LEDE Algorithms

Richard Dunks

Chase Davis

DO NOW

lede_algorithms/class2_2/DoNow_2-2.ipynb

WEEK 2
CLASS 2

Goals for today

- Review and practice calculating coefficient of correlation and coefficient of determination
- Review linear regression
- Practice performing linear regression in Python

Outcomes for today

- You will practice calculating and analyzing the coefficients of correlation and determination in Python with real-world data sets
- You will have practice creating a linear regression model in Python based on real-world data

TO REVIEW FROM LAST WEEK

Measures of Central Tendency

- Quantitative data tends to cluster around some central value
- Contrasts with the spread of data around that center (i.e. the variability in the data)
- Measurements
  - Mean is a more precise measure and more often used
  - Median is better when there are extreme outliers
  - Mode is used when the data is categorical (as opposed to numeric)

Measures of Central Tendency

- Mean

```
df.mean()
```

- Median

```
df.median()
```

- Mode

```
df.mode()
```

Measures of Variability

- Describe the distribution of our data
- Measures
  - Range (Maximum – Minimum)

```
df['Height'].max() - df['Height'].min()
```

  - Standard Deviation

```
df.std()
```

  - Variance

```
df.var()
```

  - Inter-quartile Range

```
df['height'].quantile(q=0.75) - df['height'].quantile(q=0.25)
```

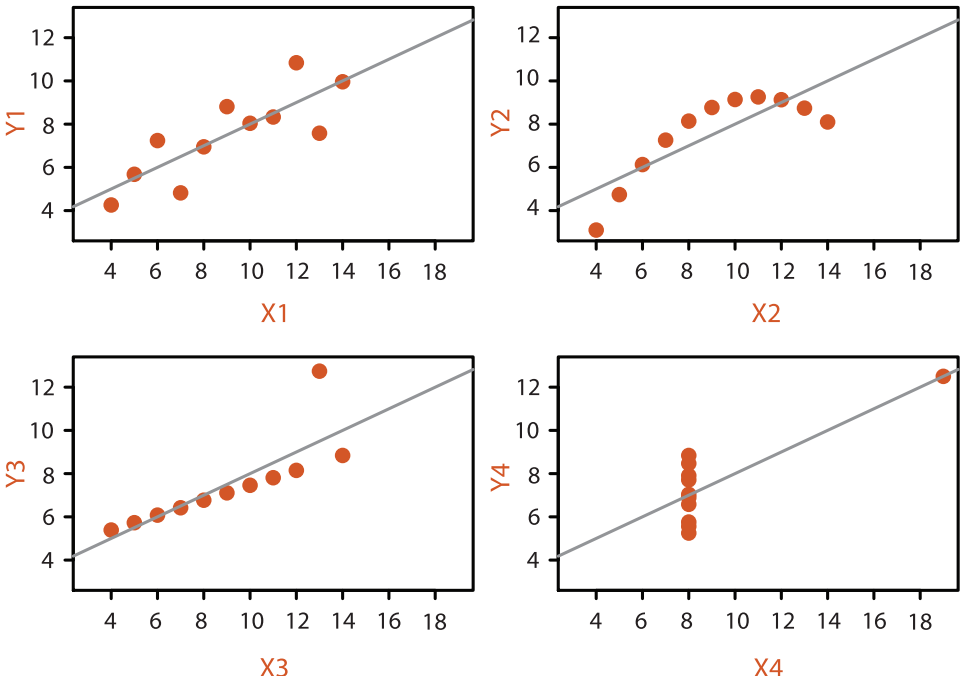
Descriptive Statistics

```
df.describe()
```

	height	weight
count	19.000000	19.000000
mean	62.336842	100.026316
std	5.127075	22.773933
min	51.300000	50.500000
25%	58.250000	84.250000
50%	62.800000	99.500000
75%	65.900000	112.250000
max	72.000000	150.000000

Anscombe's Quartet

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



Exploratory Data Analysis

- Goal -> Discover patterns in the data
- Approach
  - Understand the context
  - Summarize fields
  - Use graphical representations of the data
  - Explore outliers

Tukey, J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

CORRELATION

Correlation

- Values tend to have a relationship
- That relationship can be of several types
  - Proportional (increase in one increases the other)
  - Inversely proportional (increase in one decreases the other)
- Example
  - Height and weight

```
import pandas as pd
%matplotlib inline
```

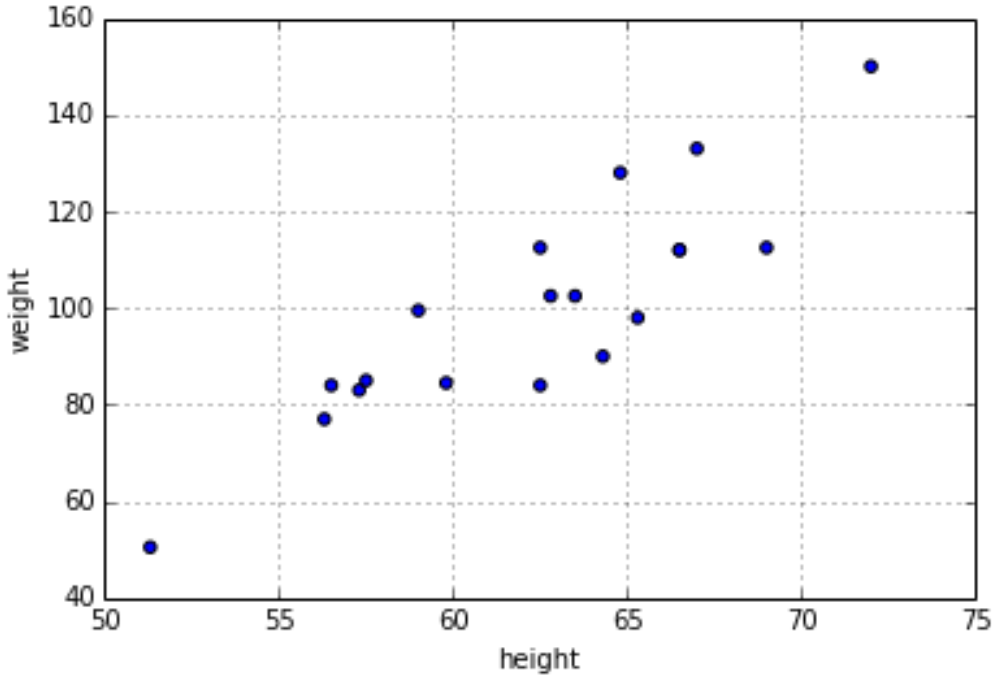
```
df = pd.read_excel("height_weight.xlsx")
```

Scatter Plot

- Plots relationship between two continuous variables

```
df.plot(kind="scatter",x="height",y="weight")
```

```
<matplotlib.axes.AxesSubplot at 0x109dec650>
```

