

Introduction to linear regression

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, better predict 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")
```

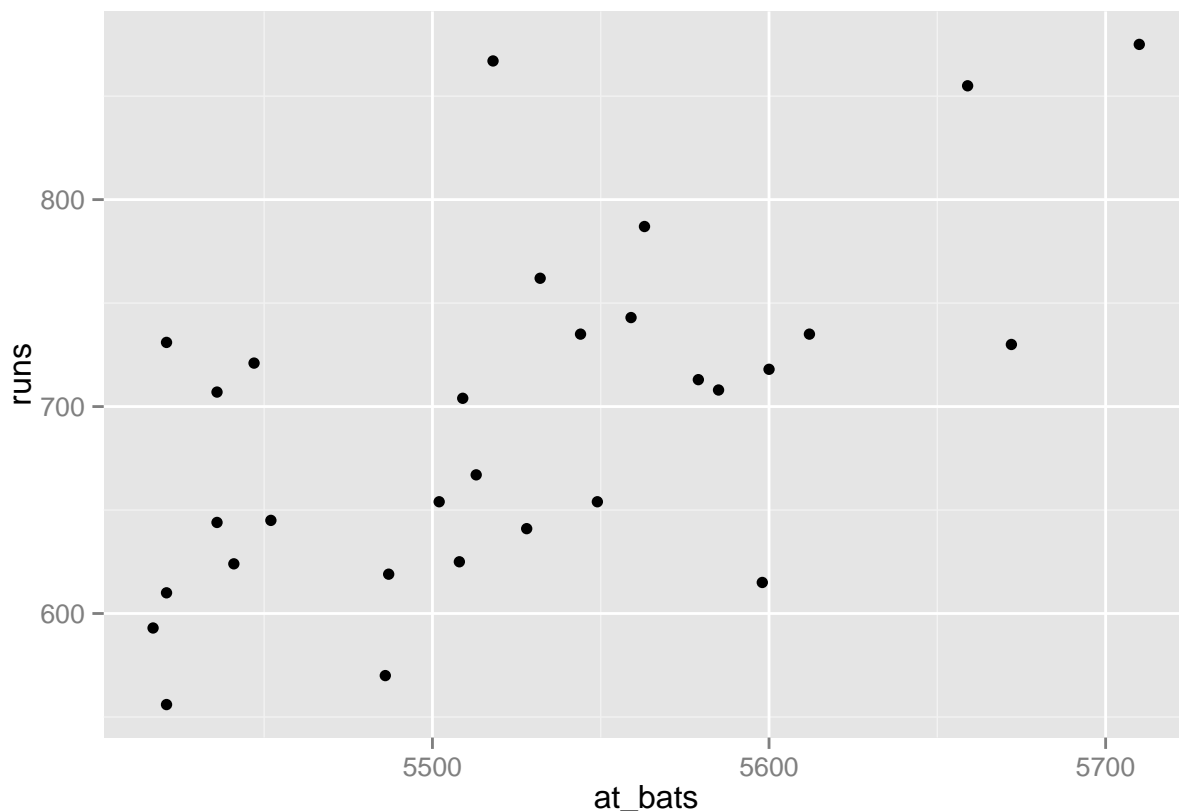
In addition to runs scored, there are seven traditionally used variables in the data set: at-bats, hits, home runs, batting average, strikeouts, stolen bases, and wins. There are also three newer variables: on-base percentage, slugging percentage, and on-base plus slugging. For the first portion of the 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?

I would use a scatter plot to make an initial comparison like this.

```
library(ggplot2)

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



A linear model would probably be the best to use for these two variables, but by eyeballing it the correlation doesn't look that high. At_bats might not be the best variable to use to predict 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

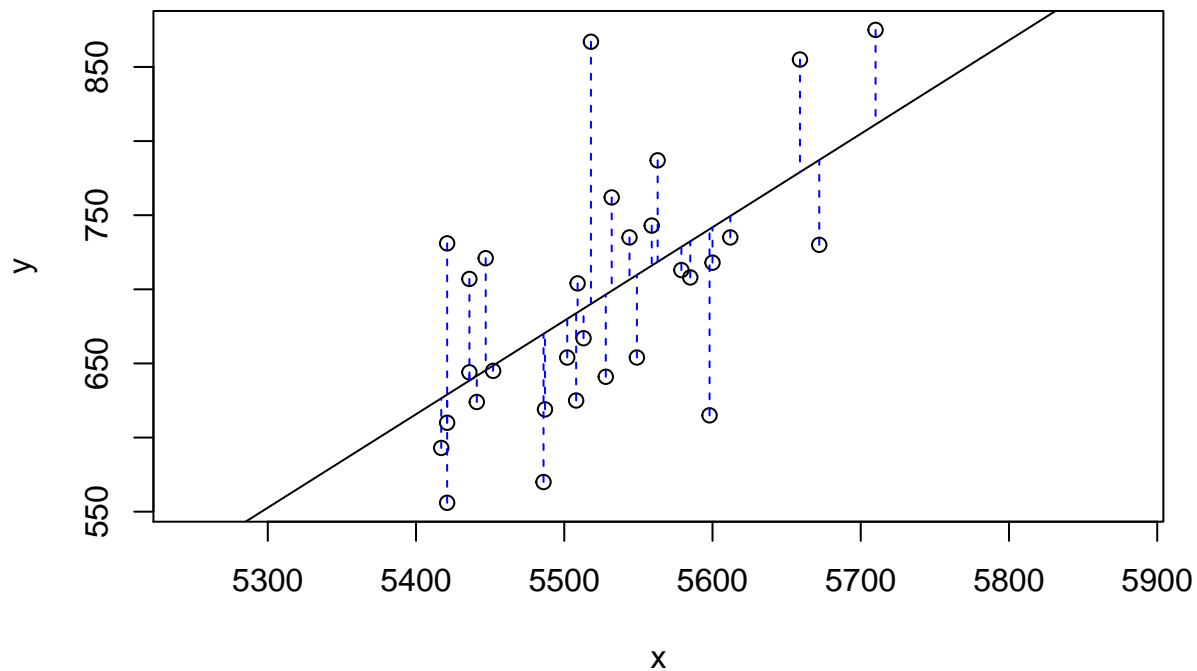
Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as `runs` and `at_bats` above.

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.

I would say it's a weak relationship (the correlation is only around 0.6). It's positive, and there may be one or two outliers. Since the relationship is weak however, it's hard to say whether these truly are outliers or if it's just not a strong relationship.

