Hao-7

```
setwd("C:/Users/bhao/Google Drive/CUNY/git/DATA606/Lab7")
library(IS606)
```

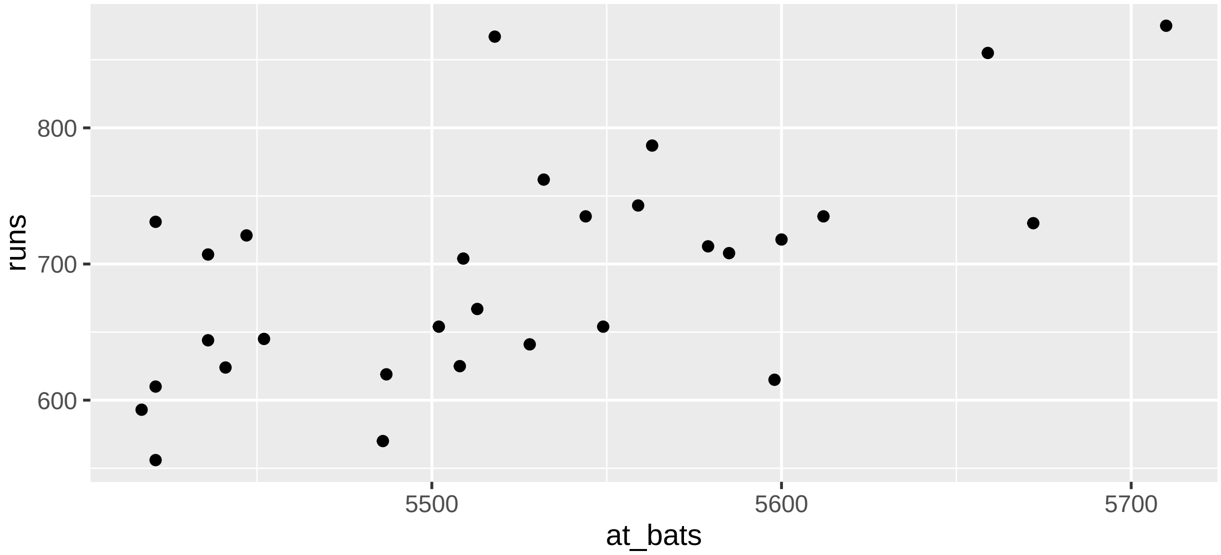
```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
library(dplyr)
library(ggplot2)
load('more/mlb11.RData')
```

Exercise1: 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?

The relationship looks approximately linear. Yes, I would be comfortable using a linear model.

```
mlb11 %>% ggplot(aes(x = at_bats, y = runs)) + geom_point()
```