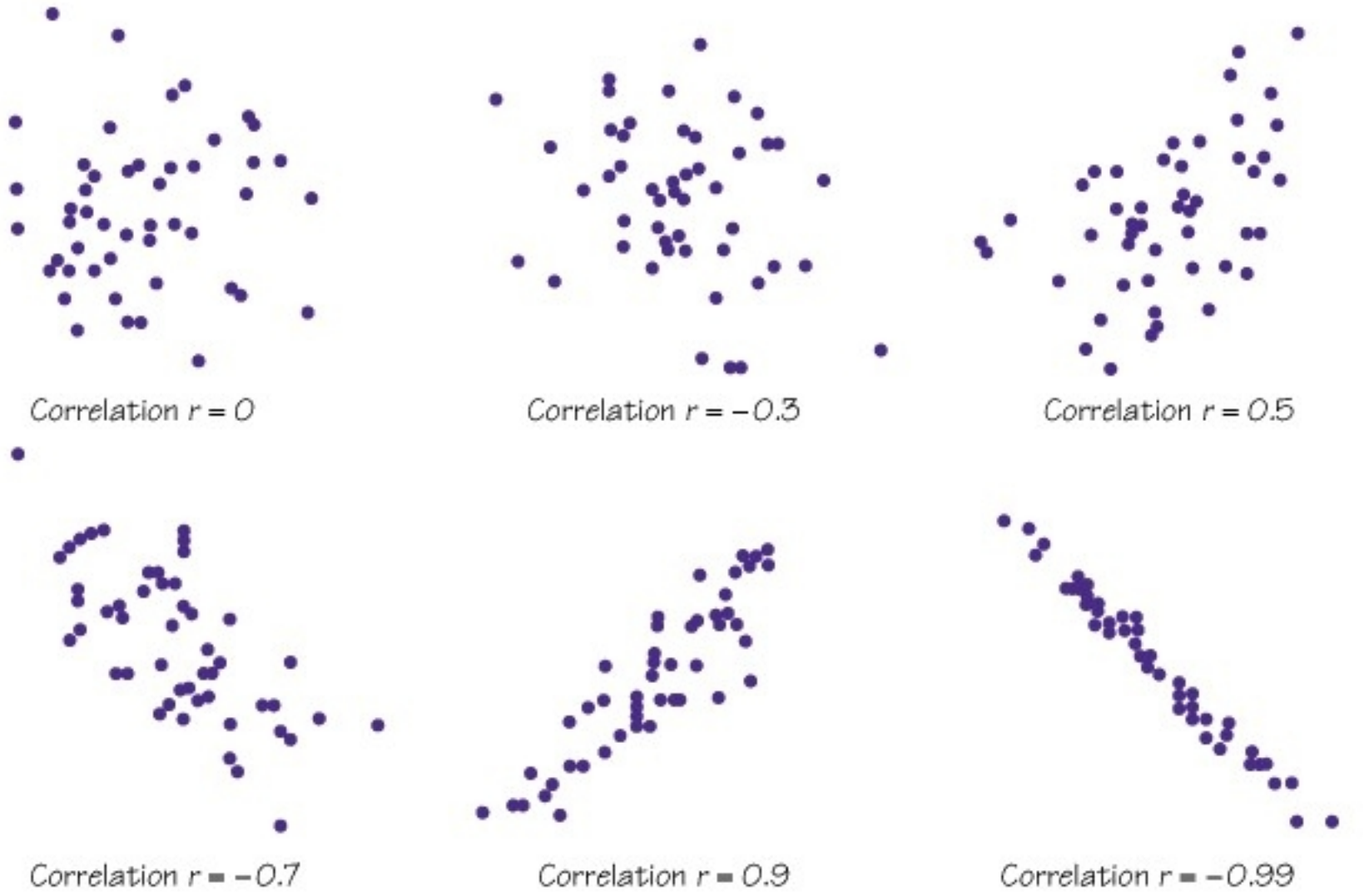


	name	height	weight
0	Joyce	51.3	50.5
1	Louise	56.3	77.0
2	Alice	56.5	84.0
3	James	57.3	83.0
4	Thomas	57.5	85.0
5	John	59.0	99.5
6	Jane	59.8	84.5
7	Jeffrey	62.5	84.0
8	Janet	62.5	112.5
9	Carol	62.8	102.5
10	Henry	63.5	102.5
11	Judy	64.3	90.0
12	Robert	64.8	128.0
13	Barbara	65.3	98.0
14	Mary	66.5	112.0
15	William	66.5	112.0
16	Ronald	67.0	133.0
17	Alfred	69.0	112.5
18	Philip	72.0	150.0

Correlation Coefficient

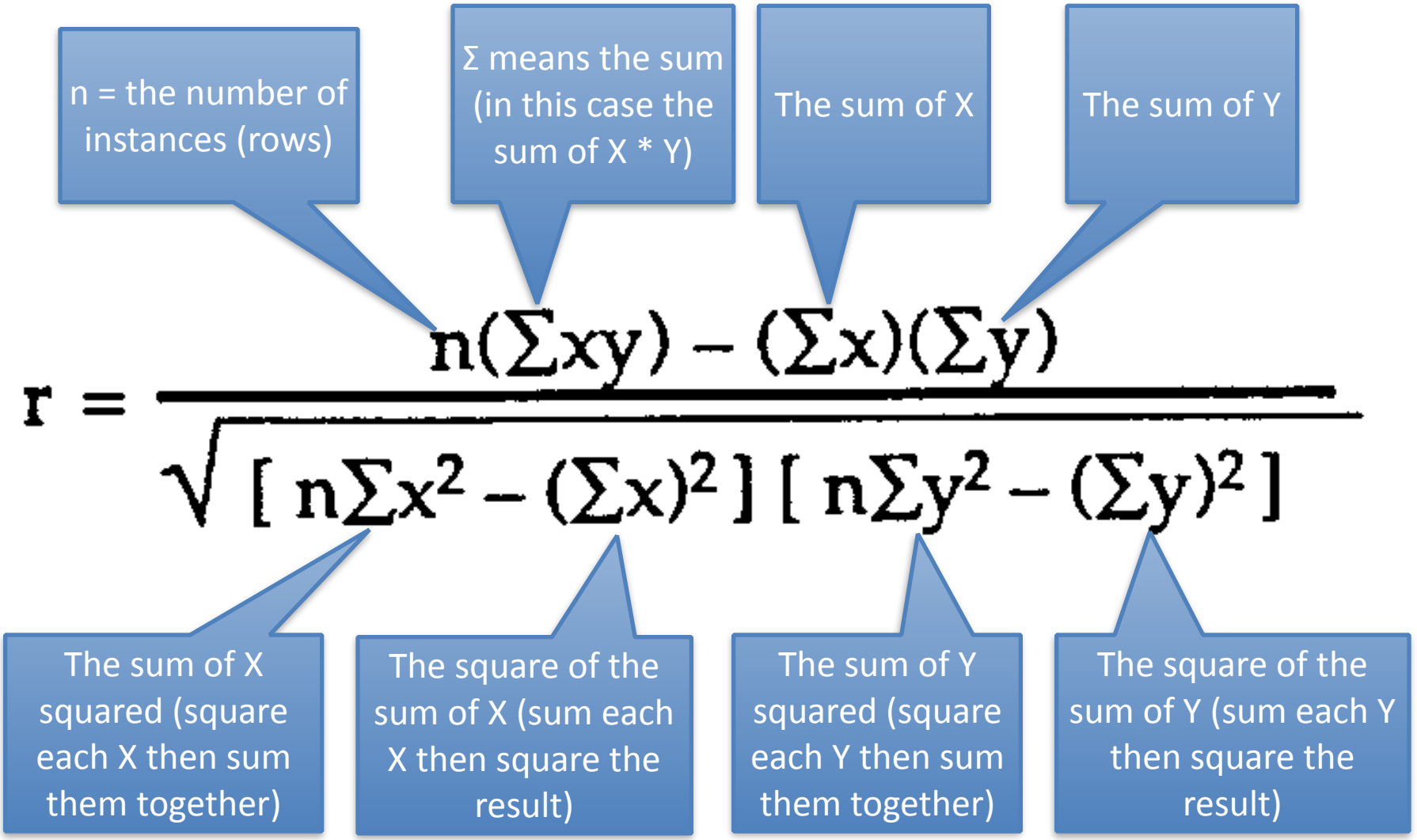
- Quantifies the amount of shared variability between variables
- Ranges between -1 and +1
  - Negative numbers are inversely proportional
  - Positive numbers are directly proportional
  - The closer to either -1 or +1, the greater the correlation

