Linear Regression - Stephen Jones Lab7

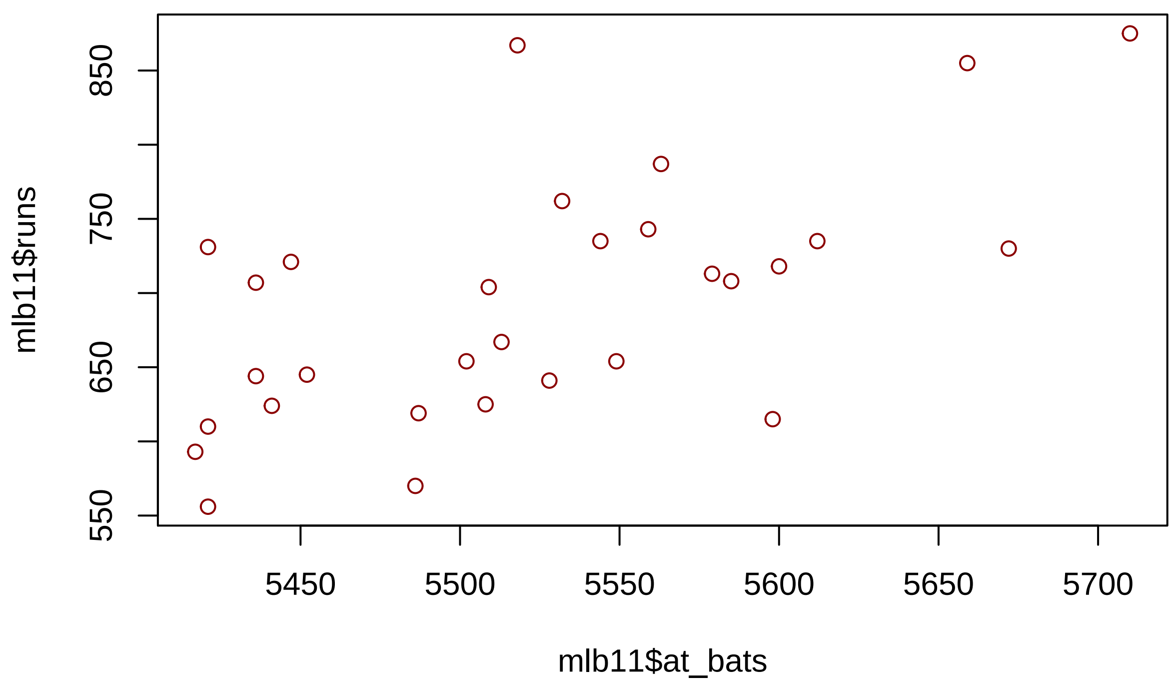
The data

```
rm(list=ls())
load("more/mlb11.RData")
```

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?

Use a scatterplot to display this relationship. The relationship appears to be approximately linear; a linear model could predict the number of runs, but with dubious accuracy.

```
plot(mlb11$at_bats,mlb11$runs,col=c("darkred"))
```




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

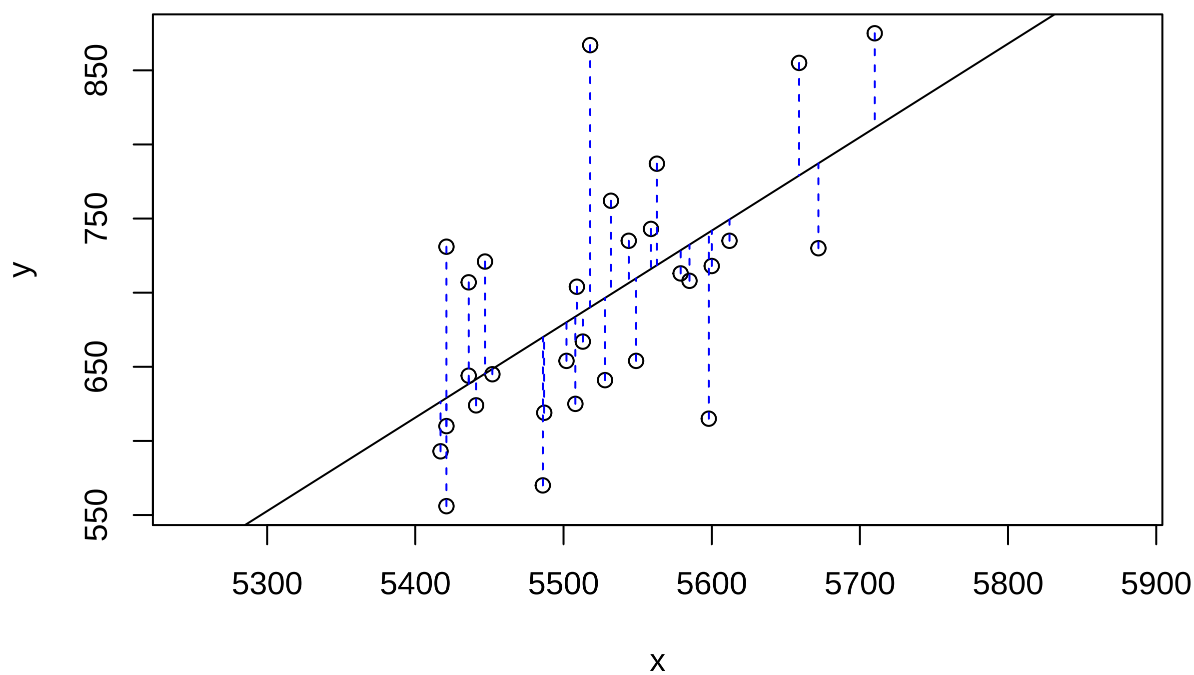
```
## [1] 0.610627
```

Sum of squared residuals

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.