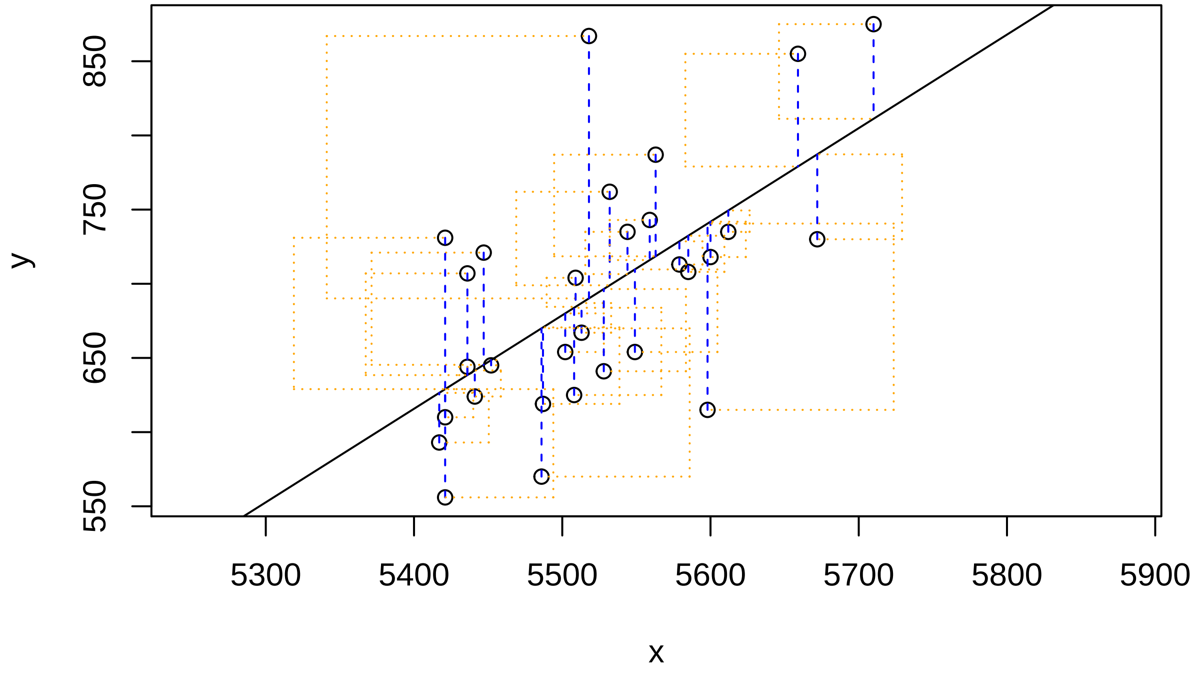
```
cor(mlb11$runs, mlb11$at_bats)
```

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

Exercise2: 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 relationship is positive, and the strength of the relationship is moderate. A few points may be considered outliers, but none appear to be influential points.

```
#plot_ss(x = mlb11$at_bats, y = mlb11$runs)
plot_ss(x = mlb11$at_bats, y = mlb11$runs, showSquares = TRUE)
```

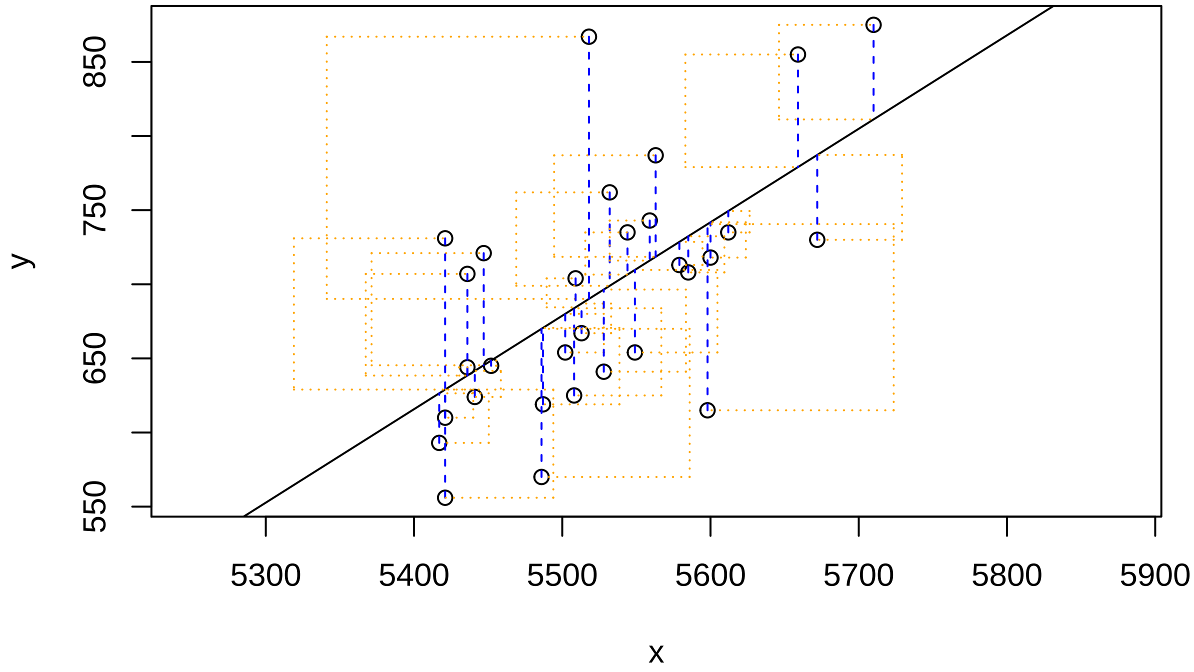


```
## Click two points to make a line.

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

Exercise3: 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, 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
```

