APPM2720 Week 7 Lecture: Intro to least squares

The key to this week is realize that one can fit equations and formulas to data to describe patterns and make predictions. The method of least squares (the algorithm) is a way to do this but it does not replace the concept that is there is an equation (the model) we are using to describe the data set. Playing off the alogithm against the model is a big theme in data science. Least squares is part of a family of algorithms that fit data by minimizing a specific criterion.

This topic will follow the Chapter 3 of *An Introduction to Statistical Learning with R* (ISLR) and the pdf for this book has been made publically available with a copy posted on the class web page.

Students might also take a look at Chapter 2 of ISLR to see if their R skills match the intro to R in this text.

A simple model for a data set

Given a set of numbers $y_1, y_2, \ldots, y_n$ how should one describe the center of these values? At this point using the **mean** or **median** may seem natural. Here is a criterion to measure how far a single number a is from all the data points.

$$S(a) = (y_1 - a)^2 + (y_2 - a)^2 + \ldots + (y_n - a)^2 = \sum_{i=1}^{n} (y_i - a)^2$$

- the mean minimizes $S(a)$ for all choices of a !
- if the square is replaced by absolute values $(y_i - a)^2 \rightarrow |y_i - a|$ then the minimium is at the median!

Fitting a line to a scatter plot.

Often a scatterplot suggests a straight line relationship between two variables in a data set

- The convention is call the variable on the x-axis X and the variable on the y-axis Y.
- For a prediction application X is chosen to be the variable that is known and Y is the variable that one wants to predict. E. g. given a specific mileage (X) what is the asking price of a used Audi A4.
- This relationship is written as an equation: $Y \approx \beta_0 + X\beta_1$. (Statisticians like Greek letters for coefficients and also like using subscripts.) But is just β_0 the intercept and β_1 the slope.
- These two parameters can be estimated from pairs of data points using least sqaures. This is done under several assumptionns -- more on this later.

Compting the least squares solution

- The criterion now depends on both the slope and intercept: $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i)^2$
- Brute force: try out all choices of the βs.
- Use a formula if it is available (exists for the simple case of slope and intercept).
- Use the R function **lsfit**

```
out<- lsfit( x, y)
out$coefficients
plot( x,y)
abline( out$coefficients)
```

- Use the formula based function **lm**

```
out<- lm( y ~ x)
out$coefficients
plot( x,y)
abline( out$coefficients)
```