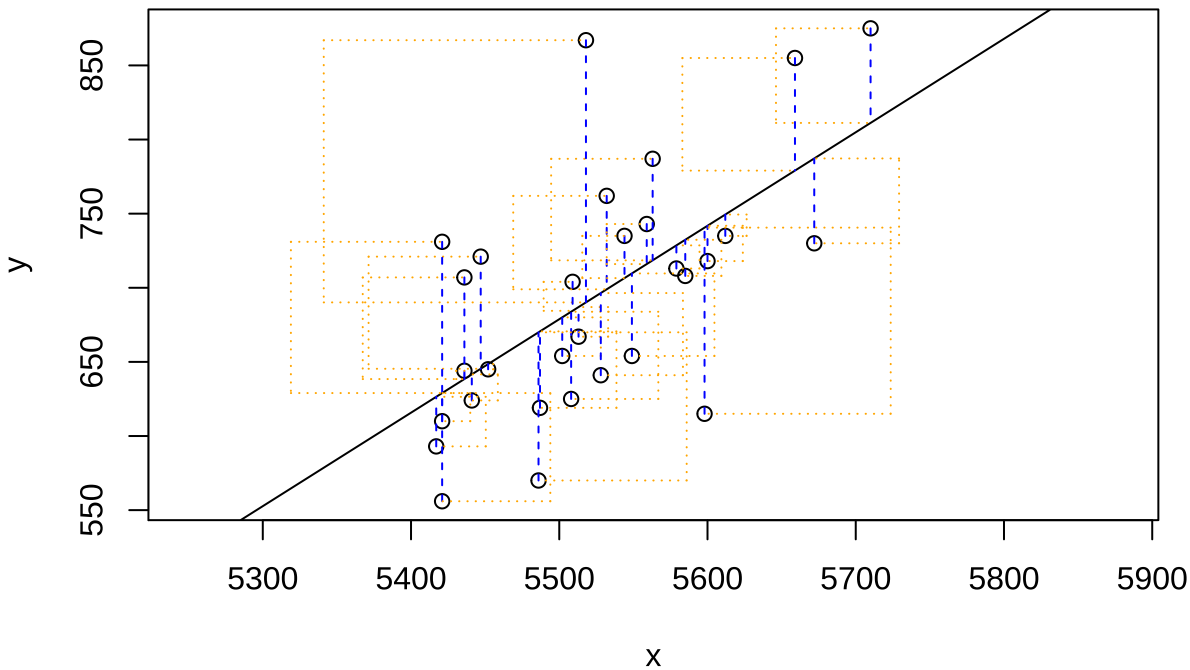
With correlation coefficient .61 we see an approximate linear relationship of moderate (positive) strength with a few outliers. Data are concentrated at fewer than 5600 at-bats, which is logical.

```
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
```

```
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
```

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?

I was not prompted to select points and I am not sure what to do here. The sum of squares is 123721.9 from the earlier plots.

The linear model

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?

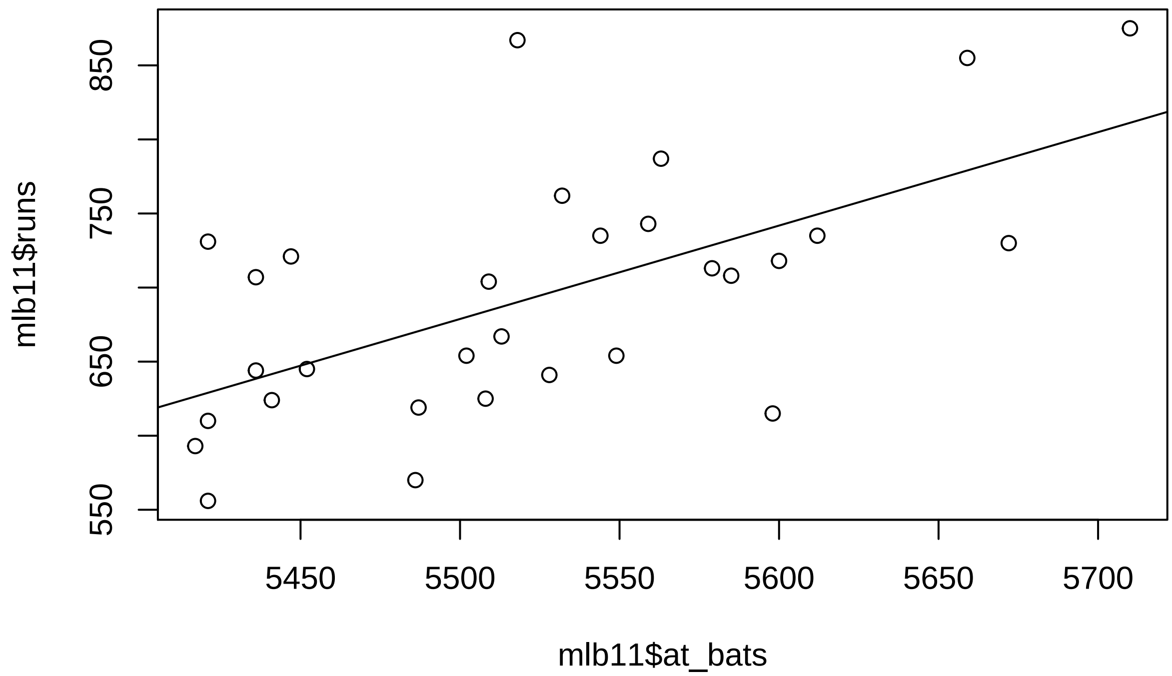
```r
m2 <- lm(runs ~ homeruns, data = mlb11)

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
```

For y = runs and x = homeruns, y = 415.2389 + 1.8345 * x is our equation.

Prediction and prediction errors

```r
plot(mlb11$runs ~ mlb11$at_bats)
m1 <- lm(runs ~ at_bats, data = mlb11)
abline(m1)
```




```r
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
```

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?

For y = runs and x = at_bats, y = -2789.2429 + 0.6305 * x is our equation.

