Introduction to probability and statistics

Alireza Fotuhi Siahpirani & Brittany Baur

sroy@biostat.wisc.edu

Computational Network Biology

Biostatistics & Medical Informatics 826

https://compnetbiocourse.discovery.wisc.edu

Sep 13th 2016

Some of the material covered in this lecture is adapted from BMI 576

Goals for today

- Probability primer
- Introduction to linear regression

A few key concepts

- Sample spaces
- Random variables
- Discrete and continuous continuous distributions
- Joint, conditional and marginal distributions
- Statistical independence

Definition of probability

- Intuitively, we use “probability” to refer to our degree of confidence in an event of an uncertain nature.

- Always a number in the interval [0,1]

  0 means “never occurs”

  1 means “always occurs”

Sample space

- *Sample space*: a set of possible outcomes for some experiment
- Examples
  - Flight to Chicago: {on time, late}
  - Lottery: {ticket 1 wins, ticket 2 wins,…,ticket *n* wins}
  - Weather tomorrow:

    {rain, not rain} or

    {sun, rain, snow} or

    {sun, clouds, rain, snow, sleet}
  - Roll of a die: {1,2,3,4,5,6}
  - Coin toss: {Heads, Tail}

Random variables

- *Random variable*: A variable that represents the outcome of a uncertain experiment
- A random variable can be
  - Discrete/Categorical: Outcomes take a fixed set of values
    - Roll of die, flight to Chicago, weather tomorrow
  - Continuous: Outcomes take continuous values
    - Height, weight

Notation

- Uppercase letters and words denote random variables
  - X, Y
- Lowercase letters and words denote values
  - x, y
- Probability that X takes value x

$$P(X = x)$$

- We will also use the shorthand form

$P(x)$ for $P(X=x)$

- For Boolean random variables, we will use the shorthand

$P(fever)$ for $P(Fever = true)$

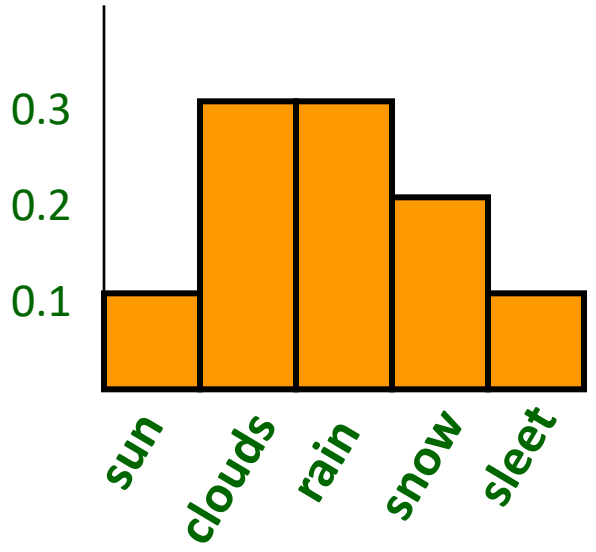
$P(\neg fever)$ for $P(Fever = false)$

Discrete probability distributions

- A probability distribution is a mathematical function that specifies the probability of each possible outcome of a random variable
- We denote this as $P(X)$ for random variable X
- It specifies the probability of each possible value of X, x
- Requirements:

$$P(x) \geq 0 \quad \text{for every } x$$

$$\sum_{x} P(x) = 1$$



Joint probability distributions

- *Joint probability distribution*: the function given by $P(X = x, Y = y)$
- Read as "X equals x and Y equals y"
- Example

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal probability distributions

- The *marginal distribution* of X is defined by

$$P(x) = \sum_{y} P(x,y)$$

"the distribution of X ignoring other variables"

- This definition generalizes to more than two variables, e.g.

$$P(x) = \sum_{y} \sum_{z} P(x,y,z)$$

Marginal distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for X

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional distributions

- The *conditional distribution* of X given Y is defined as:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

- Or in short

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

- The distribution of X given that we know the value of Y

- Intuitively, how much does knowing Y tell us about X?

Conditional distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$P(X = x \mid Y = \textit{on-time})$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

Independence

- Two random variables, X and Y, are *independent* if

$$P(x,y) = P(x) \times P(y) \quad \text{for all } x \text{ and } y$$

- Another way to think about this is knowing X does not tell us anything about Y

Independence example #1

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
on-time	0.45
late	0.55

Are X and Y independent here? NO.