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)
```



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

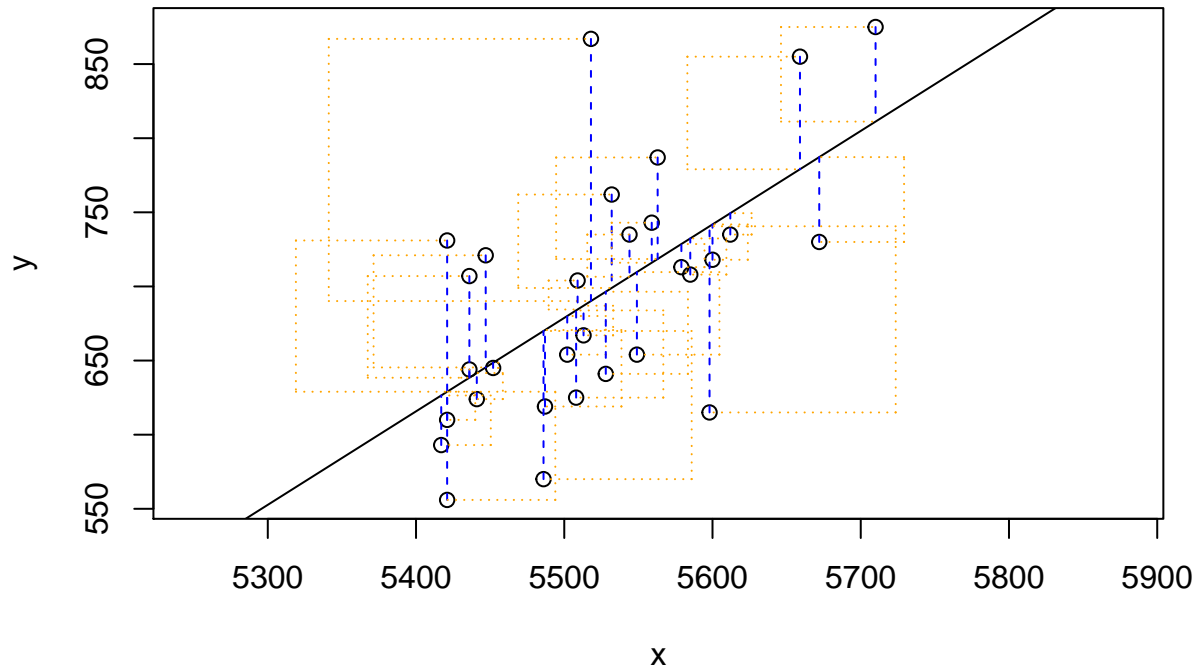
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)          x
## -2789.2429      0.6305
##
## Sum of Squares: 123721.9
```

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.

- 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?

The smallest I got by guessing was 124551.2.

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 $y \sim x$. 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.

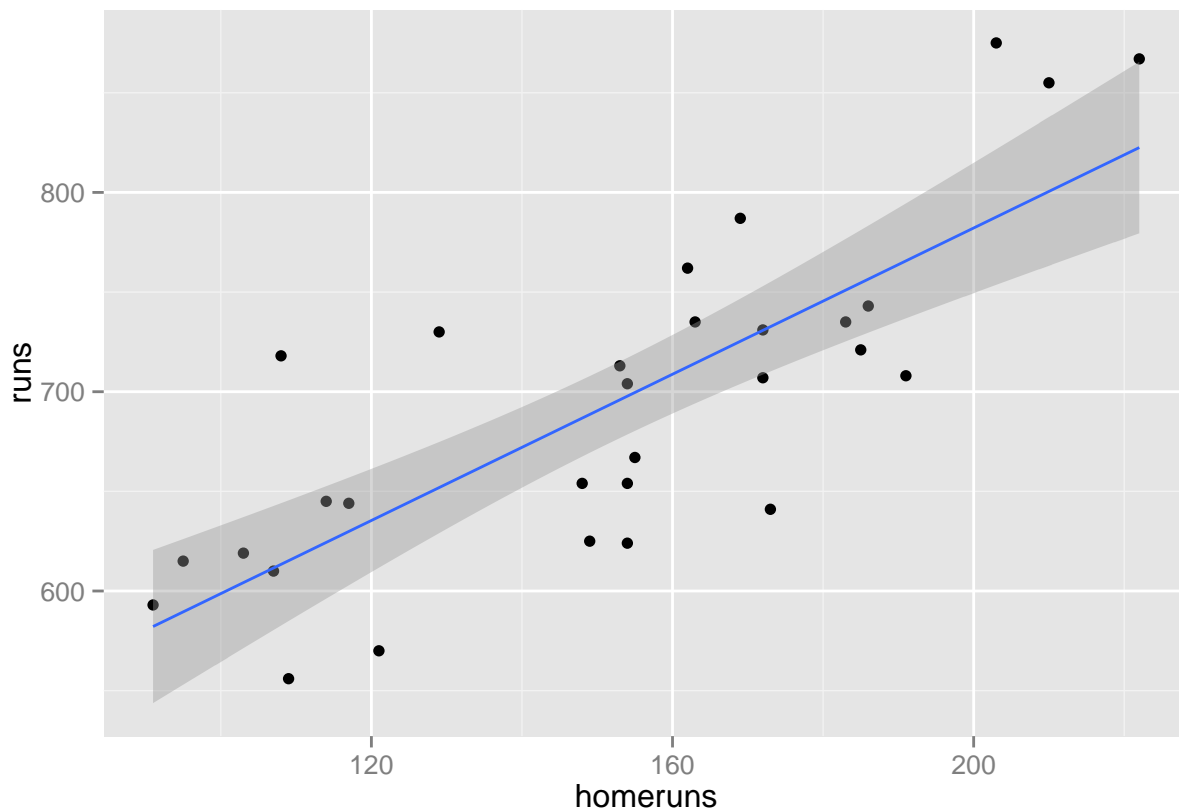
4. Fit a new model that uses `homeruns` to predict `runs`. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between success of a team and its home runs?

```
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
```

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



The equation of this line is:

