

## CSCI E-82

# Advanced Machine Learning, Data Mining & Artificial Intelligence Lecture 4

## Probability, Statistics, Regression Time Series Part I

Peter V. Henstock

Fall 2018

© 2018 Peter V. Henstock

## Administrivia

- HW2 – how is it going?
- Topic Presentations
  - Accepting groups of 2 or 3
  - Sign up using the same Google spreadsheet
  - Welcome to keep your partner or switch
    - Please talk to your existing partner regardless
  - Goal: 15 minute presentation
    - Will schedule presentations shortly but you will have a week to prepare from them
  - Need to select a moderately narrow topic
    - I will review topics on the sheet.
    - First sign-ups for a given topic win

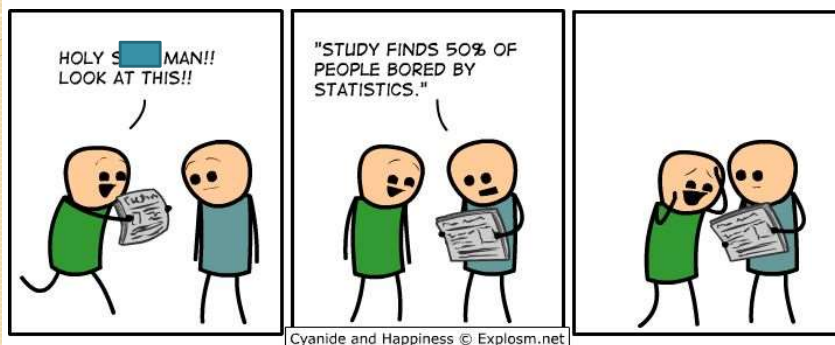
© 2018 Peter V. Henstock

# 1D Data Statistics

© 2018 Peter V. Henstock

## Let's Review Statistics

- Boringem.org



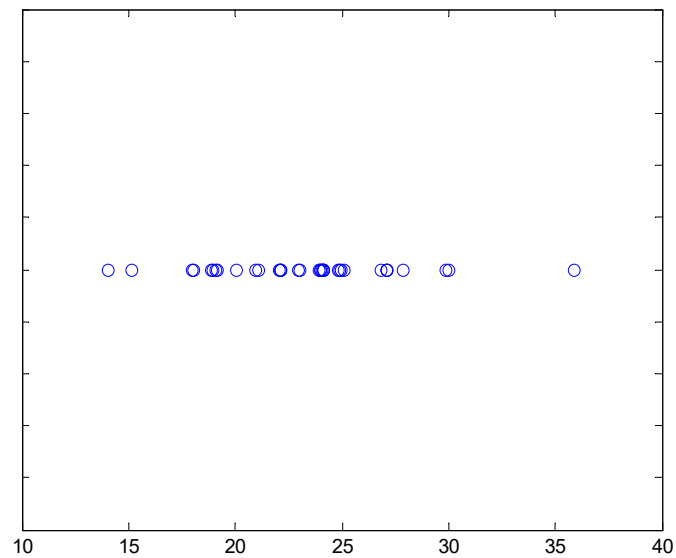
© 2018 Peter V. Henstock

## Zoom Sign-in Times (min after hour) N=36

14	23	27
15	23	27
18	23	27
18	24	27
19	24	28
19	24	30
19	24	30
20	24	36
21	24	
21	24	
22	25	
22	25	
22	25	
22	25	

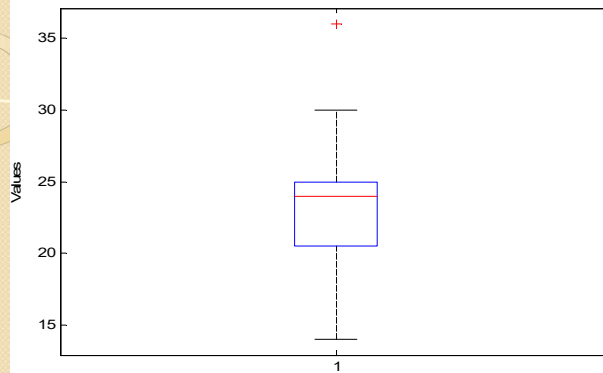
© 2018 Peter V. Henstock

## Plot (with jitter)



© 2018 Peter V. Henstock

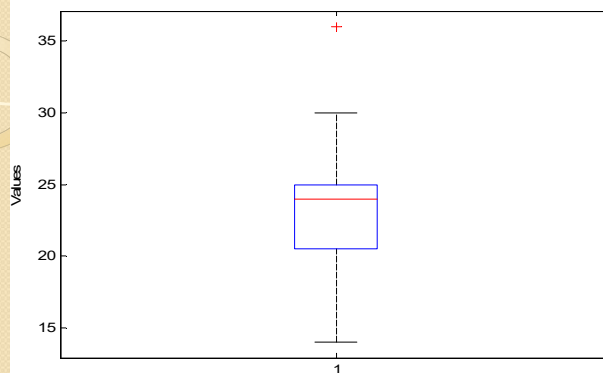
## Box-Whisker Plot



- What are the components?

© 2018 Peter V. Henstock

## Box-Whisker Plot



- Red line =  $Q2$  = median
- Box =  $Q1$  to  $Q3$
- Whiskers vary but might be:
  - $Q1 - 1.5 \text{ IQR}$  and  $Q3 + 1.5 \text{ IQR}$  where
  - $\text{IQR} = \text{Inter Quartile Range} = Q3 - Q1$

© 2018 Peter V. Henstock

## Sample Statistics & Robustness

- Mean  $\bar{x} = \frac{1}{N} \sum_{i=0}^N x_i$
- Median =
  - $x_i$  that is middle value of sorted X of odd N
  - $(x_i + x_{i+1})/2$  that is average of 2 central points
- Range =  $\max(X) - \min(X)$
- Sample variance  $s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$
- Robustness = immunity to outliers
- Which of these are robust?

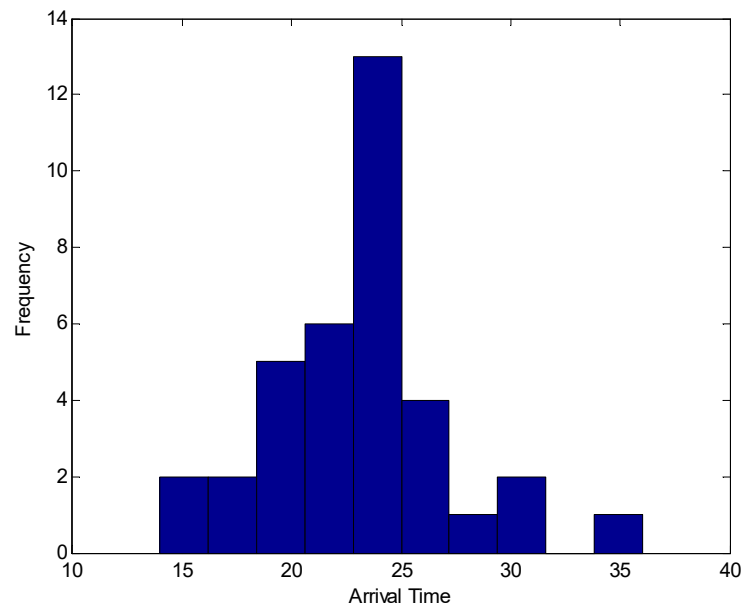
© 2018 Peter V. Henstock

## Sample Statistics?

- Mean  $\bar{x} = \frac{1}{N} \sum_{i=0}^N x_i$
- Median =
  - $x_i$  that is middle value of sorted X of odd N
  - $(x_i + x_{i+1})/2$  that is average of 2 central points
- Range =  $\max(X) - \min(X)$
- Sample variance  $s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}$
- Why are these “sample” statistics?

© 2018 Peter V. Henstock

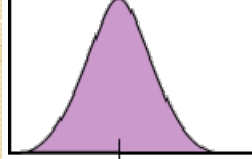
## Histogram



© 2018 Peter V. Henstock

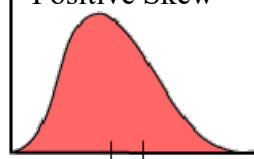
## Skewness

**Symmetric Distribution**  
Zero Skew



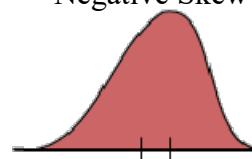
Mean = Median

**Right-Skewed Distribution**  
Positive Skew



Median Mean

**Left-Skewed Distribution**  
Negative Skew



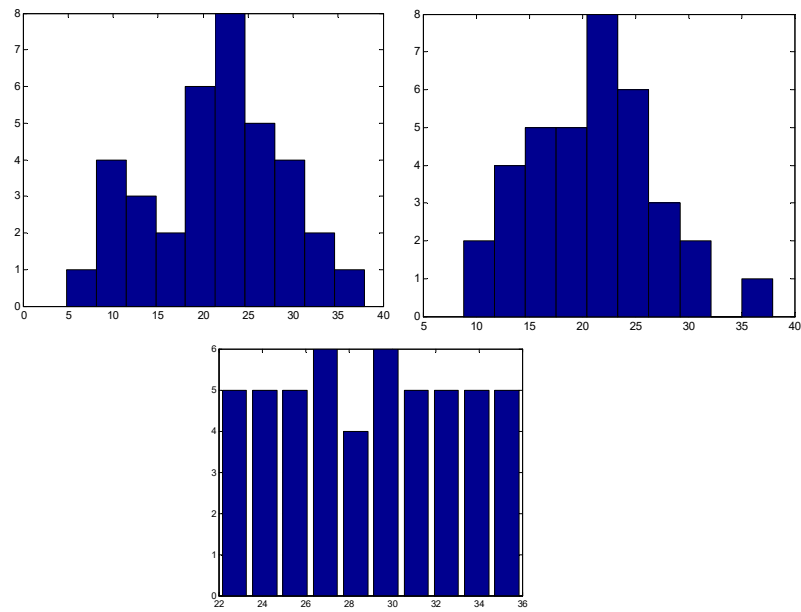
Mean Median

- $Skewness = \frac{3(mean - median)}{s}$
- $Skewness = \frac{\sum (X_i - mean)^3}{(N-1)s^3}$
- N samples of X
- s = sample standard deviation

• <https://onlinecourses.science.psu.edu/stat100/node/10>

© 2018 Peter V. Henstock

## Modality ~ # of peaks



© 2018 Peter V. Henstock

# Distributions

© 2018 Peter V. Henstock

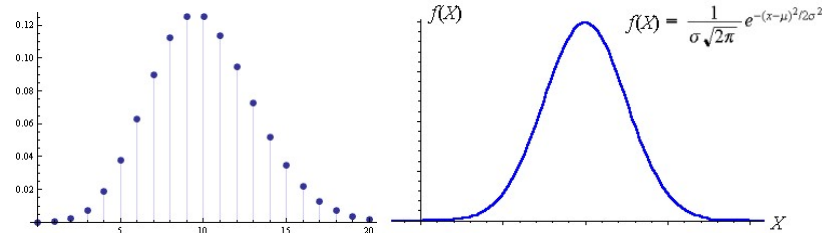
## Random Variable

- Variable that takes on a specific value
- Member of a group
- Group values are described according to a frequency distribution
- Types
  - Discrete
  - Continuous
- Frequently drawn from a distribution

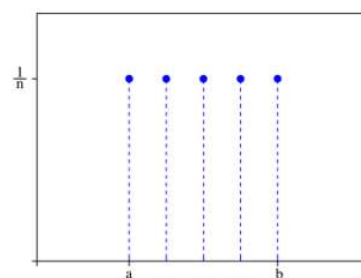
© 2018 Peter V. Henstock

## Probability Mass/Densities Function

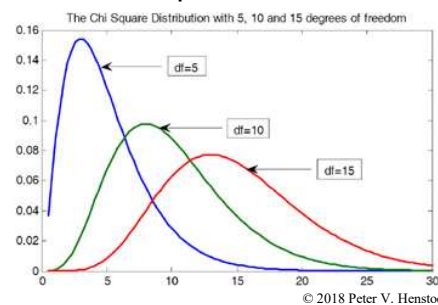
Gaussian or Normal PMF (left), PDF (right)



Uniform PMF



Chi-squared PDF



© 2018 Peter V. Henstock



## Common density functions

- Uniform(a,b)
  - $f(x) = \frac{1}{b-a}$  for  $x \in [a, b]$
- Bernoulli(p) for p in [0,1]
  - $f(0) = 1-p = \text{failure}$        $f(1) = p = \text{success}$
- Binomial(n, p) = sum n Bernoulli trials
  - $f(x) = \binom{n}{x} p^x p^{n-x}$        $x = 0, \dots, n$
- Normal (Gaussian) distribution  $N(\mu, \sigma)$ 
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

© 2018 Peter V. Henstock

## Other Discrete/Continuous densities

- <http://aleph0.clarku.edu/~djoyce/ma218/distributions.pdf>

Distribution	Type	Mass/density function $f(x)$	Mean $\mu$	Variance $\sigma^2$
UNIFORM( $n$ )	D	$1/n$ , for $x = 1, 2, \dots, n$	$(n+1)/2$	$(n^2 - 1)/12$
UNIFORM( $a, b$ )	C	$\frac{1}{b-a}$ , for $x \in [a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
BERNOULLI( $p$ )	D	$f(0) = 1-p, f(1) = p$	$p$	$p(1-p)$
BINOMIAL( $n, p$ )	D	$\binom{n}{x} p^x (1-p)^{n-x}$ , for $x = 0, 1, \dots, n$	$np$	$npq$
GEOMETRIC( $p$ )	D	$q^{x-1}p$ , for $x = 1, 2, \dots$	$1/p$	$(1-p)/p^2$
NEGATIVEBINOMIAL( $p, r$ )	D	$\binom{x-1}{r-1} p^r q^{x-r}$ , for $x = r, r+1, \dots$	$r/p$	$r(1-p)/p^2$
HYPERGEOMETRIC( $N, M, n$ )	D	$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$ , for $x = 0, 1, \dots, n$	$np$	$np(1-p)$
POISSON( $\lambda t$ )	D	$\frac{1}{x!} (\lambda t)^x e^{-\lambda t}$ , for $x = 0, 1, \dots$	$\lambda t$	$\lambda t$
EXPONENTIAL( $\lambda$ )	C	$\lambda e^{-\lambda x}$ , for $x \in [0, \infty)$	$1/\lambda$	$1/\lambda^2$
GAMMA( $\lambda, r$ )	C	$\frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x}$	$r/\lambda$	$r/\lambda^2$
GAMMA( $\alpha, \beta$ )		$= \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$ , for $x \in [0, \infty)$	$= \alpha\beta$	$= \alpha\beta^2$

