STAT 100B Lab 6

Wesley Chang

Sumemr 2020 Session B

Setup for Lab

```r
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")

attach(mlb11)
summary(mlb11)
```

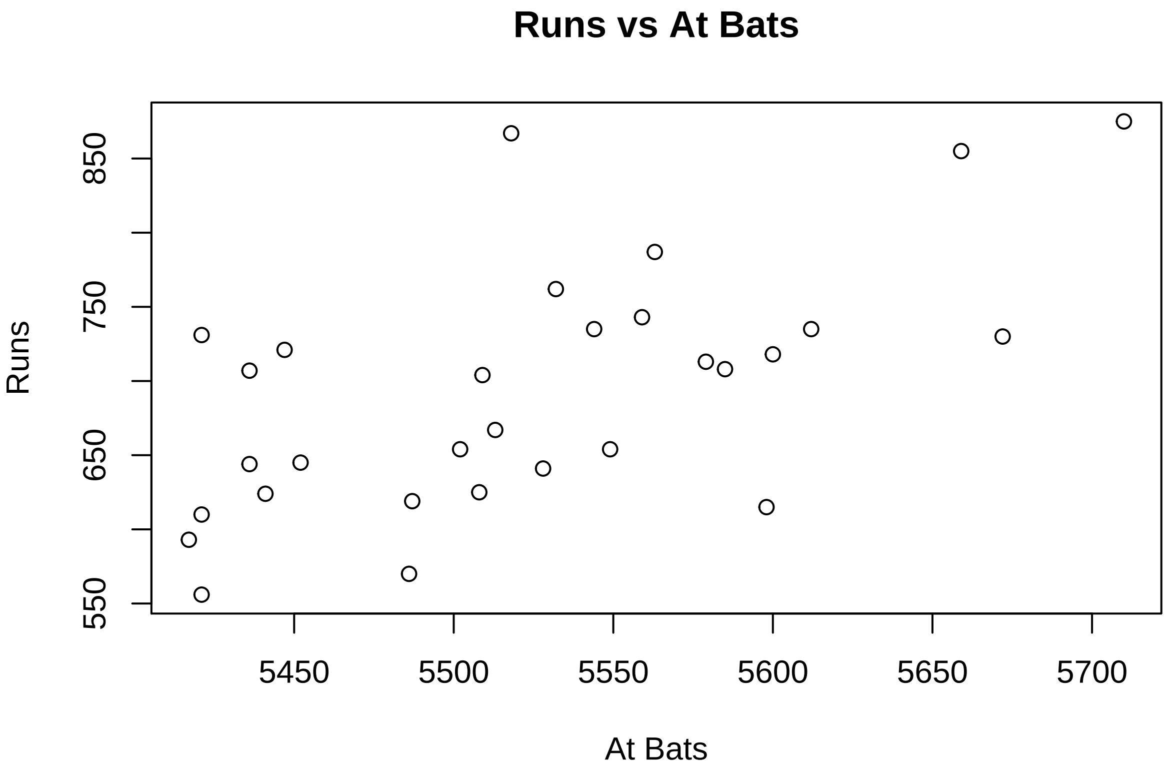
```
##                     team         runs          at_bats           hits     
##  Arizona Diamondbacks: 1   Min.   :556.0   Min.   :5417   Min.   :1263  
##  Atlanta Braves      : 1   1st Qu.:629.0   1st Qu.:5448   1st Qu.:1348  
##  Baltimore Orioles   : 1   Median :705.5   Median :5516   Median :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  
## 
```

Lab Exercises

Exercise 1

What type of plot would you use to display the relationship between `runs` *and one of the other numerical variables? Scatter plot would be a good choice. Plot this relationship using the variable* `at_bats` *as the* *predictor. Write down the R code for producing the scattter plot. The relationship should look somewhat linear. If you knew a team's* at_bats*, would you be comfortable using a linear model to predict the number of runs?*

```
plot(at_bats, runs,
      main="Runs vs At Bats",
      xlab="At Bats",
      ylab="Runs"
     )
```



Answer

If I knew a team's at_bats, I would be not be comfortable using a linear model to predict the number of runs, as the scatterplot indicates a general positive trend, but the relationship seems weak. A linear model may provide useful information about the general effect and trend that increasing at bats may have on runs, but I would not use it to predict specific values.

Exercise 2

Looking at your plot from the previous exercise, describe the relationship between these two variables. Do you see an upward trend or downward trend? Is it a strong linear relationship or a weak relationship?

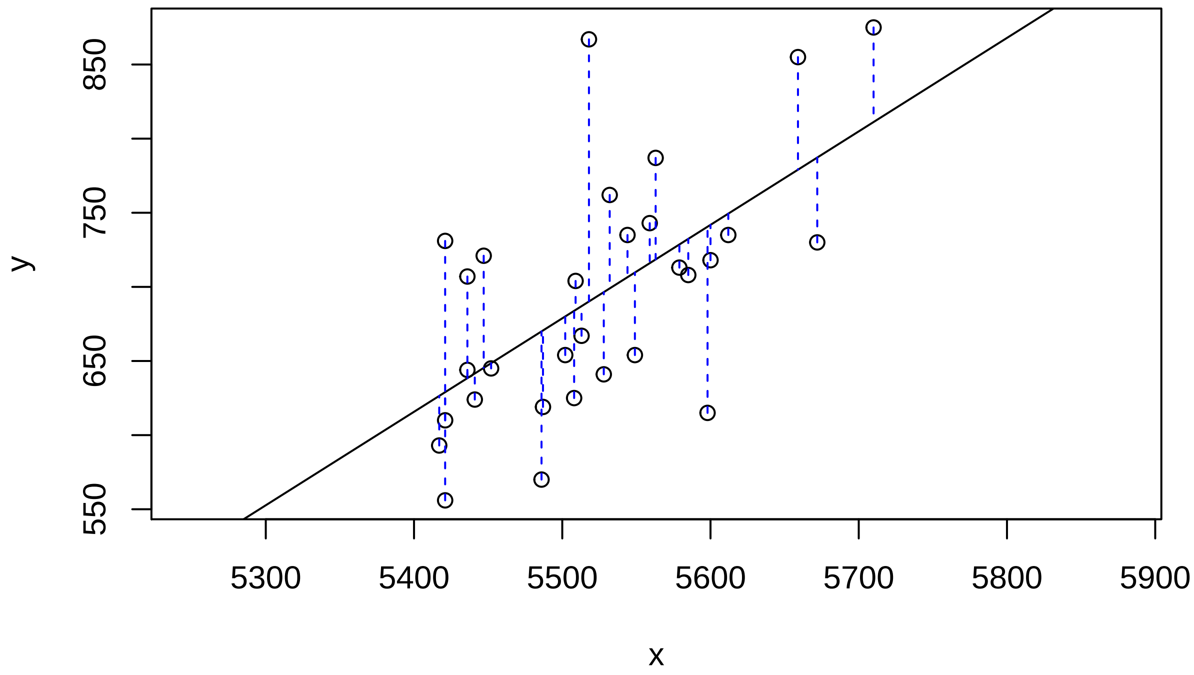
Answer

Looking at the plot, the relationship seems generally positive and upward, with a weaker linear relationship between the two variables.