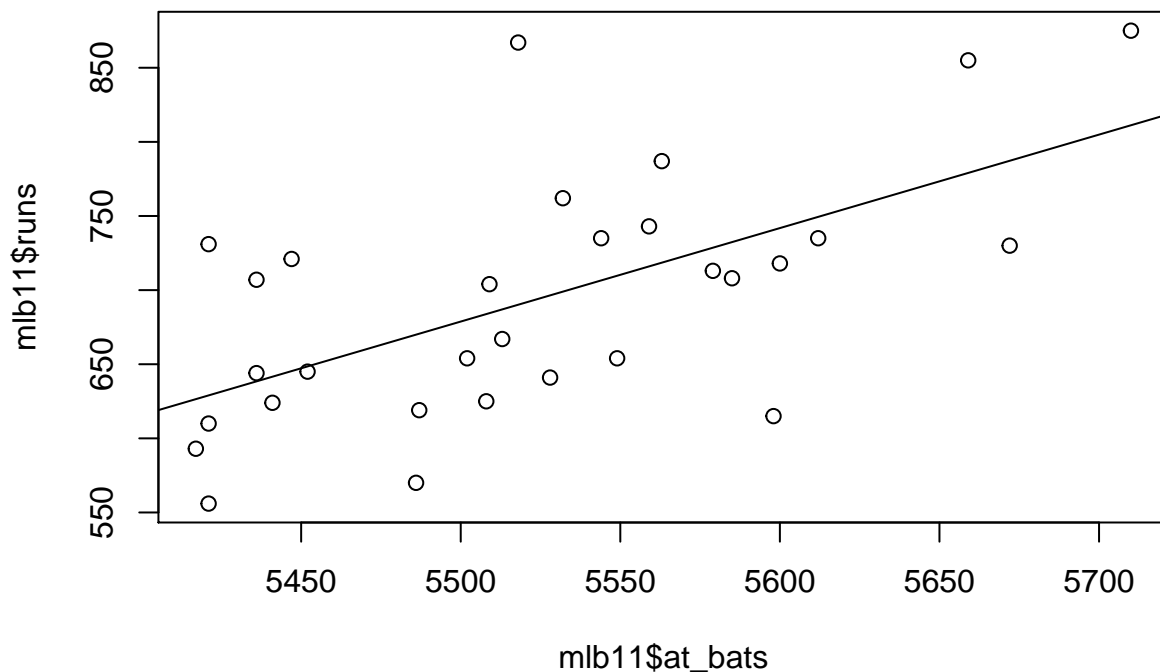
$$\hat{y} = 415.2389 + 1.8345 * \text{homeruns}$$

The slope tells us the scale of the relationship. If, for example, a change of 1 in an x variable leads to a change of 1000 in y , this would be reflected in the slope.

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?

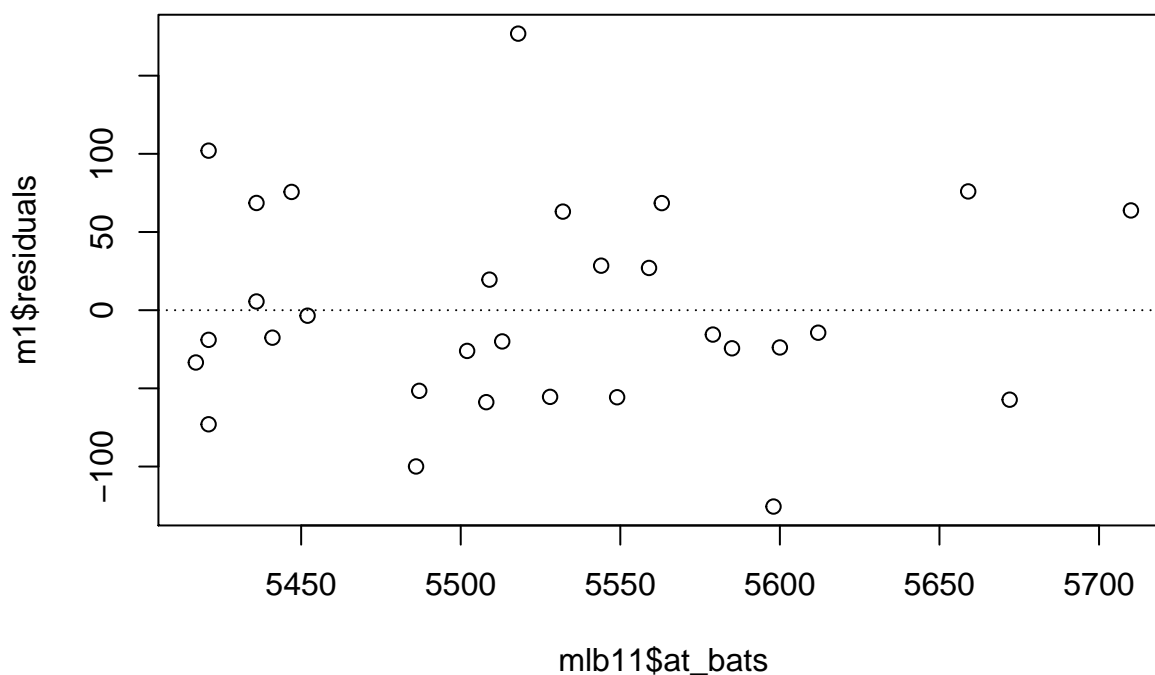
Eyeballing from the graph, one would expect about 740 runs for 5578 at bats. This would be a slight overestimate, there is a datapoint that is a data- point at about 725. The residual of this particular datapoint doesn't appear to be too large.

