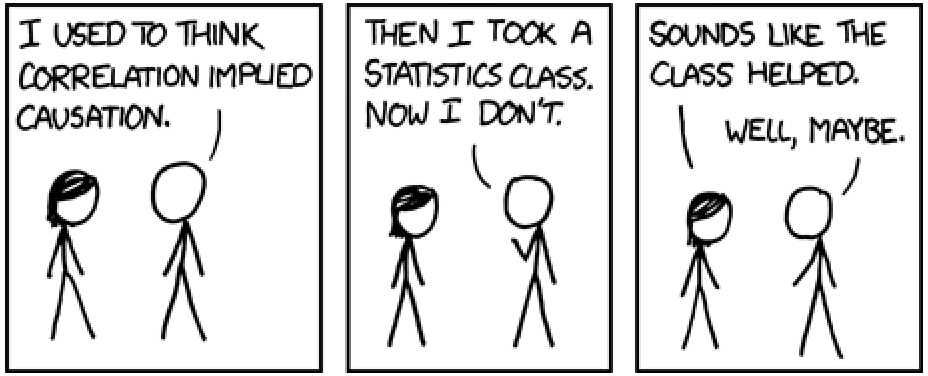
Introduction to Regressions

Cause and Effect

Correlation: Two variables are correlated when changes in one variable occur in a pattern corresponding to changes in the other.

Cause and Effect

Causation: One variable moves, and the second variable changes because of the movement of the first.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Questioning Causality

When we suspect a causal relationship (that x causes y), it is important to ask ourselves several questions:

1. Is it possible that y causes x instead?
2. Is it possible that z (a new factor that we haven't considered before) is causing both x and y?
3. Could the relationship have been observed by chance?

Establishing Causality

In order to establish causality, we need to meet several conditions:

- We can explain **why** x causes y
- We can demonstrate that **nothing else is driving the changes** (within reason)
- We can show that there is a **correlation** between x and y

Ceteris Paribus

ceteris paribus means "all else equal"

Why I stink at golf

Why am I always in the sand trap?

- Need to isolate the variables

Why I stink at golf

- Is it my club?
- My swing?
- The wind? (definitely the wind)

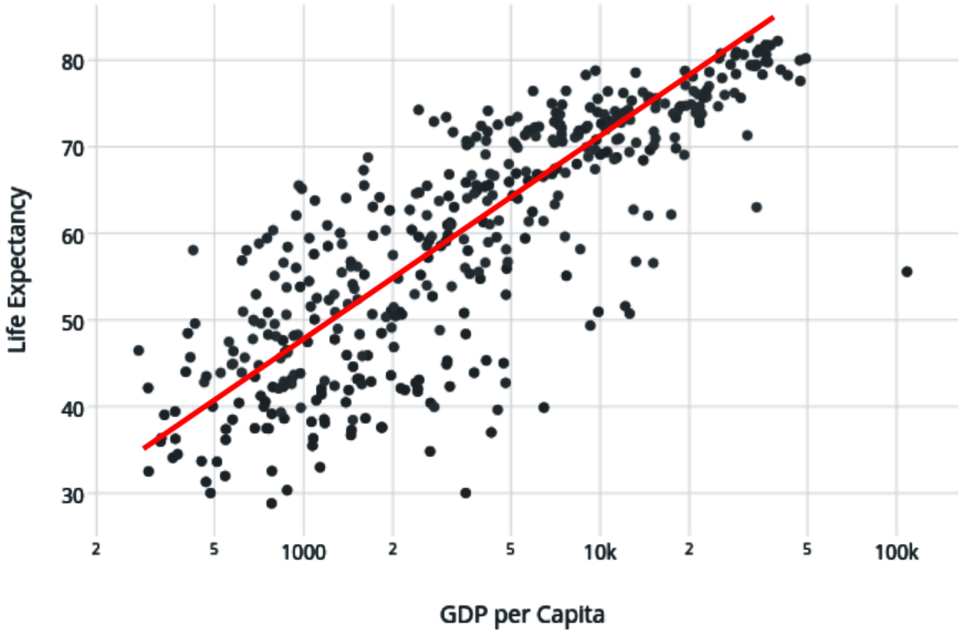
To uncover the effect

Swing my club 100 times with each golf club

- Keeping the wind, my stance, my swing, etc. consistent
- Is that even really possible?
- In many cases, no