© 2018 Peter V. Henstock

## Other Discrete/Continuous densities

- <http://aleph0.clarku.edu/~djoyce/ma218/distributions.pdf>

BETA( $\alpha, \beta$ )	C	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1},$ for $0 \leq x \leq 1$	$\frac{\alpha}{\alpha + \beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
NORMAL( $\mu, \sigma^2$ )	C	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$ for $x \in \mathbf{R}$	$\mu$	$\sigma^2$
CHISQUARED( $\nu$ )	C	$\frac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2} \Gamma(\nu/2)},$ for $x \geq 0$	$\nu$	$2\nu$
T( $\nu$ )	C	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \Gamma(\frac{\nu}{2})} (1+x^2/\nu)^{-(\nu+1)/2}$ for $x \in \mathbf{R}$	0	$\nu/(\nu-2)$
F( $\nu_1, \nu_2$ )	C	$\frac{1}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})} \frac{(\frac{\nu_1}{2})^{\nu_1/2} x^{\nu_1/2-1}}{(1+\frac{\nu_1}{\nu_2}x)^{(\nu_1+\nu_2)/2}}$ for $x > 0$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

© 2018 Peter V. Henstock

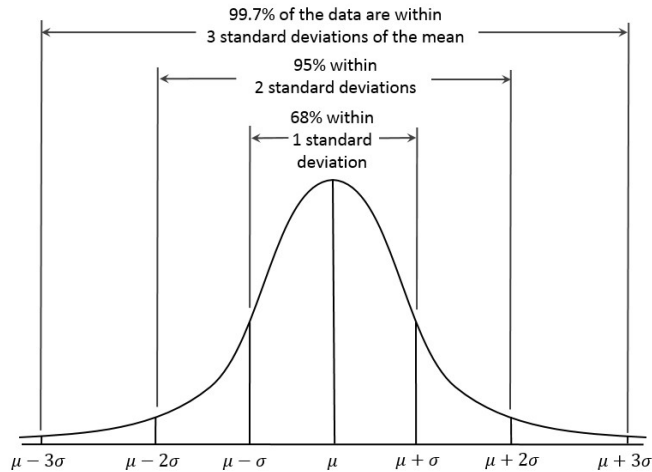
## Properties of PMF/PDF

- $\text{prob}(x) = p(x) = f(x)$  for discrete
- $p(x \in [a,b]) = \sum_{i=a}^b f(x_i)$  for discrete
- $p(x \in [a,b]) = \int_a^b f(x_i) dx$  for continuous
- $\sum_{-\infty}^{\infty} f(x_i) = 1 \quad \int_{-\infty}^{\infty} f(x_i) dx = 1$

© 2018 Peter V. Henstock

## How to use this?

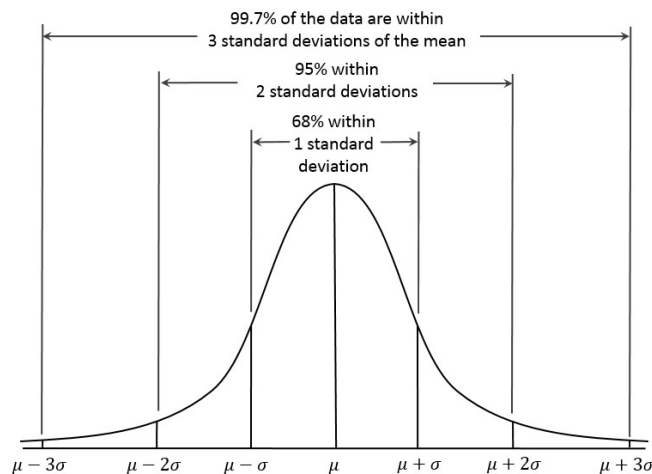
- Einstein had 160 IQ which is  $N(100, 15)$
- $P(X \geq 160) = \int_{160}^{\infty} f(x) dx$



[https://en.wikipedia.org/wiki/68%E2%80%9393%E2%80%9399.7\\_rule#/media/File:Empirical\\_Rule.PNG](https://en.wikipedia.org/wiki/68%E2%80%9393%E2%80%9399.7_rule#/media/File:Empirical_Rule.PNG) Peter V. Henstock

## How to use this?

- What is probability of being within 1 stdev of “average”?



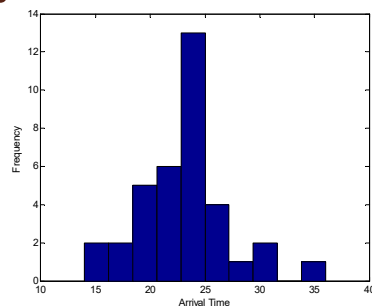
[https://en.wikipedia.org/wiki/68%E2%80%9393%E2%80%9399.7\\_rule#/media/File:Empirical\\_Rule.PNG](https://en.wikipedia.org/wiki/68%E2%80%9393%E2%80%9399.7_rule#/media/File:Empirical_Rule.PNG) Peter V. Henstock

## Statistics from PMF

- $Mean = \mu = \sum_{i=0}^N p_i x_i$
- $Var = \sigma^2 = \sum_{i=0}^N p_i (x_i - \mu)^2$

© 2018 Peter V. Henstock

## Why bother with PDF/PMFs?



- Provide an abstraction of actual sample histogram (normalized)
  - Avoids issues of outliers, etc.
- Data compression:  $N$  pts  $\rightarrow$  parameters
- Enables statistical analyses

© 2018 Peter V. Henstock

## Randomness

Characteristic	Pseudo-Random Number Generators	True Random Number Generators
Efficiency	Excellent	Poor
Determinism	Deterministic	Nondeterministic
Periodicity	Periodic	Aperiodic

- <https://www.random.org/randomness/>

© 2018 Peter V. Henstock

## Normal Distribution

- If  $X$  and  $Y$  are random variables from a normal distribution,  $X+Y$  is also normal
- If you take many independent random variables from any single distribution, the sum approximates a normal distribution
  - Central Limit Theorem (with caveats)

© 2018 Peter V. Henstock

# Probabilities

© 2018 Peter V. Henstock

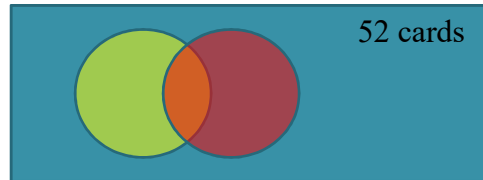
## Random Variable

- Variable that takes on a specific value
- Member of a group
- Group values are described according to a frequency distribution
- Types
  - Discrete
  - Continuous
- Frequently randomly drawn from a distribution

© 2018 Peter V. Henstock

## Probability Functions

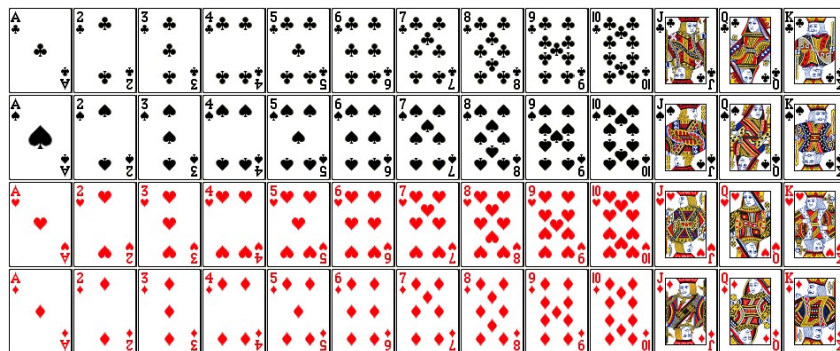
- $P(x)$  = probability of  $x$  occurring
  - Depends on the space of possibilities
- $P(x) = 1 - P(\sim X)$



- $P(\text{spades}) = ?$
- $P(\sim \text{spades}) = ?$
- $P(\text{Red "Royalty"})$

© 2018 Peter V. Henstock

## Deck of Cards

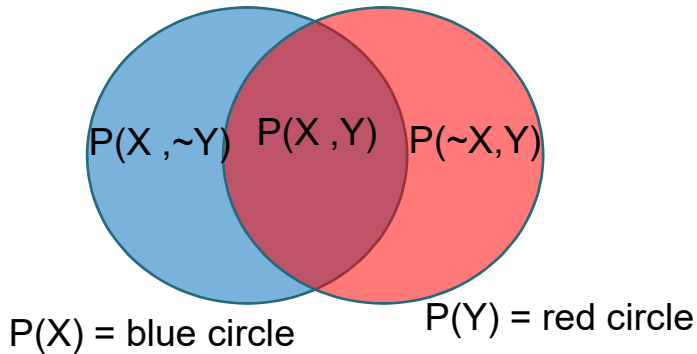


- <http://www.milefoot.com/math/discrete/counting/cardfreq.htm>

© 2018 Peter V. Henstock

## Joint Probabilities

- $P(X,Y) = P(X \text{ and } Y)$
- $P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$



© 2018 Peter V. Henstock

## Joint Probabilities

- If  $X$  and  $Y$  are independent
  - $P(X,Y) = P(X)P(Y)$
- Example 1
  - $X$  = probability of flipping coin
  - $Y$  = probability of rolling a die (dice)
- Example 2
  - $X$  = probability of grass is wet
  - $Y$  = probability of rain

© 2018 Peter V. Henstock



## What does this imply statistically?

A major airlines company received an anonymous bomb threat. To figure out ways of reducing the risk, a team was assembled. One of the members was a statistician. After careful thought and calculations, he handed a sealed bag to the airlines company and ordered them to put this bag in their plane during every journey. After a few journeys, the flight team were intrigued, and decided to open the bag. Inside, they found a bomb. They quickly confronted the statistician and asked for an explanation. He replied, "Well. Statistics show that there is a 1 in a 1,000,000 chance of a bomb being on a plane. But for 2 bombs to be on the same plane, the chances are only 1 in a 1,000,000,000,000."

© 2018 Peter V. Henstock

## What does this imply statistically?

A major airlines company received an anonymous bomb threat. To figure out ways of reducing the risk, a team was assembled. One of the members was a statistician. After careful thought and calculations, he handed a sealed bag to the airlines company and ordered them to put this bag in their plane during every journey. After a few journeys, the flight team were intrigued, and decided to open the bag. Inside, they found a bomb. They quickly confronted the statistician and asked for an explanation. He replied, "Well. Statistics show that there is a 1 in a 1,000,000 chance of a bomb being on a plane. But for 2 bombs to be on the same plane, the chances are only 1 in a 1,000,000,000,000."

**This is a joke. Neither the course staff nor Harvard endorses carrying bombs on planes**

© 2018 Peter V. Henstock

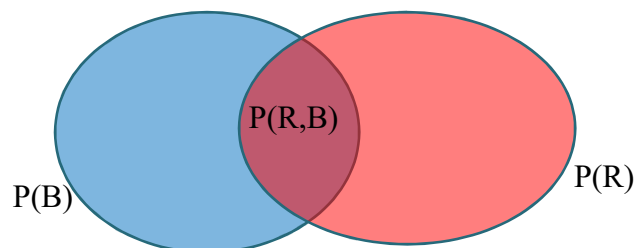
## Conditional Probabilities

- $P(X|Y)$  = Probability of X “given” Y
- $P(X|Y) = P(X,Y) / P(Y)$

© 2018 Peter V. Henstock

## Conditional Probabilities

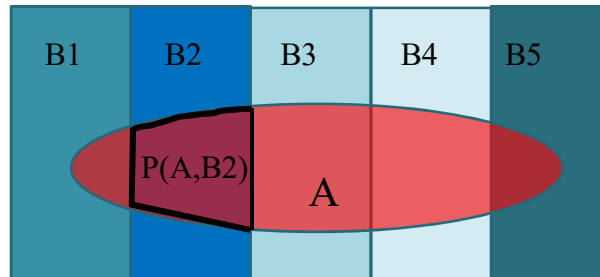
- $P(X|Y)$  = Probability of X “given” Y
- $P(X|Y) = P(X,Y) / P(Y)$
- Probability in blue given it's in red?



© 2018 Peter V. Henstock

## Visual Idea

- $P(A) = \sum_{i=0}^n P(A | B_i)P(B_i)$



- Region overlap:  $P(B2, A)$
- $P(B2, A) = P(A|B2)P(B2)$
- If we add up all intersections over B's:
  - $P(A) = \sum_{i=0}^n P(A | B_i)P(B_i)$

© 2018 Peter V. Henstock

## Basic Rules for Probabilities

- Sum Rule:
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Product Rule:
  - $P(A \cap B) = P(A, B) = P(A|B)P(B) = P(B|A)P(A)$
- Joint  $\rightarrow$  Marginal
  - $P(A) = \sum_{i=0}^n P(A, B_i)$
  - $P(A) = \sum_{i=0}^n P(A | B_i)P(B_i)$

© 2018 Peter V. Henstock

## Chain Rule



What is the probability of this whole thing?

