APPM2720 Week 7_2 Lecture: The statistical foundations of least squares

Least squares is part of a family of algorithms that fit data by minimizing a specific criterion. It is appropriate to use if the data satisfies certain assumptions. This lecture helps to explains what these assumptions are

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 freely available with a pdf copy posted on the class web page.

A simple model for a data set

Given N pairs of numbers $(x_1, y_1), (x_2, y_2), \dots, x_N, y_N)$. A useful model is to predict the Ys by a straight line in X:

$Y \approx \beta_0 + \beta_1 X$

For this to be useful we need to make some more assumptions about this idea. The assumptions taken together are called a statistical model.

- Y actually follows a linear relationship in X!
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

  The ϵ_i are called the errors and are assumed to have

1) have a histogram that follows a Gaussian(aka normal) distribution in shape, (e.g. symmetric, no outliers)

2) have a mean of zero,

3) have no obvious patterns when plotted against X_i or in any order.

The basic concept is that the ϵ_i are random, not predictable.

We will never know the values for the ϵ_i exactly! Why?

Simulating a model

Assume that $Y = 2 + 3X + \epsilon$ Generate 100 observations

```
X<- runif(100) # 100 values between [0,1]
trueErrors<- rnorm(100, mean= 0.0, sd=.5)
Y<- 2+ 3* X + trueErrors
fit<- lm( Y~X)
summary( fit)
```

We do not recover the intercept and slope exactly. and the residuals are not exactly equal to the `trueErrors` . However, the accuracy will improve as the number of observations is increased.

When using the normal an easy rule is that you expect about 95% of the values to be within 2 standard deviations of the mean. In this situation the mean is zero.

```
(trueErrors <= 2* (.5) + 0.0) &

(trueErrors >= -2* (.5) + 0.0) &
)
```

The percentage gets closer to .95 as the sample size increases. In general the probability of the true errors being in a particular interval can be computed using the `pnorm` function in R.

Checking the model -- about residuals

If you fit a line to the data find the differences

$$e_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i$$

These are just **observation - predicted value** and are called the ***residuals***.

- The least squares method actually finds the intercept and slope to minimize the sum of squares of the residuals
- The residuals are estimates of the true errors.
- The standard deviation of the residuals (aka residual standard error) is an estimate of the theoretical standard deviation of the errors.
- Examining the residuals is one of the ways to check if the assumptions of the model hold.

  Some things to try:

- Plot the residuals against the predicted values

- Plot the residuals against other important variables or ordering (e.g. time order of observations)

- Histograms or boxplots of the residuals