Exercise4: 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?

y = 415.2 + 1.835x; the slope indicates that for each homerun, the team should score on average 1.835 additional runs per season

```
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
```

Exercise5: 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?

728 runs; the closest team to 5578 at_bats was the Phillies at 5579 at_bats, who had 713 runs. This is 15 runs lower than our estimate, so the model overestimated by 15 runs.

```
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
```

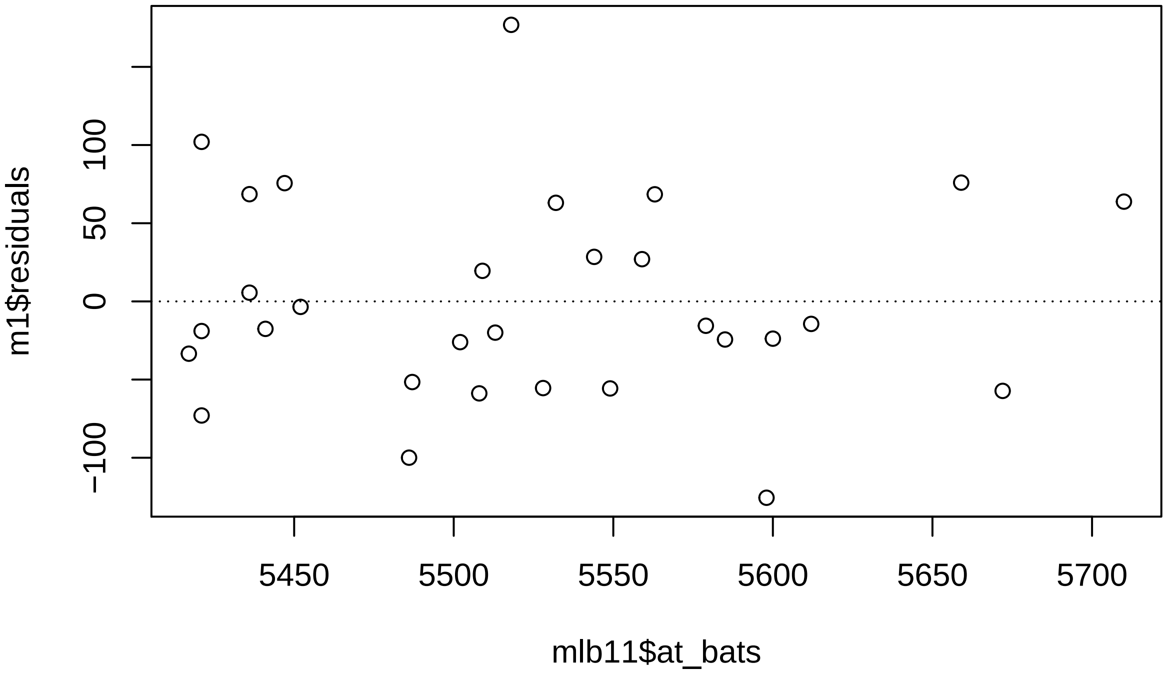
```
predict(m1, data.frame(at_bats = 5578))
```

```
##       1
## 727.965
```

Exercise6: Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

No, there is no apparently pattern in the residuals.