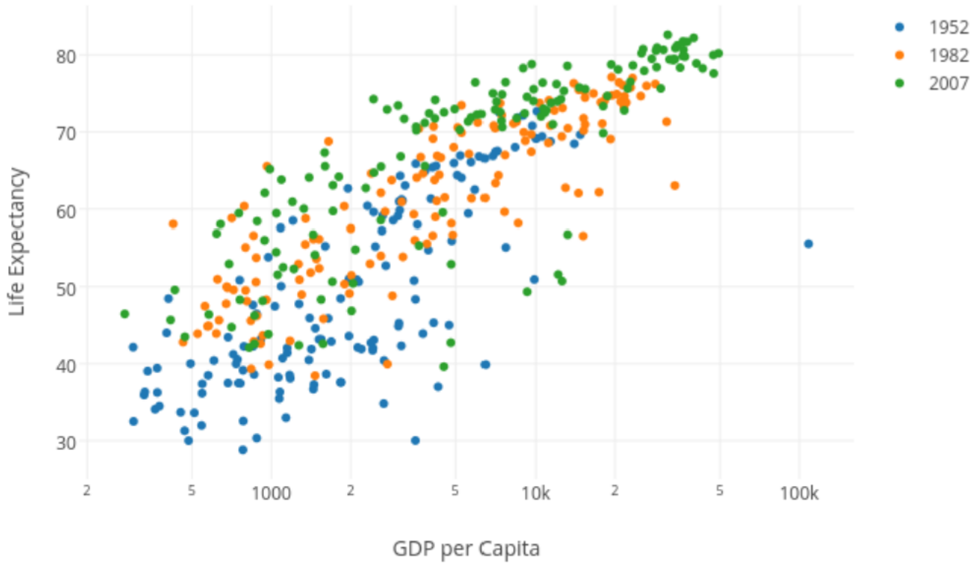
Regression analysis

- Allows us to **act as if nothing else were changing**
- Mathematicaly isolates the effect of each individual **variable** on the outcome of interest
    - Variables are the factors that we want to include in our model



Regression analysis

- Think about it like a trend line!

