Crafting with Data

Reality, Illusions, Truth & the Future

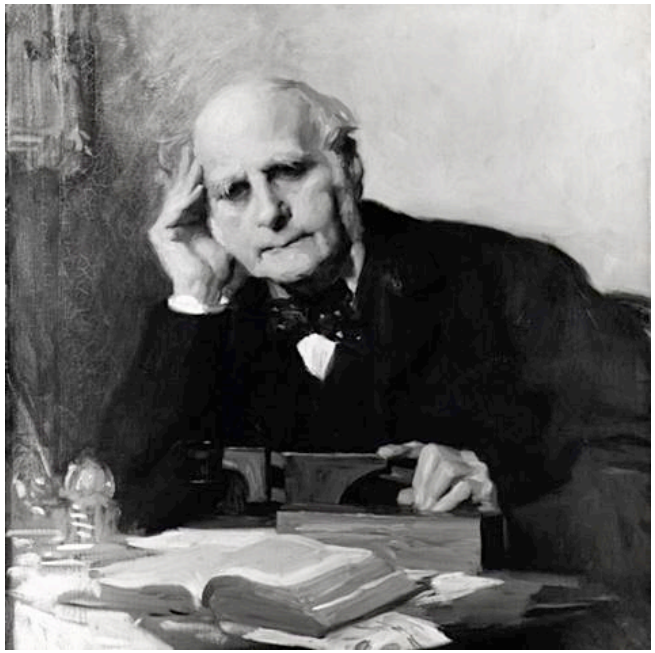
Instructor: Rob Faludi

Plan for Today

- Predictive Statistics
- Exercise One: **R**
- Readings & Assignments

Regression

- What's a better word for this?
- Galton
- author, explorer, psychologist, fingerprints, pioneer statistician



- also, founder of eugenics
- ethics aren't always obvious

Dependent vs. Independent Variables

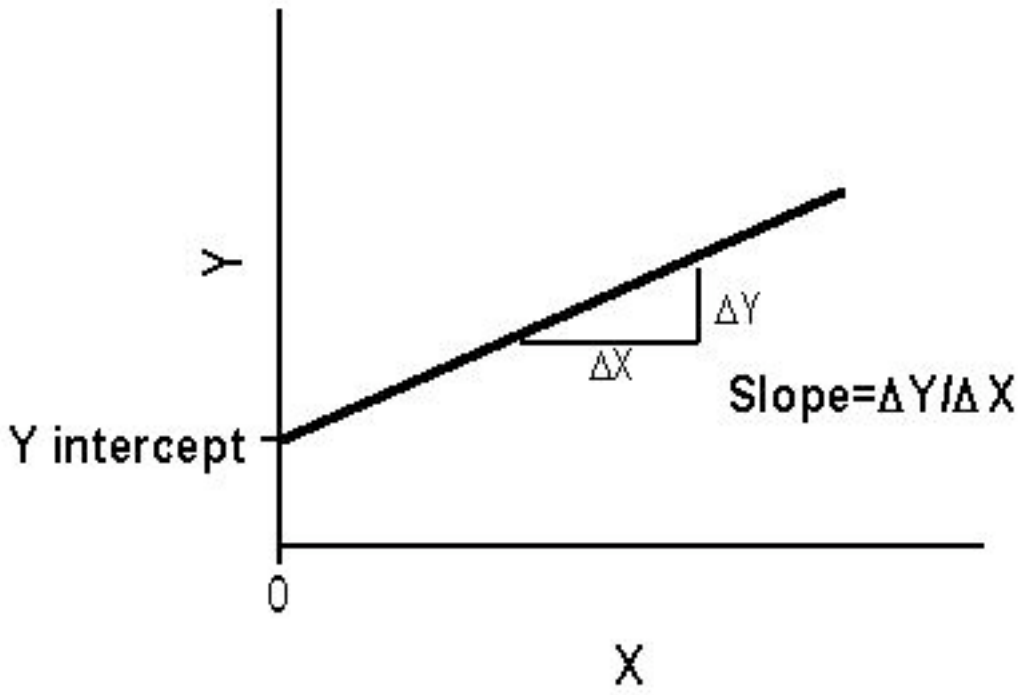
- Dependent variable is what we're interested in

- Independent will tell us where to look for it

Linear Regression

- Y' = bX + A

- where X is the variable represented on the abscissa (X-axis), b is the slope of the line, A is the Y intercept, and Y' consists of the predicted values of Y for the various values of X



Standard Error of the Estimate

- The standard error of the estimate is a measure of the accuracy of predictions made with a regression line.

