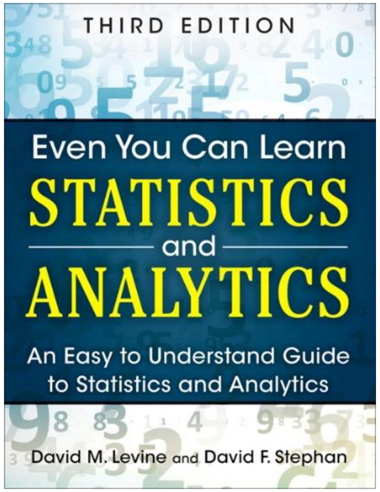
IMGD 2905

Simple Linear Regression

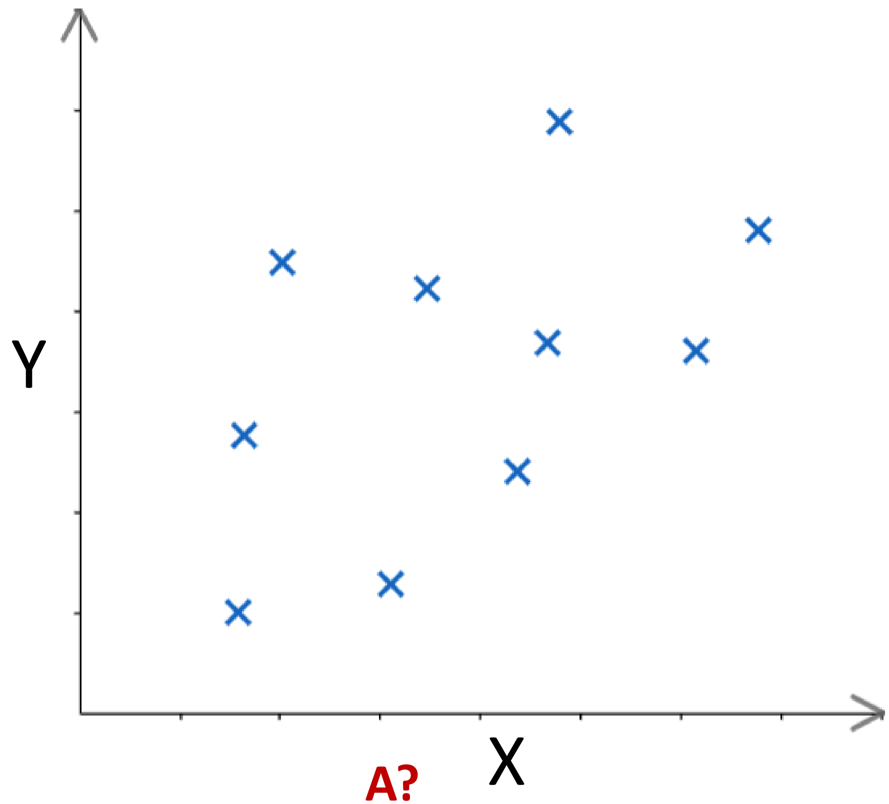
Chapter 10



Motivation

- Have data (sample, **x**'s)
- Want to know likely value of next observation (Y)
  - E.g., *playtime*
- A –

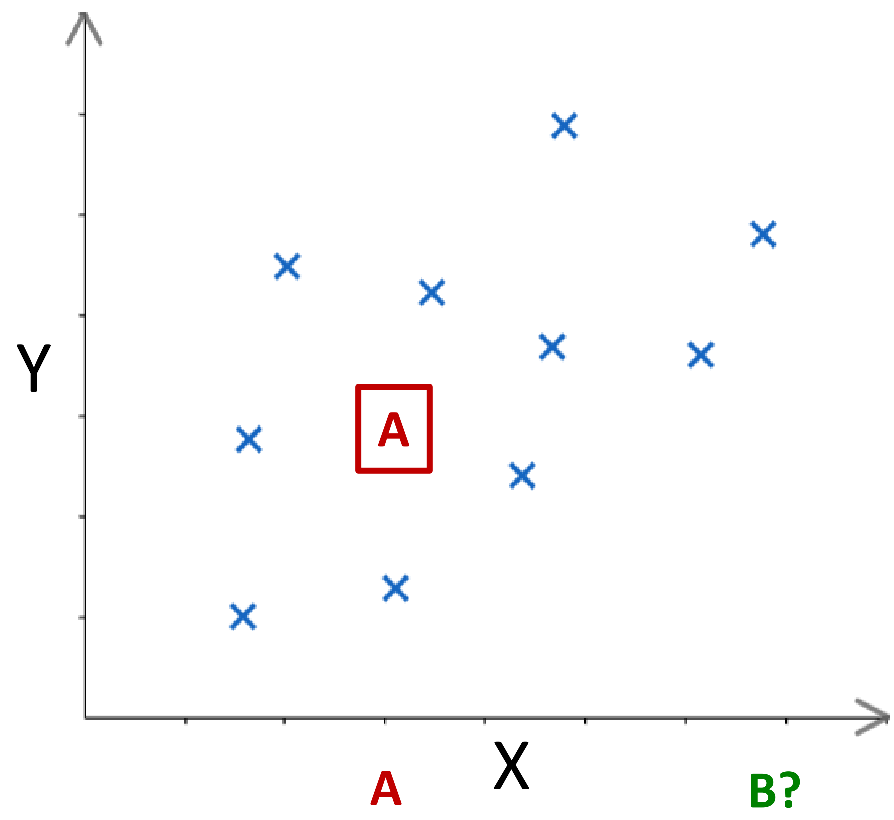
Q: Given X at A and previous Y's, what is likely next Y?



Motivation

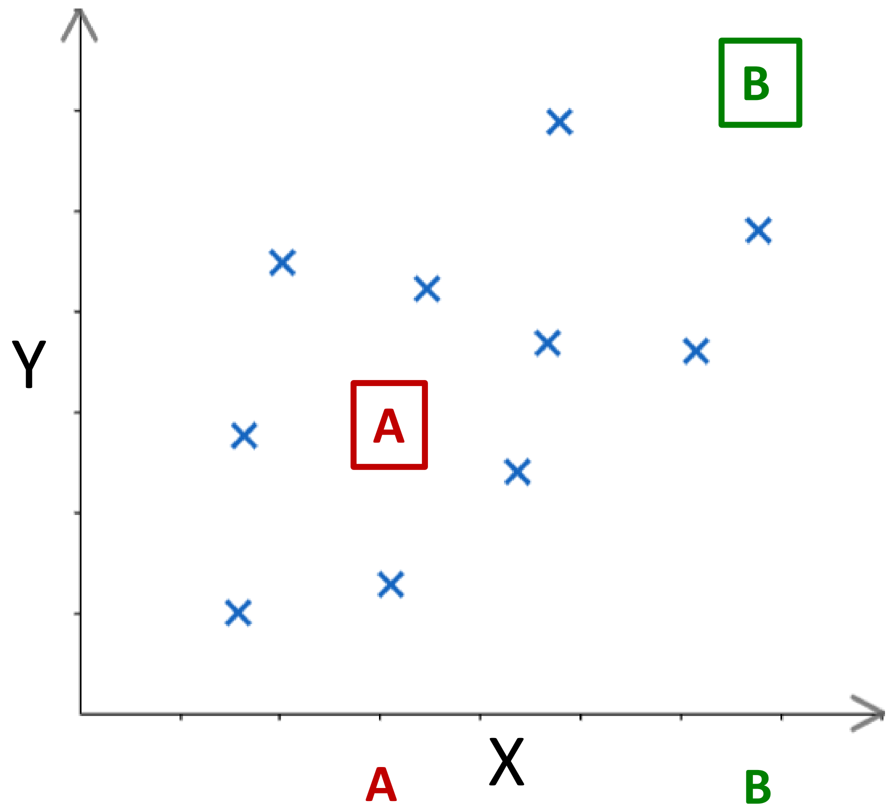
- Have data (sample, **x**'s)
- Want to know likely value of next observation (Y)
  - E.g., *playtime*
- A – reasonable to compute mean y-value (with confidence interval)
- B –

Q: Given X at B and previous Y's, what is likely next Y?



Motivation

- Have data (sample, **x**'s)
- Want to know likely value of next observation (Y)
  - E.g., *playtime* versus *skins owned*
- A – reasonable to compute mean y-value (with confidence interval)
- B – could do same, but there appears to be relationship between X and Y!

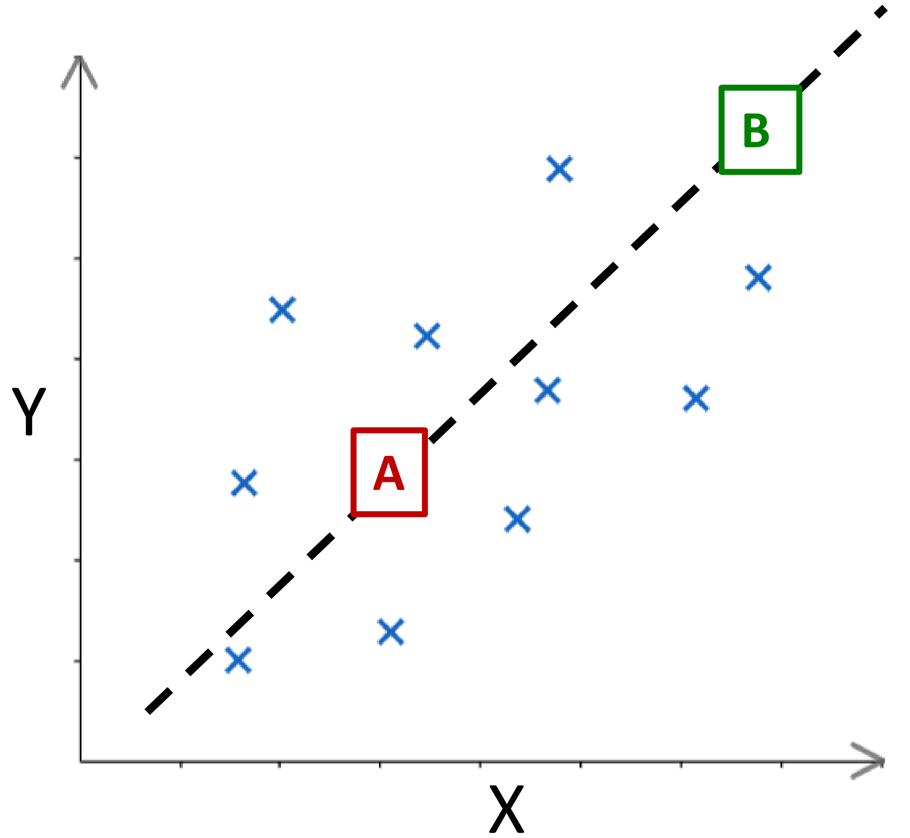


Motivation

- Have data (sample, **x**'s)
- Want to know likely value of next observation (Y)
  - E.g., *playtime* versus *skins owned*
- A – reasonable to compute mean y-value (with confidence interval)
- B – could do same, but there appears to be relationship between X and Y!

→ Predict B (here, use X data to predict Y)

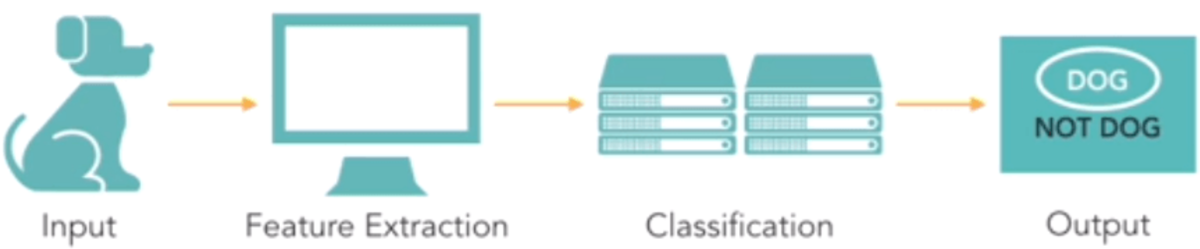
e.g., "trendline" (regression)



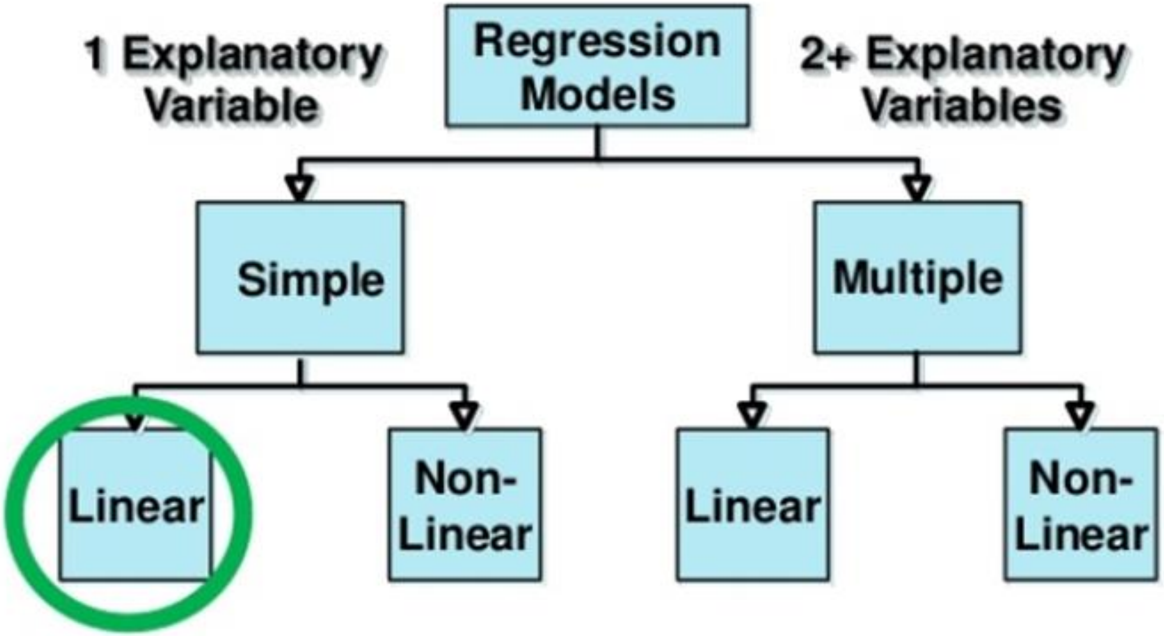
Overview

Broadly, two types of prediction techniques:

1. Regression – mathematical equation to model, use model for predictions
   - We'll discuss simple linear regression
2. Machine learning – branch of AI, use computer algorithms to determine relationships (predictions)
   - CS 4342 Machine Learning



Types of Regression Models



- Explanatory variable *explains* dependent variable
  - Variable X (e.g., skill level) explains Y (e.g., KDA)
  - Can have 1 (simple) or 2+ (multiple)
- Linear if coefficients added, else Non-linear

Outline

- Introduction (done)
- Simple Linear Regression (next)
  - Linear relationship
  - Residual analysis
  - Fitting parameters
- Measures of Variation
- Misc

Simple Linear Regression

- Goal – find a linear (line) relationship between two values
  - E.g., *KDA* and *skill*, *time* and *car speed*
- First, make sure relationship is linear! How?

Simple Linear Regression

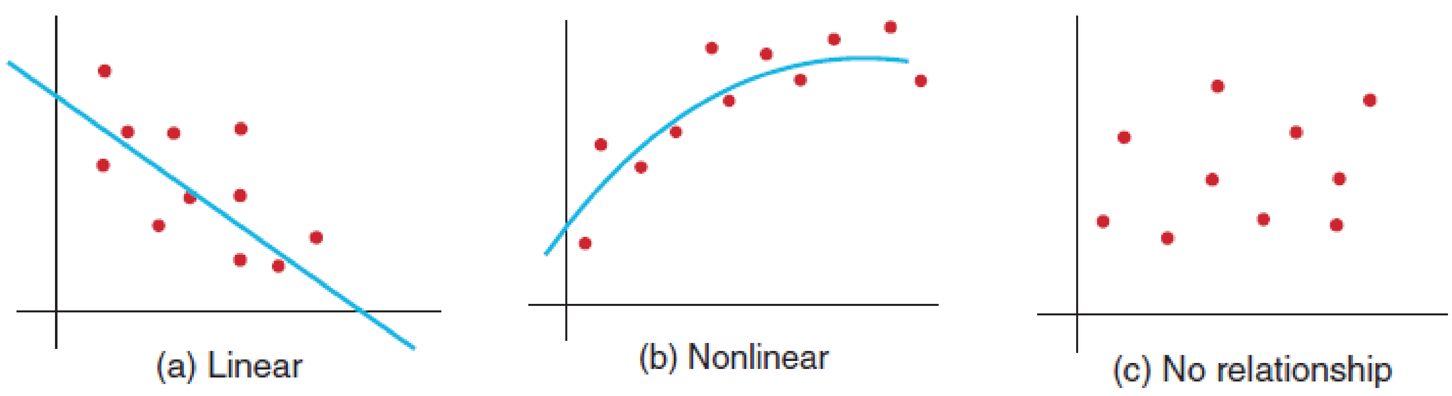
- Goal – find a linear (line) relationship between two values
  - E.g., *KDA* and *skill*, *time* and *car speed*
- First, make sure relationship is linear! How?

→ Scatterplot

(c) no clear relationship

(b) not a linear relationship

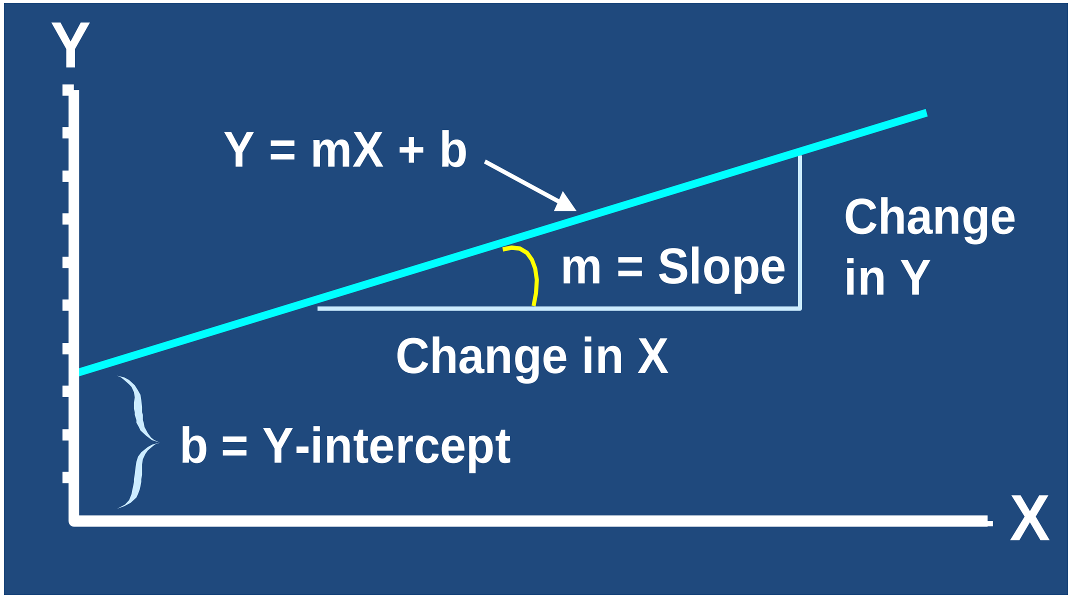
(a) linear relationship – proceed with linear regression



Linear Relationship

$Y = mX + b$

- From algebra: line in form
  - m is slope, b is y-intercept
- Slope (m) is amount *Y* increases when *X* increases by 1 unit
- Intercept (b) is where line crosses y-axis, or where y-value when $x = 0$



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Simple Linear Regression Example

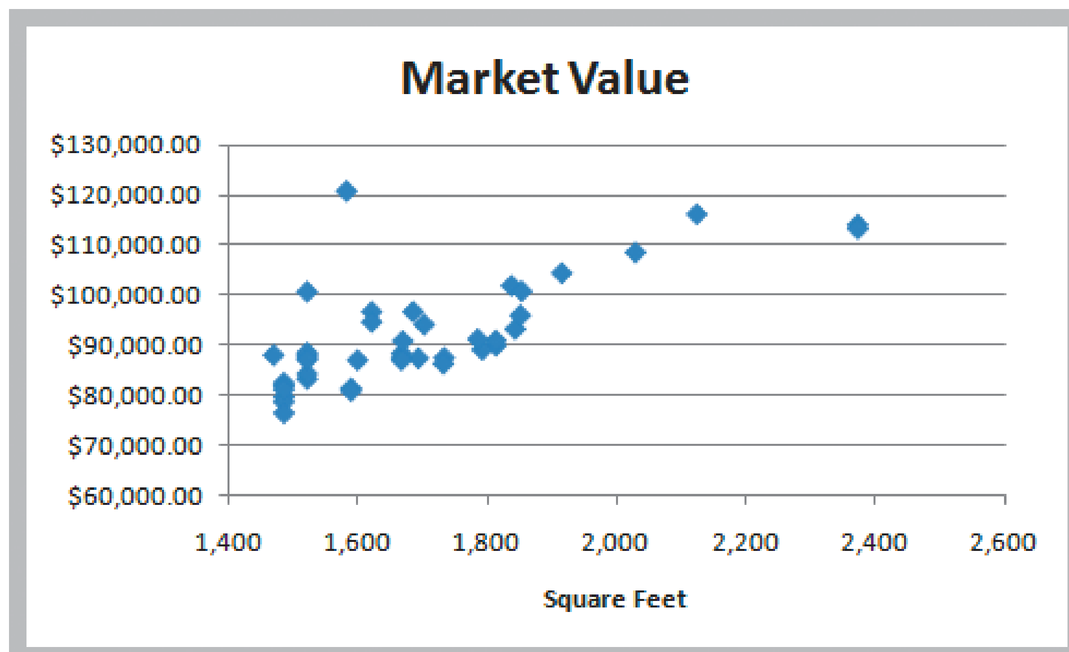
- Size of house related to its market value.

