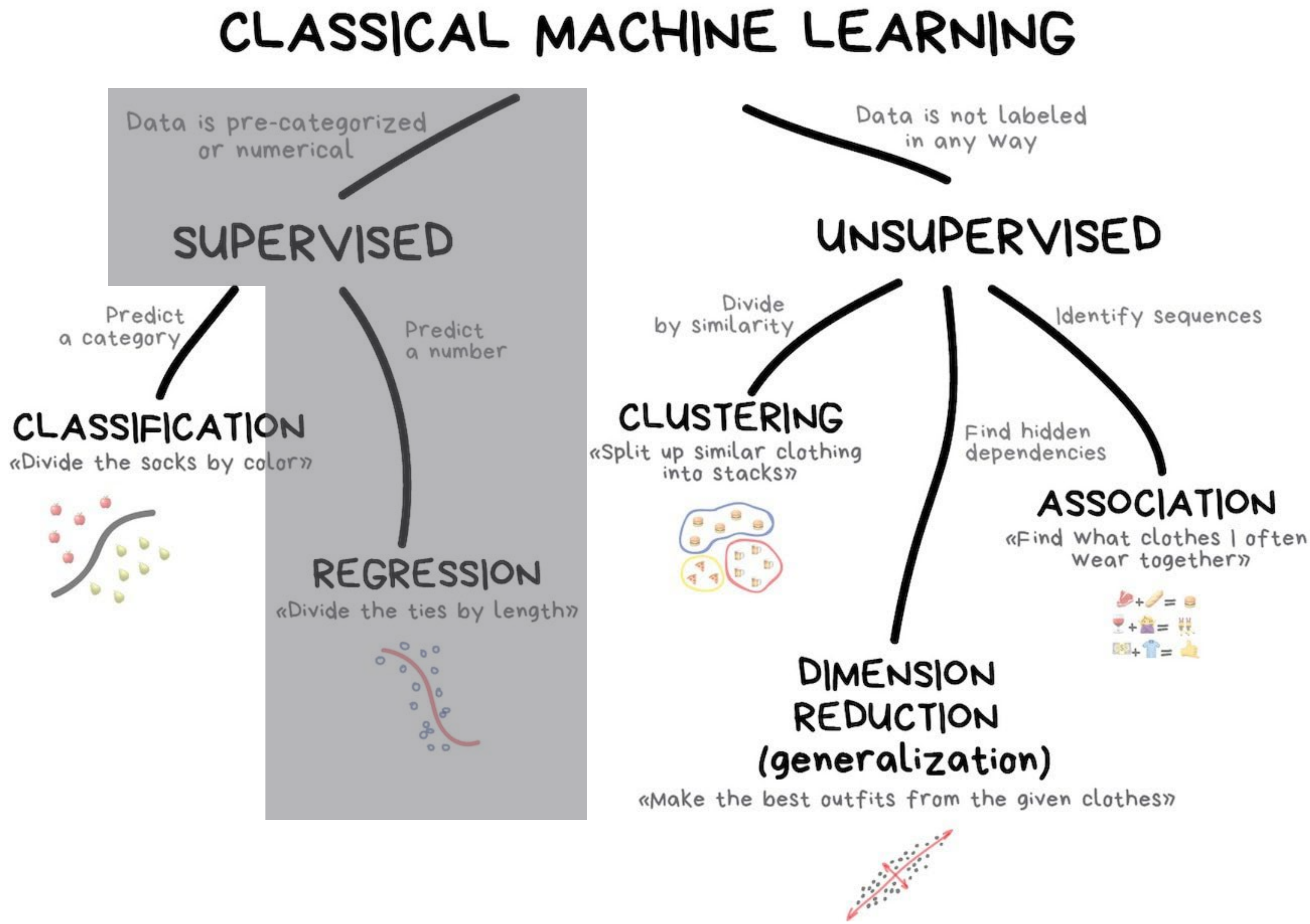
LINEAR REGRESSION

Week03

Topics Covered in This Class

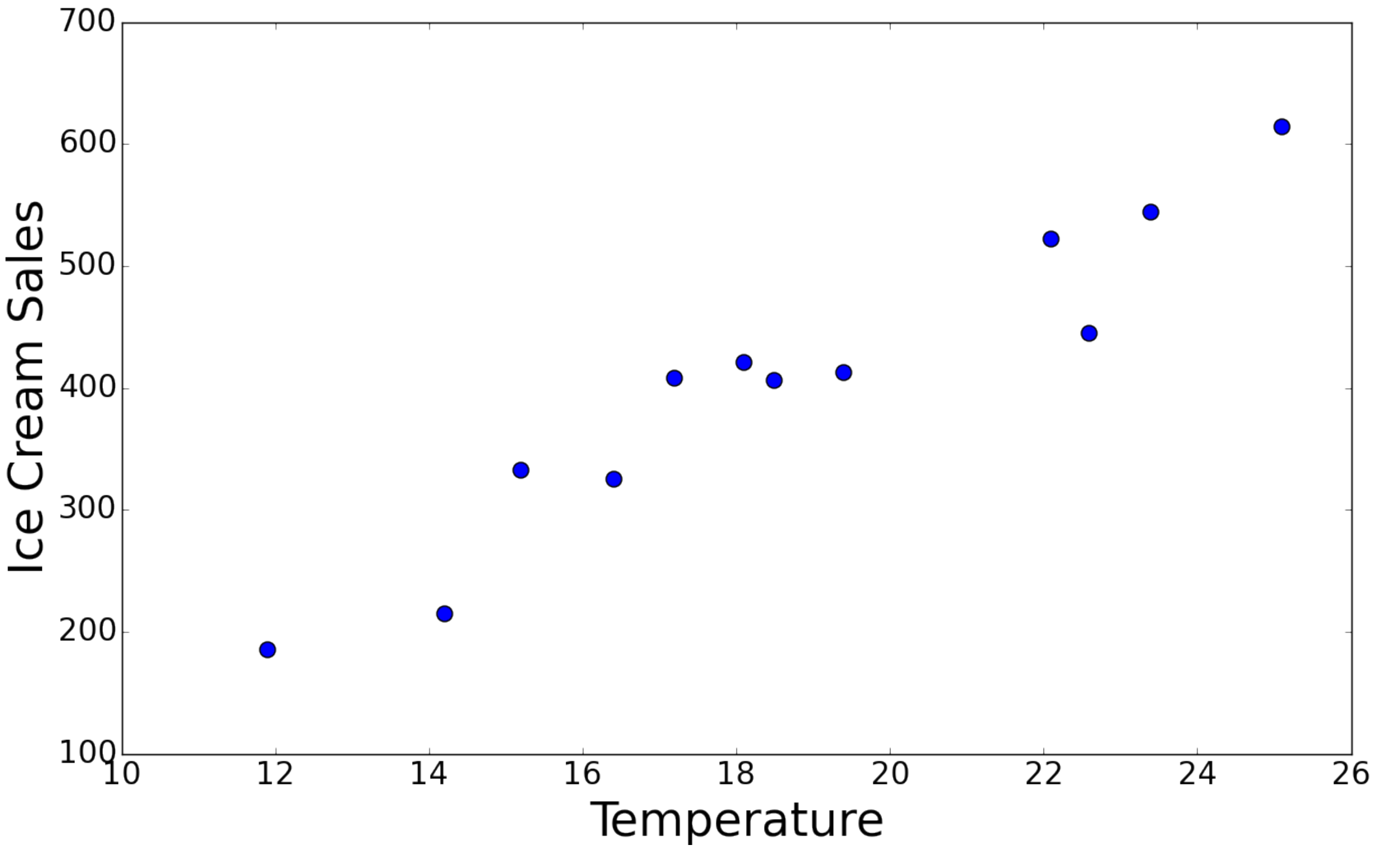


Linear Regression

Supervised Learning: Regression

- Prediction ice cream sales over given temperature

$$ice\ cream\ sales = f(temperatue)$$



Temperature (℃)	Ice Cream Sales ($)
14.2	215
16.4	325
11.9	185
15.2	332
18.5	406
22.1	522
19.4	412
25.1	614
23.4	544
18.1	421
22.6	445
17.2	408

Linear Regression

- Linear regression
  - Based on the assumption that the relationship between a scalar dependent variable y and explanatory(independent) variables X is linear
  - $X = [x_1, x_2, x_3, \dots, x_n]$
    Explanatory variables: print run(x_1), page number(x_2)