Exercise 3

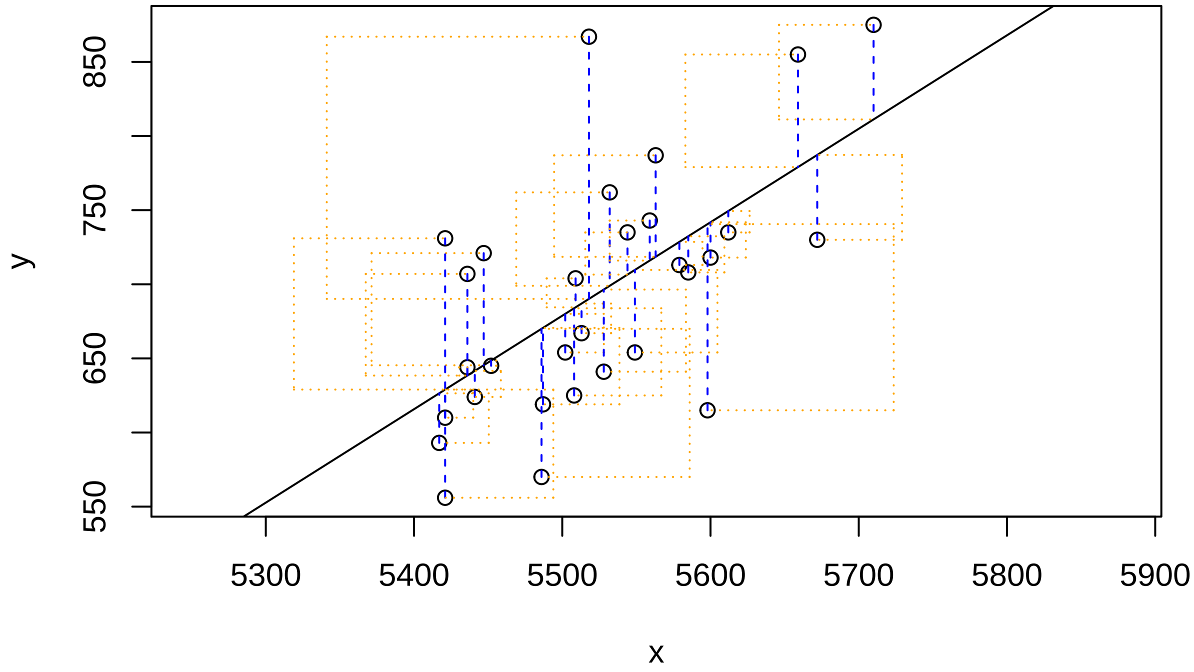
Using `plot_ss`*, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors?*

```
# scatterplot with user-inputted line
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
```

```
# scatterplot with line based on minimized SSR
```

Answer

The smallest sum of squares I got was 123729.9.

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?*

```
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
```

Answer

The equation of the regression line is $\hat{y} = 415.2389 + 1.8345x$. The slope indicates that the relationship between the success of the team and its home runs is positive.

Exercise 5

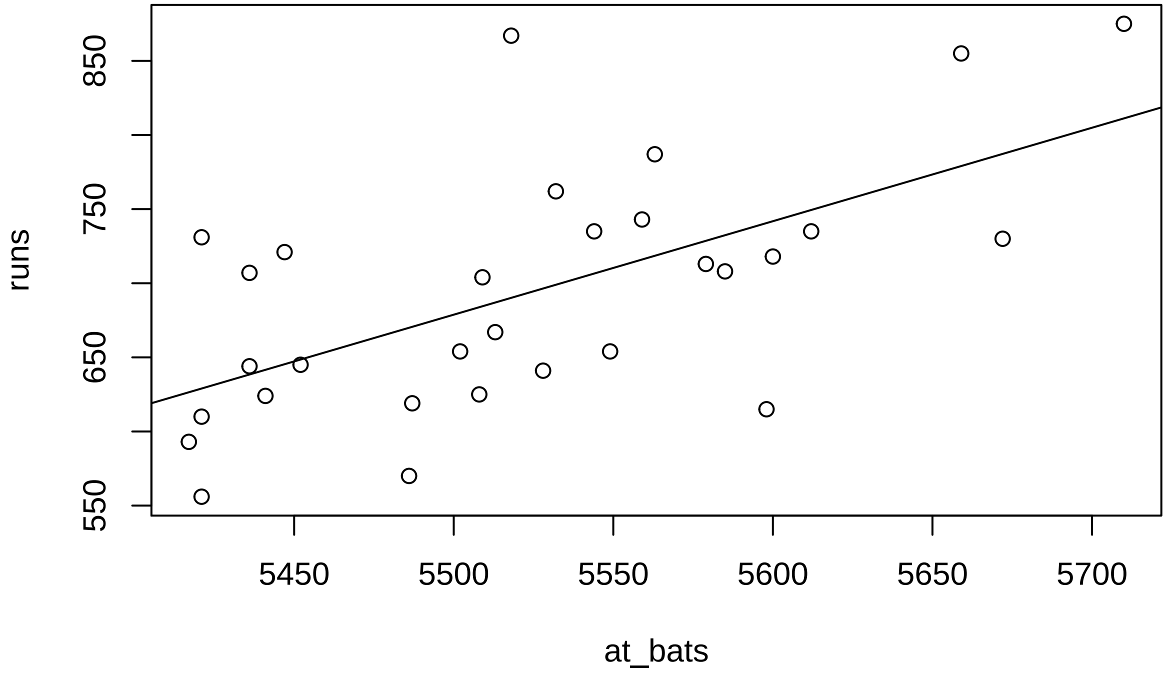
If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,578 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

Answer

```r
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
```

```
plot(runs ~ at_bats)
abline(m1)
```



```
-2789.2429+0.6305*(5578)
```

