MSDS 5043 - Assigment 6

Olivia Samples, Lipscomb University

10/20/2019

Warm Up

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" )
```

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. Your homework will focuse on the newer variables on your own.

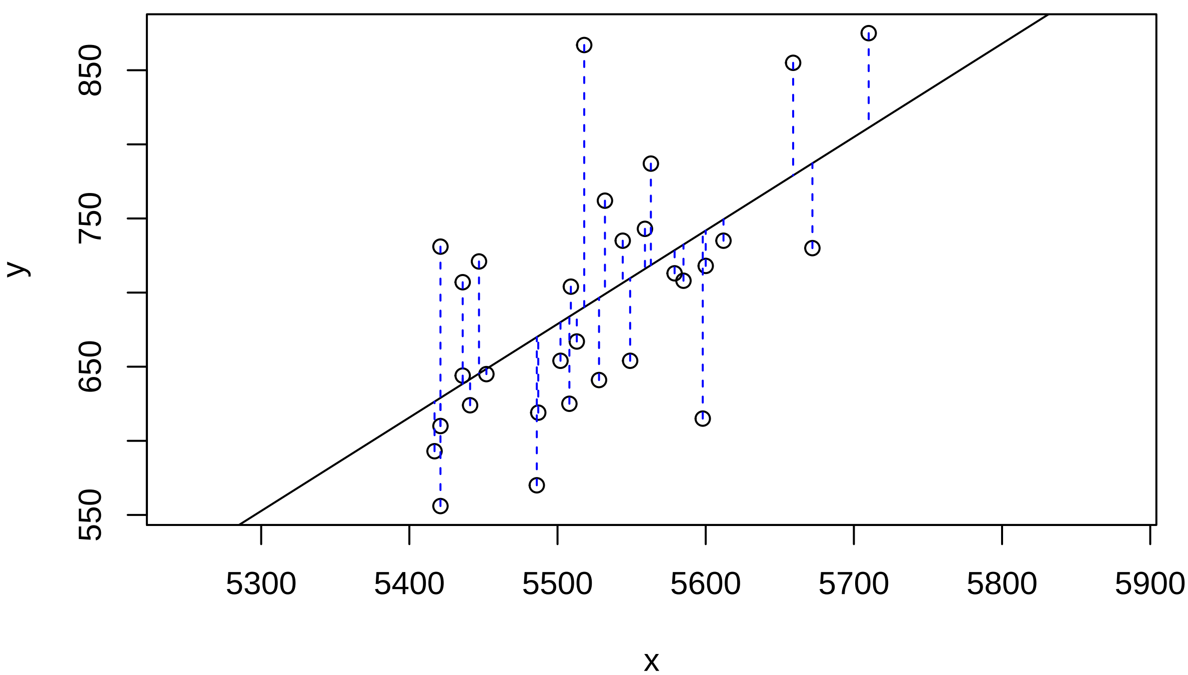
(1) Relationship between variables

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

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

Sum of squared residuals

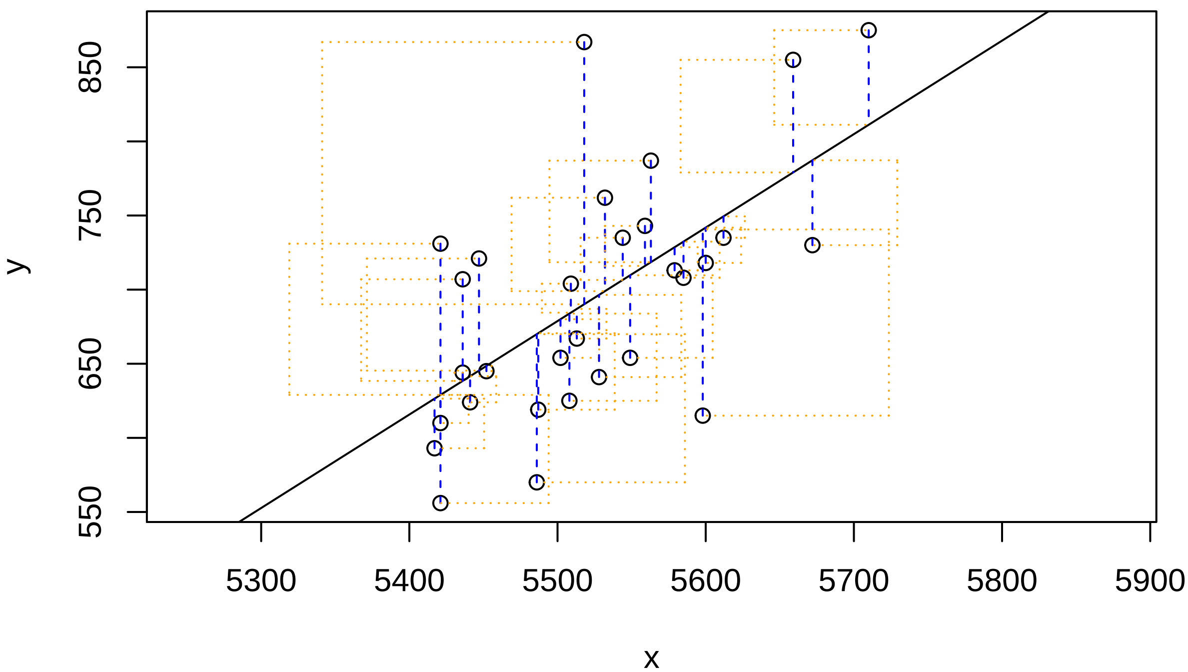
```
plot_ss(x = mlb11$at_bats, y = mlb11$runs )
```