- The sum of the errors of prediction is zero, so just like in confirmatory statistics, we square the errors so we can deal with them mathematically

- Standard Error of the Estimate: $$\sigma_{est} = \sqrt{\frac{\sum (Y-Y')^2}{N}}$$

  - ...where N is the number of pairs of (X,Y) points, Y is each dependent variable and Y' (y-prime) is the predicted value

- We don't typically know population values so: $$s_{est} = \sqrt{\frac{\sum (Y-Y')^2}{N-2}}$$

Sums of Squares

- Similar to ANOVA, we need to partition the error terms

- SSY = SSY' + SSE

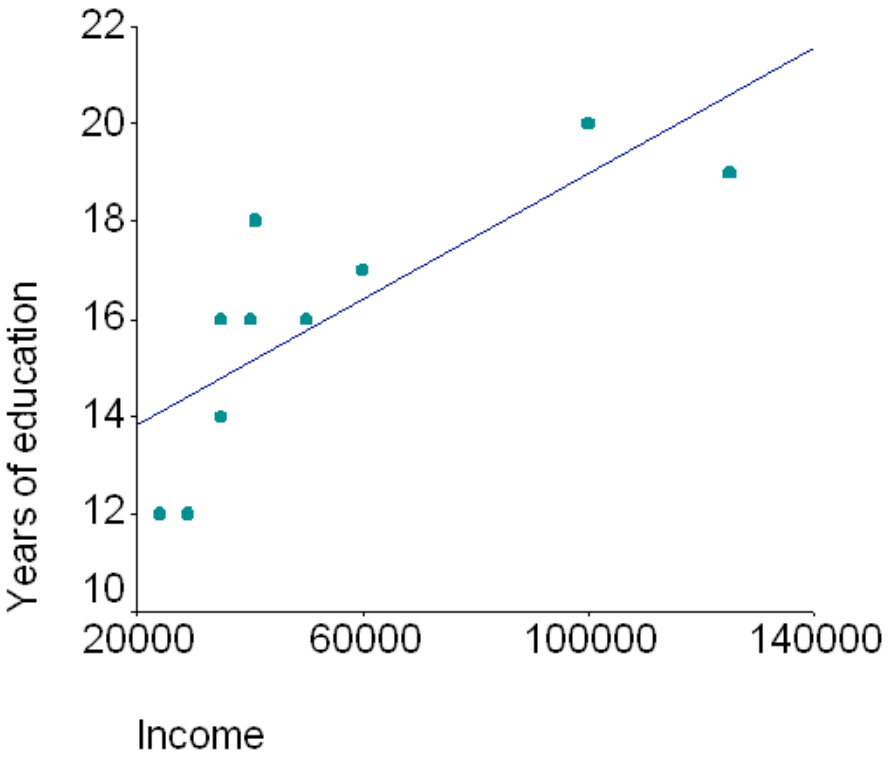
- r^2 = SSY'/SSY

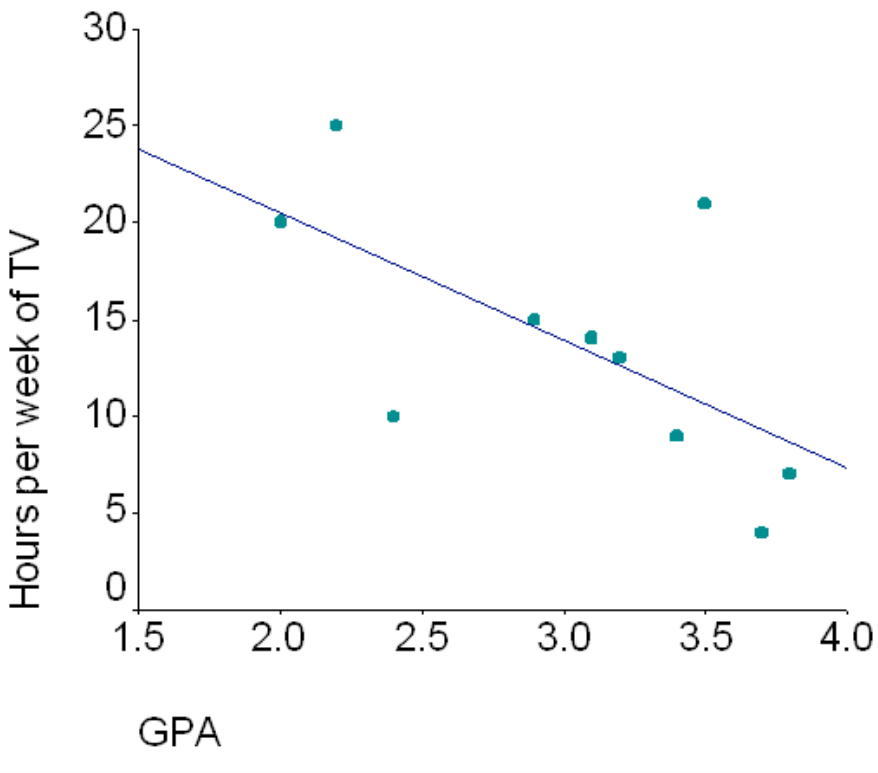
- Significance test: $$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- Look up *t* in a table and you're done.

Correlation

- how do variables relate to each other?





Correlation Coefficient

- Sample correlation: $$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

- where x-bar and y-bar are the sample means of X and Y , sx and sy are the sample standard deviations of X and Y

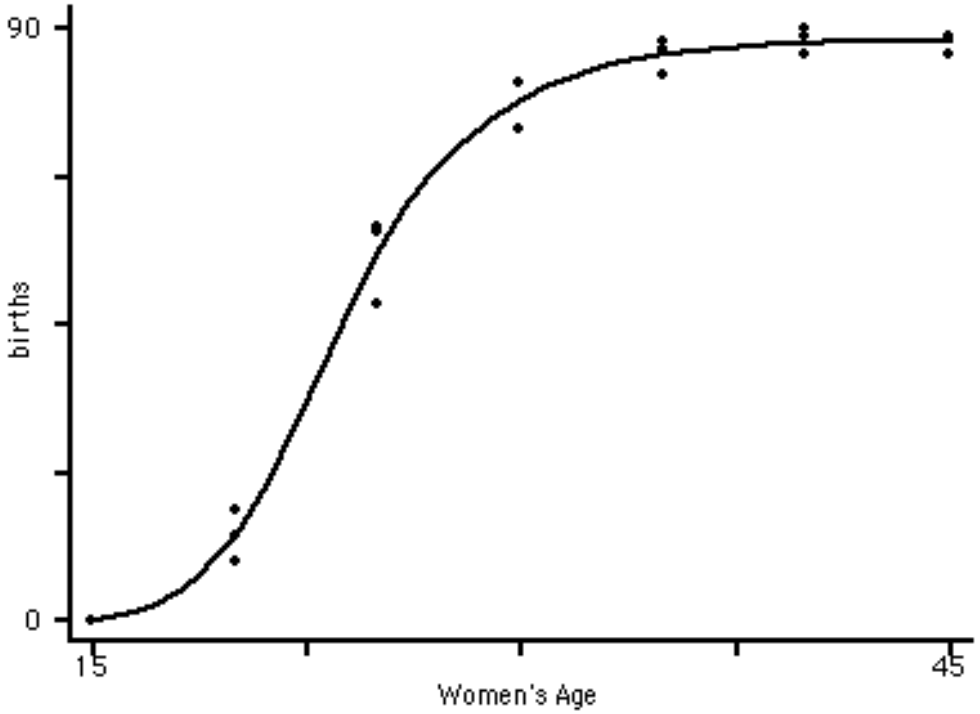
- Phew! So how do we interpret *r* ?