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

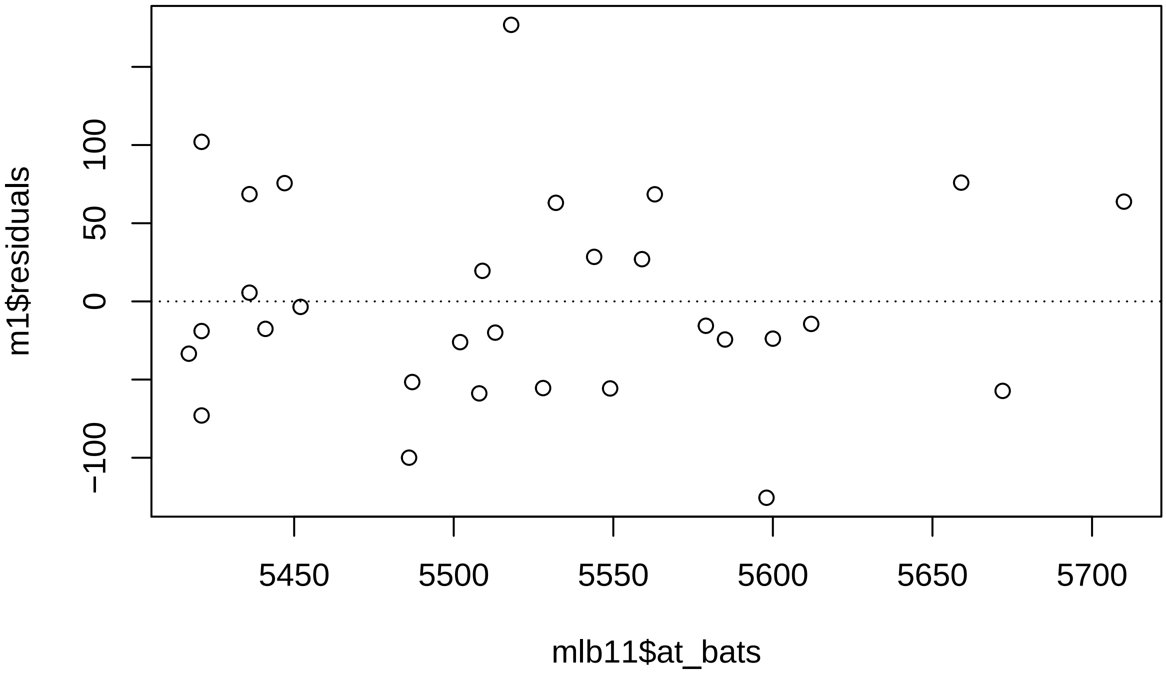
Answer

By only looking at the regression line, the manager would predict about 727.6861 runs for a team with 5,578 at-bats. The closest values in the data set that match 5,578 at-bats is 5,579 at-bats with 713 runs. Based on this, the residual would be 728-713 = 15 runs, indicating that the manager would overestimate the runs.

Exercise 6

Linearity: Is there any apparent pattern in the residuals itself? What does this indicate about the linearity of the relationship btween runs and at-bats?

```
plot(m1$residuals ~ mlb11$at_bats)
abline(h = 0, lty = 3)
```



Answer

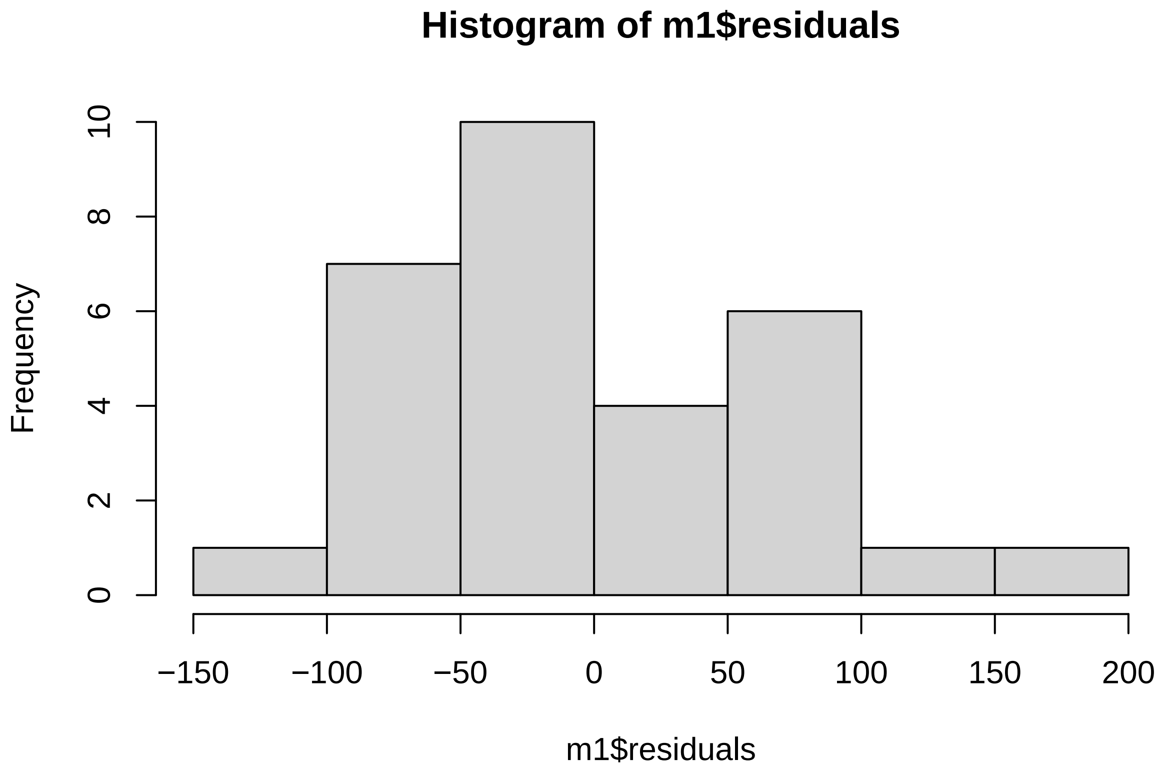
Based on the residual plot, there does not appear to be any pattern in the residuals, and the data may be skewed. However, the data still appears to be linear.