x_1	x_2
2800	22
2670	14
2800	37
2784	15
2800	38

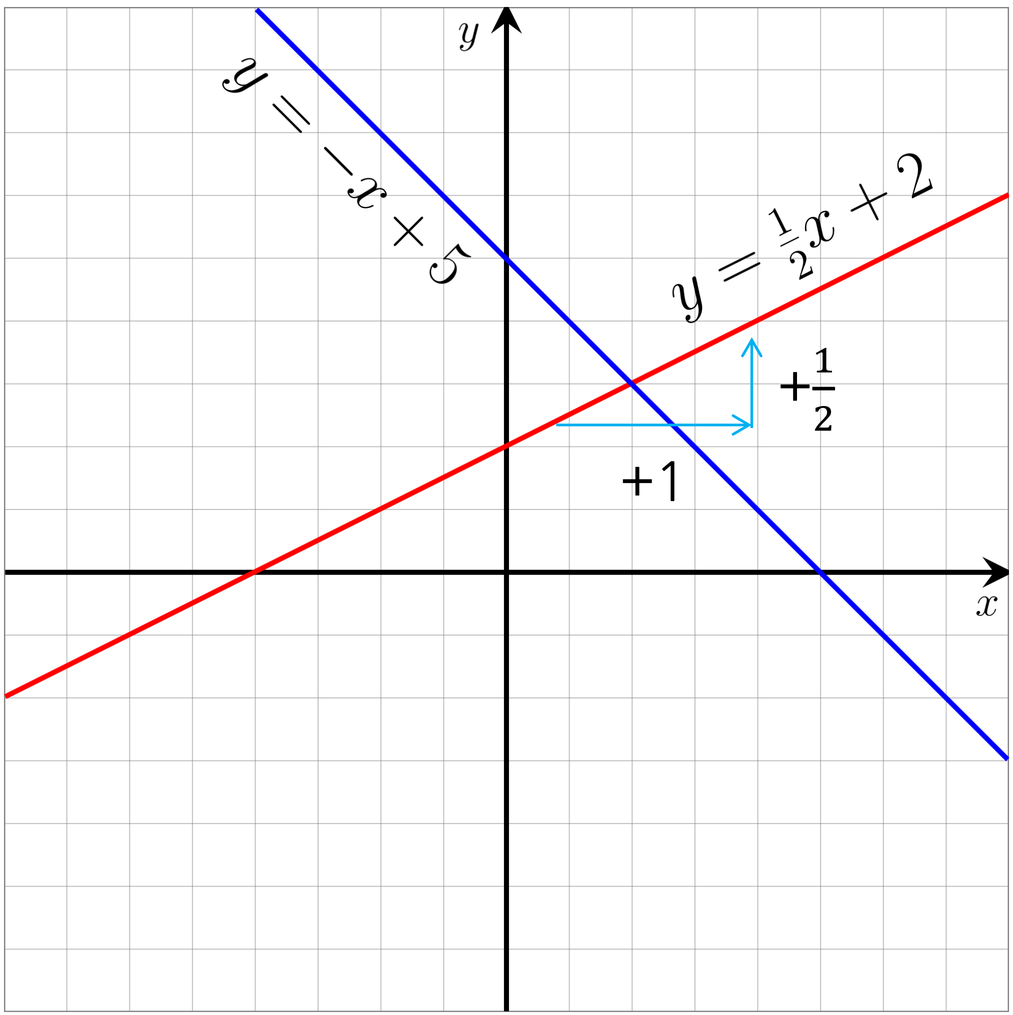
$\Rightarrow$ $\boldsymbol{X}$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

※ Linear function

- You studied linear function when you are high school student!



