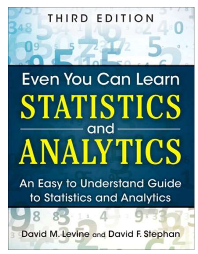
IMGD 2905

Simple Linear Regression

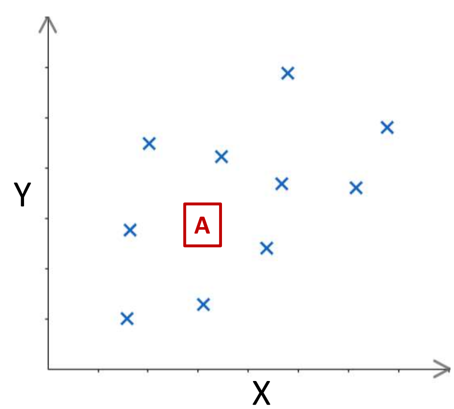
Chapter 10



1

Motivation

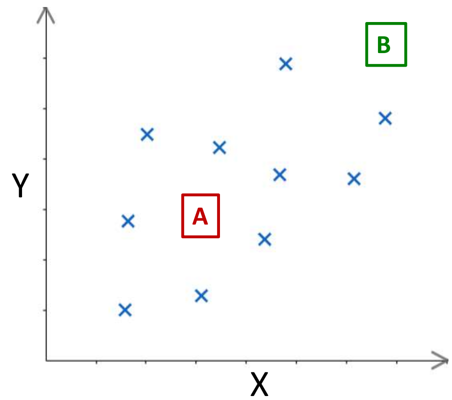
- Have data (sample, **x**'s)
- Want to know likely value of next observation
  - E.g., *playtime* versus *skins owned*
- A –



2

Motivation

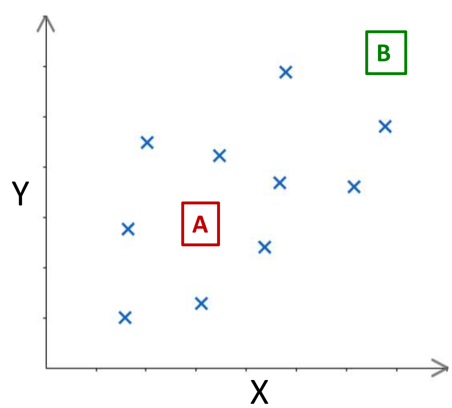
- Have data (sample, **x**'s)
- Want to know likely value of next observation
  - E.g., *playtime* versus *skins owned*
- A – reasonable to compute mean (with confidence interval)
- B –



3

Motivation

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



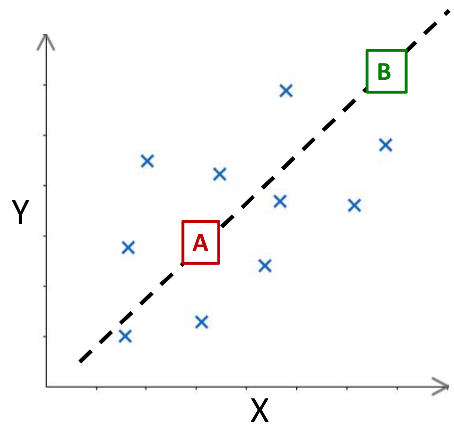
4

Motivation

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

→ Predict B

e.g., "trendline" (regression)



5

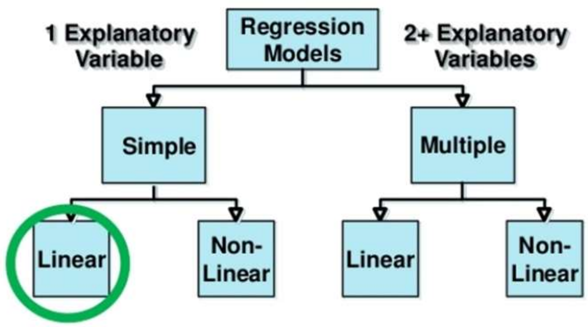
Overview

- 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 453X Machine Learning

6

Types of Regression Models



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

7

Outline

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

8

Simple Linear Regression

- Goal – find a linear relationship between to values
  - E.g., kills and skill, time and car speed
- First, make sure relationship is linear! How?

9

Simple Linear Regression

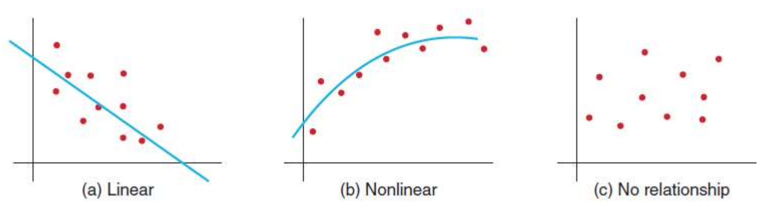
- Goal – find a linear relationship between to values
  - E.g., kills and skill, time and car speed
- First, make sure relationship is linear! How?

→ Scatterplot

(c) no clear relationship

(b) not a linear relationship

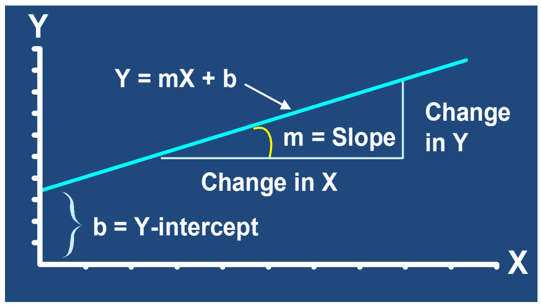
(a) linear relationship – proceed with linear regression



10

Linear Relationship

- From algebra: line in form $\boxed{Y = mX + b}$
  - 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

11

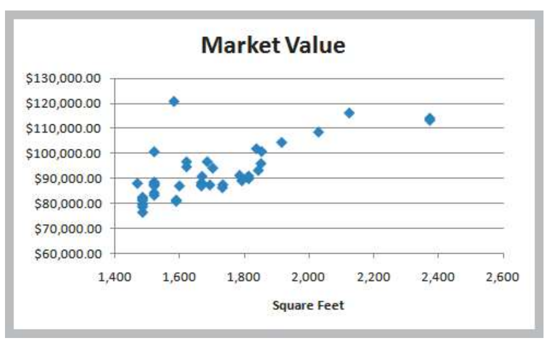
Simple Linear Regression Example

- 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

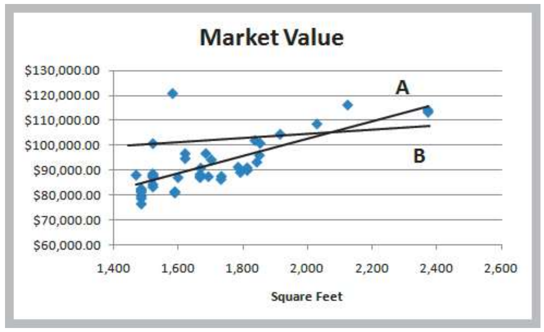


12

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?

