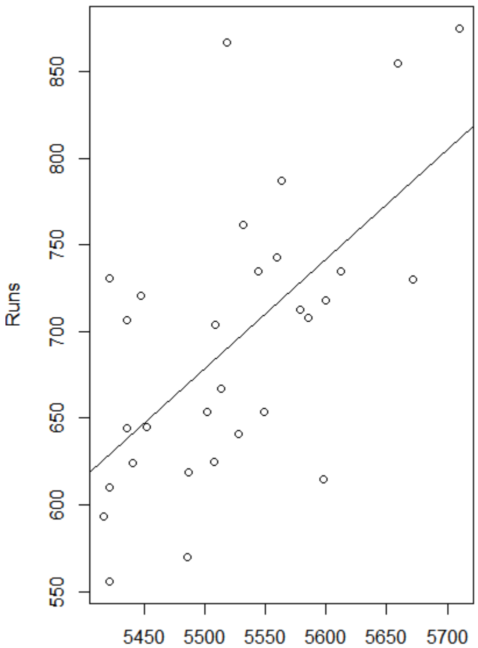
Name: Samuel L. Peoples

R Lab 7: Introduction to Linear Regression

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?

```
> plot(mlb11$runs ~ mlb11$at_bats, xlab = "At Bats", ylab = "Runs")
> abline(lm(mlb11$runs ~ mlb11$at_bats))
```



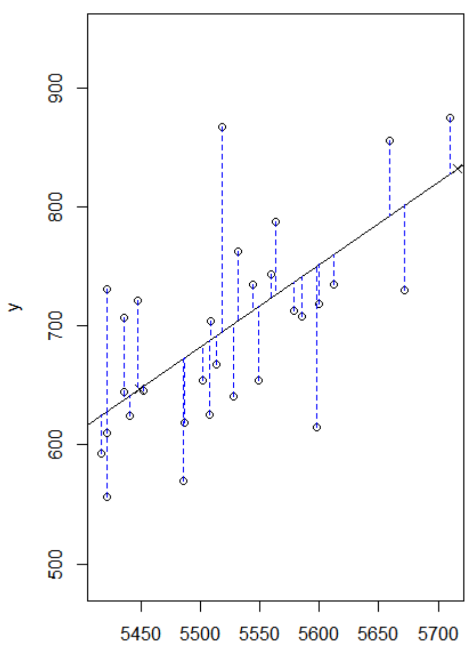
There appears to be a weak linear relationship between the two variables and does not appear to be strong enough to predict runs using a linear model.

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.

Plotting the trend line shows that there appears to be a very weak positive linear relationship, with many potential outliers falling quite far from the trend line.

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?

```
> plot_ss(mlb11$at_bats, mlb11$runs)

Call:
lm(formula = y ~ x, data = pts)

Coefficients:
(Intercept)            x
 -3122.3418       0.6918

Sum of Squares:  125157.2
```

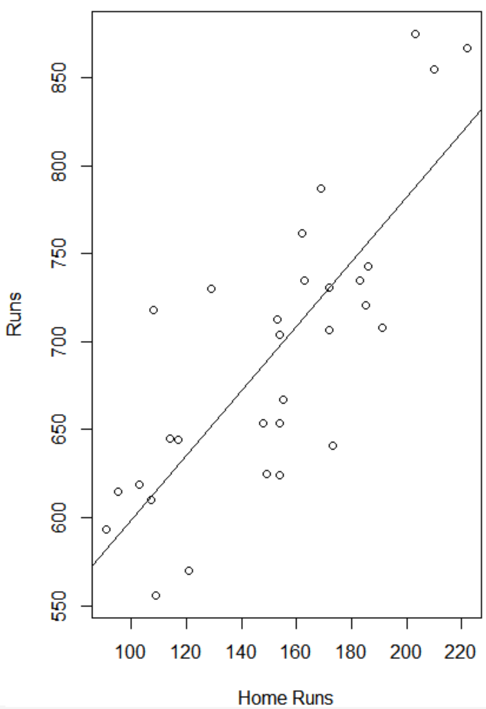


The smallest sum of squares that I was able to achieve was 125157.2

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?

```
> plot(mlb11$runs ~ mlb11$homeruns, xlab = "Home Runs", ylab = "Runs")
> abline(lm(mlb11$runs ~ mlb11$homeruns))
> cor(mlb11$runs, mlb11$homeruns)
                                                                              0.7915577
> summary(lm(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 ***
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
```

