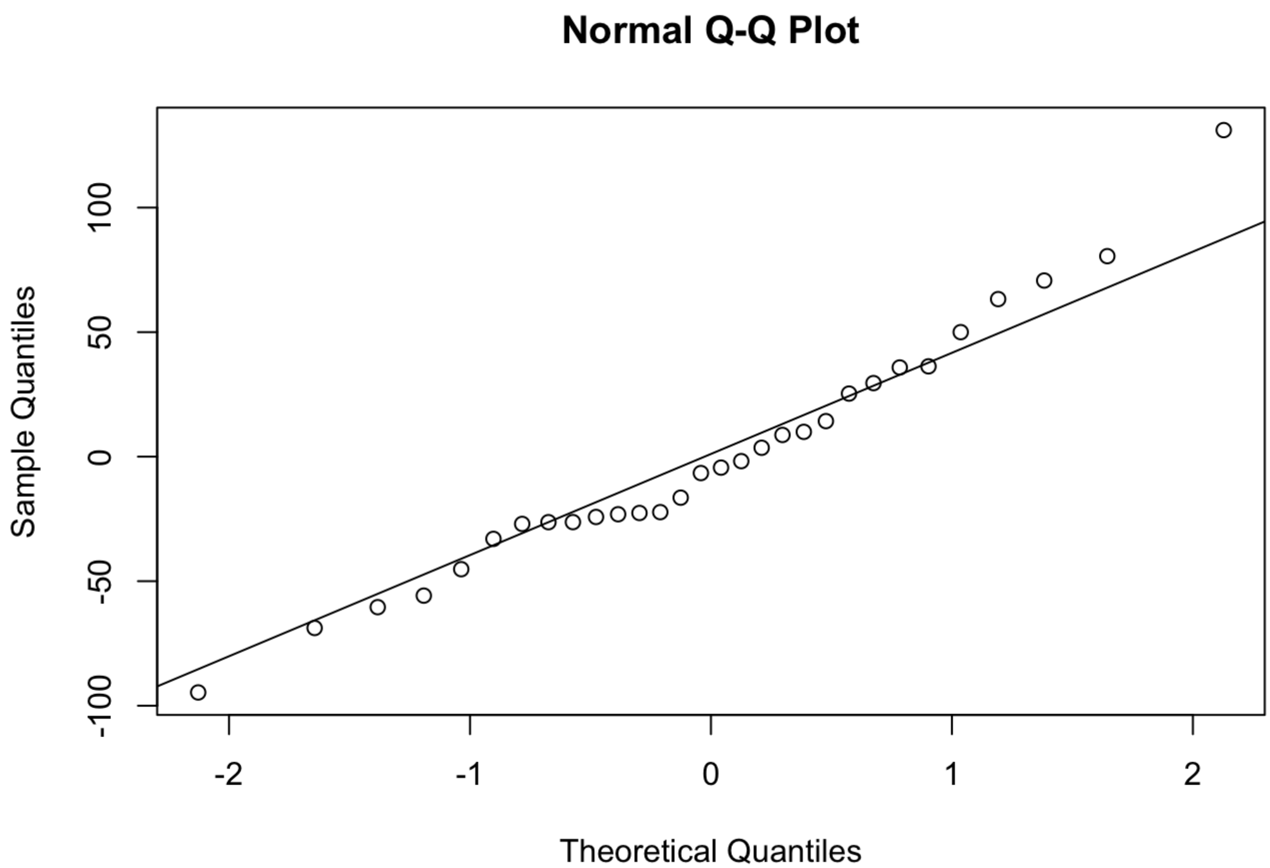
```
hist(resid(mx))
```




```
plot(mlb11$bat_avg, resid(mx))
abline(h = 0, lty = 3)
```




```
qqnorm(resid(mx))
qqline(resid(mx))
```