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
```

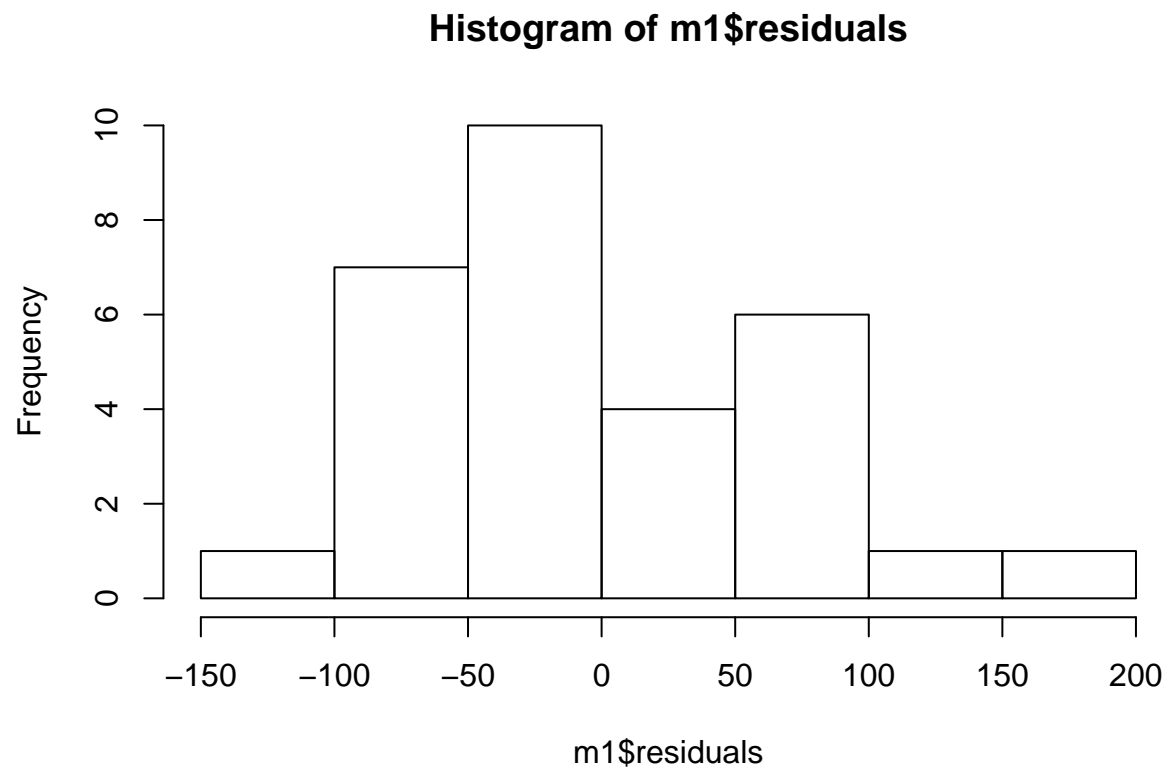


6. Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between runs and at-bats?

There appears to be no pattern in the residuals plot, indicating that there is a clear linear trend in this relationship, and any other variation is due to randomness.

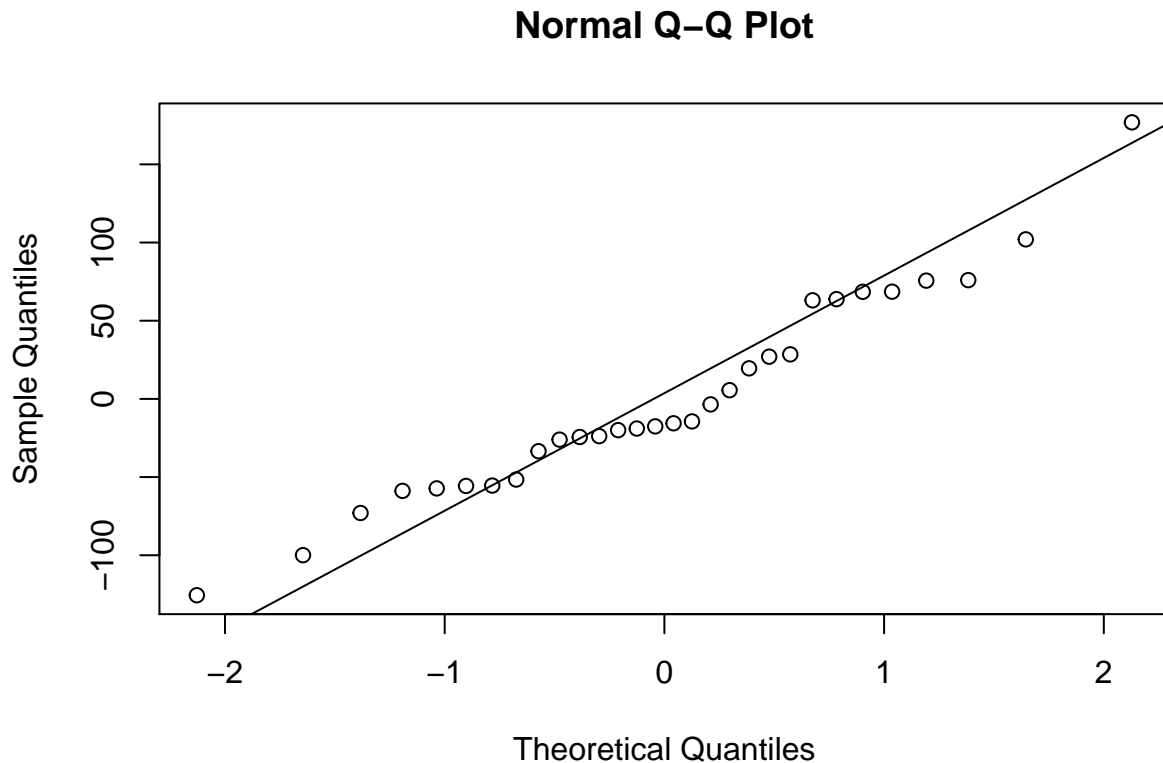
Nearly normal residuals: To check this condition, we can look at a histogram


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



or a normal probability plot of the residuals.

```
qqnorm(m1$residuals)
qqline(m1$residuals) # 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 qq plot appears to be slightly less convincing, but there doesn't appear to be a clear underlying pattern in the residuals scatterplot.

Constant variability:

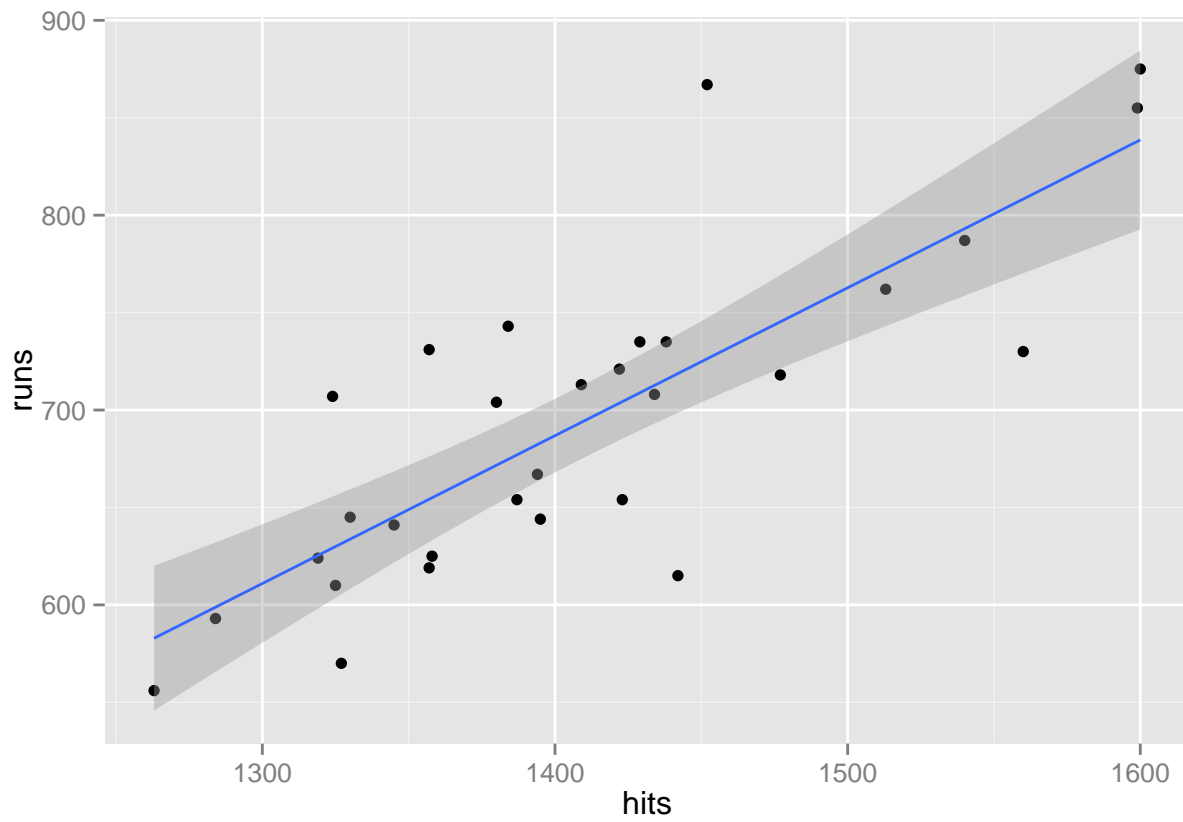
8. Based on the plot in (1), does the constant variability condition appear to be met?