```
## 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
```

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.

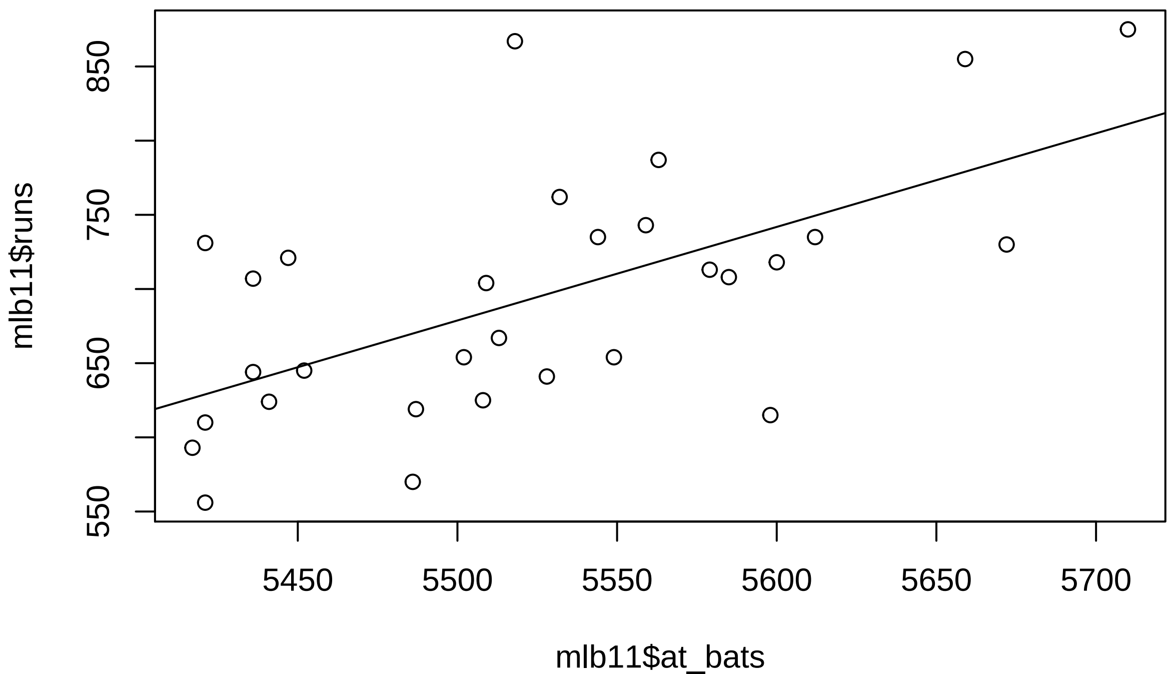
The linear model

