Stat 104: Quantitative Methods

Class 7: Regression

Regression-Better than correlation

- Regression analysis is a statistical technique that is very useful for exploring the relationship between 2 variables.
- One variable is considered the explanatory variable and the other the dependent variable.
- Regression allows one to do predictions which cannot be done with correlation.

Fun! Science! Facts!

Where did this rule come from?

> The frequency of chirping varies according to temperature. To get a rough estimate of the temperature in degrees fahrenheit, count the number of chirps in 15 seconds and then add 37. The number you get will be an approximation of the outside temperature.

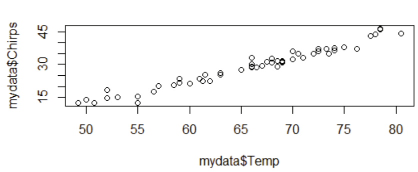
- They fit a line to the data set. This is what regression does-it relates a Y variable to an X variable.
- ***There are many ways to fit a line to data***, though one method is the most popular (but not always the best method).

Mathematically they are saying Temp = 37+Chirps…how wrong are they????

Cricket Data

- X= number of chirps per 15 seconds
- Y = Temperature

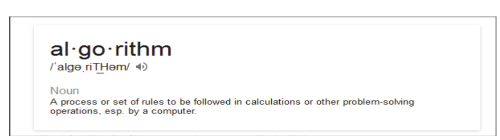
```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/chirps.csv")
```



Data from http://blog.globe.gov/sciblog/2007/10/05/measuring-temperature-using-crickets/

Fitting Line Method 1

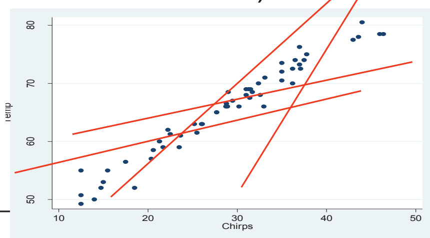
- Draw a line by hand
- Not exactly scientific..not **algorithmic**



- **We want easily reproducible results.**

Which Two Points?

- Two points define a line, but which two points (and thus which line?)



Fitting Line Method 2

- Two points define a line….so we need two points.
- Split the X axis in two, so there is a lower half and upper half group of points.
- Find two “good” points in each group
- Fit a line connecting these two points.
- I just made this method up, by the way.

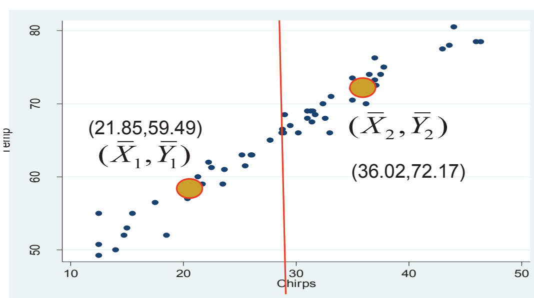
Make This an Algorithm

- Calculate the median of the X’s
- Separate the data into two groups
- Find $(\bar{X}_1, \bar{Y}_1)$ and $(\bar{X}_2, \bar{Y}_2)$
- Calculate the line between these two points
- Recall the equation of a line formula

$$Y - Y_1 = m(X - X_1)$$

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

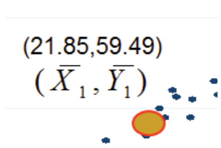
A Picture



29.5

For those who care: R Code

```
> median(mydata$Chirps)
[1] 29.5
> mean(mydata$Temp[mydata$Chirps<=median(mydata$Chirps)])
[1] 59.49107
> mean(mydata$Chirps[mydata$Chirps<=median(mydata$Chirps)])
[1] 21.84946
```



The Fitted Line

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{(72.17 - 59.49)}{(36.02 - 21.85)} = 0.90$$

$$Y - Y_1 = m(X - X_1) \Rightarrow Y = 39.8 + .9X$$

So for prediction, we say Temp = 39.8+.9(Chirps)

Interpret the Line

- How do we interpret this:

$$Temp = 39.8 + 0.9(Chirps)$$

- If chirps goes up by 1 unit (1 additional chirp per time period), temp goes up by 0.9.
- How wrong are we???