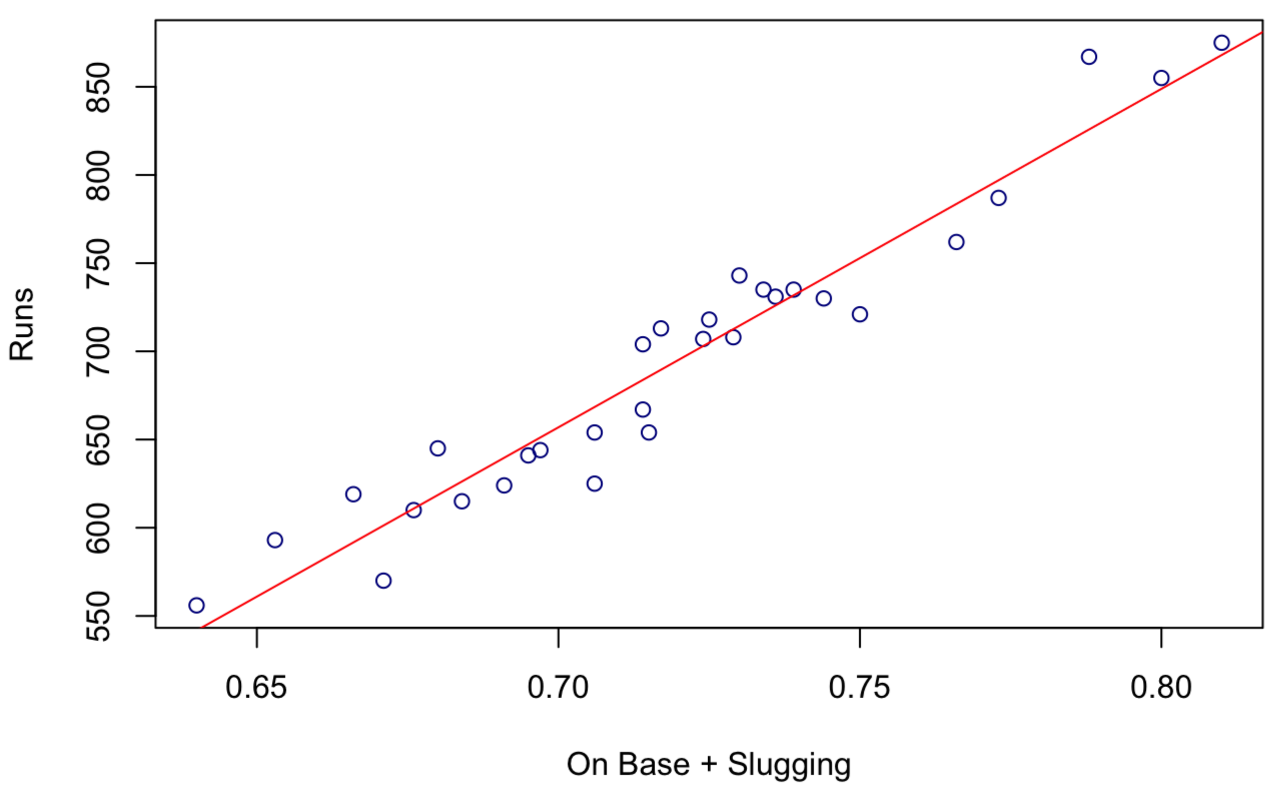


- Now examine the three newer variables. These are the statistics used by the author of *Moneyball* to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we've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

```
my <- lm(mlb11$runs ~  mlb11$new_obs)
summary(my)
```