Correlation Coefficient



http://pixshark.com/correlation-examples.htm

Calculating the Correlation Coefficient



https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

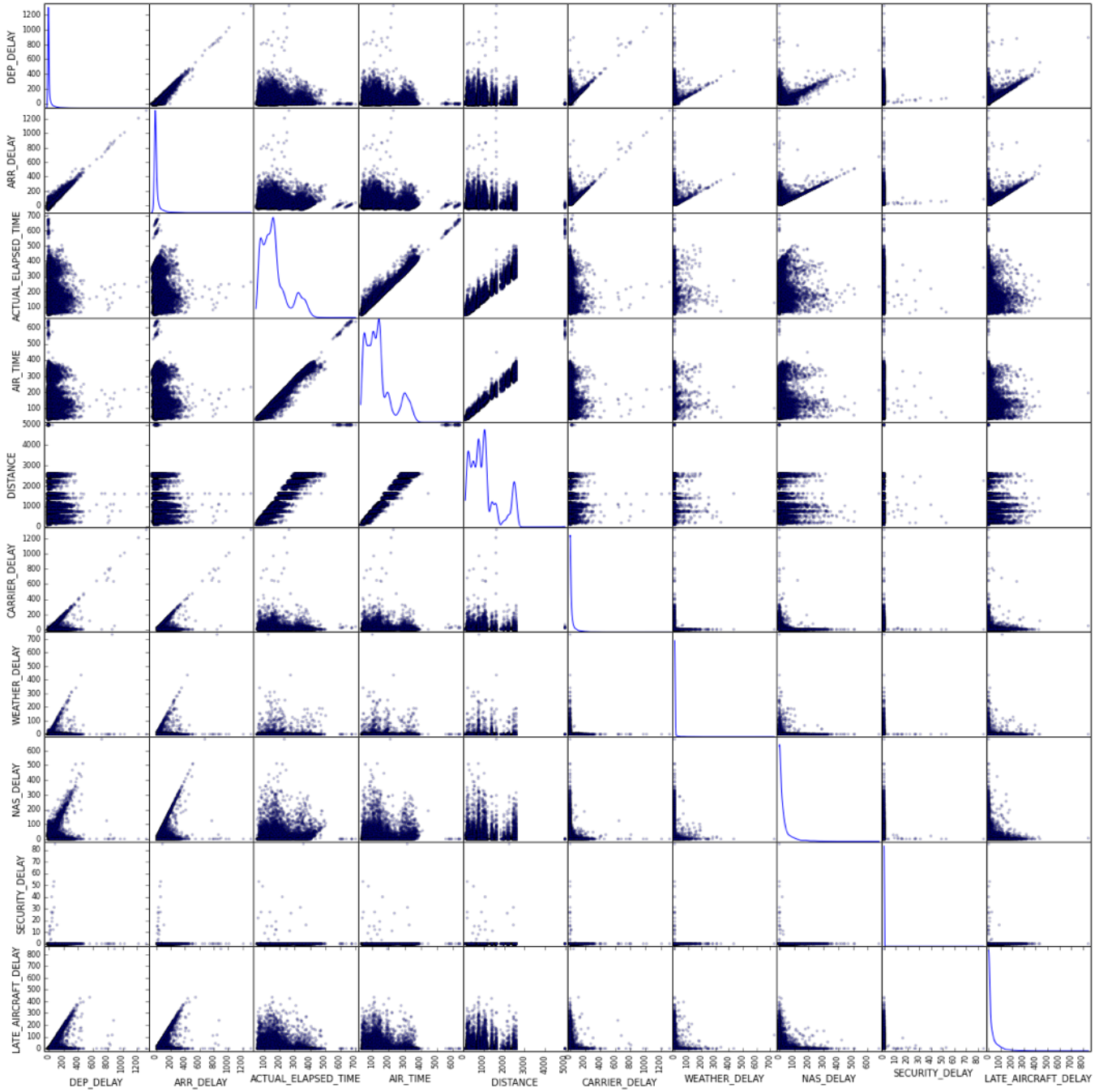
http://www.statisticshowto.com/what-is-the-correlation-coefficient-formula/

Or we could use Python

```
df[['height','weight']].corr()
```

	height	weight
height	1.000000	0.877785
weight	0.877785	1.000000

SPOTTING CORRELATIONS



http://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-scatter-matrix

Correlation of Determination

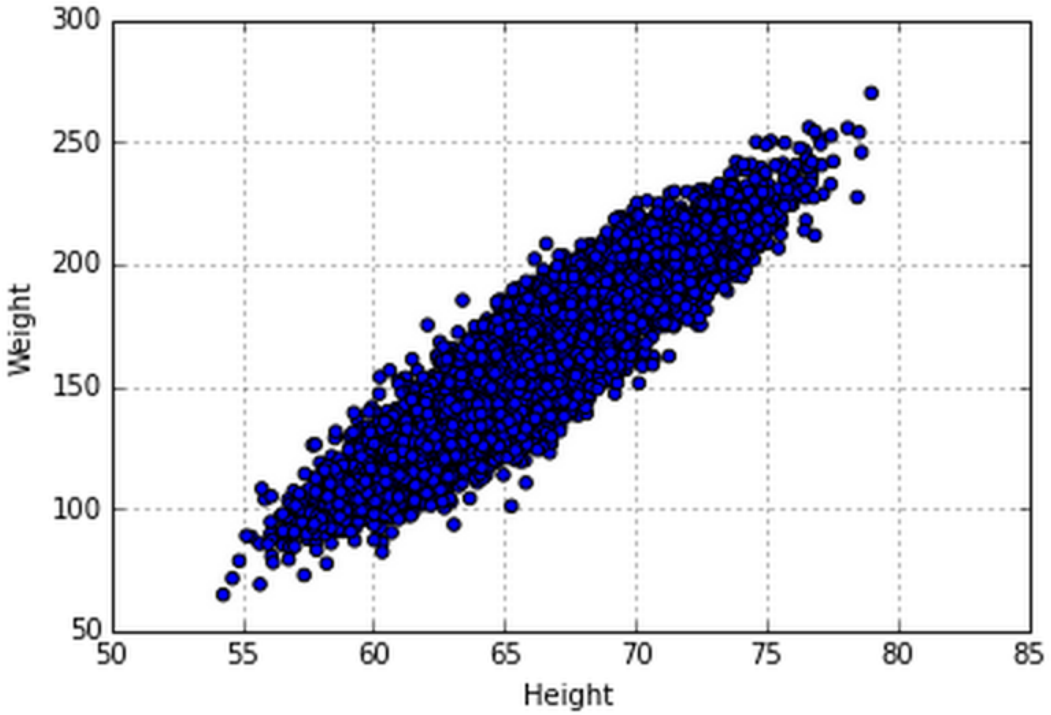
- The percentage of variance in one variable shared with the other
- More shared variability implies a stronger relationship
- Calculate by squaring the correlation coefficient
  - Ex. the correlation of determination for our height and weight dataset is 0.77


YEEEAAAAAH
I'M GONNA NEED YOU TO BRING MORE DATA
memegenerator.net

```
df2 = pd.read_csv("heights_weights_genders.csv")
```

```
df2.plot(kind="scatter",x="Height",y="Weight")
```

```
<matplotlib.axes.AxesSubplot at 0x107c93cd0>
```




```
df2[["Height","Weight"]].corr()
```

	Height	Weight
Height	1.000000	0.924756
Weight	0.924756	1.000000

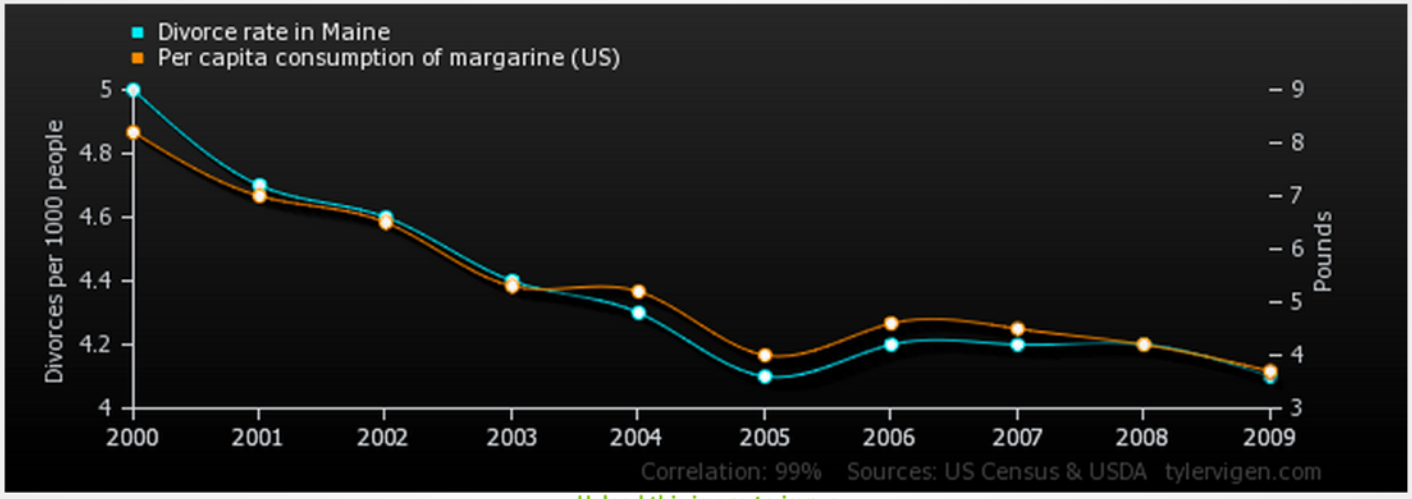
WHAT IS THE ADVANTAGE OF MORE DATA?