Pause: The Equation of a Line

English words for the French word montant
amount, figure, rising, sum

- Most Americans have been brainwashed

$$Y = mX + b$$

- (allegedly in France they use y=sx+b)
- As adults, we will now use the notation

$$Y = b_0 + b_1 X$$

https://www.math.duke.edu//education/webfeats/Slope/Slopederiv.html

Notation for **Our** Line

- We need to be able to distinguish between our observed Y values, and the Y values that our line produces.
- So given a slope and intercept, we produce what is called the fitted line:

$$\hat{Y}_i = b_0 + b_1 X_i$$

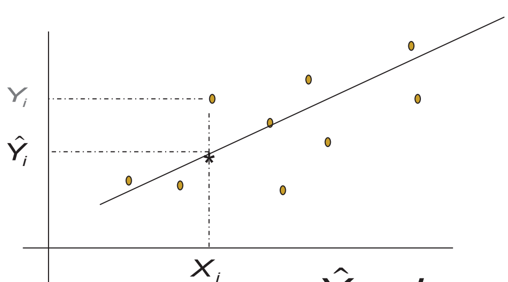
Pause: To Fit a Line to Data

- Fitting a line to data means to find "good" values of b_0 and b_1 .
- We define our fitting error as

$$e_i = Y_i - b_0 - b_1 X_i = Y_i - \hat{Y}_i$$

- Ideally, we want all the errors to be zero. Is this always possible?
- So we need a **criterion function**

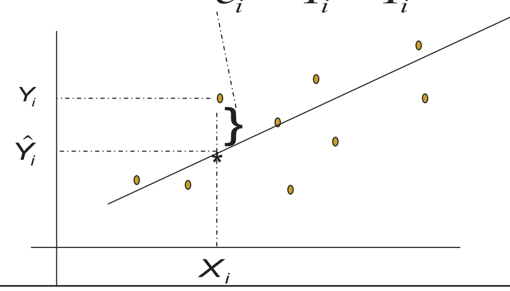
Observed versus Fitted Values



$$\hat{Y}_i = b_0 + b_1 X_i$$

The Errors

error $e_i = Y_i - \hat{Y}_i$



Criterion Function 1 (line method 3)

- Consider the line found by solving the following Criterion function

$$\min_{b_0, b_1} \sum_{i=1}^{n} | Y_i - b_0 - b_1 X_i | = \min_{b_0, b_1} \sum_{i=1}^{n} | e_i |$$

- This is called Least Absolute Deviation
- Can we use Calculus to solve this?

R Can Do This

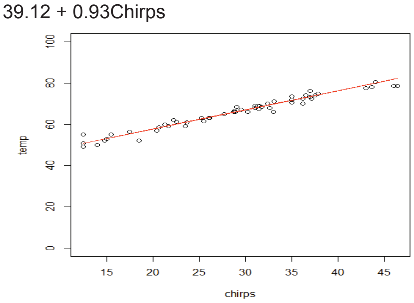
- The command is called `rq`

```
> library(quantreg)
> fit=rq(mydata$Temp~mydata$Chirps)
> summary(fit)

Call: rq(formula = mydata$Temp ~ mydata$Chirps)

tau: [1] 0.5

Coefficients:
               coefficients lower bd upper bd
(Intercept)    39.12652
mydata$Chirps   0.92988      0.81868  0.95973
```

IGNORE THIS STUFF FOR NOW

A Picture

Temp = 39.12 + 0.93Chirps



How wrong are we???????

Criterion Function 2 (line method 4)

- The most popular method of fitting a line to data is called the **least-squares method**, and involves solving the following problem

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (e_i)^2$$

- Because it is a continuous criterion function, calculus can be used to find the solution.

The values of b_0 and b_1 which minimize the residual sum of squares are:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = r \frac{s_y}{s_x}$$

Hmm relationship between r and b_1 –does that make sense?

$$b_0 = \bar{Y} - b_1 \bar{X}$$

These formulas can be derived using calculus- we pass (or not-depending on time).

These formulas are the intercept and slope for the "best fitting line".