$Y = mX + b$



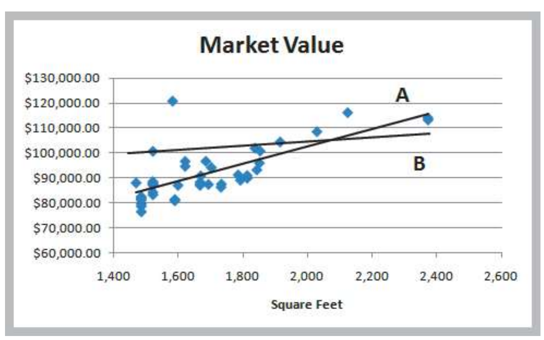
13

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?

$Y = mX + b$

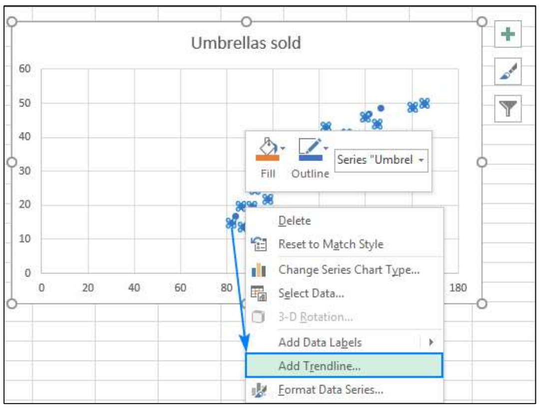
Line that gives best fit to data is one that minimizes prediction error
→ Least squares line
(more later)

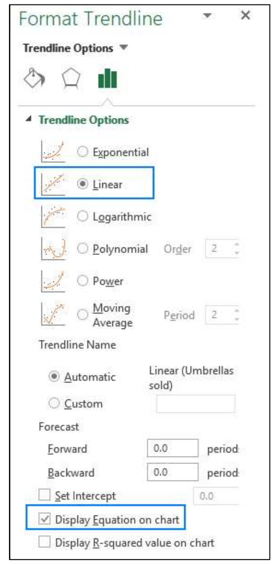


14

Simple Linear Regression Example Chart

- Scatterplot
- Right click → Add Trendline





15

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

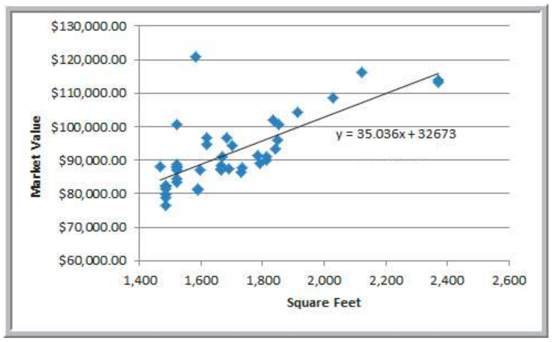
- Estimate *Y* when *X* = 1800 square feet

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

16

Simple Linear Regression Example

- Market value = 32673 + 35.036 *x* (square feet)
- Predicts market value better than just average



But before use, examine residuals

17

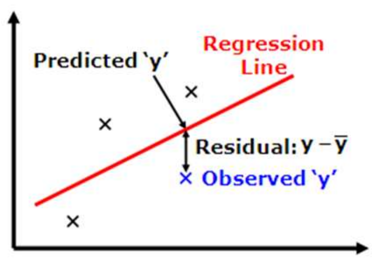
Outline

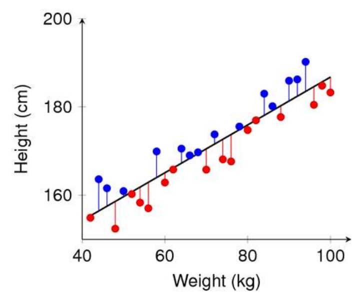
- Introduction (done)
- Simple Linear Regression
  - Linear relationship (done)
  - Residual analysis (next)
  - Fitting parameters
- Measures of Variation
- Misc

18

Residual Analysis

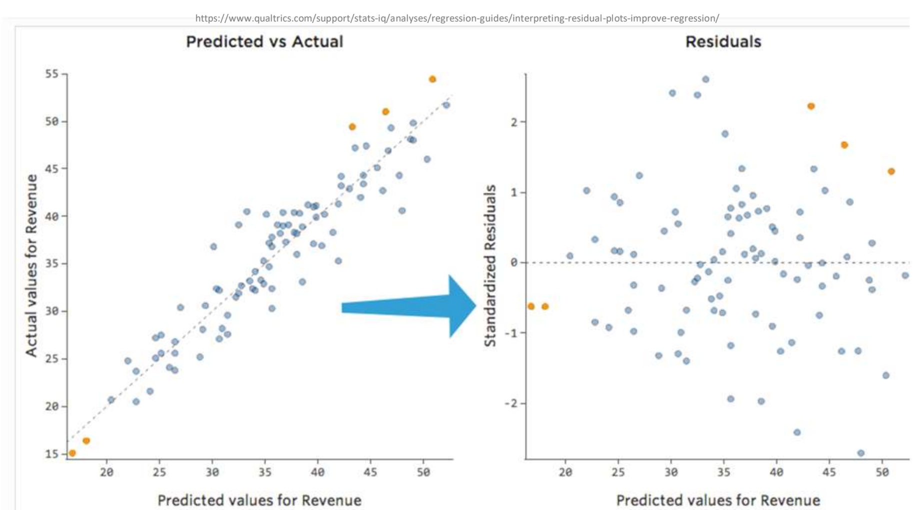
- Before predicting, confirm that linear regression assumptions hold
  - Variation around line is normally distributed
  - Variation equal for all *X*
  - Variation independent for all *X*
- How? Compute residuals (error in prediction) → Chart