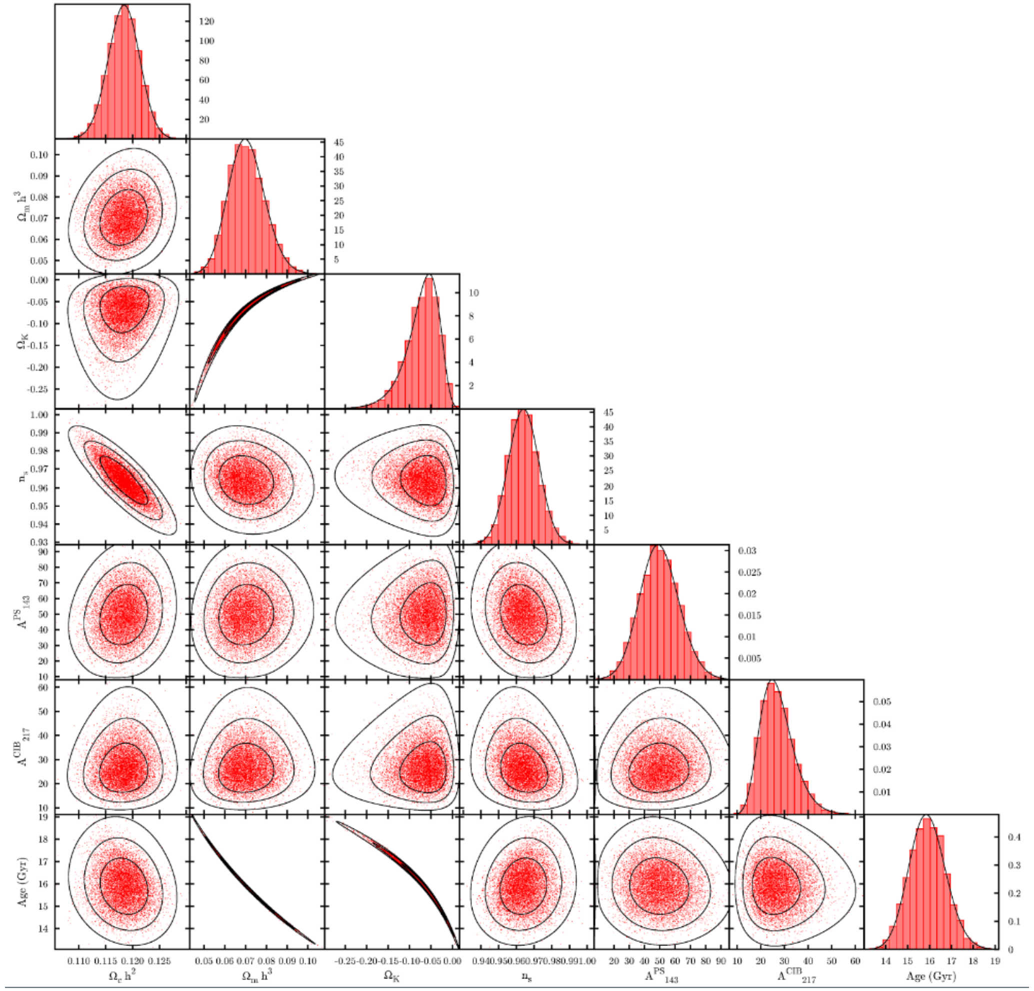


Regression analysis

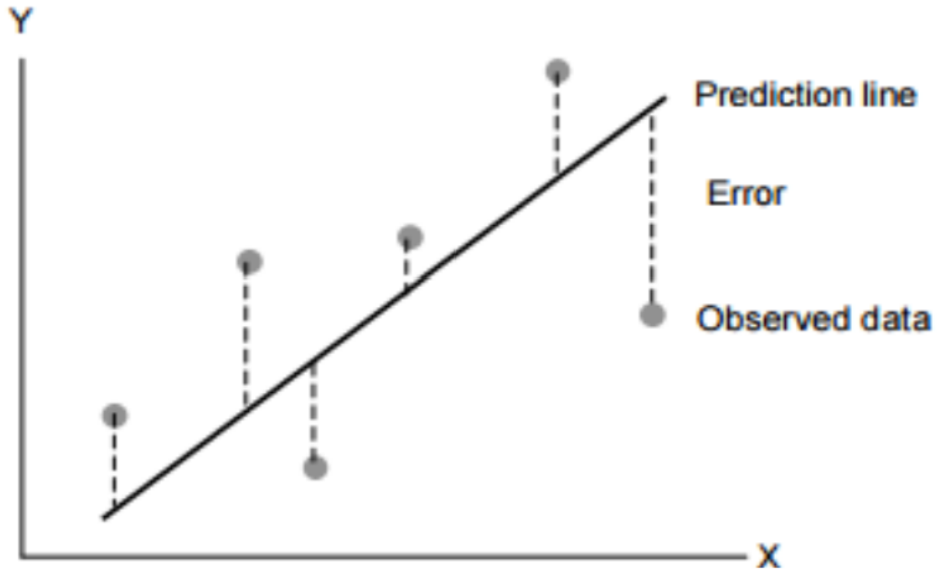