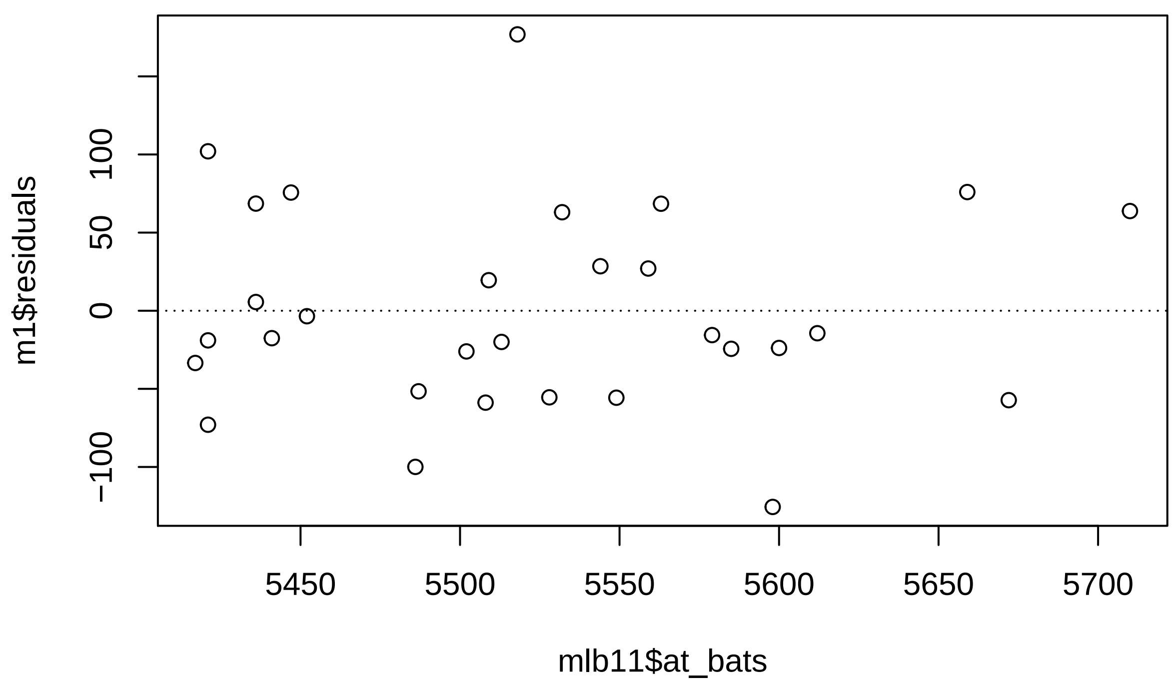
```r
runs <- -2789.2429 + (0.6305*5578)
runs
```

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

The prediction would be 728; this is an overestimate if we consider the runs by the most similar case, the Phillies, with 5579 at-bats and 713 runs. We cannot calculate a residual for 5578 runs without an observed value, but we can estimate based on the 5579 at-bats and the 713 runs of the Phillies: 728 - 713 = 15.

Model diagnostics

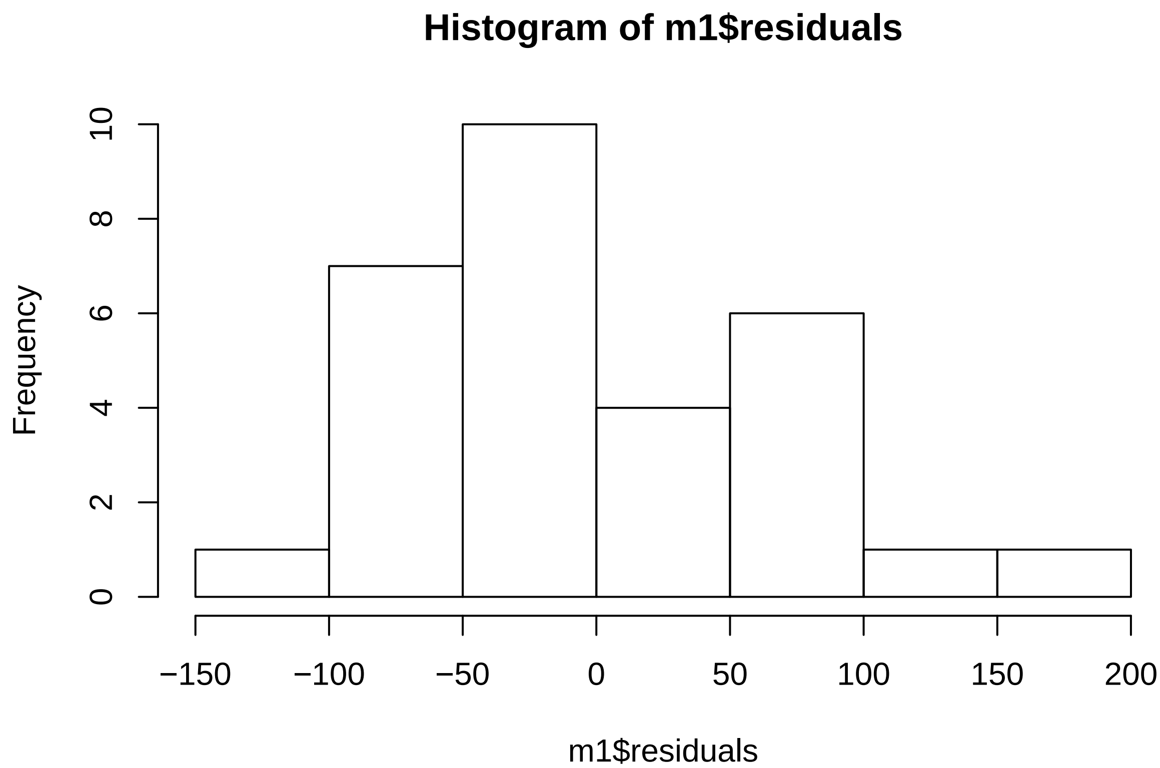
```r
plot(m1$residuals ~ mlb11$at_bats)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0
```