Using the data above, we can say that y = Estimate(Intercept) + Estimate(homeruns) (x) = 415.24 + 1.83(Homeruns)



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?

```
> summary(lm(mlb11$runs ~ mlb11$at_bats))

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
```

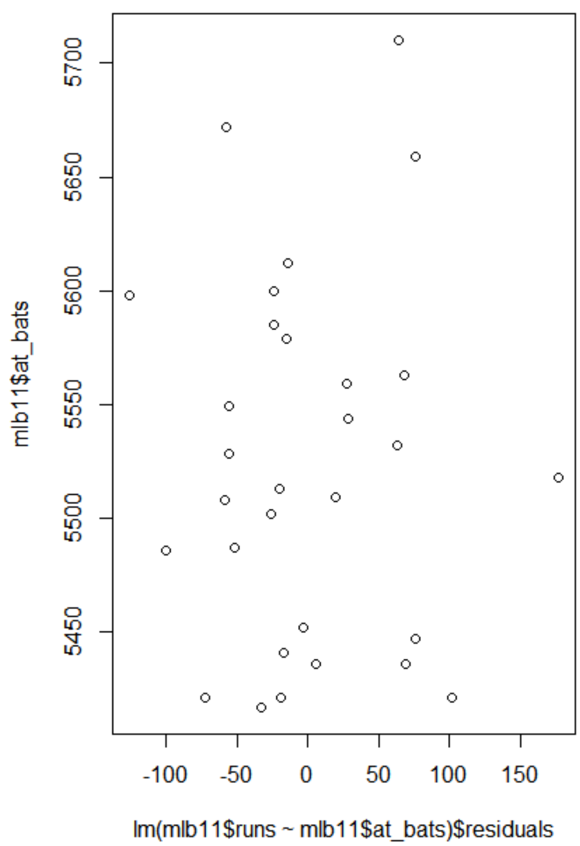
```
Residual standard error: 66.47 on 28 degrees of freedom
Multiple R-squared:  0.3729,   Adjusted R-squared:  0.3505
F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

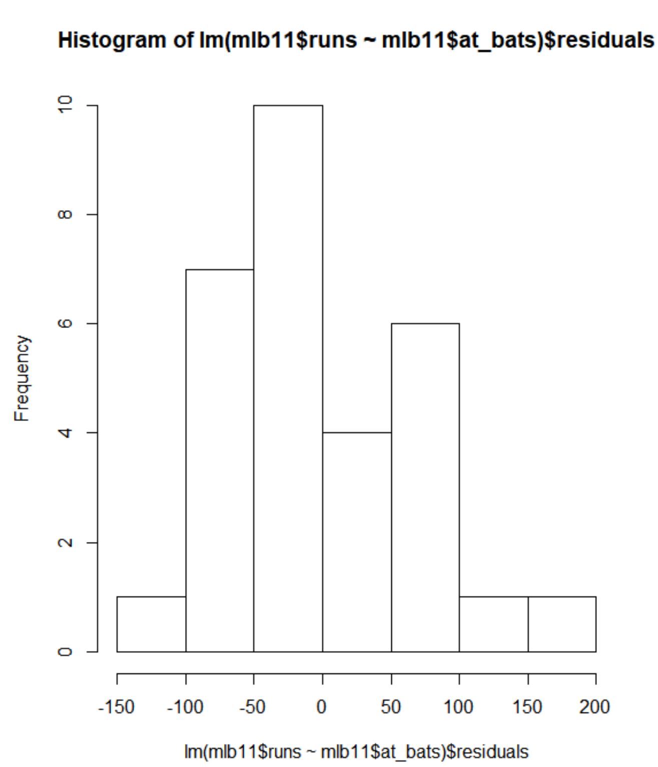
Using the equation -2789.24 + .63*5578= 727.69 we can predict that given 5578 at-bats there will be approximately 713 runs, and we can observe line 16 of the data for the "Philadelphia Phillies" has 713 runs for 5579 at-bats; providing an overestimation of 15 runs.

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?