  X = square footage

  Y = market value ($)

- Scatter plot (42 homes)
  - indicates linear trend

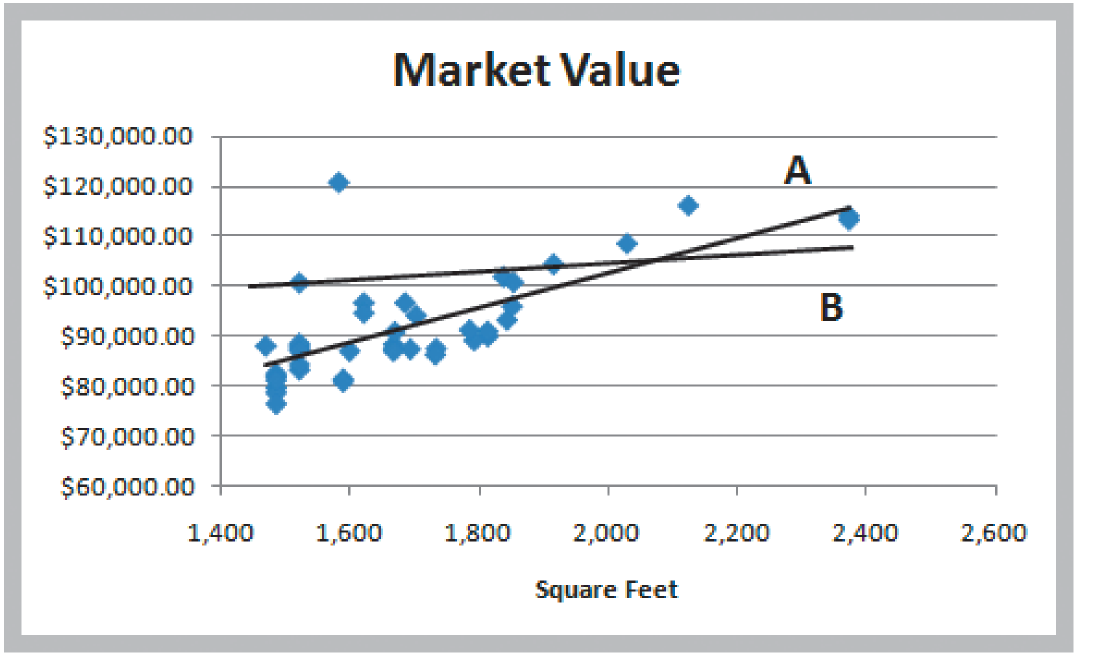
	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00



Simple Linear Regression Example

- Two possible lines shown below (A and B)
- Want to determine best regression line
- Line A looks a better fit to data
  - – But how to know?

$Y = mX + b$



Simple Linear Regression Example

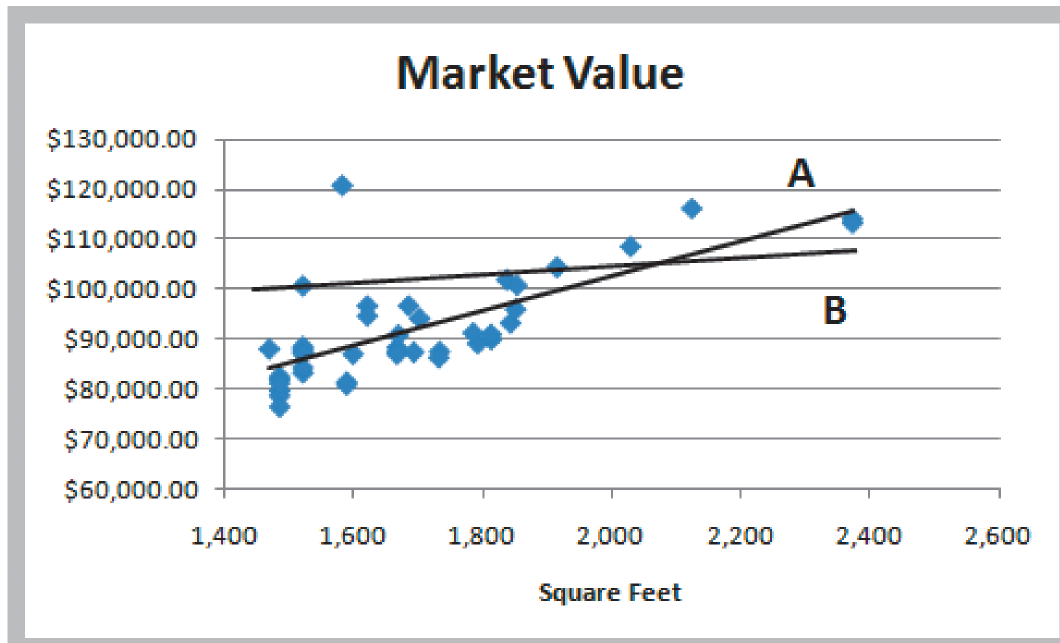
- Two possible lines shown below (A and B)
- Want to determine best regression line
- Line A looks a better fit to data
  - But how to know?

$$Y = mX + b$$

Line that gives best fit to data is one that minimizes prediction error

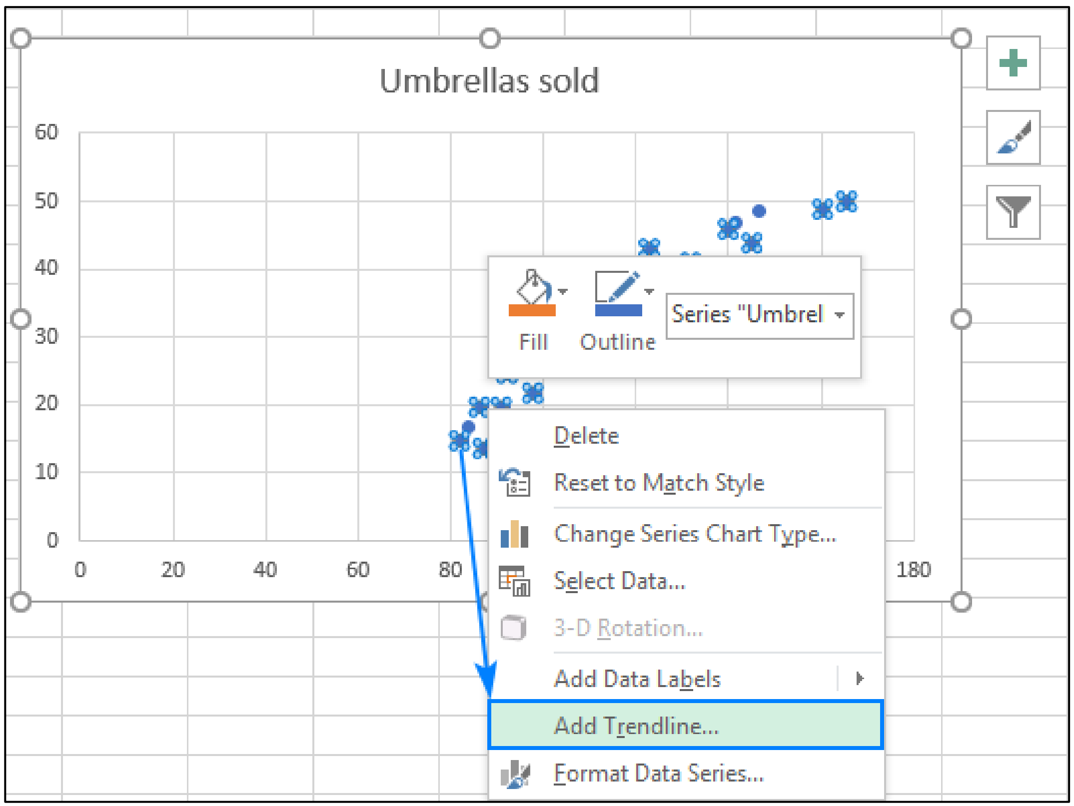
→ Least squares line

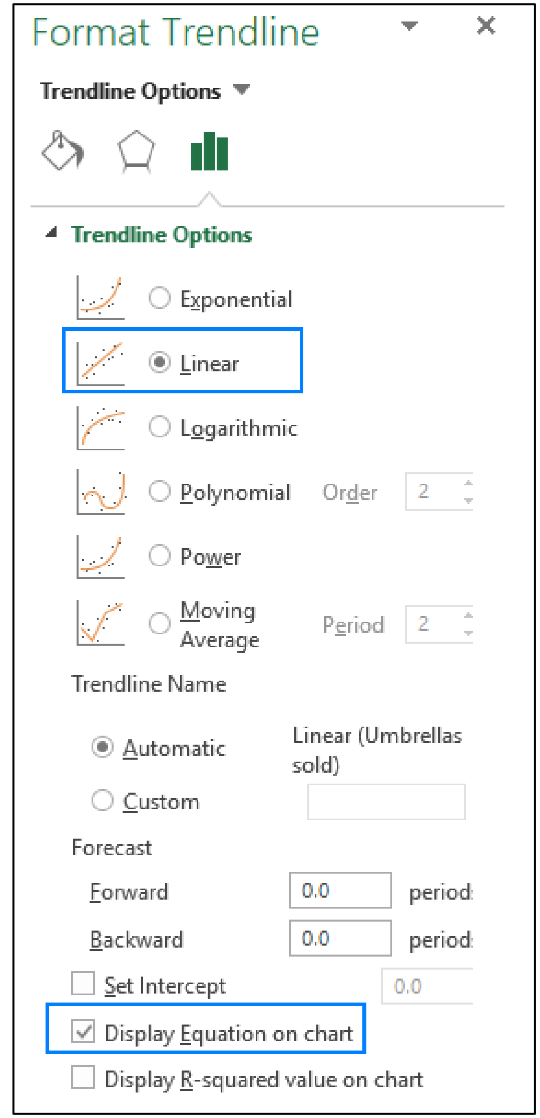
(more later)



Simple Linear Regression Example Chart

- Scatterplot
- Right click → Add Trendline





Simple Linear Regression Example Formulas

=SLOPE(C4:C45,B4:B45)

→ Slope = 35.036

=INTERCEPT(C4:C45,B4:B45)

→ Intercept = 32,673

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00

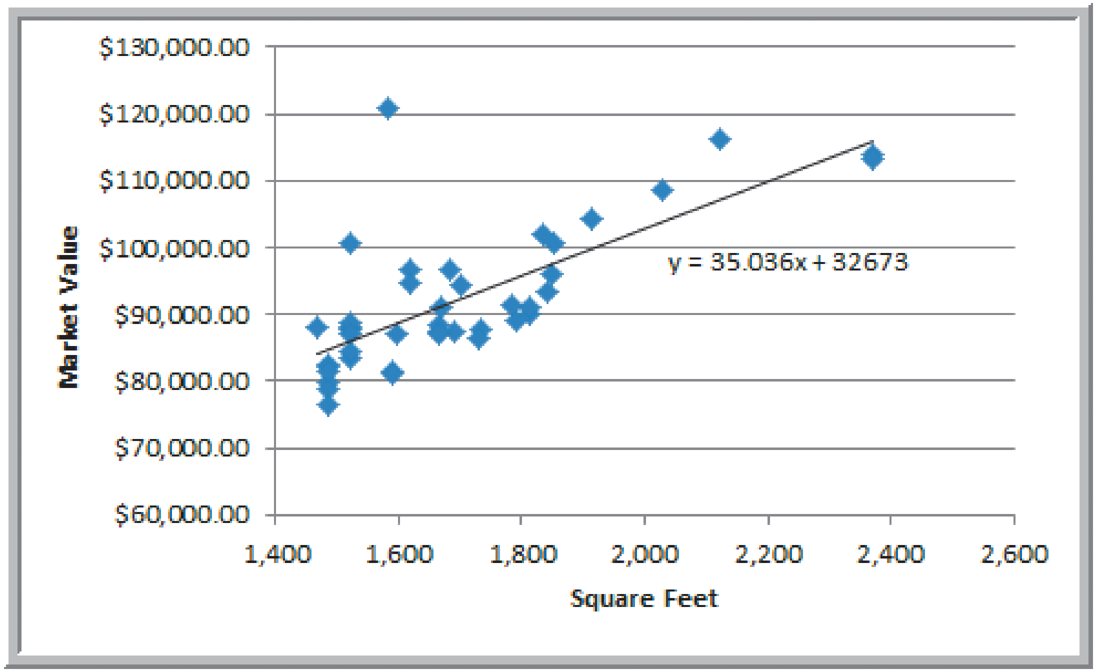
Estimate Y when X = 1800 square feet

Y = 32,673 + 35.036 x (1800) = $95,737.80

Simple Linear Regression Example

Market value = 32673 + 35.036 *x* (square feet)

Predicts market value better than just average



But before use, examine residuals

Groupwork

Simple Linear Regression

https://web.cs.wpi.edu/~imgd2905/d22/groupwork/11-regression/handout.html

Groupwork

1. In simple linear regression, the y-intercept (b) represents the:
   a. predicted value of Y
   b. change in Y per unit change in X
   c. predicted value of Y when X=0
   d. variation around the line

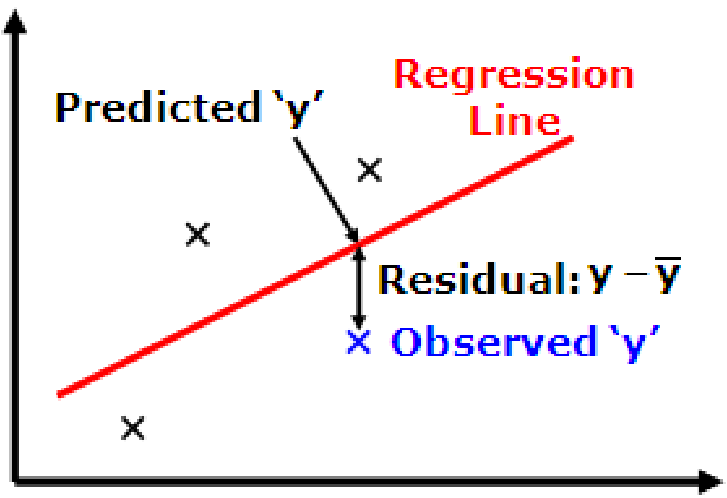
2. A simple linear regression model for predicting a player's points (Y) is 6 X + 10, where X is the player's level.
   – How many more points can a player expect to get when they level up?
   – How many points can a level 10 player expect to get?

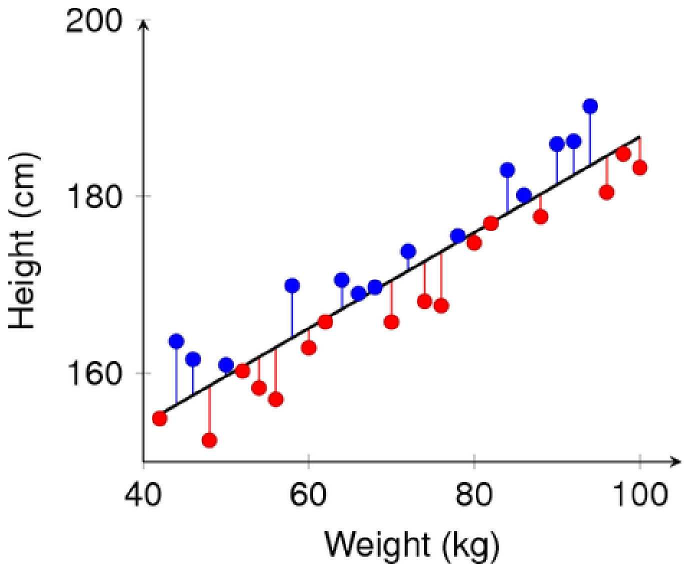
Outline

- Introduction (done)
- Simple Linear Regression
  - Linear relationship (done)
  - Residual analysis (next)
  - Fitting parameters
- Measures of Variation
- Misc

Residual Analysis

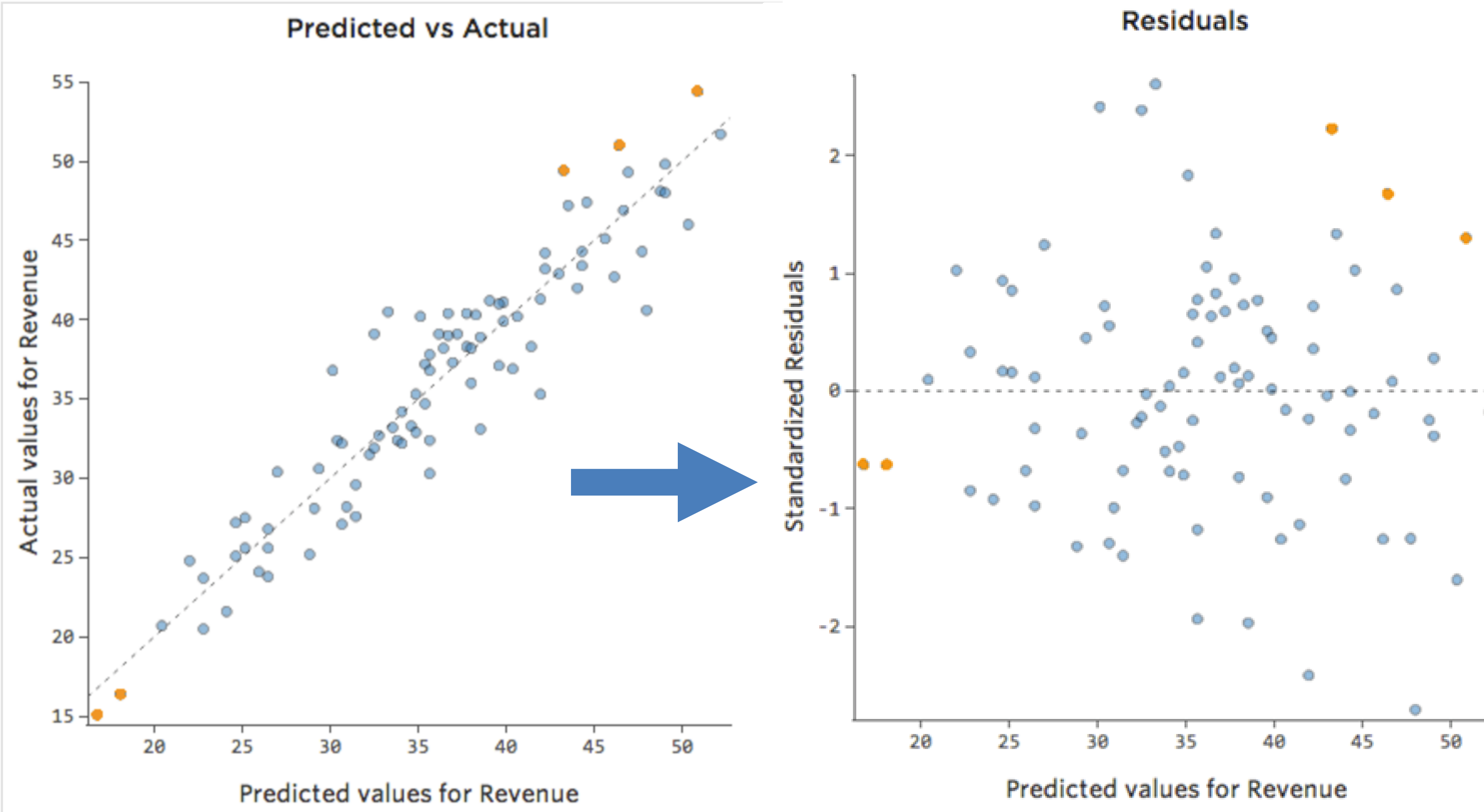
- Before predicting, confirm that linear regression assumptions hold
  - Variation around line is normally distributed
  - Variation equal for all *X*
  - Variation independent for all *X*
- How? Compute residuals (error in prediction) → Chart





Residual Analysis

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/



Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

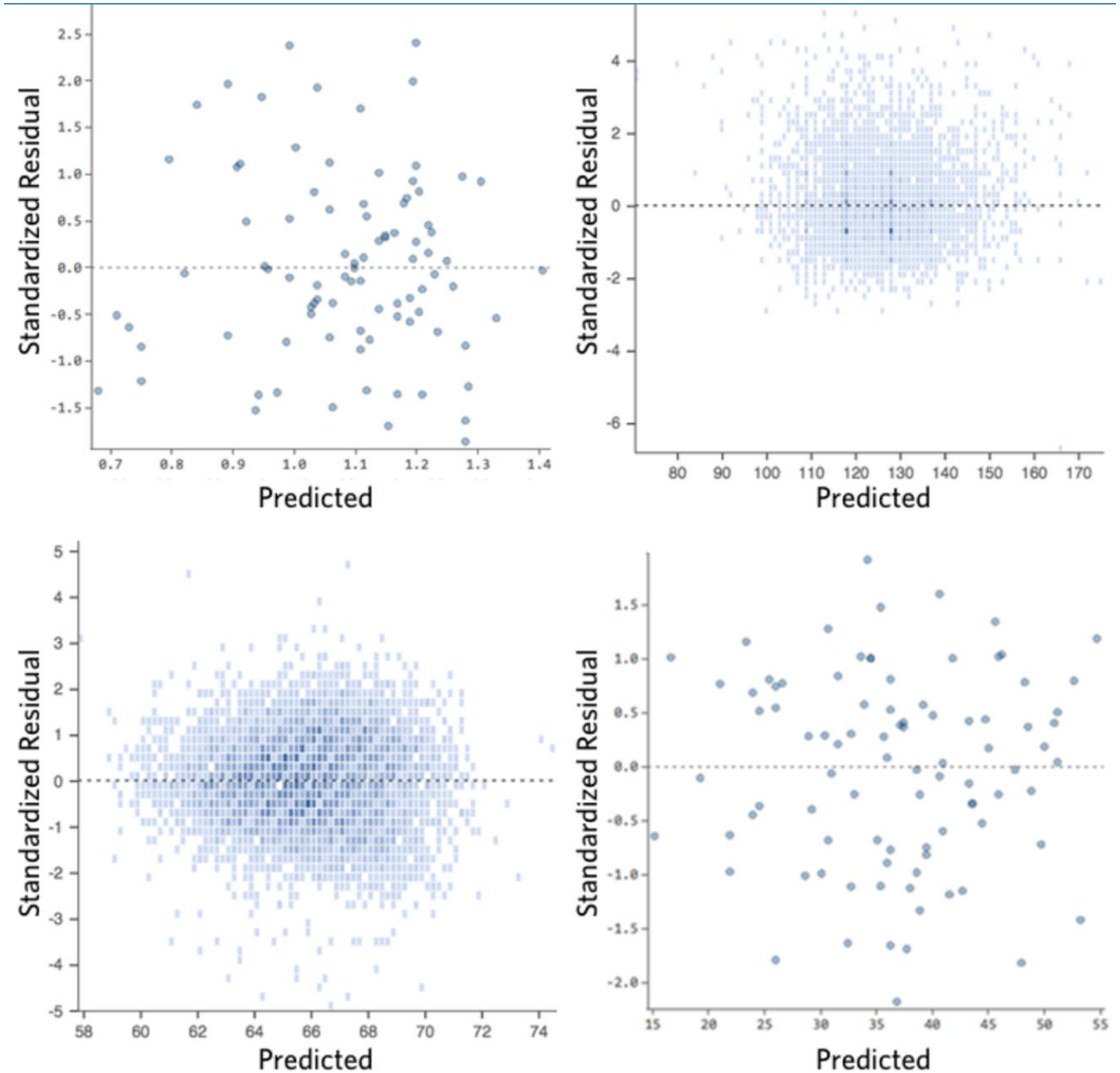
Variation around line normally distributed ?
Variation equal for all X
Variation independent for all X?