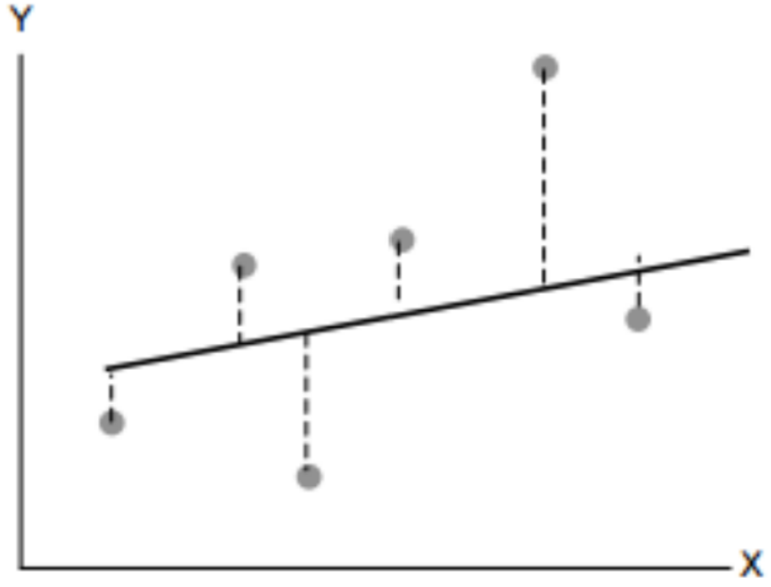
Whoops! What if there is another variable?