```
> plot(lm(mlb11$runs~mlb11$at_bats)$residuals, mlb11$at_bats)
> cor(lm(mlb11$runs~mlb11$at_bats)$residuals, mlb11$at_bats)
                                                          1.386089e-15
```

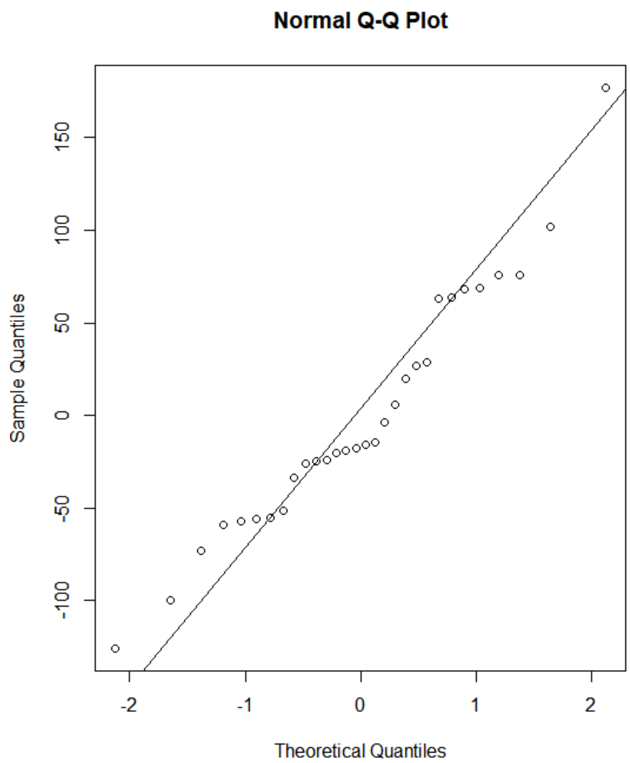
There does not seem to be a relationship or trend between the residuals, which suggests that the relationship between runs and at-bats is linearly related.





Exercise 7: Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

```
> hist(lm(mlb11$runs~mlb11$at_bats)$residuals)
> qqnorm(lm(mlb11$runs~mlb11$at_bats)$residuals)
> qqline(lm(mlb11$runs~mlb11$at_bats)$residuals)
```



The residuals appear to be distributed normally so the nearly normal residuals conditions are met.

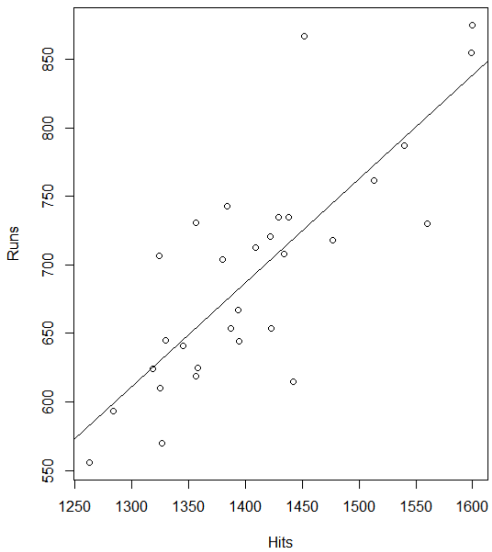
Exercise 8: Based on the plot in (1), does the constant variability condition appear to be met?

There appears to be constant variability from the least squares line.

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?