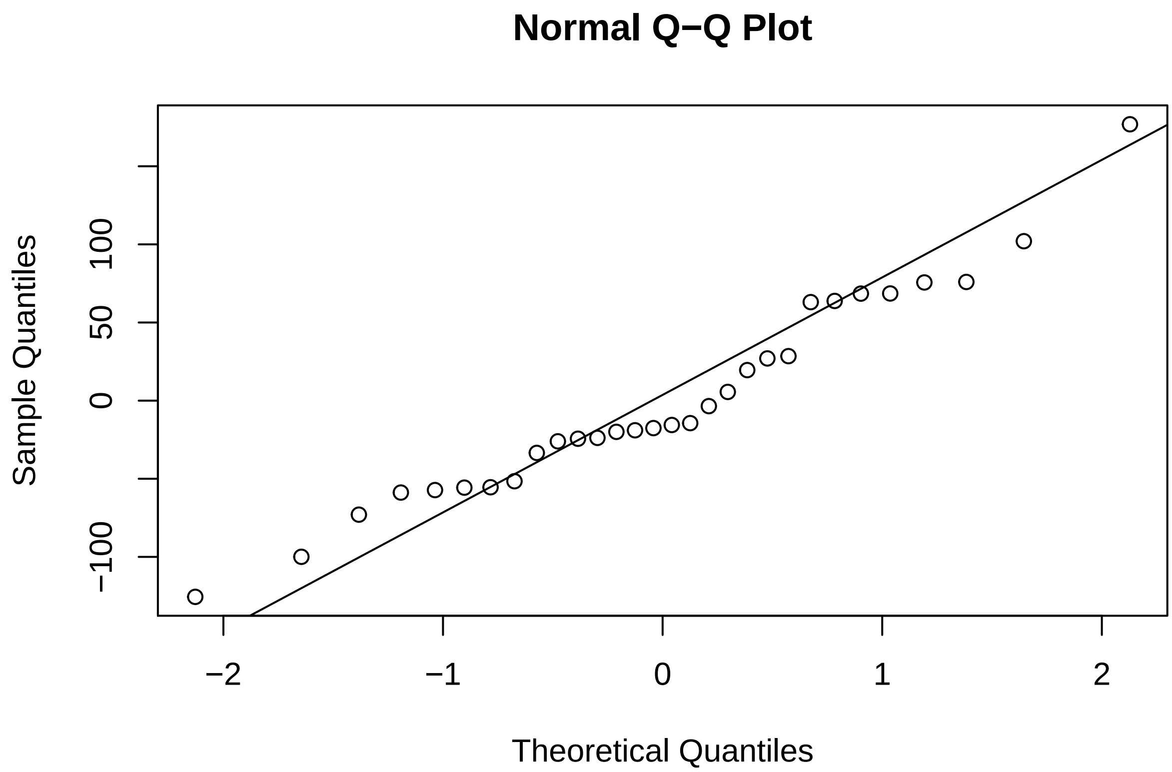
6. Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

There is no discernable pattern and the residuals appear to be evenly distributed about 0.

```r
hist(m1$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?

The histogram appears bimodal but roughly normal; otherwise the results appear approximately normal. The condition is met.

Constant variability:

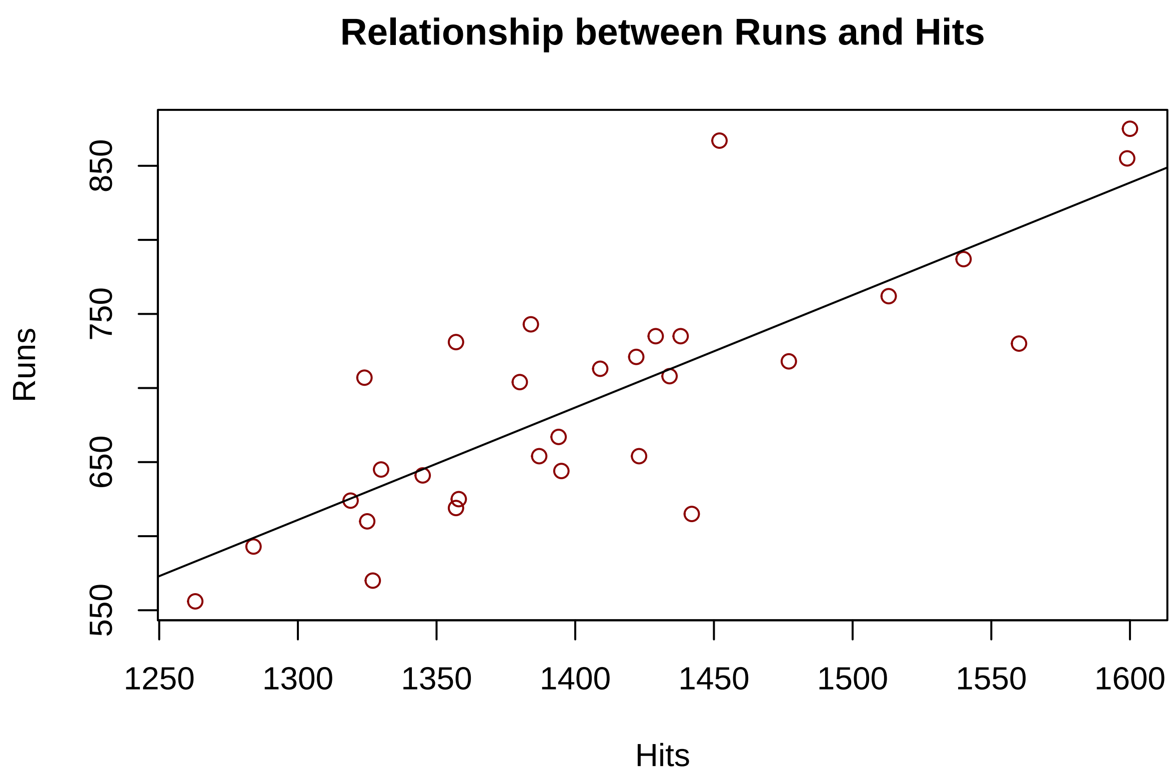
8. Based on the plot in (1), does the constant variability condition appear to be met?

This condition is met; variability around the plotline is fairly consistent among points.

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$runs ~ mlb11$hits, col=c("darkred"), main = "Relationship between Runs and Hits", xlab = "Hi
m3 <- lm(runs ~ hits, data = mlb11)
abline(m3)
```




```r
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
```

The relationship does indeed appear to be linear.

- 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?