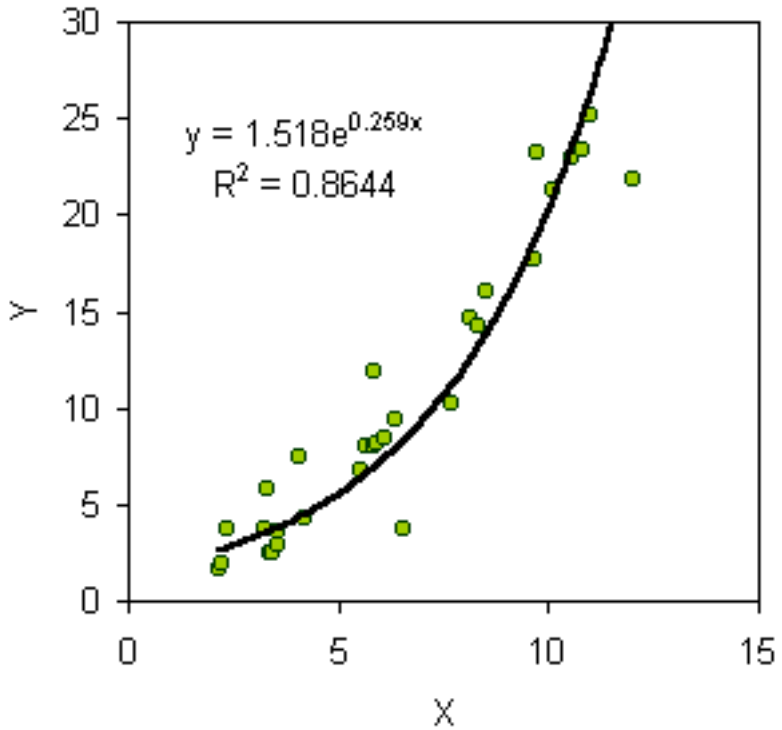
- large ≠ important, usually = trivial

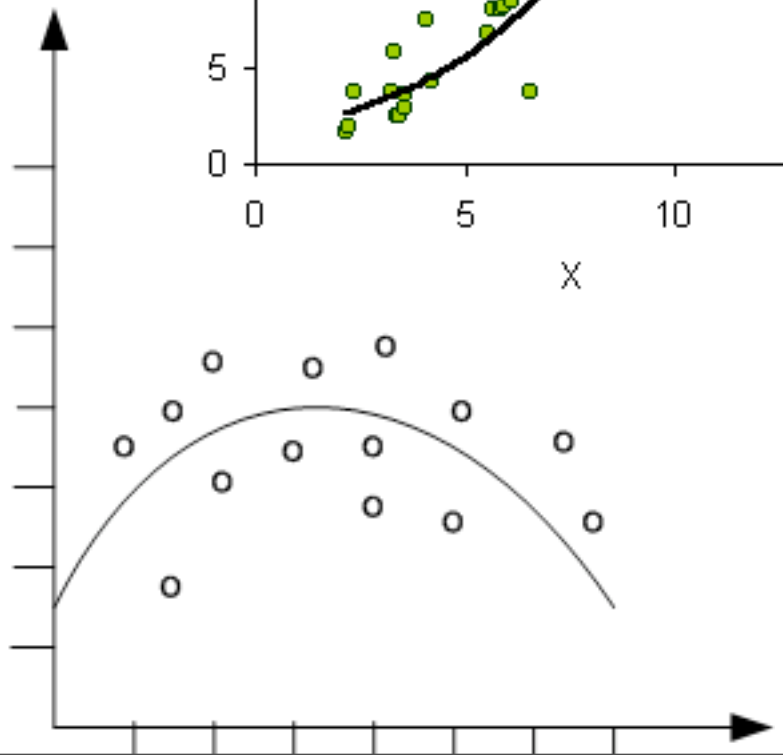
small	0.1	0.3
medium	0.3	0.5
large	0.5	1.0