Exercise 7

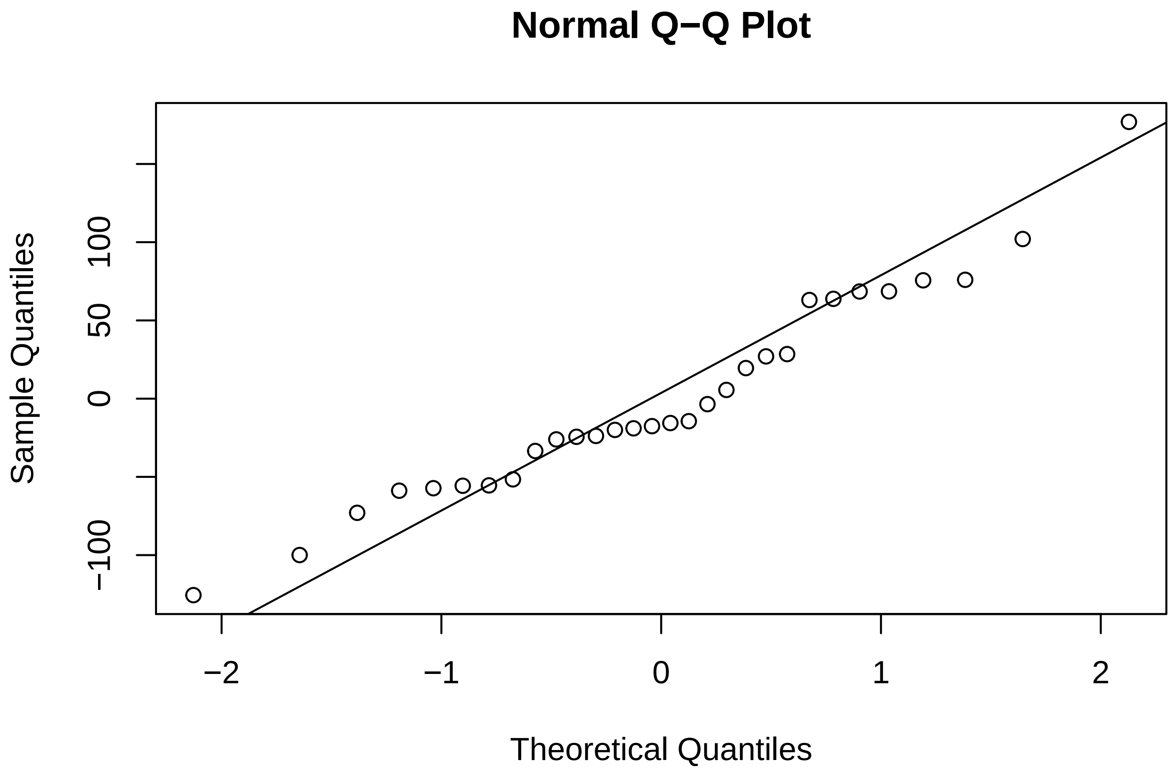
Nearly normal residuals: Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

```
## check normality plot

# method 1
hist(m1$residuals)
```



```r
# method 2
qqnorm(m1$residuals)
qqline(m1$residuals)
```



Answer

Based on the histogram and the normality plot, the distribution of residuals appears fairly normal, and that this meets the condition of nearly normal residuals.

Exercise 8

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

Answer

Based on the plot of points, the variability appears fairly constant through the data. I conclude that the condition of constant variability is met.

On Your Own

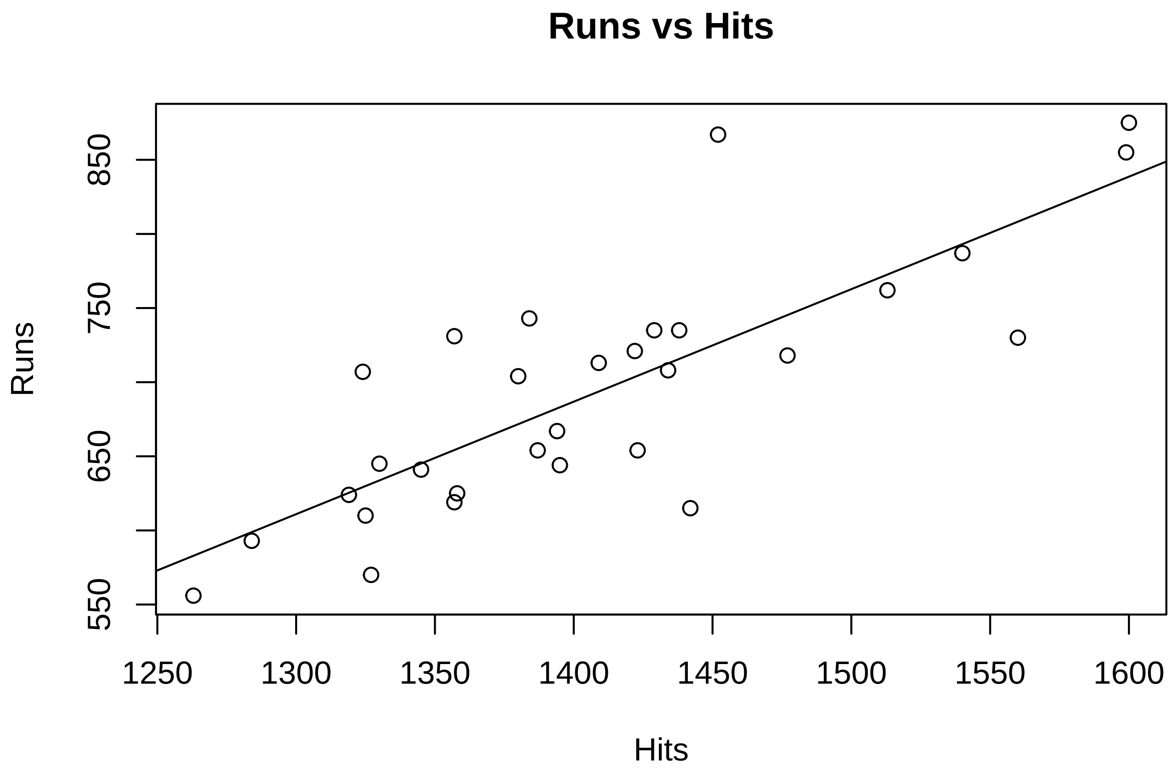
Question 1

Choose another traditional variable `hits` *from* `mlb11` *that we think might be a good predictor of* `runs`. *Show your R code to fit a linear model. At a glance, does there seem to be a linear relationship?*

```r
m3 <- lm(runs ~ hits, data = mlb11)
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
```

```
plot(hits, runs,
      main="Runs vs Hits",
      xlab="Hits",
      ylab="Runs"
     )
abline(m3)
```



Answer

At a glance, there does seem to a linear relationship, as the regression line appears to be a fairly good fit for the data.

Question 2

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.*

```
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(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
```

Answer

The R^2 for the regression between `runs` and `hits` is higher than the R^2 of the regression between `runs` and `at_bats` at 0.6419 vs 0.3729. This may mean that the new model explains 26.9% more of the variation in the data than the previous model between `runs` and `at_bats`, indicating that it may be the better model.

