# Introduction to linear regression

```
library("ggplot2")
```

## Batter up

The movie Moneyball focuses on the "quest for the secret of success in baseball". It follows a low-budget team, the Oakland Athletics, who believed that underused statistics, such as a player's ability to get on base, betterpredict the ability to score runs than typical statistics like home runs, RBIs (runs batted in), and batting average. Obtaining players who excelled in these underused statistics turned out to be much more affordable for the team.

In this lab 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.

```
load("more/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. At the end of the lab, you'll work with the newer variables on your own.

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?

We can use scatter plot to find out the relationship between `runs` and `at_bats`.

```
ggplot(mlb11,aes(x=at_bats,y=runs)) + geom_point() + geom_smooth(method="lm")
```
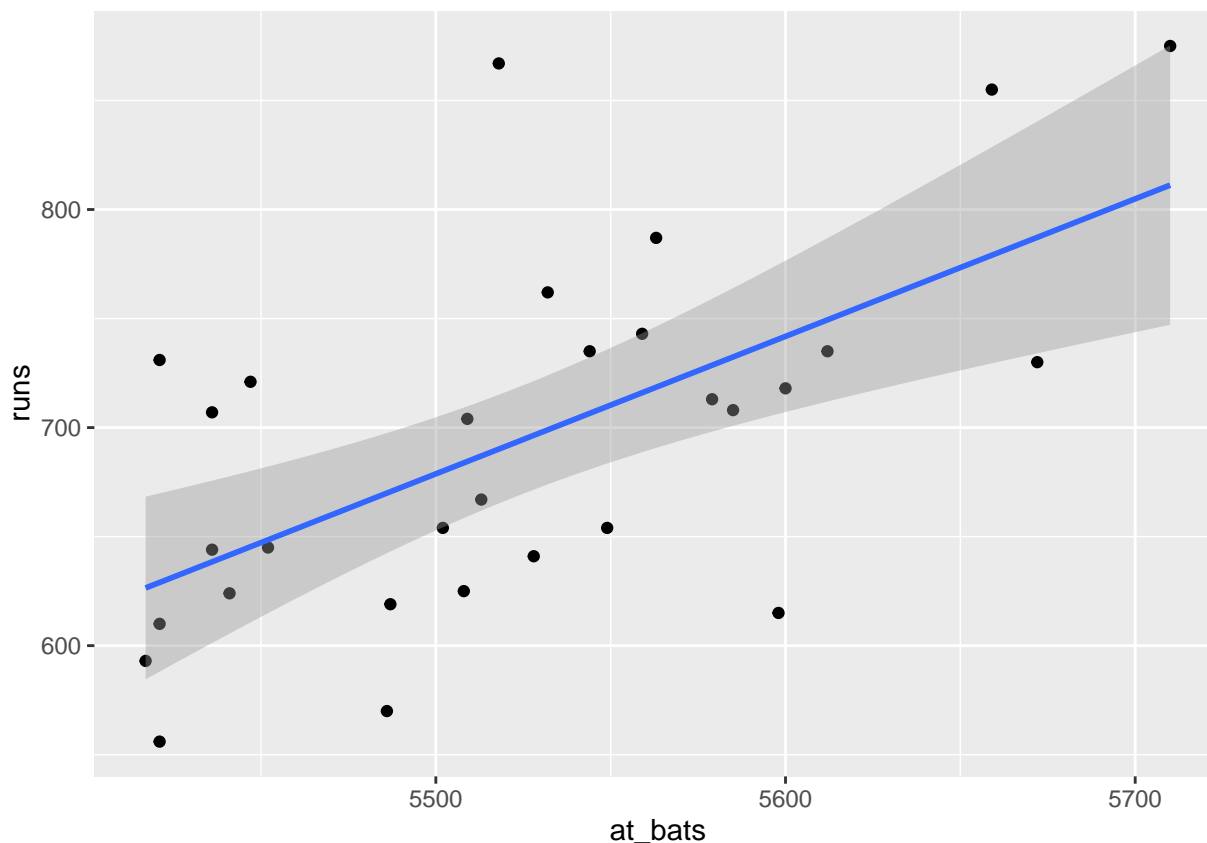
The plot shows the relationship is linear and positive. The correlation coeffiecient(R) is `0.61`. So if we know the team's `at_bats`, we can approximatly predict the number of runs.

If the relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.

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

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

## Sum of squared residuals

Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as `runs` and `at_bats` above.
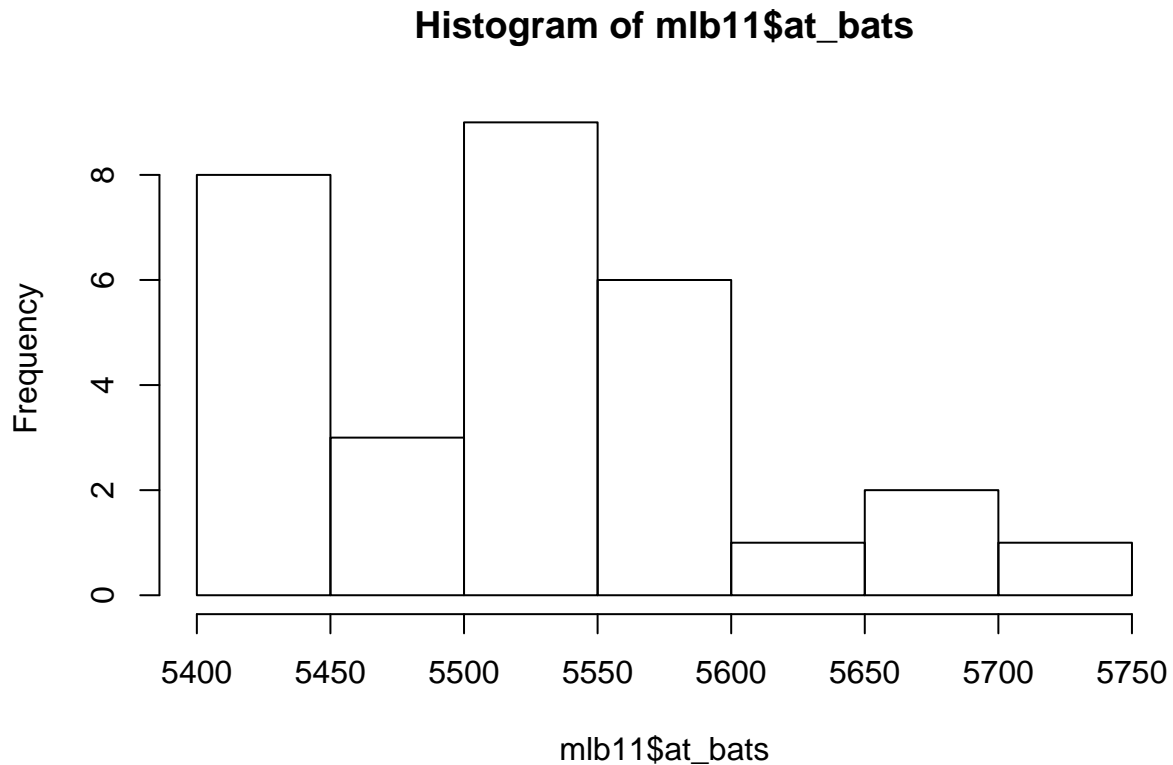
2. Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

```
#runs
hist(mlb11$runs)
```

**Histogram of mlb11$runs**
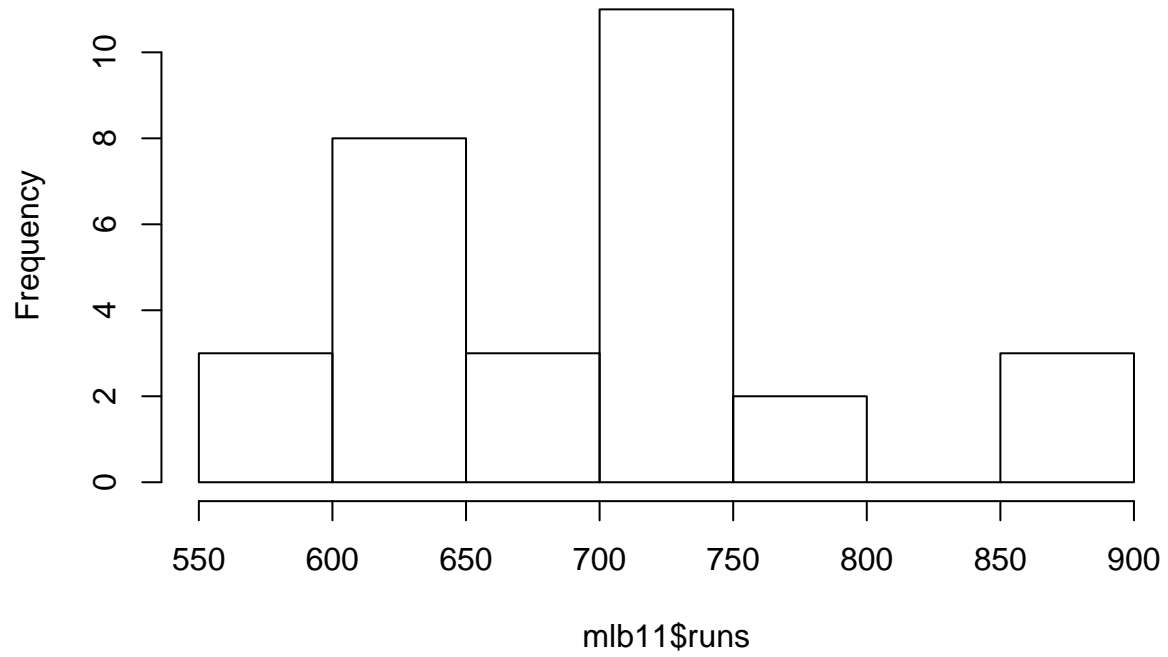


```r
#at_bats
hist(mlb11$at_bats)
```

# Histogram of mlb11$at_bats



For `runs`, the historgram looks like single peak. The distribuion is almost symmentrical. There are also an outlier.

For `at_bats`, the historgram looks like dual peak. The distribution is little bit right skewed.

As the population size is around 30, we have to make some reasonable assumptions on distributions and consider it as normal distribution.

The direction between two variables is positive. But the relationship is moderately weak. The linear regression line did not fall or fit into any actual points.

There are some influential points in the plot which alters the regression line.

Just as we used the mean and standard deviation to summarize a single variable, we can summarize the relationship between these two variables by finding the line that best follows their association. Use the following interactive function to select the line that you think does the best job of going through the cloud of points.