© 2018 Peter V. Henstock

## Chain Rule



What is the probability of this whole thing?

$P(x_1, x_2, \dots, x_n)$

How can we compute this using  
conditional probabilities?

$P(x_1)^*$

© 2018 Peter V. Henstock

## Chain Rule



What is the probability of this whole thing?

$$P(x_1, x_2, \dots, x_n)$$

How can we compute this using conditional probabilities?

$$P(X_1) * P(X_2 | X_1)$$

© 2018 Peter V. Henstock

## Chain Rule



What is the probability of this whole thing?

$$P(x_1, x_2, \dots, x_n)$$

How can we compute this using conditional probabilities?

$$P(X_1) * P(X_2 | X_1) * P(X_3 | X_1, X_2) \dots$$

...

© 2018 Peter V. Henstock

## Chain Rule



What is the probability of this whole thing?

$$P(x_1, x_2, \dots, x_n)$$

How can we compute this using conditional probabilities?

$$P(X_1) * P(X_2 | X_1) * P(X_3 | X_1, X_2) \dots \\ \dots P(X_n | X_1 \dots X_{n-1})$$

© 2018 Peter V. Henstock

## Bayes Rule

© 2018 Peter V. Henstock

## Bayes Rule

Thomas Bayes 1702-1761  
Presbyterian minister in UK



Wikipedia

- $P(h|e) = \frac{P(e|h)P(h)}{P(e)}$
- Basis of classifiers
- Basis of network inferences
- Basis of structuring information or expert systems

© 2018 Peter V. Henstock

## Bayes Rule

Thomas Bayes 1702-1761  
Presbyterian minister in UK



Wikipedia

- $P(h|e) = \frac{P(e|h)P(h)}{P(e)}$
- $P(h|e) = \frac{P(h,e)}{P(e)}$      $P(h,e) = P(e|h)P(h)$

© 2018 Peter V. Henstock

## Components of Bayes Rule

- $P(hyp|data) = \frac{P(data|hyp)P(hyp)}{P(data)}$
- $p(data|hyp) = p(data, hyp) / p(hyp)$
- What is  $p(hyp|data)$  if  $hyp$  and  $data$  are independent?

© 2018 Peter V. Henstock

## Conditional Independence

- $P(X,Y|Z) = P(X|Z)P(Y|Z)$ 
  - “X is independent of Y given Z”
  - X and Y are conditionally independent
- Clearly an extension of the general independence
  - $P(A,B) = P(A)P(B)$
- If X and Y are independent, can we say they are conditionally independent?

© 2018 Peter V. Henstock



## Conditional Independence Examples

- <https://www.quora.com/What-are-examples-of-events-that-are-independent-but-not-conditionally-independent-and-vice-versa>
- Good examples to navigate:
  - [in]dependence
  - conditional [in]dependence
  - $P(x|y)$  if independent  $p(x) \rightarrow$
  - $P(x|y) = P(x,y) / P(y) = P(x)P(y)/P(y) \rightarrow P(x)$

© 2018 Peter V. Henstock

## Components of Bayes Rule

- $$P(hyp|data) = \frac{P(data|hyp)P(hyp)}{P(data)}$$
- $P(hyp)$  = “Prior”
  - Prior probability of a hypothesis
- $P(hyp|data)$  = “posterior”
  - Probability of hypothesis given data
- $P(data|hyp)$  = “likelihood”
  - Probability of data fitting a given hypothesis
- $P(data)$  = normalizing constant
  - Data is same across all hypotheses

© 2018 Peter V. Henstock

# 2 Random Variables

© 2018 Peter V. Henstock

## Two random variables

- Mean values are the same
- Standard deviations are the same
- Relationship:
  - Correlation Coefficient
  - Covariance

© 2018 Peter V. Henstock

## Covariance

- $\text{Cov}(X, Y) = \sigma(X, Y) = E[X - E(X)] E[Y - E(Y)]$
- $\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$
- Sample Covariance(X, Y)
  - $= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$

© 2018 Peter V. Henstock

## Covariance

- $\text{Cov}(X, Y) = \sigma(X, Y) = E[X - E(X)] E[Y - E(Y)]$
- $\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$
- Sample Covariance(X, Y)
  - $= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$
- What is the  $\text{Cov}(X, X)$ ?

© 2018 Peter V. Henstock

## Covariance

- $\text{Cov}(X, Y) = \sigma(X, Y) = E[X - E(X)] E[Y - E(Y)]$
- $\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$
- Sample Covariance( $X, Y$ )
  - $= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$
- What is the  $\text{Cov}(X, X)$ ?
- If  $X$  and  $Y$  are independent, what is  $\text{Cov}(X, Y)$ ?

© 2018 Peter V. Henstock

## Covariance Matrix

- Let  $X = [X_1 \dots X_n]^T$  be a column vector
- Covariance matrix is written with  $\Sigma$
- $\Sigma_{ij} = \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

- $\Sigma = E(XX^T) - \mu\mu^T$

© 2018 Peter V. Henstock

## Covariance Matrix

- Let  $X = [X_1 \dots X_n]^T$  be a column vector
- Covariance matrix is written with  $\Sigma$
- $\Sigma_{ij} = \text{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$

$$\Sigma = \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & E[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

- If  $X_1 \dots X_n$  are all independent, what does the covariance matrix look like?

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

- Compute mean of X
- Compute mean of Y
- Sample cov of X, X
- Sample cov of Y, Y

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

- Compute mean of X
  - $\text{Sum}(X)/6 = 3.5$
- Compute mean of Y
- Sample cov of X, X
- Sample cov of Y, Y
- Sample cov of X, Y

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

- Compute mean of X
  - $\text{Sum}(X)/6 = 3.5$
- Compute mean of Y
  - $\text{Sum}(Y)/6 = 3$
- Sample cov of X, X
- Sample cov of Y, Y
- Sample cov of X, Y

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

- Compute mean of X
  - $\text{Sum}(X)/6 = 3.5$
- Compute mean of Y
  - $\text{Sum}(Y)/6 = 3$
- Sample cov of X, X
  - $[\sum (x - \text{mean}X)^2] / (N-1)$
- Sample cov of Y, Y
- Sample cov of X, Y

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

- Compute mean of X
  - $\text{Sum}(X)/6 = 3.5$
- Compute mean of Y
  - $\text{Sum}(Y)/6 = 3$
- Sample cov of X, X
  - $[\sum (x_i - \text{mean}X)^2] / (N-1) = 3.5$
  - $[\sum (x_i - 3.5)^2] / (6-1) = 3.5$
- Sample cov of Y, Y
  - $[\sum (y_i - \text{mean}Y)^2] / (N-1) = 0.4$
- Sample cov of X, Y

© 2018 Peter V. Henstock

## Sample Covariance Matrix example

X	Y
1	3
4	2
2	4
5	3
3	3
6	3

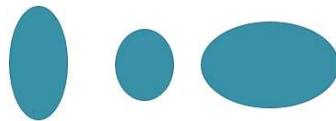
- Compute mean of X
  - $\text{Sum}(X)/6 = 3.5$
- Compute mean of Y
  - $\text{Sum}(Y)/6 = 3$
- Sample cov of X, X
  - $[\sum (x - \text{mean}X)^2] / (N-1) = 3.5$
  - $[\sum (x - 3.5)^2] / (6-1) = 3.5$
- Sample cov of Y, Y
  - $[\sum (y - \text{mean}Y)^2] / (N-1) = 0.4$
- Sample cov of X, Y
  - $[\sum (x - \text{mean}X)(y - \text{mean}Y)] / (N-1) = -0.4$

© 2018 Peter V. Henstock

## 2D Normal Distributions

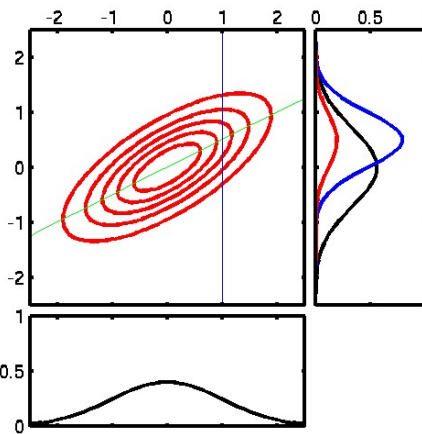
### Normal Distribution

- What can we say about the covariance of these?



### Normal Distribution

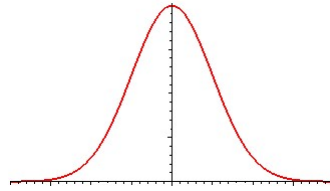
- What can we say about the covariance of these?


<http://www.atmos.washington.edu/~hakim/591/code/>

© 2018 Peter V. Henstock



## Gaussian or Normal Distribution

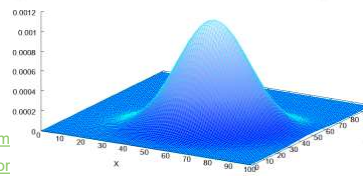


Univariate Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$

Multivariate Gaussian

$$f_{\mathbf{x}}(x_1, \dots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

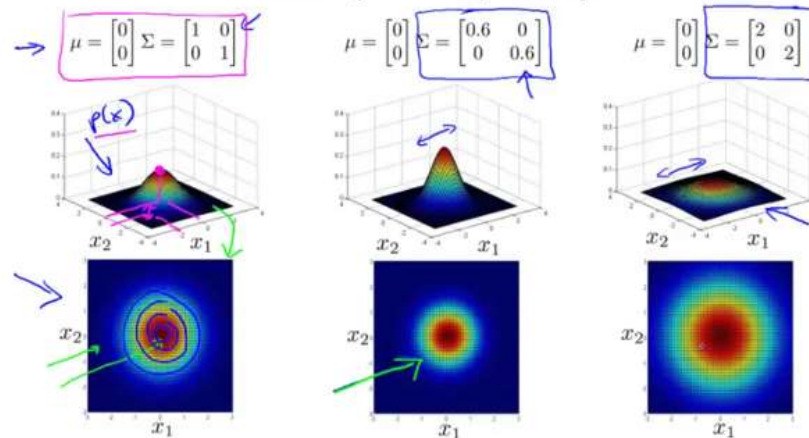


- <http://mydov.blogspot.com/2014/06/anomaly-detection.htm>
- <https://www.quora.com/What-is-an-intuitive-explanation-for-multivariate-Gaussian-distribution-aka-multivariate-normal>

© 2018 Peter V. Henstock

## Zero Covariance Term

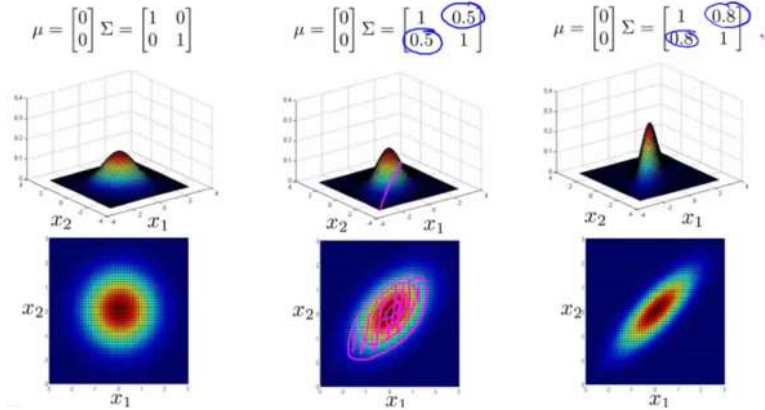
Multivariate Gaussian (Normal) examples



© 2018 Peter V. Henstock

## Non-zero Covariance Term

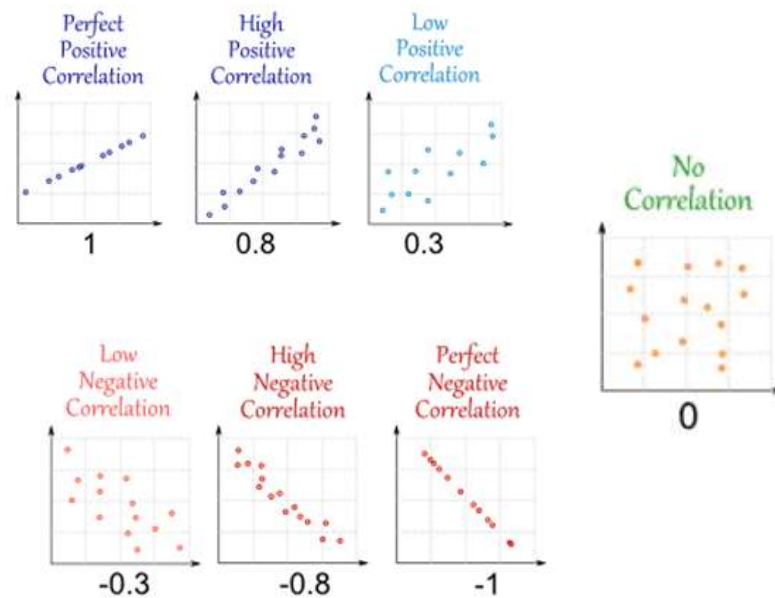
The above figure shows, that by changing the values of  $\mu$ , the overall position of the contour profile



© 2018 Peter V. Henstock

## Correlation

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



© 2018 Peter V. Henstock

## Correlation coefficient

- $r = \text{CorrCoef}(X,Y) = \text{cov}(X,Y) / (s_X s_Y)$ 
  - Technically, the Pearson corr. Coef.