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

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

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

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

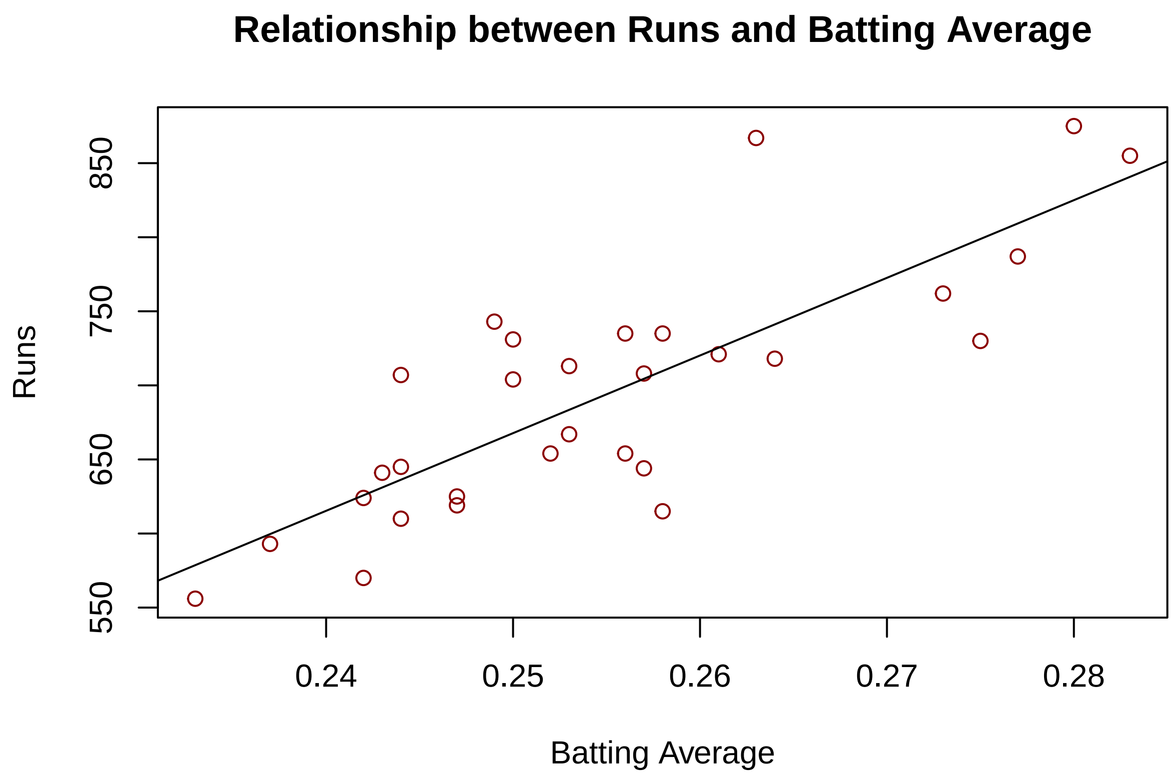
With R^2 value of .8012108, it appears that hits are a better predictor than at_bats; the higher the value (and closer to 1) the more accurate the predictor.

- 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).

```r
catbat<-cor(mlb11$runs, mlb11$at_bats)
chits<-cor(mlb11$runs, mlb11$hits)
chruns<-cor(mlb11$runs, mlb11$homeruns)
cbatavg<-cor(mlb11$runs, mlb11$bat_avg)
cskout<-cor(mlb11$runs, mlb11$strikeouts)
cstbase<-cor(mlb11$runs, mlb11$stolen_bases)
cwins<-cor(mlb11$runs, mlb11$wins)

lmcatbat<-summary(lm(mlb11$runs~mlb11$at_bats))$r.squared
lmchits<-summary(lm(mlb11$runs~mlb11$hits))$r.squared
lmchruns<-summary(lm(mlb11$runs~mlb11$homeruns))$r.squared
lmcbatavg<-summary(lm(mlb11$runs~mlb11$bat_avg))$r.squared
lmcskout<-summary(lm(mlb11$runs~mlb11$strikeouts))$r.squared
lmcstbase<-summary(lm(mlb11$runs~mlb11$stolen_bases))$r.squared
lmcwins<-summary(lm(mlb11$runs~mlb11$wins))$r.squared

rsq<-c(lmcatbat,lmchits,lmchruns,lmcbatavg,lmcskout,lmcstbase,lmcwins)
cor<-c(catbat,chits,chruns,cbatavg,cskout,cstbase,cwins)
name<-c("at_bats","hits","homeruns","bat_avg","strikeouts","stolen_bases","wins")
c<-cbind(name,cor,rsq)

#return the highest correlation coefficient.
answer<-c[order(-cor),]
answer
```

```
##      name           cor                   rsq
## [1,] "bat_avg"      "0.809985885461508"   "0.656077134646863"
## [2,] "hits"         "0.801210813231711"   "0.641938767239419"
## [3,] "homeruns"     "0.791557685558218"   "0.626563569566283"
## [4,] "at_bats"      "0.610627046720669"   "0.372865390186805"
## [5,] "wins"         "0.600808771113306"   "0.360971179446681"
## [6,] "stolen_bases" "0.0539814103796295" "0.00291399266657394"
## [7,] "strikeouts"   "-0.411531204450297" "0.169357932236313"
```

```r
plot(mlb11$runs ~ mlb11$bat_avg, col=c("darkred"), main = "Relationship between Runs and Batting Average
m4 <- lm(runs ~ bat_avg, data = mlb11)
abline(m4)
```




```r
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
```

Batting average appears to be the best predictor of runs. The batting average correlation coefficient is highest among the traditional variables.