```
plot_ss(x = mlb11$at_bats, y = mlb11$runs)
```
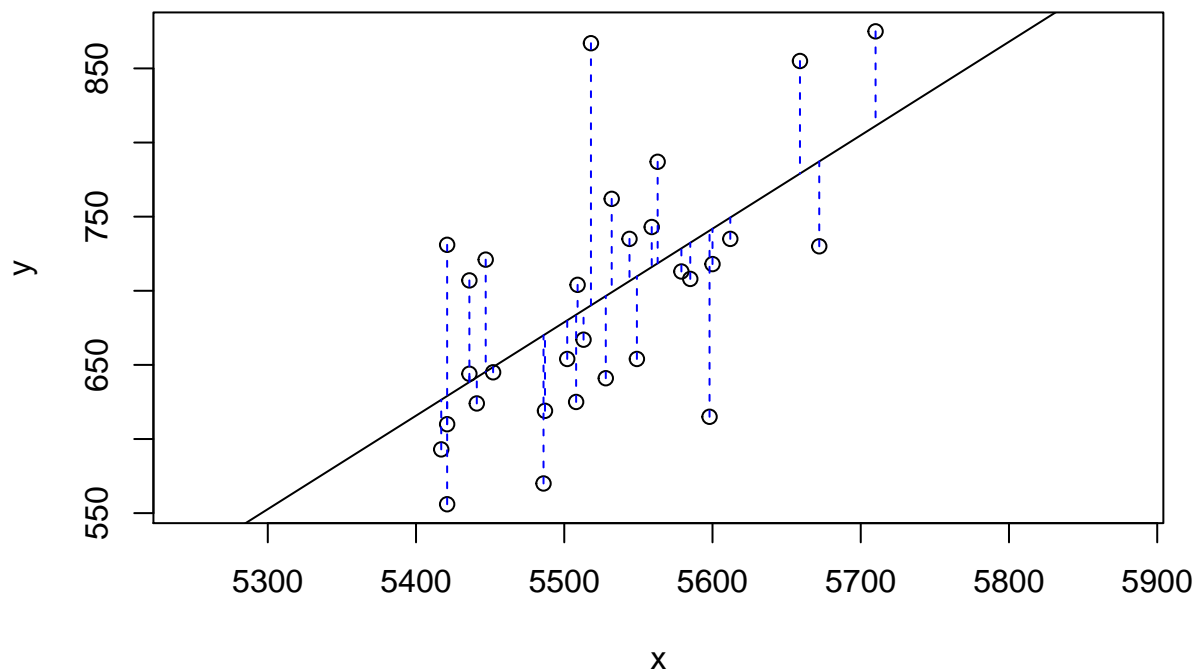
4

```
## 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
```

After running this command, you'll be prompted to click two points on the plot to define a line. Once you've done that, the line you specified will be shown in black and the residuals in blue. Note that there are 30 residuals, one for each of the 30 observations. Recall that the residuals are the difference between the observed values and the values predicted by the line:

$$e_i = y_i - \hat{y}_i$$

The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. To 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)
```
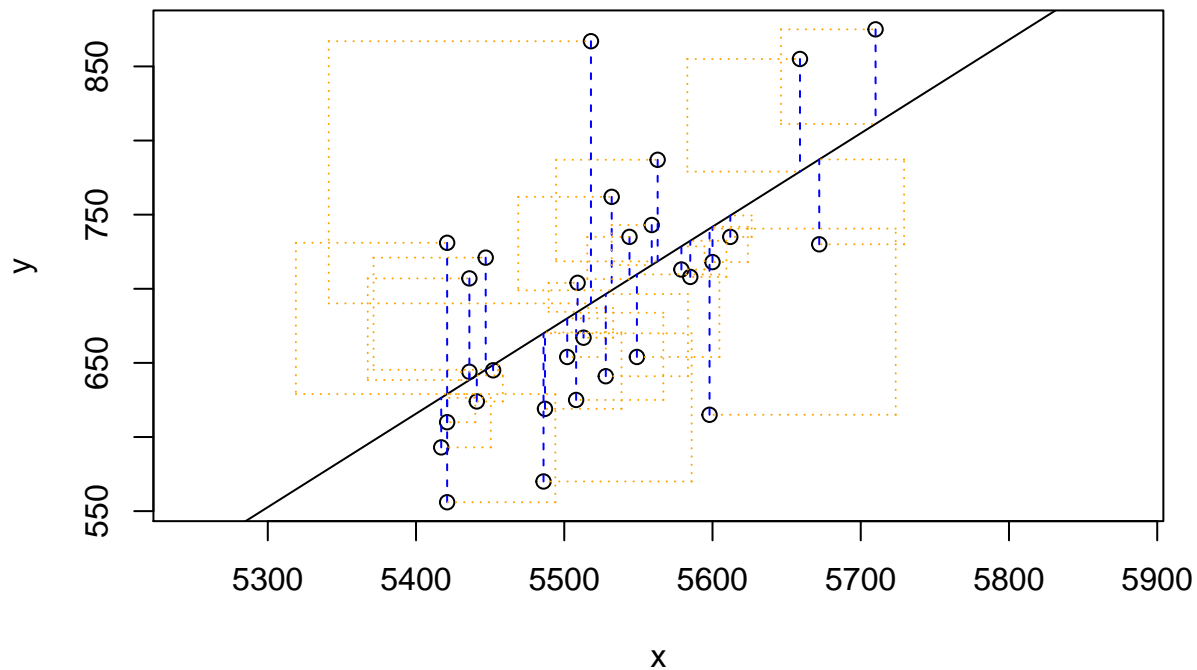
```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##  -2789.2429       0.6305
##
## Sum of Squares:   123721.9
```
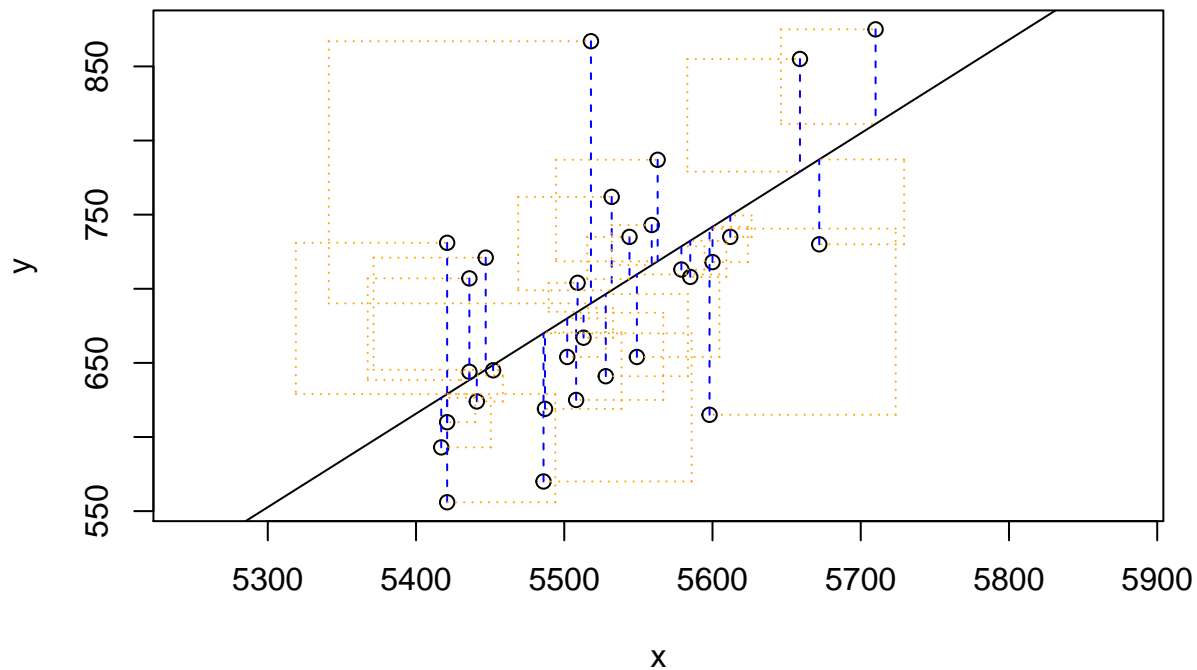
Note that the output from the `plot_ss` function provides you with the slope and intercept of your line as well as the sum of squares.

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?

```r
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
```

*#143146.7*
*#140789.6*
*#133267*
*#129510.2*
*#126750.9*

The lowest sum of squares that I got is `126750`. Each regression line has different sum of squares and different line of fit. The line of best fit has lower sum of squares.

## The linear model

It is rather cumbersome to try to get the correct least squares line, i.e. the line that minimizes the sum of squared residuals, through trial and error. Instead we can use the `lm` function in R to fit the linear model (a.k.a. regression line).

```
m1 <- lm(runs ~ at_bats, data = mlb11)
```

The first argument in the function `lm` is a formula that takes the form `y ~ x`. Here it can be read that we want to make a linear model of `runs` as a function of `at_bats`. The second argument specifies that R should look in the `mlb11` data frame to find the `runs` and `at_bats` variables.

The output of `lm` is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

```
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
```

Let's consider this output piece by piece. First, the formula used to describe the model is shown at the top. After the formula you find the five-number summary of the residuals. The "Coefficients" table shown next is key; its first column displays the linear model's y-intercept and the coefficient of `at_bats`. With this table, we can write down the least squares regression line for the linear model:

$$\hat{y} = -2789.2429 + 0.6305 * atbats$$

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, $R^2$. The $R^2$ value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 37.3% of the variability in runs is explained by at-bats.