- $$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Range of values for correlation is  $[-1, 1]$
- No units

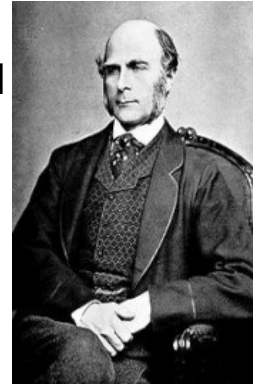
© 2018 Peter V. Henstock

# Regression

© 2018 Peter V. Henstock

## Sir Francis Galton

- 1822-1911 Victorian England
- Statistician & scientist
  - Regression
  - Standard deviation
  - Correlation
  - Psychometrics
  - Fingerprint classification
  - Weather map and scientific meteorology
- Coined terms in our lexicon:
  - Eugenics
  - Nature vs. Nurture



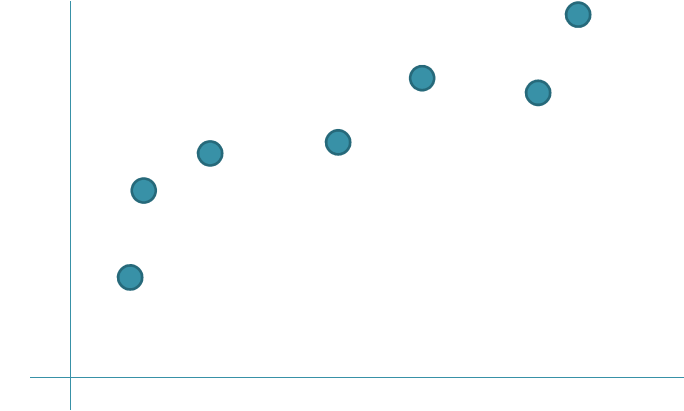
© 2018 Peter V. Henstock

## Types of Learning

- Supervised
  - Provide output labels/values & input features
- Unsupervised
  - Provide only input features
- Semi-supervised learning
- Reinforcement learning

© 2018 Peter V. Henstock

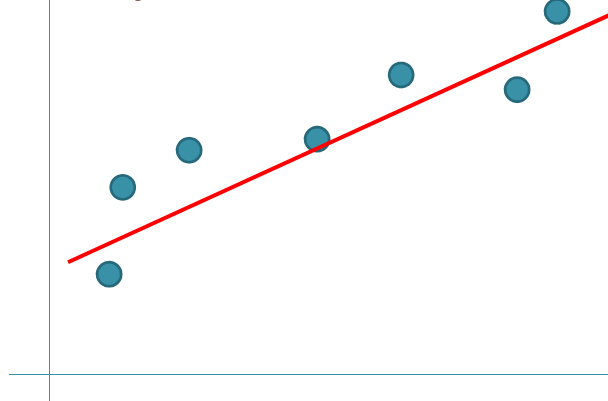
## Let's model data



- What would be the simplest model for this type of result

© 2018 Peter V. Henstock

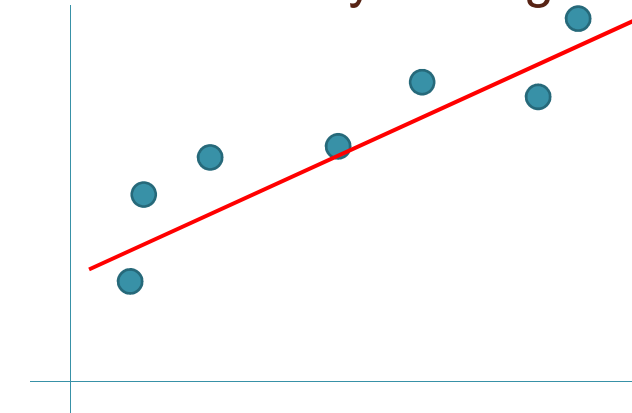
## Let's eyeball draw a line



- How is this model?
- Does it fit the data well?

© 2018 Peter V. Henstock

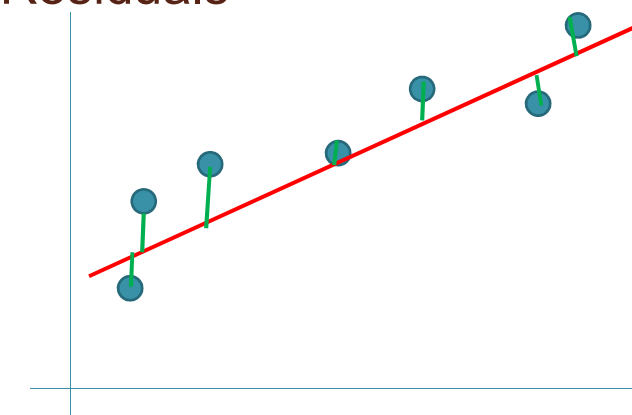
## Alternative to eyeballing it



- How could we come up with a description of what a good linear fit would be for this set of blue dots?

© 2018 Peter V. Henstock

## Residuals



- Residual = estimate on line – actual blue
- Residuals are drawn with green lines

© 2018 Peter V. Henstock

## Regression vs. PCA

- Is this regression line fit the same one as the principal component?

© 2018 Peter V. Henstock

## Generative Model

- What is it?

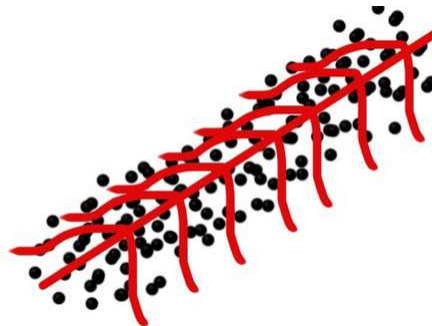
© 2018 Peter V. Henstock

## Generative Model

- What does generative mean?
  - Could try to model the residuals and come up with some useful ideas on model
  - Instead we are going to create a model that assumes the points were distributed by a generating model
  - Specifically, we start with points on a line and randomly shift the points according to a distribution
- $y_i = w_0 + w_1 x_i + \varepsilon_i$
- $\varepsilon_i \sim N(\text{mean}=0, \text{stdev}=\sigma)$
- What does this have to do with residuals?

© 2018 Peter V. Henstock

## Residuals



- Residuals come from a normal distribution centered on the line
- <http://stats.stackexchange.com/questions/148803/how-does-linear-regression-use-the-normal-distribution>

© 2018 Peter V. Henstock



## Residuals → Best Fit Model

- How can we use the residuals to come up with a best fit line?
- What would be a good function to optimize?
  - a)  $\sum \text{residuals}$
  - b)  $\sum |\text{residuals}|$
  - c)  $\sum [\text{residuals}]^2$
  - d)  $\sum [y_i - \text{residuals}]^2$
  - e)  $\sum [\text{residuals} / y_i]^2$

© 2018 Peter V. Henstock

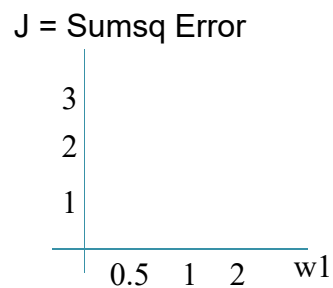
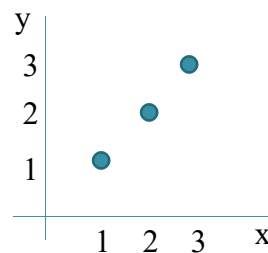
## Simple case

- Optimize the least squares function
- $\text{SumsqError} = \sum_{i=1}^N \epsilon_i^2$
- But  $y_i = w_0 + w_1 x_i + \epsilon_i$  but we don't know the true weights so we estimate
- so  $\epsilon_i = y_i - \widehat{w}_0 - \widehat{w}_1 x_i$
- $\text{SumsqError} = \sum_{i=1}^N (y_i - \widehat{w}_0 - \widehat{w}_1 x_i)^2$
- How can we minimize the error?
- How can we find the best  $\widehat{w}_0$  and  $\widehat{w}_1$ ?

© 2018 Peter V. Henstock

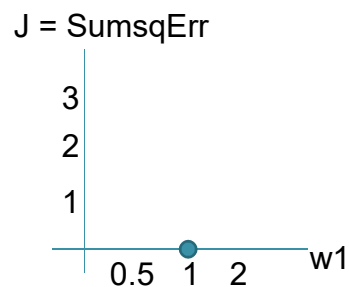
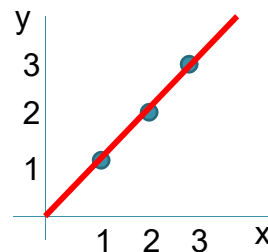
## Sanity check example

- Sumsq error = cost function =  $J$
- Goal is to minimize  $J$
- $J$  = sum-square distance between
  - predicted  $y$  values of line given parameters
  - the actual  $y$  values
  - at each  $x$  point



© 2018 Peter V. Henstock

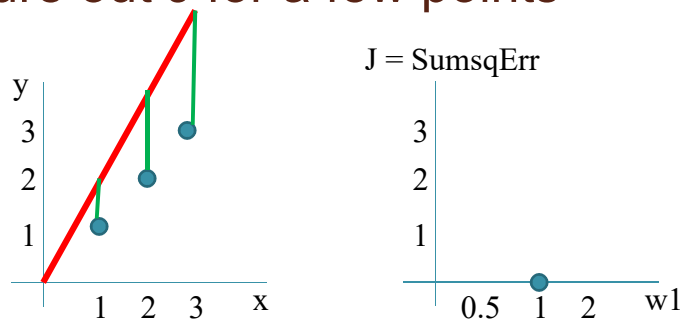
## Figure out $J$ for a few points



- Assume  $w_0 = 0$  (zero y-intercept)
- For  $w_1 = 1$ :

© 2018 Peter V. Henstock

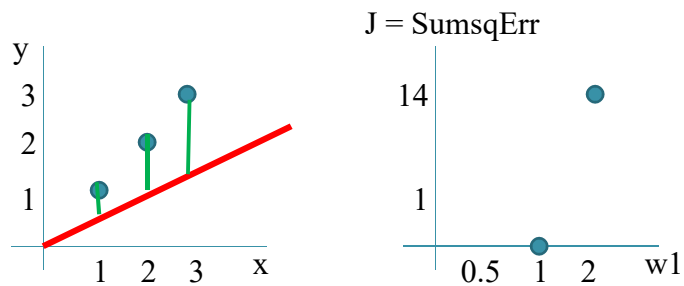
## Figure out J for a few points



- Assume  $w_0 = 0$  (zero y-intercept)
- For  $w_1 = 1$ :  $J = 0$
- For  $w_1 = 2$ :

© 2018 Peter V. Henstock

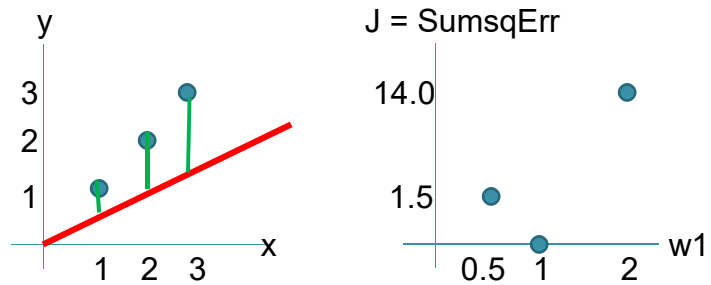
## Figure out J for a few points



- Assume  $w_0 = 0$  (zero y-intercept)
- For  $w_1 = 1$ :  $J = 0$
- For  $w_1 = 2$ :  $J = (2-1)^2 + (4-2)^2 + (6-3)^2 = 14$
- For  $w_1 = 0.5$ :

© 2018 Peter V. Henstock

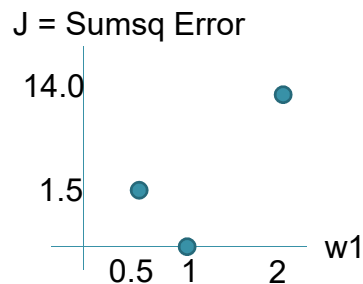
## Figure out J for a few points



- Assume  $w_0 = 0$  (zero y-intercept)
- For  $w_1 = 1$ :  $J = 0$
- For  $w_1 = 2$ :  $J = (2-1)^2 + (4-2)^2 + (6-3)^2 = 14$
- For  $w_1 = 0.5$ :  $J = (1-1/2)^2 + (2-1)^2 + (1.5-1)^2 = 1.5$

© 2018 Peter V. Henstock

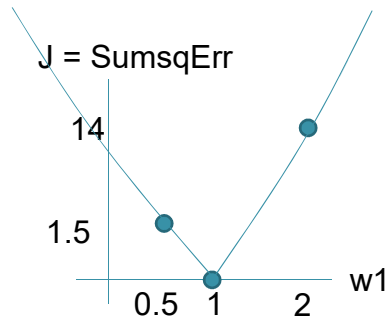
## Modeling the error



© 2018 Peter V. Henstock

## What shape is $J$ vs. $w_1$ ?

- Parabola (badly drawn below)



- Min value will be 0 in this case—why?
- How could we optimize the  $J$  as a function of  $w_1$ ?

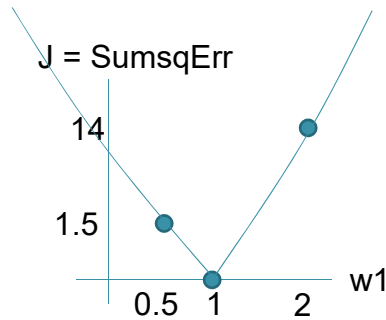
© 2018 Peter V. Henstock

## Derivative Formulation

© 2018 Peter V. Henstock

## What shape is J vs. w1?

- Parabola (badly drawn below)



- Min value will be 0 in this case—why?
- How could we optimize the J as  $f(w_1)$ ?

© 2018 Peter V. Henstock

## Take derivative and set to 0

- $\sum_{i=1}^N (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$
- Differentiate with respect to....?