19

Residual Analysis

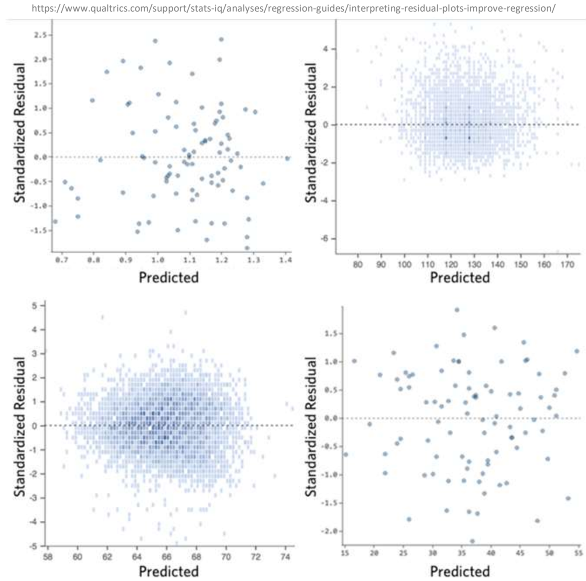


Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

20

Residual Analysis – Good

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



Clustered towards middle

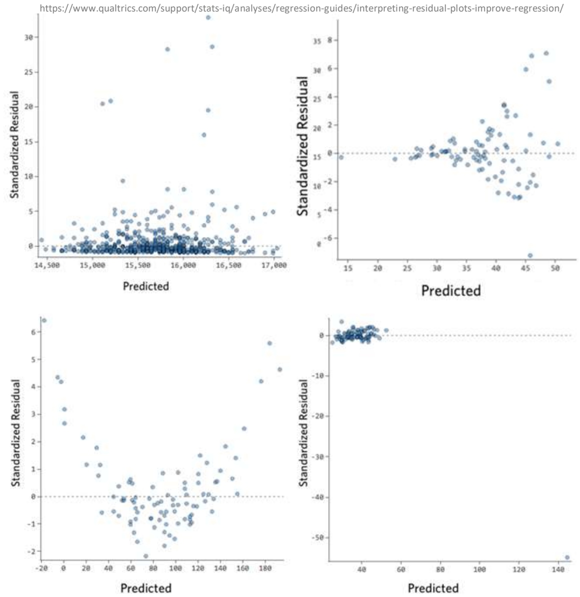
Symmetrically distributed

No clear pattern

21

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)

22

Residual Analysis – Summary

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

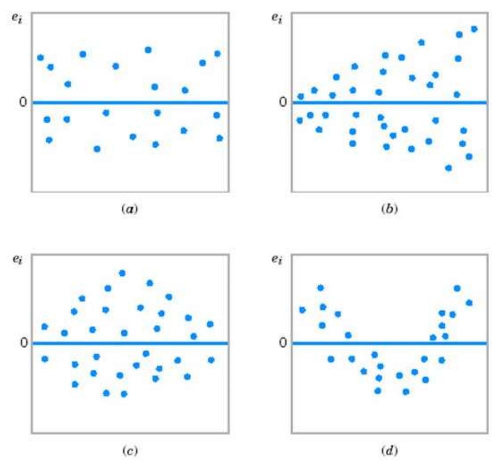
____

(a) ok

(b) funnel

(c) double bow

(d) nonlinear



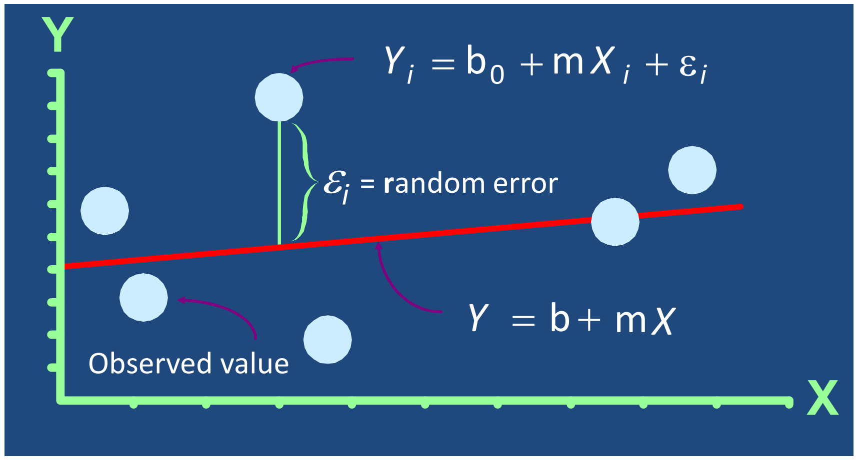
23

Outline

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

24

Linear Regression Model



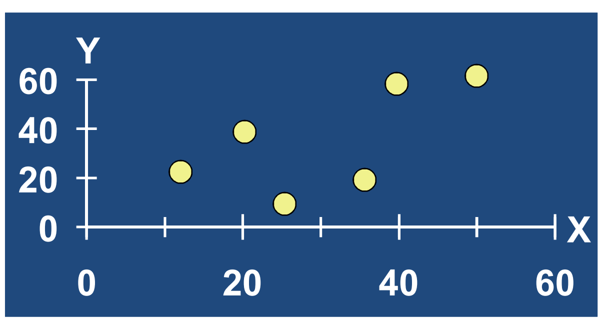
https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