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?

```
ggplot(mlb11,aes(x=homeruns,y=runs)) + geom_point() + geom_smooth(method="lm")
```
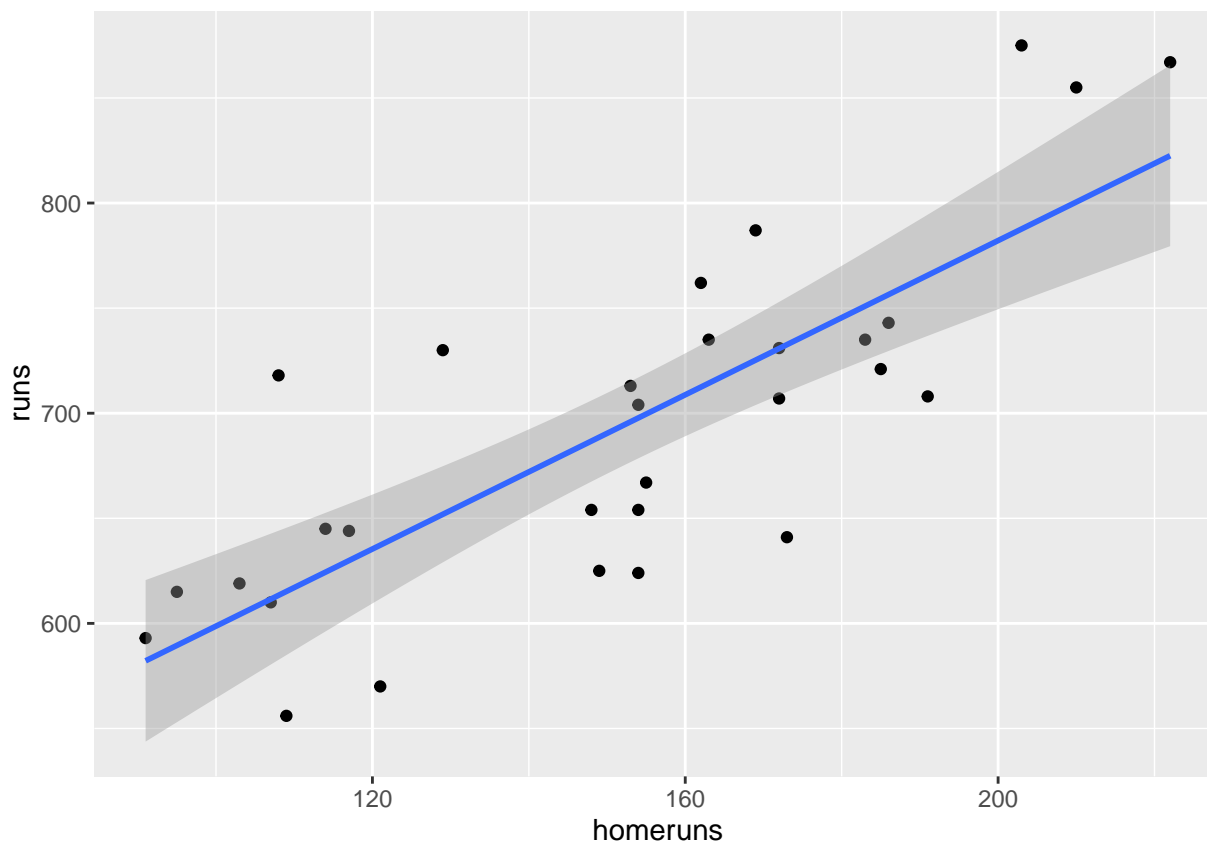
```r
m2 <- lm(runs ~ homeruns,mlb11)
summary(m2)
```

```
##
## 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
```

From the plot, is has a strong positive linear relationship between `homeruns` and `runs`. It also has some outliers, but it is not so influential.

From the summary, it has an y intercept of 415.23. The slope is 1.8345 with the standard error of 0.2677. $R^2$ value is 62.66%. So the slope tells that the if the team has more homeruns, their total runs will be more.

## Prediction and prediction errors

Let's create a scatterplot with the least squares line laid on top.

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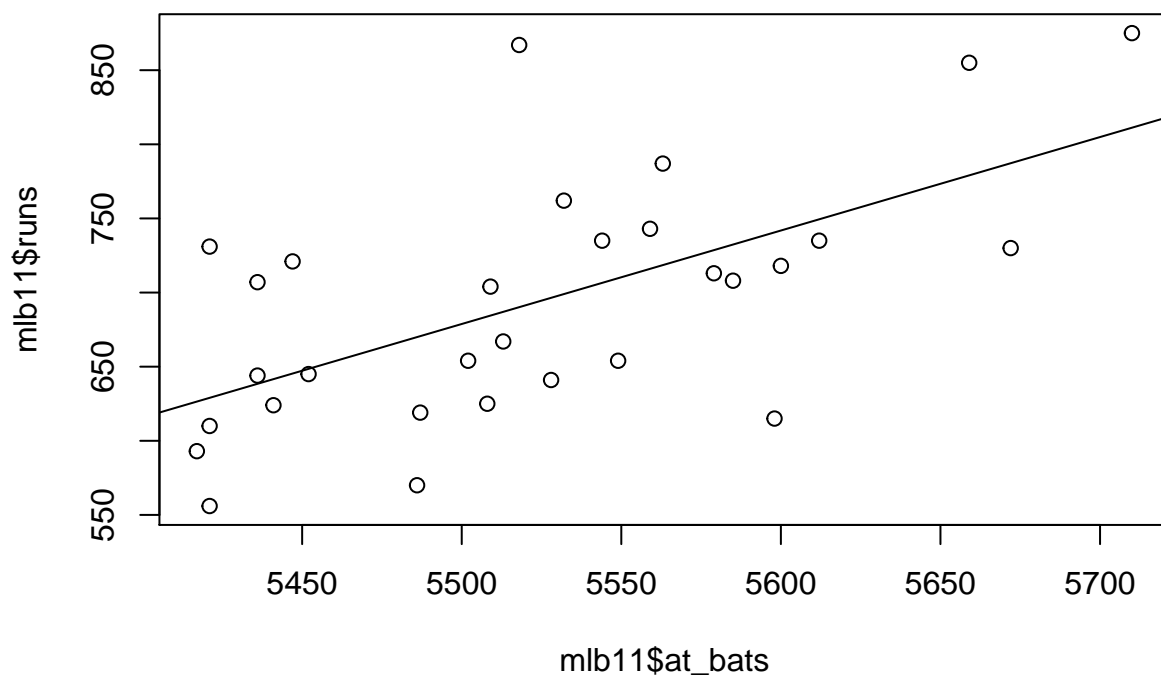


The function `abline` plots a line based on its slope and intercept. Here, we used a shortcut by providing the model `m1`, which contains both parameter estimates. This line can be used to predict $y$ at any value of $x$. When predictions are made for values of $x$ that are beyond the range of the observed data, it is referred to as *extrapolation* and is not usually recommended. However, predictions made within the range of the data are more reliable. They're also used to compute the residuals.

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?

```
plot(mlb11$runs ~ mlb11$at_bats)
abline(v=5578,lty = 3)  # adds a horizontal dashed line at y = 0
abline(m1)
abline(h=728,lty = 3)
abline(h=715,lty = 3)
```
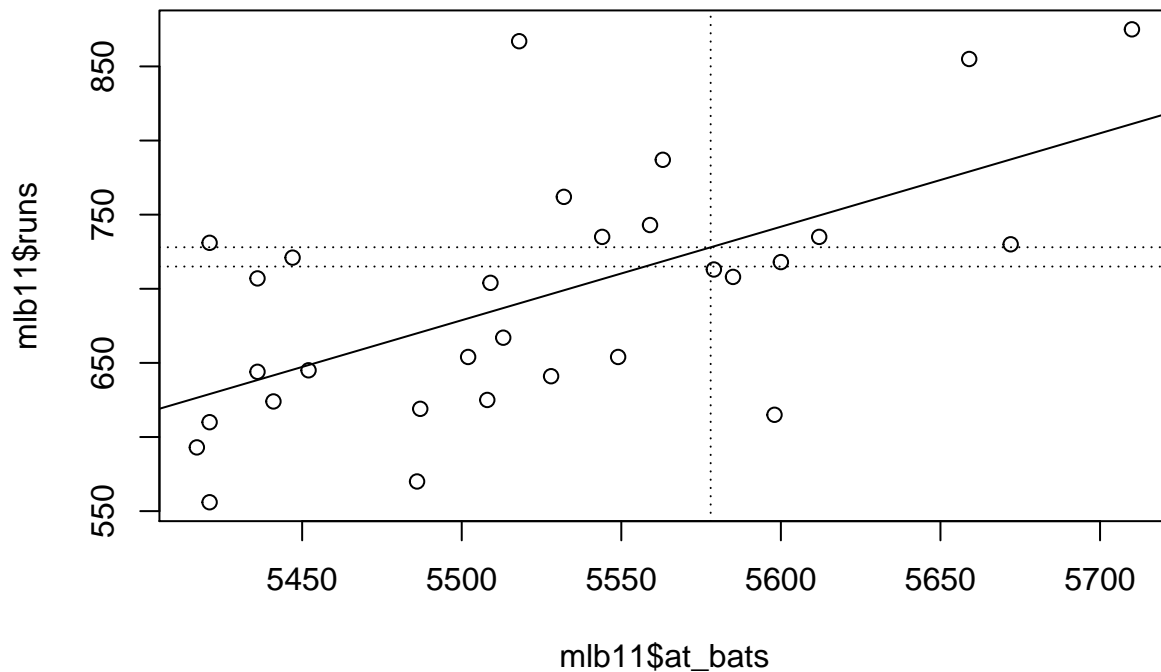
```
y <- -2789.2429+0.6305*5578
```

The runs which she will predict for 5578 is 728 from the least squares regression line. This is an over estimate from the actual value 713. There is a difference or the residual is 15.

## Model diagnostics

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

*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.