Residual Analysis – Good

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Symmetrically distributed

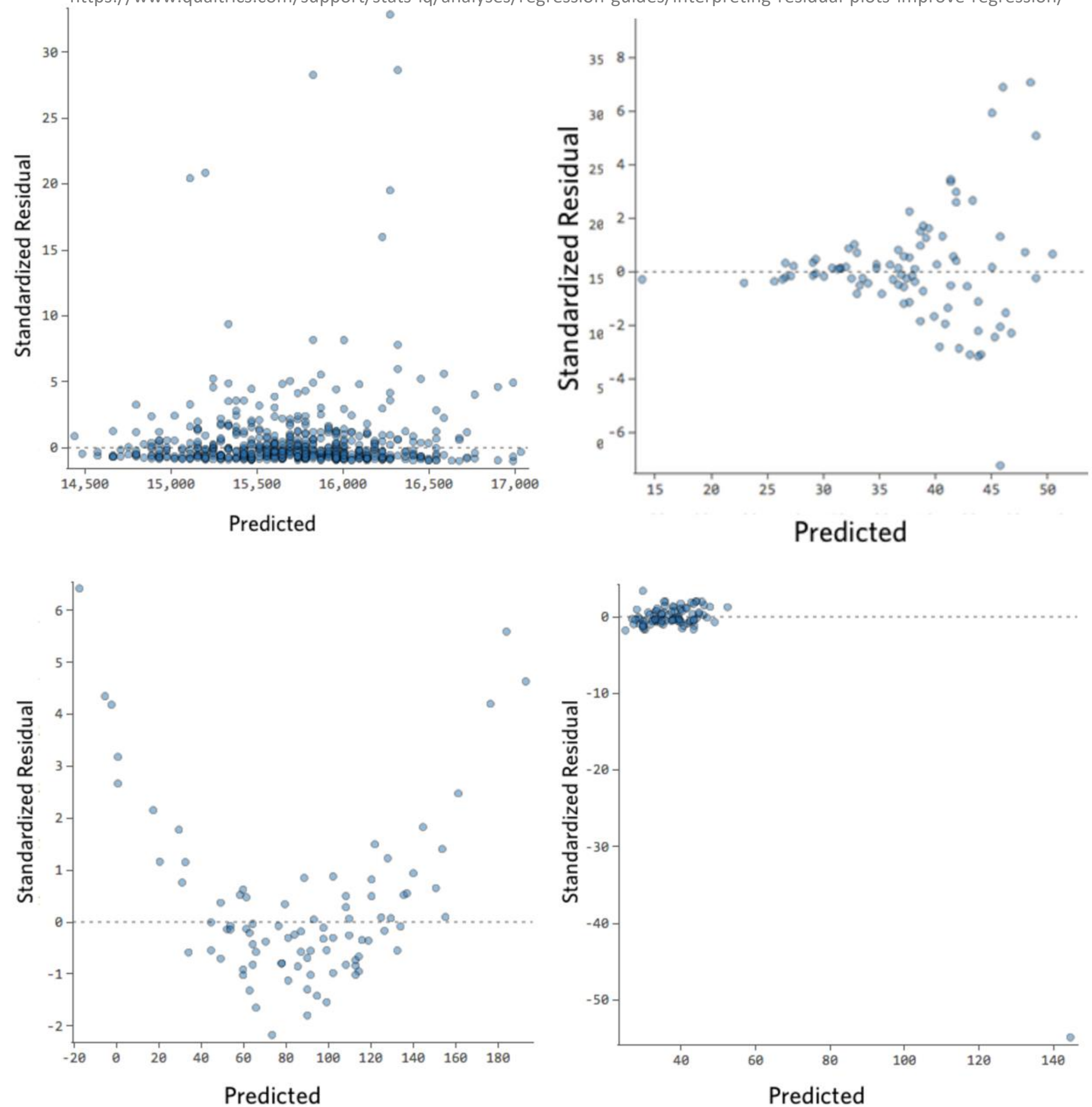
No clear pattern



Clustered towards middle, ok

Residual Analysis – Bad

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/



Clear shape

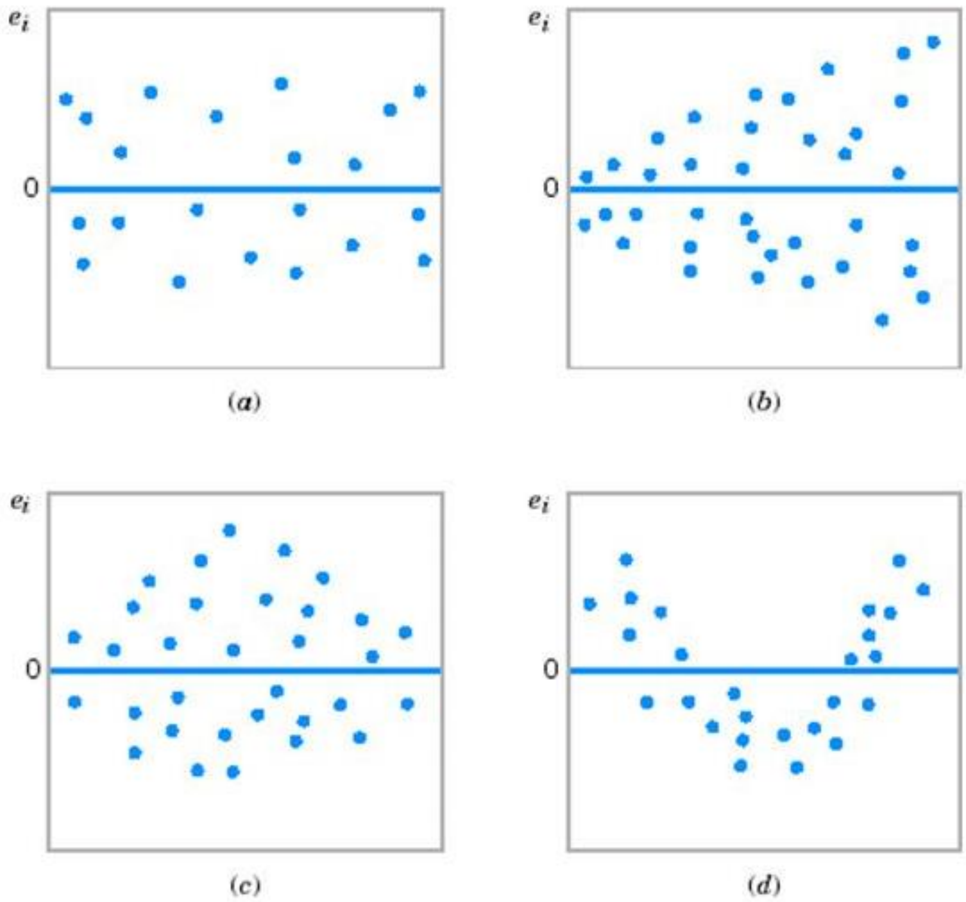
Outliers

Patterns

Note: could do normality test (QQ plot)

Residual Analysis – Summary

- Regression assumptions:
  - Normality of variation around regression
  - Equal variation for all y values
  - Independence of variation



(a) good

(b) funnel

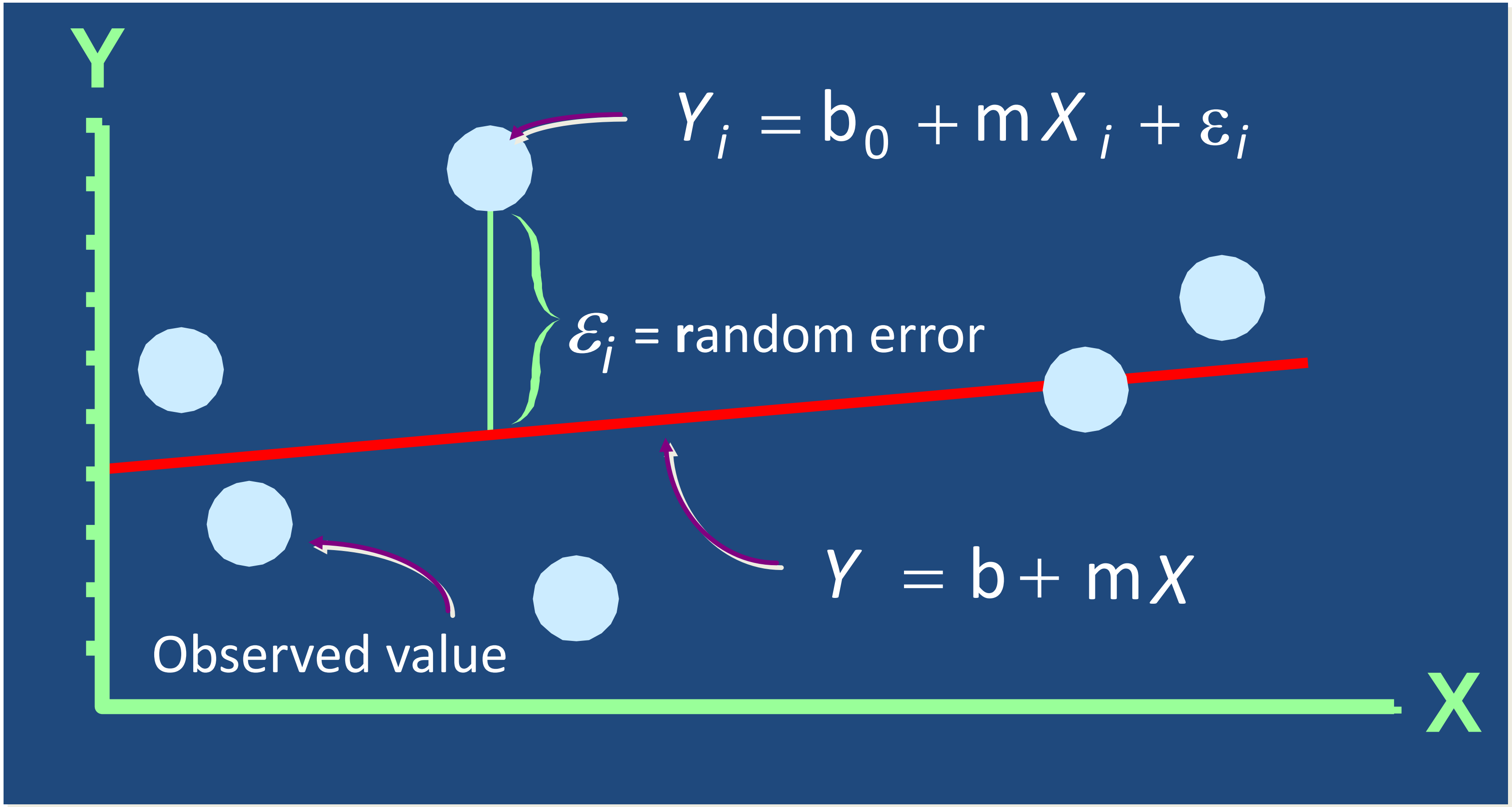
(c) double bow

(d) nonlinear

Outline

- Introduction (done)
- Simple Linear Regression
  - Linear relationship (done)
  - Residual analysis (done)
  - Fitting parameters (next)
- Measures of Variation
- Misc

Linear Regression Model

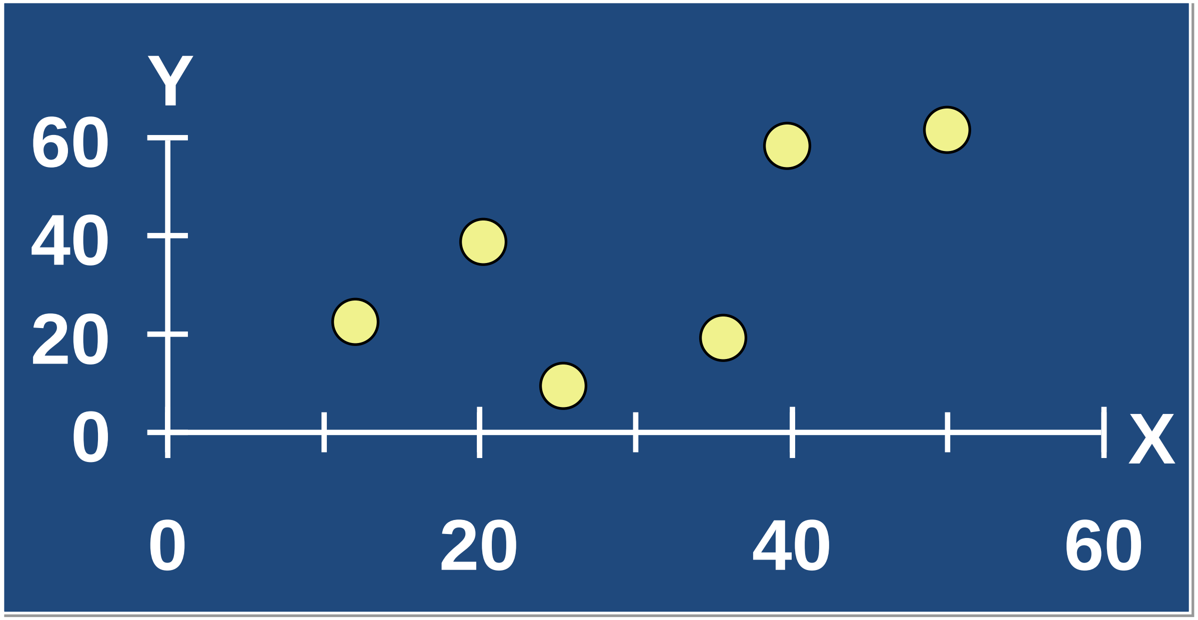


https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Random error associated with each observation (Residual)

Fitting the Best Line

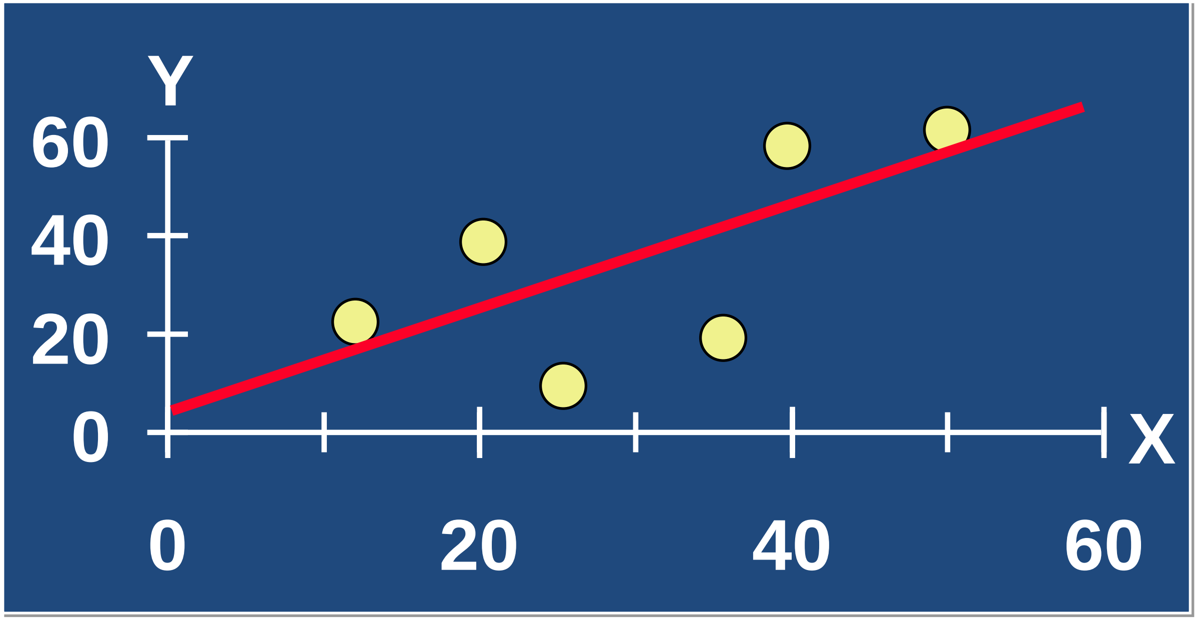
- Plot all (X_i, Y_i) Pairs



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Fitting the Best Line

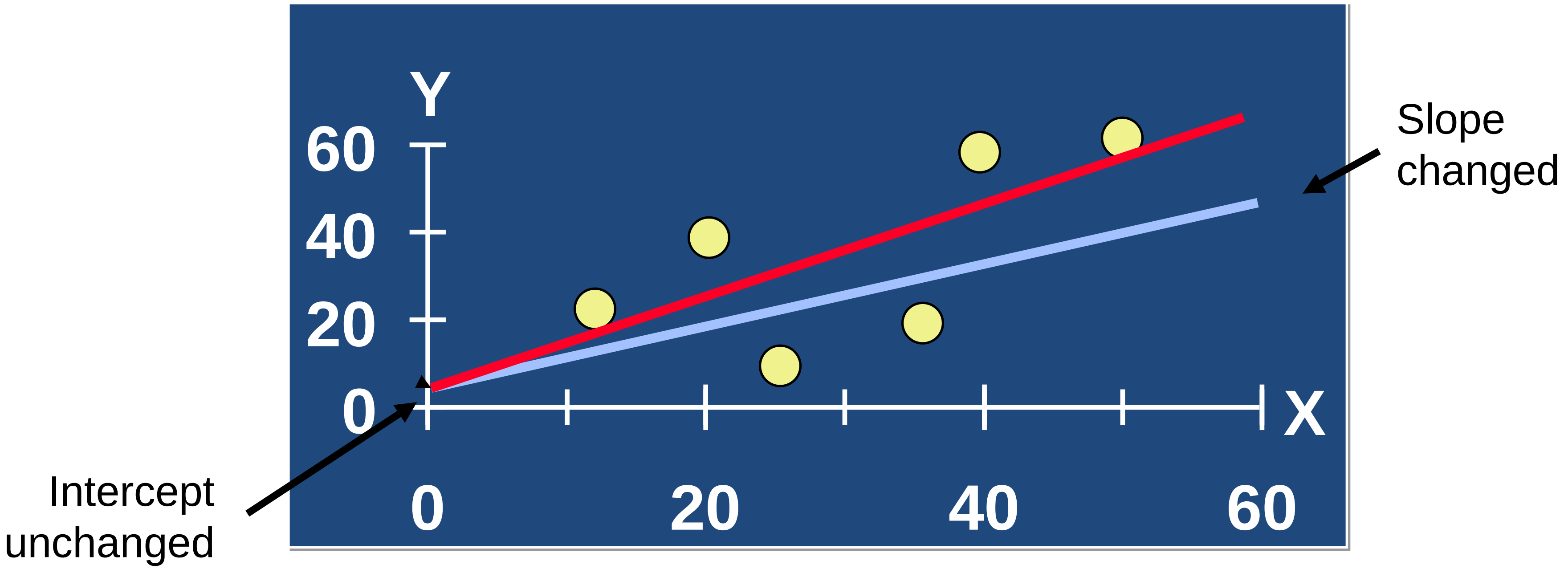
- Plot all (X_i, Y_i) Pairs
- Draw a line. But how do we know it is best?



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Fitting the Best Line

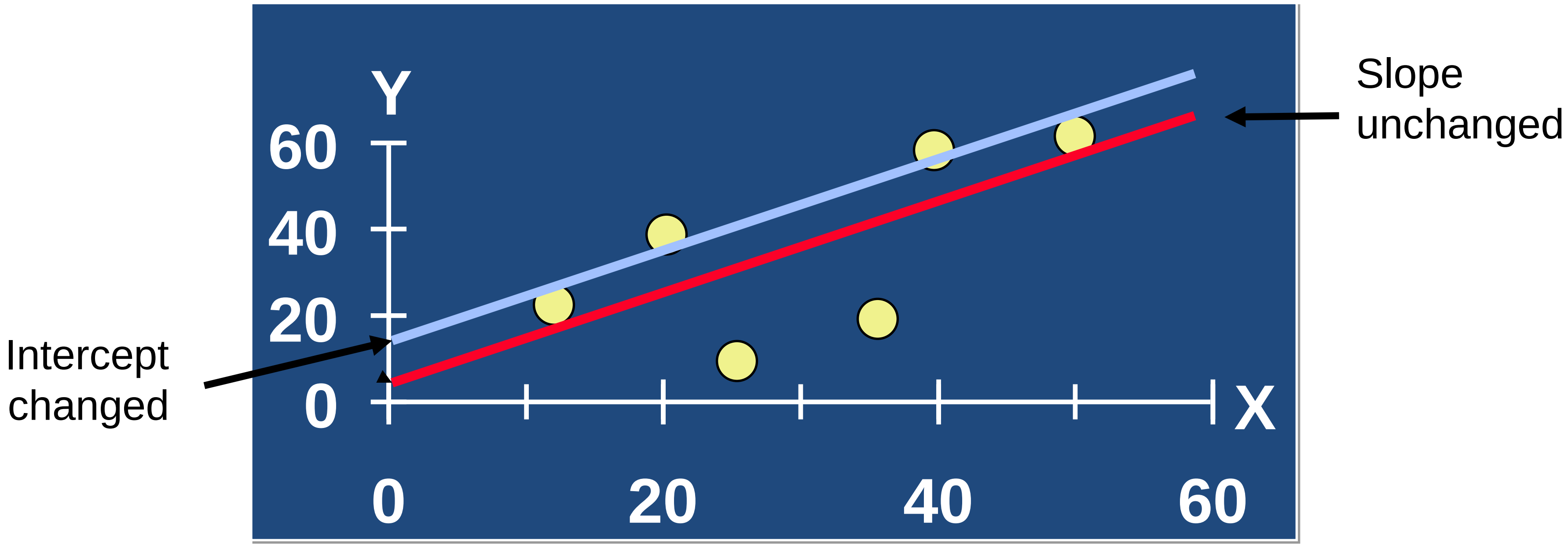
- Plot all (X_i, Y_i) Pairs
- Draw a line. But how do we know it is best?