spurious correlations

Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

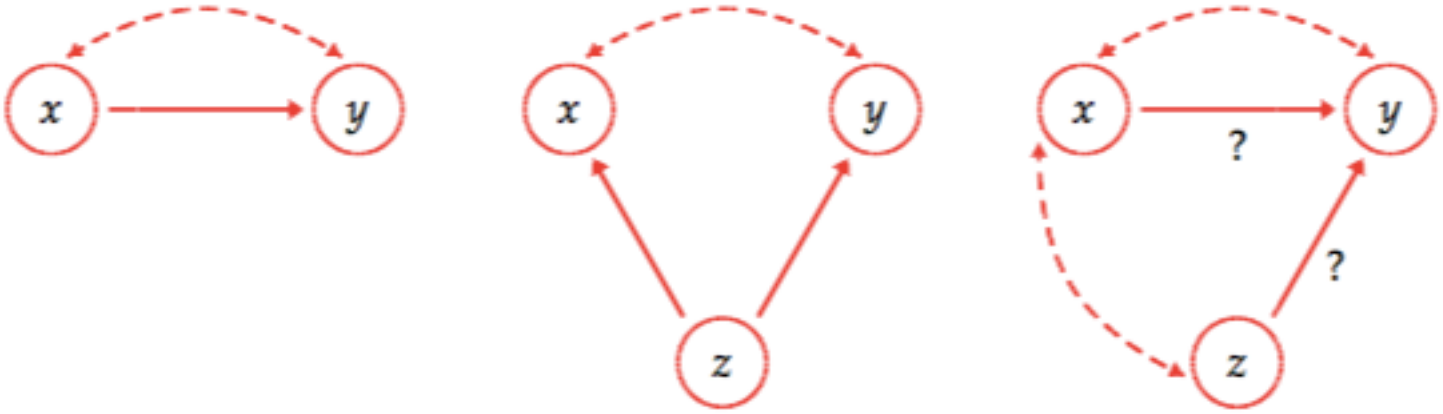
Correlation: 0.992558

Permalink - Mark as interesting (9,989) - Not interesting (3,911)
View all correlations - Discover a new correlation

Re-Chart

http://tylervigen.com/view_correlation?id=1703

Causation



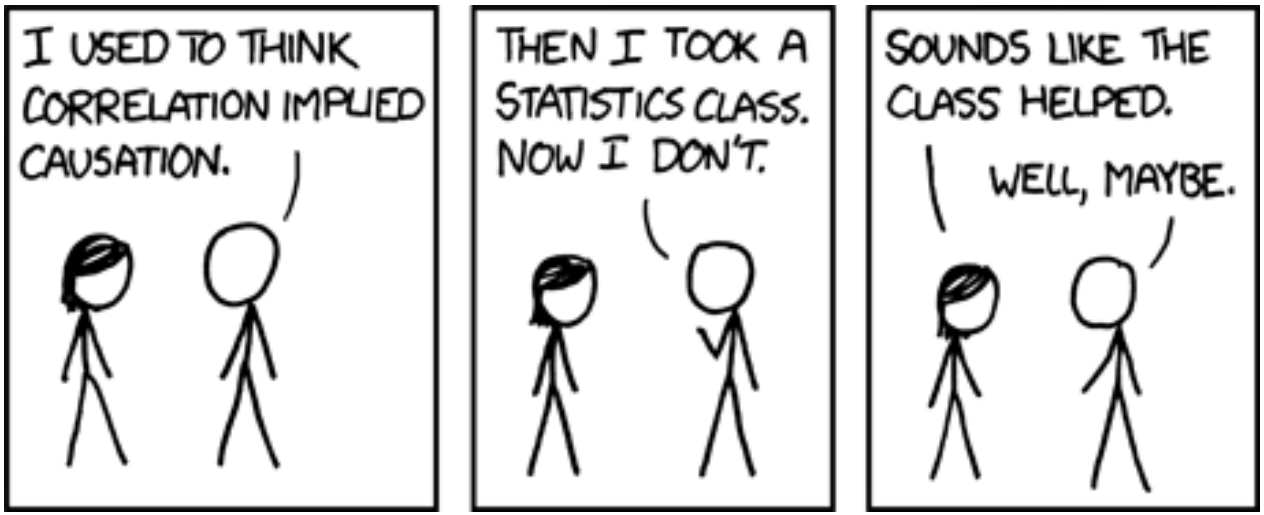
Causation
(a)

Common response
(b)

Confounding
(c)

Correlation does not imply causation

10 MIN BREAK



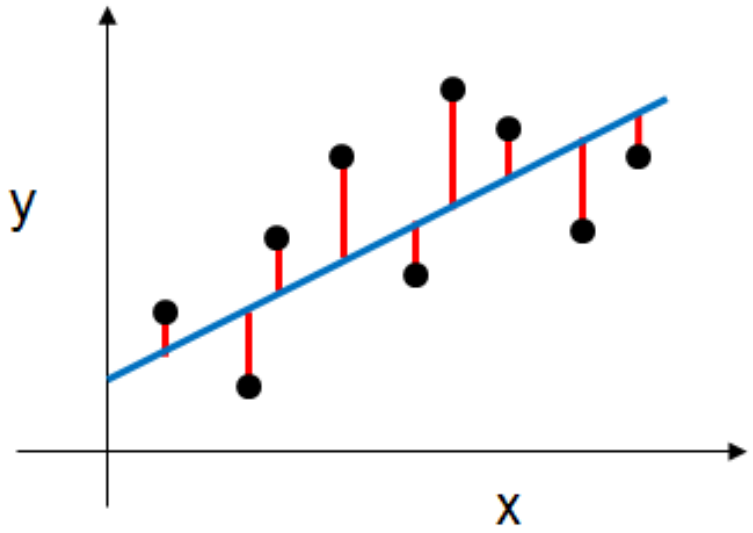
https://xkcd.com/552/

Linear Regression

- Using the known relationship between continuous variables, we can predict unseen values
- Assumes relationship is linear (but doesn't need to be)

Linear Regression

- Draw a line that minimizes the distance between each point
  - "Line of best fit" -> minimizes the sum of squared residuals



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

http://nbviewer.ipython.org/github/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb

Linear Regression

- Characteristics of the line defines the relationship
  - Slope -> relationship between independent and dependent variable (how Y increases per unit of X)
  - Intercept -> expected mean value of Y at X=0
- Values along the line are the predicted values for any given value X