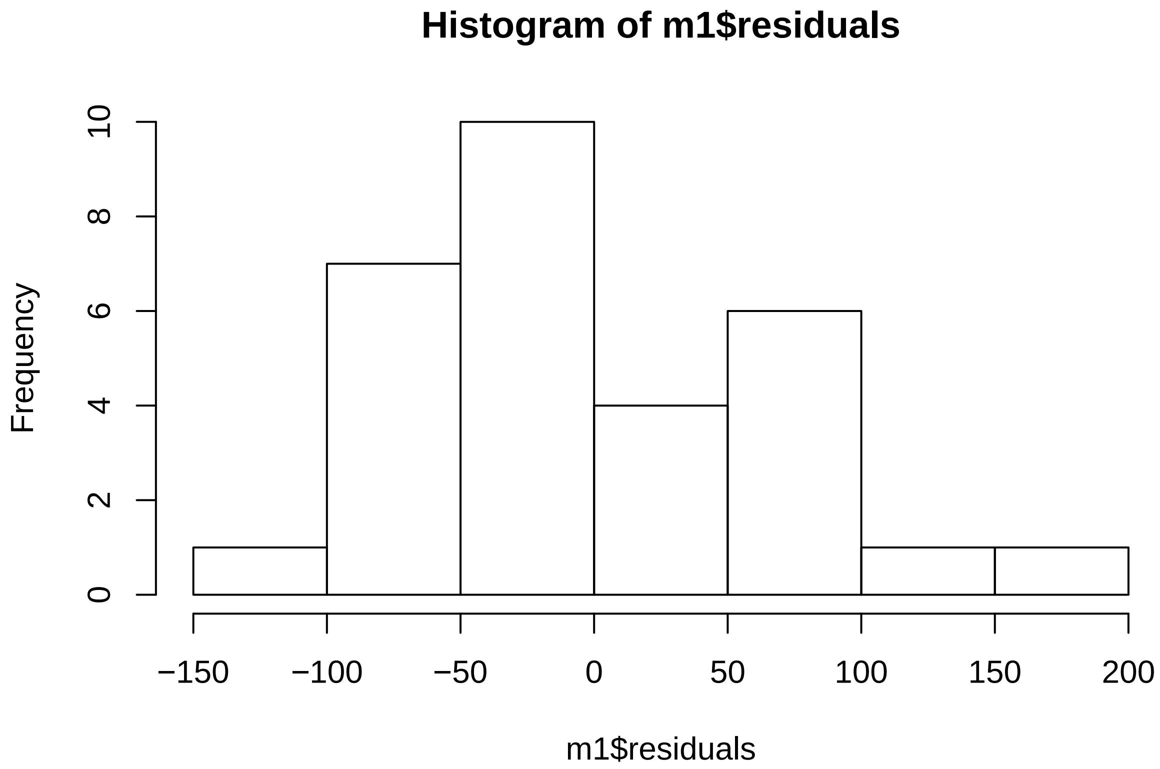
```
plot(m1$residuals ~ mlb11$at_bats)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0
```