Random error associated with each observation

25

Fitting the Best Line

- Plot all (X_i, Y_i) Pairs

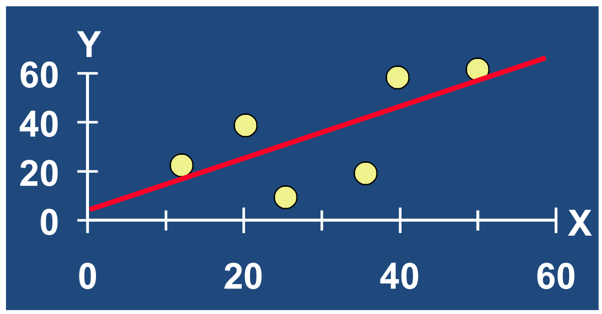


https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

26

Fitting the Best Line

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

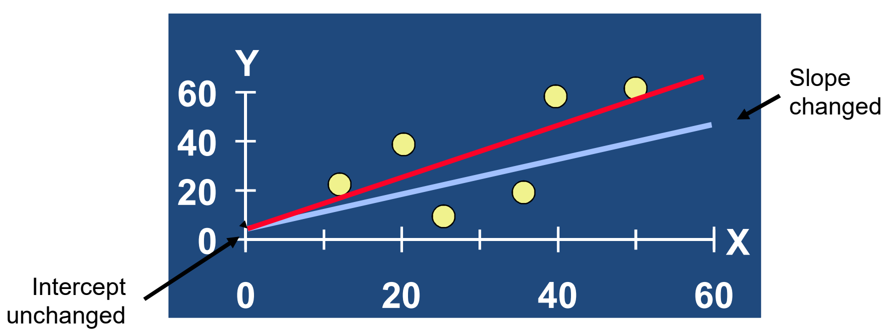


https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

27

Fitting the Best Line

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

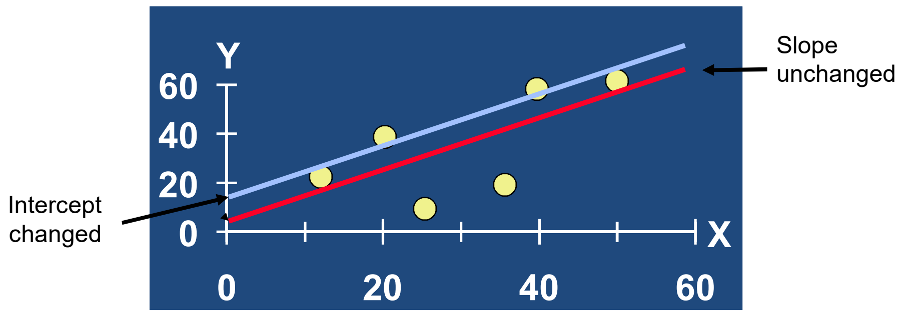


https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

28

Fitting the Best Line

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

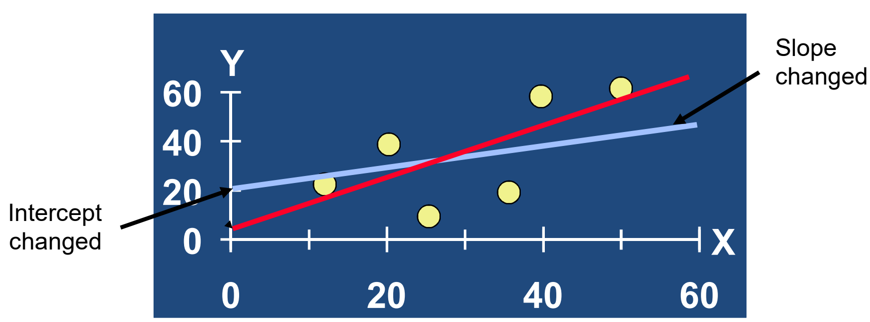


https://www.scribd.com/presentation/230686725/Fu-Ch11-Linear-Regression

29

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

30

Linear Regression Model

- Relationship between variables is linear function

Population Y-Intercept

Population Slope

Random Prediction Error

Want error as small as possible

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

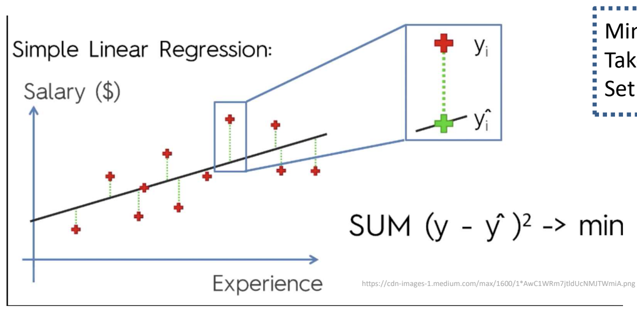
Dependent (response) Variable (e.g., kills)

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

31

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)

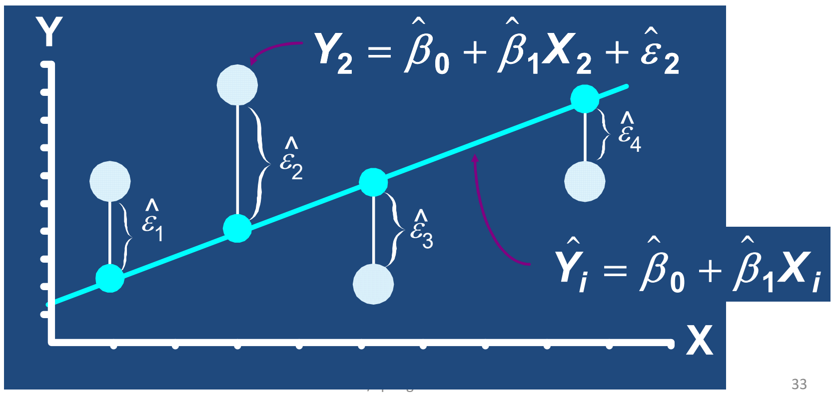


Minimize: $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$
Take derivative
Set to 0 and solve

32

Least Squares Line Graphically

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

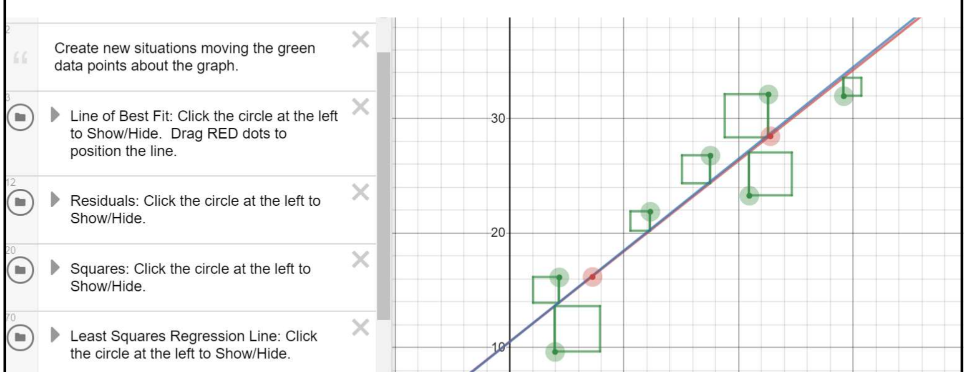


33

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

33

Least Squares Line Graphically



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

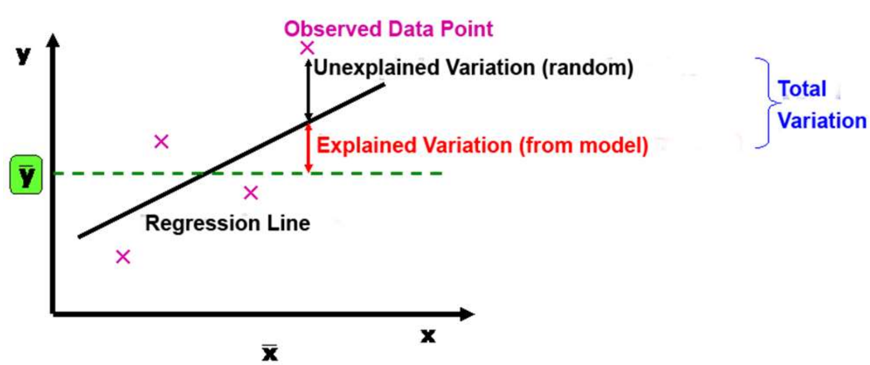
34

Outline

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

35

Measures of Variation

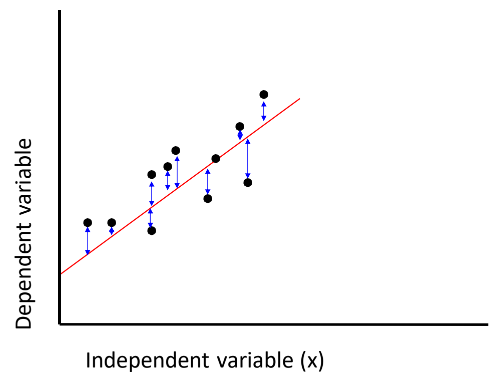


- Several sources of variation in y
  - Error in prediction (unexplained)
  - Variation from model (explained)