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Fitting the Best Line

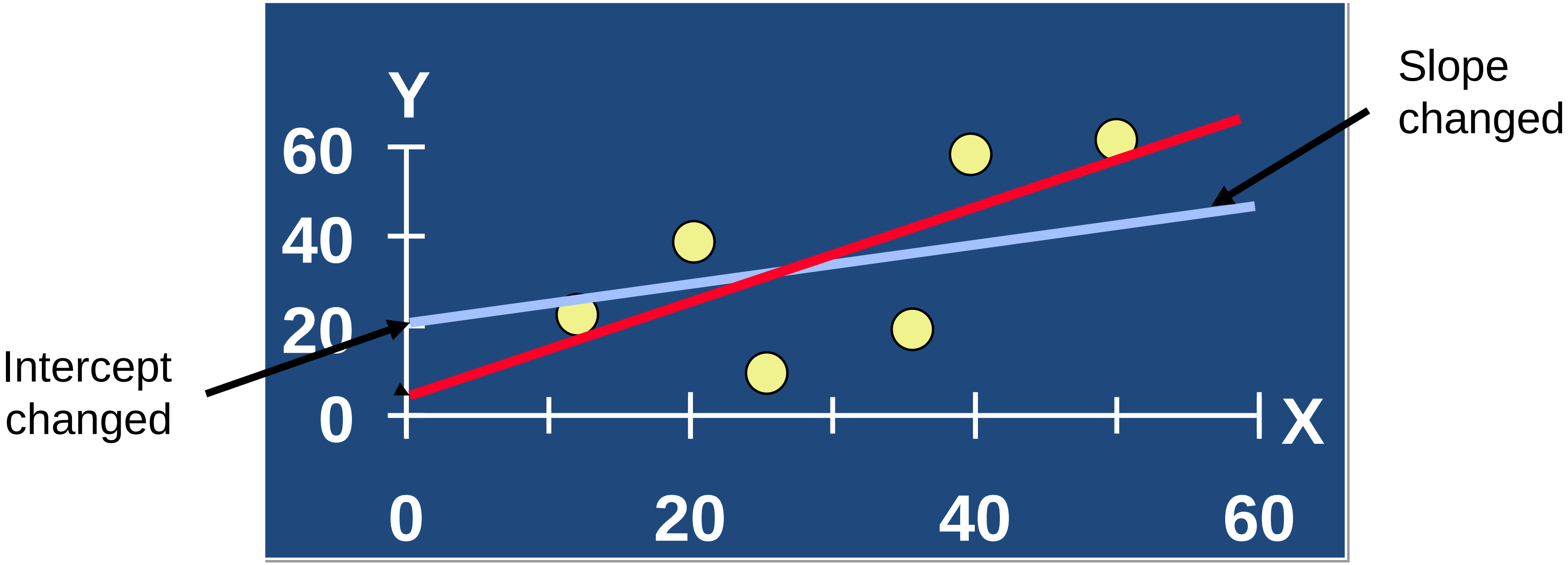
- Plot all (X_i, Y_i) Pairs
- Draw a line. But how do we know it is best?



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Fitting the Best Line

- Plot all (X_i, Y_i) Pairs
- Draw a line. But how do we know it is best?



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Linear Regression Model

- Relationship between variables is linear function

Population Y-Intercept

Population Slope

Random Prediction Error

Want error as small as possible

$$Y_i = b_0 + mX_i + \varepsilon_i$$

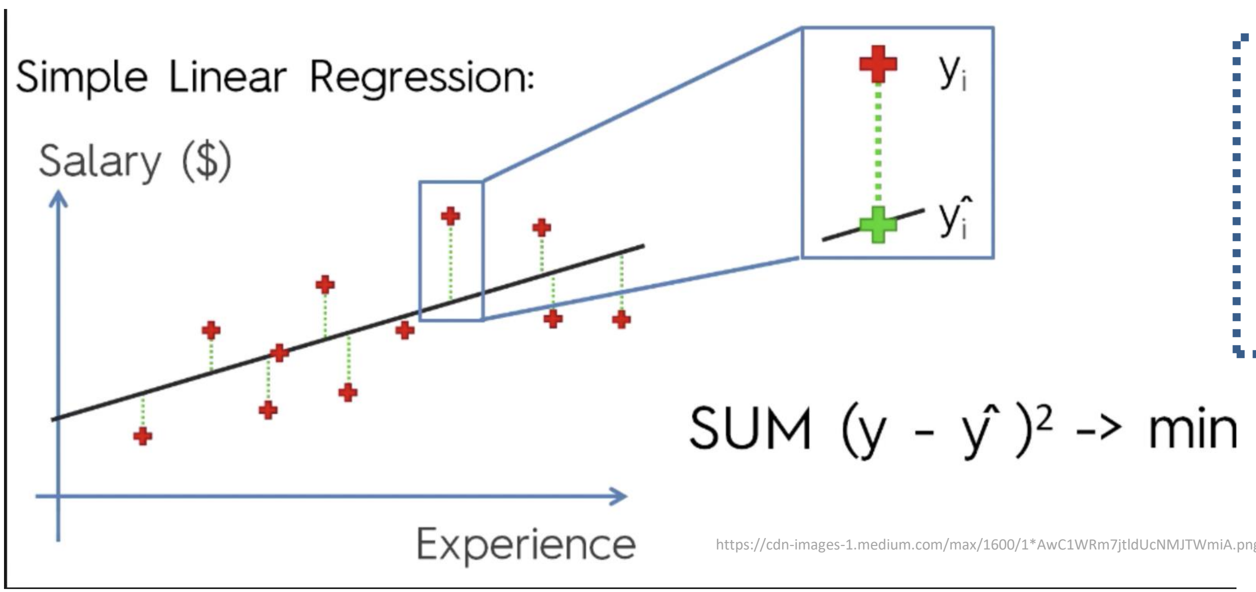
Dependent (response) Variable (e.g., kills)

Independent (explanatory) Variable (e.g., skill level)

Least Squares Line

- Want to minimize difference between actual y and predicted ŷ
  - Add up ε_i for all observed y's
  - But positive differences offset negative ones
  - (remember when this happened for variance?)

  → Square the errors! Then, minimize (using Calculus)



Minimize:

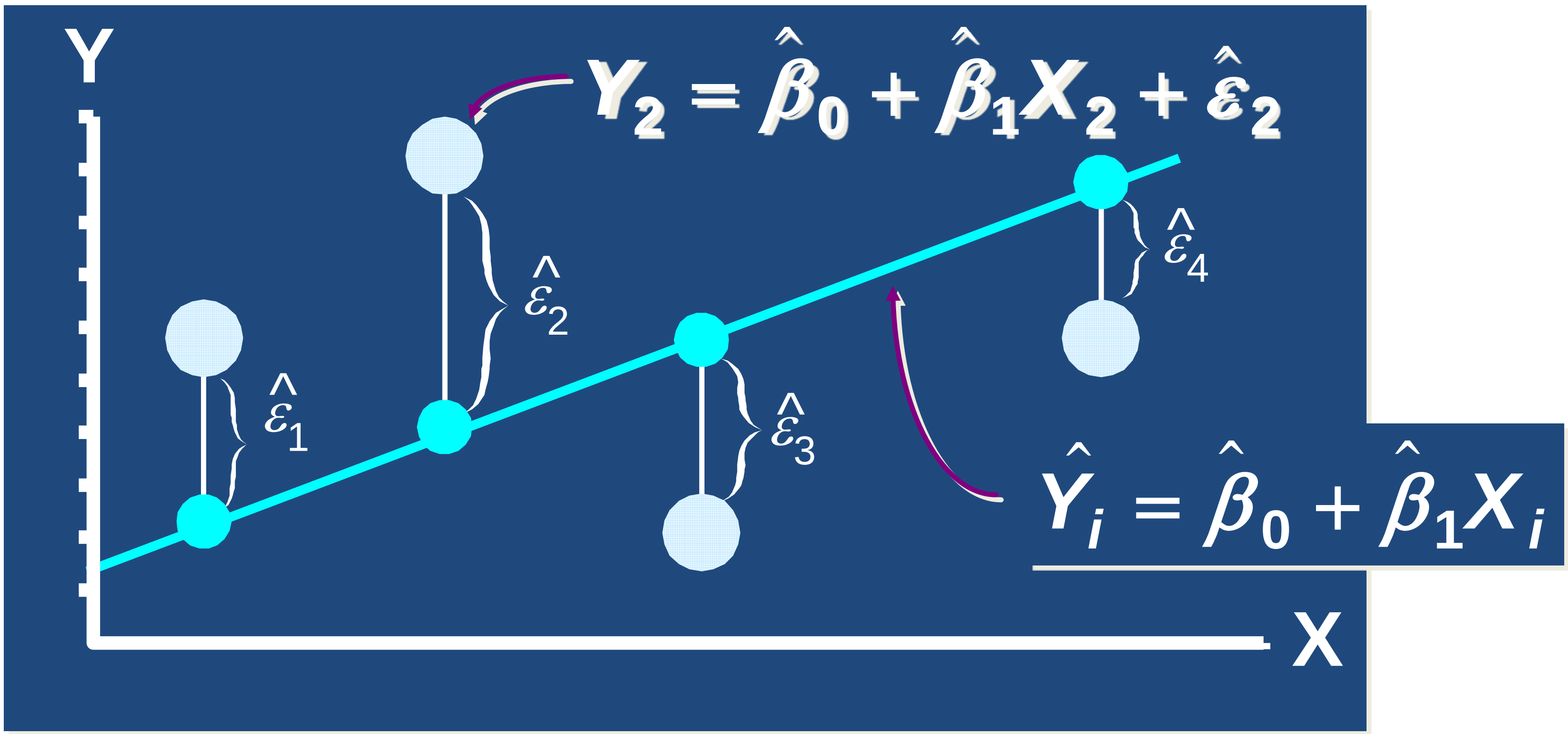
$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

Take derivative

Set to 0 and solve

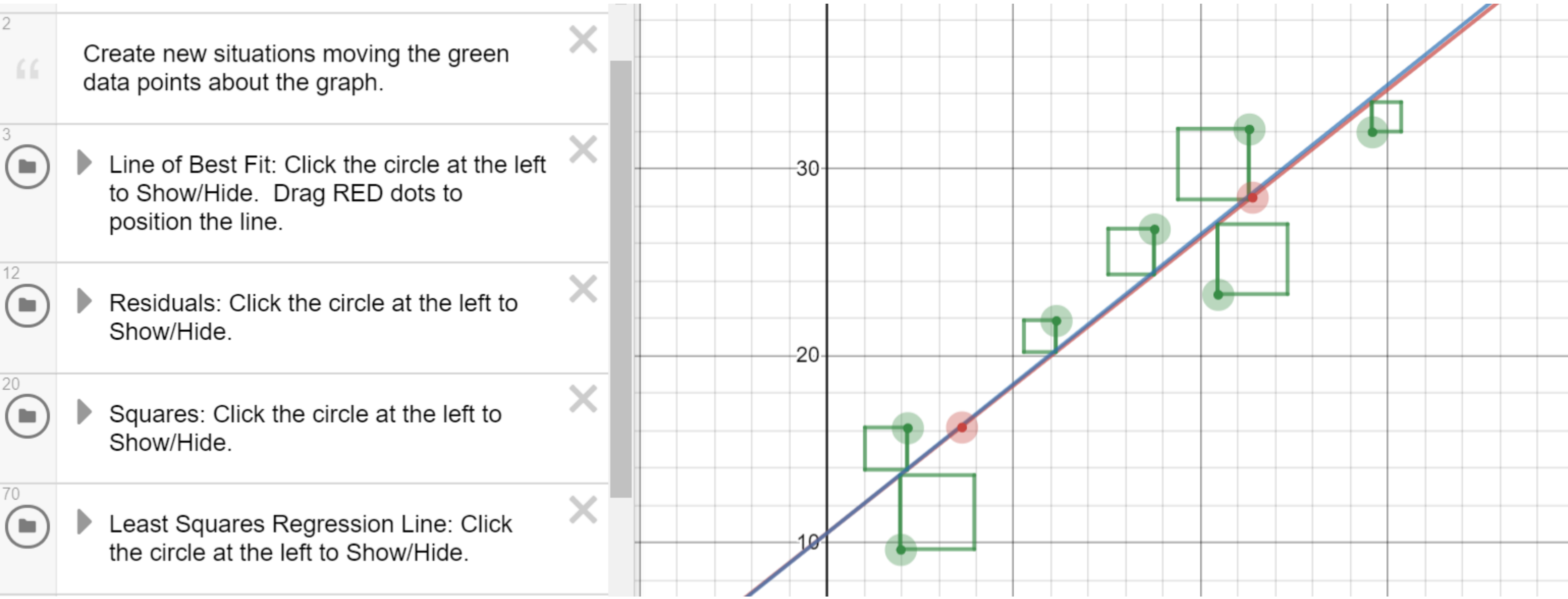
Least Squares (LS) Line Graphically

$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$$



https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Least Squares Line Graphically – Interactive Demo

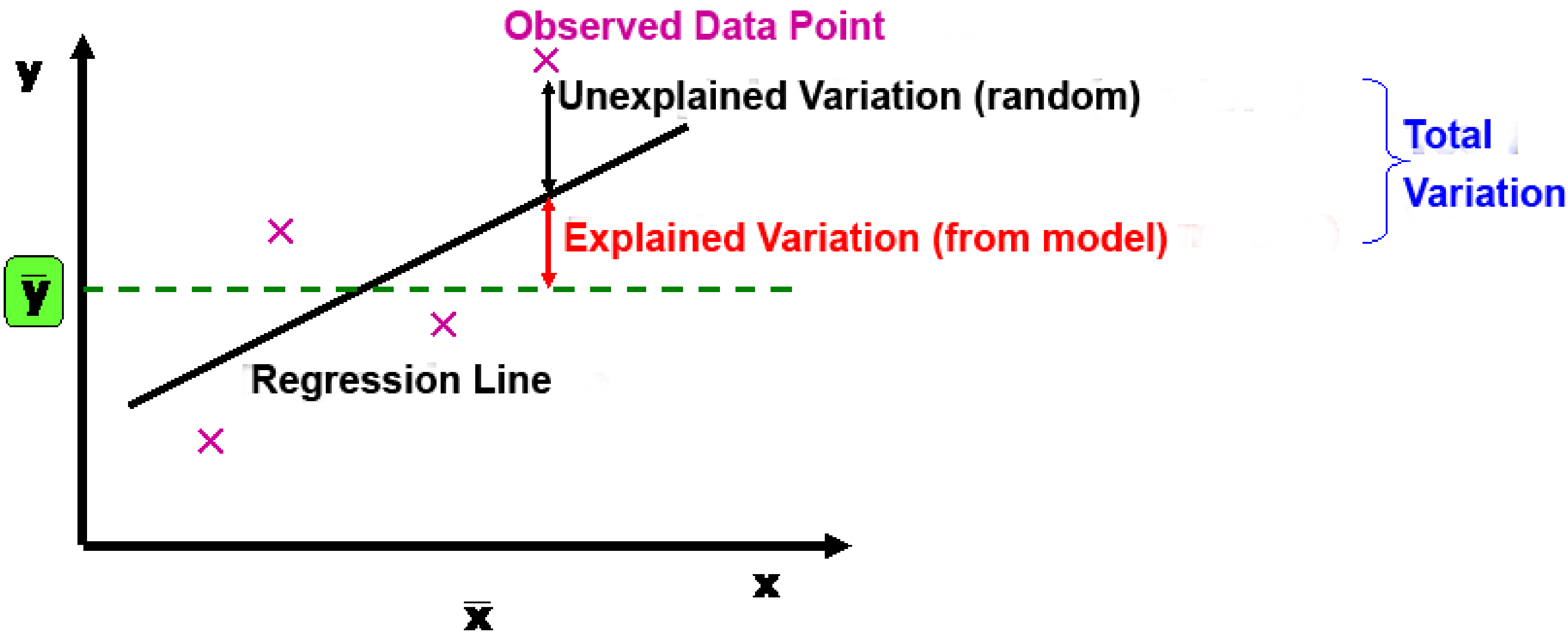


https://www.desmos.com/calculator/zvrc4lg3cr

Outline

- Introduction (done)
- Simple Linear Regression (done)
- Measures of Variation (next)
  - Coefficient of Determination
  - Correlation
- Misc

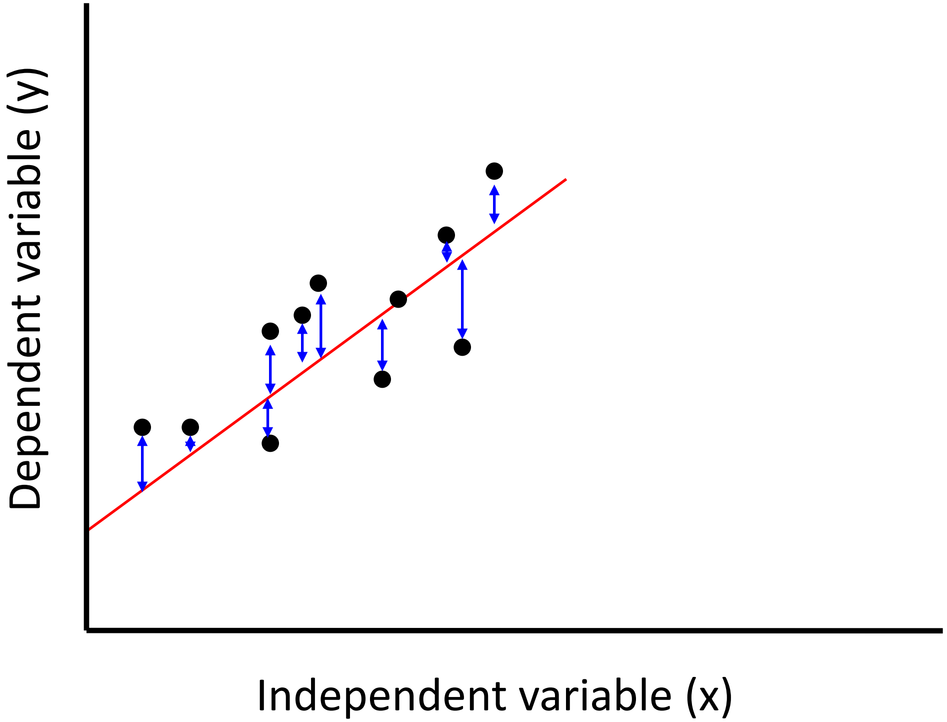
Measures of Variation



- Several sources of variation in y
  - Error in prediction (unexplained)
  - Variation from model (explained)

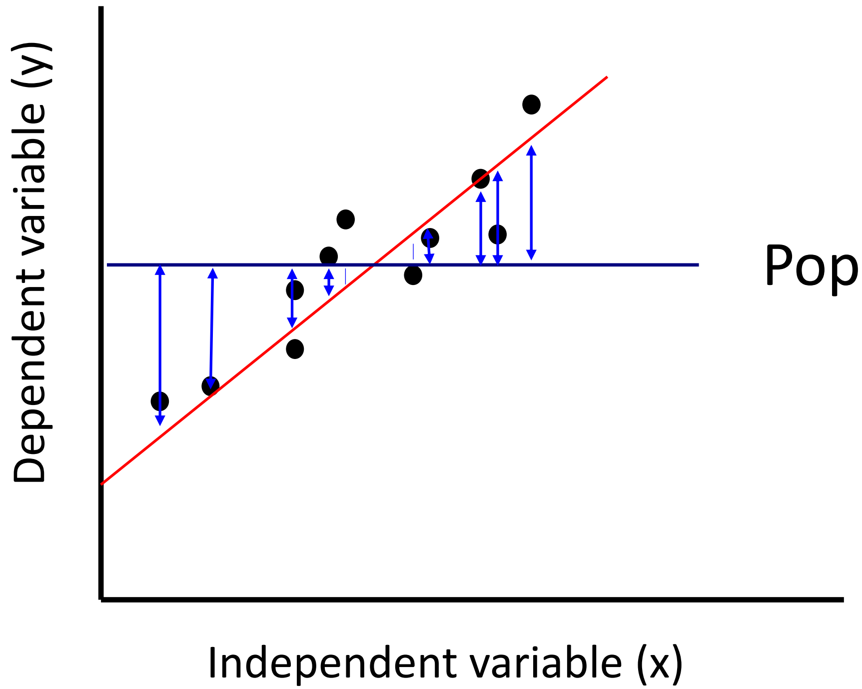
Break this down (next)

Sum of Squares of Error (SSE)



- Least squares regression selects line with lowest total sum of squared prediction errors
- Sum of Squares of Error, or SSE
- Measure of unexplained variation

Sum of Squares Regression (SSR)