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

11
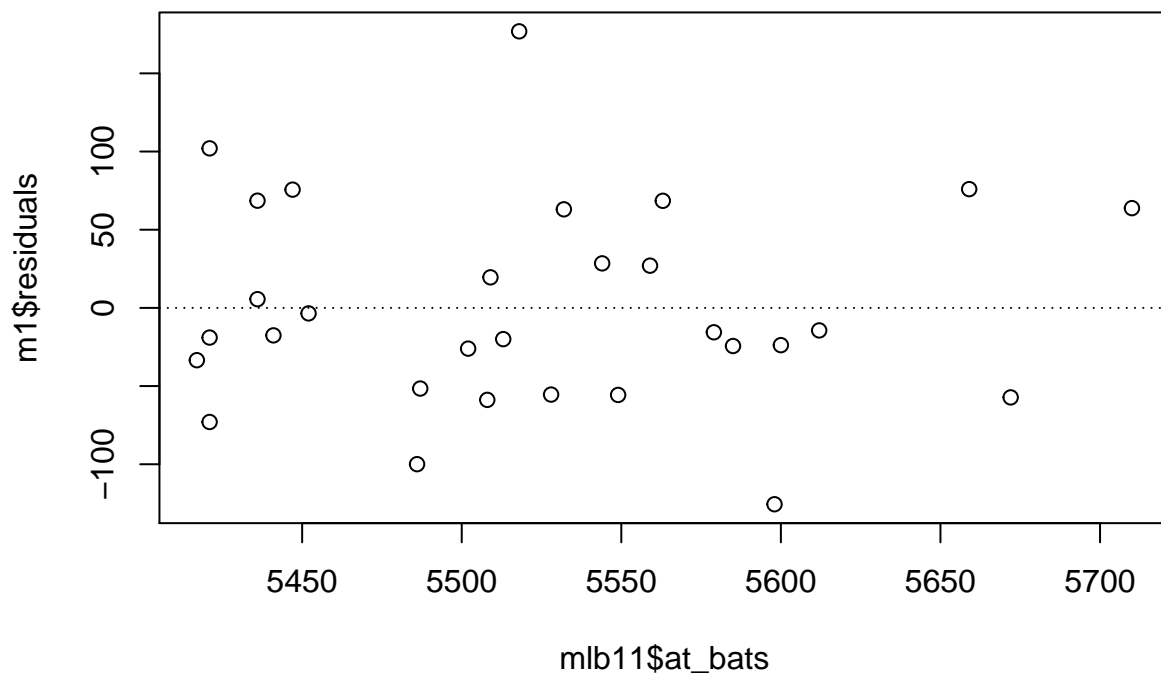
6. Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

On a high-level, the residual plot pattern appears to be non-linear. But we cannot derive a conclusion based on it.
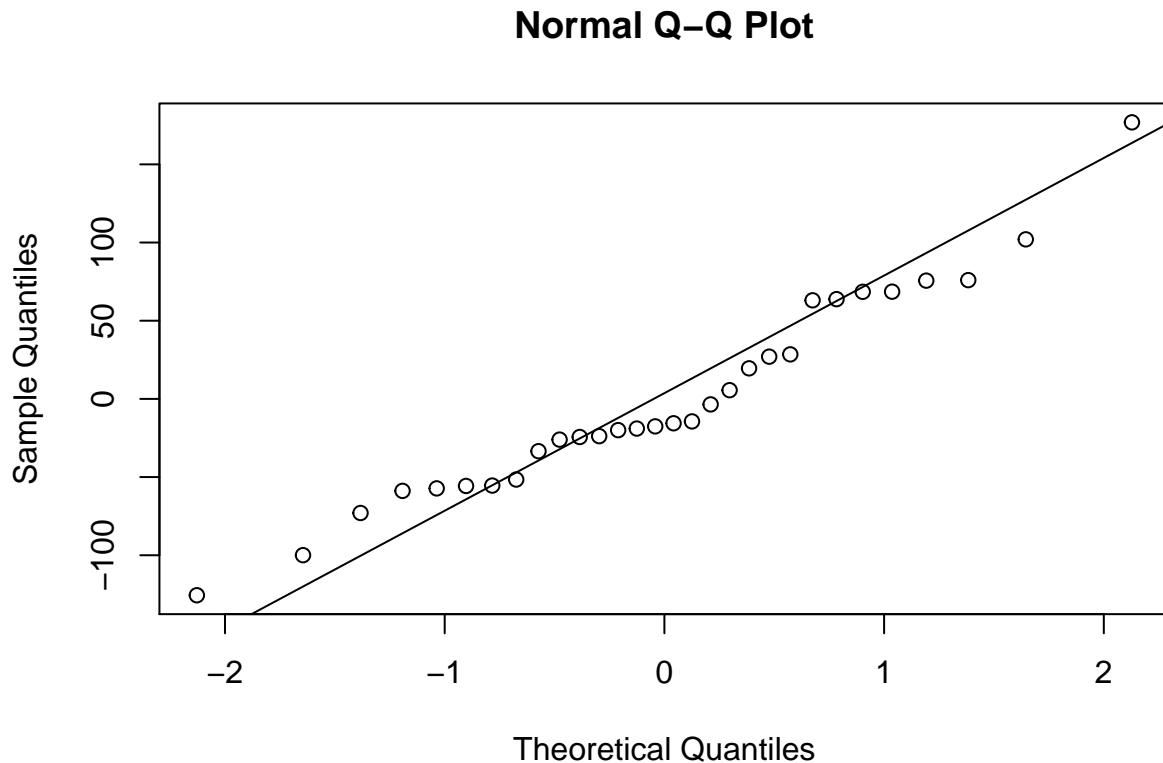
*Nearly normal residuals*: To check this condition, we can look at a histogram

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

## Histogram of m1$residuals
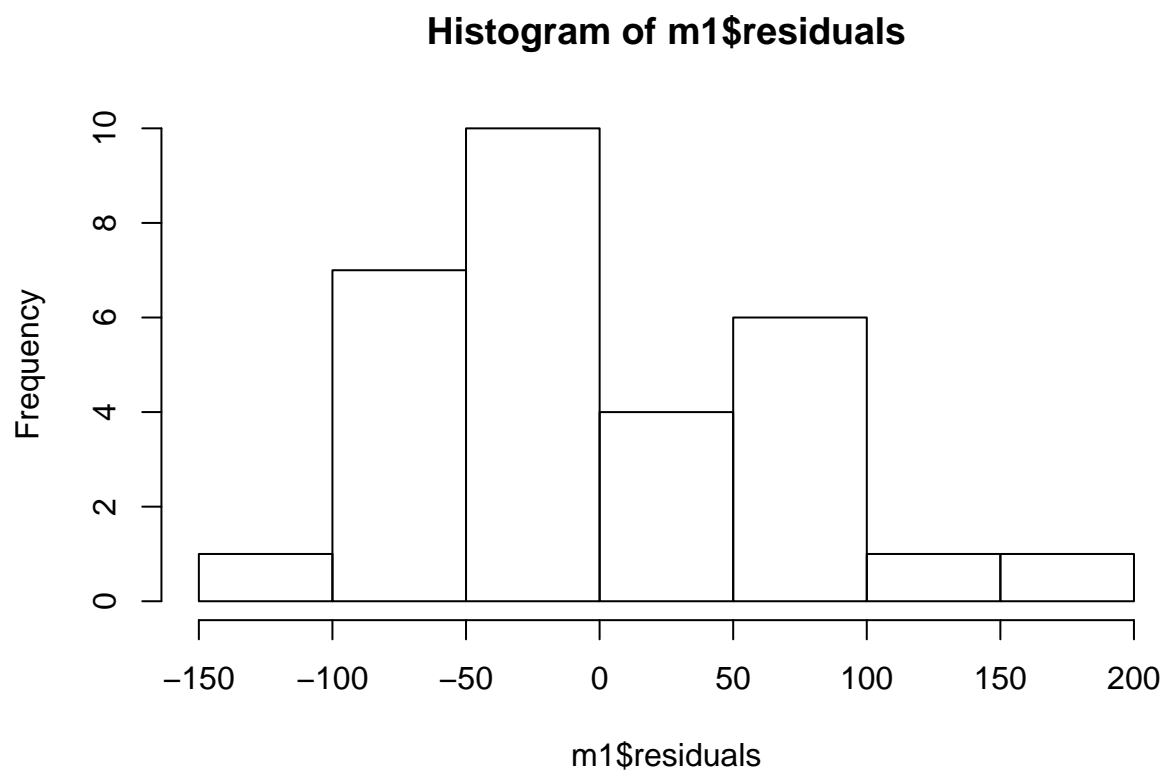


or a normal probability plot of the residuals.

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

## Normal Q-Q Plot



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

Yes. Based on the histogram and normal probability plot it looks like the residuals condition appear to be met.

*Constant variability*:

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

Yes. plot it looks like the constant variability condition to be met.