Independence example #2

joint distribution

x, y	$P(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
fly-United	0.9
fly-Northwest	0.1

Are X and Y independent here?

YES.

Conditional independence

- Two random variables X and Y are *conditionally independent* given Z if

$$P(X \mid Y,Z) = P(X \mid Z)$$

"once you know the value of Z, knowing Y doesn't tell you anything about X"

- Alternatively

$$P(x,y \mid z) = P(x \mid z) \times P(y \mid z) \quad \text{for all } x,y,z$$

Conditional independence example

Flu	Fever	Headache	P
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Are Fever and Headache independent? NO.

e.g. $P(fever, headache) \neq P(fever) \times P(headache)$

Conditional independence example

Flu	Fever	Headache	P
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Are Fever and Headache conditionally independent given Flu: YES.

$$P(fever, headache \mid flu) = P(fever \mid flu) \times P(headache \mid flu)$$
$$P(fever, headache \mid \neg flu) = P(fever \mid \neg flu) \times P(headache \mid \neg flu)$$
etc.

Chain rule of probability

- For two variables

$$P(X,Y) = P(X \mid Y) \times P(Y)$$

- For three variables

$$P(X,Y,Z) = P(X \mid Y,Z) \times P(Y \mid Z) \times P(Z)$$

etc.

- To see that this is true, note that

$$P(X,Y,Z) = \frac{P(X,Y,Z)}{P(Y,Z)} \times \frac{P(Y,Z)}{P(Z)} \times P(Z)$$

Example discrete distributions

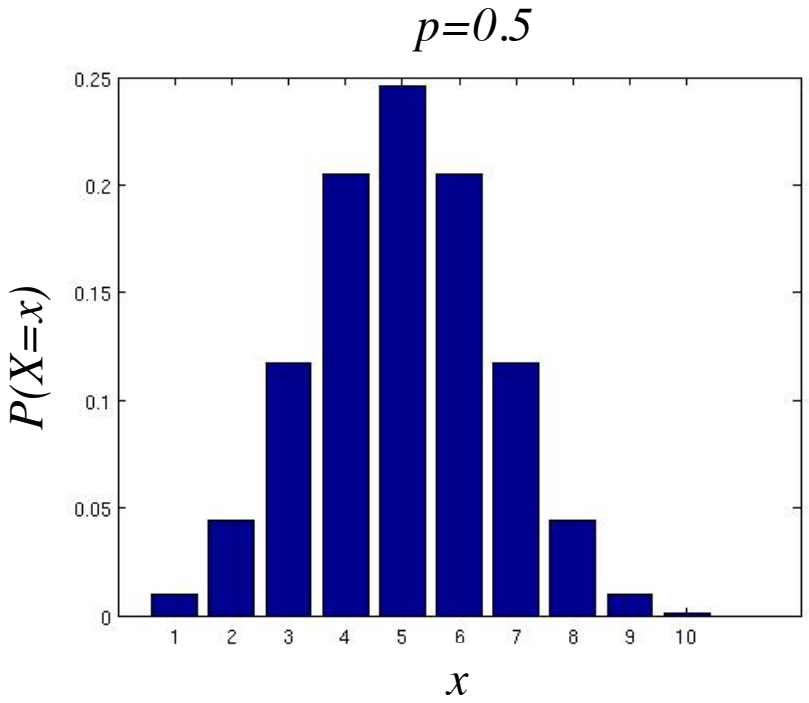
- Binomial distribution
- Multinomial distribution

