Moneyball

Part 1

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 exercise 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 (and load up `mosaic` while we're at it!).

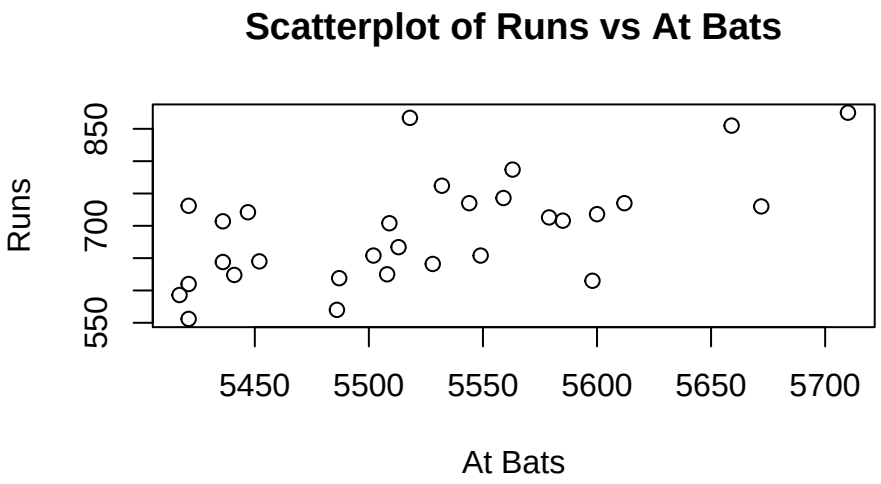
In addition to runs scored, there are seven traditionally used variables in the data set: at-bats, hits, home runs, batting average, strikeouts, stolen bases, and wins. There are also three newer variables: on-base percentage, slugging percentage, and on-base plus slugging. For the first portion of the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

SOLUTION:

The best type of plot in my Opinion would be a Scatterplot as we're displaying the relationshio between 'runs' and a numerical variable.

```
plot(mlb11$at_bats, mlb11$runs, main="Scatterplot of Runs vs At Bats", xlab="At Bats", ylab="Runs")
```



After plotting, this I found that htere doesn't seem to be a linear relationship between the two variables. I would not be comfortable using a linear model to predict the number of runs.

Here is the result for correlation test used in this exercise SOLUTION:

```
cor(runs ~ at_bats, data=mlb11)
```

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

0.610627 is the Correlation that is said to be found.

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 between two quantitative 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.

SOLUTION:

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 represents 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)
```

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. What are residuals?

The most common way to do linear regression is to select the **line that minimizes the sum of squared residuals**.

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?

YOU CAN SKIP THIS PROBLEM

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 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)
sum(residuals(m1)^2)
```

With this table, what is the least squares regression line?

SOLUTION:

runs = -2789.2429 + 0.6305 * at_bats is what the Least Squares Regression Line is.

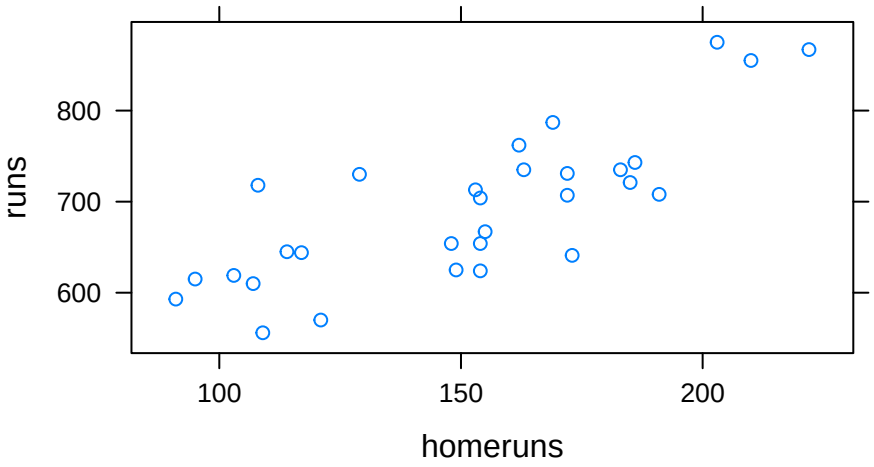
$$\widehat{runs} = -2789.24 + 0.63 at\_bats$$

fitted runs = -2789.24 + 0.63at_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?

SOLUTION:

```
xyplot(runs ~ homeruns, data=mlb11)
```




```
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
```

The Slope is 1.8345. In relation to the model, this means that every home run is associated with an increase of about 1.8345 runs.

Prediction and prediction errors

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

```
xyplot(runs ~ at_bats, data=mlb11, type=c("p", "r"))
```

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 5,579 at-bats? Is this an overestimate or an underestimate, and by how much?