Formula for a Line

$$y = mx + b$$

slope y-intercept

$$y = 3x - 5$$

slope y-intercept

http://www.algebra-class.com/slope-formula.html

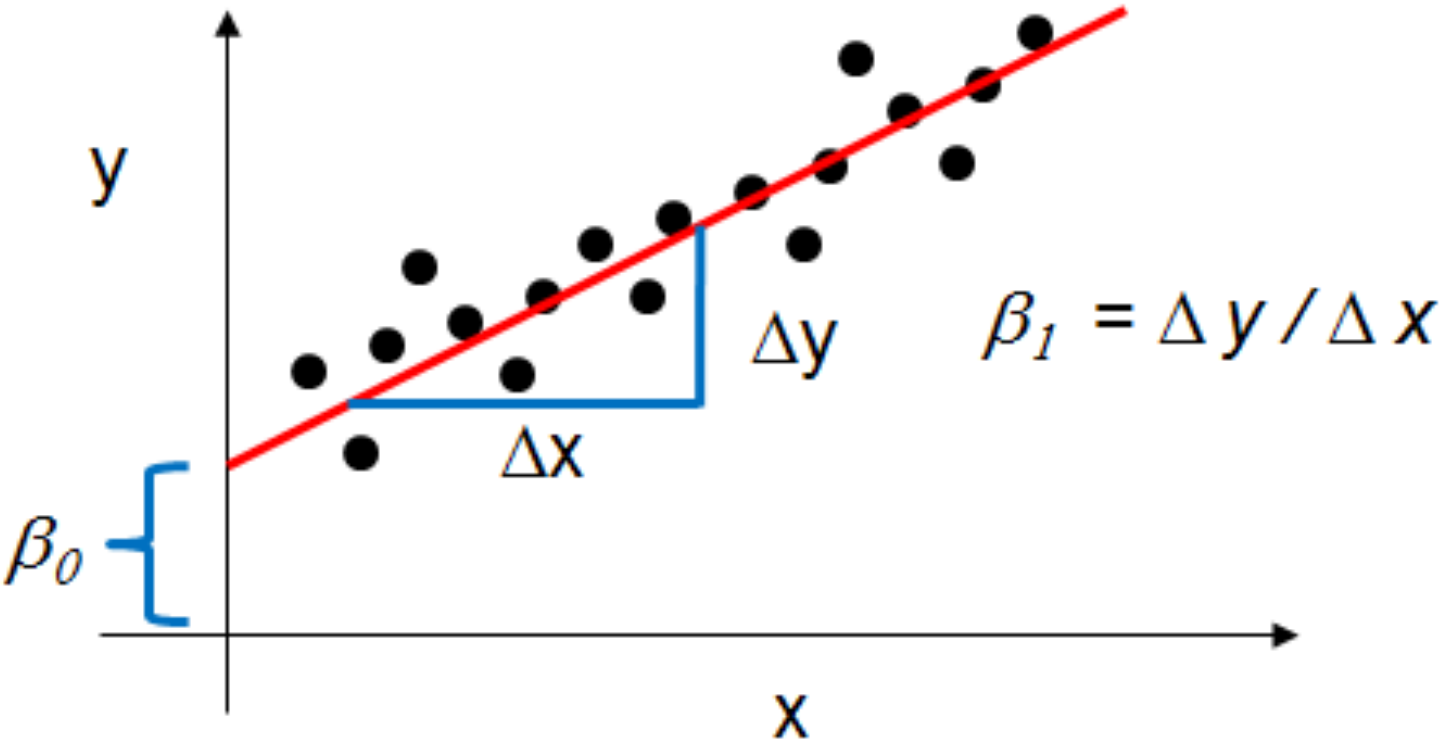
Linear Regression Line

$$y = \beta_0 + \beta_1 x$$

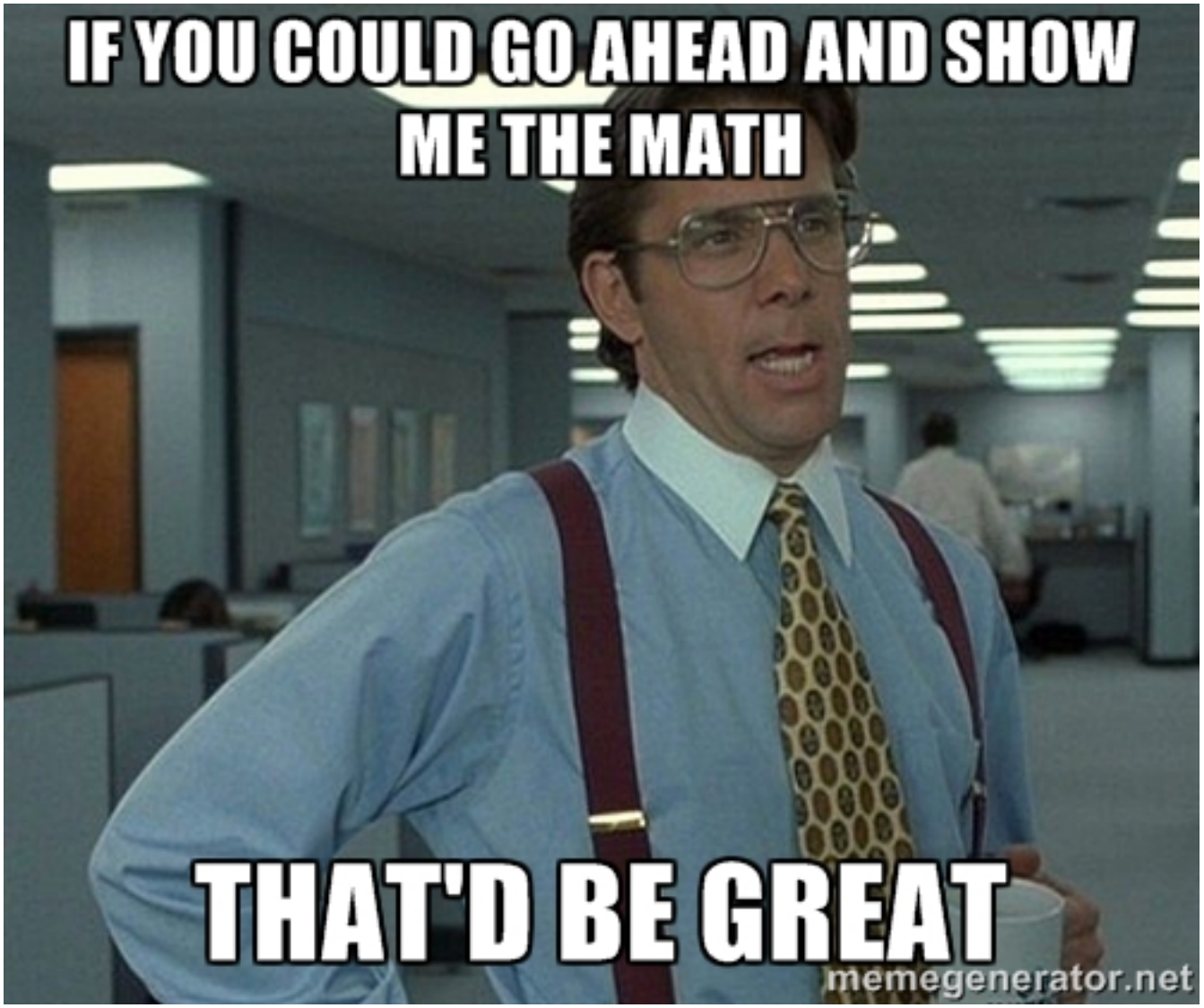
What does each term represent?

- y is the response
- x is the feature
- β_0 is the intercept
- β_1 is the coefficient for x

Slope



http://nbviewer.ipython.org/github/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb



Andrew Ng Machine Learning Class: http://cs229.stanford.edu/notes/cs229-notes1.pdf

For the Rest of Us

- The fit function calculates the coefficients through a relatively simple calculation that approximates the values very closely
- This makes the calculation relatively quick, even for larger datasets

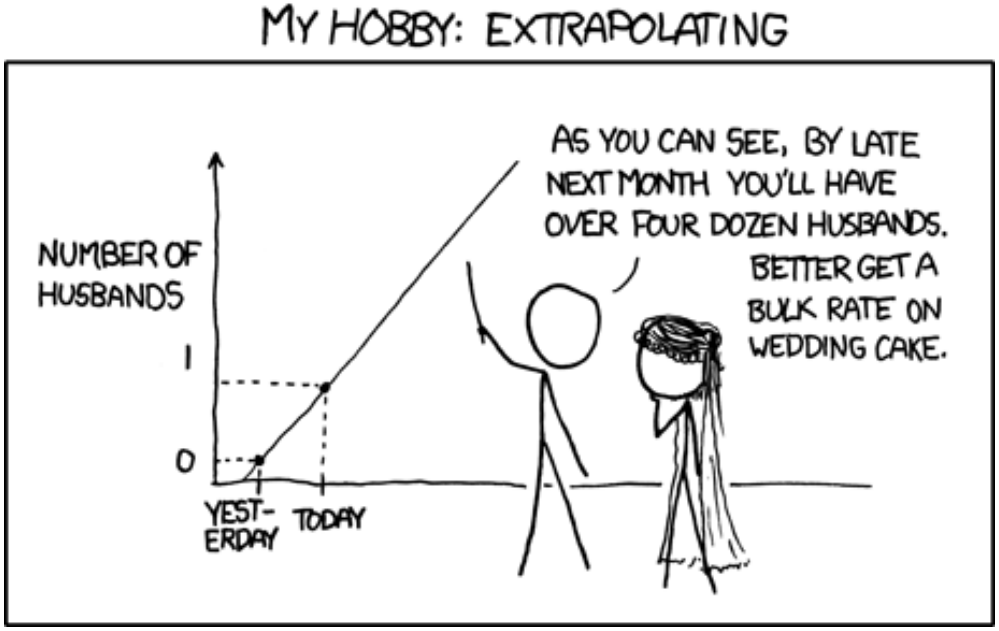
Prediction

- Using the relationship between variables, we can predict values based on the relationship
- Can estimate the magnitude as well as the general trend
- More data points, the better the prediction

