Math 455 Class 1 January 21

Textbook: Linear Models with R 2nd edition

NOT 1st or 3rd edition.

If you get the 1st or 3rd edition the homework will not make sense. Problems are different.

Textbook website has scripts for all R code in the book. This is useful for the homework. Most problems in Chapter X can be solved by looking at examples in Chapter X and making appropriate adaptations.

Course Website:

people.math.binghamton.edu/dikran/455/

Notes posted here.

Prerequisites: Math 447 & 448

(or equivalent)

This is exactly the problem that you spent most of Math 448 studying. So a review of what it means for an estimator to be "unbiased" is not out of place.

In case you need a quick review:

On the course web page there are summaries of Math 447 and 448.

Something which is not technically a prereq but is extremely useful: Linear Algebra.

See, for example from Faraway p. 33

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

an equation in linear algebra.

Q: Why Regression Analysis?

A: All of social science!

Many if not all social science questions are answered by gathering data and doing a regression analysis

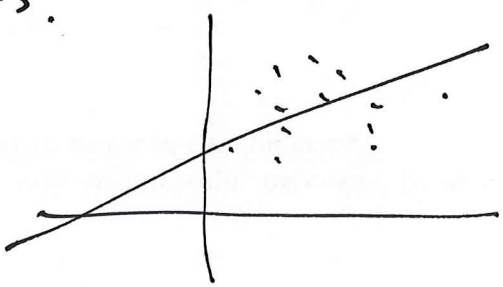
Another answer: How do we make health care decisions / ~~off~~ how does the FDA decide which drugs to approve?

Regression Analysis.

Short version: we are using regression for so many things that it is worth understanding very well.

What is regression analysis?

Regression analysis is drawing a line through a cloud of points.



This is meant very loosely: we could also be drawing a 4-dimensional hyperplane through a cloud of points in 5-dimensional space.

Where would such a space come from?

When we have patients and we measure

height weight age blood pressure (systolic, diastolic)

each patient is now a point in a 5-dim'l space, or higher-dim'l, if we measure more things.

Most regressions we look at will have many variables. Example, Faraway p. 18

The command to do a linear regression in R is "lm". The syntax of this command uses Wilkinson-Rogers Notation.

Wilkinson - Rogers Notation:

We want to have a linear equation

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p + a_0$$

$Y, X_1, X_2, \ldots X_p$ will have names.

in the example, Faraway p. 18

Species, Area, Elevation, etc.

We dont want to have to write out the full equation and name the ~~variables~~ coefficients every time.

Solution: Wilkinson - Rogers Notation.

Many syntactic tricks; e.g.

lmod ← lm (Species ~ .)

This period means "all other variables in the current data set."