Regression analysis

Or lots of variables??

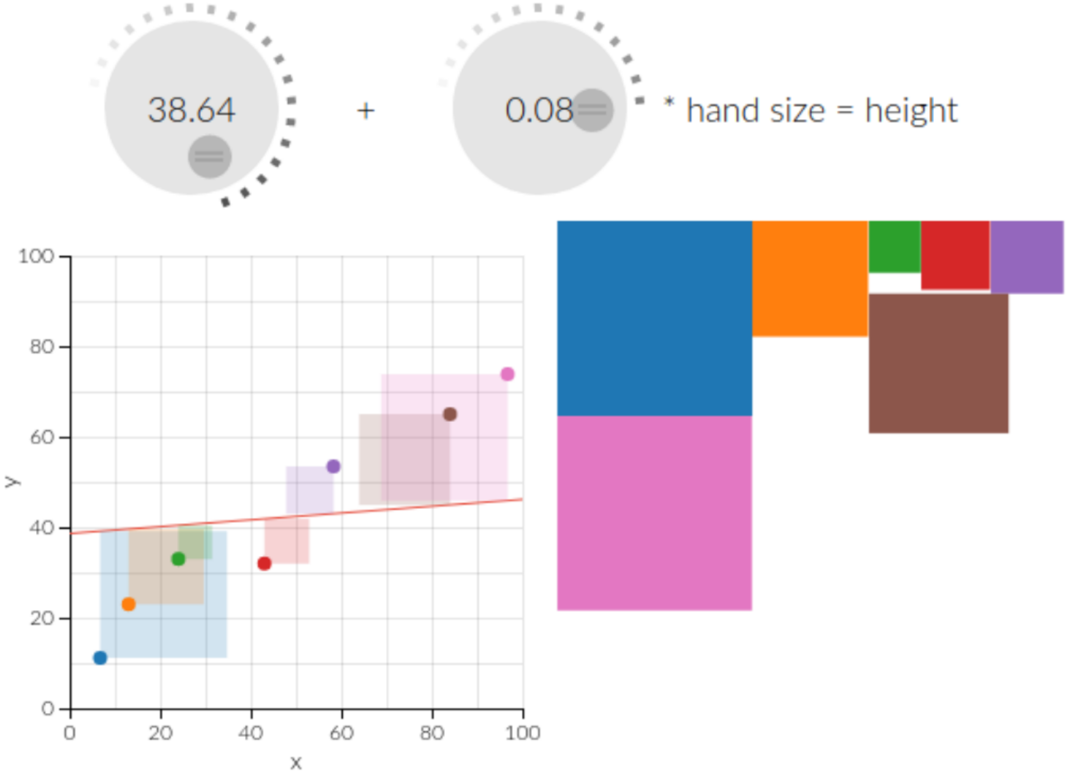


Minimize Errors and Best Fit Lines

Best Fit	Something Else
	

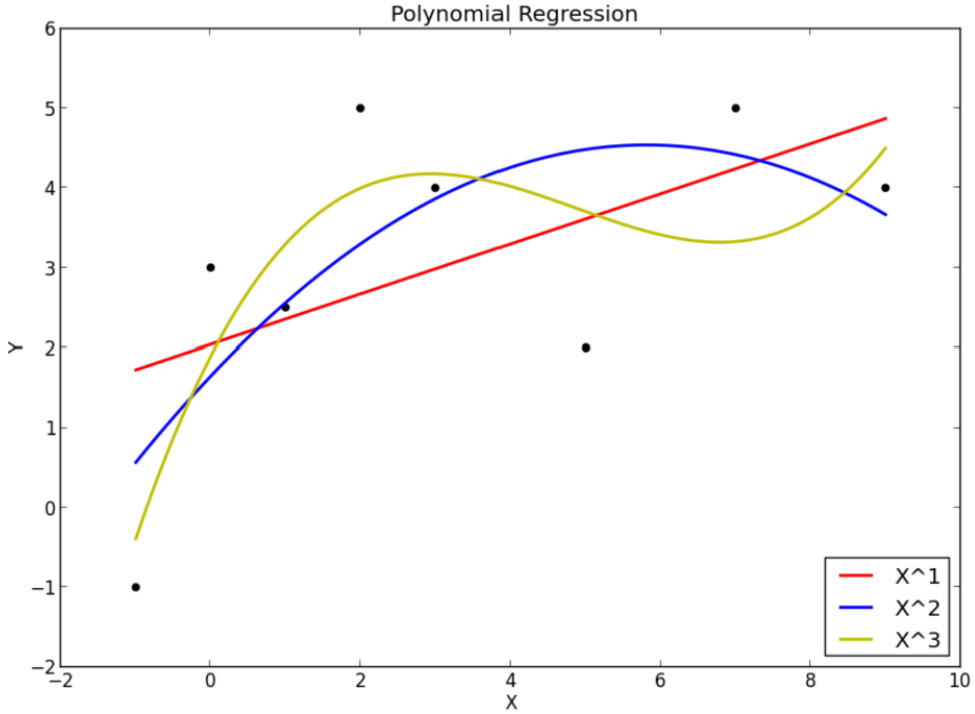
Minimize Errors and Best Fit Lines

Try it by hand!



Why LINEAR regression?