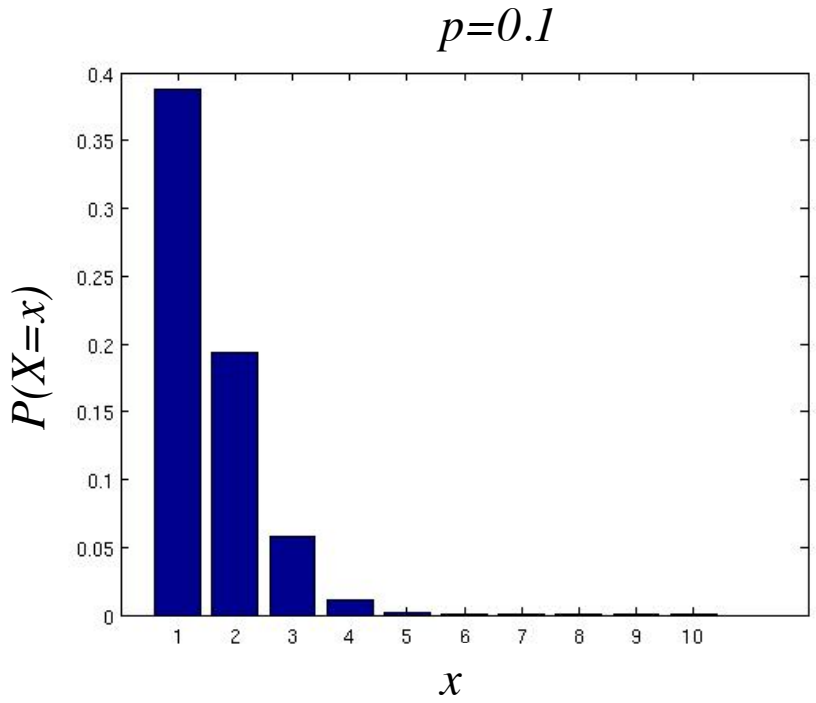
The binomial distribution

- Two outcomes per trial of an experiment
- Distribution over the number of successes in a fixed number n of independent trials (with same probability of success p in each)

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- e.g. the probability of x heads in n coin flips





The multinomial distribution

- A generalization of the binomial distribution to more than two outcomes
- Provides a distribution of the number of times any of the outcomes happen.
- For example consider rolling of a die n = 100 times. Each time we can have one of k = 6 outcomes, $\{1, .., 6\}$
- X_i is the variable representing the number of times the die landed on the i^{th} face, $i \in \{1, .., 6\}$
- p_i is the probability of the die landing on the i^{th} face

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

$$\sum_{i=1}^{k} x_i = n$$

Continuous random variables

- When our outcome is a continuous number we need a continuous random variable
- Examples: Weight, Height
- We specify a density function for random variable X as

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Probabilities are specified over an interval
- Probability of taking on a single value is 0.

Continuous random variables contd

- To define a probability distribution for a continuous variable, we need to integrate *f(x)*

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

$$P(b \leq X \leq a) = \int_{b}^{a} f(x)dx$$

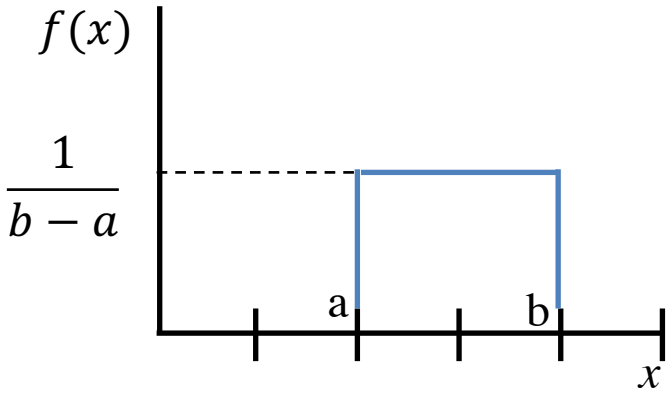
Example continuous distributions

- Uniform distribution
- Gaussian distribution
- Exponential distribution

Uniform distribution

- A variable X is said to have a uniform distribution, between $[a, b]$, where, $a < b$, if