Built into R

- These equations are built into R (and Excel and many other packages) and are what R uses when you call the `lm` command

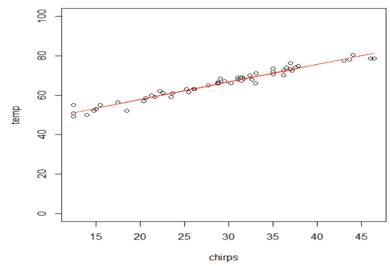
```
> fit=lm(mydata$Temp~mydata$Chirps)
> summary(fit)

Call:
lm(formula = mydata$Temp ~ mydata$Chirps)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    40.02525
mydata$Chirps   0.89180
```

IGNORE FOR NOW

Note: the `lm` commands gives a lot of output that I deleted for now.

Nice Picture

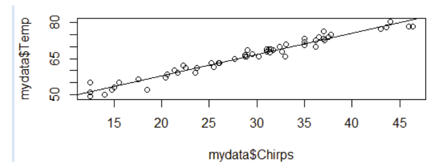
Temp = 40.02 + 0.89Chirps



How wrong are we???????

Getting the fitting line plot in R

```
> plot(mydata$Chirps,mydata$Temp)
> fit=lm(mydata$Temp~mydata$Chirps)
> abline(fit)
```

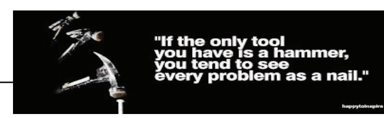


Summary

- Method 1: too stupid to use
- Method 2: temp = 39.8+0.9(chirps)
- Method 3: temp = 39.12+0.93(chirps)
- Method 4: temp = 40.02+0.89(chirps)
- Why do we care? If we get the same answer different ways-maybe it's a good answer.
- What if there are major differences? Which answer is correct? Hmmm.
- How wrong are we???

Least Squares and Your Toolbox

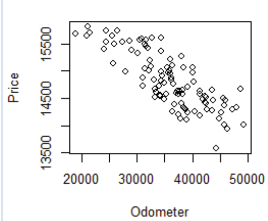
- People love least squares-it's the most popular way to fit a line to data, and the method we spend a lot of time examining in this course.
- But it has issues; it used to be the easiest to compute but that's not an issue anymore.
- Always remember



Regression Example

- Predicting used Honda Accord price from mileage:

```
> head(mydata)
  Price Odometer
1 14636    37388
2 14122    44758
3 14016    45833
4 15590    30862
5 15568    31705
6 14718    34010
```




```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/accordprices.csv")
```


Just getting the coefficients

```
> fit=lm(Price~Odometer)

> coef(fit)
  (Intercept)      Odometer
 1.706677e+04 -6.231548e-02
```

b_0 b_1

Lots of Regression Output in R

```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)
```

Ignore

```
Residuals:
    Min      1Q  Median      3Q     Max
-730.32 -235.01    1.31  187.75  691.25
```

Ignore

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.707e+04  1.690e+02  100.97   <2e-16 ***
Odometer    -6.232e-02  4.618e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b_0 (Intercept), b_1 (Odometer)

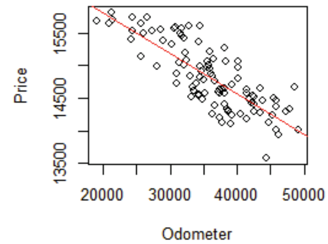
Ignore

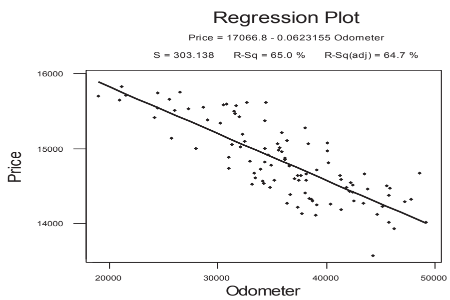
```
Residual standard error: 303.1 on 98 degrees of freedom
Multiple R-squared:  0.6501,    Adjusted R-squared:  0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

We will eventually explain this complete printout-ignore most of it for now.

Fitted Line Plot in R