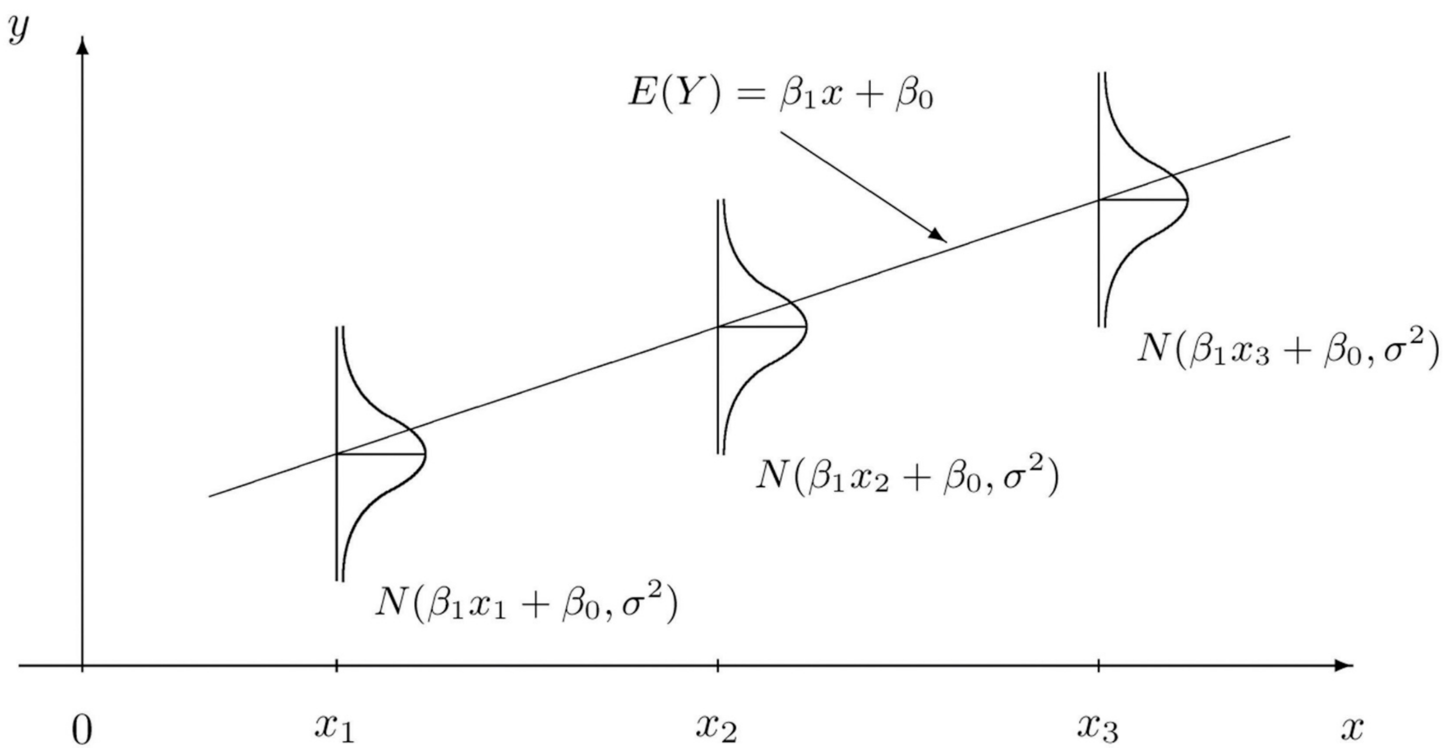
$$y = \frac{1}{2}x + 5$$

slope → coefficient

intercept

Main Assumptions of Linear Regression

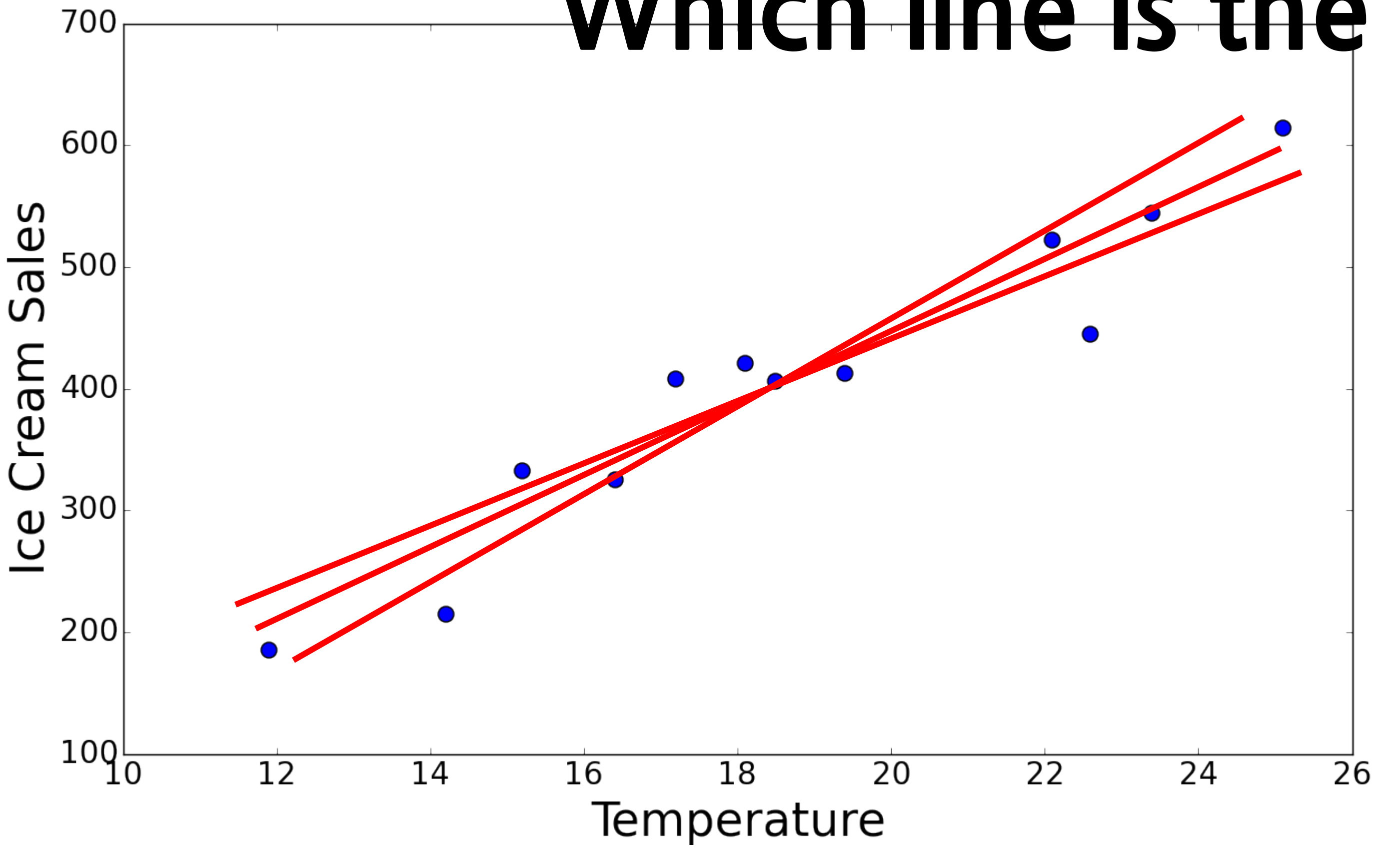
- Linear regression analysis makes several key assumptions
  - Linear relationship
  - Homoscedasticity
  - Normality
  - No or little multicollinearity