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

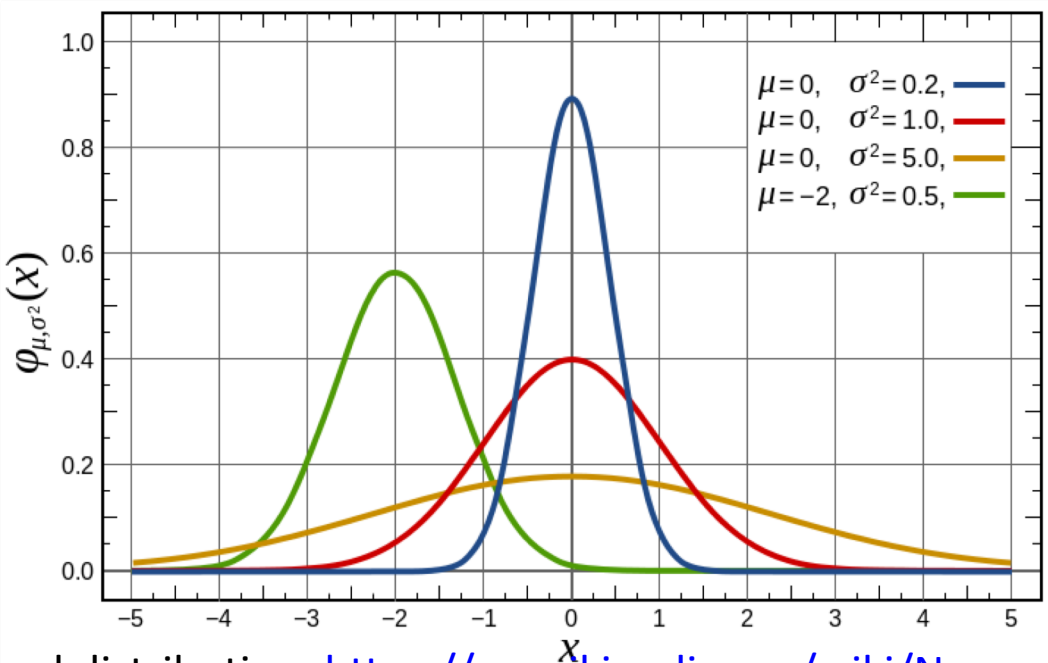


Adapted from Wikipedia

Gaussian Distribution

- The univariate Gaussian distribution is defined by two parameters, Mean: μ and Standard deviation: σ

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

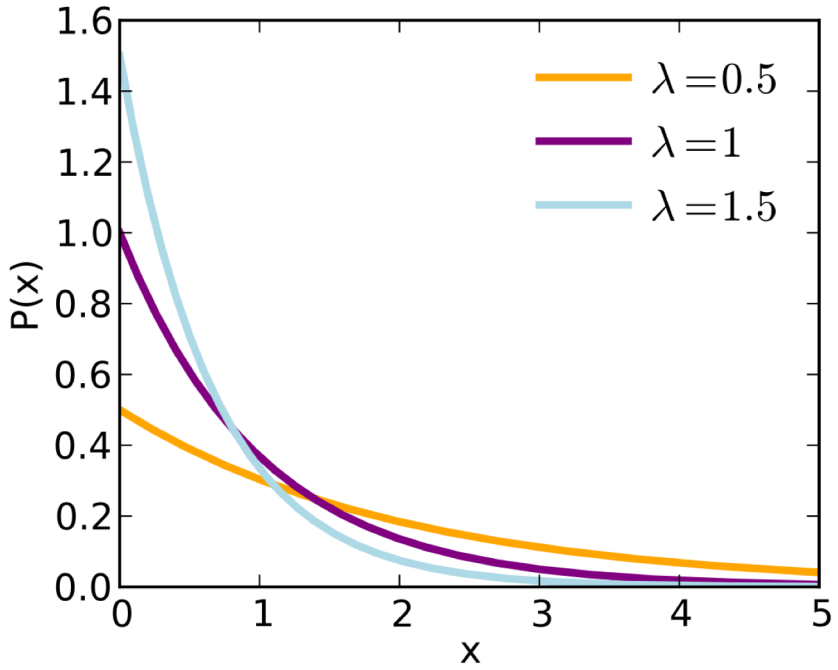


From Wikipedia: Normal distribution, https://en.wikipedia.org/wiki/Normal_distribution

Exponential Distribution

- Exponential distribution for a random variable X is defined as

$$f(x) = \lambda e^{-\lambda x}$$

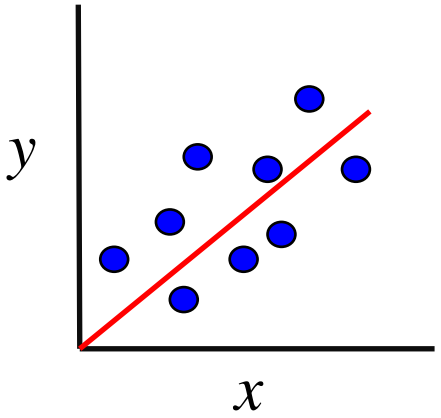


From Wikipedia: Exponential distribution, https://en.wikipedia.org/wiki/Exponential_distributio