```
> plot(Odometer,Price)
> abline(fit,col="red")
```





Interpretation of the slope: For each additional mile on the odometer, the price *decreases* by an average of $0.062

Do not interpret the intercept as cars that have not been driven cost $17066.8

Prediction

- The regression line says that

$$price = 17066.8 - 0.0623(odometer)$$

- So the predicted price for an Accord with 40000 miles is

$$price = 17066.8 - 0.0623(40000) = \$14574.8$$

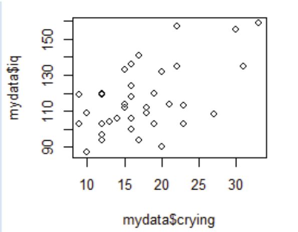
- How wrong are we???

Example: Crying Babies

- Babies who cry a lot may be more easily stimulated than other babies, and this may be an indication of higher IQ. Karelitz, et al. (1964) studied the association between IQ and crying frequency with 37 babies.
- The researchers caused the babies to cry by snapping a rubber band on the sole of their foot (bastards...).
- They recorded the frequency of cries as the number of peak cries (example: WAAAHHHHH-WAAAAHHHH is two peaks) in the most active 20 seconds of crying. Three years later, they measured the babies' IQs.

The data

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/crying.csv")
> names(mydata)
[1] "crying" "iq"
> plot(mydata$crying,mydata$iq)
```



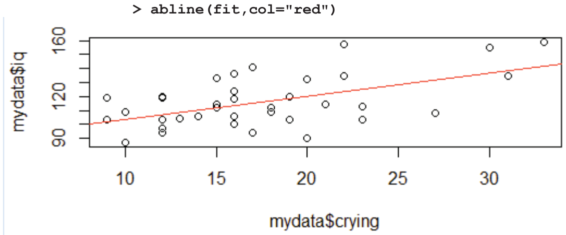
Regression Output

```
> fit=lm(mydata$iq~mydata$crying)
> summary(fit)

Call:
lm(formula = mydata$iq ~ mydata$crying)

Residuals:
    Min      1Q  Median      3Q     Max
-30.192  -9.791  -3.619  11.808  33.458

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    86.6898     7.9650  10.884 8.83e-13 ***
mydata$crying   1.6751     0.4313   3.884 0.000436 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 35 degrees of freedom
Multiple R-squared:  0.3012,    Adjusted R-squared:  0.2812
F-statistic: 15.09 on 1 and 35 DF,  p-value: 0.000436
```

Interpretation?

The Fitted Line Graph

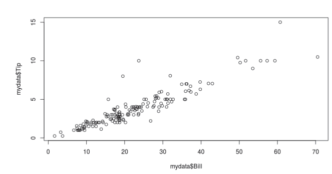
```
> plot(mydata$crying,mydata$iq)
> abline(fit,col="red")
```



Example: Restaurant Tips

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/RestaruantTips.csv")
```

Bill	Tip	Credit	Guests	Day	Server	PctTip
23.70	10.00	n	2	f	A	42.2
36.11	7.00	n	3	f	B	19.4
31.99	5.01	y	2	f	A	15.7
17.39	3.61	y	2	f	B	20.8
15.41	3.00	n	2	f	B	19.5
18.62	2.50	n	2	f	A	13.4
21.56	3.44	n	2	f	B	16.0
19.58	2.42	n	2	f	A	12.4
23.59	3.00	n	2	f	A	12.7

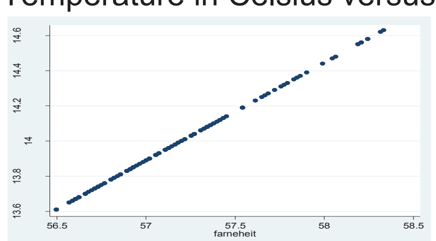


How do we interpret this output?

```
> describe(mydata$PctTip)
   vars   n  mean   sd median trimmed  mad min  max range skew kurtosis   se
X1    1 157 16.62 4.39   16.2   16.29 2.82 6.7 42.2  35.5 2.48    12.24 0.35