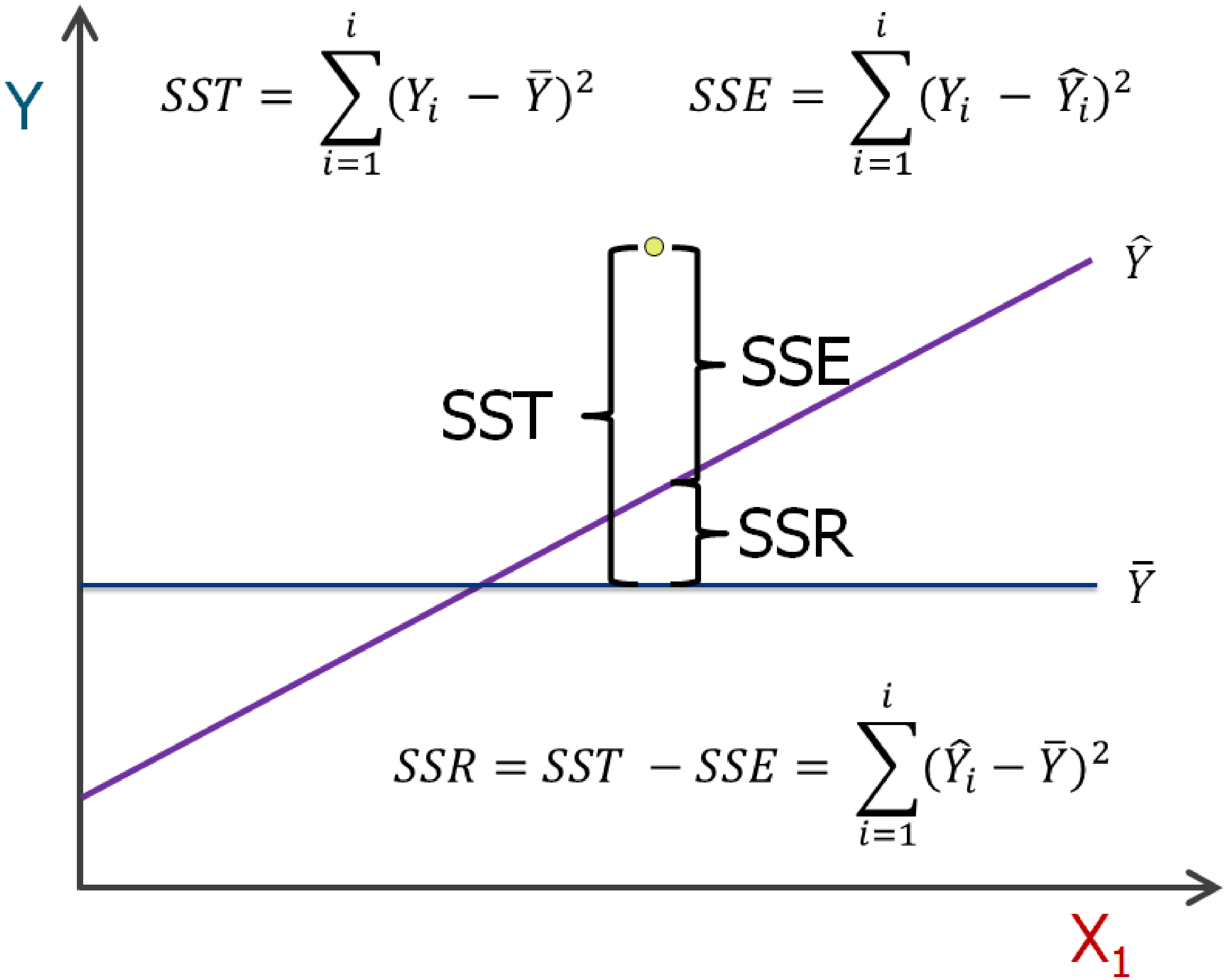


Population mean: $\bar{y}$

- Differences between prediction and population mean
  - Gets at variation due to X & Y
- Sum of Squares Regression, or SSR
- Measure of explained variation

Sum of Squares Total

- Total Sum of Squares, or SST = SSR + SSE



Coefficient of Determination

- Proportion of total variation (SST) explained by the regression (SSR) is known as the Coefficient of Determination (R^2)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

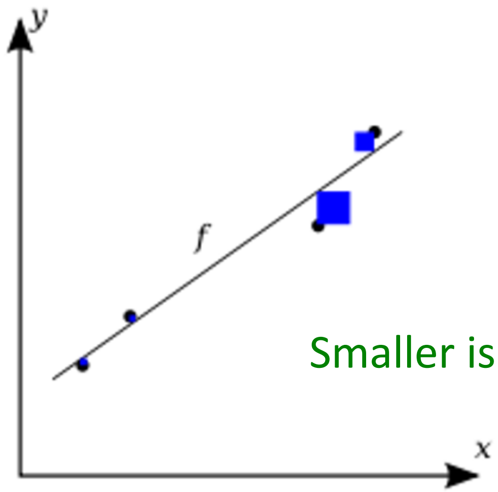
- Ranges from 0 to 1 (often said as a percent)

  1 – regression explains all of variation

  0 – regression explains none of variation

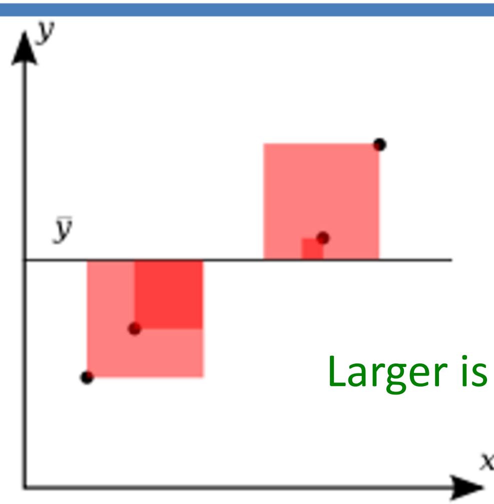
Coefficient of Determination – Visual Representation

$R^2 = 1 -$



Smaller is better

Variation in observed data model cannot explain (error)

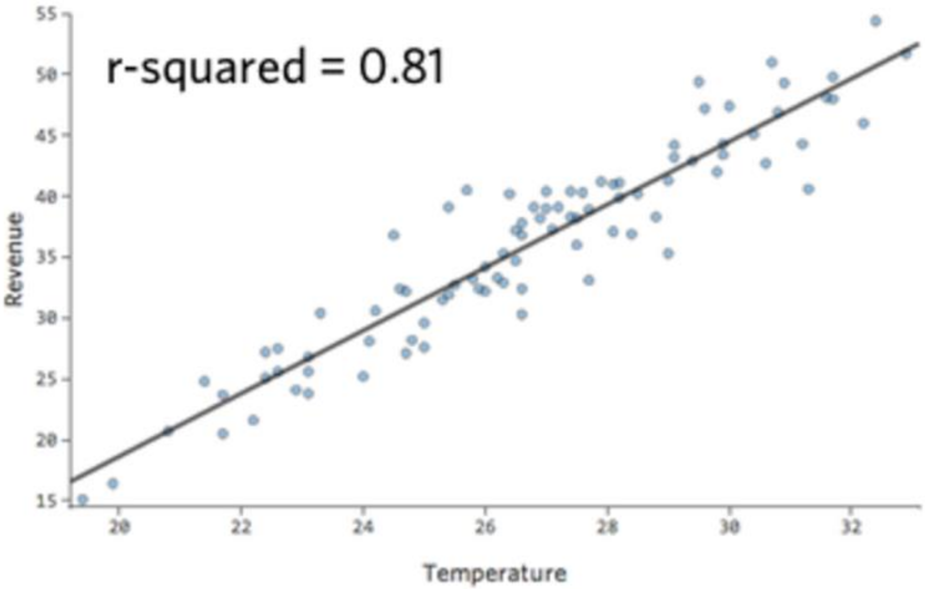


Larger is better

Total variation in observed data

https://upload.wikimedia.org/wikipedia/commons/thumb/8/86/Coefficient_of_Determination.svg/400px-Coefficient_of_Determination.svg.png

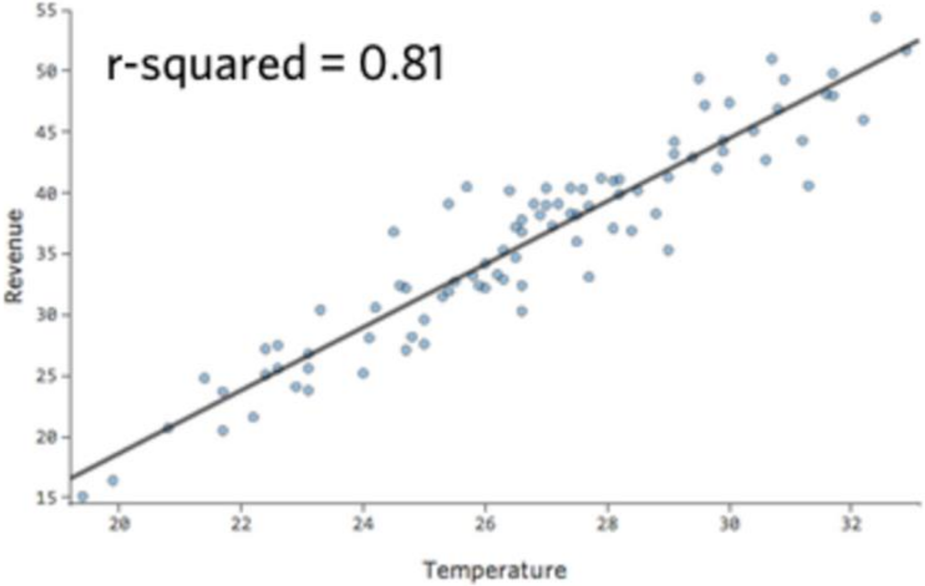
Coefficient of Determination Example

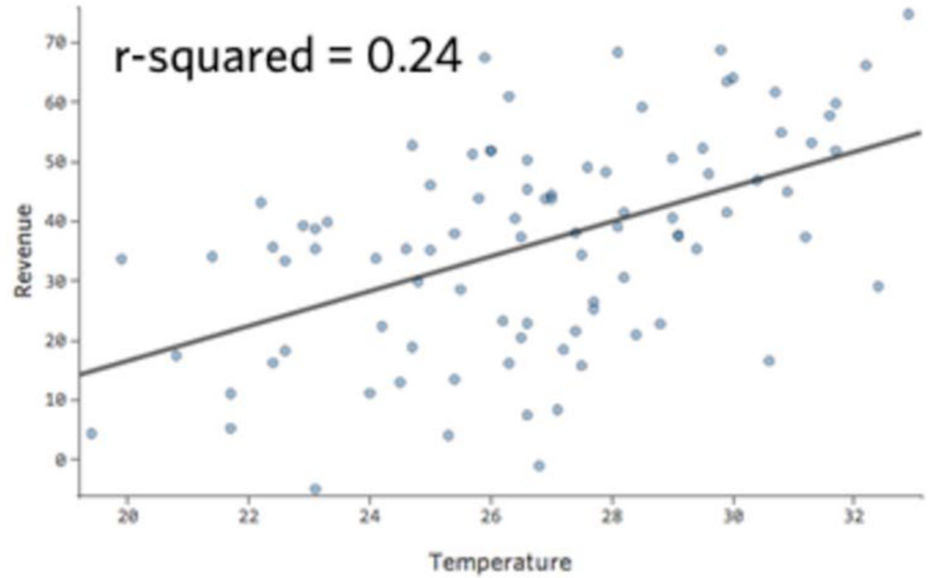


- How "good" is regression model? Roughly:

  $0.8 <= R^2 <= 1$ strong

Coefficient of Determination Example



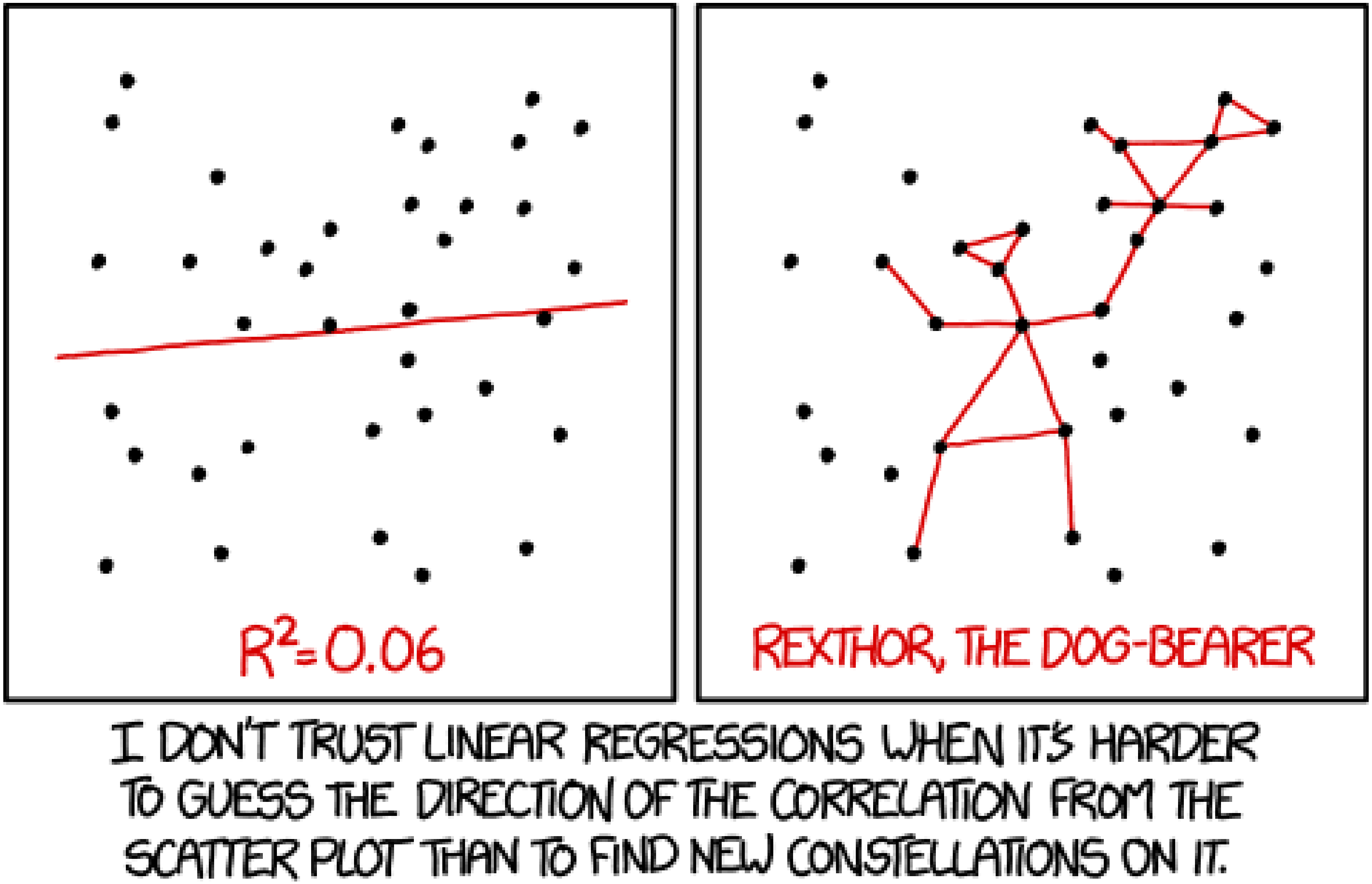


- How “good” is regression model? Roughly:

  0.8 <= R^2 <= 1 — strong

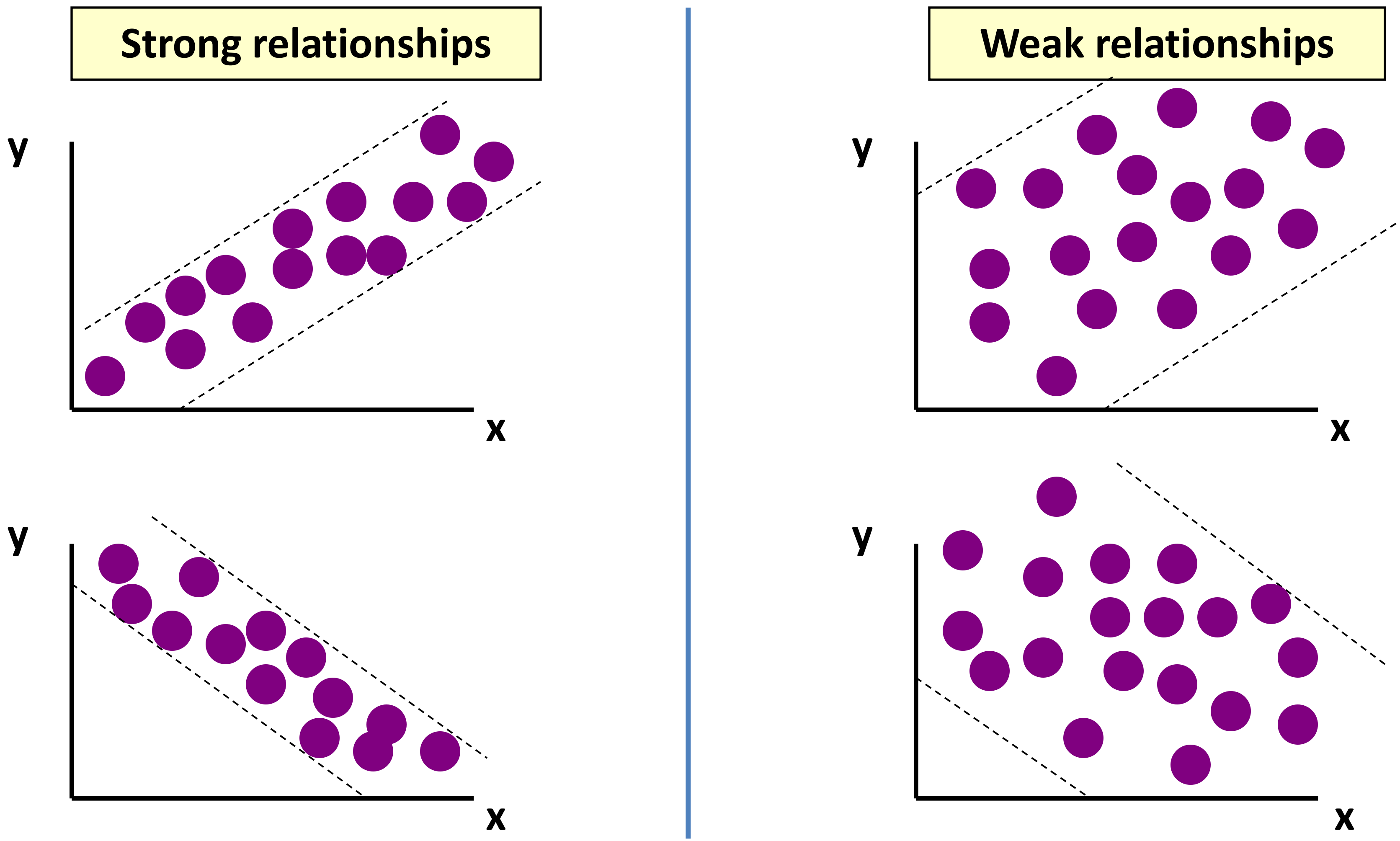
  0 <= R^2 < 0.5 — weak

How “good” is the Regression Model?



https://xkcd.com/1725/

Relationships Between X & Y



Relationship Strength and Direction – Correlation

- Correlation measures strength and direction of linear relationship
  - -1 perfect neg. to +1 perfect pos.
  - – Sign is same as regression slope
  - – Denoted R. Why? Square R = R^2

Pearson's Correlation Coefficient

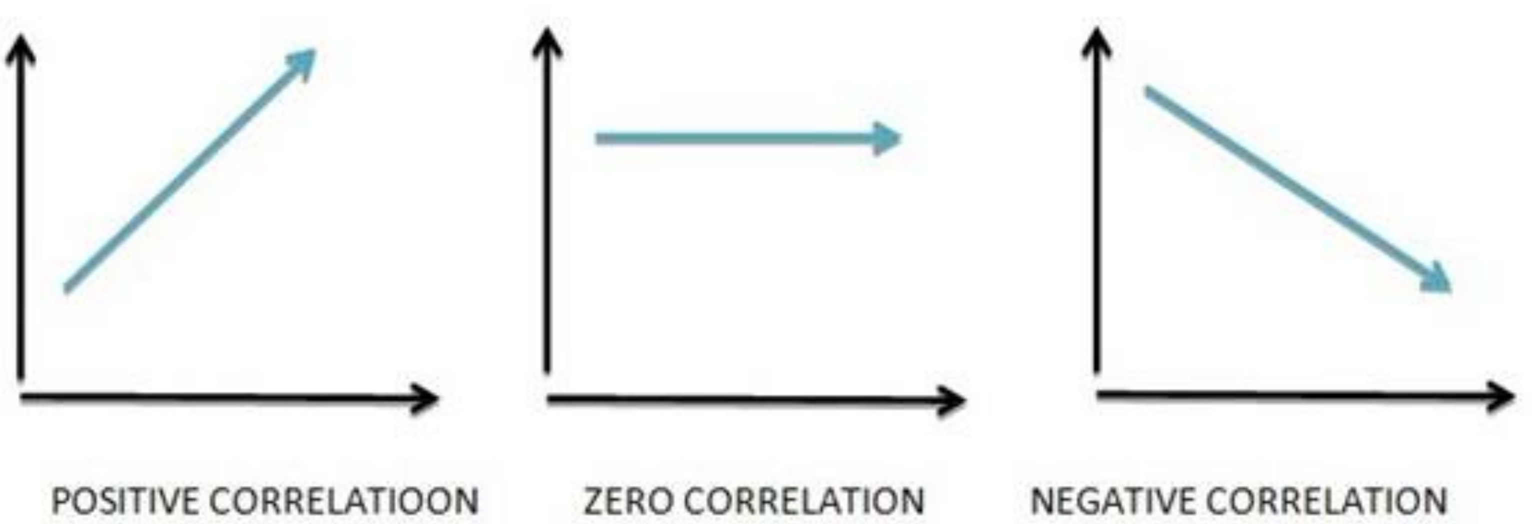
$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Vary together (numerator)

Vary Separately (denominator)