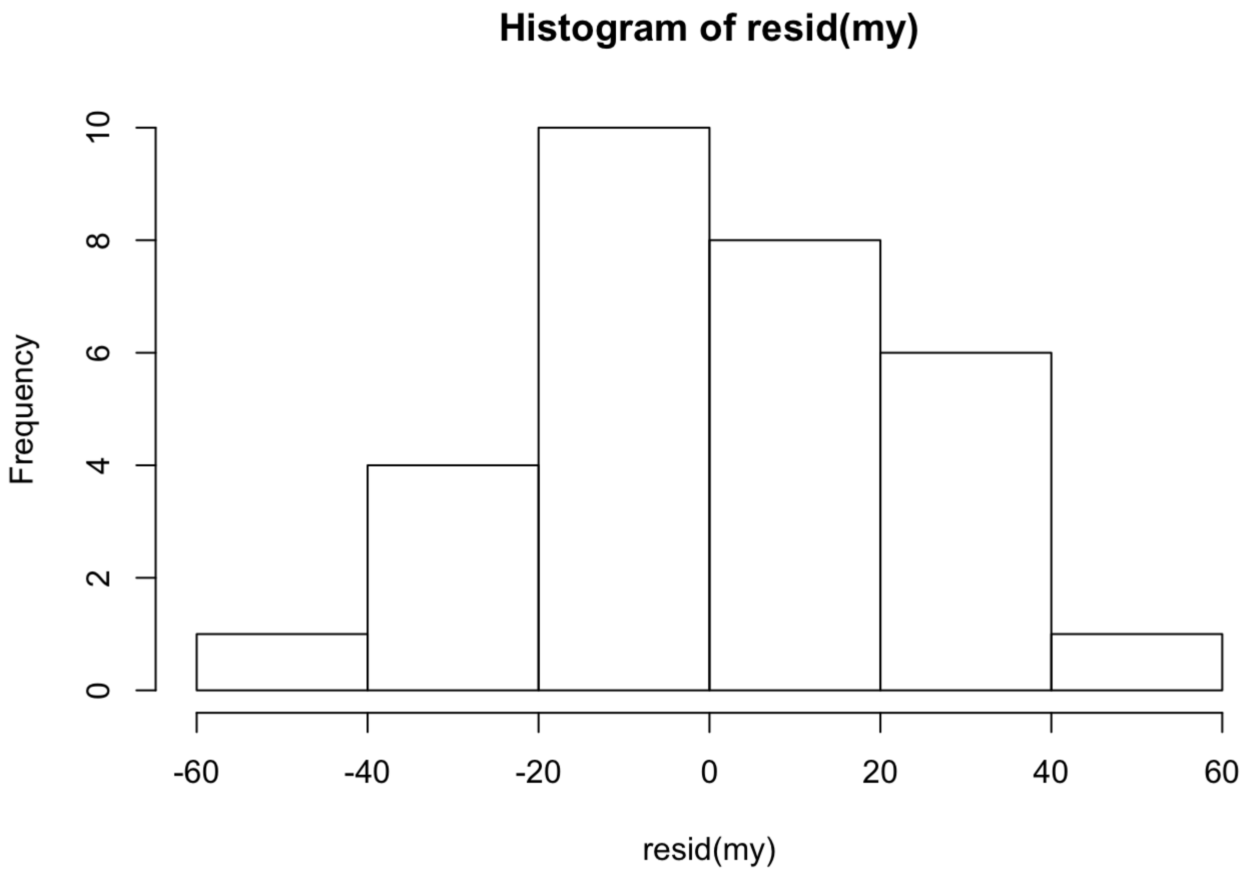
```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$new_obs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.456 -13.690   1.165  13.935  41.156
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -686.61      68.93  -9.962 1.05e-10 ***
## mlb11$new_obs  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 = 'd
arkblue')
abline(my, col = 'red')
```