Break this down (next)

36

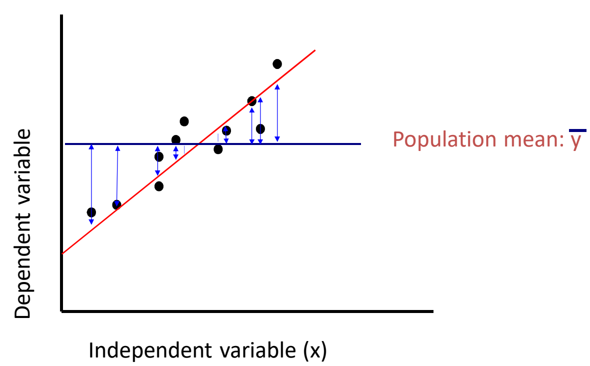
Sum of Squares of Error



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

37

Sum of Squares Regression

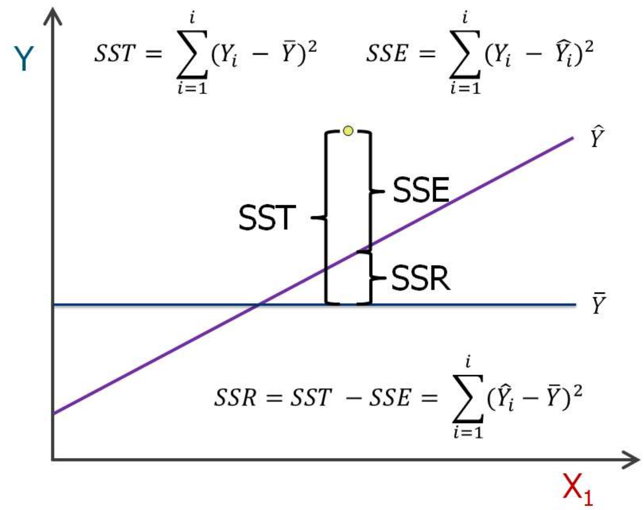


- Differences between prediction and population mean
  - Gets at variation due to X & Y
- Sum of Squares Regression, or SSR
- Measure of explained variation

38

Sum of Squares Total

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



39

Coefficient of Determination

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

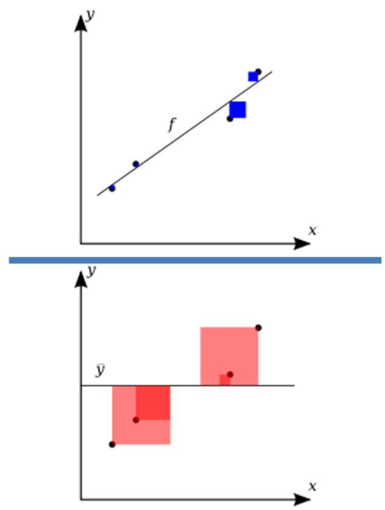
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Ranges from 0 to 1 (often said as a percent)
  - 1 – regression explains all of variation
  - 0 – regression explains none of variation

40

Coefficient of Determination – Visual Representation

$R^2 = 1 -$



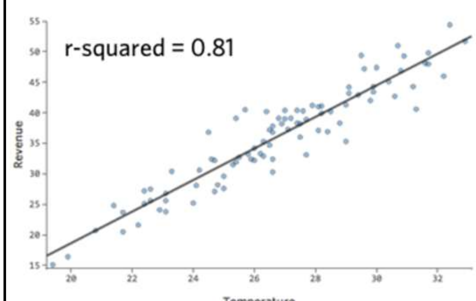
Variation in observed data model cannot explain (error)

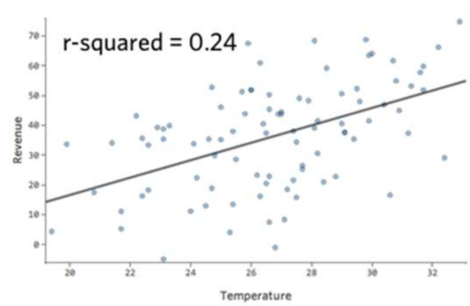
Total variation in observed data

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

41

Coefficient of Determination Example



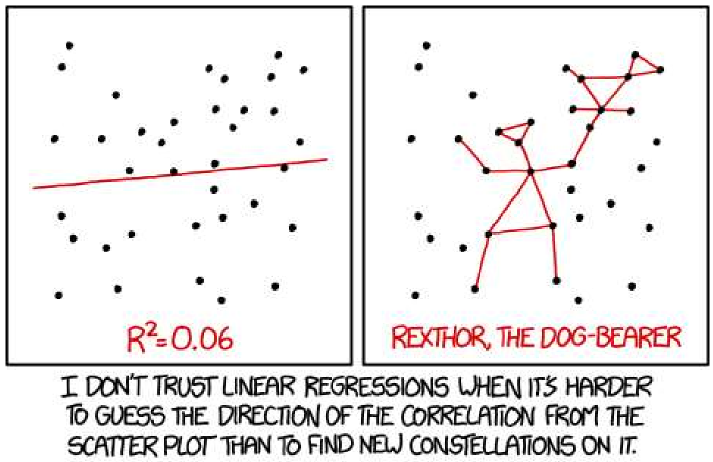


- How “good” is regression model? Roughly:

$0.8 <= R^2 <= 1$	strong
$0.5 <= R^2 < 0.8$	medium
$0 <= R^2 < 0.5$	weak

42

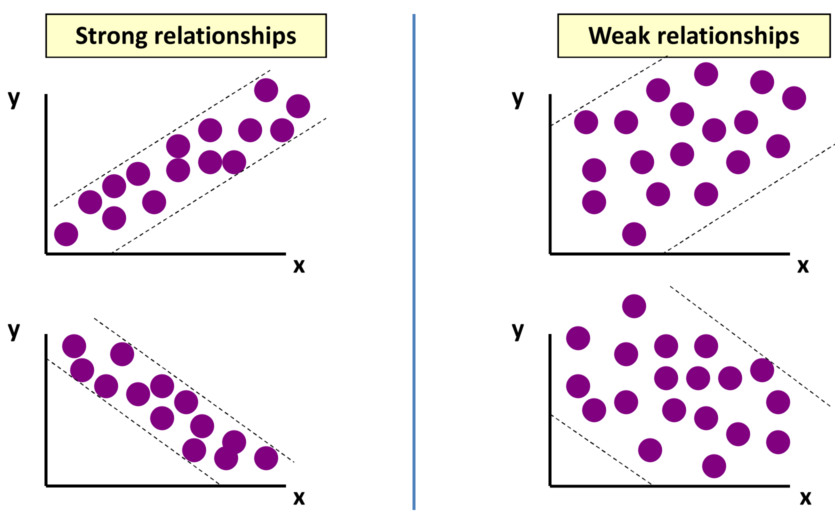
How “good” is the Regression Model?