© 2018 Peter V. Henstock

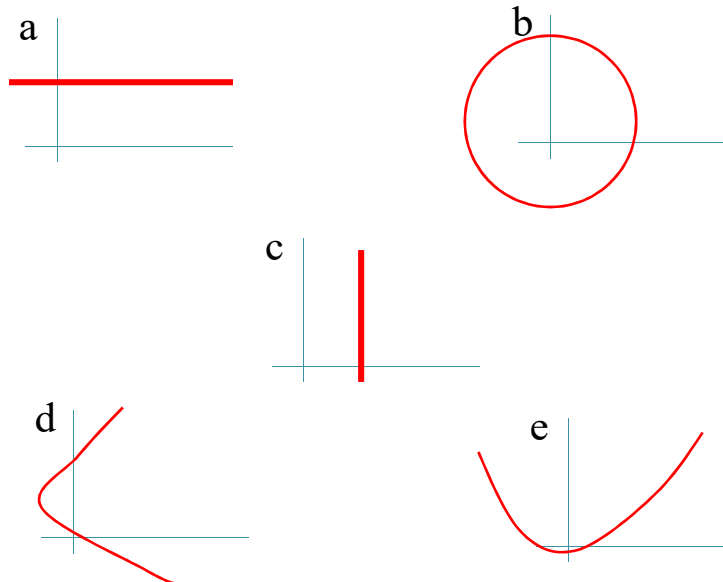
## Take derivative and set to 0

- $\sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$
- Differentiate with respect to the  $w_0$  &  $w_1$
- $\frac{\partial y}{\partial w} \sum_{i=1}^N (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2 =$
- $2 \sum_{i=1}^N y_i - 2N\hat{w}_0 - \hat{w}_1 \sum_{i=1}^N x_i$
- $\hat{w}_0 = \frac{1}{N} [\sum_{i=1}^N y_i - \hat{w}_1 \sum_{i=1}^N x_i]$
- $\hat{w}_1 = \frac{\sum_{i=1}^N y_i x_i - \frac{\sum_{i=1}^N y_i \sum_{i=1}^N x_i}{N}}{\sum_{i=1}^N x_i^2 - \frac{(\sum_{i=1}^N x_i)^2}{N}}$

<http://isites.harvard.edu/fs/docs/icb.topic515975.files/OLSDerivation.pdf>

© 2018 Peter V. Henstock

## Which work for linear regression?



© 2018 Peter V. Henstock

## Ordinary Least Squares vs. MLE

- Previous derived a generative model that was based on least-squares
- Reasonable approach
- MLE = maximum likelihood estimation
- Different framework for optimizing
- Likelihood =  $P(\text{Data} \mid \text{hyp})$
- MLE =  $\arg \max P(\text{Data} \mid \text{hyp})$ 
  - Select the parameters that maximize this quantity as our estimates

© 2018 Peter V. Henstock

## MLE Estimate

- Assume  $P(\text{data}_i \mid \text{hyp})$  are independent
- MLE =  $\operatorname{argmax} \prod_{i=1}^N P(\text{data}_i \mid \text{hyp})$
- What is this  $P(\text{data}_i \mid \text{hyp})$  using our previous assumptions?
- $\operatorname{argmax} \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$
- What are we optimizing over in  $\operatorname{argmax}$ ?

© 2018 Peter V. Henstock



## MLE Optimizing

- Argmax  $\prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$
- To make the math easy, take a log
- To optimize, take derivative with respect to the betas
- Result in this case will be identical to the least squares approach
  - MLE estimates equal to OLS in this case

© 2018 Peter V. Henstock

## MLE Estimate

- Assume  $P(\text{data}_i | \text{hyp})$  are independent
- $\text{MLE} = \text{argmax} \prod_{i=1}^N P(\text{data}_i | \text{hyp})$
- What is this  $P(\text{data}_i | \text{hyp})$  using our previous assumptions?

© 2018 Peter V. Henstock

# Multiple Linear Regression

© 2018 Peter V. Henstock

## Multiple Linear Regression

- Previously had  $y = w_0 + w_1x_1$
- What if we had 3 or 10 variables?

- We can still use regression

- Here is a simple case:

$$Y = w_0 + w_1x_1 + w_2x_2 + \dots w_kx_k = \sum w_ix_i = w^T x$$

© 2018 Peter V. Henstock

## Multiple Regression with Matrices

- $Y = XW + e$
- $Y$  is  $N \times 1$  matrix
  - What do we usually call  $Y$ ?

© 2018 Peter V. Henstock

## Multiple Regression with Matrices

- $Y = XW + e$
- $Y$  is  $N \times 1$  matrix
- $X$  is  $N \times (k+1)$  include a column of 1s
- $W$  is  $(k+1) \times 1$
- What do we usually call  $Y$ ?

© 2018 Peter V. Henstock

## Multiple Regression with Matrices

- $Y = XW + e$
- $Y$  is  $N \times 1$  matrix
- $X$  is  $N \times (k+1)$  include a column of 1s
- $W$  is  $(k+1) \times 1$
- What do we usually call  $Y$ ?
- What do we usually call  $X$ ?

© 2018 Peter V. Henstock

## Multiple Regression with Matrices

- $Y = XW + e$
- $Y$  is  $N \times 1$  matrix (column vector)
- $X$  is  $N \times (k+1)$  include a column of 1s
- $W$  is  $(k+1) \times 1$  matrix
- $e$  is  $N \times 1$  column vector
- What do we usually call  $Y$ ?
- What do we usually call  $X$ ?
- What do we usually call  $W$ ?

© 2018 Peter V. Henstock

## Multiple Regression with Matrices

- $Y = XW + e$
- $Y$  is  $N \times 1$  matrix (column vector)
- $X$  is  $N \times (k+1)$  include a column of 1s
- $W$  is  $(k+1) \times 1$  matrix
- $e$  is  $N \times 1$  column vector
- What are the rows and column referring to for the  $X$ ?

© 2018 Peter V. Henstock

## Matrix Inverse

- For numbers:
  - Consider  $x * 1/x \rightarrow 1$
  - In a sense,  $1/x$  is inverse of  $x$
  - In a sense,  $x$  is inverse of  $1/x$
- For matrices
  - $AA^{-1} = I$  and  $A^{-1}A = I$
  - Fairly easy for  $2 \times 2$  matrices
  - Possible for  $3 \times 3$  matrices by hand
  - Challenging for anything larger so use any linear algebra software or python

© 2018 Peter V. Henstock

## Matrix Inverse

- Do all numbers have an inverse?
- Do all matrices have an inverse?

© 2018 Peter V. Henstock

## Matrix Inverse

- Do all numbers have an inverse?
- Do all matrices have an inverse?
  - Officially, has to be a square matrix
  - (Note: there are pseudo inverses otherwise)
  - If it's square it also has to have "full rank"
    - Can't have rows that are products of each other
    - Can't have zero rows
    - Can't have rows that are weighted sums of any other rows
    - Determinant cannot be 0

© 2018 Peter V. Henstock

## Least Squares Estimate

- $SSE = e^T e$ 
  - Is this an inner product or outer product?
- $e = Y - XW$
- $SSE = (Y - XW)^T (Y - XW)$
- $SSE = (Y^T - W^T X^T) (Y - XW)$ 
  - Note that  $(AB)^T = B^T A^T$
- $SSE = (Y^T Y - W^T X^T Y + W^T X^T X W - Y^T X W)$ 
  - The 2<sup>nd</sup> and 4<sup>th</sup> term are equivalent
    - Not obvious from the matrix format
    - True when you multiply them out

© 2018 Peter V. Henstock

## Least Squares Estimate

- $SSE = e^T e$ 
  - Is this an inner product or outer product?
- $e = Y - XW$
- $SSE = (Y - XW)^T (Y - XW)$
- $SSE = (Y^T - W^T X^T) (Y - XW)$ 
  - Note that  $(AB)^T = B^T A^T$
- $SSE = (Y^T Y - W^T X^T Y + W^T X^T X W - Y^T X W)$ 
  - The 2<sup>nd</sup> and 4<sup>th</sup> term are equivalent
- $SSE = (Y^T Y - 2Y^T X W + W^T X^T X W)$

© 2018 Peter V. Henstock

## Least Squares Estimate

- $SSE = (Y^T Y - 2Y^T XW + W^T X^T XW)$
- How to minimize this?
- Derivative wrt  $W$  and set to 0
- $d/dW \rightarrow 0 - 2X^T Y + 2X^T XW = 0$ 
  - Note that  $d/dW$  of  $U^T V W = V^T U$
  - Last term with the  $W$ 's on either side is a quadratic which yields 2 and removal of  $W^T$
  - $X^T X W = X^T Y \rightarrow W = (X^T X)^{-1} X^T Y$
  - $(X^T X)^{-1} X^T$  is the Moore-Penrose Pseudoinverse of  $X$
- Why can't we just do an inverse of  $X$ ?

© 2018 Peter V. Henstock

## Observation

- From before:
  - $Y = w_0 + w_1 x_1 + w_2 x_2 + \dots w_k x_k = \sum w_i x_i$
- What does  $x_1$  actually look like?
- If we modified it to be  $3x_1$ , what would happen to the overall equation?
- What if we used  $x_1^2$  or cubed it to  $x_1^3$ ?

© 2018 Peter V. Henstock



## Observation

- From before:
  - $Y = w_0 + w_1x_1 + w_2x_2 + \dots w_kx_k = \sum w_ix_i$
- What does  $x_1$  actually look like?
- If we modified it to be  $3x_1$ , what would happen to the overall equation?
- What if we used  $x_1^2$  or cubed it to  $x_1^3$ ?
  - Obtain polynomial regression

© 2018 Peter V. Henstock

## Full multiple linear regression

- $Y = w_0 + w_1x_1 + w_2x_2 + \dots w_kx_k = \sum w_ix_i$
- $Y = w_0 + w_1x_1 + w_2x_2 + \dots w_kx_k + b_0x_1x_2 + b_1x_1x_3 + \dots b_{zx_{k-1}}x_k + c_0x_1x_2x_3 + \dots$
- $Y =$  sum of
  - Independent factors +
  - 2-way interactions +
  - 3-way interactions +
  - K-way interactions
- Can include squared, cubed, power terms

© 2018 Peter V. Henstock

## Hierarchy Principles

- Only found this is statistics and never machine learning
- Need to include all lower terms if you include a higher order term
- If you have  $x_1^3$ 
  - Need to include  $x_1^2$  as well as  $x_1$
- If you have  $x_1^2x_2$ 
  - Need to include  $x_1^2$ ,  $x_1x_2$ ,  $x_1$ ,  $x_2$

© 2018 Peter V. Henstock

## Further options

- $Y = \text{sqrt}(X_1^5X_2X_3^3X_4)$
- Can multiple linear regression work for this kind of model?
- What does Y look like?
- What is  $\log(ab)$ ?  $\rightarrow \log(a) + \log(b)$
- $\log(\text{sqrt}(x))$ ?  $\rightarrow 1/2\log(x)$

© 2018 Peter V. Henstock

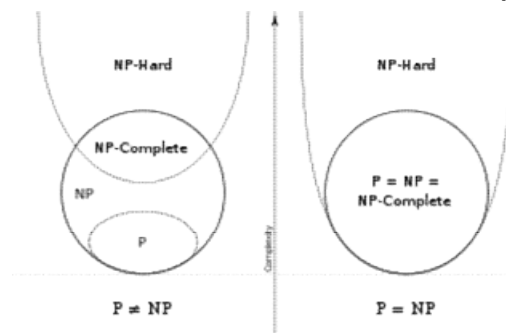
## Big O notation

- Algorithms tend to be characterized by their approximate computational load
- Computation is a function of a parameter like  $n$  = size
- Sum of  $N$  numbers is  $O(N)$
- Tree search is  $O(\log N)$
- We tend to round down so if the actual computation was  $O(5N^3 + 4N + 17)$ 
  - Use just  $O(N^3)$

© 2018 Peter V. Henstock

## Complexity

- $P$  = polynomial
- $NP$  = nondeterministic polynomial time
- $NP$ -complete =  $NP$  and  $NP$ -hard
- $NP$ -hard =
  - at least as hard as hardest problem in  $NP$



[https://en.wikipedia.org/wiki/NP-hard#/media/File:P\\_np\\_np-complete\\_np-hard.svg](https://en.wikipedia.org/wiki/NP-hard#/media/File:P_np_np-complete_np-hard.svg)

© 2018 Peter V. Henstock

## Computational aspects

- Pseudoinverse
- Solution parameters =  $(X^T X)^{-1} X^T y$
- What is the computational requirement?
- Matrix addition?
- Matrix multiplication?
- Matrix inverse?

© 2018 Peter V. Henstock

# Gradient Descent

© 2018 Peter V. Henstock

## Gradient Descent

- Problem:
  - You are placed on a hill in fog
- Goal:
  - Find the bottom of the hill
- Method
  - How would you do that?

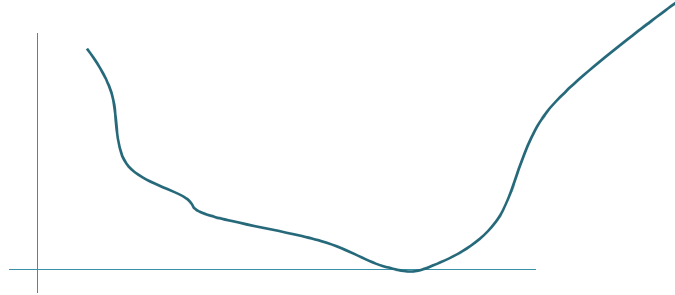
© 2018 Peter V. Henstock

## Gradient Descent

- Idea: you are placed on a hill in fog
  - Find the bottom of the hill
- do {
  - Figure out which way is “down”
  - Move a certain distance in that direction
- } until at lowest spot
- What assumptions are we making?