---

## 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?

```
ggplot(mlb11,aes(x=hits,y=runs)) + geom_point() + geom_smooth(method="lm")
```
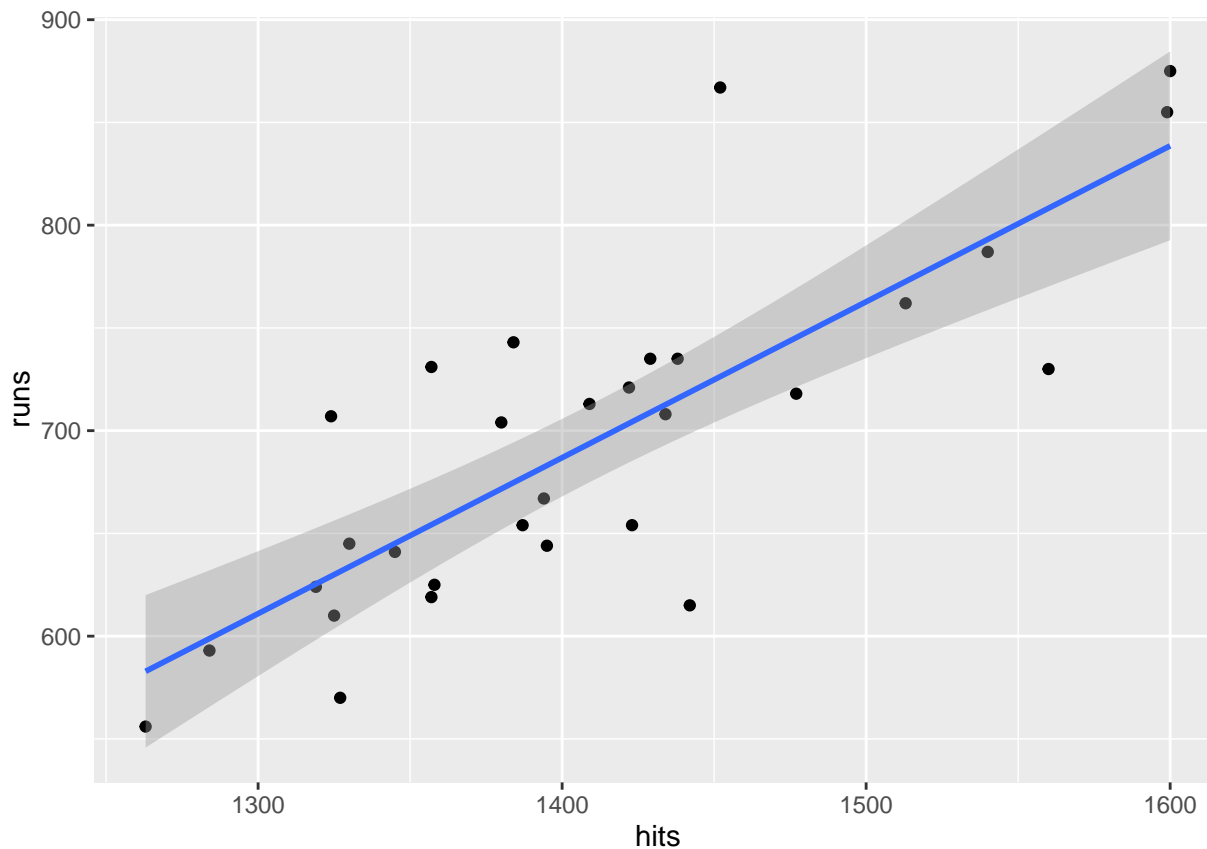
The plot shows the relationship between `hits` and `runs`. There is a strong positive correlation between the two variables. There are some outliers which influence the regression line.

- 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?

```
ggplot(mlb11,aes(x=hits,y=runs)) + geom_point() + geom_smooth(method="lm")
```
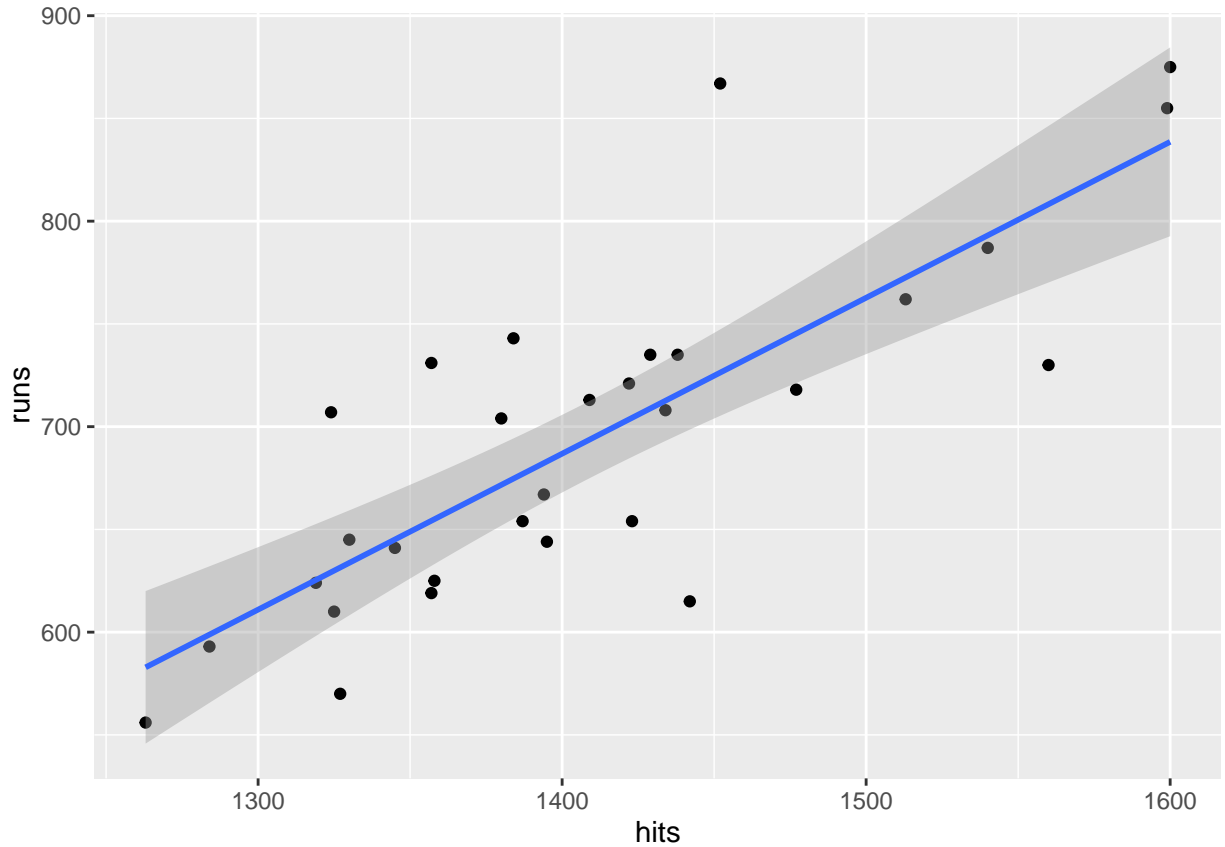
```r
m3 <- lm(runs~hits,mlb11)

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

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

```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
```

**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
```

Both the plots have positive correlation. The intercept is -375.56 and the slope is 0.7589

y = -375.56 + 0.7589x

$R^2$ for `at_bats` is 0.3729

$R^2$ for `hits` is 0.6419. This is higher than the previous model.

- 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).
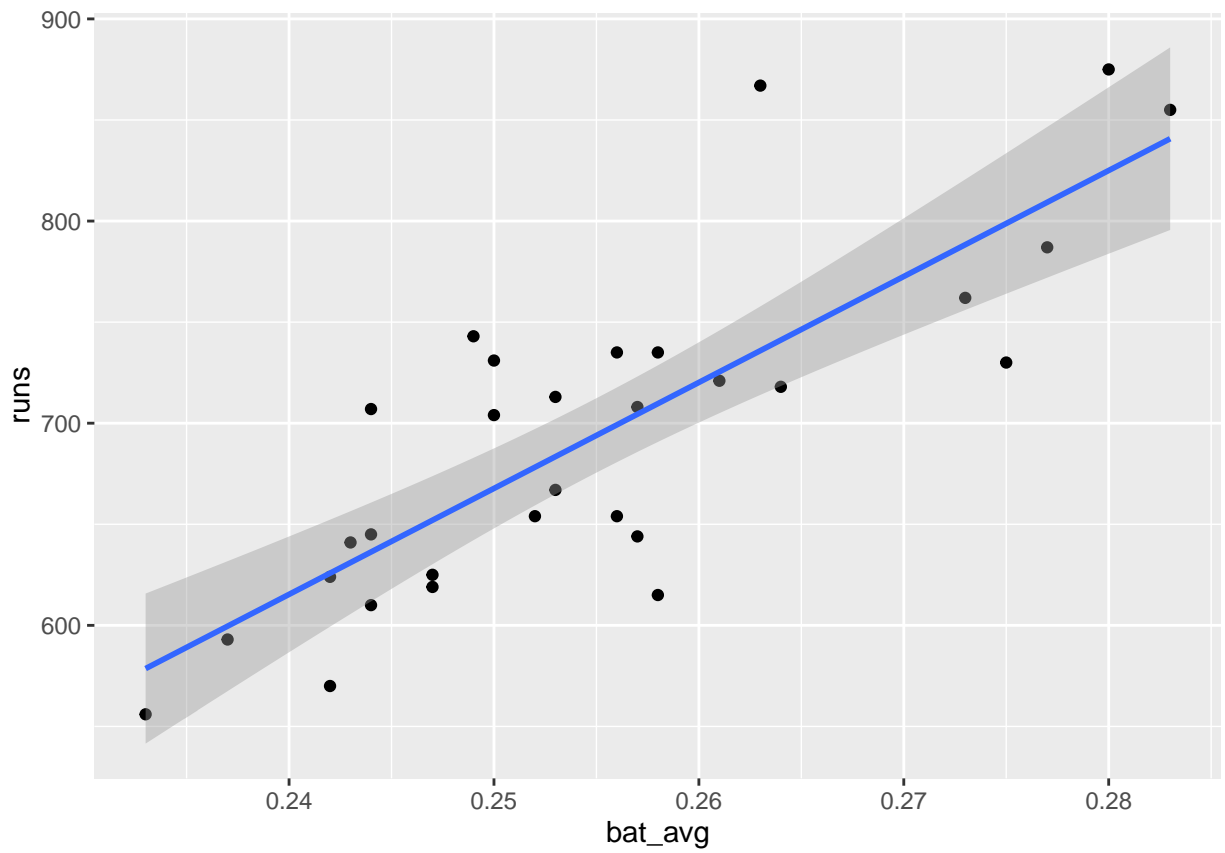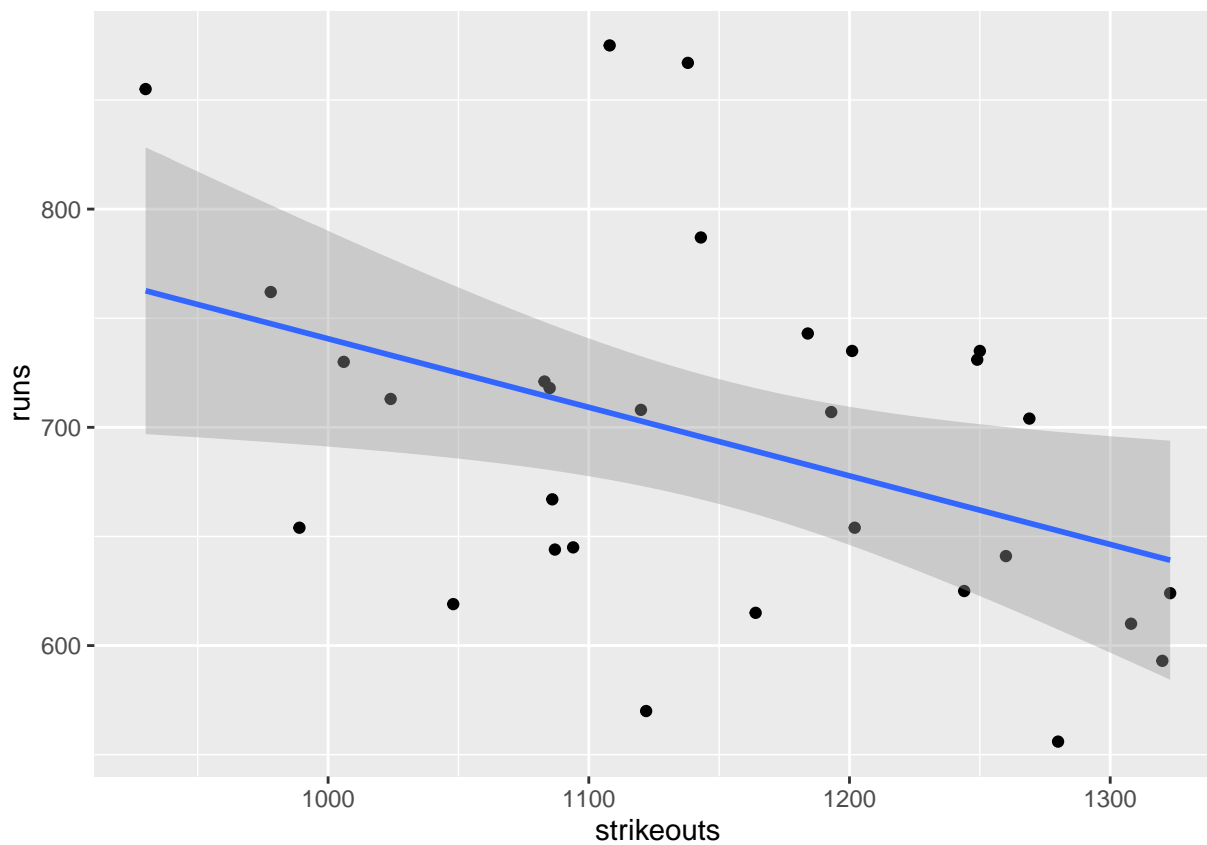
**cor**(mlb11$hits,mlb11$runs)