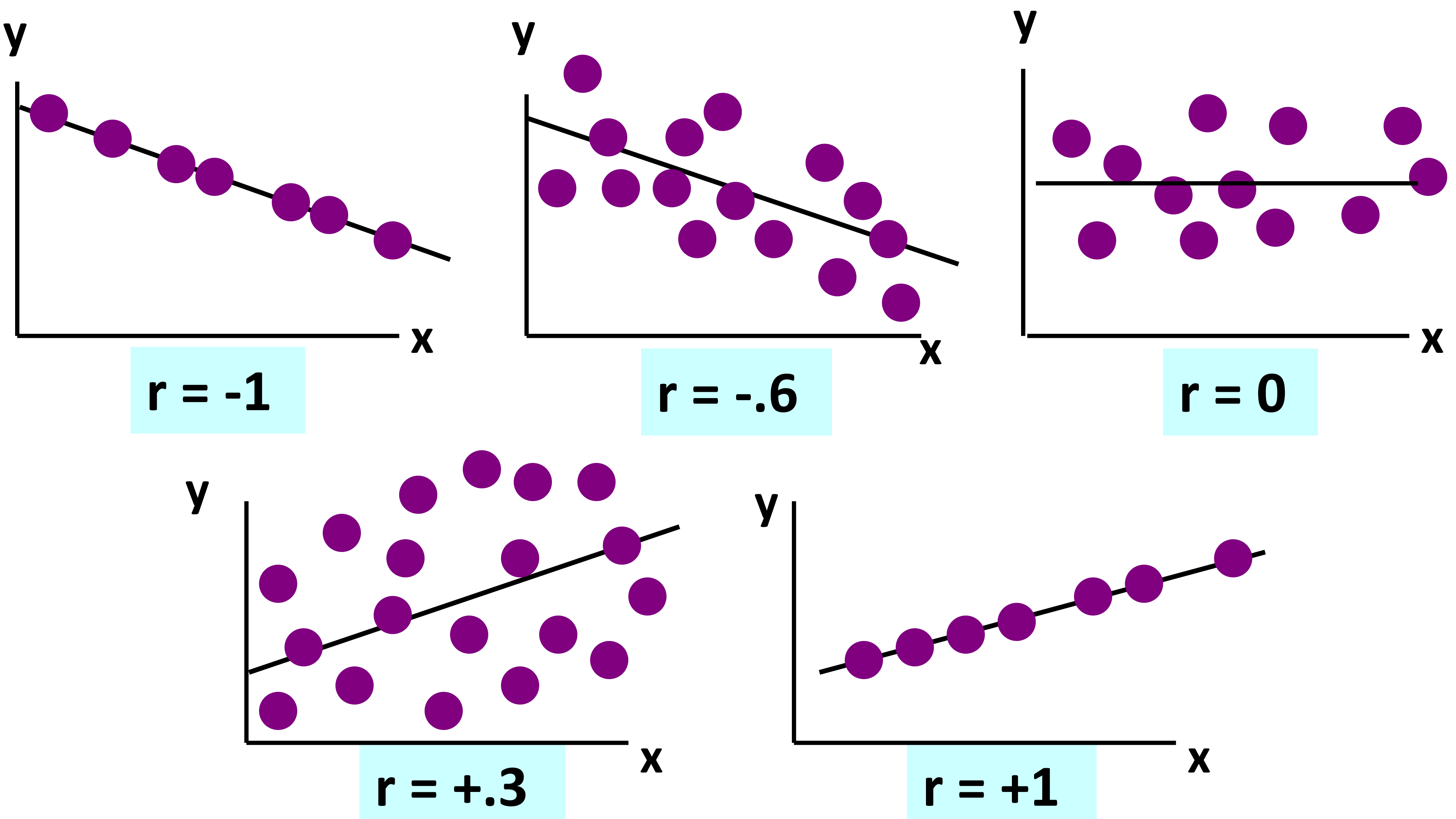
Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable



https://www.mbaskool.com/2013_images/stories/dec_images/pearson-coeff-bcon.jpg

Correlation Examples

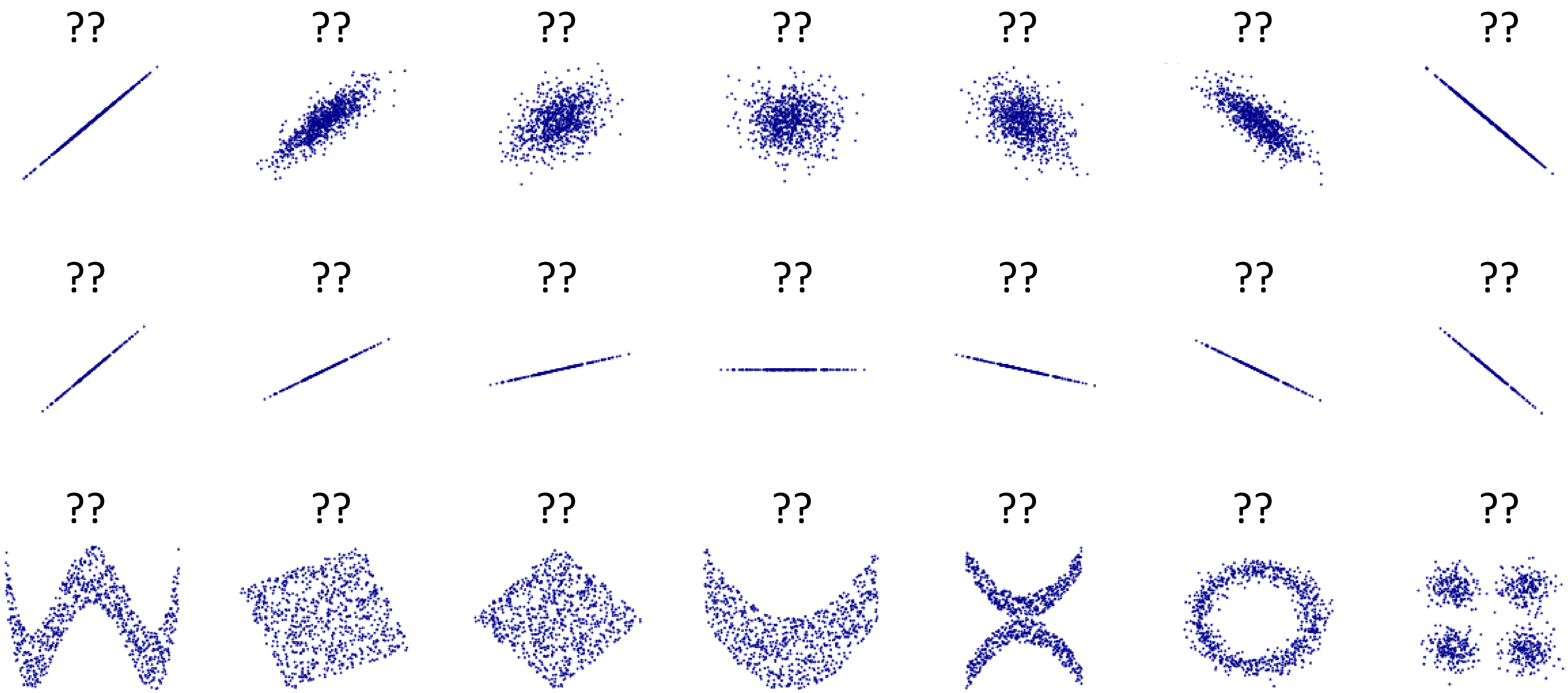


Groupwork

- Introduction
  - **Icebreaker**: What game are you looking forward to playing this summer?
- Groupwork
  - Think, discuss, write down – qualtrics
- Correlation
  - Consider scatterplots
  - Estimate correlation

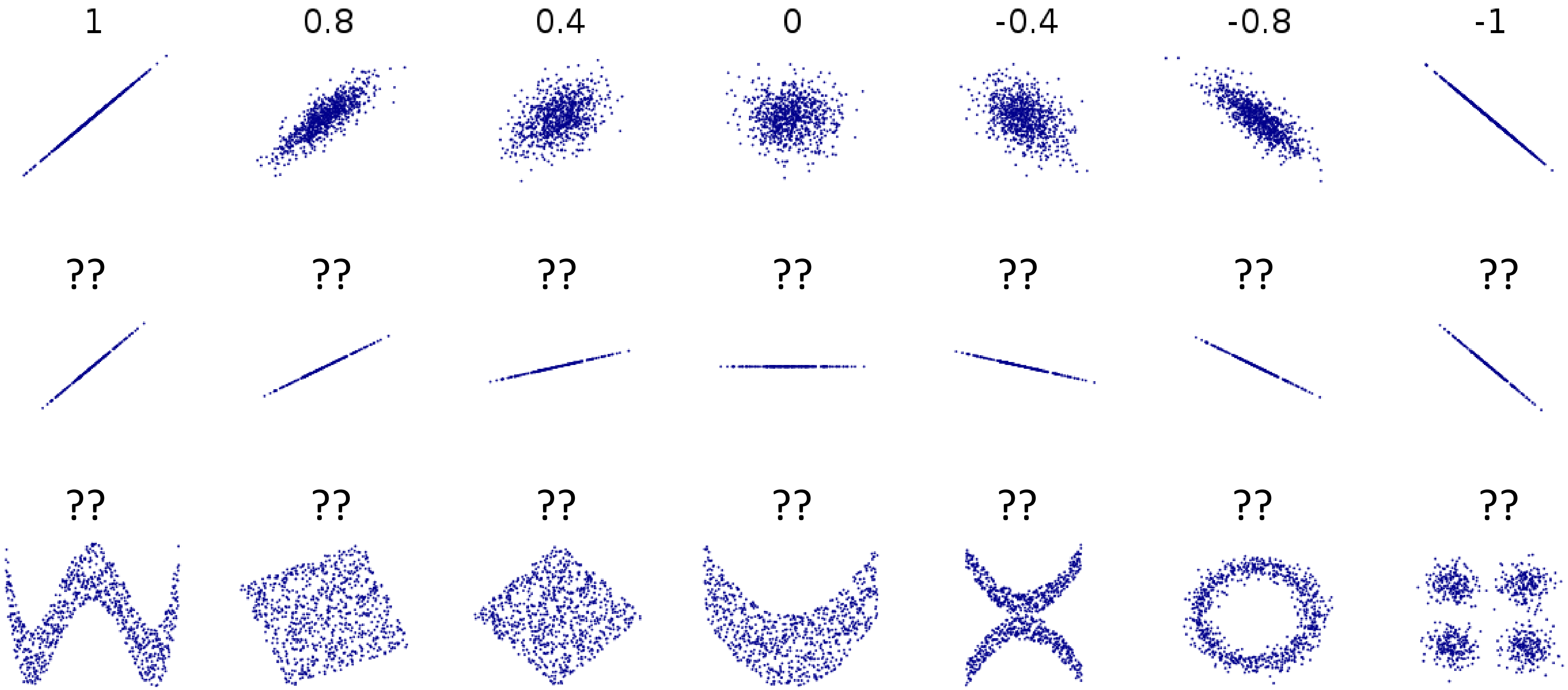
https://web.cs.wpi.edu/~imgd2905/d22/groupwork/12-correlation/handout.html

Correlation Examples



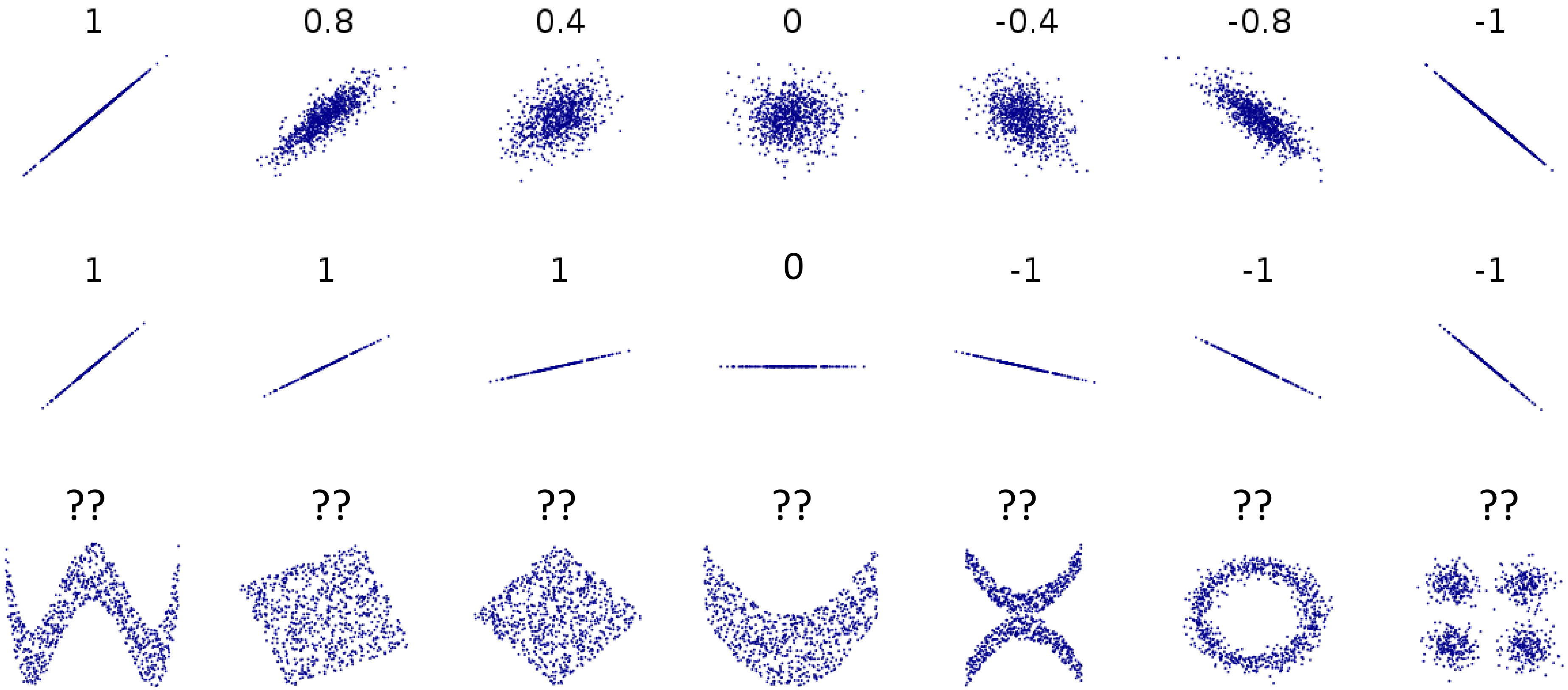
https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png

Correlation Examples



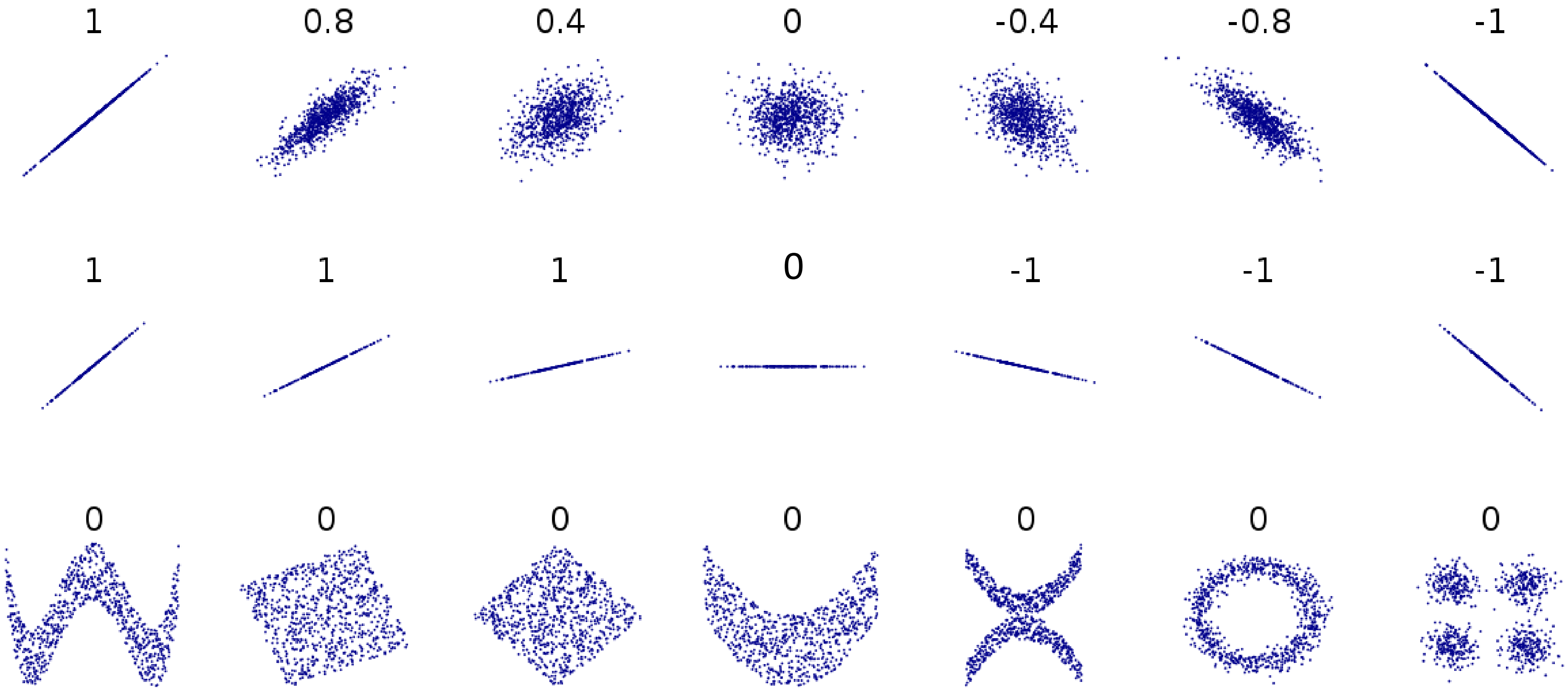
https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png

Correlation Examples



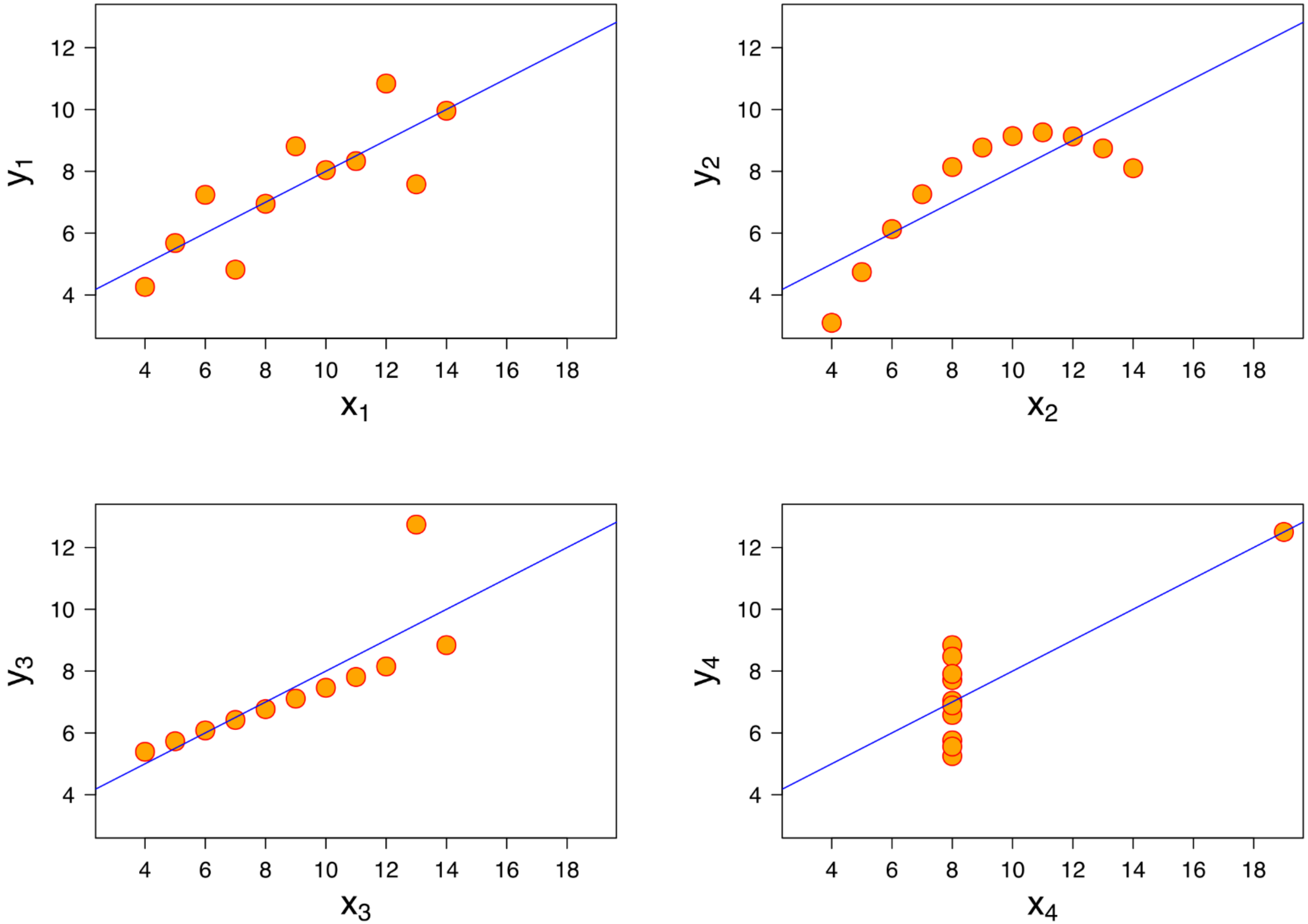
https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png

Correlation Examples



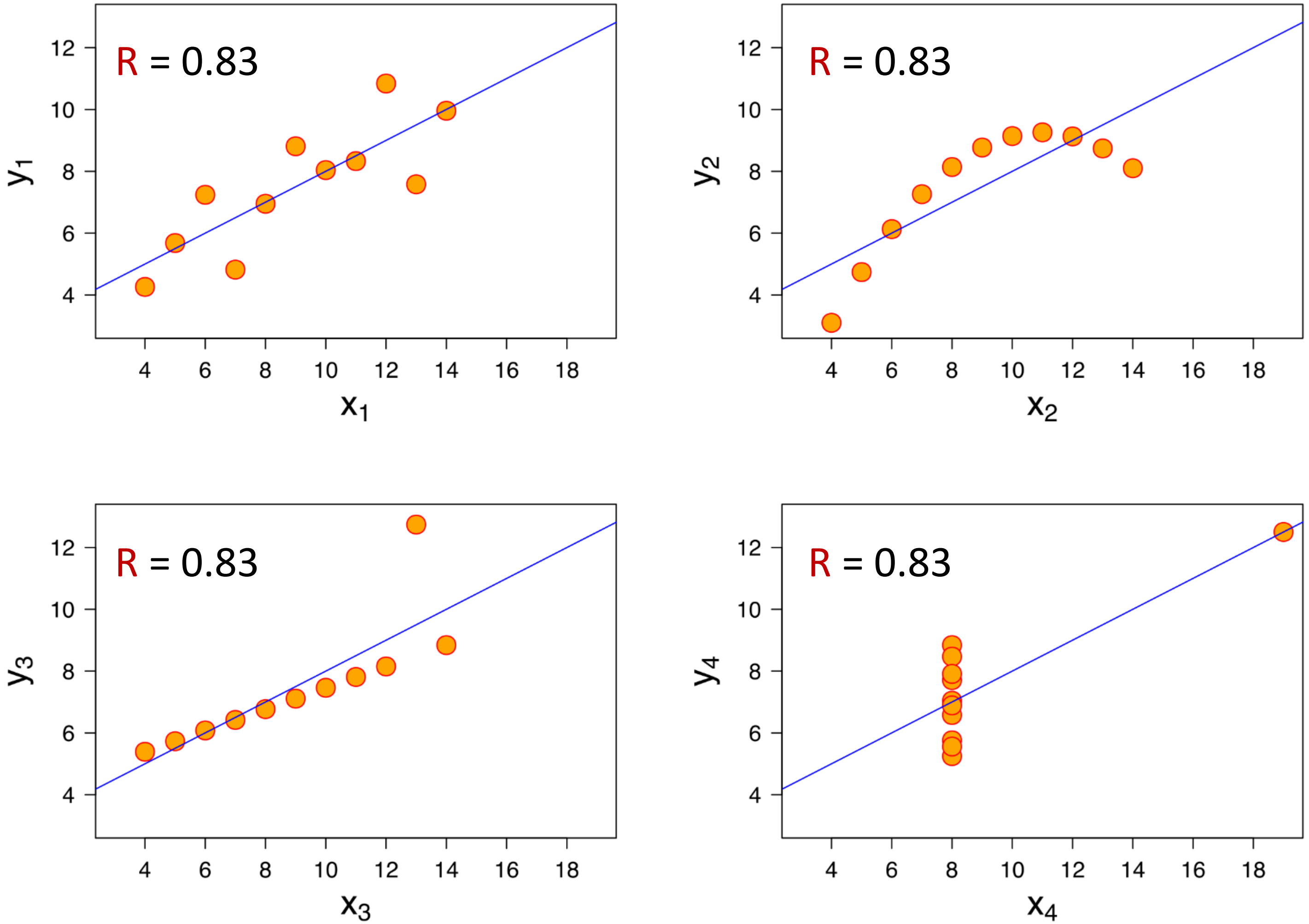
https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1200px-Correlation_examples2.svg.png

Correlation Examples



(Note, would want to use residual analysis before using predictions!)