© 2018 Peter V. Henstock

## Calculus Approach

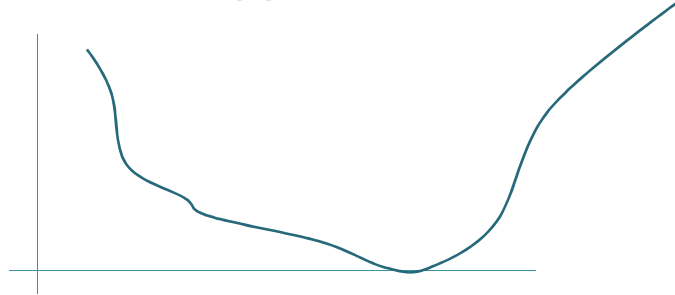


- Calculus tells us if that we have a function  $f$  then we need to do 2 things to find the minimum

- 1)
- 2)

© 2018 Peter V. Henstock

## Calculus Approach

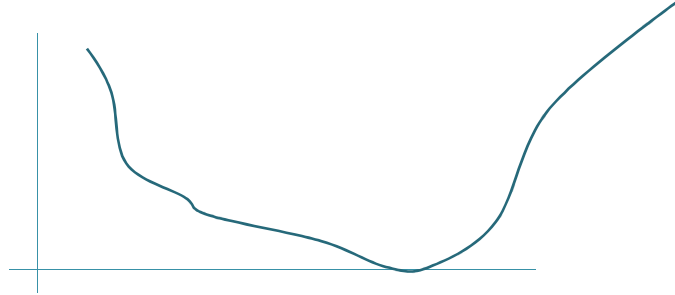


- Calculus tells us if that we have a function  $f$  then we need to do 2 things to find the minimum

- 1) Compute the derivative
- 2) Check the concavity using 2<sup>nd</sup> derivative

© 2018 Peter V. Henstock

## Calculus Approach



- Calculus tells us if that we have a function  $f$  then we need to do 2 things to find the minimum
    - 1) Compute the derivative
    - 2) Check the concavity using 2<sup>nd</sup> derivative
- How do we actually locate the minimum?

© 2018 Peter V. Henstock

## General Gradient Descent

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n), \quad n \geq 0.$$

- For a partial derivative of 1 variable
  - $w_{n+1} = w_n - \alpha \frac{\partial}{\partial w_n} J(w_n)$
  - $\alpha$  is the learning rate
  - $J()$  is the cost function
  - Iteratively adjusting  $w_n$  with better estimates using the gradient of the cost function as the direction.



© 2018 Peter V. Henstock

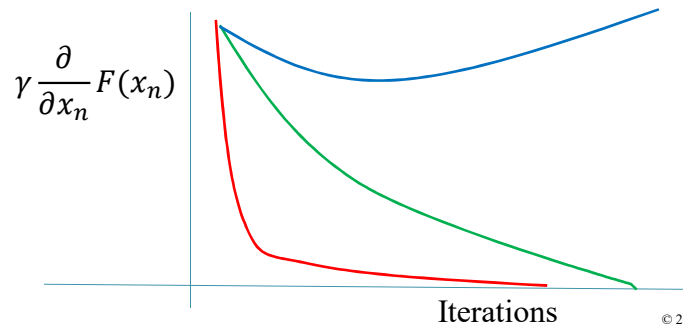
## Diagnostics for Gradient Descent

- What should the gradient look like as a function of time (iterations)?
- Hint:
  - When should gradient descent end?

© 2018 Peter V. Henstock

## Diagnostics for Gradient Descent

- What should the gradient look like as a function of time (iterations)?
- Hint:
  - When should gradient descent end?

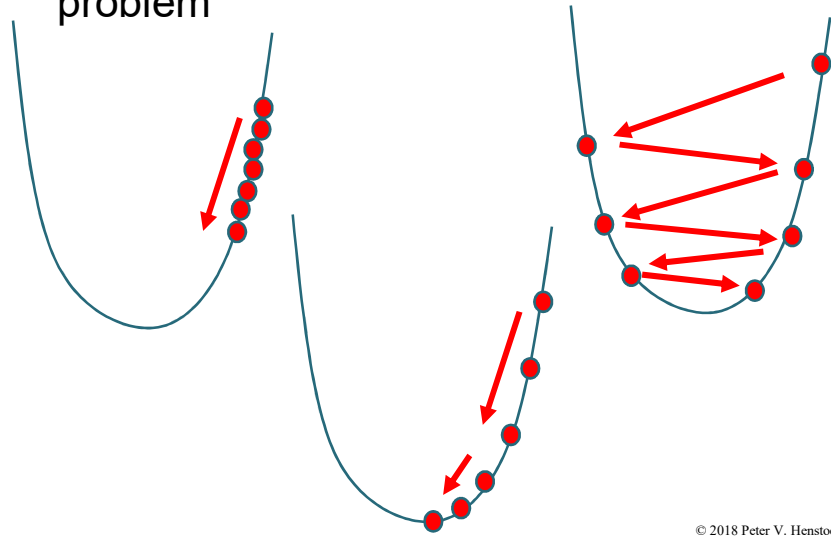


© 2018 Peter V. Henstock



## How large a learning rate?

- It depends...because it's a Goldilocks problem

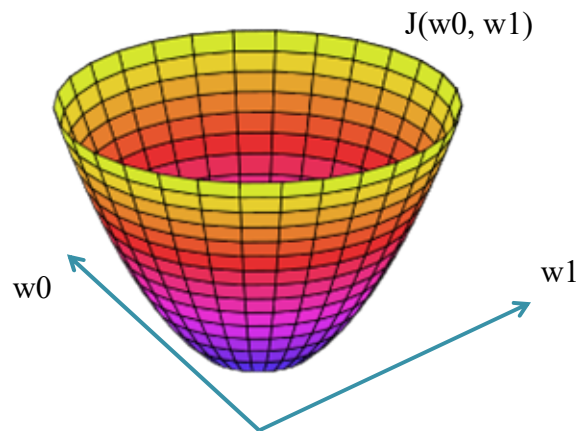


## Gradient Descent

- Have an optimization function  $J$  we are trying to minimize
- Have 2 parameters  $w_0$  and  $w_1$
- Start with an initial guess of  $w_0'$  and  $w_1'$
- Iteratively shift  $w_0$  and  $w_1$  in the direction to make  $J$  better until “done”

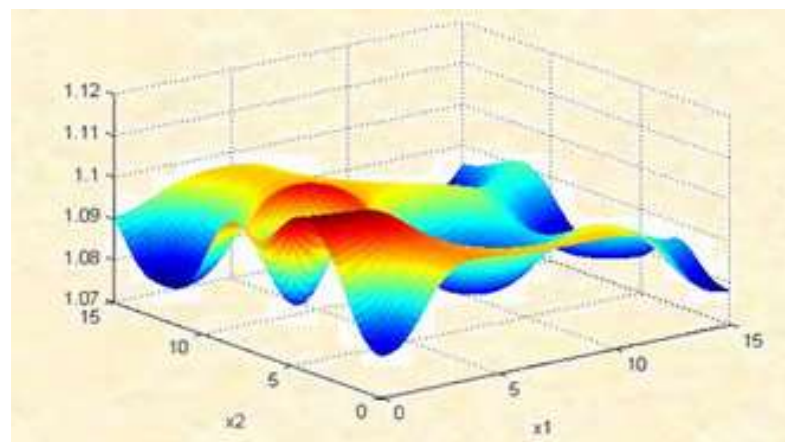
© 2018 Peter V. Henstock

2D problem is same idea



© 2018 Peter V. Henstock

Will Gradient Descent Work?



© 2018 Peter V. Henstock

## 2D Gradient Descent

- For each loop
  - $w_1 \leftarrow w_1 - \alpha \partial J / \partial w_1$
  - $w_0 \leftarrow w_0 - \alpha \partial J / \partial w_0$
- Note: J is a function of  $w_0$  and  $w_1$ 
  - Be sure to update both new values using the older values
  - Do not update  $w_1$  and use the updated  $w_1$  to update  $w_0$

© 2018 Peter V. Henstock

## 2D Gradient Descent

- For each loop
  - $w_1 \leftarrow w_1 - \alpha \partial J / \partial w_1$
  - $w_0 \leftarrow w_0 - \alpha \partial J / \partial w_0$
- $\partial J / \partial w_0 = 2 \sum_i (w_1 x_i + w_0 - y_i)$
- $\partial J / \partial w_1 = 2 \sum_i (w_1 x_i + w_0 - y_i) x_i$

© 2018 Peter V. Henstock

## How to ensure achieved optimal?

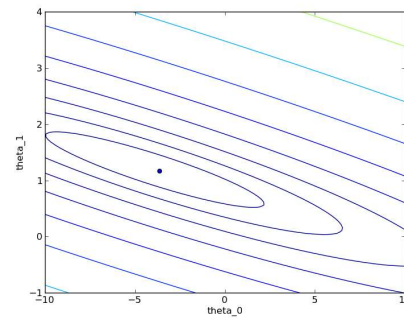
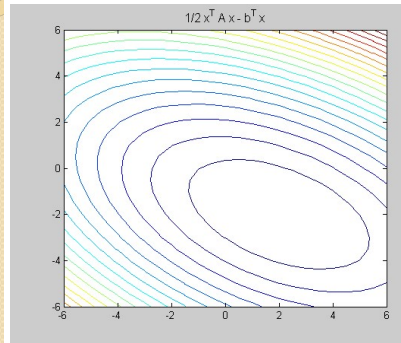
© 2018 Peter V. Henstock

## How to ensure achieved optimal?

- Start with multiple starting points
- Typically try a few different learning rates and plot the change of gradient to ensure things are going the right direction

© 2018 Peter V. Henstock

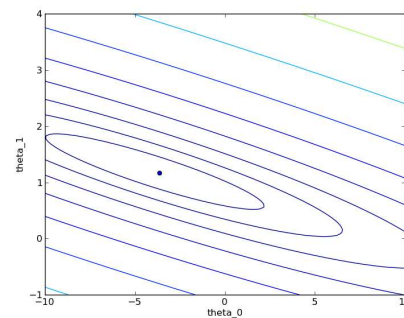
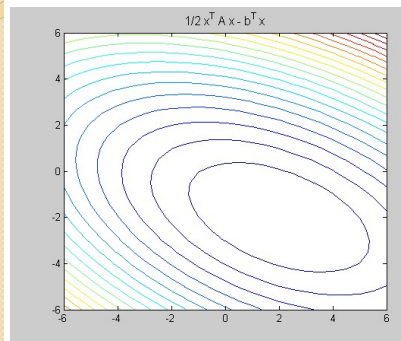
## Which is easier for gradient descent?



- [http://www.informatik.uni-konstanz.de/fileadmin/informatik/ag-saupe/Webpages/lehre/na\\_08/Lab1/10\\_CG/html/myCG.html](http://www.informatik.uni-konstanz.de/fileadmin/informatik/ag-saupe/Webpages/lehre/na_08/Lab1/10_CG/html/myCG.html)
- <http://aimotion.blogspot.com/2011/10/machine-learning-with-python-linear.html>

© 2018 Peter V. Henstock

## How to solve this problem?



- [http://www.informatik.uni-konstanz.de/fileadmin/informatik/ag-saupe/Webpages/lehre/na\\_08/Lab1/10\\_CG/html/myCG.html](http://www.informatik.uni-konstanz.de/fileadmin/informatik/ag-saupe/Webpages/lehre/na_08/Lab1/10_CG/html/myCG.html)
- <http://aimotion.blogspot.com/2011/10/machine-learning-with-python-linear.html>

© 2018 Peter V. Henstock

# Regression Diagnostics

© 2018 Peter V. Henstock

## Regression Process

- We load the data
- We perform a fit
- We're done!
- What could go wrong?

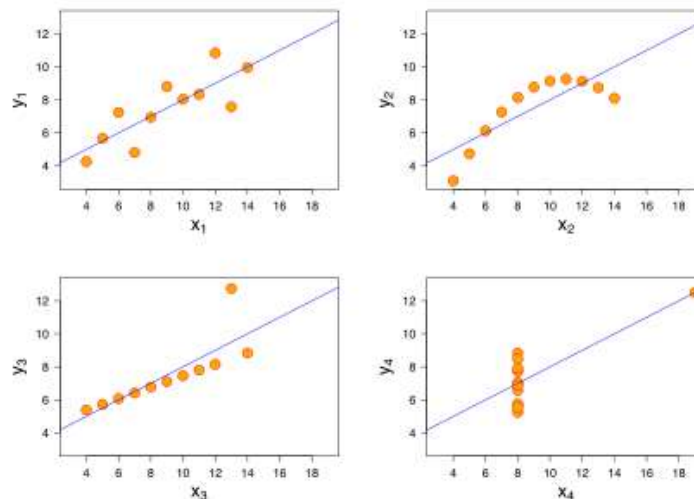
© 2018 Peter V. Henstock

## Regression Process

- We load the data
- We perform a fit
- We're done!
  
- What could go wrong?
- Assumptions:
  - iid errors = independent & identically distributed
  - Common variance
  - $N(0, \text{variance})$
  - Y and X are related

© 2018 Peter V. Henstock

## Classic Anscombe Example



[https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet) 1973  
 Anscombe, F. J. 1973. Graphs in statistical analysis. The American Statistician, 27, pp 17-21.

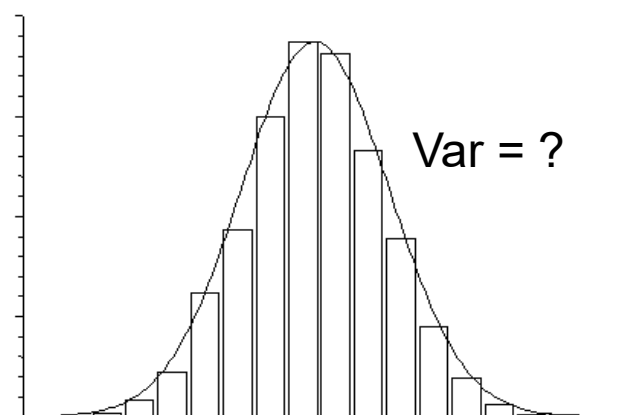
© 2018 Peter V. Henstock

## How can we diagnose the issues?

- Residuals should have a common variance
- Residuals should have a  $N(0, \text{variance})$
- Residuals should be iid

© 2018 Peter V. Henstock

## Histogram of Residuals



Var = ?