Question 3

Now that you can summarize the linear relationship between two variables, investigate the relationships between `runs` *and each of the other four traditional variables:* `bat_avg`*,* `strikeouts`*,* `stolen_bases`*, and* `wins`*. 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 four).*

```
summary(lm(runs ~ bat_avg, data = mlb11))$r.squared
```

```
## [1] 0.6560771
```

```
summary(lm(runs ~ strikeouts, data = mlb11))$r.squared
```

```
## [1] 0.1693579
```

```
summary(lm(runs ~ stolen_bases, data = mlb11))$r.squared
```

```
## [1] 0.002913993
```

```
summary(lm(runs ~ wins, data = mlb11))$r.squared
```

```
## [1] 0.3609712
```

```
# The data with the highest R-squared is the regression between runs and batting average, 0.6560771

## support conclusion using graphical and numerical methods

# 1: look at correlation
cor(runs, bat_avg)
```

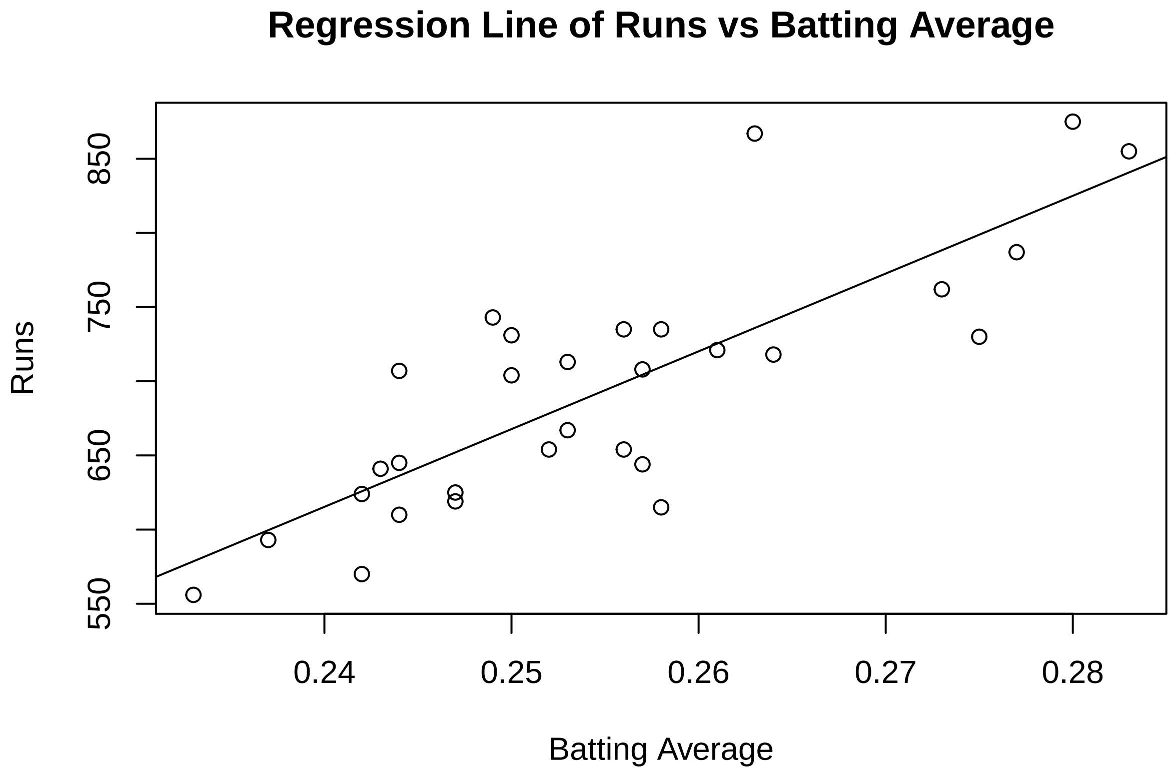
```
## [1] 0.8099859
```

```
# 2: look at scatterplot and regresion line
m4 <- lm(runs~bat_avg,data=mlb11)
summary(m4)
```

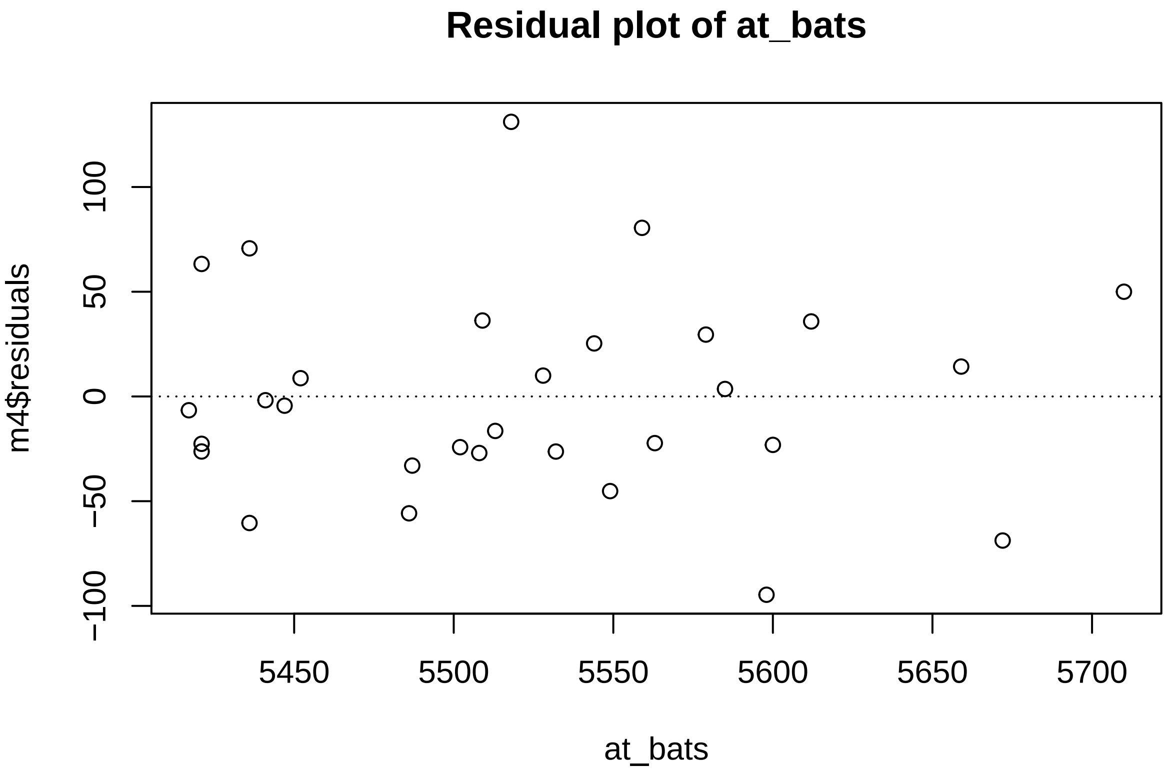
```
##
## Call:
## lm(formula = runs ~ bat_avg, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.676 -26.303  -5.496  28.482 131.113
```

```
##
## 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
```