```
> plot(mlb11$runs ~ mlb11$hits, xlab = "Hits", ylab = "Runs")
> abline(lm(runs ~ hits, data = mlb11))
> summary(lm(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
(Intercept) -375.5600   151.1806  -2.484
mlb11$hits     0.7589     0.1071   7.085
            Pr(>|t|)
(Intercept)   0.0192 *
mlb11$hits  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
```



We can see a weak linear relationship between runs and hits, with linear regression formula of -375.56 + .7589 * Hits

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?

The values for Runs and At_Bats is

```
Multiple R-squared:  0.3729, Adjusted R-
squared:  0.3505
```

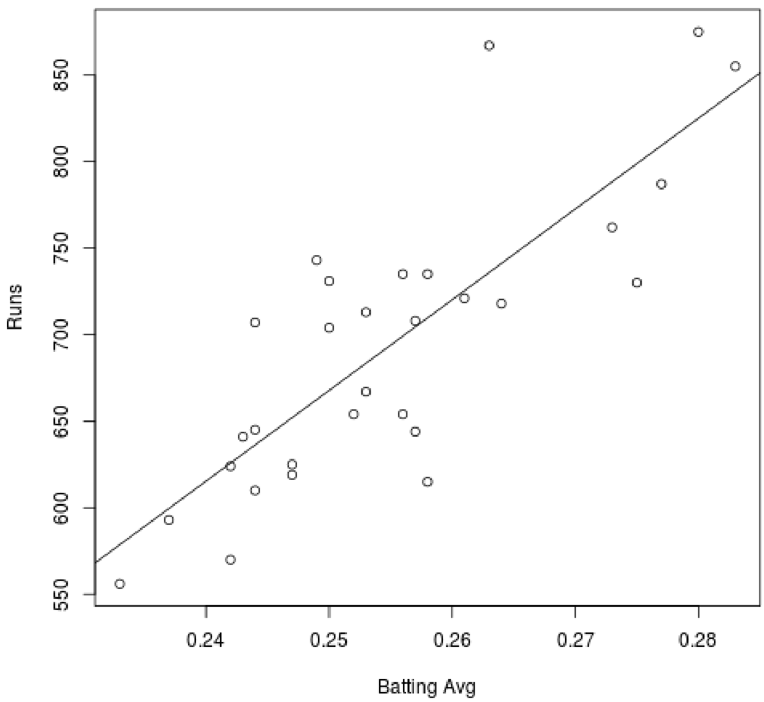
Where the values for Runs and Hits is

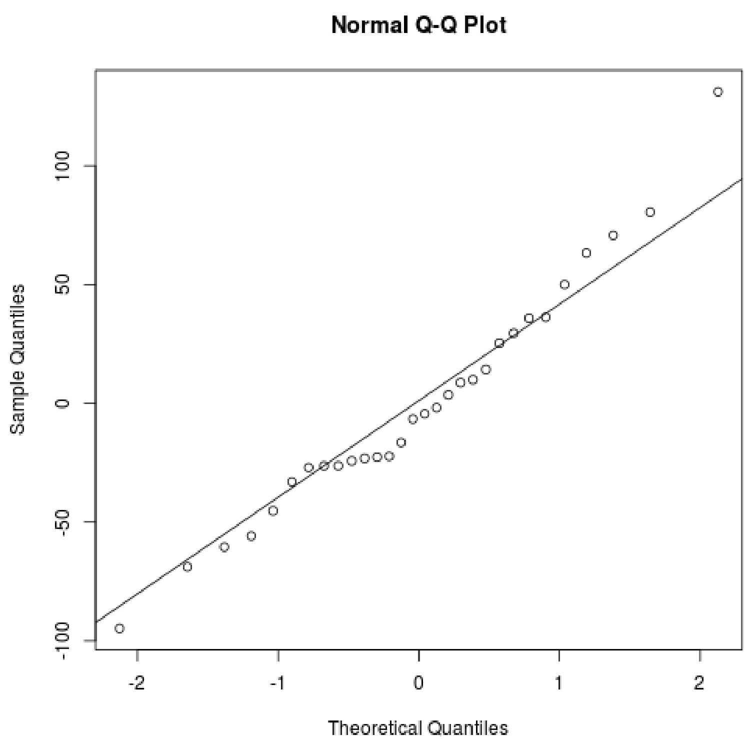
```
Multiple R-squared:  0.6419, Adjusted R-
squared:  0.6292
```

The relationship is nearly twice as strong with the Runs and Hits model, which suggests that this model predicts runs with more accuracy than Runs and At_Bats.

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).