https://xkcd.com/1725/

43

Relationships Between X & Y



44

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? R = $\sqrt{R^2}$

Pearson's Correlation Coefficient

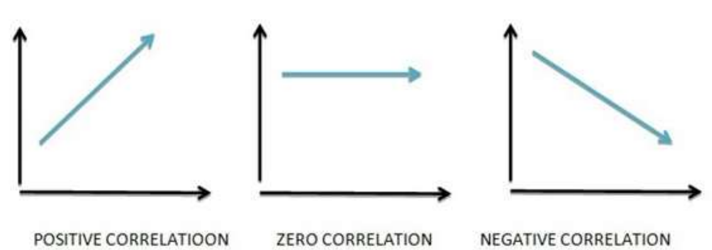
$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Vary together

Vary Separately

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

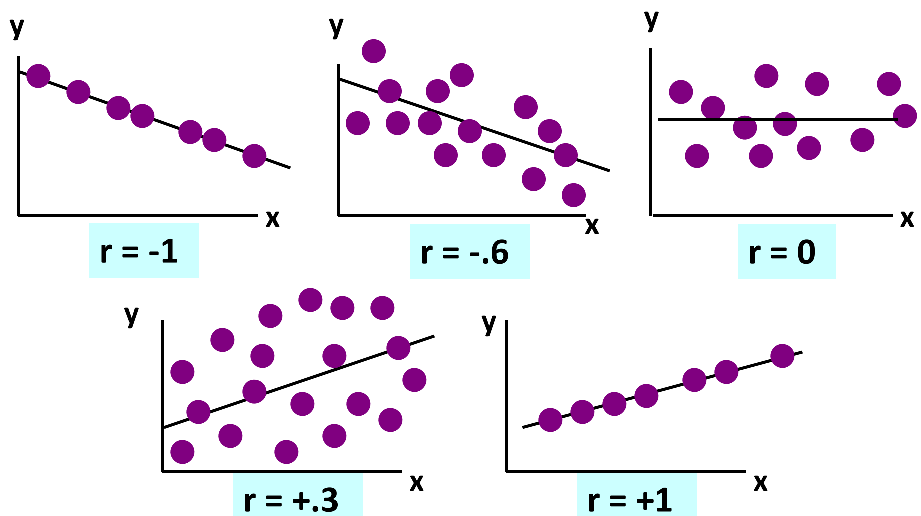
$\overline{Y}$ = mean of Y variable



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

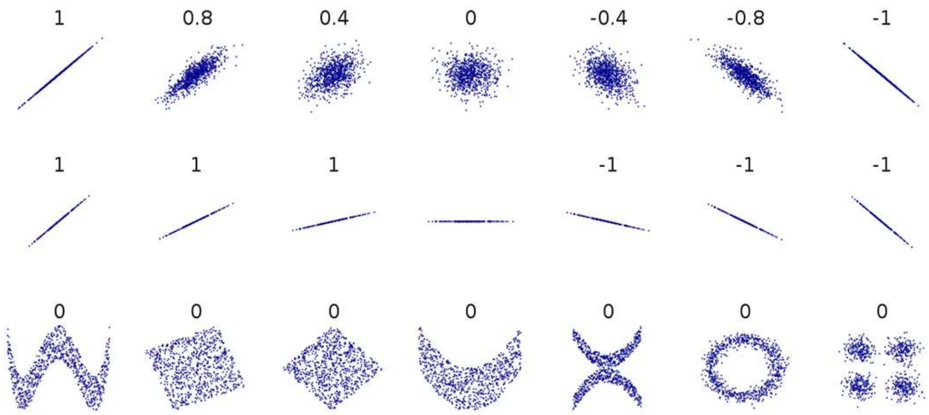
45

Correlation Examples (1 of 3)



46

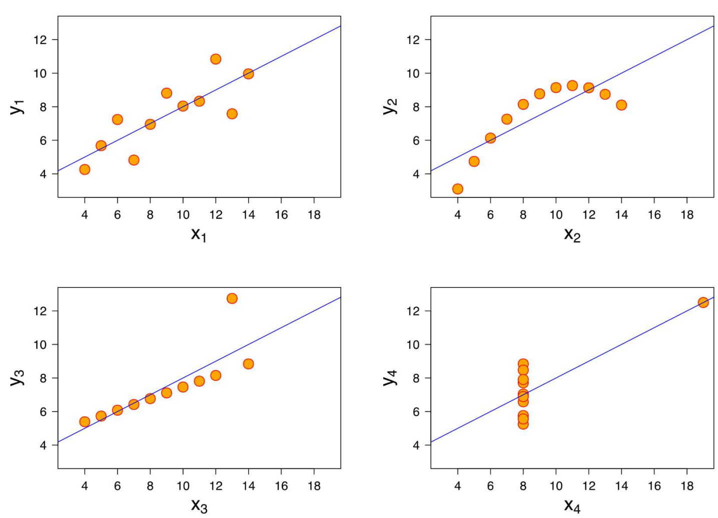
Correlation Examples (2 of 3)



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

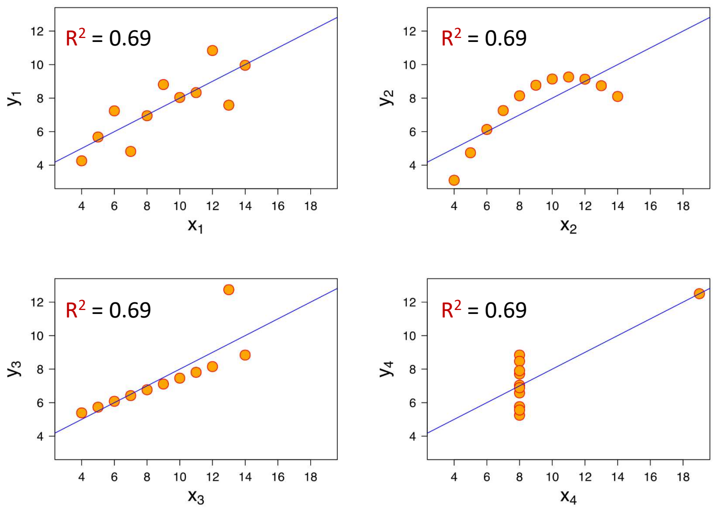
47

Correlation Examples (3 of 3)



48

Correlation Examples (3 of 3)