LET'S PREDICT

Simple_Linear_Regression.ipynb

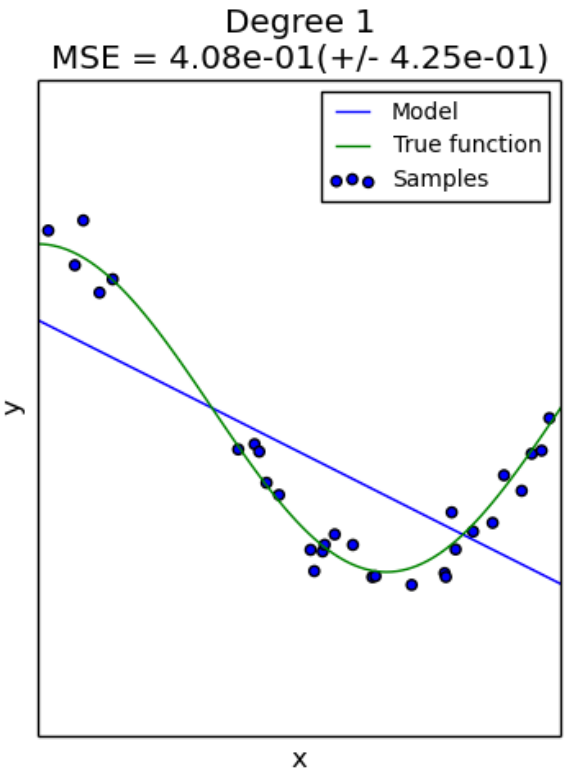
5 MIN BREAK

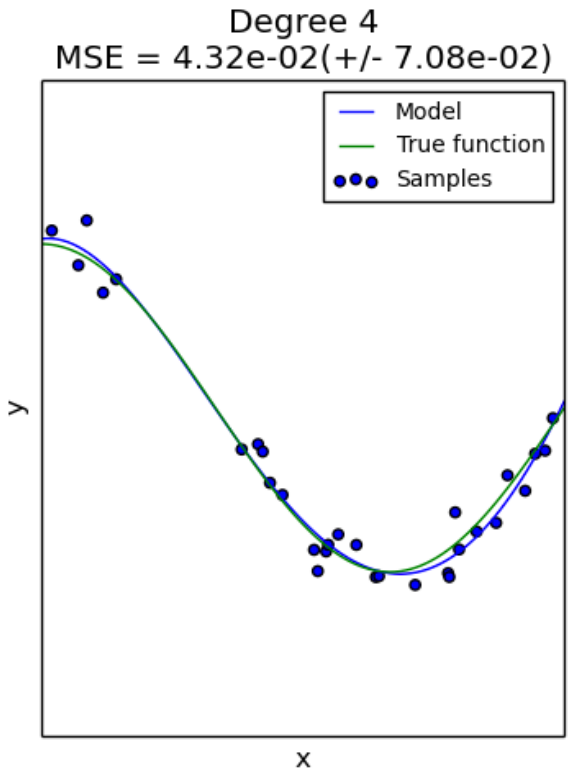


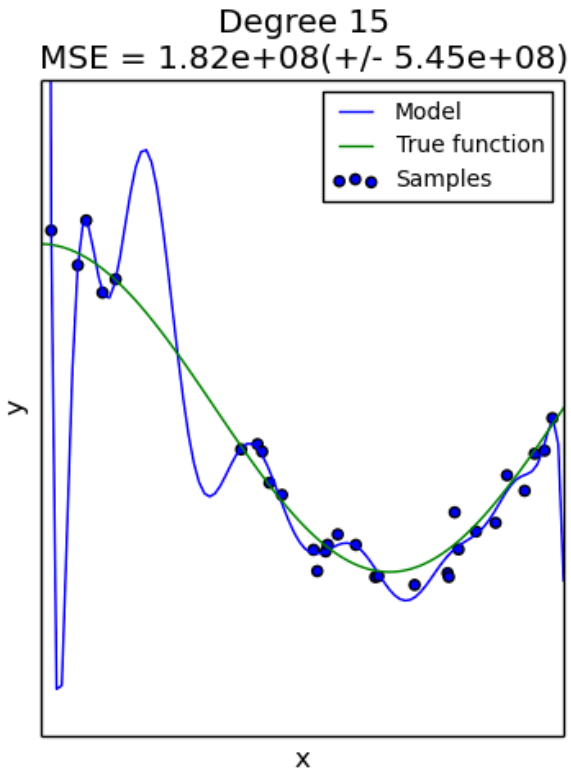
https://xkcd.com/605/

ISSUES

Underfitting/Overfitting





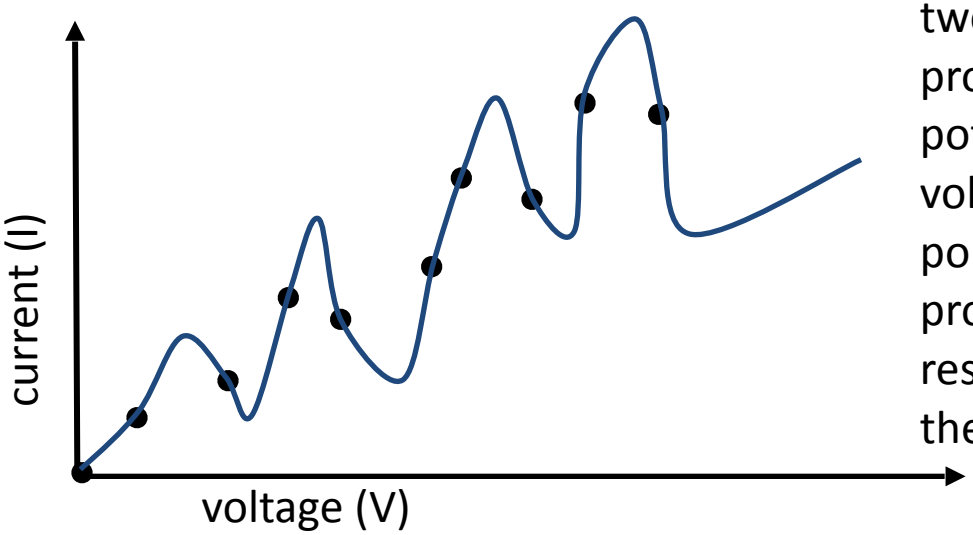


http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Overfitting Example

Experimentally measure 10 points

Fit a curve to the Resulting data.



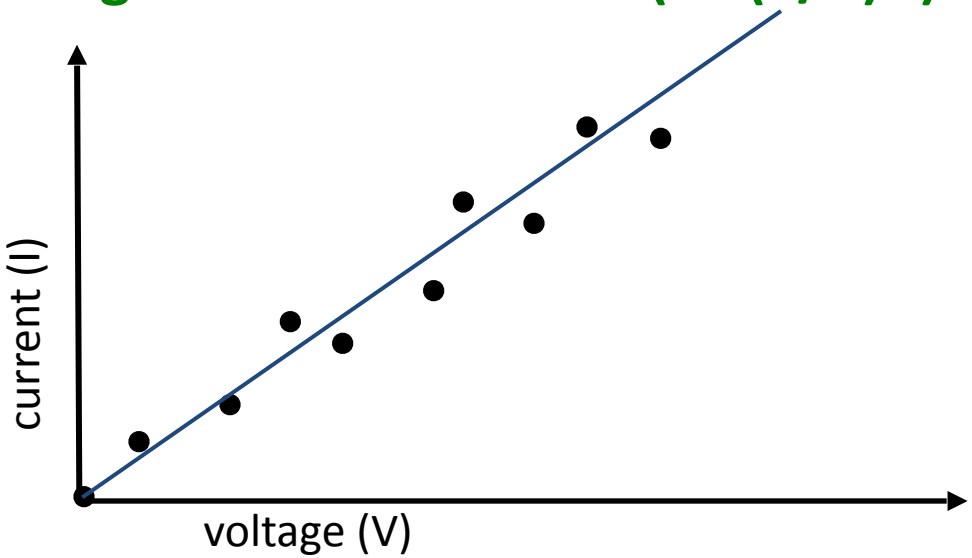
In electrical circuits, Ohm's law states that the current through a conductor between two points is directly proportional to the potential difference or voltage across the two points, and inversely proportional to the resistance between them.

Perfect fit to training data with an 9^{th} degree polynomial
(can fit *n* points exactly with an *n*-1 degree polynomial)

Ohm was wrong, we have found a more accurate function!