```
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
```

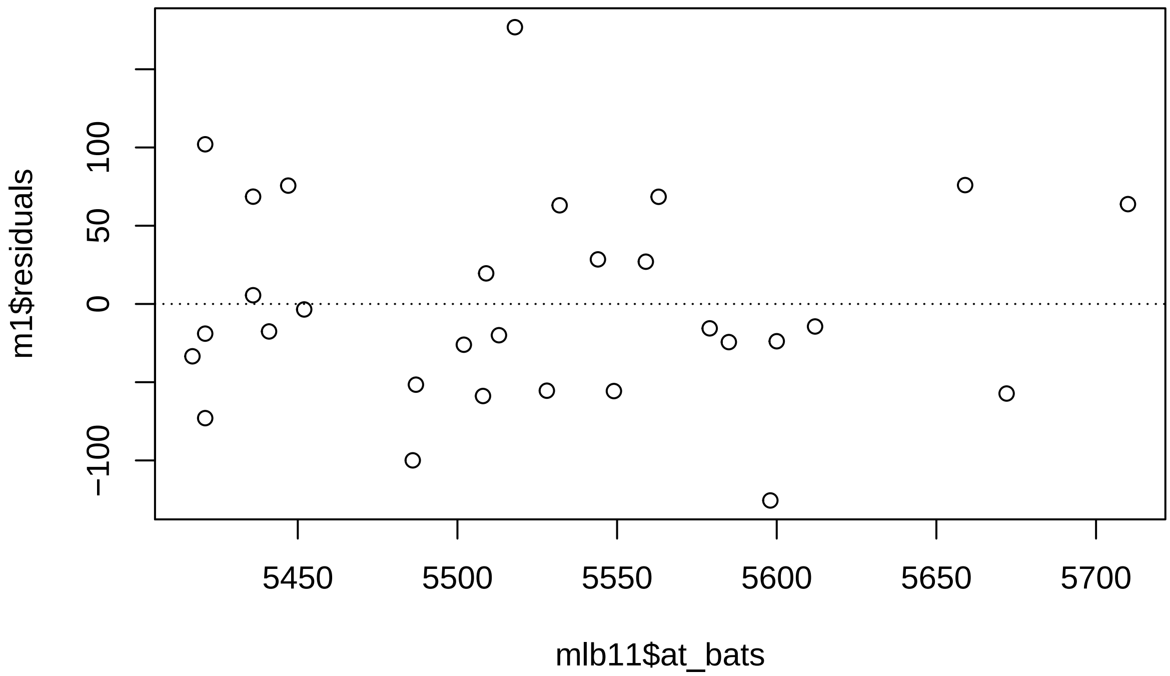
```r
plot(mlb11$runs ~ mlb11$at_bats )
abline(m1)
```



To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

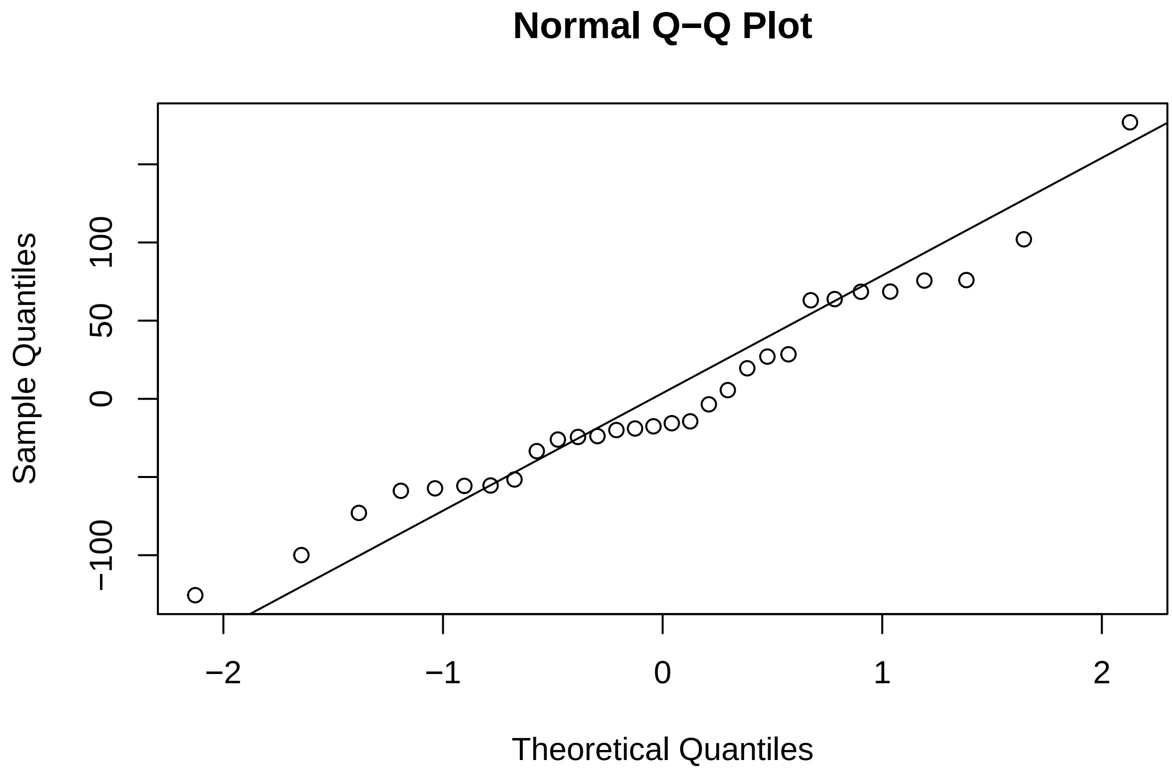
(1) Linearity: You already checked if the relationship between runs and at-bats is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. at-bats. Recall that any code following a # is intended to be a comment that helps understand the code but is ignored by R.

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



(2) 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 )
```