Anscombe's Quartet

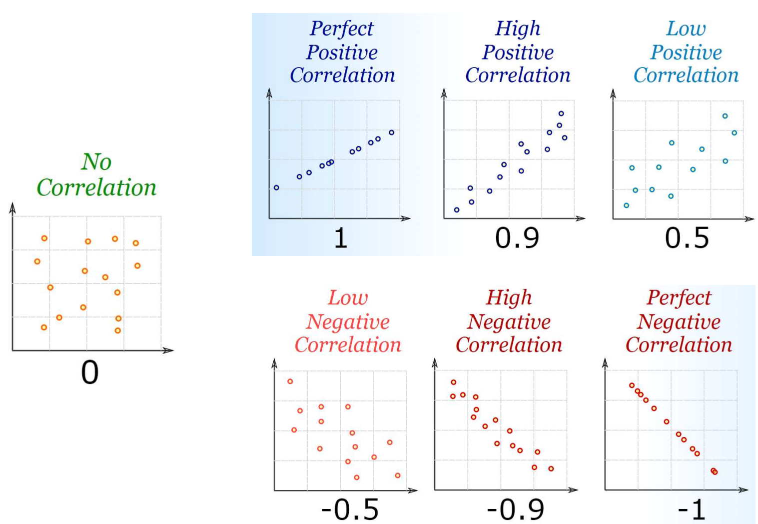
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Summary stats:

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

49

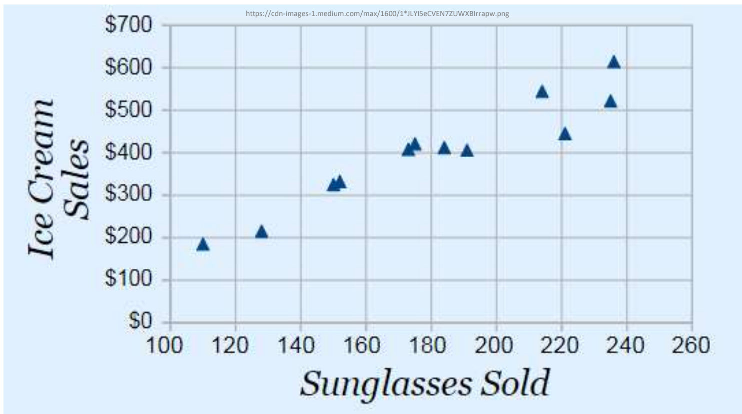
Correlation Summary



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

50

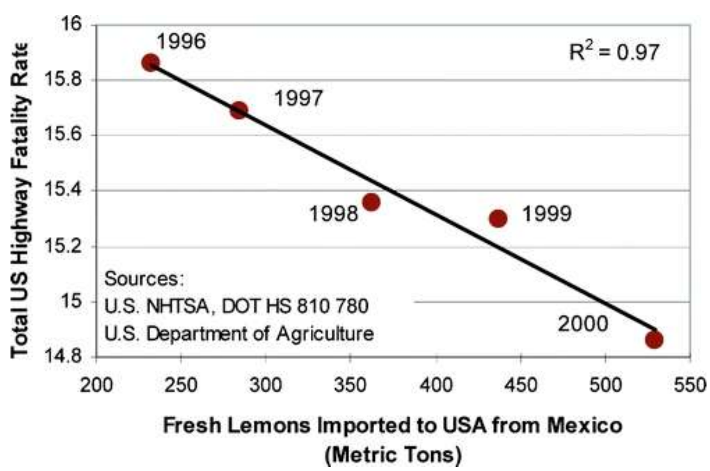
Correlation is not Causation



Buying sunglasses *causes* people to buy ice cream?

51

Correlation is not Causation

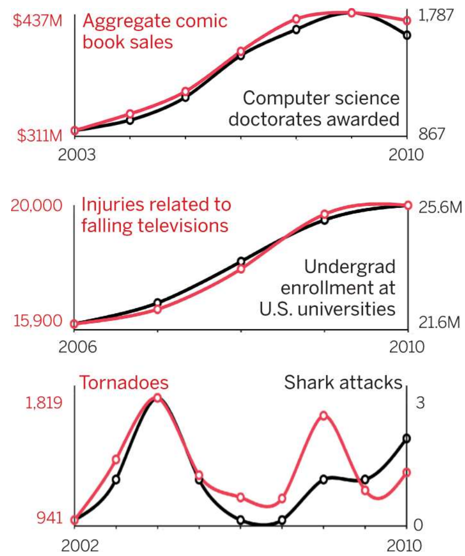


Importing lemons causes fewer highway fatalities?

52

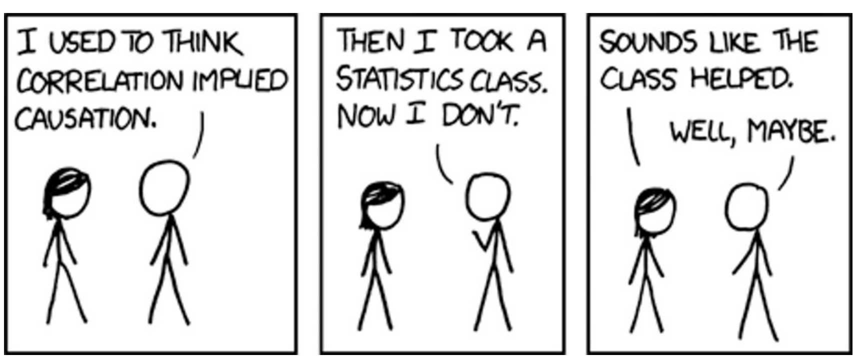
Correlation is not Causation

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



53

Correlation is not Causation



https://xkcd.com/552/

54

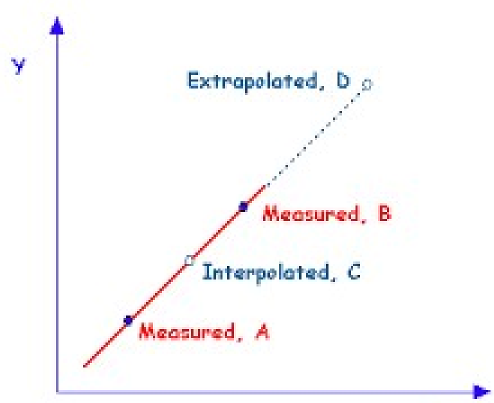
Outline

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

55

Extrapolation versus Interpolation

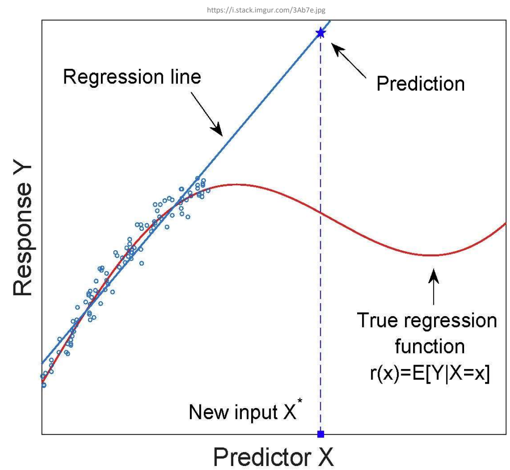
- Prediction
  - Interpolation – within measured X-range
  - Extrapolation – outside measured X-range