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

**ggplot**(mlb11,**aes**(x=homeruns,y=runs)) + **geom_point**() + **geom_smooth**(method="lm")
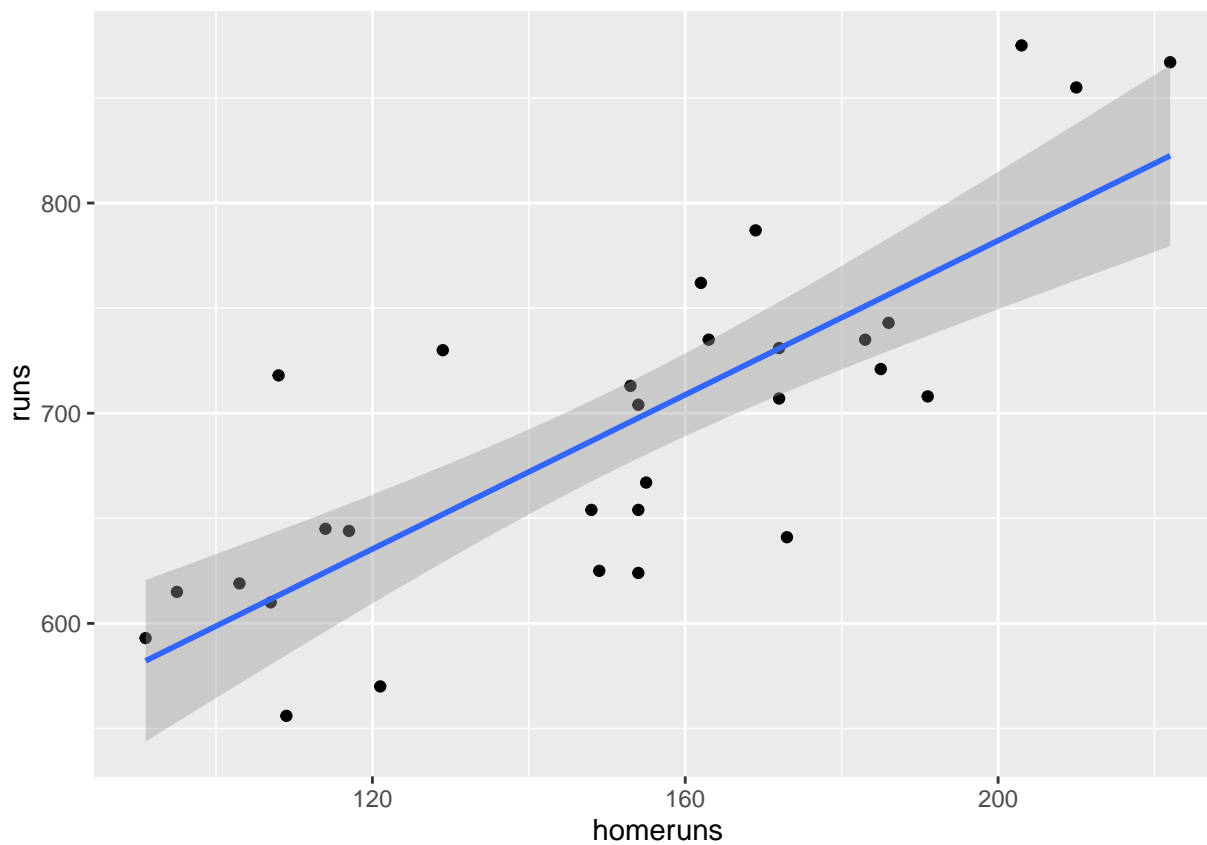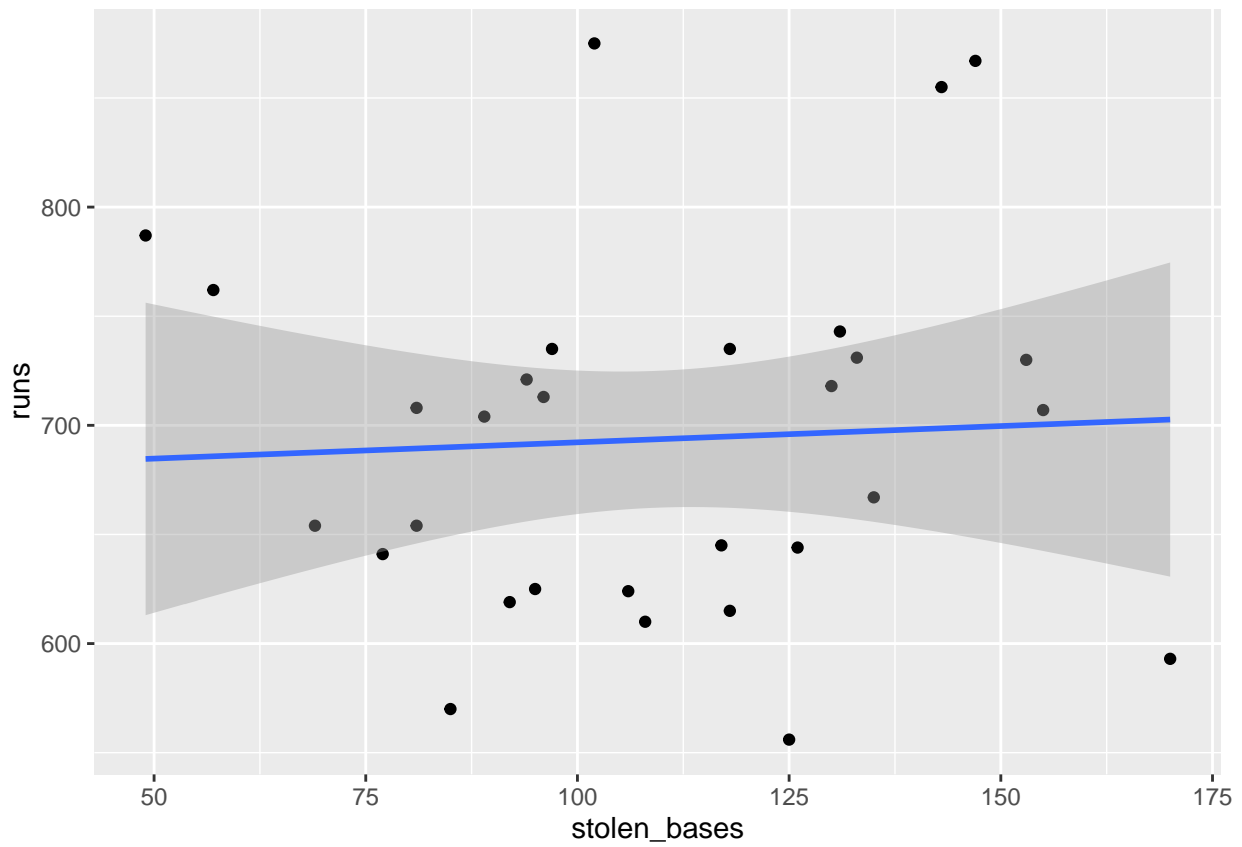
```
ggplot(mlb11,aes(x=bat_avg,y=runs)) + geom_point() + geom_smooth(method="lm")
```
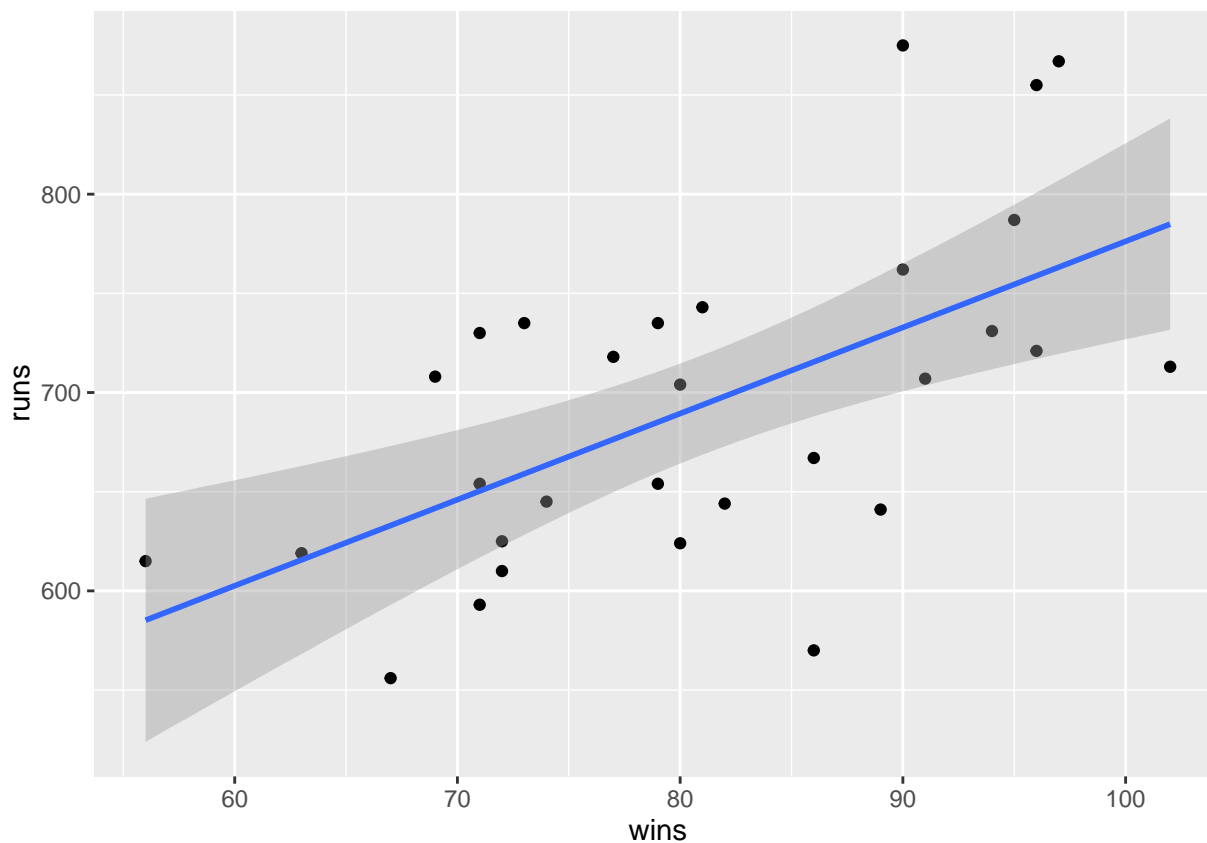
```
ggplot(mlb11,aes(x=strikeouts,y=runs)) + geom_point() + geom_smooth(method="lm")
```

```
ggplot(mlb11,aes(x=stolen_bases,y=runs)) + geom_point() + geom_smooth(method="lm")
```

```
ggplot(mlb11,aes(x=wins,y=runs)) + geom_point() + geom_smooth(method="lm")
```