How to Determine Relationship between x and y

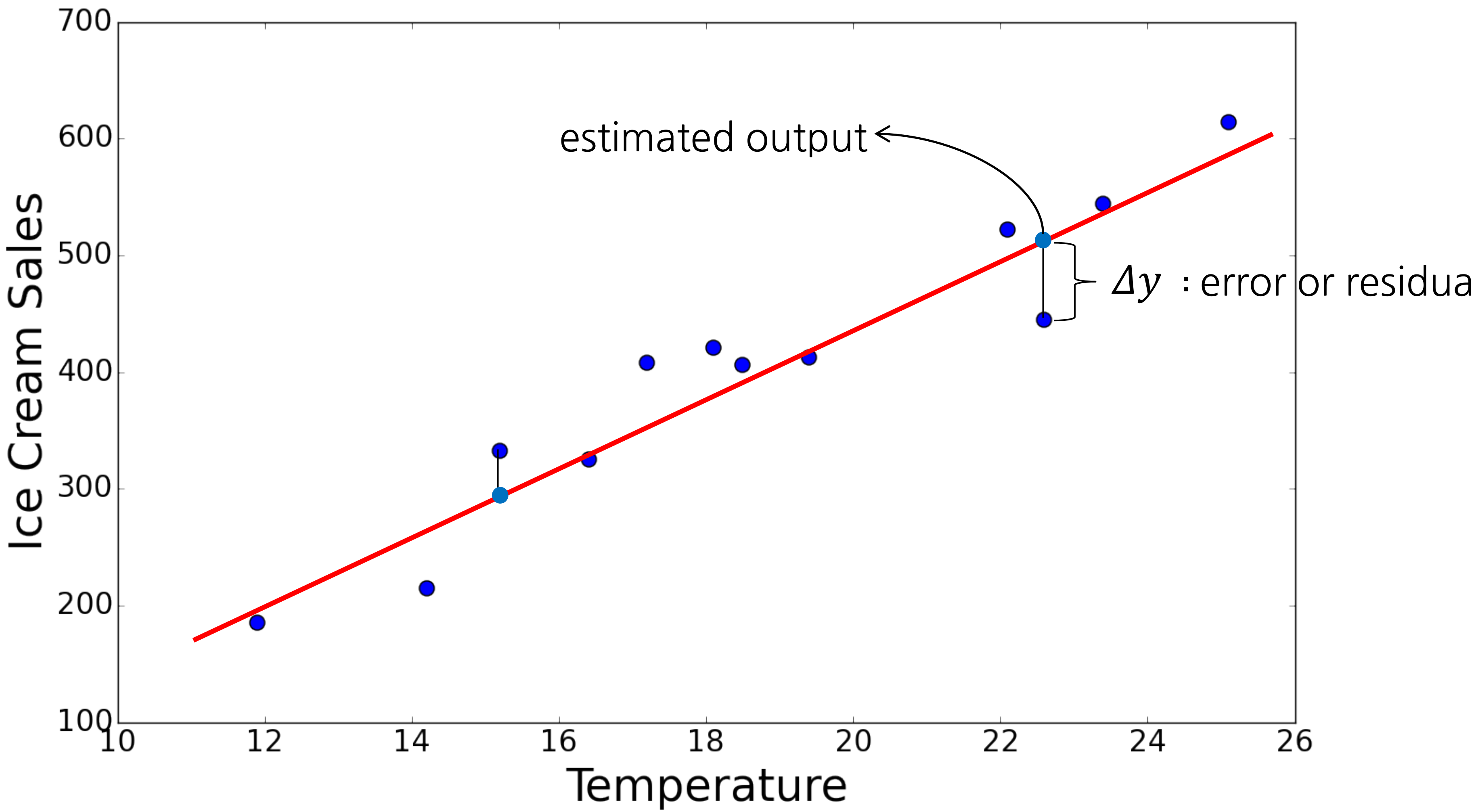
- Need a criterion

Which line is the best?