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?

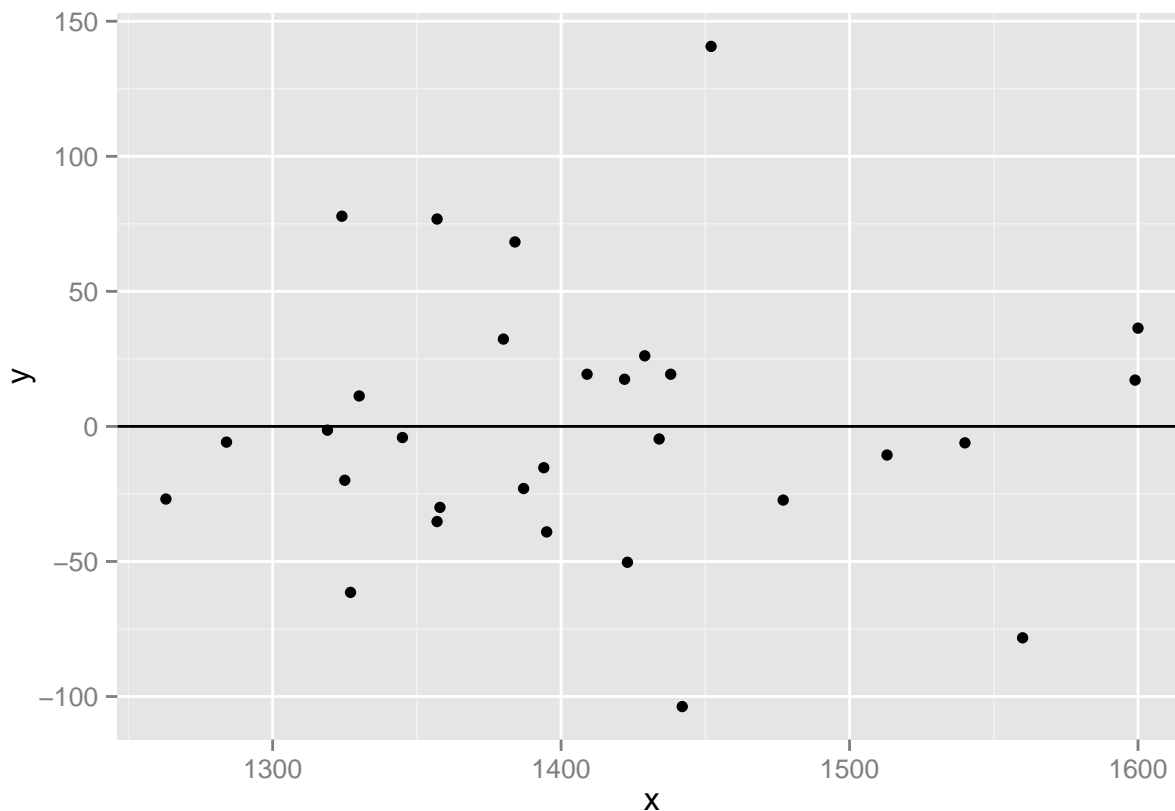
Lets see if the number of hits predicts the number of runs

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



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

ggplot(data.frame(x=mlb11$hits,y=m3$residuals),aes(x=x, y=y)) +
  geom_point() + geom_hline(x=0)
```



At a glance, there appears to be a linear relationship. Looking at the graph of the residuals, there appears to be no hidden trend.

- 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?

```
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
```

```
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
```

The relationship with hits seems to be slightly more significant than the relationship with at_bats, based on the slightly higher R-squared.

- 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).

```
mhits <- lm(runs~hits,data=mlb11)
mbatavg <- lm(runs~bat_avg,data=mlb11)
mstrikeouts <- lm(runs~strikeouts,data=mlb11)
mstolenbases <- lm(runs~stolen_bases,data=mlb11)
mwins <- lm(runs~wins,data=mlb11)
```

```
summary(mhits)$r.squared
```

```
## [1] 0.6419388
```

```
summary(mbatavg)$r.squared
```

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