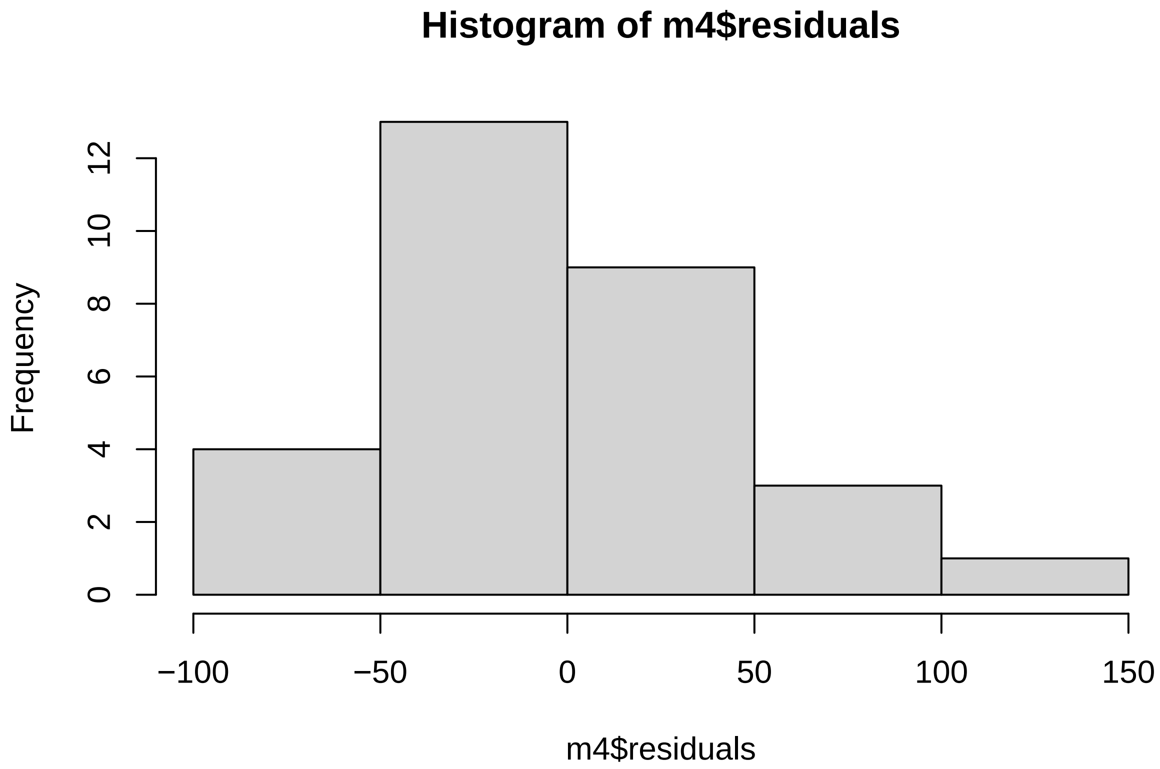
```
plot(runs~bat_avg,
     main="Regression Line of Runs vs Batting Average",
     xlab = "Batting Average",
     ylab = "Runs"
     )
abline(m4)
```



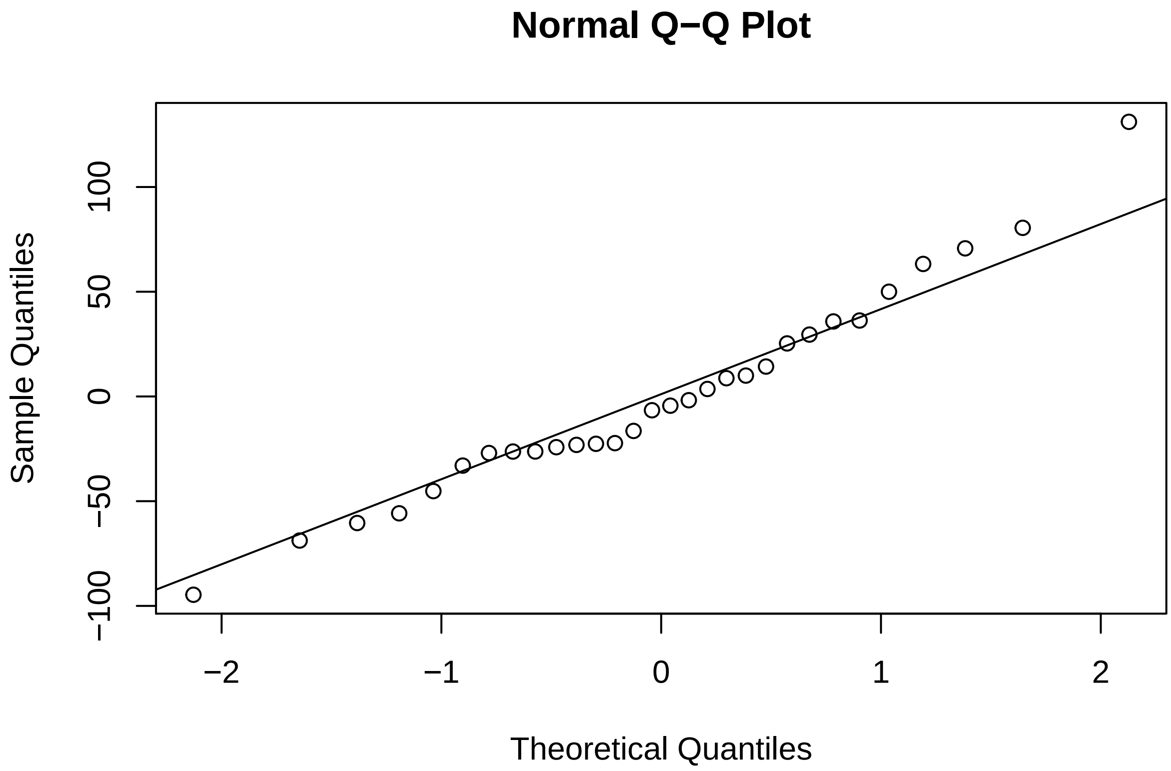
```
# 3: checking linearity using residuals
plot(m4$residuals~at_bats,
     main="Residual plot of at_bats"
     )
abline(h = 0, lty = 3)
```



```
# 4: Checking for nearly normal residuals
hist(m4$residuals)
```



```
qqnorm(m4$residuals)
qqline(m4$residuals)
```



Answer

From comparing the R^2 of the different variables, we see that batting average has the highest value, 0.6561. From the graphs we generated, we can see that the data for batting average supports the conditions needed for linear regression.