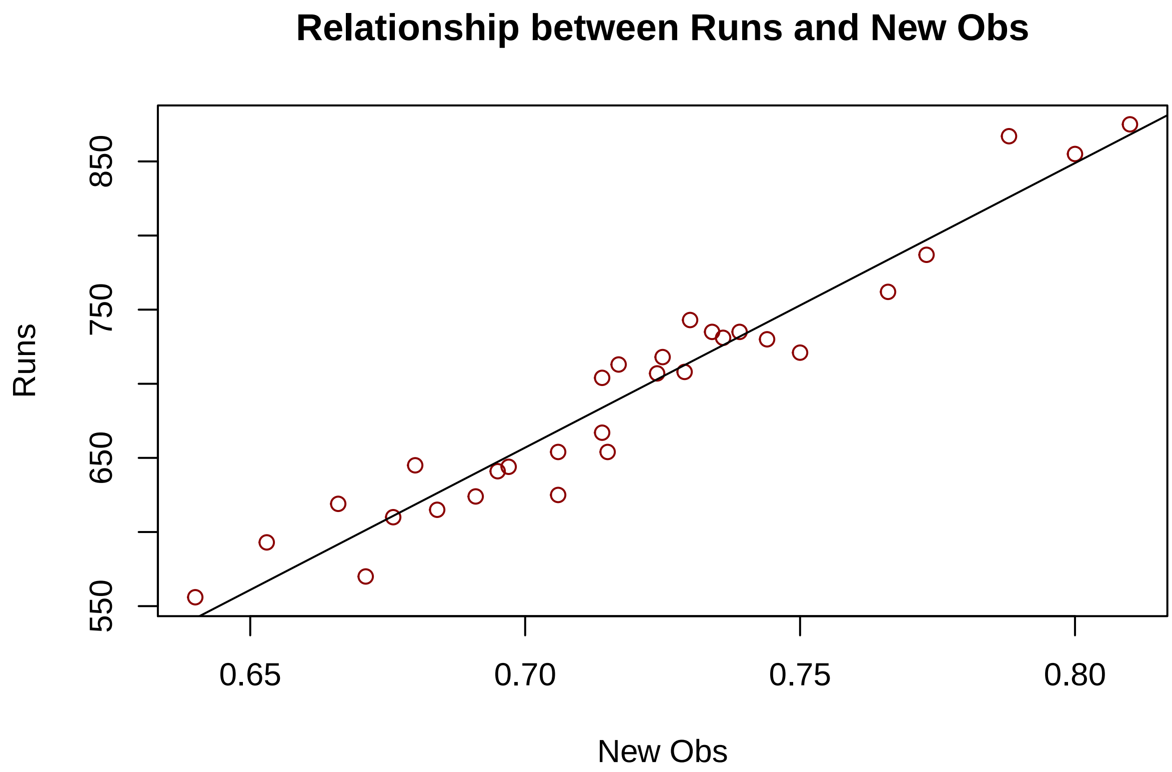
- Now examine the three newer variables. These are the statistics used by the author of *Moneyball* to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we've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

```
cnob<-cor(mlb11$runs, mlb11$new_onbase)
cnslu<-cor(mlb11$runs, mlb11$new_slug)
cnobs<-cor(mlb11$runs, mlb11$new_obs)

lmnonb<-summary(lm(mlb11$runs~mlb11$new_onbase))$r.squared
lmnslu<-summary(lm(mlb11$runs~mlb11$new_slug))$r.squared
lmnobs<-summary(lm(mlb11$runs~mlb11$new_obs))$r.squared

rsq2<-c(lmcatbat,lmchits,lmchruns,lmcbatavg,lmcskout,lmcstbase,lmcwins,lmnonb,lmnslu,lmnobs)
cor2<-c(catbat,chits,chruns,cbatavg,cskout,cstbase,cwins,cnob,cnslu,cnobs)
name2<-c("at_bats","hits","homeruns","bat_avg","strikeouts","stolen_bases","wins","new_onbase","new_slu
c2<-cbind(name2,cor2,rsq2)

#return the highest correlation coefficient.
answer2<-c2[order(-cor2),]
answer2
```

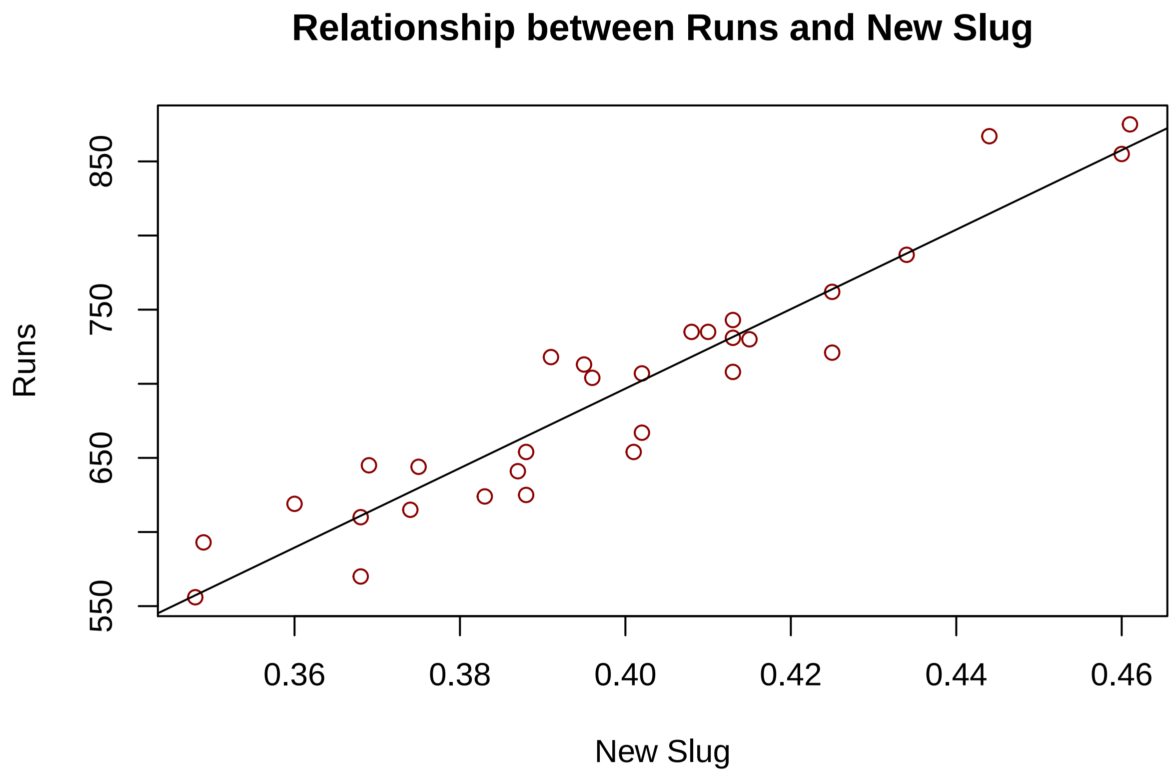
```
##       name2          cor2                 rsq2
##  [1,] "new_obs"      "0.966916297490023"  "0.934927126351814"
##  [2,] "new_slug"     "0.947032400929154"  "0.896870368409638"
##  [3,] "new_onbase"   "0.921469072430616"  "0.849105251446139"
##  [4,] "bat_avg"      "0.809985885461508"  "0.656077134646863"
##  [5,] "hits"         "0.801210813231711"  "0.641938767239419"
##  [6,] "homeruns"     "0.791557685558218"  "0.626563569566283"
##  [7,] "at_bats"      "0.610627046720669"  "0.372865390186805"
##  [8,] "wins"         "0.600808771113306"  "0.360971179446681"
##  [9,] "stolen_bases" "0.0539814103796295" "0.00291399266657394"
## [10,] "strikeouts"   "-0.411531204450297" "0.169357932236313"
```

Looking solely at correlation coefficients (and confirming with R-squared for each variable pair), the new statistics are better predictors overall.