Non-linear regression

- deals with curves



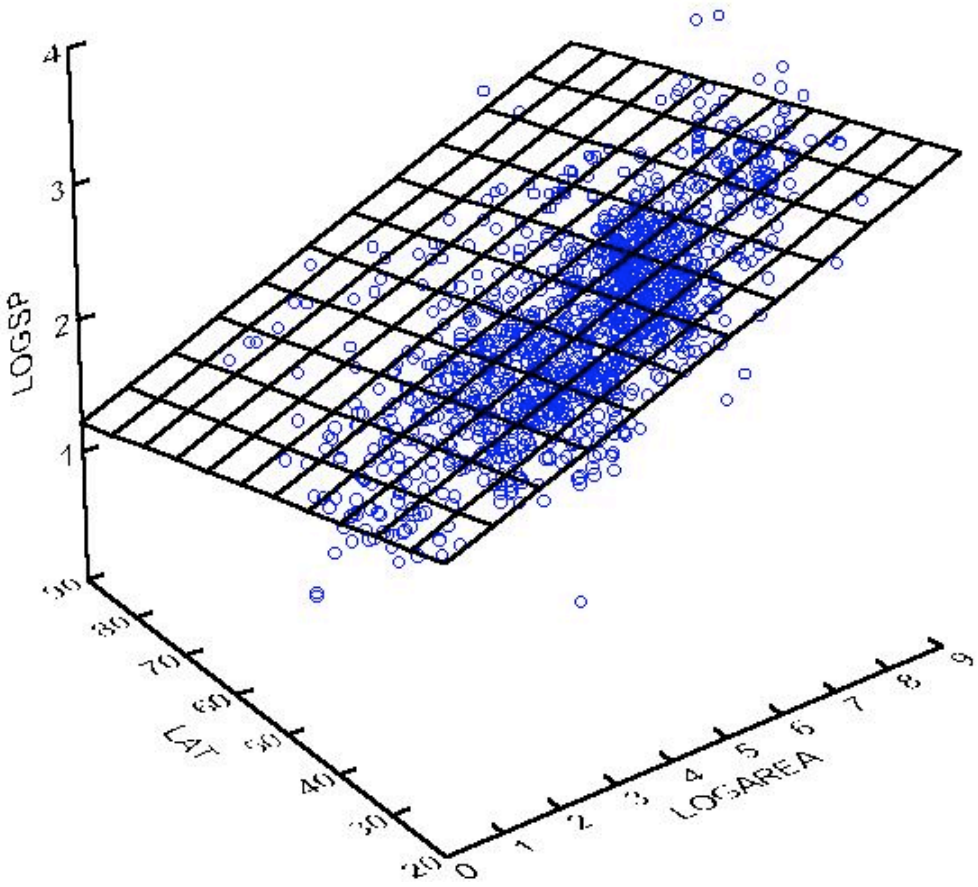




Multiple regression

- Deals with multiple independent variables

- fitting a plane



R

http://www.r-project.org/

The Predictive Analytics Company

Free Statistics

www.freestatistics.info

Google Docs BETA

R

http://www.r-project.org/

R Facts

- Interpreted
    - interactive!
    - can be slower for large data sets
- Dynamically typed
- Free version of S
- Extensible

Downloading R

- http://cran.r-project.org/mirrors.html
    - Slovenia? New Zealand? Brazil?

Install and Launch R

- For your platform

R Help

- help()
- help.start()
- help(t.test)
- ?t.test
- help("stem")