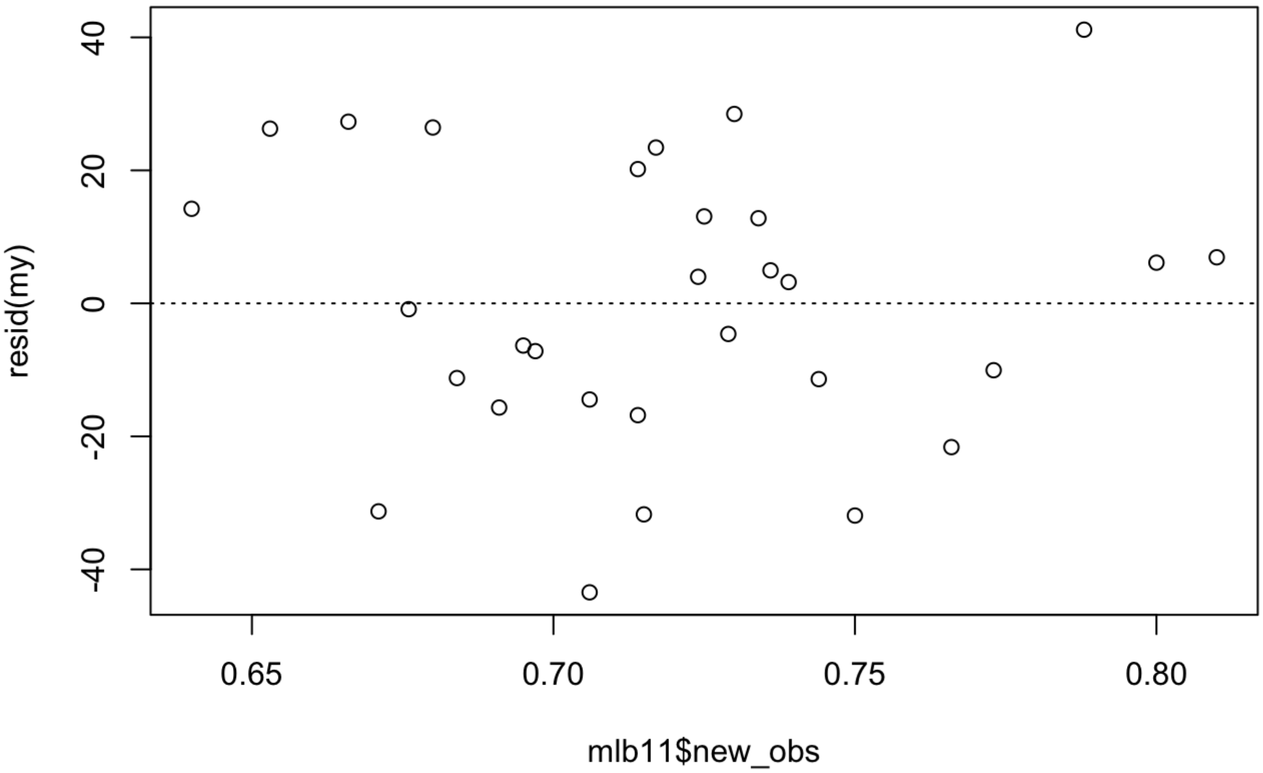


- Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

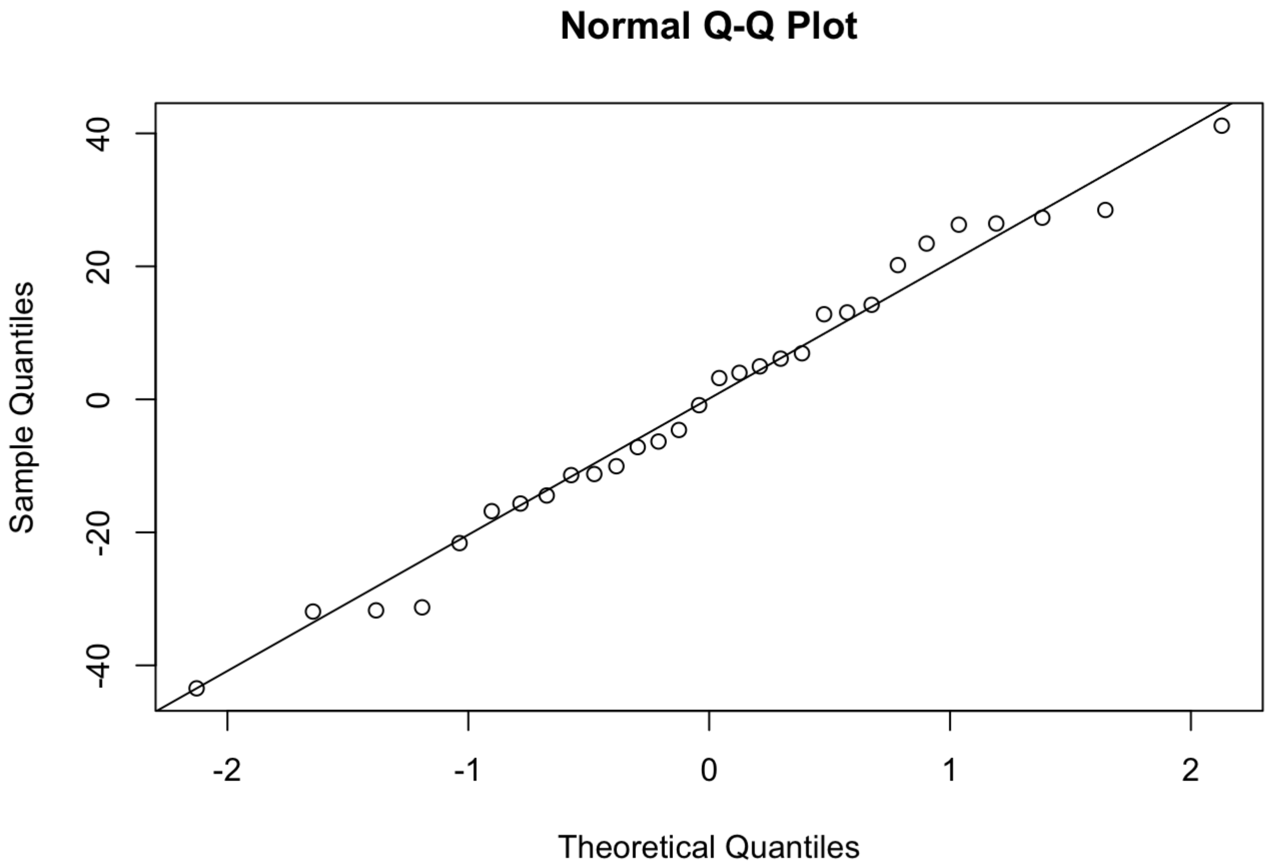
```
hist(resid(my))
```




```
plot(mlb11$new_obs, resid(my))
abline(h = 0, lty = 3)
```




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


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-40
-20
0
20
40
-2
-1
0
1
2