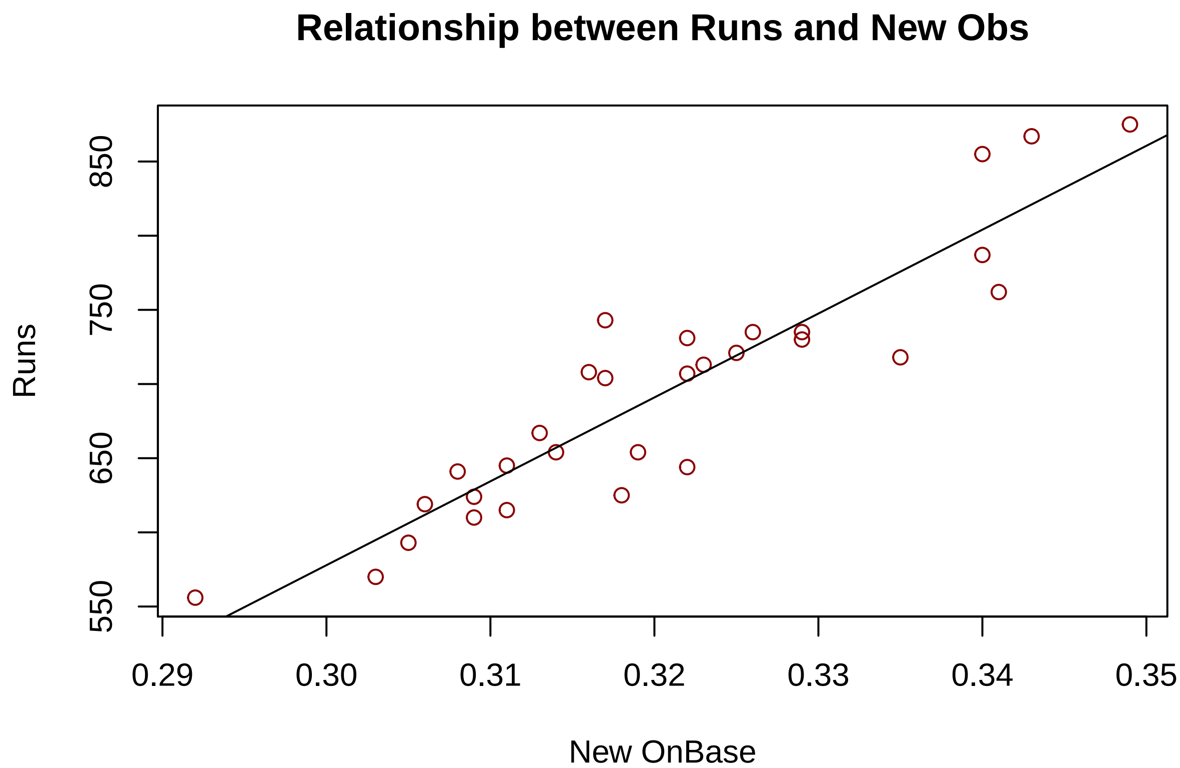
```
plot(mlb11$runs ~ mlb11$new_obs, col=c("darkred"), main = "Relationship between Runs and New Obs", xlab
m5 <- lm(runs ~ new_obs, data = mlb11)
abline(m5)
```




```r
plot(mlb11$runs ~ mlb11$new_slug, col=c("darkred"), main = "Relationship between Runs and New Slug", xla
m6 <- lm(runs ~ new_slug, data = mlb11)
abline(m6)
```



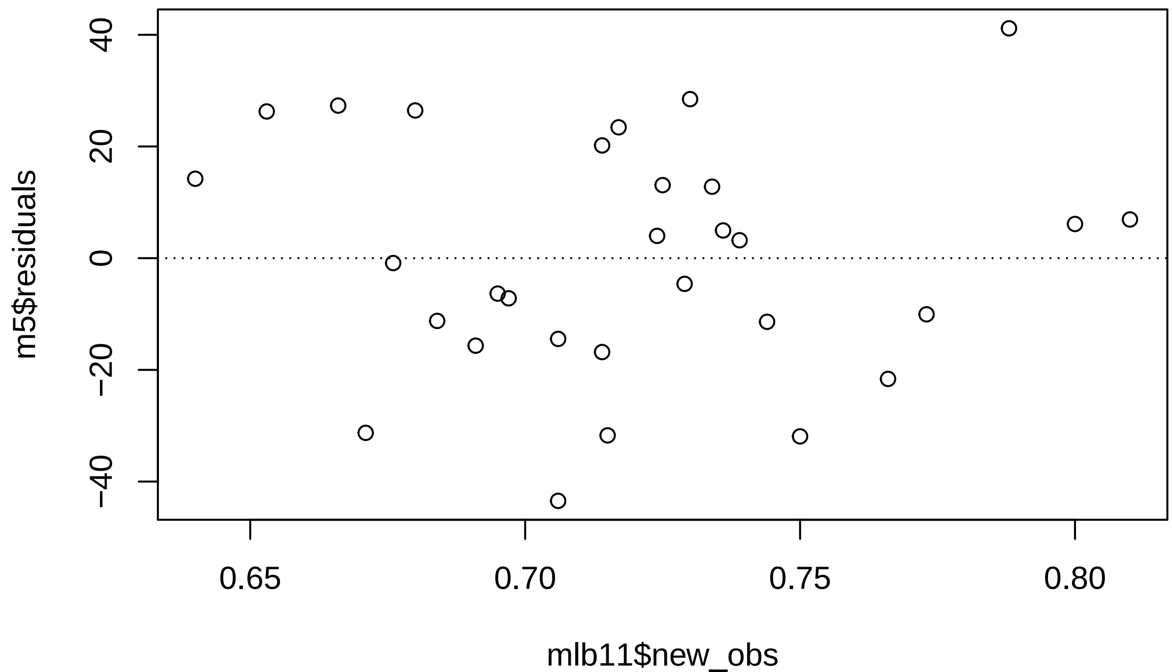

```
plot(mlb11$runs ~ mlb11$new_onbase, col=c("darkred"), main = "Relationship between Runs and New Obs", xl
m7 <- lm(runs ~ new_onbase, data = mlb11)
abline(m7)
```



Plots of each new variable vary from the regression line less than the traditional variables. I am not sure what variable "new_obs" is, but it is the best predictor of runs.