Question 4

Now examine the three newer variables, `new_slug`*,* `new_obs`*,* `new_onbase`*. These are the statistics used by the author of Moneyball to predict a team's success. Of all ten variables we've analyzed, which seems to be the best predictor of* `runs`*?*

```
# examine correlation
cor(runs, new_slug)
```

```
## [1] 0.9470324
```

```
cor(runs, new_obs)
```

```
## [1] 0.9669163
```

```
cor(runs, new_onbase)
```

```
## [1] 0.9214691
```

```
# compare R-squared values
summary(lm(runs~new_slug, data = mlb11))$r.squared
```

```
## [1] 0.8968704
```

```
summary(lm(runs~new_obs, data = mlb11))$r.squared
```

```
## [1] 0.9349271
```

```
summary(lm(runs~new_onbase, data = mlb11))$r.squared
```

```
## [1] 0.8491053
```

Answer

Based on R^2 values, the regression model with `new_obs` seems to be the best predictor of `runs`, with the largest value of 0.9349.

Question 5

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

```
# The data with the highest R-squared is the regression between runs and new_obs, 0.9349

## support conclusion using graphical and numerical methods

# 1: look at correlation
cor(runs, new_obs)
```

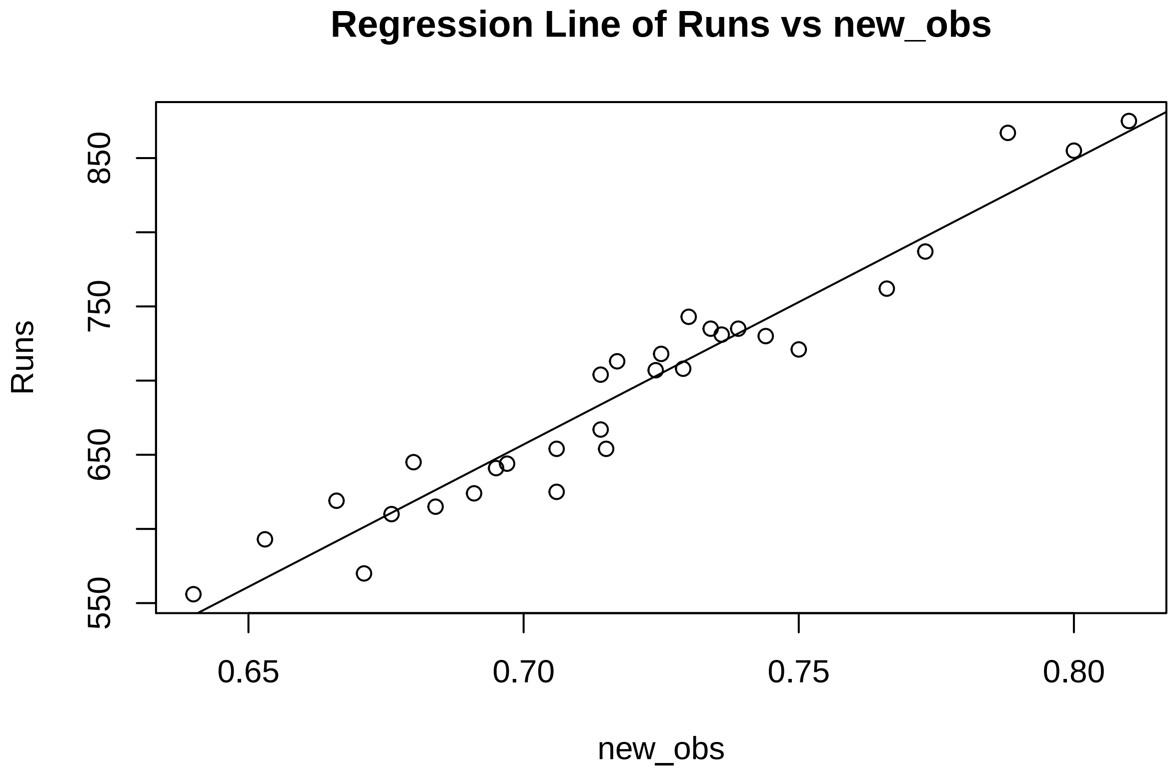
```
## [1] 0.9669163
```

```
# 2: look at scatterplot and regresion line
m5 <- lm(runs~new_obs,data=mlb11)
summary(m5)
```

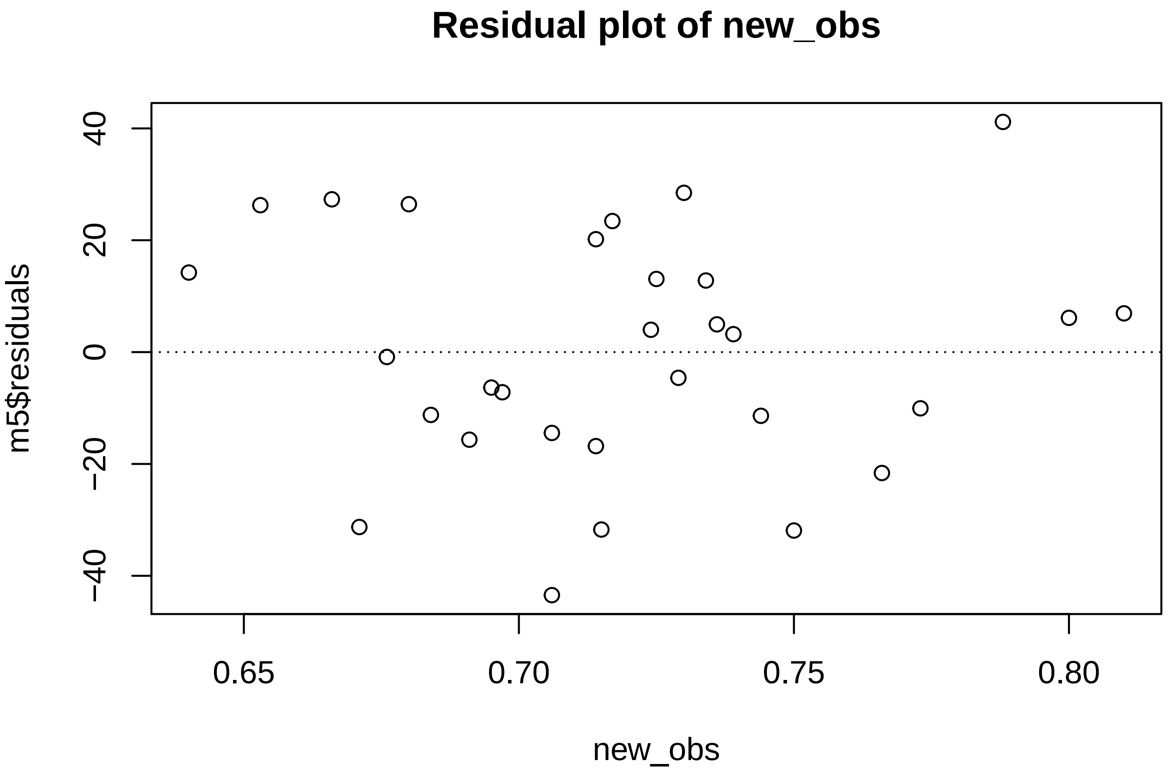
```
##
## Call:
## lm(formula = runs ~ new_obs, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.456 -13.690   1.165  13.935  41.156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -686.61      68.93  -9.962 1.05e-10 ***
## new_obs      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
```