Least Square Method

- Minimize summation of squared error
  - Squared error=(estimated output-real output)^2=error^2



Least Square Method

- Minimize summation of squared error

estimated output

Sum for all data points in train set $\Leftarrow$ $$\sum_i (y_i - \widehat{y_i})^2$$

real output

- In the simple case: Only one independent variable
  - Estimated output $\widehat{y_i} = \beta_0 + \beta_1 x_i$

Unknown values

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Known values

※ Summary for Notation

- Hat, ($\hat{}$)
  - Represents estimation
  - β_1 is unknown true value, $\widehat{\beta_1}$ is estimation for β_1 through model learning
  - y_i is known output value of $i-\mathrm{th}$ sample, $\hat{y}_i$ is estimated output by learned model

- Bar, ($\bar{}$)
  - Represents sample mean
  - Arithmetic average of the observed values of variable

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - If the number of input variables is more than one, elements of sample mean vector consist of average of each variable

$$\bar{\mathbf{x}} = \left(\frac{\sum_{i=1}^{n} x_{1i}}{n}, \frac{\sum_{i=1}^{n} x_{2i}}{n}, \cdots, \frac{\sum_{i=1}^{n} x_{pi}}{n}\right) = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}$$

- Bold character usually represents vector

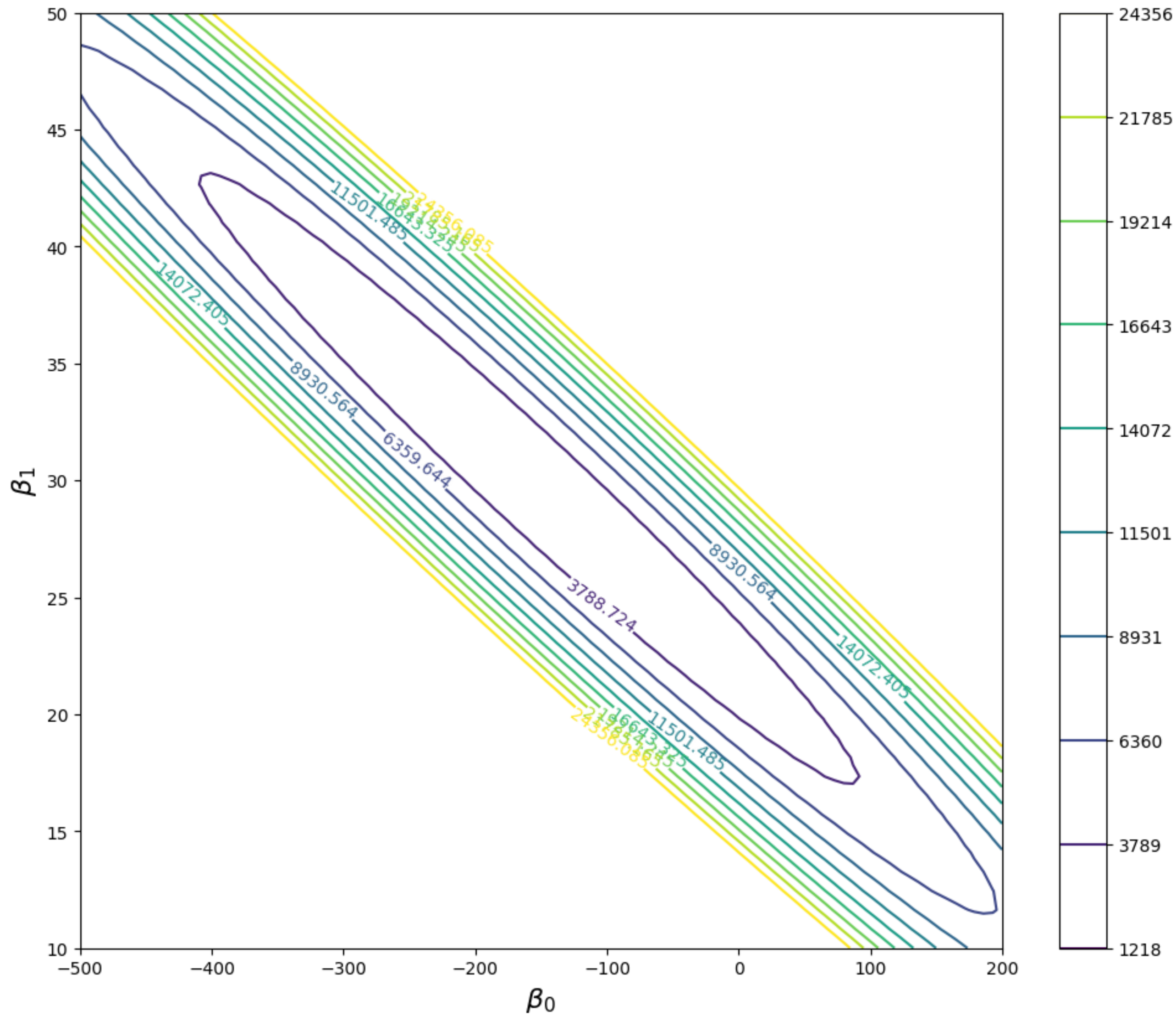
Least Square Method

- Which one is better?

Temperature (℃)	Ice Cream Sales ($)	$\beta_1 = 20, \beta_0 =$-80		$\beta_1 = 30, \beta_0 = -160$	
		Estimated Sales	Squared error	Estimated Sales	Squared error
14.2	215	204	121	266	2601
16.4	325	248	5929	332	49
11.9	185	158	729	197	144
15.2	332	224	11664	296	1296
18.5	406	290	13456	395	121
22.1	522	362	25600	503	361
19.4	412	308	10816	422	100
25.1	614	422	36864	593	441
23.4	544	388	24336	542	4
18.1	421	282	19321	383	1444
22.6	445	372	5329	518	5329
17.2	408	264	20736	356	2704
	sum		174901		14594