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

56

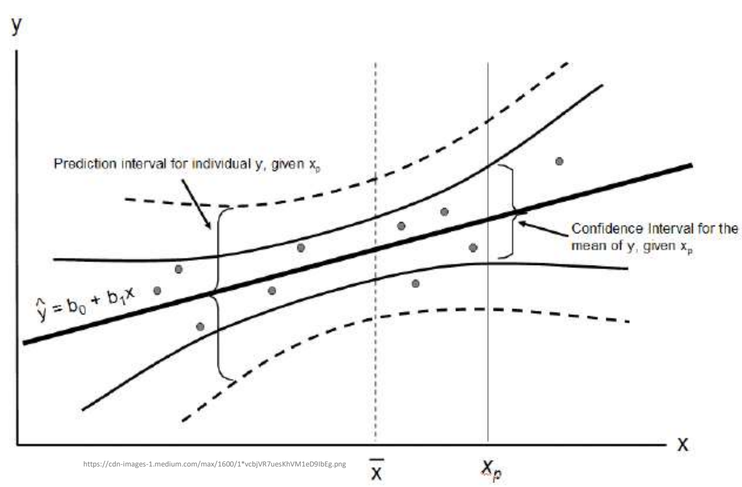
Be Careful When Extrapolating



If extrapolate, make sure have reason to assume model continues

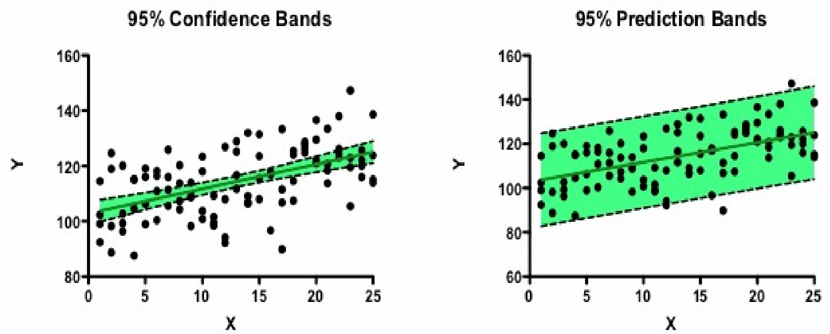
57

Prediction and Confidence Intervals (1 of 2)



58

Prediction and Confidence Intervals (2 of 2)

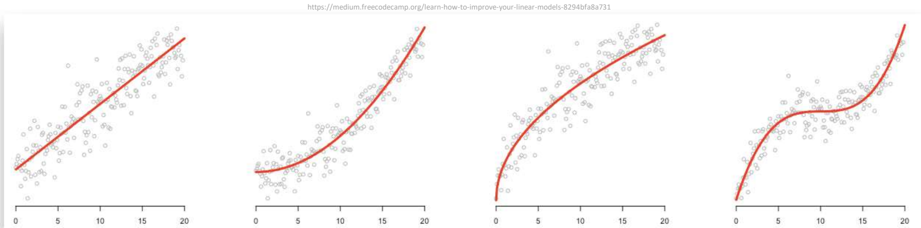


https://www.graphpad.com/guides/prism/7/curve-fitting/reg_mostpointsareoutsideconfidencebands.png

59

Beyond Simple Linear Regression

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

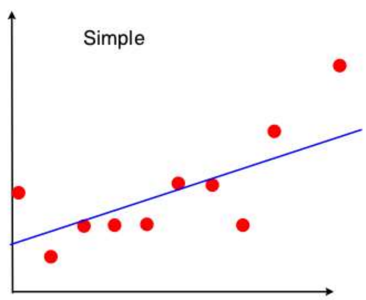


Linear Quadratic Root Cubic

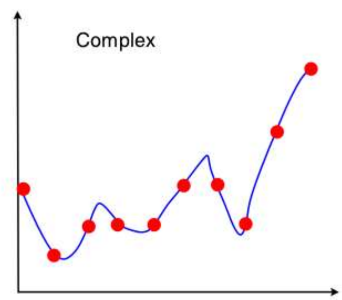
- Multiple regression – more parameters beyond just X
  - Book Chapter 11
- More complex models – beyond just $Y = mX + b$

60

More Complex Models



$y = 12x + 9$



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

- Higher order polynomial model has less error

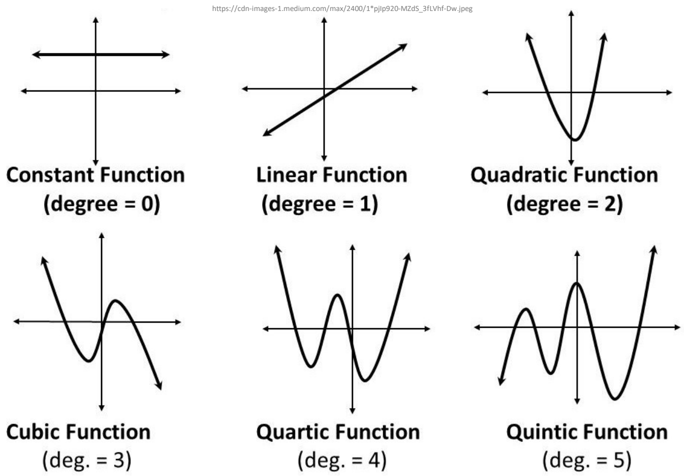
→ A "perfect" fit (no error)

- How does a polynomial do this?

61

Graphs of Polynomial Functions

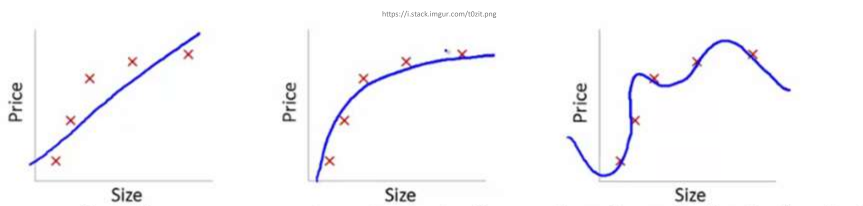
https://cdn-images-1.medium.com/max/2400/1*pjIp920-MZdS_3fLVhf-Dw.jpeg



Higher degree, more potential "wiggles"
But should you use?

62

Underfit and Overfit



$\theta_0 + \theta_1 x$ | $\theta_0 + \theta_1 x + \theta_2 x^2$ | $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

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 model do not predict well (i.e., for non-observed values)

- Just right – fit data well “enough” with as few parameters as possible

63