Homework

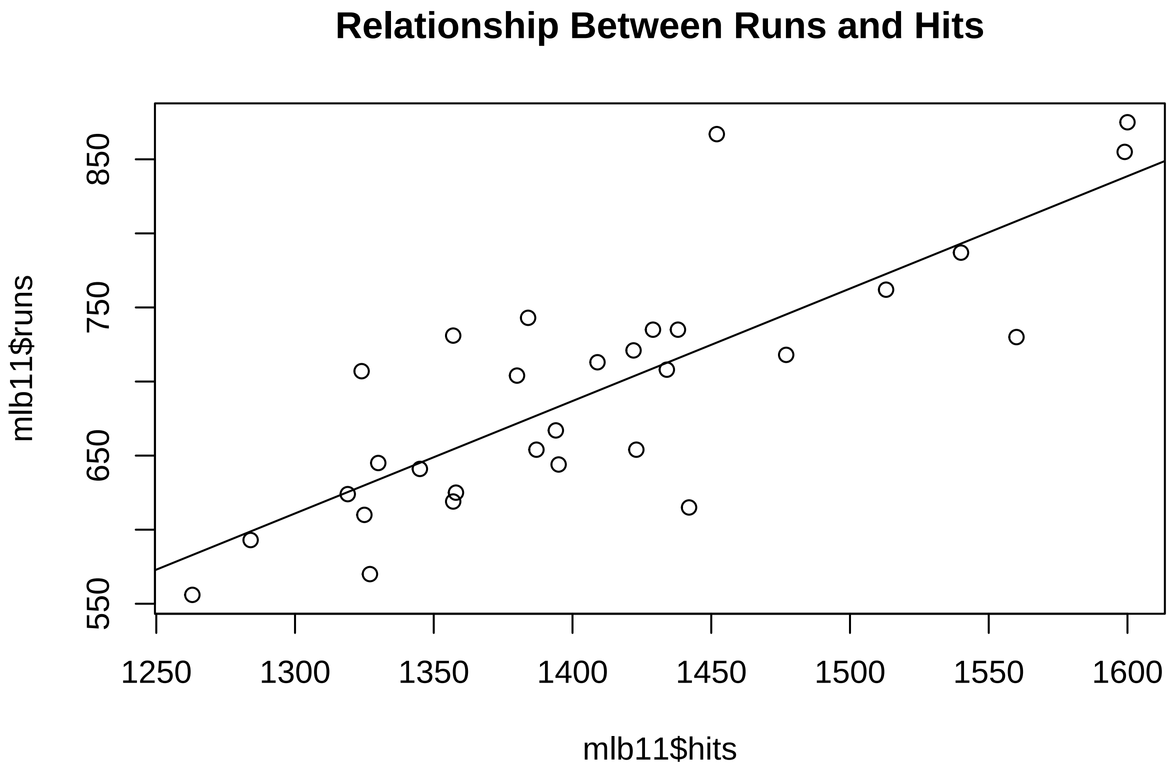
(1) Choose another traditional variable from mlb11 that you think might be a good predictor of runs. Produce a scatterplot of the two variables and fit a linear model. Does there seem to be a linear relationship?

```
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
##
```