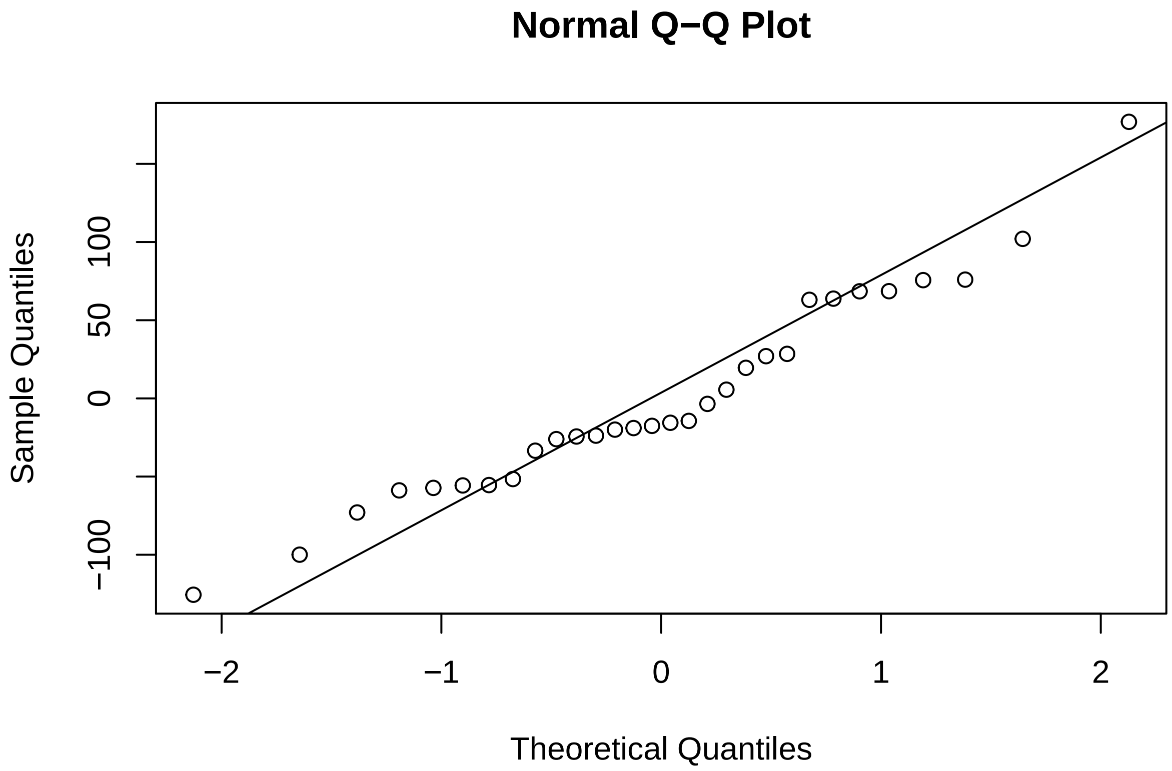
Exercise7: Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

Yes

```
hist(m1$residuals)
```



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



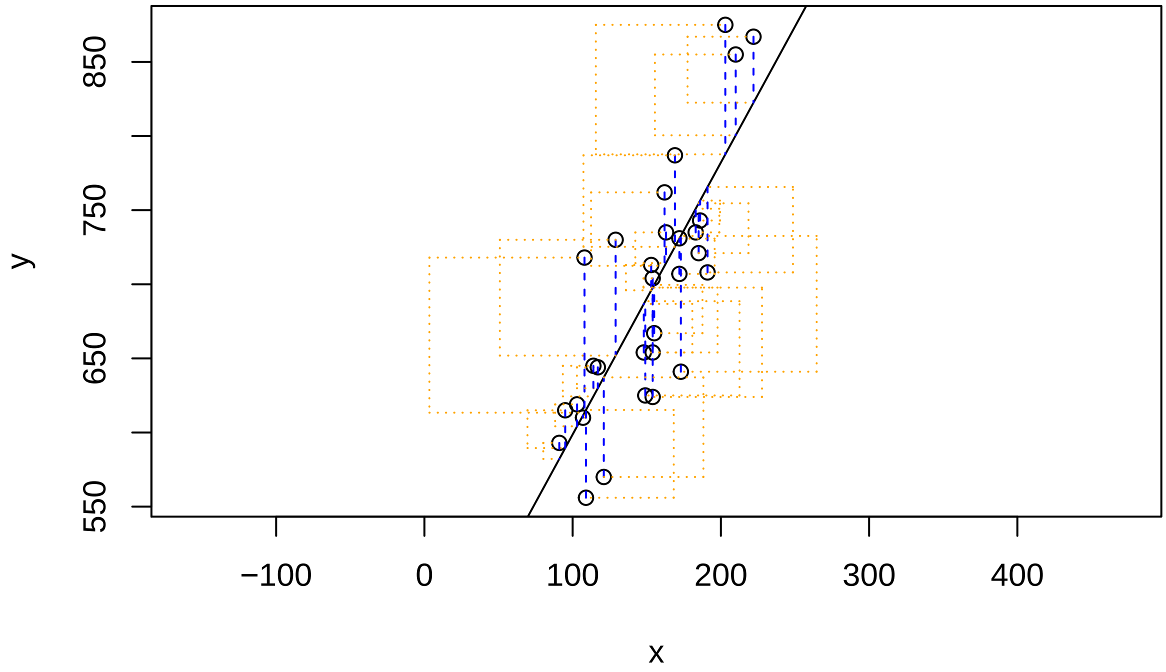
Exercise8: Based on the plot in (1), does the constant variability condition appear to be met?