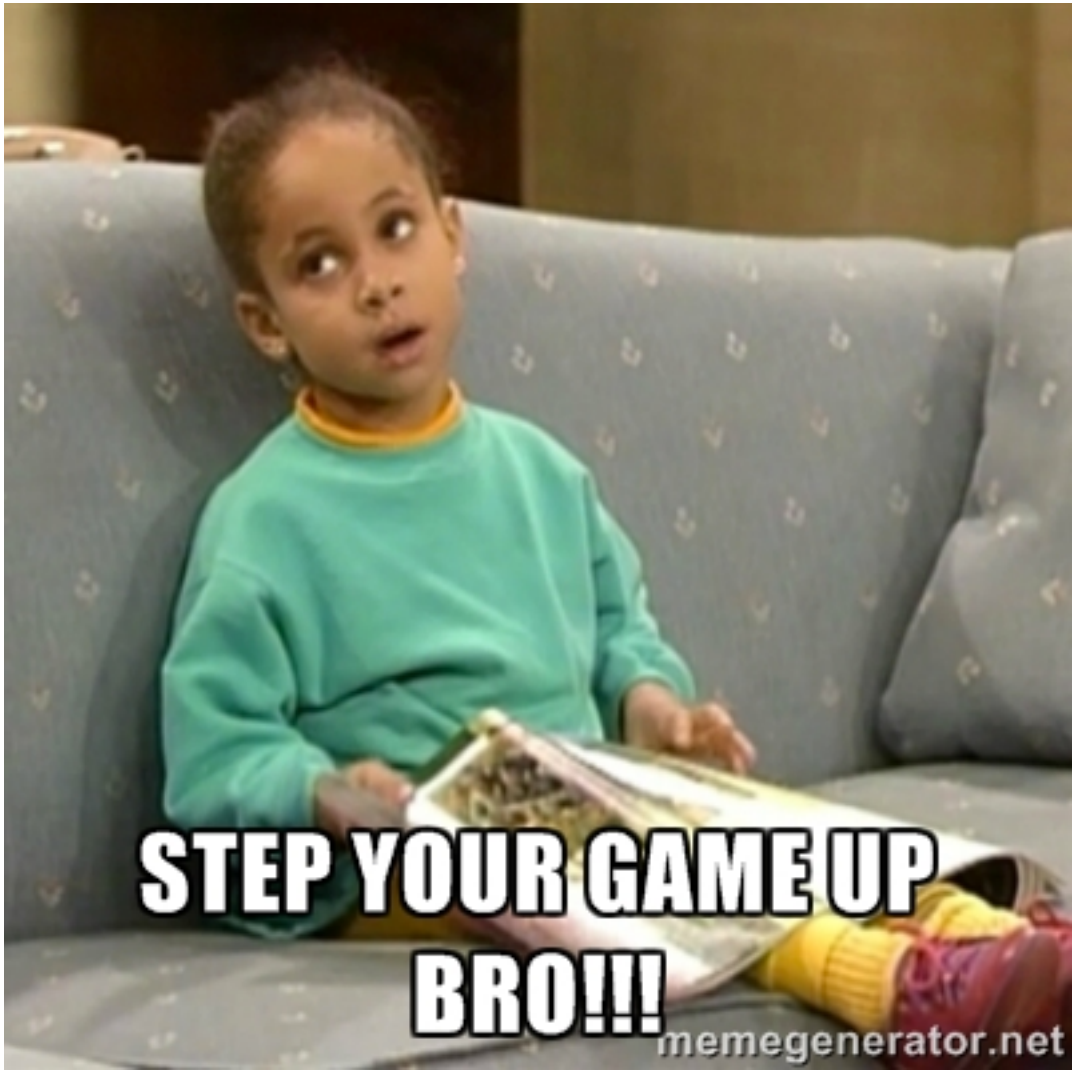
Overfitting Example

Testing Ohms Law: V = IR (I = (1/R)V)



Better generalization with a linear function that fits training data less accurately.

Multiple Linear Regression



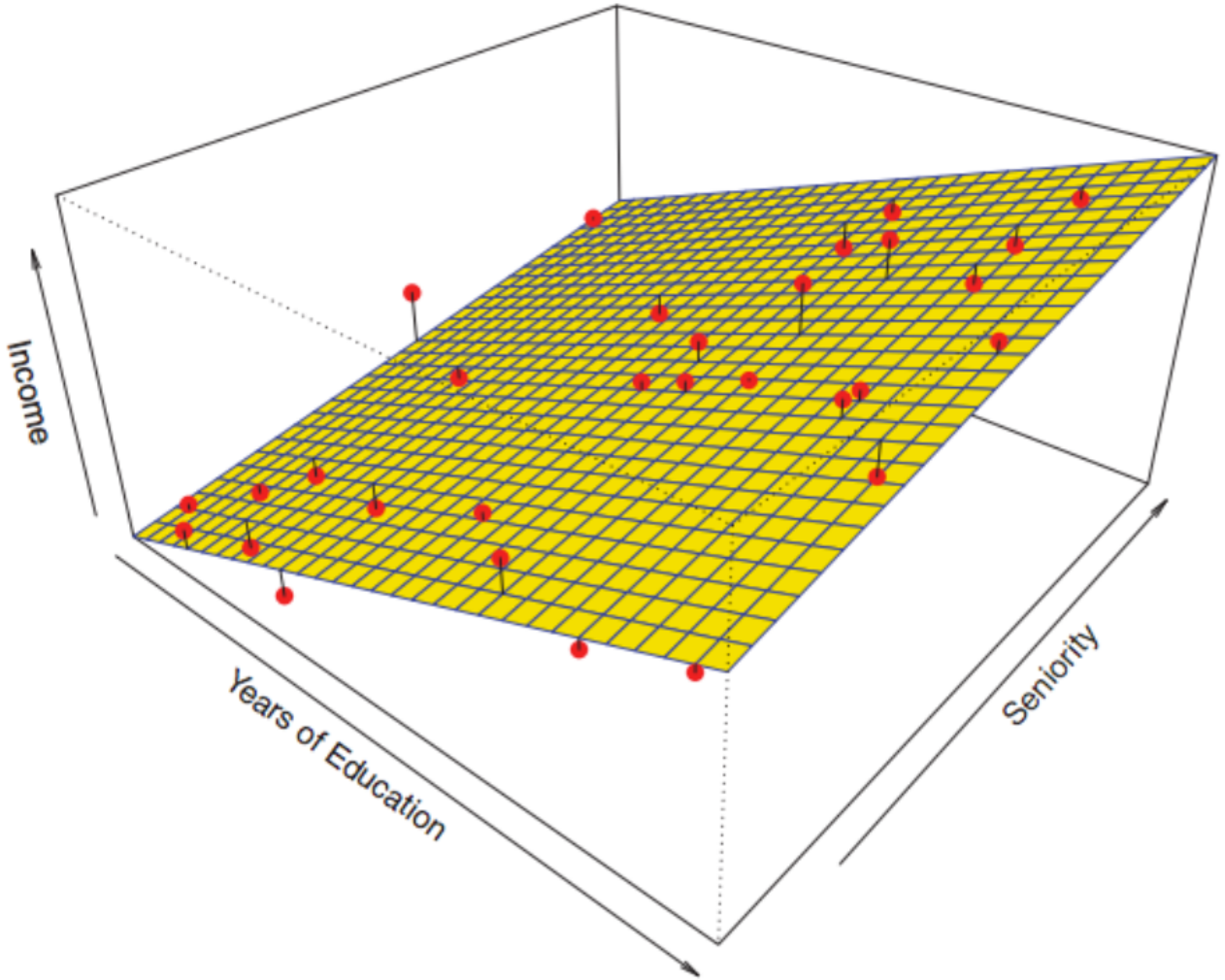
Multiple Linear Regression

- Instead of one feature we have several features
- Each feature has their own coefficient