- Faster
- More honest



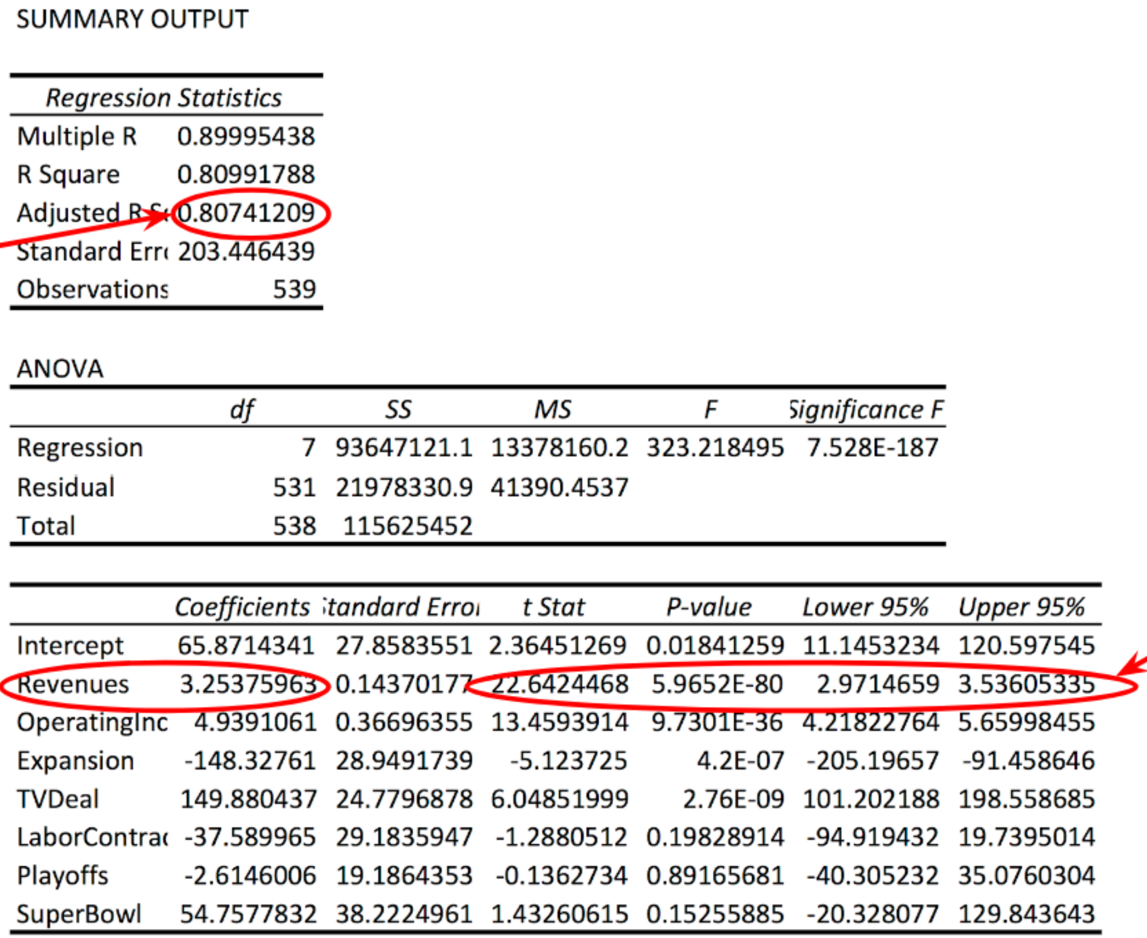
Regression in Excel - Disclaimer

First, when doing regression in the real world™, don't use Excel.

- If you need to do real regressions for a project, let me know and we can talk about appropriate tools

Now that we have that out of the way, let's do regression in Excel!

Regression in Excel



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89995438
R Square	0.80991788
Adjusted R S	0.80741209
Standard Err	203.446439
Observations	539

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	93647121.1	13378160.2	323.218495	7.528E-187
Residual	531	21978330.9	41390.4537		
Total	538	115625452			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	65.8714341	27.8583551	2.36451269	0.01841259	11.1453234	120.597545
Revenues	3.25375963	0.14370177	22.6424468	5.9652E-80	2.9714659	3.53605335
OperatingInc	4.9391061	0.36696355	13.4593914	9.7301E-36	4.21822764	5.65998455
Expansion	-148.32761	28.9491739	-5.123725	4.2E-07	-205.19657	-91.458646
TVDeal	149.880437	24.7796878	6.04851999	2.76E-09	101.202188	198.558685
LaborContra	-37.589965	29.1835947	-1.2880512	0.19828914	-94.919432	19.7395014
Playoffs	-2.6146006	19.1864353	-0.1362734	0.89165681	-40.305232	35.0760304
SuperBowl	54.7577832	38.2224961	1.43260615	0.15255885	-20.328077	129.843643

How well our model works (between 0 and 1)

The effect of a one unit change in Revenues

Whether or not effect is significant (or should be attributed to chance)

Regression terms

- **Coefficient**: This is the effect of changing a variable by one unit (from “untreated” to “treated”)
- **Standard Error (Standard Deviation)**: Measures how noisy the effect of the independent variable is on the dependent variable
    - Larger numbers indicate more noise

Regression terms

- **Confidence Interval**: Assuming our regression analysis includes all relevant information, we expect that the true coefficient (treatment effect) lies within this range 95% of the time (for a 95% confidence interval)
- **Statistical Significance**: When the Average Treatment Effect has a confidence interval (at 95% or 99% levels, typically) that does not include 0

What we assume

1. Effects are Linear (there are some workarounds)
2. Errors are normally distributed (bell-shaped)
3. Variables are not Collinear
4. No Autocorrelation (problematic for time series data)
5. Homoskedasticity (errors are shaped the same across all observations)

What we assume

All of these assumptions can be modified, but not by Excel. We **almost always** violate at least one assumption with any given dataset

When should we use regression, then?

- Regression Analysis is most useful when you care about WHY
- If you want to just predict WHAT will happen next, we have better tools for you!
  (We will spend the rest of the course looking at them)

For Lab

Work with your group to analyze your data from previous labs using regression analysis. Use the scientific method:

- Write down your hypothesis (what you believe should be the relationship between variables and why you think that is true)
- Organize the data
- Implement the regression(s)
- Decide whether or not the regression results support your hypothesis, and what this means for your conclusions and visuals