The best variable that predicts runs is `bat_avg`. It is followed by `hits`.

For `bat_avg`, below are the summary statistics

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

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

```
m4 <- lm(runs~bat_avg,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
```

- 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?
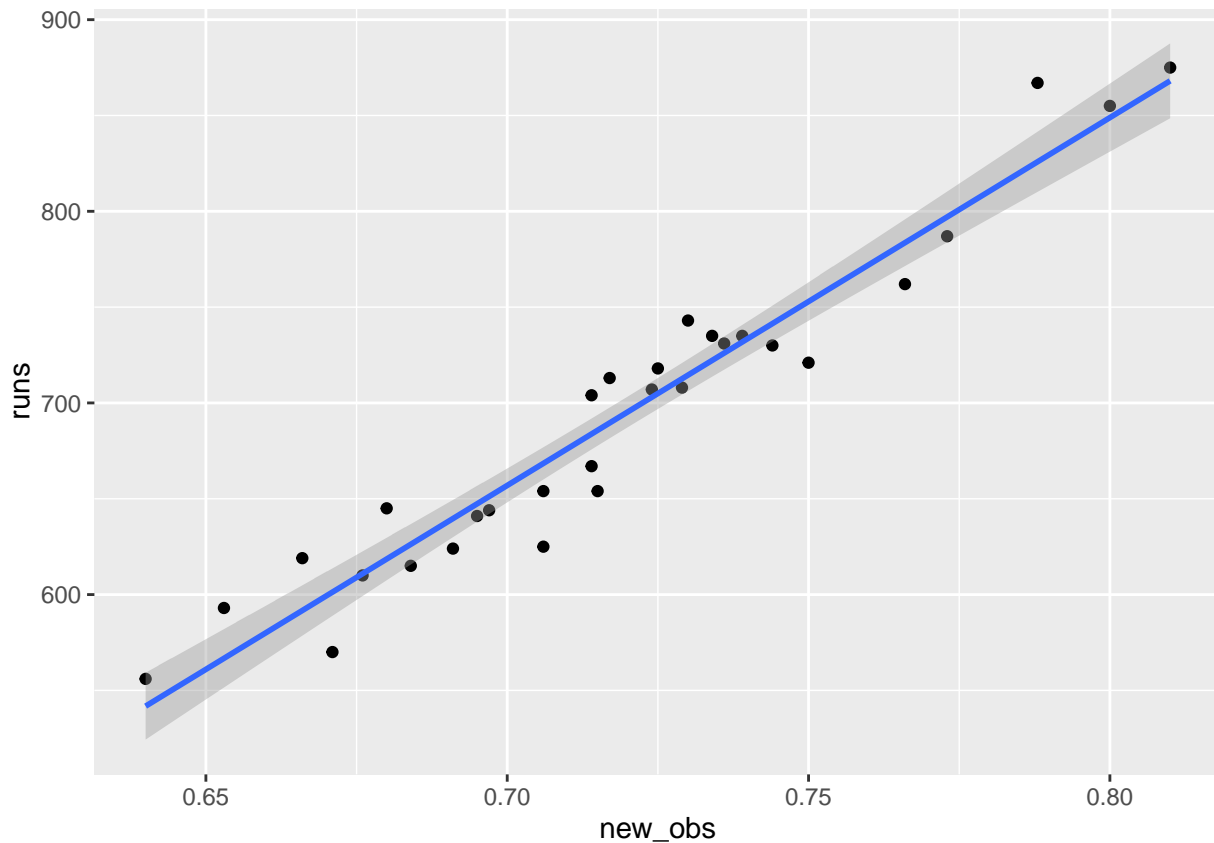
```
ggplot(mlb11,aes(x=new_onbase,y=runs)) + geom_point() + geom_smooth(method="lm")
```
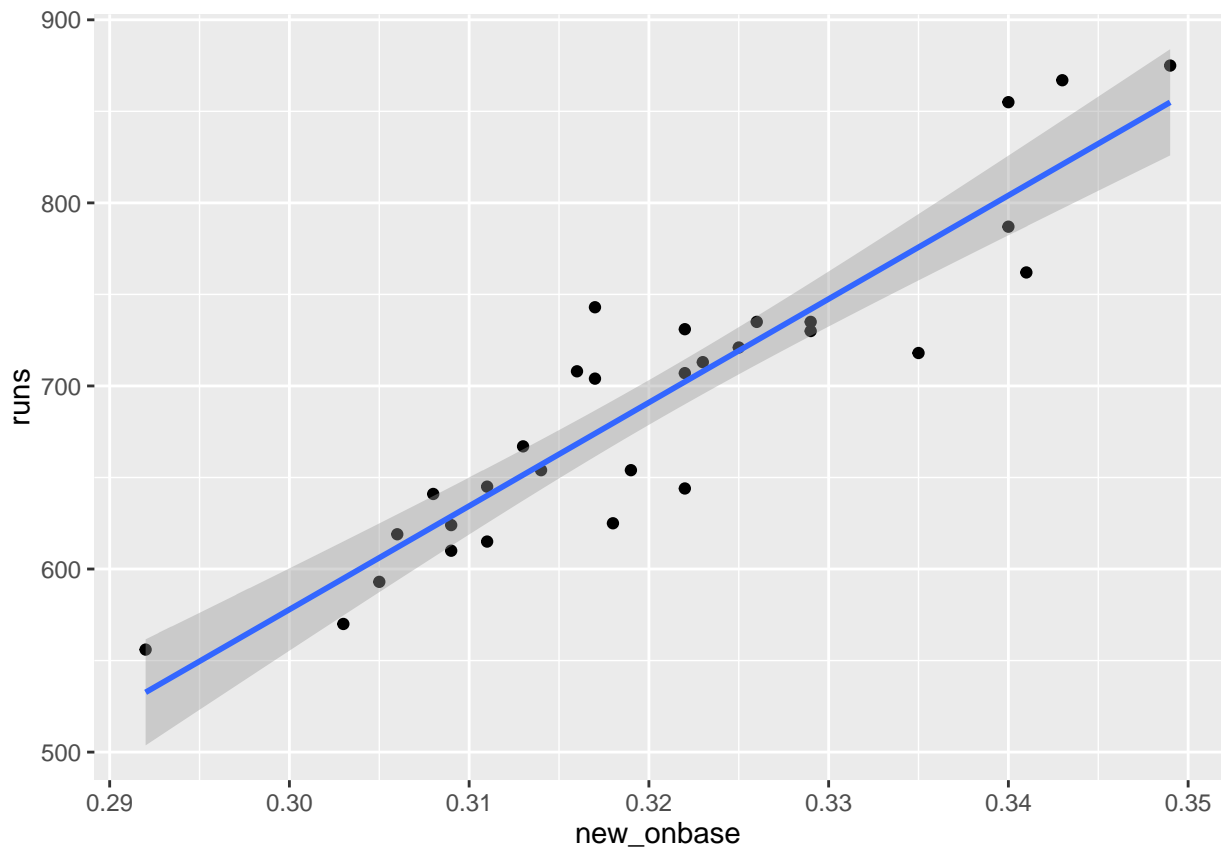


```
ggplot(mlb11,aes(x=new_slug,y=runs)) + geom_point() + geom_smooth(method="lm")
```
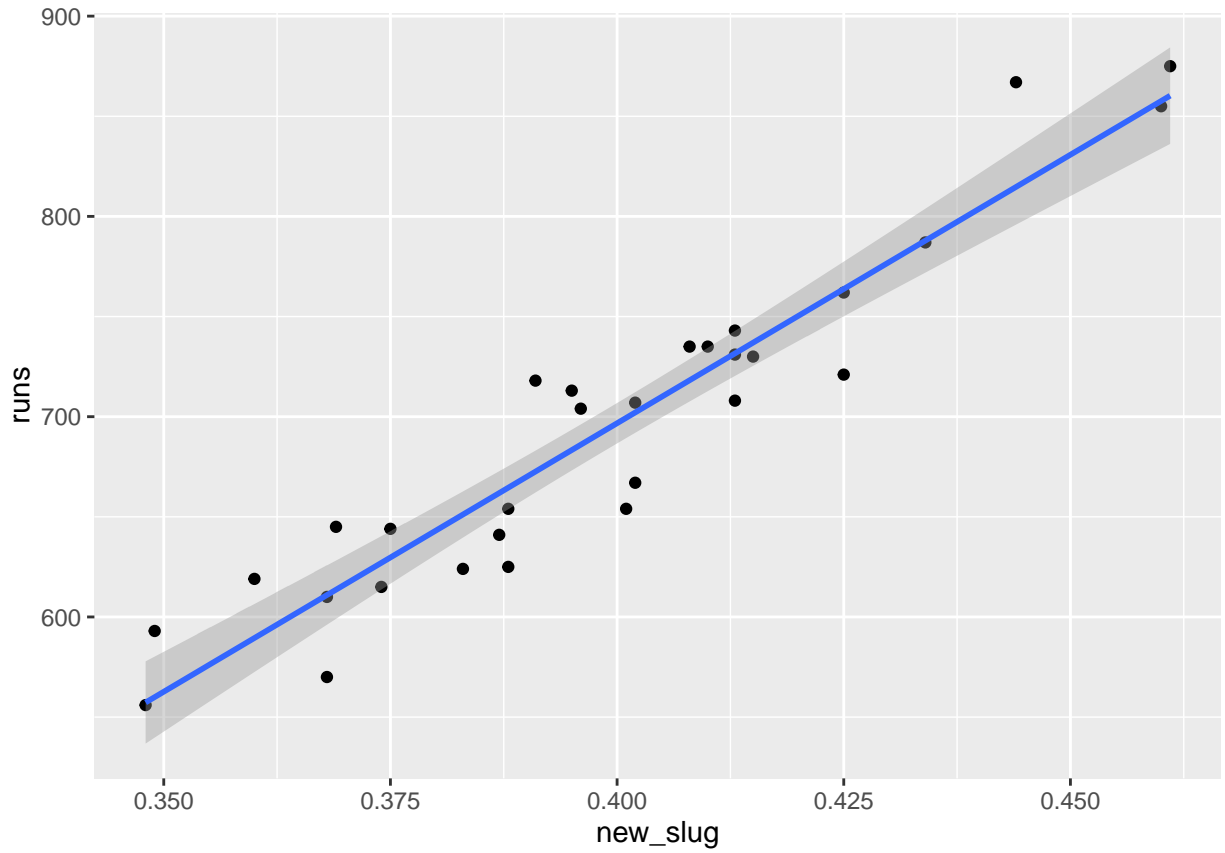
```
ggplot(mlb11,aes(x=new_obs,y=runs)) + geom_point() + geom_smooth(method="lm")
```