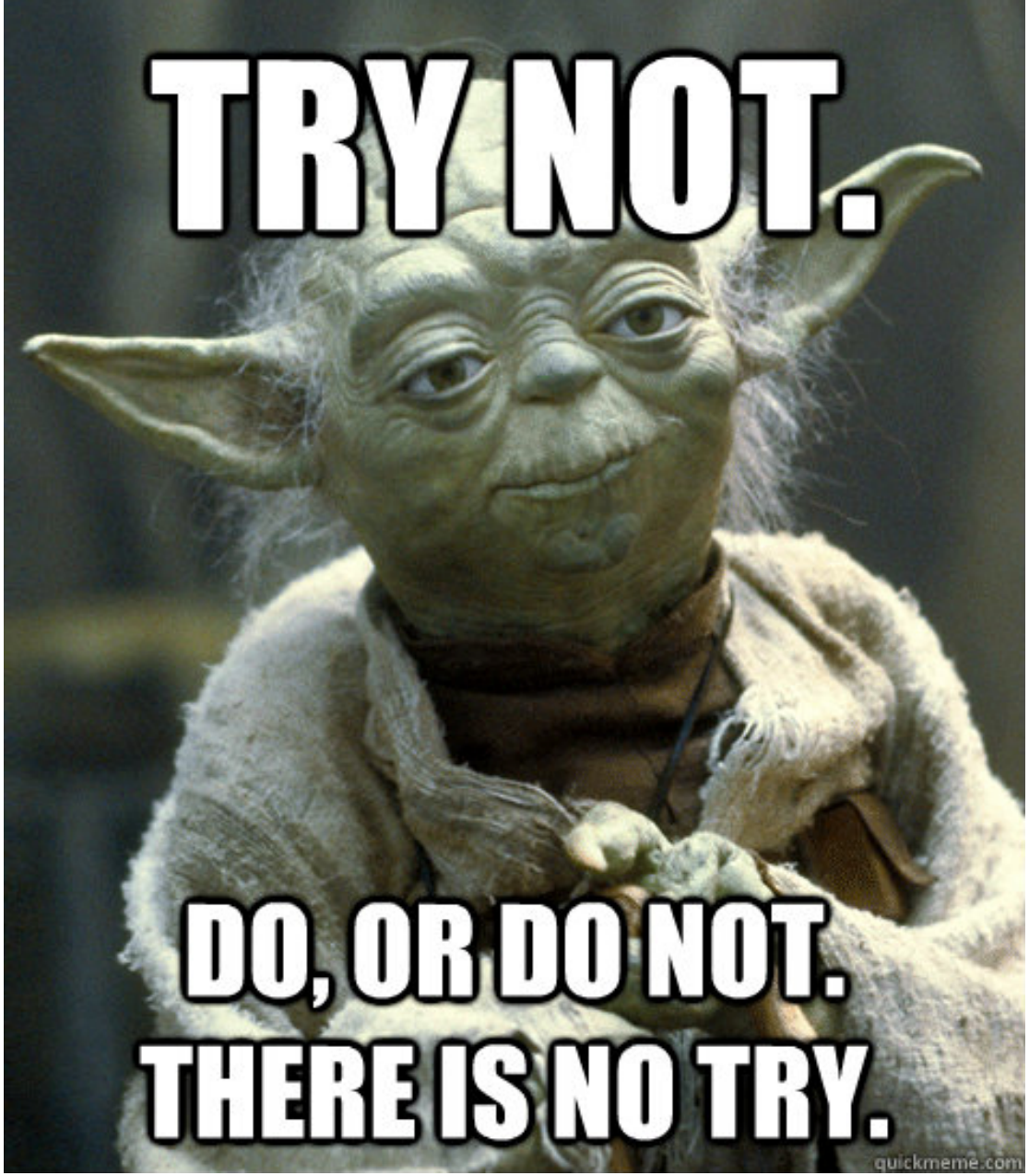
$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

http://nbviewer.ipython.org/github/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb

Multiple Linear Regression



Introduction to Statistical Learning
http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf



Multiple_Variable_Regression.ipynb


NOT SURE IF I UNDERSTAND THE MATERIAL
OR IF I OVER SIMPLIFIED EVERYTHING
memegenerator.net

WE'VE OBVIOUSLY OVERSIMPLIFIED

There's much more to validating and tuning regression models, but don't worry about that for now.

Supervised Learning

- The target value (**y**) is known in advance
- We use the relationship between the features (**x**) and the target value for instances where **y** is known to create a predictor for instances where **y** isn't known
- We'll explore this more next week

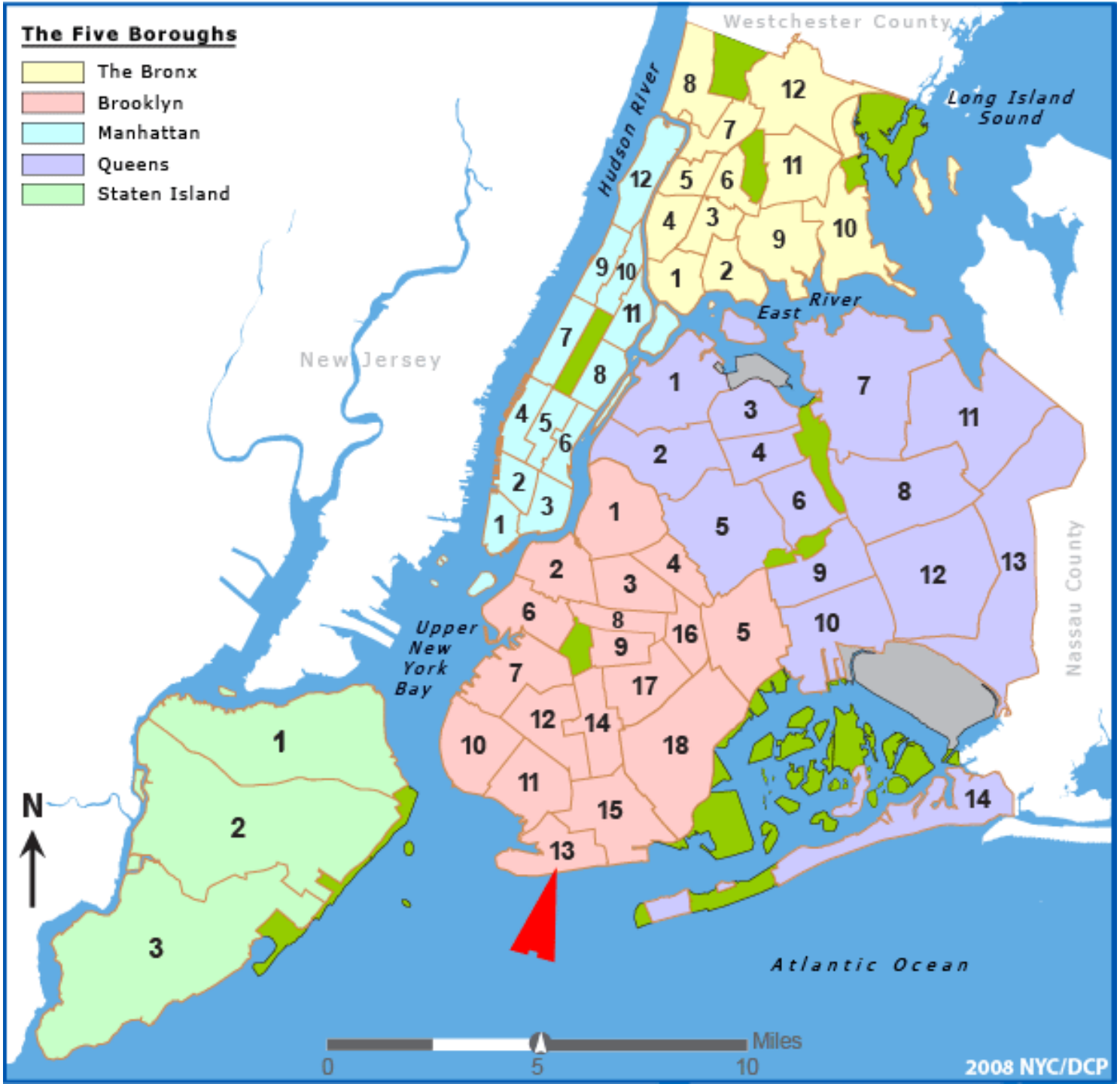
WRAP-UP

Exercises

1. Write code necessary to analyze the relationship between median income and recycling rate in New York City Community Boards (using 2013_NYC_CD_MedianIncome_Recycle.xlsx). Calculate:
   - coefficient of correlation
   - coefficient of determination
2. What is the relationship between these two variables? Write a short Tumblr post outlining the relationship based on your findings

Median Income and Recycling

http://iquantny.tumblr.com/post/79846201258/the-huge-correlation-between-median-income-and


The Five Boroughs
The Bronx
Brooklyn
Manhattan
Queens
Staten Island
Westchester County
Long Island Sound
Hudson River
East River
New Jersey
Upper New York Bay
Nassau County
Atlantic Ocean
N
Miles
0
5
10
2008 NYC/DCP

Exercises

3. Based on the outputs from Exercise 1, create a function that takes in a median income and outputs an estimated recycling rate.

Exercises

4. Using the height_weight_gender.csv data from class, filter the data by gender and create models for each gender (male and female). Write a function that takes in a person's height and gender and outputs an estimated weight with appropriate error estimation.

Exercises

5. Using data from the FiveThirtyEight post http://53eig.ht/1e2aV6U, write code to calculate the correlation of the responses from the poll.
6. Write a short Tumblr post describing the results of your analysis