Correlation Examples



(Note, would want to use residual analysis before using predictions!)

Anscombe's Quartet

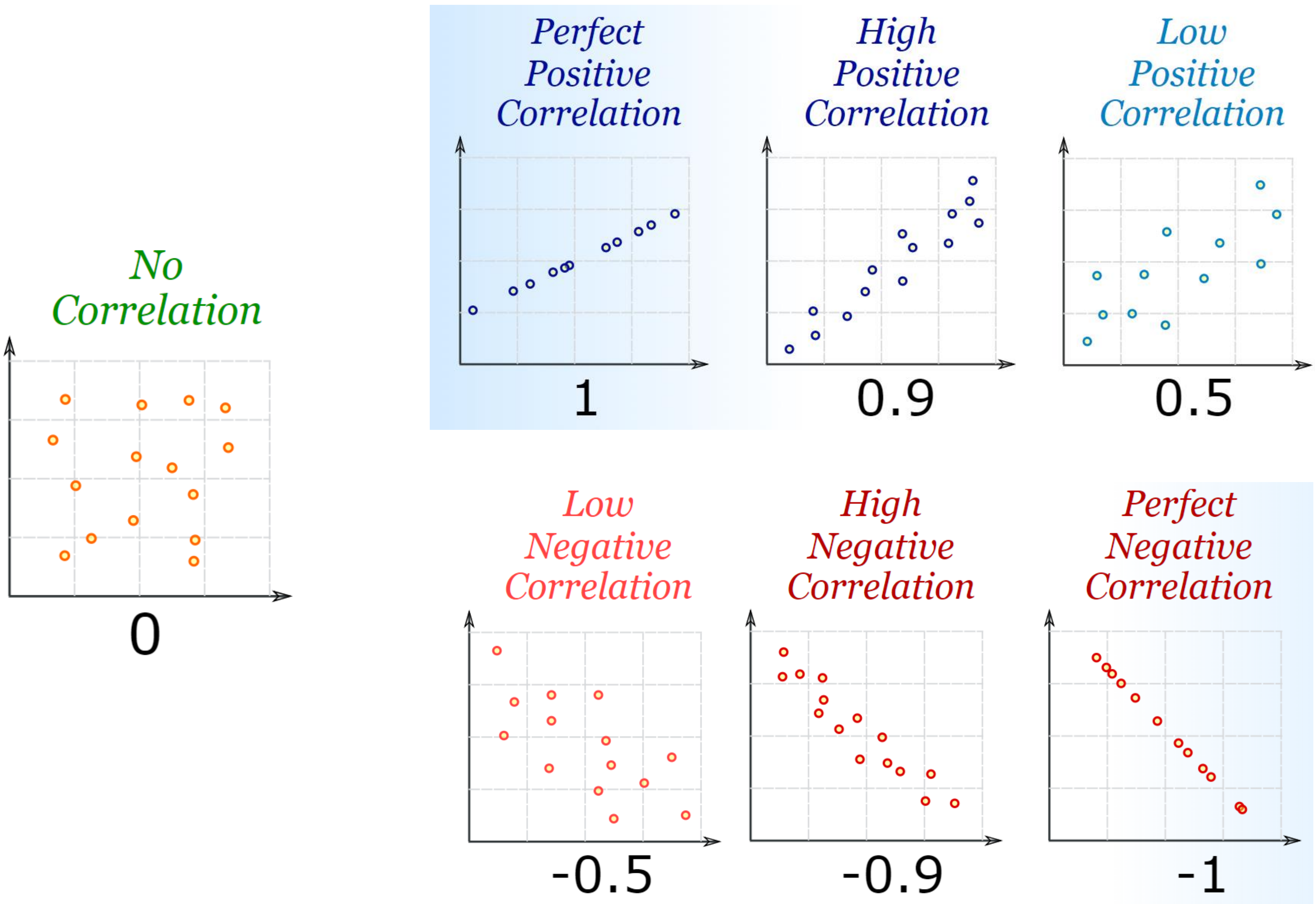
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Summary stats:

$Mean_x$	9
$Mean_y$	7.5
Var_x	11
Var_y	4.125
Model:	y=0.5x+3

$R^2 = 0.69$

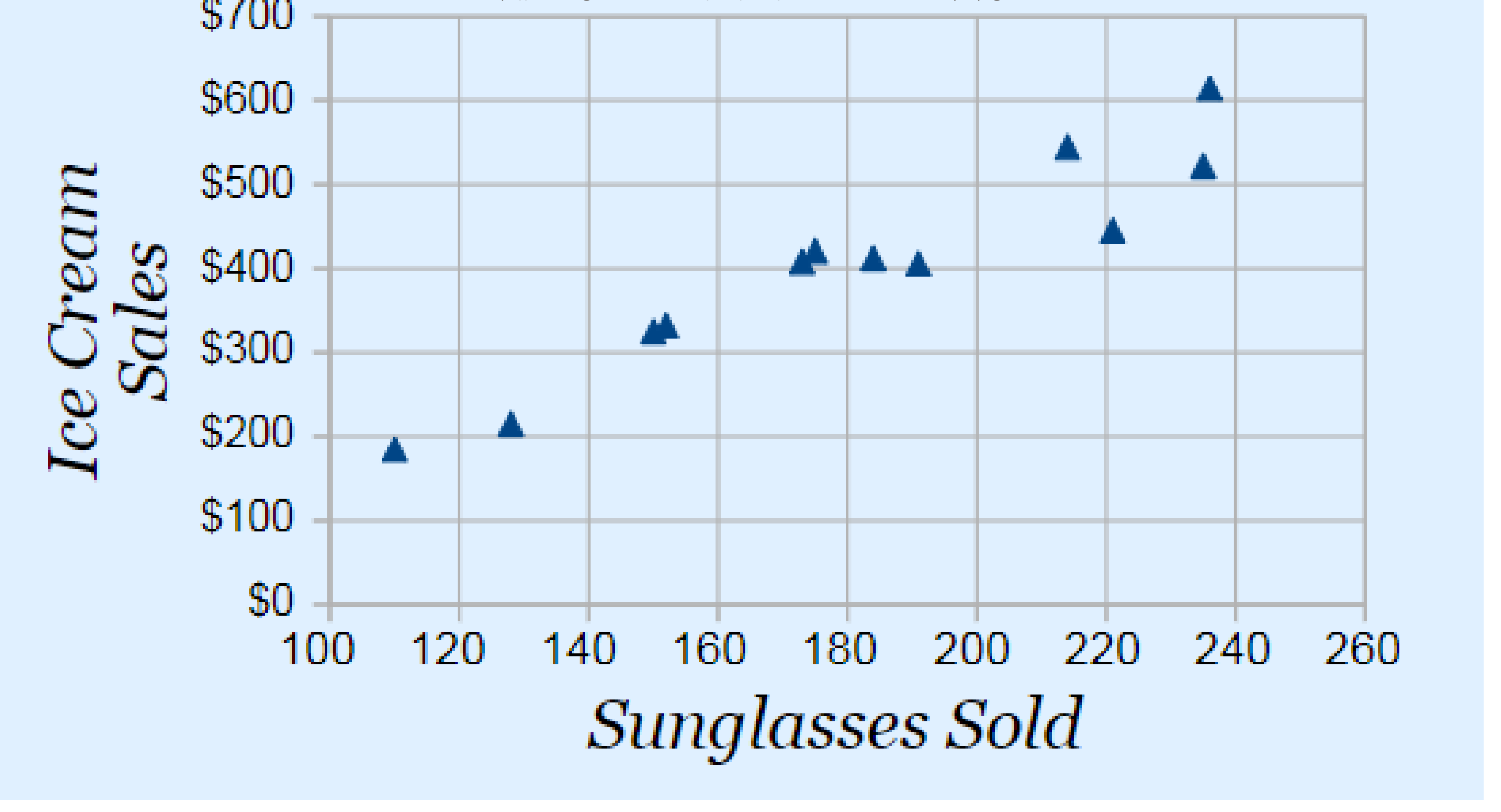
Correlation Summary



https://www.mathsisfun.com/data/correlation.html

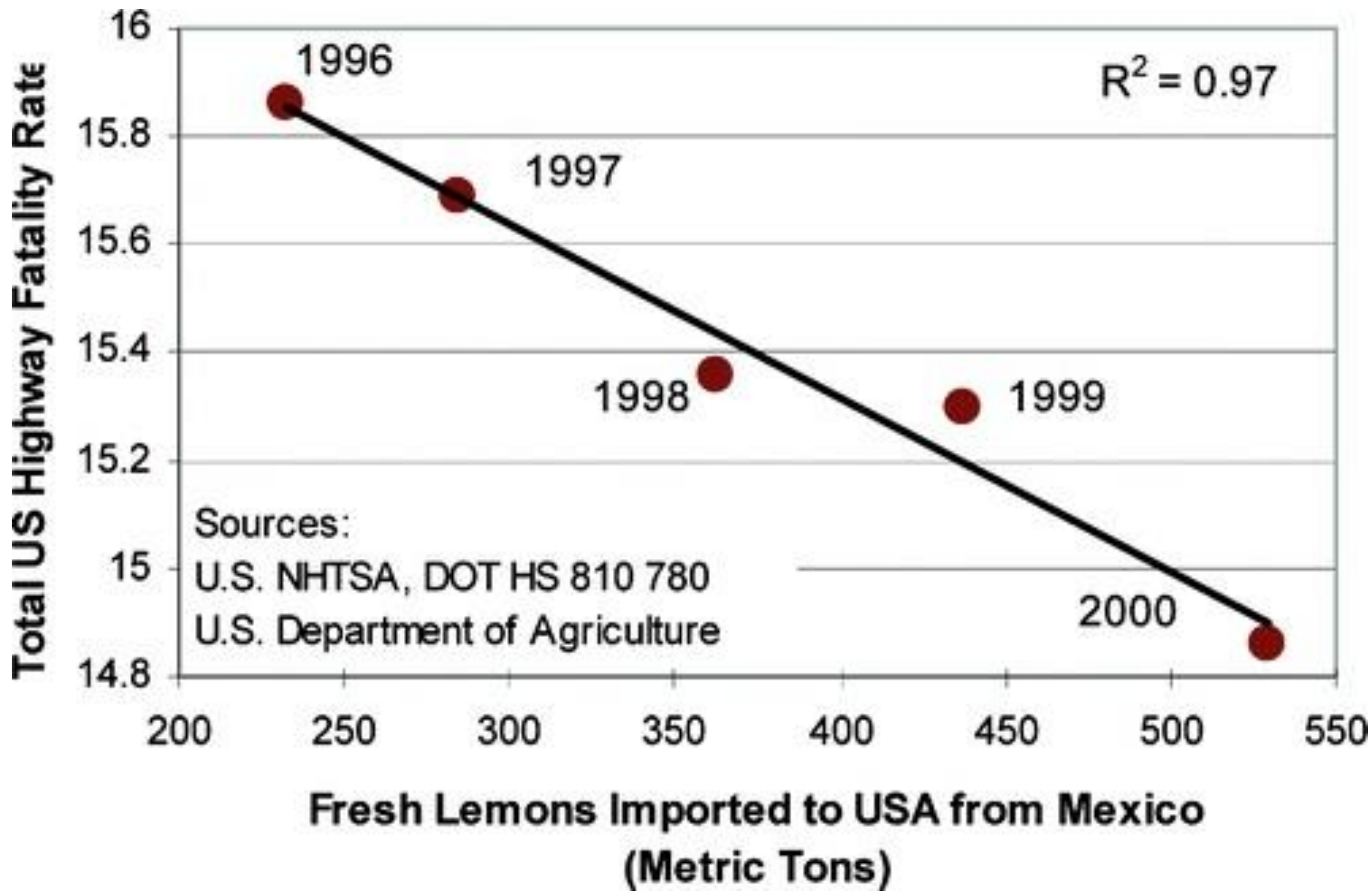
Correlation is not Causation

https://cdn-images-1.medium.com/max/1600/1*JLYI5eCVEN7ZUWXBIrrapw.png



Buying sunglasses *causes* people to buy ice cream?

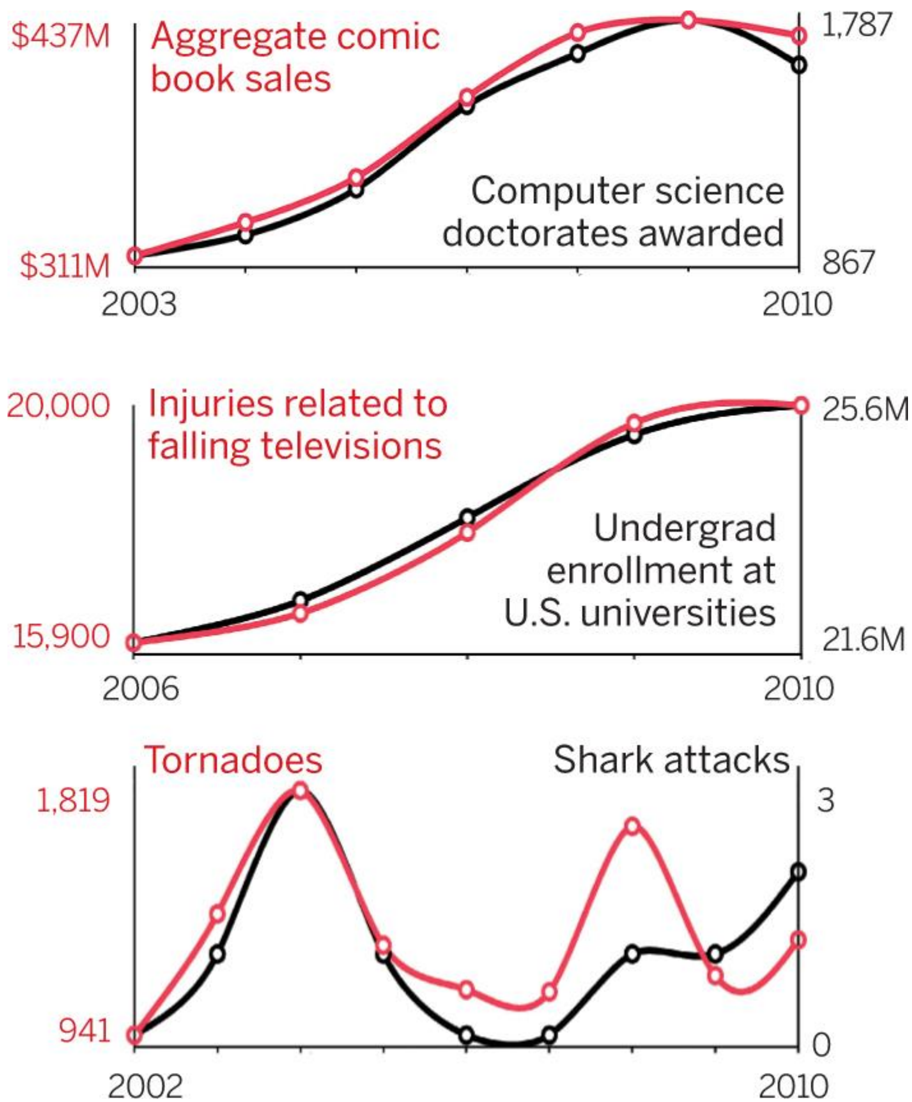
Correlation is not Causation



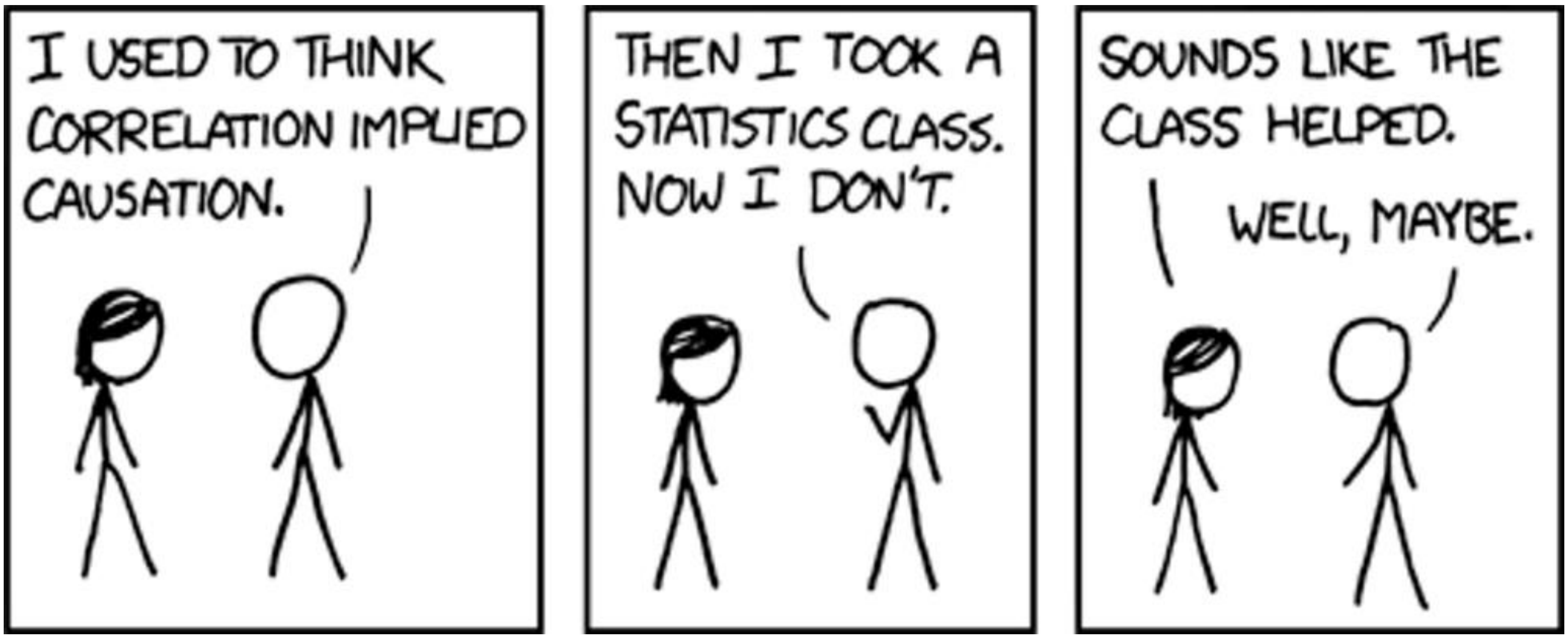
Importing lemons causes fewer highway fatalities?