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

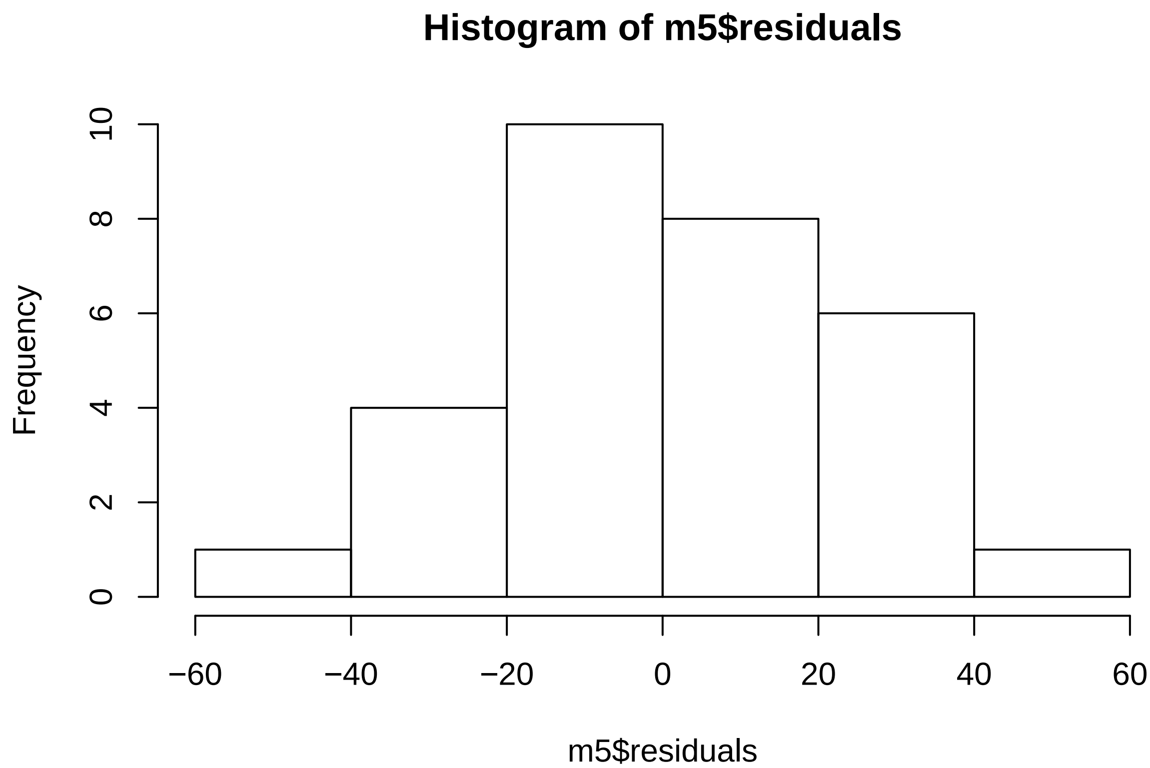
```
plot(m5$residuals ~ mlb11$new_obs)+
abline(h = 0, lty = 3)
```




```
## integer(0)
```

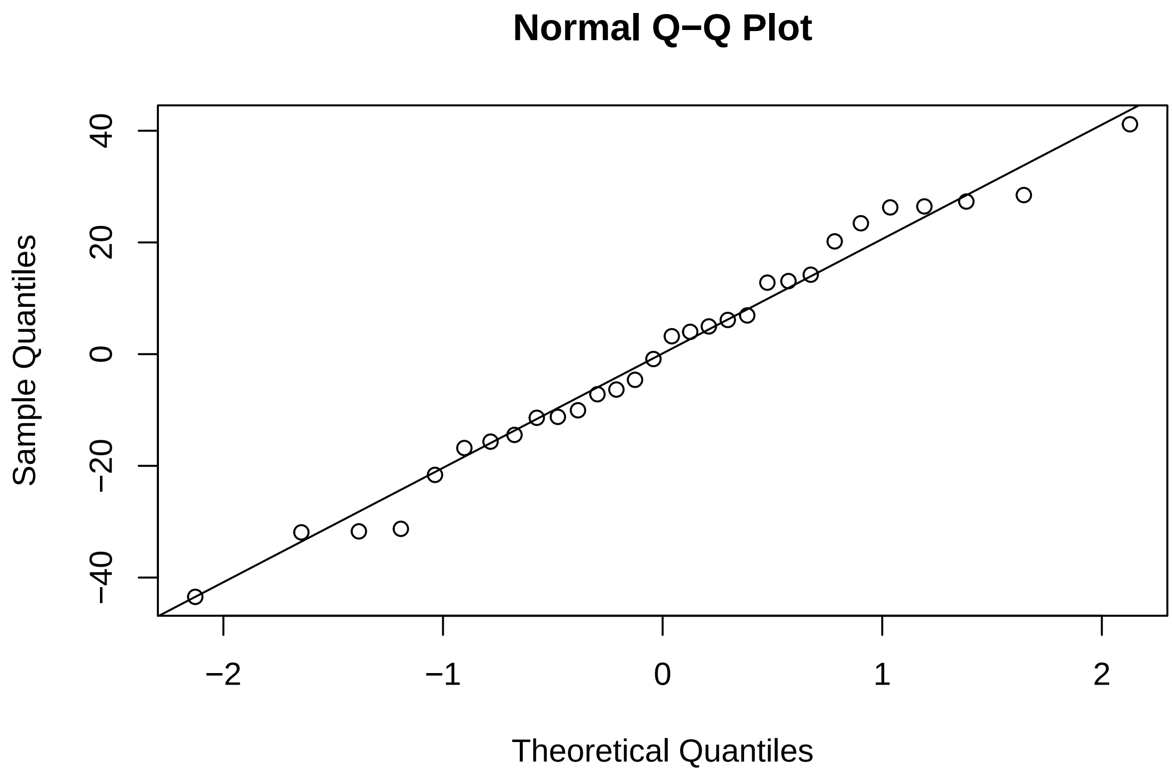
The relationship appears linear; variability is constant and distribution appears to be normal.

```
hist(m5$residuals)
```



Histogram indicates a nearly normal distribution, which is confirmed by the QQ pot below. Variability in distance to the line below is approximately constant; conditions are met.

```
qqnorm(m5$residuals)
qqline(m5$residuals)
```


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