I will choose hits.

```
plot(mlb11$runs ~ mlb11$hits, main = "Relationship Between Runs and Hits")
lm <- lm(mlb11$runs ~ mlb11$hits)
abline(lm)
```




```r
summary(lm)
```

```
##
## Call:
## lm(formula = 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 Pr(>|t|)
## (Intercept) -375.5600   151.1806  -2.484   0.0192 *
## mlb11$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
```

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

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

Here, the correlation is high which means the variables are strongly correlated. Similarly, the linear model statistics and plot show that positive and linear.

(2) How does this relationship compare to the relationship between runs and at_bats ? Use the R2 values from the two model summaries to compare. Does your variable seem to predict runs better than at bats ? How can you tell?

We have that the R^2 statistic from Runs~At_Bats is 0.3729 and the R^2 statistic from Runs~Hits is 0.6419. This means that the relationship between runs and hits is 26.9% more significant than the relationship between runs and at_bats. We know this because the R^2 of a linear model describes the amount of variation in the dependent variable. The relationship associated with the greater percentage describes the linear model better.

(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

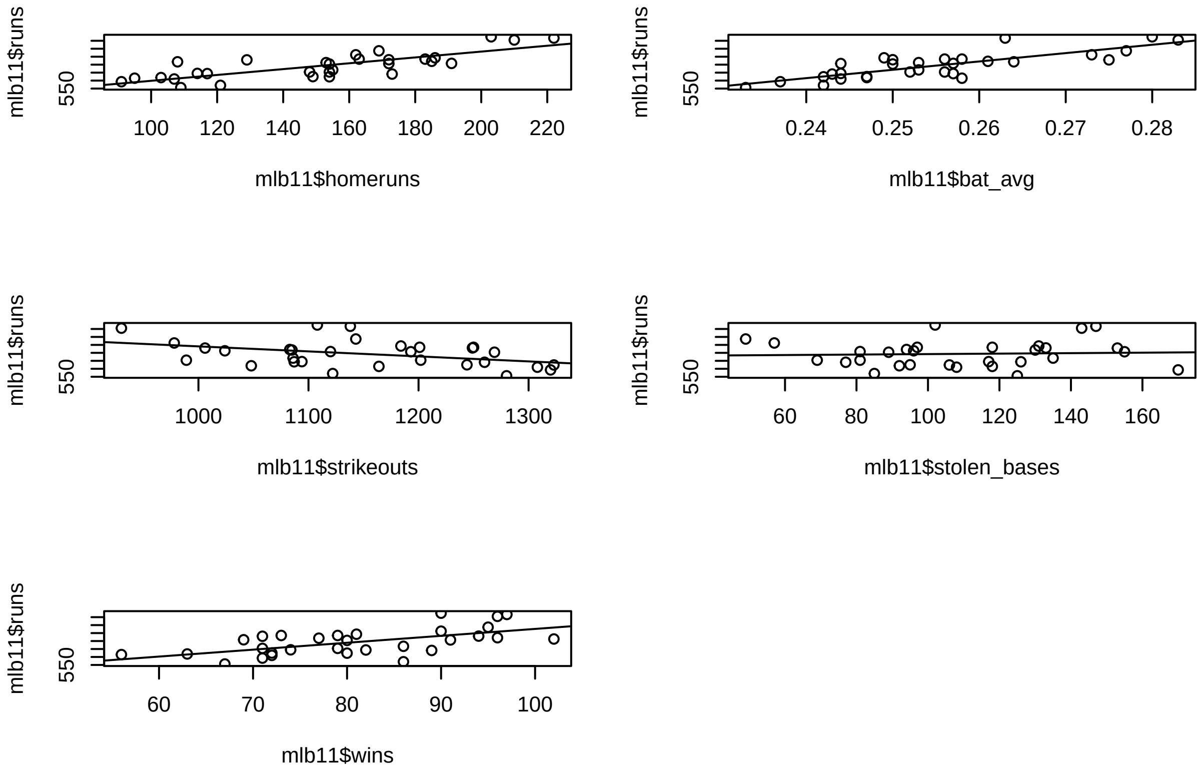
```
lm1 <- lm(mlb11$runs ~ mlb11$homeruns + mlb11$bat_avg + mlb11$strikeouts + mlb11$stolen_bases + mlb11$wi
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$homeruns + mlb11$bat_avg + mlb11$strikeouts +
##     mlb11$stolen_bases + mlb11$wins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.455 -24.247   2.674  21.418  41.078
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -701.78982  198.06819  -3.543  0.00166 **