SOLUTION:

```
pred1 <- makeFun(m1)
pred1(at_bat=5579)
filter(mlb11, at_bats==5579)
```

Calculate the residual:

```
713 - pred1(at_bat=5579)
```

The team predicted 728 runs but actually made 713, as such there is an overestimate of 15.6 runs.

Model Diagnostics

To assess whether the linear model is reliable, we need to check for Linearity, Independence, Normal errors, and Equal Variance.

- **Linearity**: You already checked if the relationship between runs and at-bats is linear using a scatterplot.
- **Equal Variance**: We want look at a plot of the residuals against the fitted values. If we see a change in the spread of the residuals for larger values of the fitted values, we would be uncomfortable with using a linear regression model as is.

```
xyplot(resid(m1) ~ fitted(m1), data=mlb11, type=c("p", "r"))
```

When looking at the plot, I see an equal variance of the residuals being distriubted. That means that model's assumptions are met.

6. Based on this, does the equal variance condition appear to be met?

Yes , we can assume that the equal variance condition is met.

- **Independence**: We don't really have independence here, since each observation in the dataset reflects on how a particular team did when playing against the other teams. We'll overlook this bit.

- **Normal errors**: To check this condition, we can look at a histogram of the residuals or a normal quantile plot:

```
histogram(~residuals(m1), width=50)
```

The Histogram shows that the residuals are normally distributed, as the Histogram shows a bell curve.

```
qqmath(~resid(m1))
ladd(panel.qqmathline(resid(m1)))
```

The Q-Q plot shows that the residuals are normally distributed.

7. Do the residuals appear normally distributed?

Yes based on the Q-Q and Histogram the residuals appear normally distributed.

8. Do there appear to be any outliers?

SOLUTION:

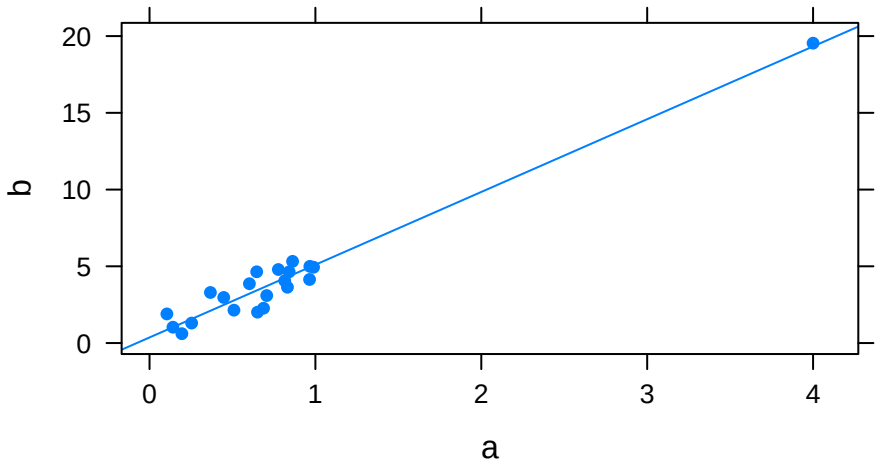
Yes there does seem to be one Outlier, when the Resiedual reaches 200.

9. Unusual points such as this can be:

YOU CAN SKIP THIS PROBLEM

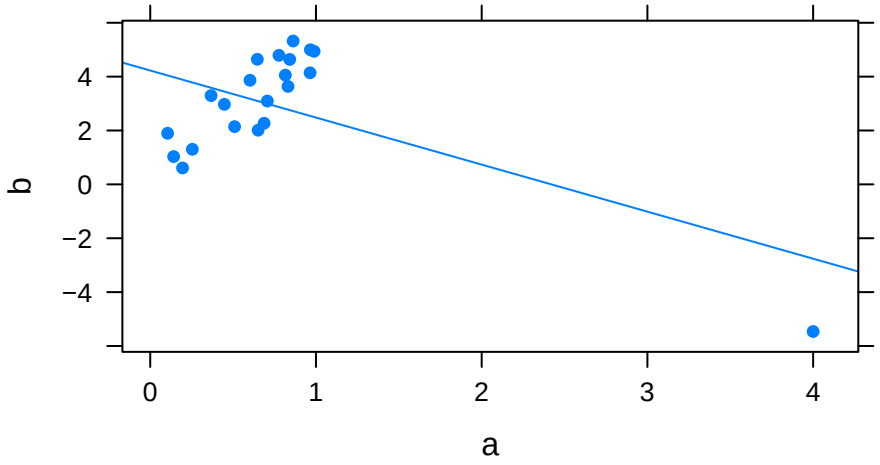
- high leverage: the observation is far away from the other points in the x-direction.

```
set.seed(15)
a <- c(runif(20), 4)
b <- a*5+rnorm(21)
xyplot(b~a, pch=16, type=c("p", "r"))
```