```
plot(mlb11$runs ~ mlb11$bat_avg, xlab = "Batting Avg", ylab = "Runs")
abline(lm(mlb11$runs ~ mlb11$bat_avg))
qqnorm(lm(mlb11$runs ~ mlb11$bat_avg)$residuals)
qqline(lm(mlb11$runs ~ mlb11$bat_avg)$residuals)
summary(lm(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 **
 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
```





After verifying that there is a linear relationship between Batting Average and Runs, we can see an R value of .6561, which is the most accurate model for predicting runs.

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?

```
> summary(lm(mlb11$runs ~ mlb11$new_onbase))

Call:
lm(formula = mlb11$runs ~ mlb11$new_onbase)

Residuals:
    Min      1Q  Median      3Q     Max 
-58.270 -18.335   3.249  19.520  69.002 

Coefficients:
                 Estimate Std. Error t value
(Intercept)       -1118.4      144.5  -7.741
mlb11$new_onbase   5654.3      450.5  12.552
                 Pr(>|t|)    
(Intercept)      1.97e-08 ***
mlb11$new_onbase 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(lm(mlb11$runs ~ mlb11$new_slug))

Call:
lm(formula = mlb11$runs ~ mlb11$new_slug)

Residuals:
   Min     1Q Median     3Q    Max 
-45.41 -18.66  -0.91  16.29  52.29 

Coefficients:
               Estimate Std. Error t value
(Intercept)     -375.80      68.71   -5.47
mlb11$new_slug  2681.33     171.83   15.61
               Pr(>|t|)    
(Intercept)    7.70e-06 ***
mlb11$new_slug 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(lm(mlb11$runs ~ mlb11$new_obs)
```

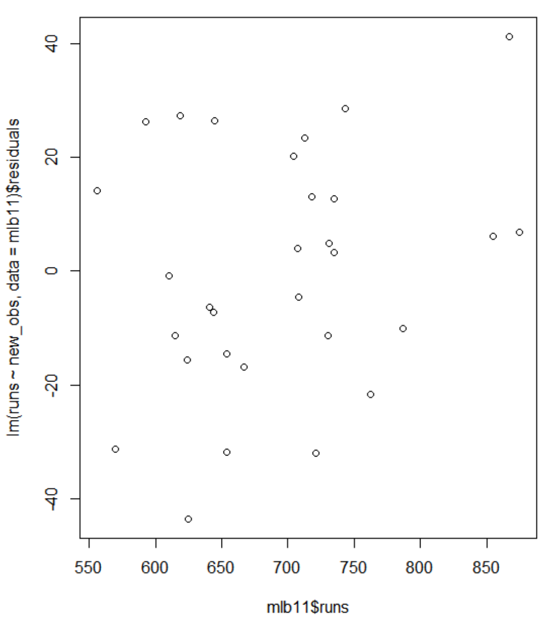
```
+ )