## mlb11$homeruns        1.05296    0.20479   5.142 2.90e-05 ***
## mlb11$bat_avg      4081.54263  570.10770   7.159 2.12e-07 ***
## mlb11$strikeouts      0.06283    0.06167   1.019  0.31843
## mlb11$stolen_bases    0.51121    0.16575   3.084  0.00508 **
## mlb11$wins            0.82739    0.59097   1.400  0.17429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.38 on 24 degrees of freedom
## Multiple R-squared:  0.9154, Adjusted R-squared:  0.8977
## F-statistic: 51.91 on 5 and 24 DF,  p-value: 4.339e-12
```

From these linear model statistics, we can see that the variable with the largest coefficient is batting average. This also has the smallest p-value which suggests that it is the most statistically significant. We can further highlight the significance of the batting average by comparing the plots of each of the scatter plots.

```
par(mfrow = c(3, 2))
plot(mlb11$runs ~ mlb11$homeruns)
abline(lm(mlb11$runs ~ mlb11$homeruns))
plot(mlb11$runs ~ mlb11$bat_avg)
abline(lm(mlb11$runs ~ mlb11$bat_avg))
plot(mlb11$runs ~ mlb11$strikeouts)
abline(lm(mlb11$runs ~ mlb11$strikeouts))
plot(mlb11$runs ~ mlb11$stolen_bases)
abline(lm(mlb11$runs ~ mlb11$stolen_bases))
plot(mlb11$runs ~ mlb11$wins)
abline(lm(mlb11$runs ~ mlb11$wins))
```



And from comparing these scatter plots, we can see that the bat_avg vs. runs plot has the strongest positively linear relationship.

(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 than 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?

```
#?cor
cor(mlb11$runs, mlb11[sapply(mlb11, is.numeric)])
```

```
##      runs  at_bats      hits  homeruns   bat_avg strikeouts stolen_bases
## [1,]    1 0.610627 0.8012108 0.7915577 0.8099859 -0.4115312   0.05398141
##           wins new_onbase  new_slug   new_obs
## [1,] 0.6008088  0.9214691 0.9470324 0.9669163
```

Here, the result shows that the three newer variables have the three highest correlation coefficients. And so, in general we can say that these newer variables are more significant than the older variables We will compare the R^2 statistic of these three newer variables to see which is the best predictor of runs.

```
lm2 <- lm(mlb11$runs ~ mlb11$new_obs)
summary(lm2)
```

```
##
## 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
```

```r
lm3 <- lm(mlb11$runs ~ mlb11$new_onbase)
summary(lm3)
```