Yes

On Your Own1: 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?

Using homeruns, there does appear to be a linear relationship.

```
plot_ss(x = mlb11$homeruns, y = mlb11$runs, showSquares = TRUE)
```



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

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##     415.239        1.835
##
## Sum of Squares:  73671.99
```

Own Your Own2: 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 r-squre values for the at_bats- and homeruns-based linear models were 0.37 and 0.63, respectively. Therefore, it would seem that homeruns are a better predictor of runs than at_bats are.

```
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
```

```
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
```

On Your Own3: 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).

Based on the r-square values calculated below, batting average has the highest r-square and thus best predicts runs.

```
mlb11 %>% summarise(at_bats = cor(runs, at_bats)^2,
                    hits = cor(runs, hits)^2,
                    homeruns = cor(runs, homeruns)^2,
                    bat_avg = cor(runs, bat_avg)^2,
                    strikeouts = cor(runs, strikeouts)^2,
                    stolen_bases = cor(runs, stolen_bases)^2,
                    wins = cor(runs, wins)^2)
```

```
##     at_bats      hits  homeruns   bat_avg strikeouts stolen_bases
## 1 0.3728654 0.6419388 0.6265636 0.6560771  0.1693579  0.002913993
##        wins
## 1 0.3609712
```

```
m3 = lm(runs ~ bat_avg, data = mlb11)
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
```

On Your Own4: 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?

In general, the new metrics with r-square values from 0.85 to 0.93 are much more effective at predicting runs. Of all 10 variables analyzed, the combined on-base and slugging (new_obs) metric exhibits the highest r-square value at 0.93. Based on limited baseball knowledge, this makes sense as it combines the frequency players get on base and the effectiveness of hitting.

```
mlb11 %>% summarise(new_onbase = cor(runs, new_onbase)^2,
                    new_slug = cor(runs, new_slug)^2,
                    new_obs = cor(runs, new_obs)^2)
```