```
summary(mstrikeouts)$r.squared
```

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

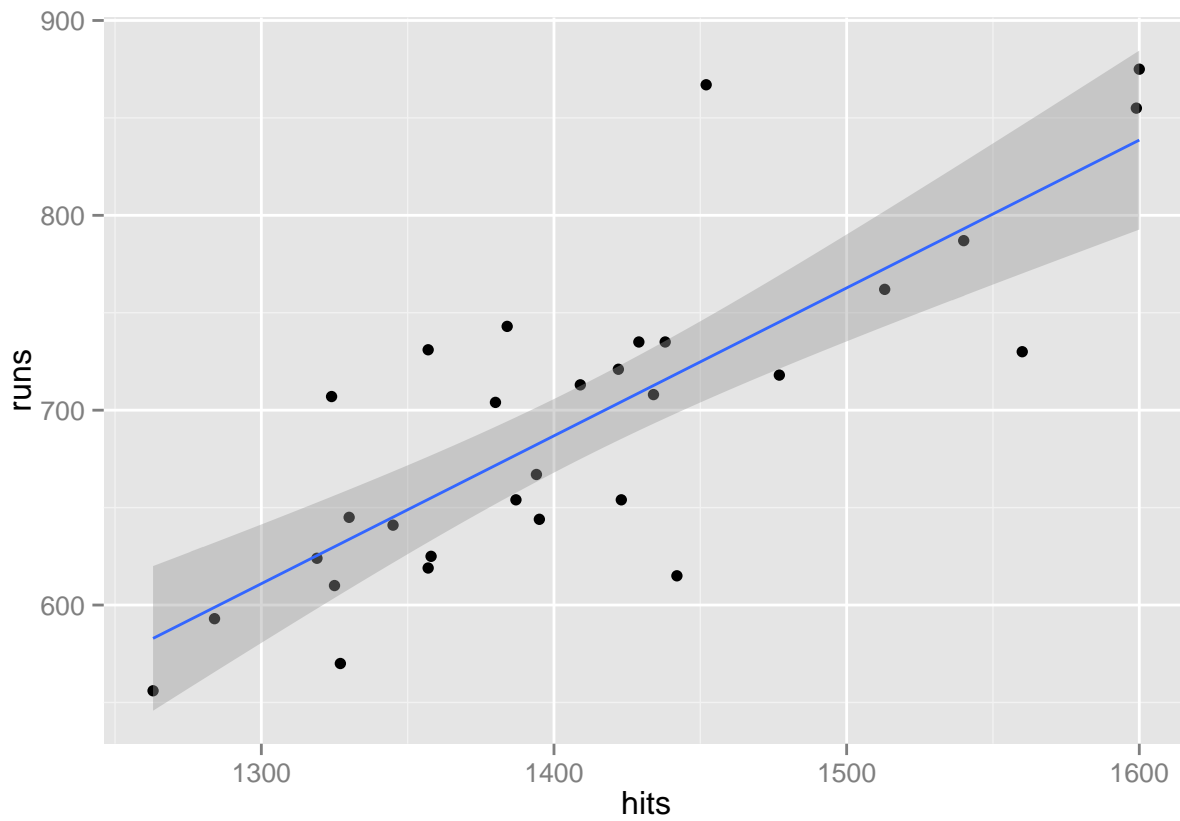
```
summary(mstolenbases)$r.squared
```

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

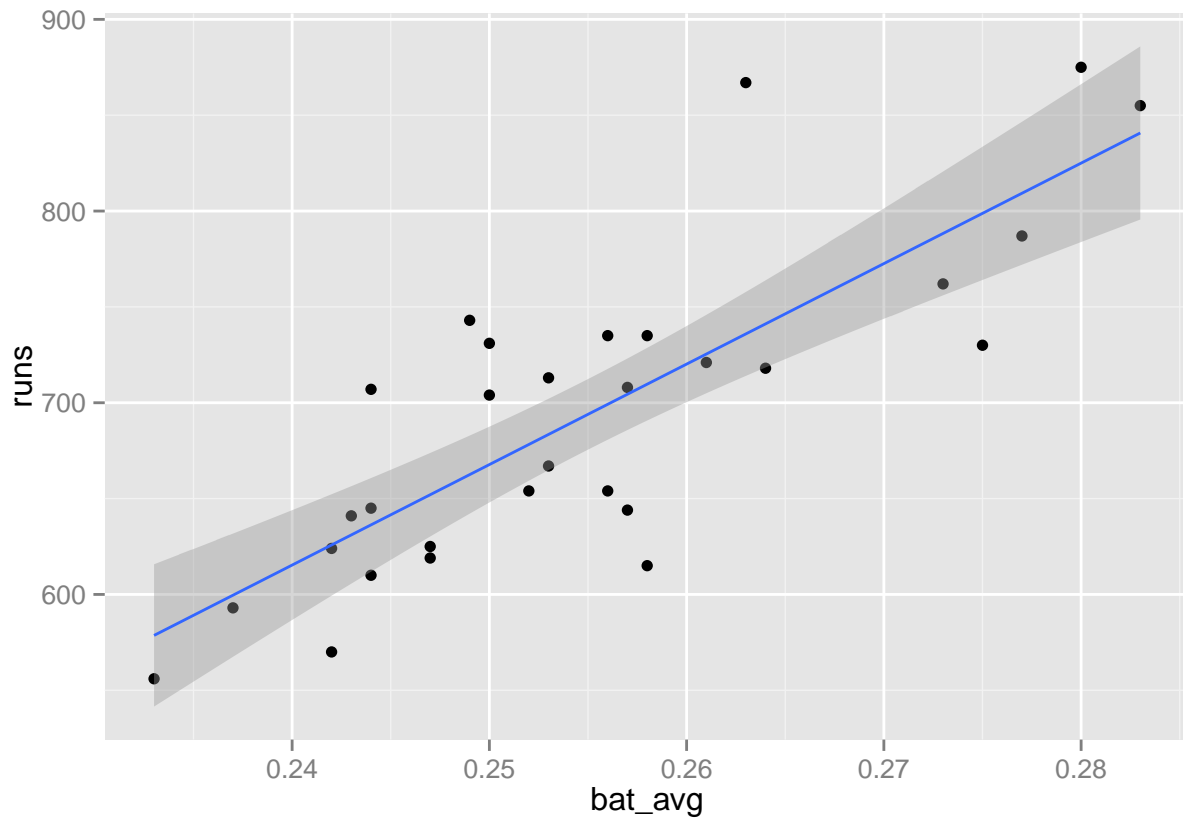
```
summary(mwins)$r.squared
```

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

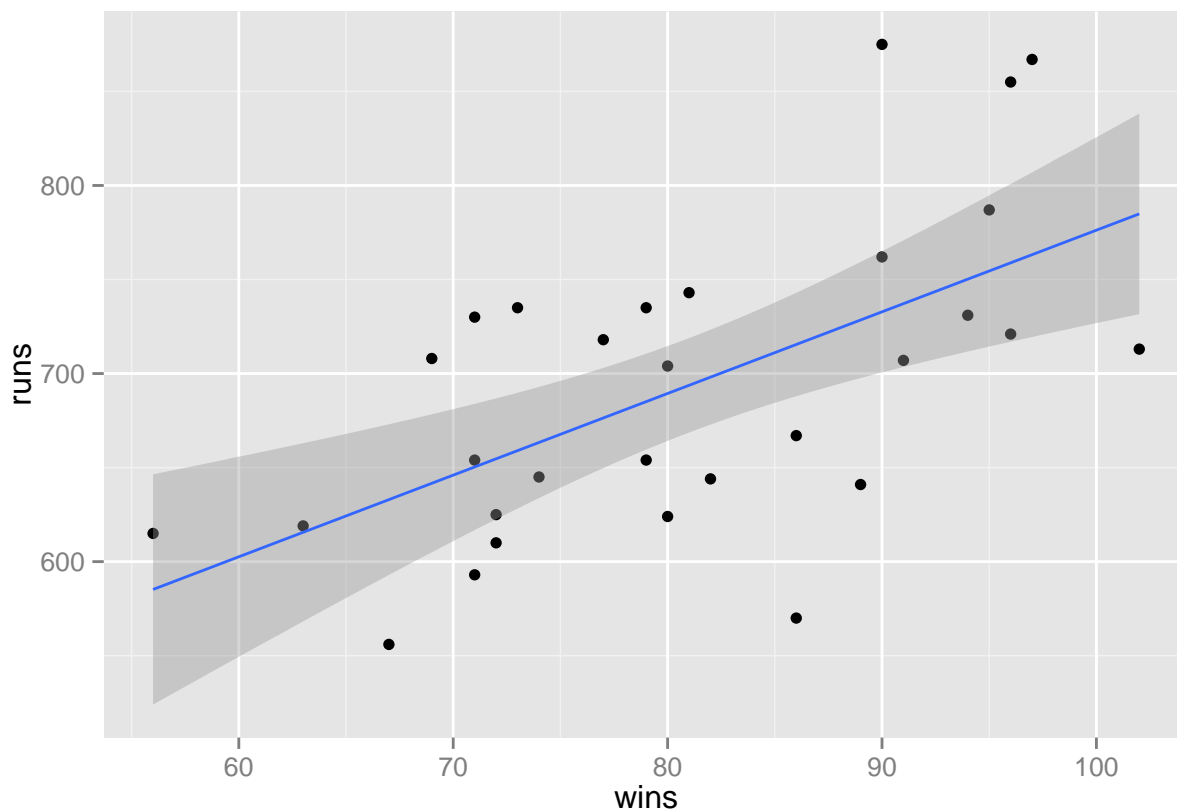
```
ggplot(mlb11, aes(x=hits,y=runs)) + geom_point() + stat_smooth(method="lm")
```



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



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



Three of the best variables are graphed. Hits and batting average are pretty good at predicting runs, and wins are a bit less predictive. There is almost no relationship between stolen bases and strikeouts and runs (although, to be expected, there is a negative relationship between strikeouts and runs).

- Now examine the three newer variables. These are the statistics used by the author of *Moneyball* to predict a team's 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?