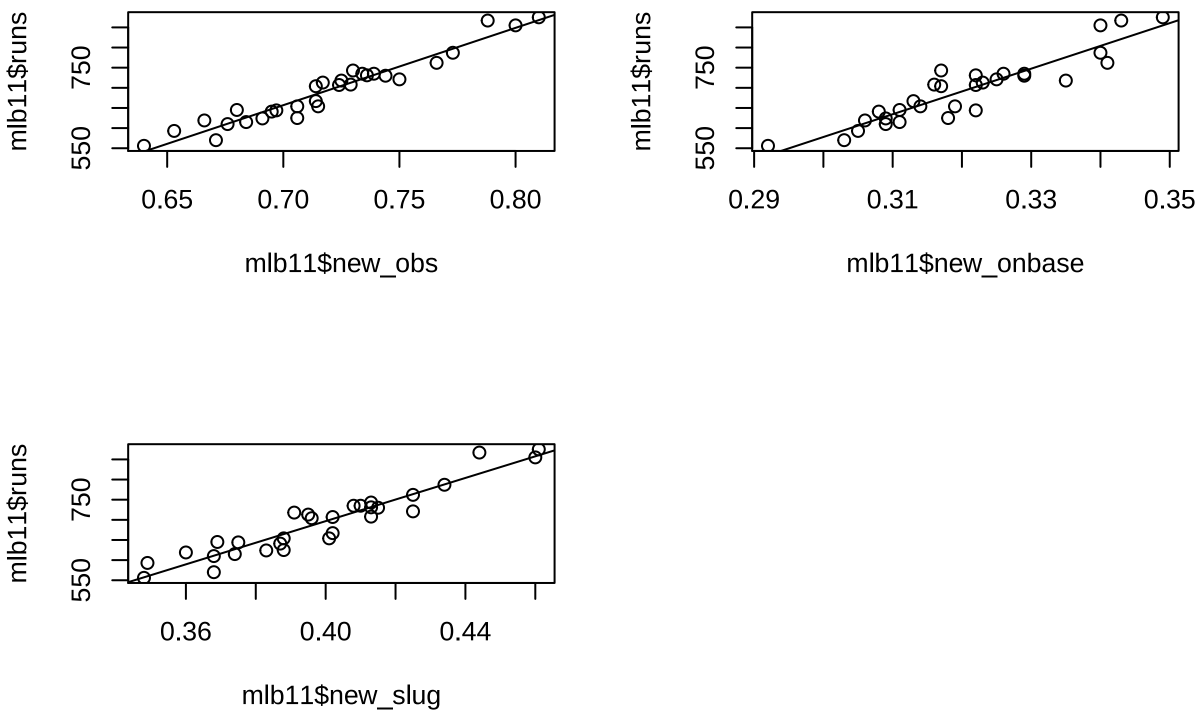
```
##
## Call:
## lm(formula = mlb11$runs ~ mlb11$new_onbase)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.270 -18.335   3.249  19.520  69.002
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1118.4      144.5  -7.741 1.97e-08 ***
## mlb11$new_onbase   5654.3      450.5  12.552 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
```

```r
lm4 <- lm(mlb11$runs ~ mlb11$new_slug)
summary(lm4)
```

```
##
## 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 Pr(>|t|)
## (Intercept)      -375.80      68.71   -5.47 7.70e-06 ***
## mlb11$new_slug  2681.33     171.83   15.61 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
```

Here, the R^2 statistic of new_obs is the largest. Similarly, its correlation coefficient was the largest. We can conclude then that the variable, new_obs, is the best predictor of runs from all 10 variables. We can confirm this by looking at the plots of each. When comparing the scatter plot with the fitted line of each, we can see that the runs vs. new_obs plot has the best fit line of the three.

