Lab 6: Introduction to linear regression

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform. Note that the order of the choices in multiple choice questions may be different on the Duke Coursera platform than the order in this document.

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. (If you're not familiar with the movie, you can read more about it and watch the trailer here (http://www.imdb.com/title/tt1210166/).)

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 the data for the 2011 season.

```
load(url("http://www.openintro.org/stat/data/mlb11.RData"))
```

In addition to runs scored, there are seven traditionally used variables in the data set: `at_bats`, `hits`, `homeruns`, `bat_avg` (batting average), `strikeouts`, `stolen_bases`, and `wins`. Even 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). There are also three newer variables: `new_onbase` (on base percentage), `new_slug` (slugging percentage), and `new_obs` (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.

Question 1 What type of plot would you use to display the relationship between `runs` and one of the other numerical variables?

A) histogram

B) box plot
C) scatterplot
D) bar plot

Question 2 Plot the relationship between `runs` and `at_bats`, using `at_bats` as the explanatory variable.
The relationship appears to be …

A) linear
B) negative
C) horseshoe-shaped (∩)
D) u-shaped (∪)

Exercise: If you knew a team's `at_bats`, would you be comfortable using a linear model to predict their 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)
```

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.

Question 3 Looking at your plot from the previous exercise, which of the following best describe the relationship between these two variables?

A) The relationship is negative, linear, and moderately strong. One of the potential outliers is a team with approximately 5520 at bats.
B) The relationship is positive, linear, and moderately strong. One of the potential outliers is a team with approximately 5520 at bats.
C) The relationship is positive, linear, and very weak. There are no outliers.
D) The relationship is positive, linear, and very weak. One of the potential outliers is a team with approximately 5520 at bats.

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)
```

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)
```

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.

Exercise: Using `plot_ss`, choose a line that does a good job of minimizing the sum of squares. Run the function several times. Report your smallest 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)
```

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.

Question 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?

A) For each additional home run, the model predicts 1.83 more runs, on average.
B) Each additional home run increases runs by 1.83.
C) For each additional home run, the model predicts 1.83 fewer runs, on average.
D) For each additional home run, the model predicts 415.24 more runs, on average.
E) For each additional home run, the model predicts 415.24 fewer runs, on average.

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.

Exercise: 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?

To find the observed number of runs for the team with 5,579 at bats you can use the following:

```
mlb11$runs[mlb11$at_bats == 5579]
```

This code subsets the vector of `mlb11$runs` for the case where `mlb11$at_bats` equals 5,579.

Question 5 What is the residual for the prediction of runs for a team with 5,579 at-bats? Choose the closest answer.

A) -15.32
B) 15.32
C) 713
D) 5,579

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.

(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.

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

Question 6 Which of the following statements about the residual plot is false?

A) The residuals appear to be randomly distributed around 0.
B) The residuals show a curved pattern.
C) The plot is indicative of a linear relationship between runs and at-bats.
D) The team with a very high residual compared to the others appears to be an outlier.

(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 prob plot
```

Question 7 Which of the following is true?

A)The residuals are extremely right skewed, hence the normal distribution of residuals condition is not met.
B) The residuals are extremely left skewed, hence the normal distribution of residuals condition is not met.
C) The residuals are perfectly symmetric, hence the normal distribution of residuals condition is met.
D) The residuals are fairly symmetric, with only a slightly longer tail on the right, hence it would be appropriate to deem the the normal distribution of residuals condition met.

(3) Constant variability:

Question 8 Based on the plot in (1), the constant variability condition appears to be met.

A) True
B) False

Exercise: 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?

Exercise: 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?

Question 9

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`?

A) at bats
B) hits
C) wins
D) batting average

Question 10

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`?
A) on-base plus slugging (`new_obs`)
B) slugging percentage (`new_slug`)
C) on-base percentage (`new_onbase`)

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