```
mnewonbase <- lm(runs~new_onbase, data=mlb11)
mnewslug <- lm(runs~new_slug, data=mlb11)
mnewobs <- lm(runs~new_obs, data=mlb11)
```

```
summary(mnewonbase)$r.squared
```

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

```
summary(mnewslug)$r.squared
```

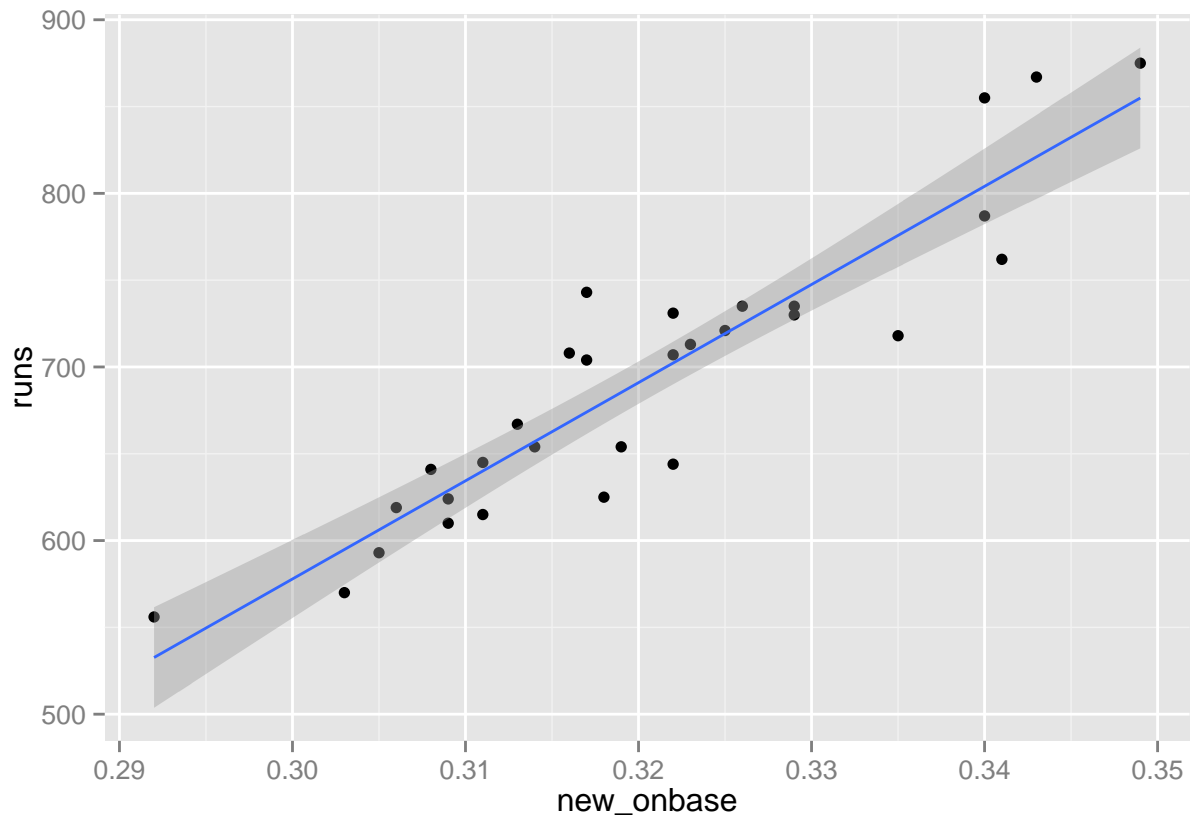
```
## [1] 0.8968704
```