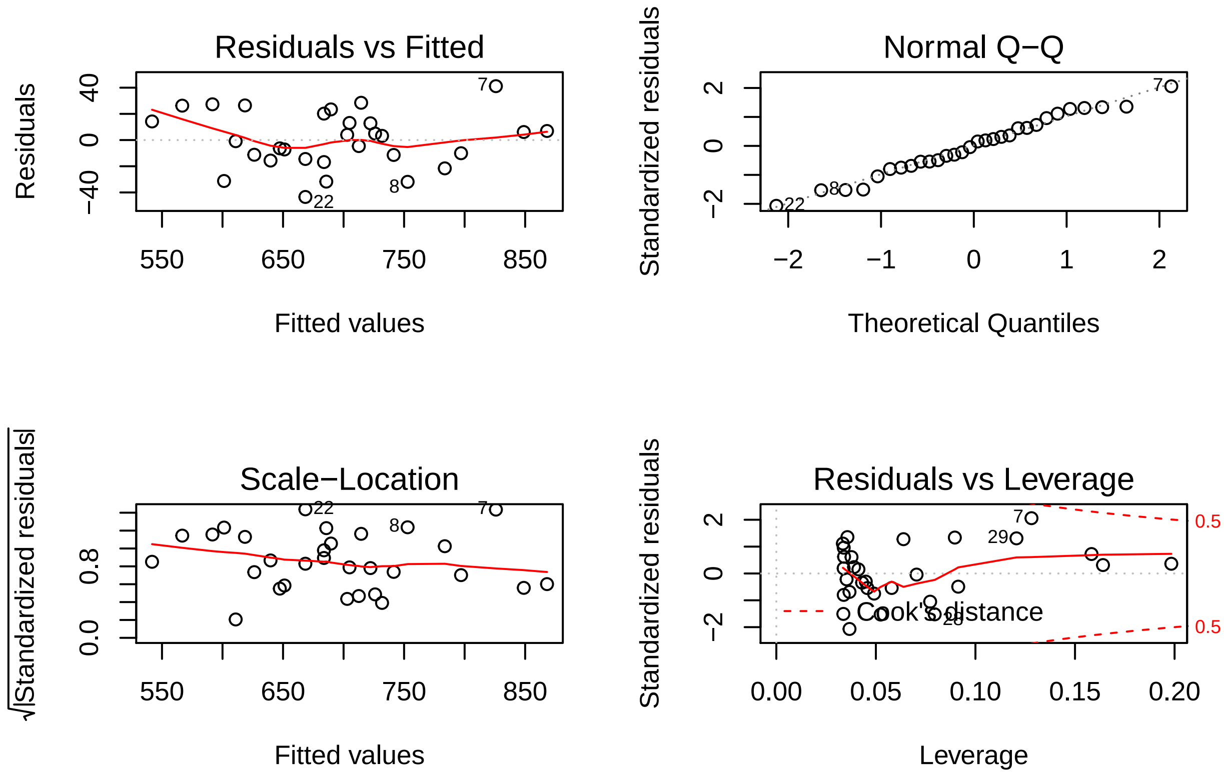
```
par(mfrow = c(2, 2))
plot(mlb11$runs ~ mlb11$new_obs)
abline(lm2)

plot(mlb11$runs ~ mlb11$new_onbase)
abline(lm3)

plot(mlb11$runs ~ mlb11$new_slug)
abline(lm4)
```



That would make sense because your OBG stands for on-base-percentage, and the more often you are on base, the more often you get runs. Also, in order to get a run, you have to be on base.

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

```
par(mfrow = c(2, 2))
plot(lm2)
```



(a) Linearity: First we will check linearity using Residuals vs. Fitted plot.

This plot gives us approximately constant variability of residuals without strong curves or indications of non-normality. However, there are some bends because n is not large.

```
length(mlb11$runs)
```

```
## [1] 30
```

(b) Nearly Normal Residuals: Now we will test to make sure the residuals are nearly normal using a normal probability plot.

We have that our Normal Q-Q plot results in our residuals follow the straight line which indicates that we can assume the residuals are fairly normal.

(c) Constant variability: Finally, we will reference our Scale-Location plot to see if the variability is fairly constant.

If the line is horizontal and contains equally spread points, then we can assume constant variability. Our plot is mostly fitting these requirements with a fairly horizontal line, but a slight slope is noticeable possibly due to the small n.