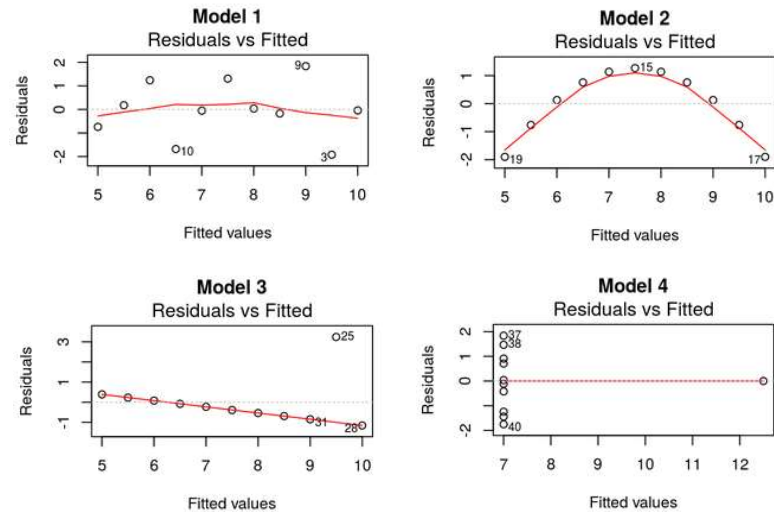
Center = ?

Symmetry = ?

© 2018 Peter V. Henstock



## Residuals vs. X



© 2018 Peter V. Henstock

## Heteroscedascity

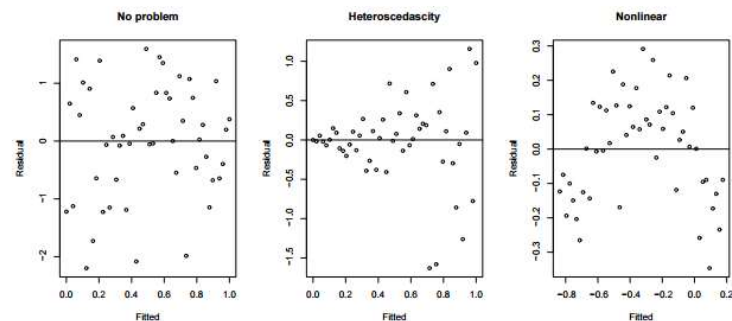


Figure 7.5: Residuals vs Fitted plots - the first suggests no change to the current model while the second shows non-constant variance and the third indicates some nonlinearity which should prompt some change in the structural form of the model

- Faraway: “Practical Regression and ANOVA using R” July 2002

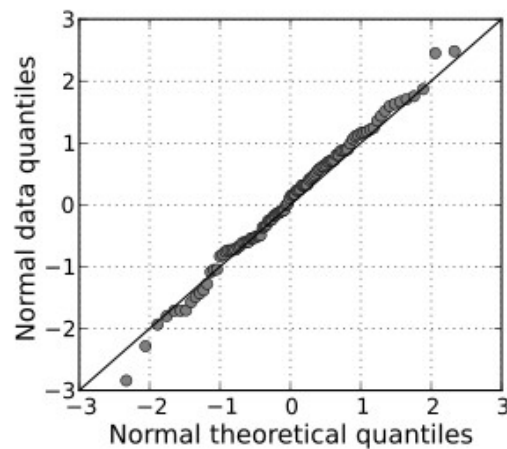
© 2018 Peter V. Henstock

## Data Transformations

- Common transformations include:
  - Log
  - Sqrt
  - Arcsin
- What do we apply these to?
  - Y?
  - X?

© 2018 Peter V. Henstock

## Quantile-Quantile (Q-Q) Plot

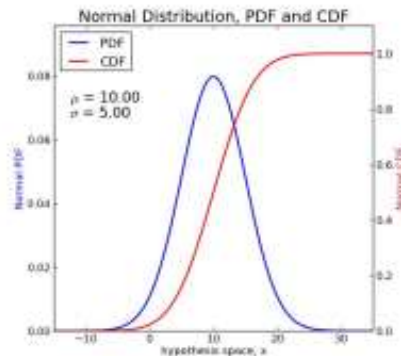


- Look for linear relationship
- Should have at least  $[-1 \ 1]$  linear

© 2018 Peter V. Henstock

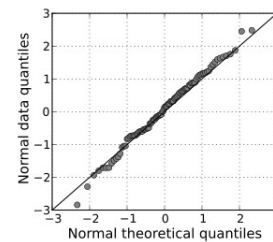
## CMF = cumulative mass function

- $\text{Prob}(a \text{ to } b) = \sum_a^b pmf$
- $\text{Cmf}(x) = \text{cumulative pmf} = \sum_{-\infty}^x pmf$ 
  - Maps from  $x \rightarrow [0, 1]$  probability
- $\text{InvCmf}(p)$  maps  $[0, 1]$  to an  $x$ -value

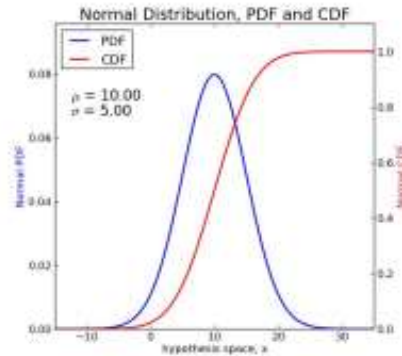


## Quantile-Quantile (Q-Q) Plot

- Sort the residuals (y-axis)
  - Use Order Statistics
- Sample invnormal cdf evenly
  - Sample at probabilities  $1/(n+1), 2/(n+1) \dots n/(n+1)$
  - Convert the probabilities back to  $x$ -values for the  $x$ -axis
- Plot the sampled invcdf vs. residuals



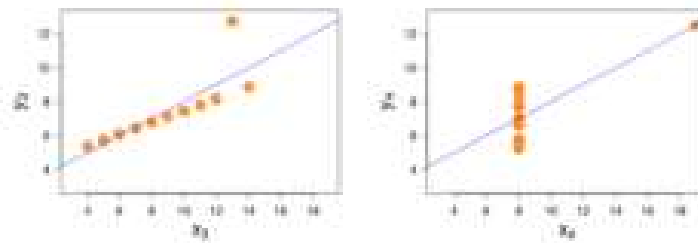
## Q-Q plot intuition



- Central value of the distribution (50%)
  - Should match the mean
- Value  $\text{mean} \pm \text{stdev}$  should line up
- Method works for other distributions too

© 2018 Peter V. Henstock

## What about these cases?



© 2018 Peter V. Henstock

## Leverage Points

- A single isolated point might have a lot of influence on the overall fit
- What characteristics might go into such a point?

© 2018 Peter V. Henstock

## Leverage Points

- A single isolated point might have a lot of influence on the overall fit
- What characteristics might go into such a point?
  - 1) Shifts the line
  - 2) Isolated in x from other points

© 2018 Peter V. Henstock

## Leverage Points

- 1) Shifts the line
- 2) Isolated in x from other points

- $$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

- High leverage point if  $> 4/n$
- Bad leverage point: distorts line (outlier)
- Good leverage point: consistent with line

© 2018 Peter V. Henstock

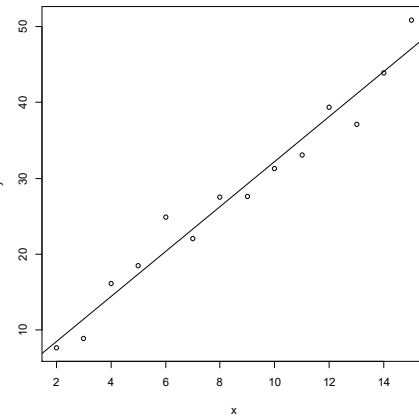
## R Example for Regression

© 2018 Peter V. Henstock

## What does R do?

```
x <- 2:15
y <- 1:length(x)*3 + 5 +
  rnorm(length(x))*2.5
plot(x,y)
```

```
fit <- lm(y ~ x)
summary(fit)
hist(resid(fit))
plot(fit)
```



© 2018 Peter V. Henstock

## R call

```
Call:
lm(formula = y ~ x)

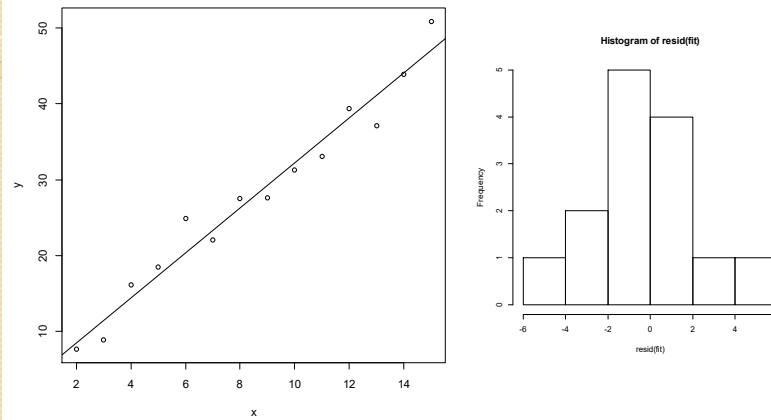
Residuals:
    Min       1Q   Median       3Q      Max
-2.2088 -1.3598  0.3992  1.3213  1.6921

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0393     0.9639   1.078   0.302
x              3.0456     0.1025  29.724 1.31e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.545 on 12 degrees of freedom
Multiple R-squared:  0.9866,    Adjusted R-squared:  0.9855
F-statistic: 883.5 on 1 and 12 DF,  p-value: 1.314e-12
```

© 2018 Peter V. Henstock

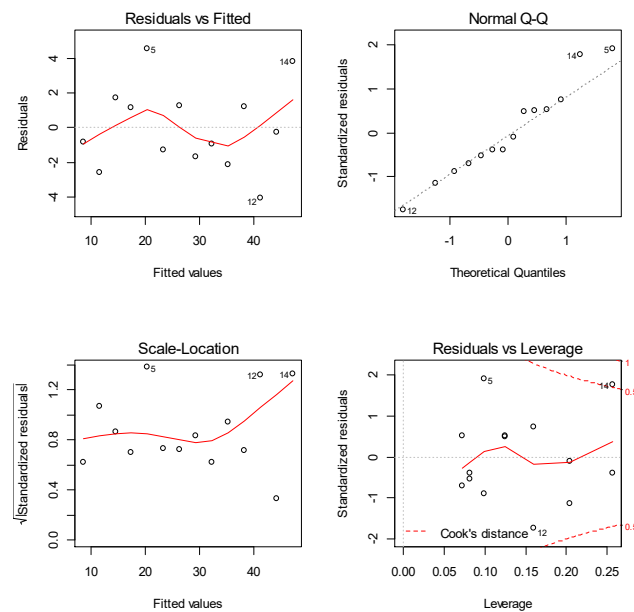
## Original data with fit



- Checking fit and residual distribution

© 2018 Peter V. Henstock

## Regression diagnostics



© 2018 Peter V. Henstock



## Actual Fit Values

```

Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-2.2088 -1.3598  0.3992  1.3213  1.6921

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0393     0.9639   1.078   0.302
x             3.0456     0.1025  29.724 1.31e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.545 on 12 degrees of freedom
Multiple R-squared:  0.9866,    Adjusted R-squared:  0.9855
F-statistic: 883.5 on 1 and 12 DF,  p-value: 1.314e-12

```

© 2018 Peter V. Henstock

## Time Series Modeling



[http://www.artnet.com/artists/ert%C3%A9/complete-numbers-suite-set-of-10-HLI\\_PTOIsn\\_8aldkjAWWWw2](http://www.artnet.com/artists/ert%C3%A9/complete-numbers-suite-set-of-10-HLI_PTOIsn_8aldkjAWWWw2)

© 2018 Peter V. Henstock

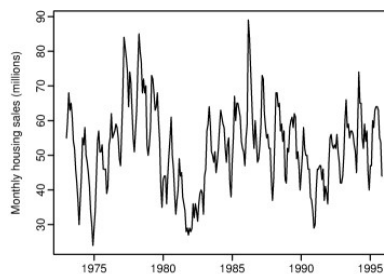
## Why Do Time Series Analysis?

- Model of data
- Interpretation
- Forecasting
- Control
- Hypothesis testing
- Simulation

© 2018 Peter V. Henstock

## Time Series Modeling

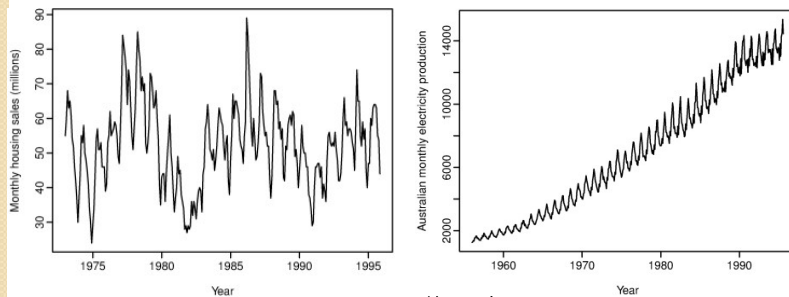
- Previously with regression:
  - $Y \sim f(X)$
  - $X$  samples are i.i.d.
  - Generative model:  $Y$  is  $f(x) + N(0, \sigma)$
- Why don't just use  $Y \sim f(t)$  ?