> coef(lm(mydata$Tip~mydata$Bill))
(Intercept) mydata$Bill
 -0.2922675   0.1822147
```


Example: Global Temperature

- Temperature in Celsius versus Fahrenheit.



Wait-do we even need regression here??

Regression Example: Market Model

In finance, a popular model is to regress stock returns against returns of some market index, such as the S&P 500.

The slope of the regression line, referred to as "beta", is a measure of how sensitive a stock is to movements in the market.

$$Stockreturn_t = \alpha + \beta Indexreturn_t$$

Market Model

$$Stockreturn_t = \alpha + \beta Indexreturn_t$$

Beta=0 : cash under the mattress

Beta=1 : same risk as the market

0<Beta<1 : safer than the market

Beta >1: riskier than the market

Beta < 0 : what would this mean???

These 5 stocks are strictly for the bulls

Alex Rosenberg | @AcesRose
Wednesday, 24 Feb 2016 | 12:08 PM ET

CNBC

If you think the S&P is primed to bounce, you might want to take a look at high-beta stocks.

These are the names that tend to track the S&P most excitedly — climbing the most when the market rises and falling the furthest when the market skids.

Given that stocks as a whole have suffered this year, it's little surprise that these high-beta stocks have had an especially rough go of it.

Dividing all of the stocks in the S&P 1500 with market values above $500 million into quintiles based on their beta measures, one finds that the average 2016 performance for the highest-beta stocks is a 13.7 percent drop. Meanwhile, the average stock in the lowest-beta quintile of the market is up 1.2 percent, based on a CNBC analysis of figures from FactSet.

STOCKS FOR THE BULLS
SUPER-HIGH-BETA STOCKS

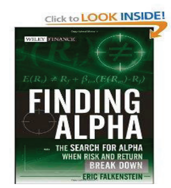
COMPANY	YTD PERF.	BETA
Four Corners	-32.4%	9.1
Lannett Co.	-36.7%	3.3
Wisdomtree	-25.4%	3.2
TripAdvisor	-26.0%	3.1

The Search for Alpha

- In the market model, what is the stock return if the index does nothing?

$$Stockreturn_t = \alpha + \beta Indexreturn_t$$

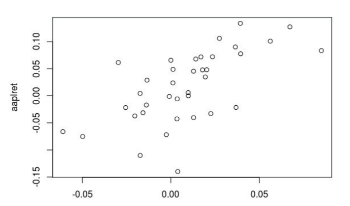
People talk about "buying someone's alpha"; i.e. what does the fund manager bring to the table above the index returns.



Finding the beta for $AAPL

- $AAPL is Apple

Beta is usually calculated using three years of monthly returns



Getting stock data into R (advanced)

```
> library(quantmod)

> getSymbols("AAPL",from="2014-06-01")
[1] "AAPL"
> getSymbols("SPY",from="2014-06-01")
[1] "SPY"

> aaplret=as.numeric(monthlyReturn(Ad(AAPL)))
> spyret=as.numeric(monthlyReturn(Ad(SPY)))

> plot(spyret,aaplret)
```

Using Regression to find Beta

```
> fit=lm(aaplret~spyret)
> coef(fit)

(Intercept) spyret
0.004931447 1.416238655
```

The Beta for $AAPL is 1.42; its considered “riskier” than the market

Compare with finance.yahoo.com

Apple Inc. (AAPL)
NasdaqGS - NasdaqGS Delayed Price. Currency in USD
Add to watchlist

158.63 -2.63 (-1.63%)
At close: September 8 4:00PM EDT

Summary | Conversations | Statistics | Profile | Financials

Previous Close	161.26	Market Cap	819.36B
Open	160.86	Beta	1.43
Bid	158.58 x 100	PE Ratio (TTM)	18.01
Ask	158.65 x 1400	EPS (TTM)	8.81
Day's Range	158.53 - 161.15	Earnings Date	Oct 23, 2017 - Oct 27, 2017
52 Week Range	102.53 - 164.94	Dividend & Yield	2.52 (1.56%)
Volume	28,611,535	Ex-Dividend Date	2017-08-10
Avg. Volume	26,685,771	1y Target Est	171.41

Trade prices are not sourced from all markets

There are no Stat Police

- Beta depends on the data you use!

Apple Inc. (NASDAQ:AAPL)

158.63 -2.63 (-1.63%)
Sep 8 - Close
NASDAQ real-time data - Disclaimer
Currency in USD

Range	158.53 - 161.15	Div/yield	0.63/1.59
52 week	102.53 - 164.94	EPS	8.79
Open	160.86	Shares	5.17B
Vol / Avg.	28.61M/27.21M	Beta	1.30
Mkt cap	819.36B	Inst. own	63%
P/E	18.04		

From google finance

Today’s Tools

- New toolbox additions
  - Covariance and Correlation-measures of association
  - Fitting a line to data (least squares method)

Things you should know

- The least squares estimates
- There are many ways to fit a line to data
- The least squares method is the most popular way to fit a line to data
- The Market Model is a popular model in finance for assessing the risk of a stock relative to the market as a whole.