

## Lab 7: Introduction to linear regression

### Batter up

The recently released movie *Moneyball* focuses on the “quest for the secret of success in baseball”. It follows a low-budget team, the Oakland Athletics, in the early 2000’s, who believed that underused statistics, such as a player’s ability to get on base, actually better predicts the ability to score runs than typical statistics like homeruns, RBIs (runs batted in), and batting average. In fact, 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, best helps us predict a team’s runs scored in a season.

### The data

Let’s load up the batting data for the 2009 season.

```
source("http://www.openintro.org/stat/data/mlb09.csv", destfile = "mlb09.csv")
```

In addition to runs scored, there are seven traditionally-used variables in the data set: at-bats, hits, homeruns, batting average, strikeouts, walks and stolen bases.<sup>1</sup> The last three variables in the data set are on-base and slugging percentage and on base plus slugging. For the first portion of the analysis we’ll stick with the traditionally used variables. The last question in the on your own part is about the newer variables.

**Exercise 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 explanatory variable. Does the relationship look linear? If you knew a team’s at-bats, would you feel confident in your ability to make a good prediction of their runs?

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

```
cor(mlb09$runs, mlb09$at_bats)
```

---

<sup>1</sup>Though it’s not necessary for this lab, if you’d like a refresher in the rules of baseball and a description of these statistics visit [http://en.wikipedia.org/wiki/Baseball\\_rules](http://en.wikipedia.org/wiki/Baseball_rules) and [http://en.wikipedia.org/wiki/Baseball\\_statistics](http://en.wikipedia.org/wiki/Baseball_statistics).

## Sum of squared residuals

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

**Exercise 2** Looking at your plot from the previous exercise, describe the relationship between these two variables? Make sure to discuss the form, direction, and the strength of the relationship as well as any unusual observations.

Just as we used the mean and standard deviation to summarize a single variable, we can summarize the relationship between these two variables most simply 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.

```
plotSS(x = mlb09$at_bats, y = mlb09$runs)
```



After running this line of code, you'll be prompted to click two points on the plot to define a line. Once you've done that, it will plot the line in black and the residuals in red. The blue boxes represent the squared residuals.

You'll recall that the most common way to do linear regression is to select the line that minimizes the sum of squared residuals, where a residual is defined as the difference between the observed y-value and the predicted y-value:

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

Note that the output from the `plotSS()` function provides you with the slope and intercept of your line as well as the sum of squares.

**Exercise 3** Using `plotSS()`, choose a line that does the best job of minimizing the sum of squares. Run the function several times to get the lowest possible value. What is that value? How does it compare to your neighbors?

## 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 one of the built in functions in R, `lm()` to fit the regression line.

```
m1 <- lm(runs ~ at_bats, data = mlb09)
```

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 `mlb09` dataframe to find these 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 pull up the basic information.

```
summary(m1)
```

At the top of the output is the formula that you specified, runs as a function of at-bats. Below that is a table that contains summary statistics on the residuals and below that is the regression output. The coefficient estimate of the intercept is shown in the first row, next to **(Intercept)**, and the estimate of the slope is shown in the second line, next to **at\_bats**.

**Exercise 4** Using the estimates from the R output write the equation of the regression line for predicting runs from at-bats. What is the slope? What does the slope tell us in the context of the relationship between success of a team and its at-bats?

Another piece of useful information given in the regression output is  $R^2$ , **Multiple R-squared**. This value represents the proportion of variability in the response variable explained by the explanatory variable. In this case, 36.2% of the variability in runs is explained by at-bats. The remainder is due to other factors related to the team's success.

## Prediction and prediction errors

Let's overlay the regression line on the scatterplot of runs vs. at-bats.

```
plot(x = mlb09$at_bats, y = mlb09$runs)
abline(m1)
```

**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 that had 5,578? Is this an overestimate or an underestimate, and by how much? In other words, what is the error for this prediction?

## Model diagnostics

In order to check if the conditions for a linear regression are met we should 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 make a residuals plot of the residuals vs. at-bats to check this condition.

```
plot(x = mlb09$at_bats, y = m1$residuals)
abline(h = 0, lty = 3) # adds a horizontal dashed line at y = 0
```

**Exercise 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?

- (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) # adds diagonal line to the normal probability plot
```



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

(3) Constant variability:

**Exercise 8** Based on the residuals plot from earlier, does the constant variability condition appear to be met?



## On Your Own

1. Choose another *traditional* variable from `mlb09` 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?
2. How does this relationship compare to the relationship between `runs` and `at_bats`? Using the  $R^2$  value from the summaries of the two models. Based on your comparison, what do you think the  $R^2$  value represents? Does your variable seem to predict `runs` better than `at_bats`? Why?
3. Now that you can summarize the linear relationship between two variables, investigate the relationships between `runs` and all other *traditional* variables. Which variable best predicts `runs`? Support your conclusion using the graphical and numerical methods we've discussed and describe the relationships you explore in context of the problem. How well does this variable predict `runs`? 
4. 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?
5. Check the model diagnostics for the regression model with the predictor you decided was the best predictor for runs.
6. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere, e.g. lecture, discussion section, previous labs, or homework problems? Be specific in your answer. 

## Notes

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.