Goals for today

- Probability primer
- **Introduction to linear regression**

Linear regression



$$\widehat{y} = \beta x + \beta_0$$

Slope Intercept

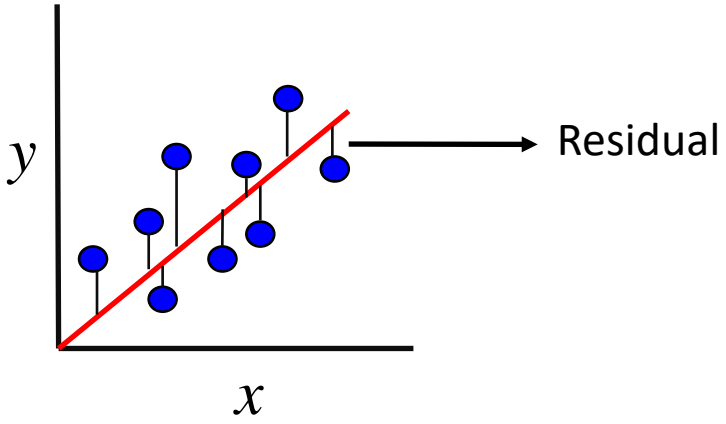
Linear regression assumes that output (y) is a linear function of the input (x)

Given:

Data= $\{(x_1, y_1), \ldots, (x_N, y_N)\}$

Estimate: $\beta = \{\beta_0, \beta_1\}$

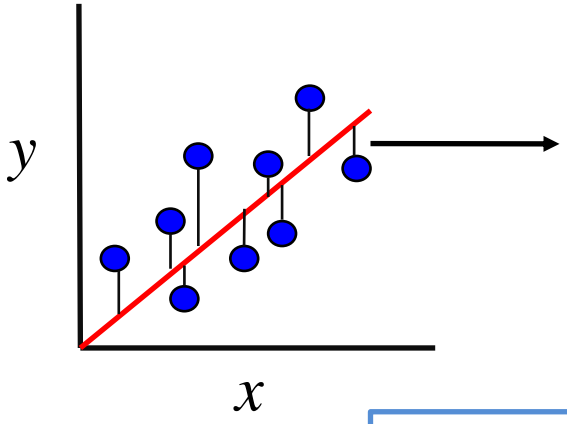
Residual Sum of Squares (RSS)



$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

To find the β, we need to minimize the Residual Sum of Squares

Minimizing RSS



$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial}{\partial \beta_0} RSS(\beta) = \sum_{i=1}^{N} -2(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_0} RSS(\beta) = 0 \quad \Longrightarrow \quad \beta_0 = \frac{\sum_{i=1}^{N}(y_i - \beta_1 x_i)}{N}$$

$$\frac{\partial}{\partial \beta_1} RSS(\beta) = \sum_{i=1}^{N} -2x_i(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_1} RSS(\beta) = 0 \quad \Longrightarrow \quad \beta_1 = \frac{\sum_{i=1}^{N} x_i(y_i - \beta_0)}{\sum_{i=1}^{N} x_i^2}$$

Linear regression with p inputs

- Y: output
- Inputs: $\{X_1, \cdots, X_p\}$

$$y = f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

intercept

Parameters/coefficients

Given: Data= $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$

Estimate: $\beta = \{\beta_0, \beta_1, \cdots, \beta_p\}$

Ordinary least squares for estimating β

- Pick the β that minimizes the residual sum of squares RSS

$$RSS(\beta) = \Sigma_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2$$

$$RSS(\beta) = \Sigma_{i=1}^{N}(y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij}\beta_j)^2$$

How to minimize RSS?

- Easier to think in matrix form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$Y = \mathbf{X}\beta$$

$$RSS(\beta) = (Y - \mathbf{X}\beta)^T (Y - \mathbf{X}\beta)$$

This is the square of y-Xb in matrix world

Simple matrix calculus

1. $$\partial(X^T) = (\partial X)^T$$
2. $$\partial(X + Y) = \partial X + \partial Y$$
3. $$\partial(XY) = (\partial X)Y + X(\partial Y)$$

The Matrix Cookbook
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Estimating β by minimizing RSS

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T(Y - \mathbf{X}\beta)$$

$$2\mathbf{X}^T(Y - \mathbf{X}\beta) = 0$$

$$\mathbf{X}^T Y - \mathbf{X}^T\mathbf{X}\beta = 0$$

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T Y$$

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

Works well when $(X^TX)^{-1}$ is invertible. But often this is not true. Need to regularize or add a prior

References

- Chapter 2, Probabilistic Graphical Models. Principles and Techniques. Friedman & Koller.
- Slides adapted from Prof. Mark Craven's Introduction to Bioinformatics lectures.
- All of Statistics, Larry Wasserman.
- Chapter 3, The Elements of Statistical Learning, Hastie, Tibshirani, Friedman.