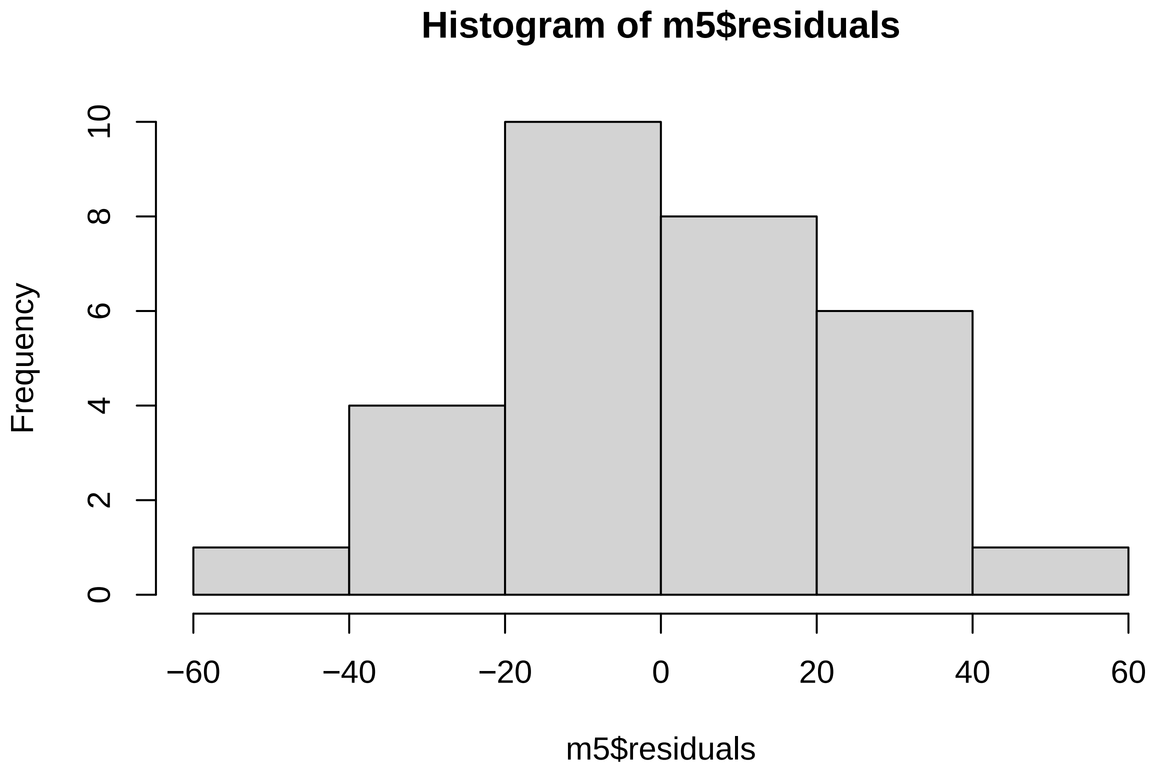
```
plot(runs~new_obs,
     main="Regression Line of Runs vs new_obs",
     xlab = "new_obs",
     ylab = "Runs"
     )
abline(m5)
```



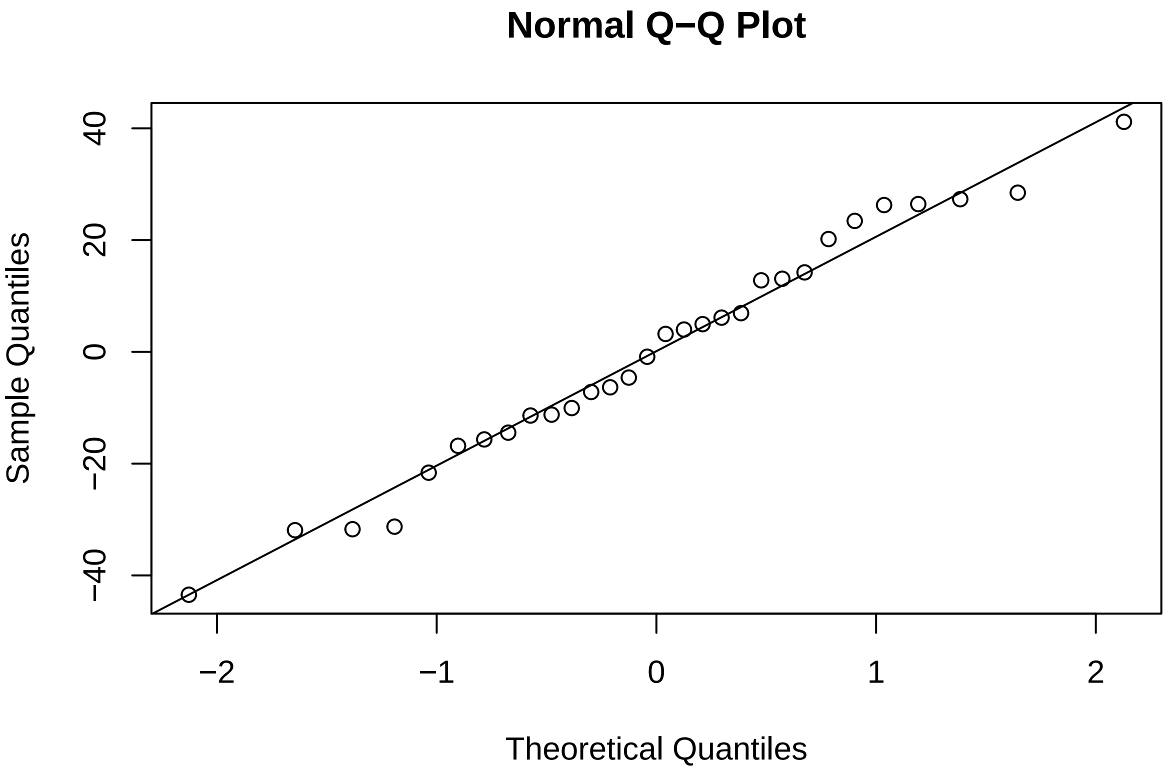
```
# 3: checking linearity using residuals
plot(m5$residuals~new_obs,
     main="Residual plot of new_obs"
     )
abline(h = 0, lty = 3)
```



```
# 4: Checking for nearly normal residuals
hist(m5$residuals)
```



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



Answer

Based on the model diagnostics for `new_obs`, the data seems to meet all of the conditions for simple linear regression.