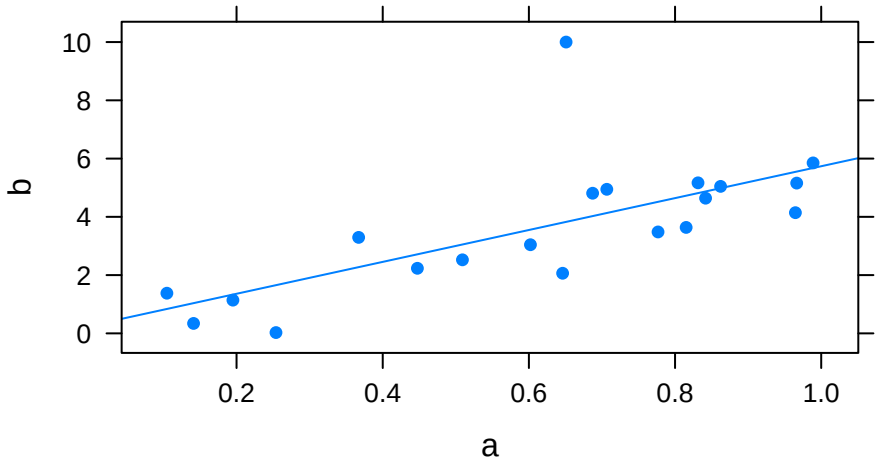
- high influence: the observation, when included, will significantly move the regression line.

```
b[21] <- b[21] -25
xyplot(b~a, pch=16, type=c("p", "r"))
```



- neither: the observation appears to be really far away from the rest of the observations, but due to its position, is unlikely to move the regression line substantially.

```
set.seed(15)
a <- sort(runif(20))
b <- a*5 + rnorm(20)
b[10] <- 10
xyplot(b~a, pch=16, type=c("p", "r"))
```



10. Is the point that you've identified high leverage, high influence, or both?

SOLUTION:

From first plot, the last data point (the point on the far right) is a high leverage point because it's an extreme value on the predictor (x) axis (the a variable). High leverage points have the potential to influence the regression line, but whether they actually do so depends on whether they also have a large residual.

In the second plot, the b value of the high leverage point was adjusted to make it also have a large residual. When this point is included, the regression line is significantly moved. Hence, it is now both a high leverage and a high influence point

In the third plot, the point b[10] is far away from the rest of the observations in the vertical (y) direction, but it's not at an extreme a value, so it's not a high leverage point. It also doesn't significantly affect the slope of the regression line, so it's not a high influence point either.

Part 2

Now ,this is your turn (working in groups is a nice solution)

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. At a glance, does there seem to be a linear relationship? '

```
plot(mlb11$hits, mlb11$runs)
model_hits <- lm(runs ~ hits, data = mlb11)
abline(model_hits)
```

When graphing this as a Scatterplot, I find that there is a Linear relationship between the two variables.

' 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. Does your variable seem to predict `runs` better than `at_bats`? How can you tell?

```
summary(model_hits)$r.squared
model_at_bats <- lm(runs ~ at_bats, data = mlb11)
summary(model_at_bats)$r.squared
```

Using the R^2 values, I find that the variable hits is a better predictor of runs.

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 we've discussed (for the sake of conciseness, only include output for the best variable, not all five).

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

1. The Resuiduals vs Fitted Graph show that nonlinearity is present.
2. However, the Normal Q-Q Graph shows that the residuals are normally distributed.
3. The Scale-Location Graph shows that the residuals are not equally distributed.
4. The Residuals vs Leverage Graph shows that there are no high leverage points.
5. 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 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?

```
# Create linear regression models
model_new_onbase <- lm(runs ~ new_onbase, data = mlb11)
model_new_slug <- lm(runs ~ new_slug, data = mlb11)
model_new_obs <- lm(runs ~ new_obs, data = mlb11)

# Print summary statistics including R-squared values
summary(model_new_onbase)$r.squared
summary(model_new_slug)$r.squared
summary(model_new_obs)$r.squared

# Create scatterplots with regression lines
par(mfrow = c(1, 3)) # setting the plot window to have one row and three columns

# Scatterplot for new_onbase
plot(mlb11$new_onbase, mlb11$runs, main = "New On Base", xlab = "New On Base", ylab = "Runs")
abline(model_new_onbase, col="red")
```

```r
# Scatterplot for new_slug
plot(mlb11$new_slug, mlb11$runs, main = "New Slug", xlab = "New Slug", ylab = "Runs")
abline(model_new_slug, col="blue")

# Scatterplot for new_obs
plot(mlb11$new_obs, mlb11$runs, main = "New OBS", xlab = "New OBS", ylab = "Runs")
abline(model_new_obs, col="green")
```

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

Based on the R-squared values, the new_onbase variable is the best predictor of runs. Also, looking the visualization of models, the new_obs variable has the best line of fit.