© 2018 Peter V. Henstock

## Signal vs. Noise Issue

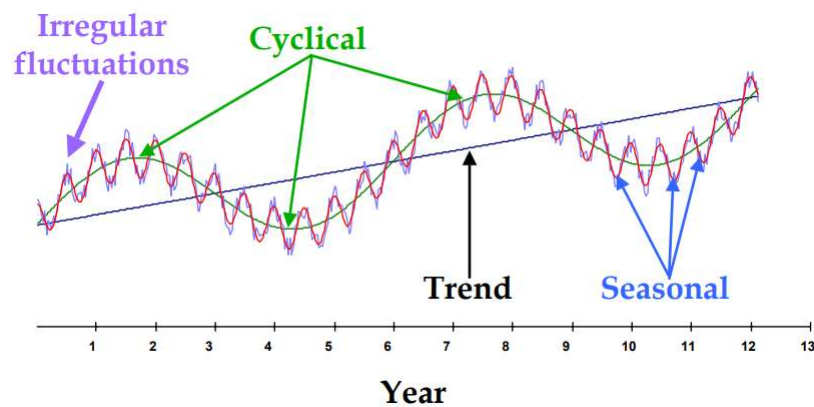
- Regression model would probably smooth the data to provide a good trend
- Decomposition:
  - Could extract the trend and seasonality
  - Are they additive or multiplicative?
  - Then model the remainder which is what?



• <https://www.otexts.org/sites/default/files/tpp/images/decomp1.png>

© 2018 Peter V. Henstock

## Time Series Components



• [http://www.statistik.wiso.uni-goettingen.de/veranstaltungen/graduateseminar/SmoothingMethods\\_Narodzinek-Karpowska.pdf](http://www.statistik.wiso.uni-goettingen.de/veranstaltungen/graduateseminar/SmoothingMethods_Narodzinek-Karpowska.pdf)

© 2018 Peter V. Henstock

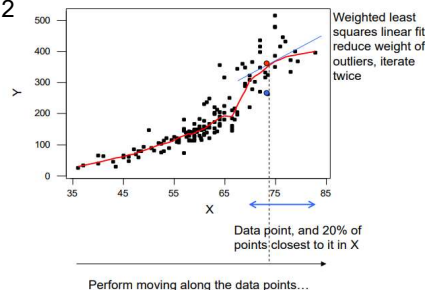
## Smoothing

- Moving Average: average over last  $m$ 
  - $x(n) = \frac{1}{m} \sum_{k=0}^{m-1} x(n-k)$
- Weighted Moving Average: avg over last  $m$ 
  - $x(n) = \frac{1}{\sum_{k=0}^m w_k} \sum_{k=0}^m w_k x(n-k)$
  - Weights usually sum to 1
  - Use this to heavily weight more recent
- Exponential Smoothing
  - Single:  $s(n) = \alpha x(n) + (1-\alpha)s(n-1)$
  - Double:  $s(n) = \alpha x(n) + (1-\alpha)[s(n-1) + b(n-1)]$ 
    - $b(n) = \beta[s(n)-s(n-1)] + (1-\beta)b(n-1)$

© 2018 Peter V. Henstock

## Lowess smoothing

- Lowest (weighted) regression
- Non-parametric model
- [http://sites.stat.psu.edu/~fxc11/Stat462\\_STABLE/Lect12\\_lowess.pdf](http://sites.stat.psu.edu/~fxc11/Stat462_STABLE/Lect12_lowess.pdf)
- Smoothing param 0.2  
= fraction of points
- Degree 1 = linear



Weighted least squares linear fit

$$\min_{\beta_0, \beta_1} \sum_{j \in N_i} w_j (y_j - (\beta_0 + \beta_1 x_j))^2$$

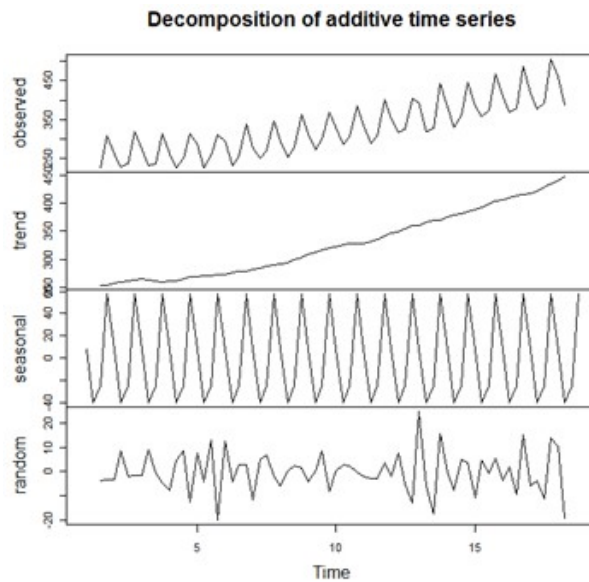
F. Chiaromonte

Weight function

$$w_i = \left( 1 - \left( \frac{d(x_j, x_i)}{\max_{\ell \in N_i} d(x_\ell, x_i)} \right)^3 \right)^3$$

© 2018 Peter V. Henstock

## Signal Decomposition

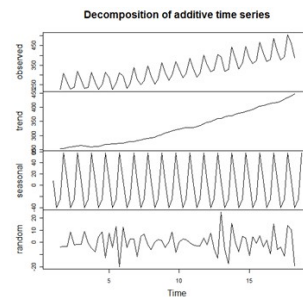


<https://onlinecourses.science.psu.edu/stat510/node/69>

© 2018 Peter V. Henstock

## In this lecture and next

- Remove the trend
  - Figure out seasonality\*
  - Figure out frequency\*
  - Model the “random”
- \* = next lecture



© 2018 Peter V. Henstock

## Time Series Framework

- Need a model to explain random signals
- Existing models may not be explainable by the  $Y \sim f(t) + \text{noise}$
- Existing models may not be predictable
- But, we know something about  $Y$
- Framework: random signals
  - $Y(t)$  characterized by a distribution at each  $t$

© 2018 Peter V. Henstock

## Framework is Random Sequence

- $x_1, x_2, \dots, x_n$  (Changed notation to  $x$ )
- Need to exploit the order of sequence
- As random, it cannot be predicted
- Deterministic process: known / predictable
- Stochastic or Random process:
  - System that generates all possible random sequences of which we have a realization  $x_1 \dots x_n$  with the specific random values
  - Goal is to model the process
  - Model process as random since it's either random or we do not have a better model

© 2018 Peter V. Henstock

## Time Series Review

- Random Process

- $w(k) = 0, 2, 0, -1, 2, -1, 2, 1, 0, 0, 1, 2, -1, -2$

- AR Model

- $x(k) = \phi_1 x(k-1) + \phi_2 x(k-2) + \dots \phi_p x(k-p) + w(k)$

- Let's use a 1st order model  $\phi_1 = 0.9$

- $x(0) = 0$

- $x(1) = 0.9 \cdot 0 + 2 = 2$

- $x(2) = 0.9 \cdot 2 + 0 = 1.8$

- $x(3) = 0.9 \cdot 1.8 + -1 = 0.62$

© 2018 Peter V. Henstock

## Time Series Review

- Random Process

- $w() = 0, 2, 0, -1, -2, -1, 2, 1, 0, 0, 1, 2, -1, -2$

- MA Model (clearer notation)

- $x(k) = w(k) + \theta_1 w(k-1) + \theta_2 w(k-2) + \dots \theta_p w(k-p)$

- Let's use a 2nd order model  $\theta_1 = 0.7, \theta_2 = 0.3$

- Assume  $x(-1) = 0$

- $x(0) = 0$

- $x(1) = 2 + 0.7(0) = 2$

- $x(2) = 0 + 0.7(2) + 0.3(0) = 1.4$

- $x(3) = -1 + 0.7(0) + 0.3(2) = -0.4$

© 2018 Peter V. Henstock



## Ergodic Process

- Time estimate converges to the true estimate as  $N \rightarrow \text{infinity}$ 
  - Ergodic mean
    - True if  $1/N \sum x(k) = E[x(k)] = \mu$  as  $N \rightarrow \text{infinity}$
- If we take a reasonably long sample,
  - Then we can infer the statistical properties of the whole sequence

© 2018 Peter V. Henstock

## Strict Stationary

- If joint distribution of  $N$  observations is invariant to time shifts
- $F_{x(k_1)}(x_1) = F_{x(k_1+T)}(x_1)$ 
  - All have same distribution regardless of  $T$

© 2018 Peter V. Henstock



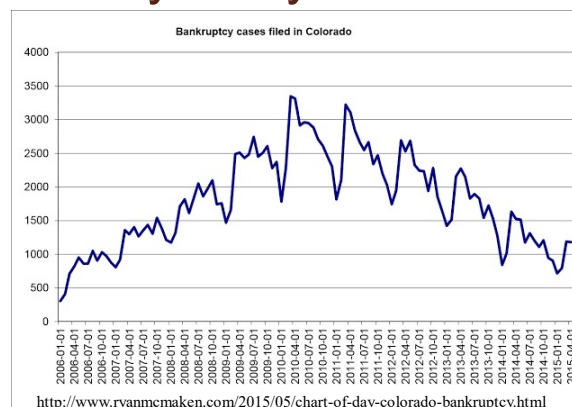
## (Wide Sense or Weak) Stationarity

- Few real systems are strictly stationary
  - Joint distribution of all vectors of  $d$  dimensions remains the same for any fixed  $d$
- Weaker definition is WSS
  - First order stationary:  $E[X(t)]$  is same for all  $t$
  - Second order:
    - 1) First order stationary and
    - 2)  $\text{Cov}[X(t), X(t-\text{lag})]$  is function of only lag

Note: Gaussian processes described by mean & cov

© 2018 Peter V. Henstock

## Exploratory analysis



Constant mean or trend?

Constant variance?

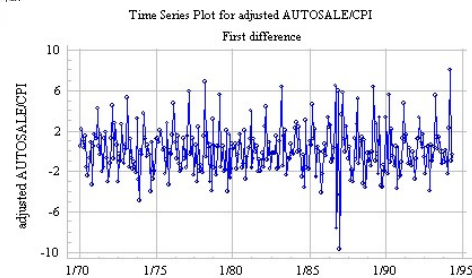
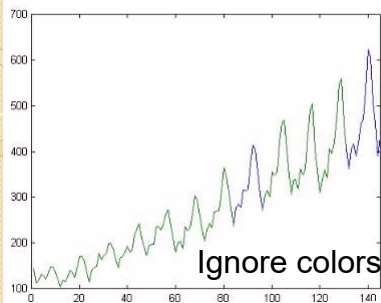
Seasonality?

Outliers?

Cycles?

© 2018 Peter V. Henstock

## Which is stationary? Which order?



<http://www.es.lancs.ac.uk/cres/captain/dhrdemo.html>  
<https://people.duke.edu/~rnau/411diff.htm>

© 2018 Peter V. Henstock

## Mathematical idea of stationary

$x(k) = A \cos(wk + u)$  where  $u$  is  $U(0, \pi)$

Is this stationary?

How would we determine the answer?

© 2018 Peter V. Henstock

## Mathematical idea of stationary

$x(k) = A \cos(wk + u)$  where  $u$  is  $U(0, \pi)$

Is this stationary?

How would we determine the answer?

$$E(x|k) = ?$$

© 2018 Peter V. Henstock

## Mathematical idea of stationary

$x(k) = A \cos(wk + u)$  where  $u$  is  $U(0, \pi)$

Is this stationary?

How would we determine the answer?

$$E(x|k) = \int_0^\pi [?] f(u) du \text{ where } f(u) = ?$$

© 2018 Peter V. Henstock

## Mathematical idea of stationary

$x(k) = A \cos(wk + u)$  where  $u$  is  $U(0, \pi)$

Is this stationary?

How would we determine the answer?

$$E(x|k) = \int_0^\pi A \cos(wk + u) f(u) du$$

$$E(x|k) = A \int_0^\pi \cos(wk + u) \frac{1}{\pi} du = \frac{-2}{\pi} \sin(wk)$$

So what can you conclude?

© 2018 Peter V. Henstock

## Mathematical idea of stationary

$x(k) = A \cos(wk + u)$  where  $u$  is  $U(0, \pi)$

Is this stationary?

How would we determine the answer?

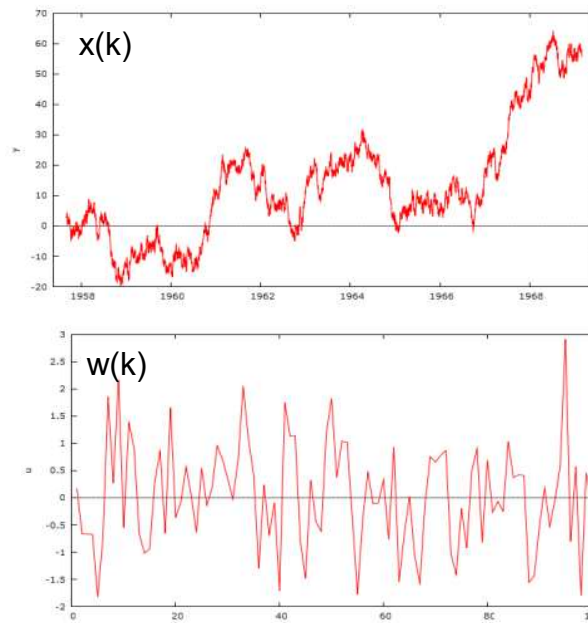
$$E(x|k) = \int_0^\pi A \cos(wk + u) f(u) du$$

$$E(x|k) = \frac{A}{\pi} \int_0^\pi \cos(wk + u) 1 du = \frac{-2A}{\pi} \sin(wk)$$

$E(x|k)$  depends on  $k$  so not stationary

© 2018 Peter V. Henstock

## Random Walk



© 2018 Peter V. Henstock