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
(Intercept)    -686.61      68.93  -9.962
mlb11$new_obs  1919.36      95.70  20.057
              Pr(>|t|)
(Intercept)   1.05e-10 ***
mlb11$new_obs  < 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
```

Multiple R-squared onbase	Multiple R-squared slug	Multiple R-squared obs
.8941	.8969	.9349

These models are much more accurate at predicting runs, where the new_obs variable has the closest linear relationship.

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

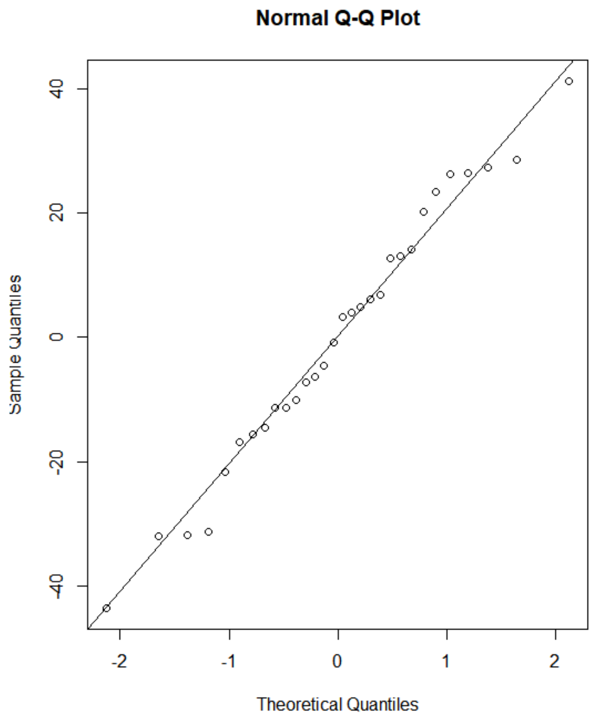
```
> plot(lm(runs ~ new_obs, data = mlb11)$residuals ~ mlb11$runs)
```

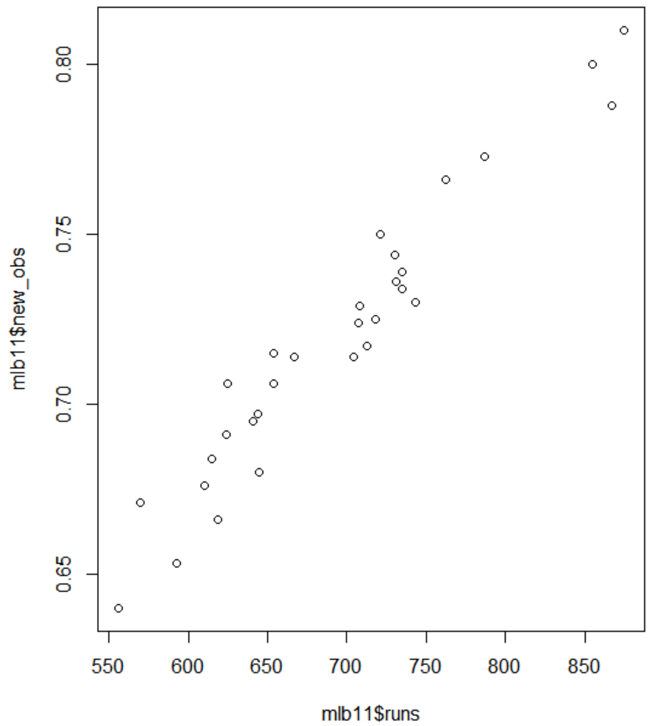


There does not appear the be any relationship between the residuals, so the relationship between runs and new obs are linearly related.

```
> plot(mlb11$runs, mlb11$new_obs)
> qqnorm(lm(runs ~ new_obs, data =
mlb11)$residuals)
> qqline(lm(runs ~ new_obs, data =
mlb11)$residuals)
```

The relationship between the residuals appears to be normally distributed, and the variablility appears to be constant.


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


mlb11$new_obs
mlb11$runs
0.80
0.75
0.70
0.65
550
600
650
700
750
800
850