Correlation is not Causation

https://science.sciencemag.org/content/sci/348/6238/980.2/F1.large.jpg?width=800&height=600&carousel=1



Correlation is not Causation



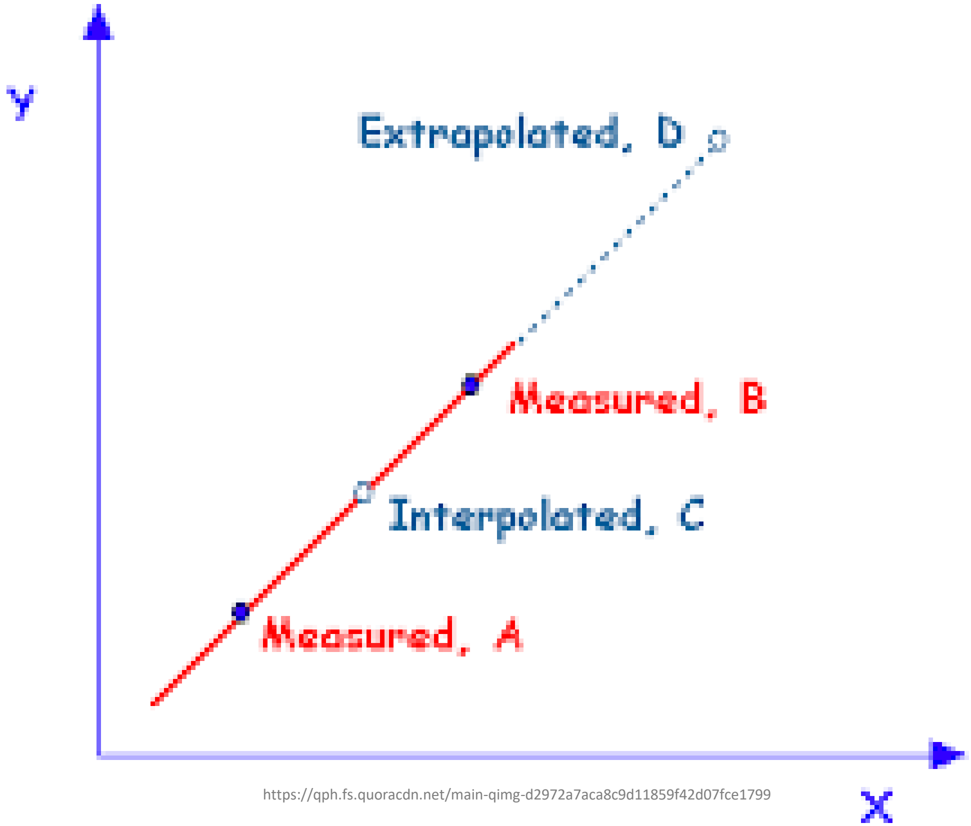
https://xkcd.com/552/

Outline

- Introduction (done)
- Simple Linear Regression (done)
- Measures of Variation (done)
- Misc (next)

Extrapolation versus Interpolation

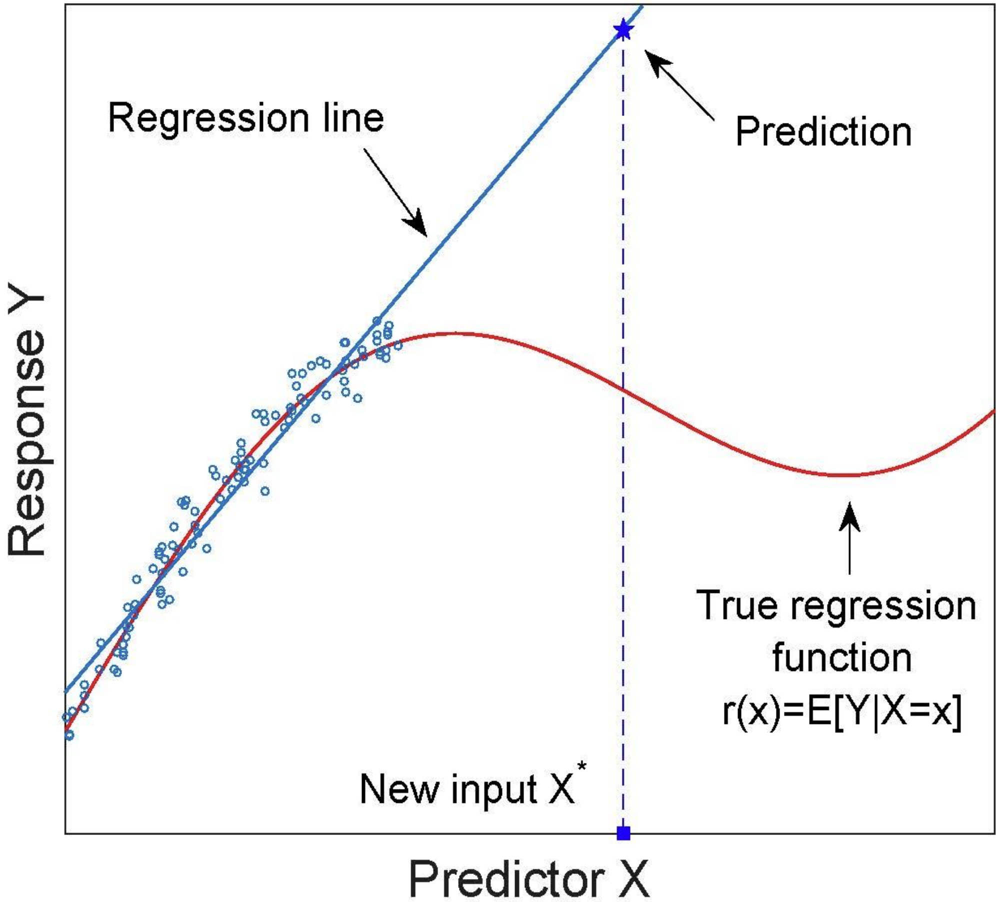
- Prediction
  - Interpolation – within measured X-range
  - Extrapolation – outside measured X-range



https://qph.fs.quoracdn.net/main-qimg-d2972a7aca8c9d11859f42d07fce1799

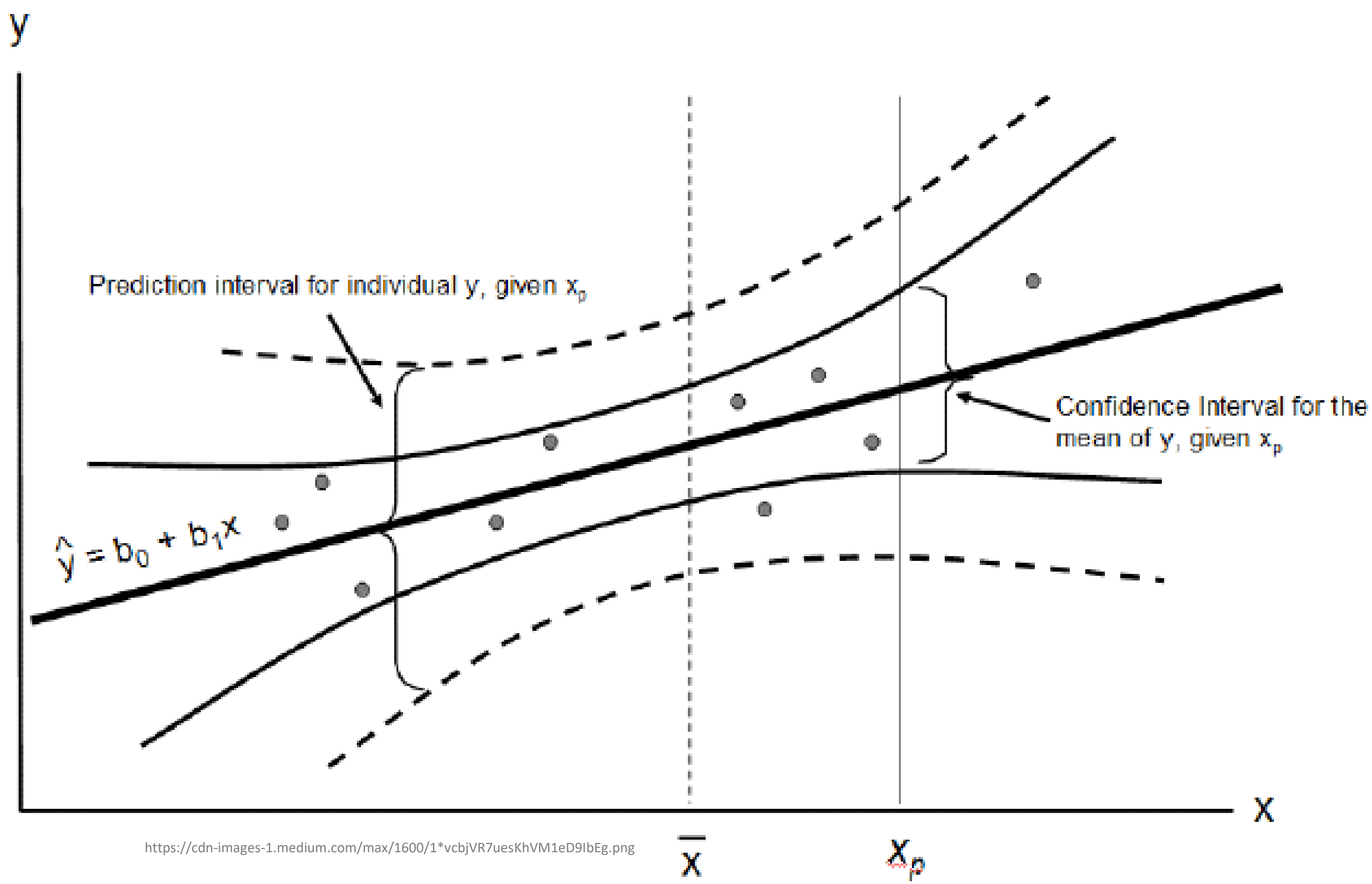
Be Careful When Extrapolating

https://i.stack.imgur.com/3Ab7e.jpg



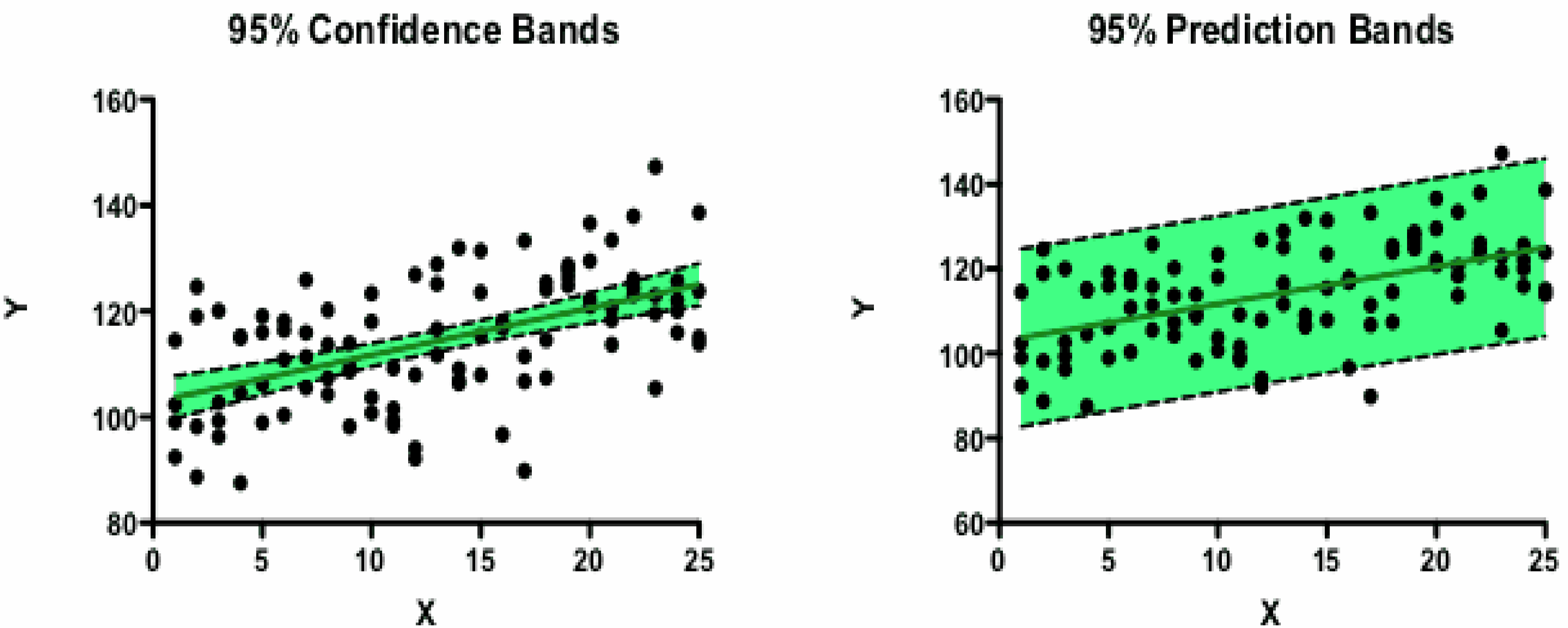
If extrapolate, make sure have reason to assume model continues

Prediction and Confidence Intervals (1 of 2)



https://cdn-images-1.medium.com/max/1600/1*vcbjVR7uesKhVM1eD9IbEg.png

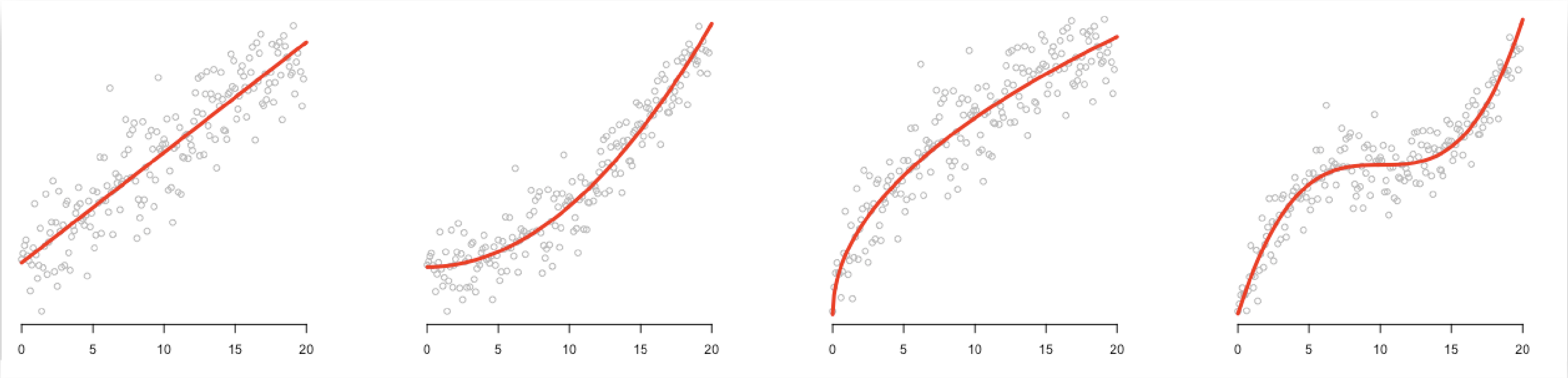
Prediction and Confidence Intervals (2 of 2)



https://www.graphpad.com/guides/prism/7/curve-fitting/reg_mostpointsareoutsideconfidencebands.png

Beyond Simple Linear Regression

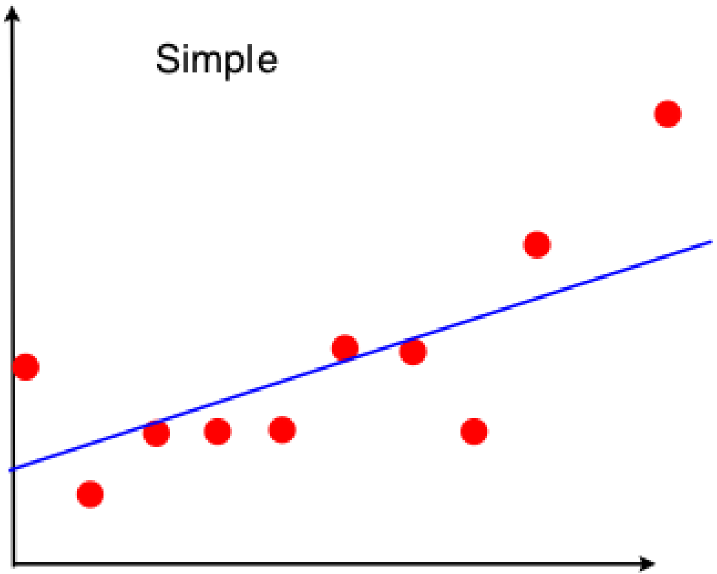
https://medium.freecodecamp.org/learn-how-to-improve-your-linear-models-8294bfa8a731



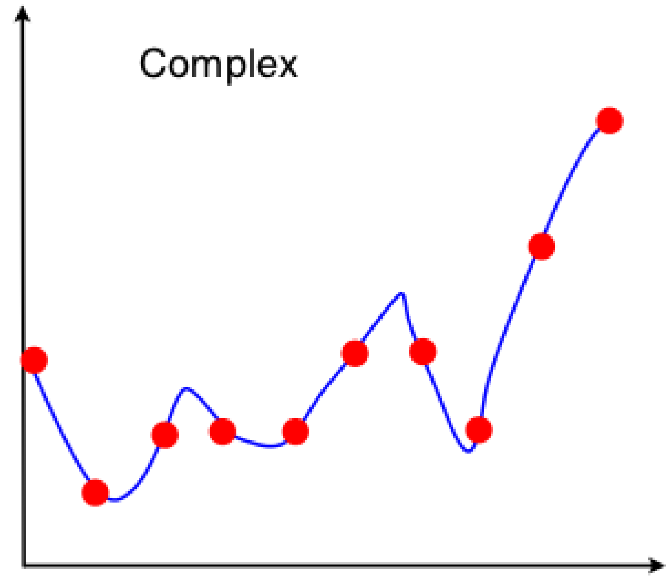
Linear Quadratic Root Cubic

- Multiple regression – more parameters beyond just X
  - Book Chapter 11
- More complex models – beyond just $Y = mX + b$

More Complex Models



$y = 12x + 9$



$y = 18x^4 + 13x^3 - 9x^2 + 3x + 20$

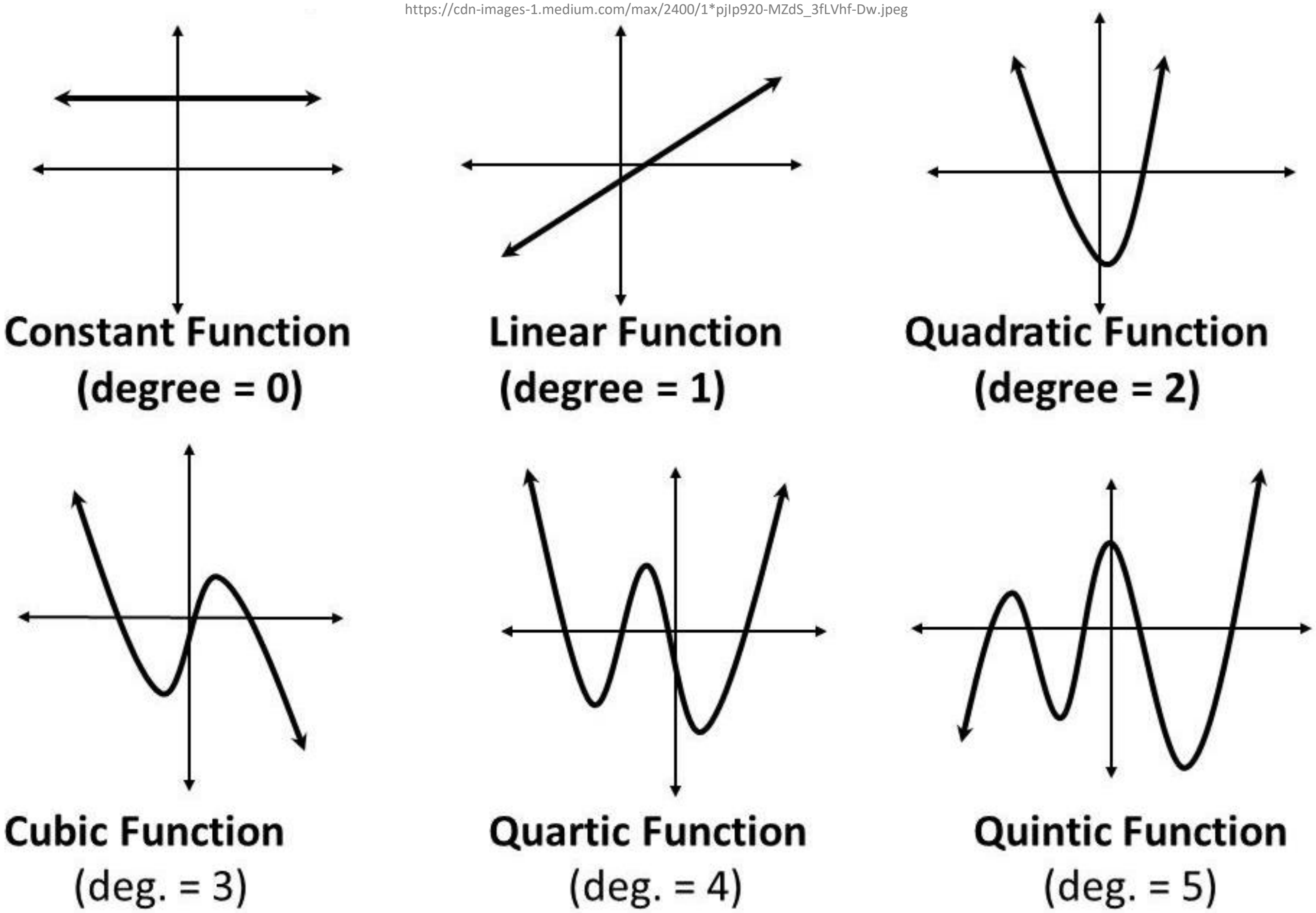
- Higher order polynomial model has less error

→ A "perfect" fit (no error)

- How does a polynomial do this?

Graphs of Polynomial Functions

https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg

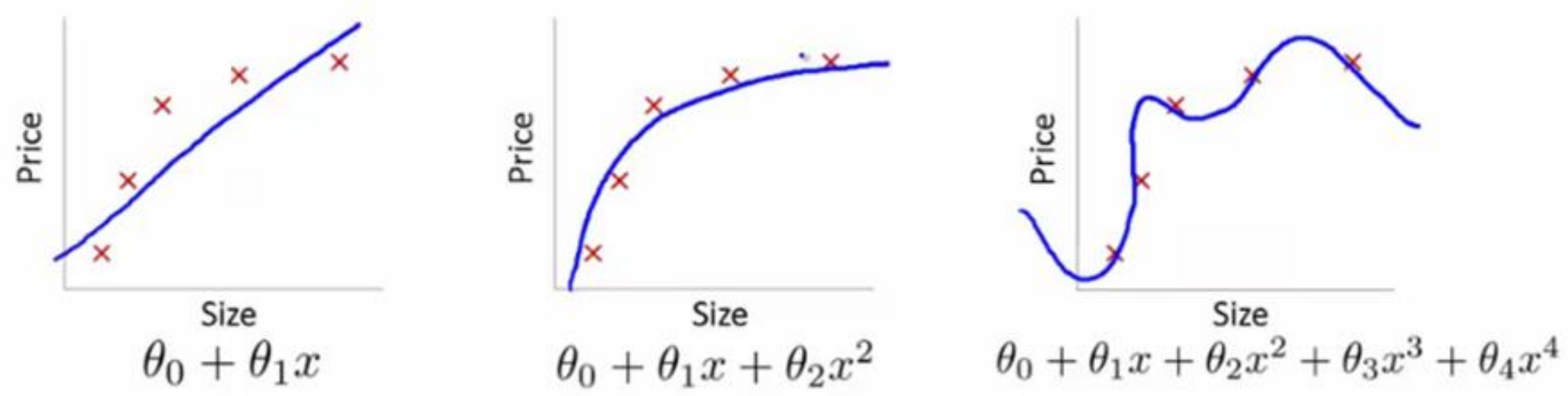


Higher degree, more potential “wiggles”

But should you use?

Underfit and Overfit

https://i.stack.imgur.com/t0zit.png



Underfit | Just Right | Overfit

- Overfit analysis matches data too closely with more parameters than can be justified
- Underfit analysis does not adequately match data since parameters are missing

→ Both models fit well, but do not *predict* well (i.e., for non-observed values)

- Just right – fit data well "enough" with as few parameters as possible (*parsimonious* - desired level of prediction with as few terms as possible)

Summary

- Can use regression to predict un-measured values
- Before fit
  - Visual relationship (scatter plot) and residual analysis
- Strength of fit – R^2 and correlation (R)
- Beware
  - Correlation is not causation
  - Extrapolation
- Higher order, more complex models can fit better
  - Beware of overfit → less predictive power

https://d3h0owdjgzys62.cloudfront.net/images/3963/live_cover_art/thumb2x/summary_final.png