```
##   new_onbase  new_slug   new_obs
## 1  0.8491053 0.8968704 0.9349271
```

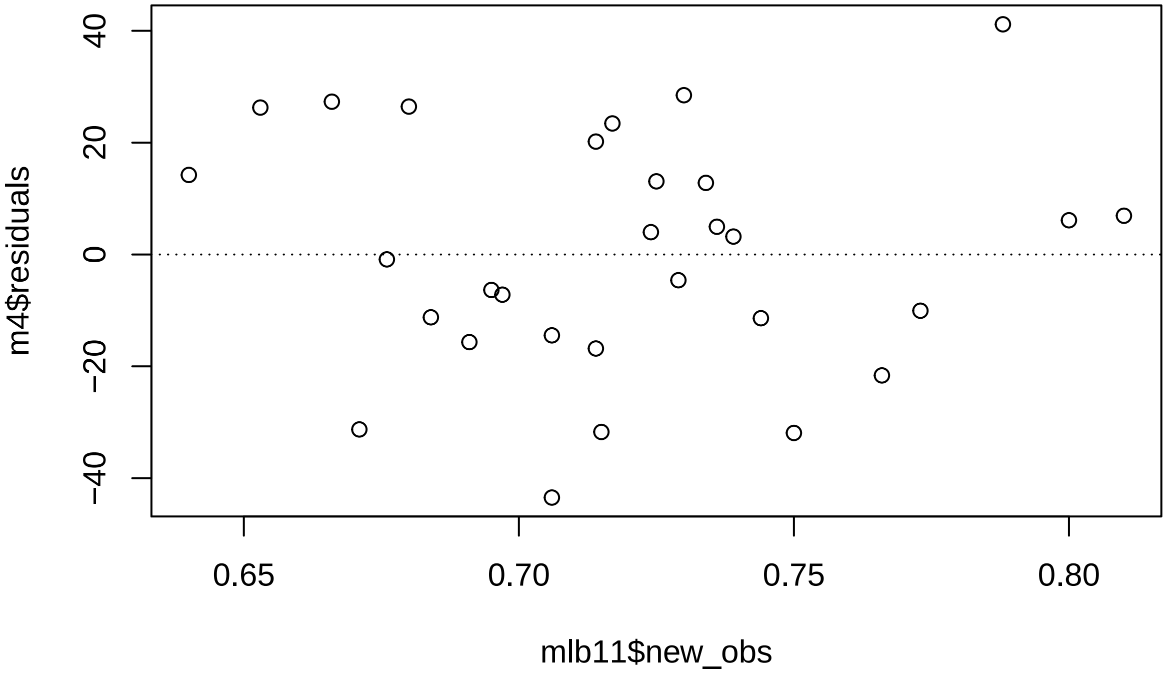
On Your Own5: Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

The model diagnostics look good. The residuals plot appears random, and the residuals appear to be normally distributed.

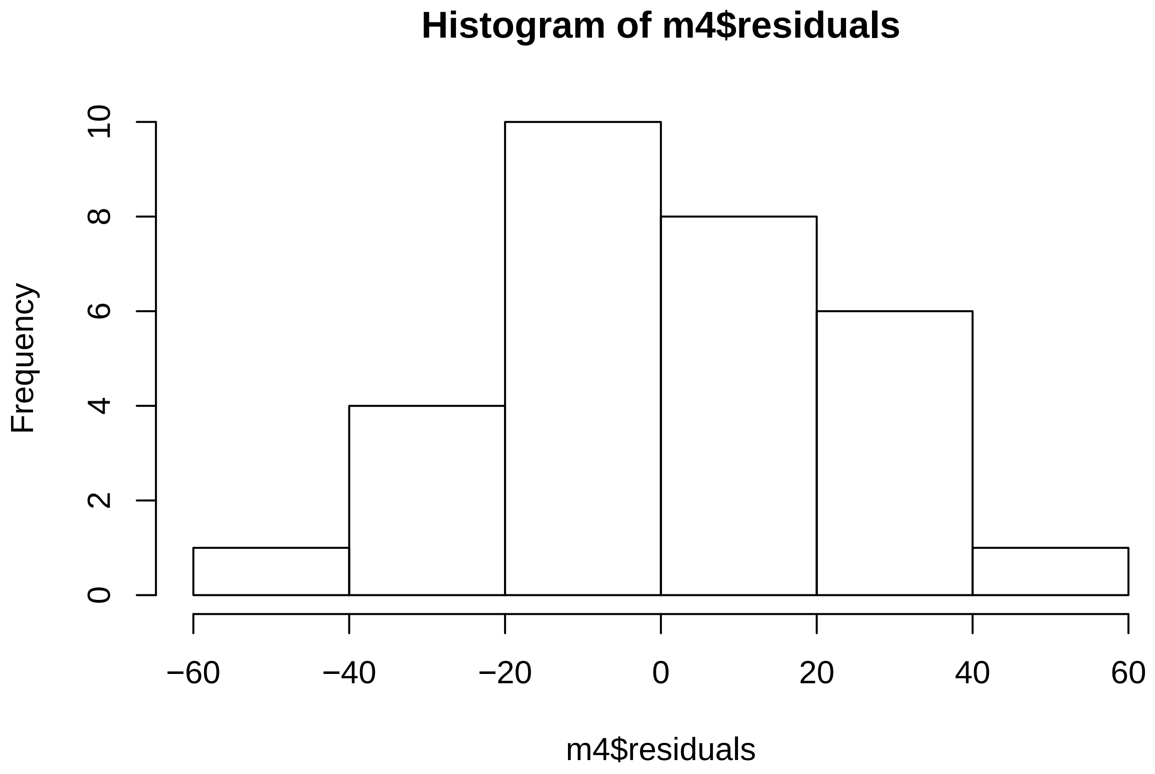
```
m4 = lm(runs ~ new_obs, data = mlb11)
summary(m4)
```