```
summary(mnewobs)$r.squared
```

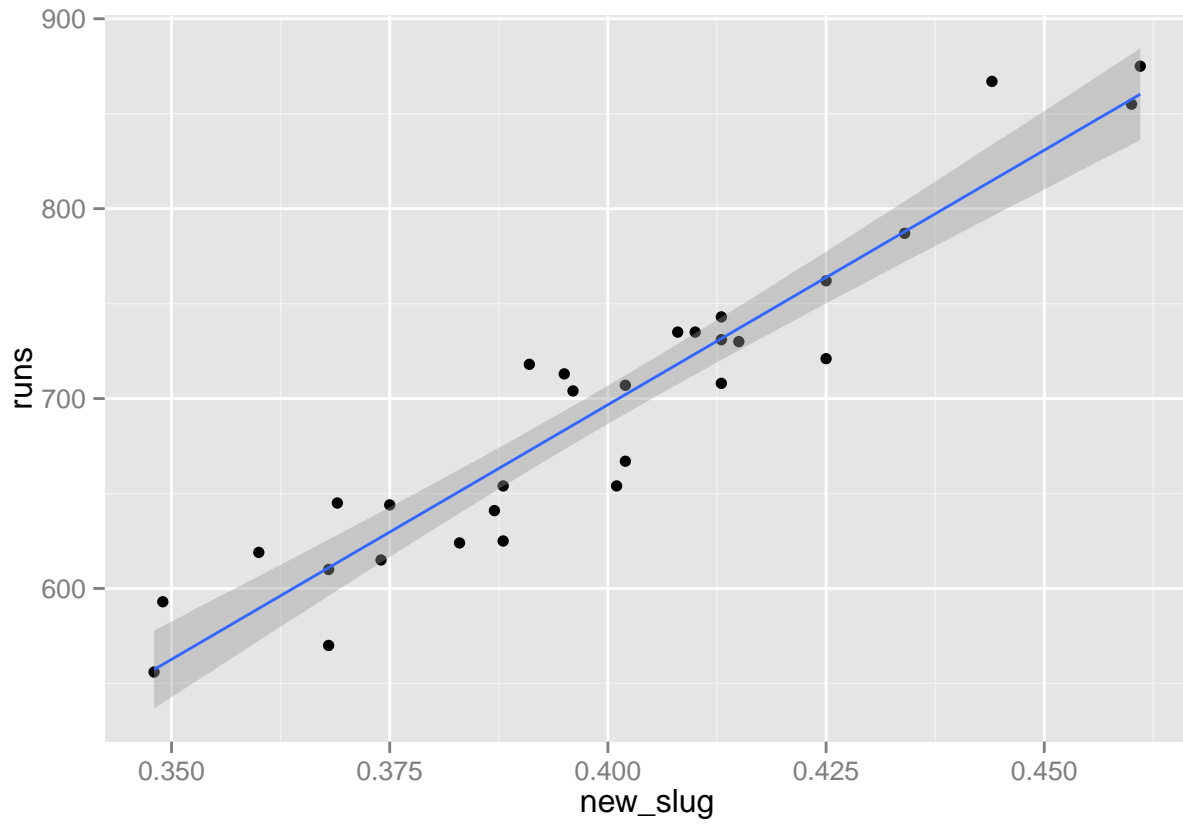


```
## [1] 0.9349271
```

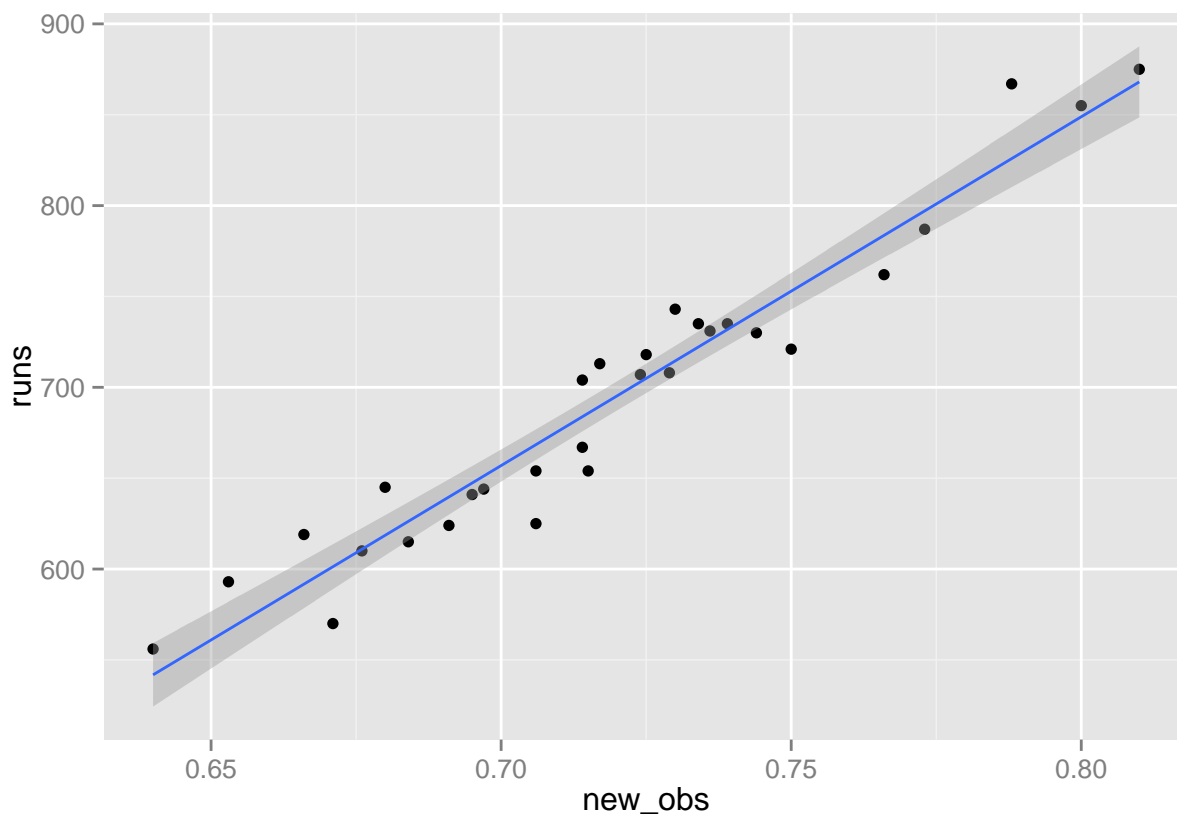
```
ggplot(mlb11,aes(x=new_onbase,y=runs)) + geom_point() + stat_smooth(method="lm")
```



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



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



These have way better relationships! This does make intuitive sense... Being on base makes it easier to score, and slugging makes it easier to round the bases.

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

This is a product of OpenIntro that is released under a [Creative Commons Attribution-ShareAlike 3.0 Unported](#). This lab was adapted for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel from a lab written by the faculty and TAs of UCLA Statistics.