The above shown new variables shows a good prediction of `runs` than the traditional variables. These are more effective predictors. Below are summary statistics.

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

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

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

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

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

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

```
lm(runs~new_onbase,mlb11)
```

```
##
## Call:
## lm(formula = runs ~ new_onbase, data = mlb11)
##
## Coefficients:
## (Intercept)    new_onbase
##       -1118          5654
```

```r
lm(runs~new_slug,mlb11)
```

```
##
## Call:
## lm(formula = runs ~ new_slug, data = mlb11)
##
## Coefficients:
## (Intercept)      new_slug
##      -375.8        2681.3
```

```r
m5 <- summary(lm(runs~new_obs,mlb11))
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
```

Of all the variables analyzed, `new_obs` is the best predictor of runs. It has an $R^2$ of 93.49.
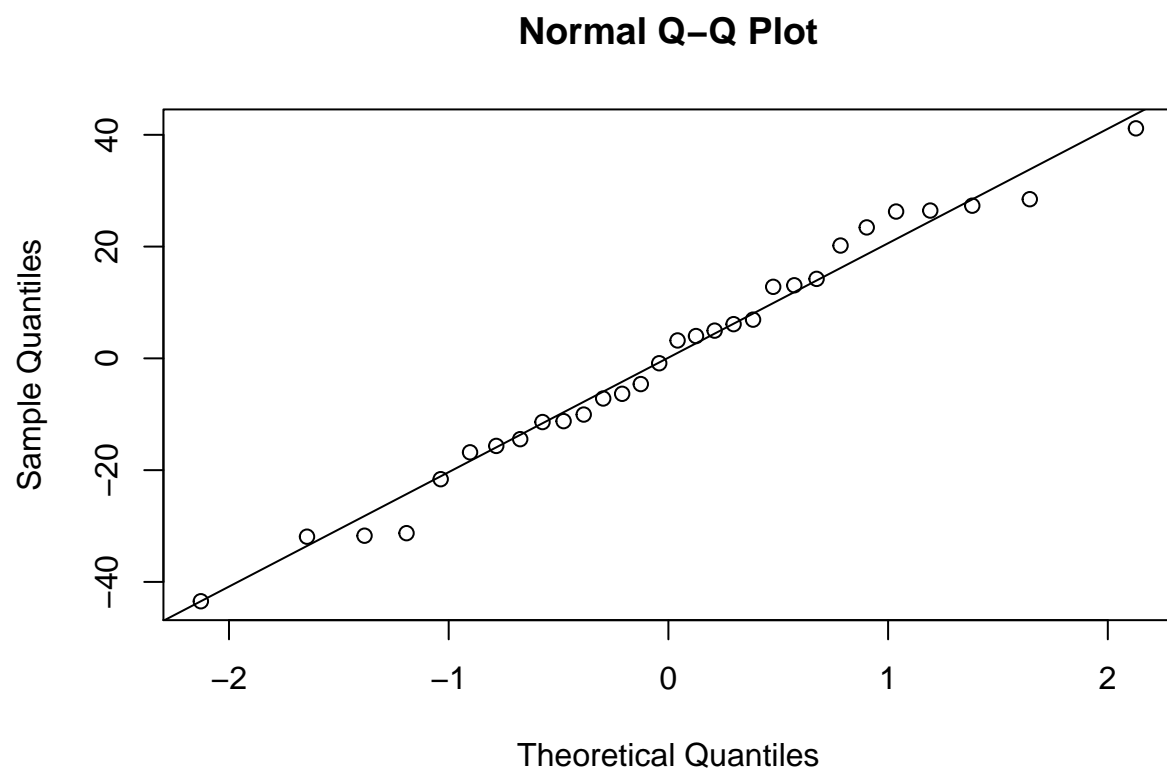
- Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

```r
hist(mlb11$new_obs)
```

**Histogram of mlb11$new_obs**
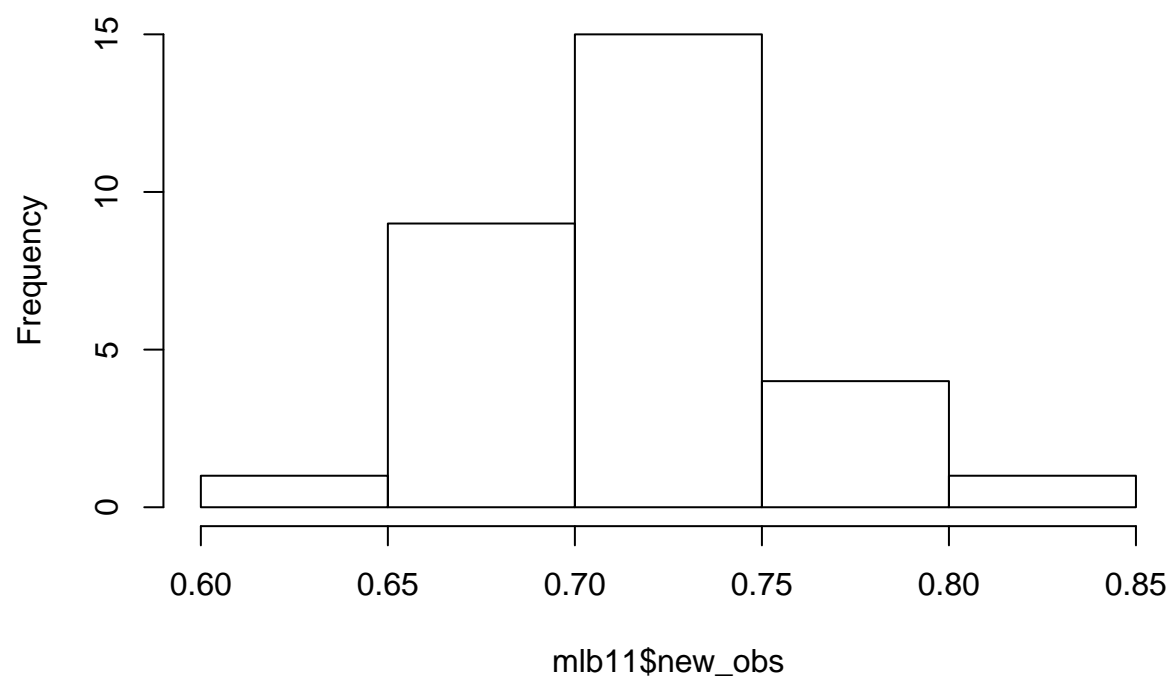


```
qqnorm(m5$residuals)
qqline(m5$residuals)
```

## Normal Q–Q Plot



As the variable is normal and has constant variablilty. And plots show that the variable `new_obs` is the best predictor of runs.