R Variables

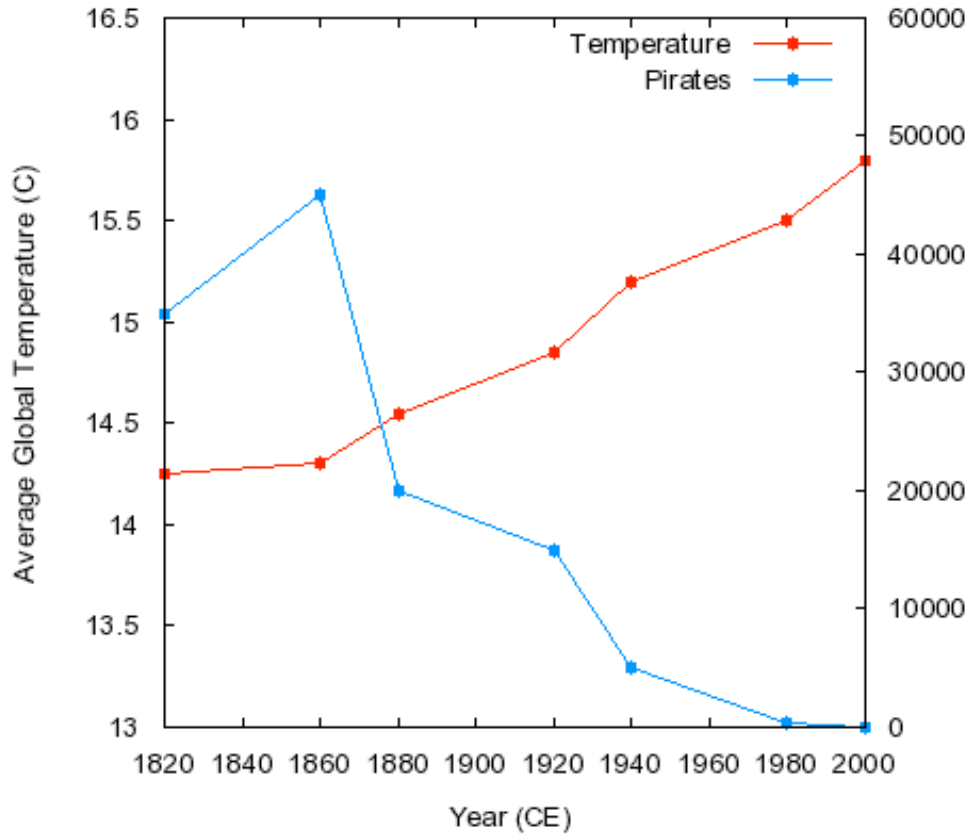
- x <- 3 #comments go after the pound sign
- x <- ‘foop’
- x <- 1:20
- x * 50
- x[4]
- x <- c(27, 23, 31, 42, 15, 28)

R Descriptives

- mean(x)
- median(x)
- mode(x) #explain
- table(x)
- var(x)
- sd(x)

R Graphs

- y <- rnorm(100)
- hist(y)
- z <- runif(100)
- plot(y,z) #scatterplot
- barplot(y)
- plot(y, type='o', col='blue')
- pie(x)
- pie(x, main='Rob\'s Data', col=rainbow(length(x)))
- stem(y)



R More

- objects() #what did we make?
- rm(x) # goodbye to x
- x <- ‘pirates are scary’ # strings
- command line:
  - R
  - /Library/Frameworks/R.framework/Resources/R
  - q() or Ctrl-D to exit!

R Statistics

- binom.test(40,100,p=0.5)
- a <- c(14,23,50,20)
- b<- c(12,20,56,10)
- chisq.test(a, p = b, rescale.p= TRUE)
- women <-c (5.5,5.3,6.0,5.8,5.2,5.6)
- men <-c (6,5.9,6.1,5.5,6.2,5.9)
- t.test(men,women)

R More Stats

- women <- c(5.5,5.3,6.0,5.8,5.2,5.6)
- men <- c(6,5.9,6.1,5.5,6.2,5.9)
- t.test(men,women)
- boxplot(men,women)

R Files

- http://faludi.com/dowloads/data.tab
- myData = read.table(file="~/Desktop/data.tab", header=TRUE)
- t.test(myData['Men'],myData['Women'])
- boxplot(myData)

Men	Women
67.28	66.11
69.08	61.11
70.61	65.17
72.39	66.11
67.61	61.11
76.44	65.17

Readings and Assignments

- Readings
  - How to Lie with Maps, *selected chapters for Wednesday*