```
##
## 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
```

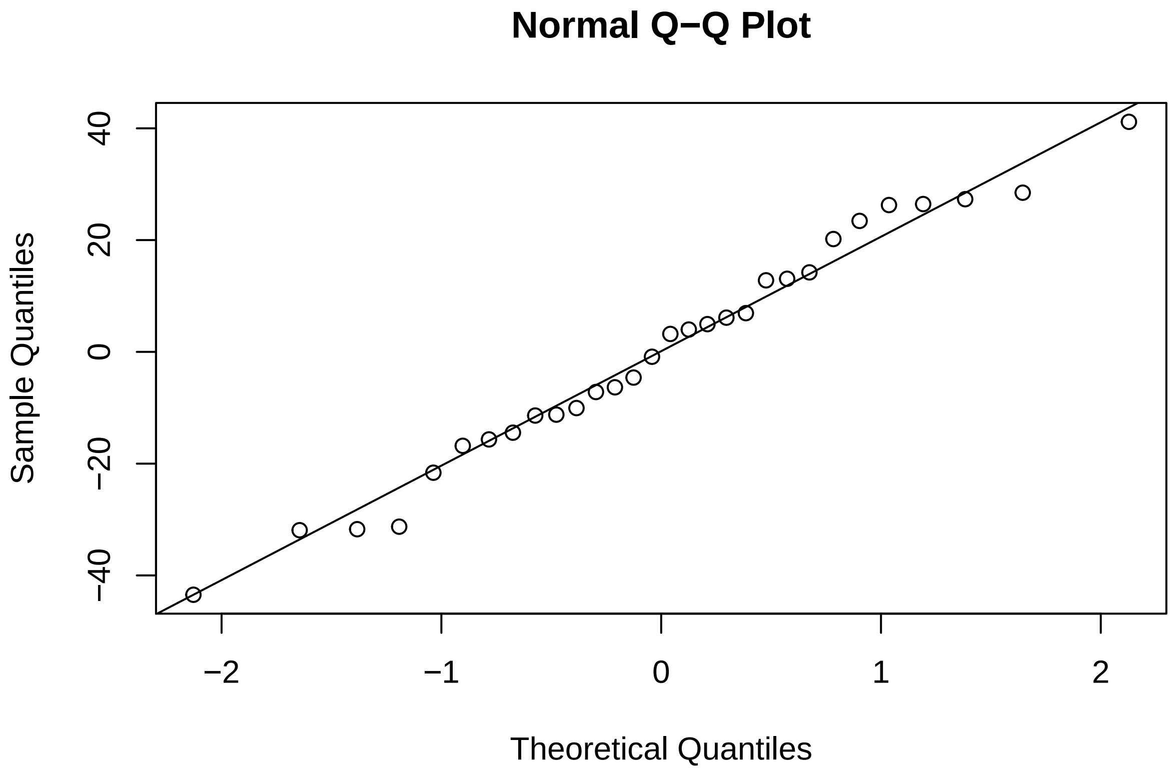
```
plot(m4$residuals ~ mlb11$new_obs)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0
```



```
hist(m4$residuals)
```



```
qqnorm(m4$residuals)
qqline(m4$residuals)  # adds diagonal line to the normal prob plot
```


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