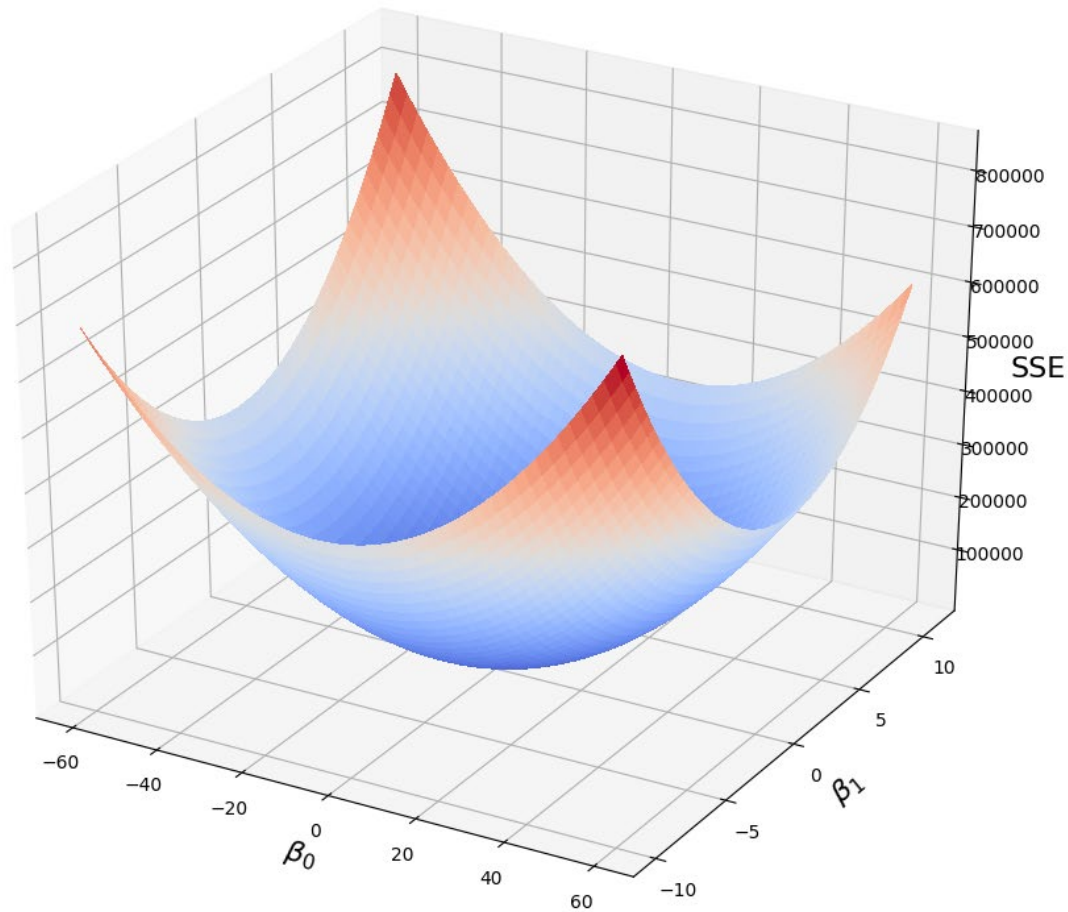
Least Square Method

- Summation of squared error with different β_0 and β_1



Least Square Method

- Summation of squared error with different β_0 and β_1 for simulated data from $y = x + 1$



Optimization for Linear Regression

- Variables to be determined

$$\beta_0, \beta_1$$

- Objective function

$$\min f(\beta_0, \beta_1) = \min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- Constraints
  - No constraint

Optimization for Linear Regression

- Solution
  - Calculate partial derivatives with respect to β_0, β_1

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

  - Solve linear equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Multiple Input Variables

- More than one input variable
  - Want to predict consumption of petrol

Petrol Tax($)	Average Income ($)	Paved Highways (miles)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	1976	0.525	541
9	4092	1250	0.572	524
9	3865	1586	0.58	561
7.5	4870	2351	0.529	414
8	4399	431	0.544	410
10	5342	1333	0.571	457
8	5319	11868	0.451	344
8	5126	2138	0.553	467
8	4447	8577	0.529	464
7	4512	8507	0.552	498
…	…	…	…	…

Multiple Input Variables

- Estimation based on petrol tax, average income, length of paved highways, proportion of population with driver's license

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

  - y =consumption of petrol
  - x_1=petrol tax
  - x_2=average income
  - x_3=length of paved highways
  - x_4=proportion of population with driver's license
  - ϵ is random error which follows Gaussian distribution with 0 mean, σ^2 variance

➡

$$\min \sum_i (y_i - \hat{y_i})^2$$

Same as the simple case!

Optimization for Linear Regression: Multivariate

- Multivariate linear regression

$$\min f(\beta_0, \dots, \beta_p) = \min \sum_i \left(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right)^2$$

  - Estimated parameters are obtained by setting partial derivatives zero

$$\frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}) = 0$$

$$\frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_1} = \sum_{i=1}^{n} -2x_{1i}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}) = 0$$

$$\vdots$$

$$\frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_p} = \sum_{i=1}^{n} -2x_{pi}(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}) = 0$$

Multiple Input Variables

- Matrix approach to multiple regression model

$$
\begin{aligned}
y_1 &= \beta_0 \cdot 1 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{k1} \\
y_2 &= \beta_0 \cdot 1 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{k2} \\
&\vdots \\
y_n &= \beta_0 \cdot 1 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_p x_{kn}
\end{aligned}
$$

n samples, p input variables

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}
$$

$$
\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

$$
= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \boldsymbol{I})
$$

$$
E = \left\| \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right\|^2
$$

Optimization for Linear Regression: Multivariate

$$\min E = \min \|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\|^2$$

- Solution is obtained by setting $\frac{\partial E}{\partial \boldsymbol{\beta}} = 0$

$$\frac{\partial (\mathbf{x} - A\mathbf{s})^T W (\mathbf{x} - A\mathbf{s})}{\partial \mathbf{s}} = -2A^T W (\mathbf{x} - A\mathbf{s})$$

$$\Downarrow$$

$$\frac{\partial (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathrm{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T (\mathbf{y} - \mathrm{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$\Downarrow$$

$$X^T \mathrm{X}\boldsymbol{\beta} - X^T \mathbf{y} = 0$$

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

Estimation of Regression Coefficients: Multivariate

- Use least square methods as same as simple linear regression

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

  - $\mathbf{X}^{\mathbf{T}}$: transpose matrix of $\mathbf{X}$
  - $\mathbf{X}^{-\mathbf{1}}$: inverse matrix of $\mathbf{X}$

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

  - Residual(error) terms

$$\boldsymbol{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = [e_1\ e_2 \cdots\ e_n]^T$$

  - SSE

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 = \boldsymbol{e}^T\boldsymbol{e}$$

Is the Regression Model Significant?

- Modeling learning is not the end of the analysis
  - Check overall significance in regression models
    - Whether the regression model is overall significant for predicting a target

  - Check significance of regression coefficients
    - Whether the specific variable is significant for predicting a target

- In the case of simple linear regression, testing overall significance of the model is the same as testing significance of regression coefficients
  - Because only one explanatory variable is used

Test of Model Significance

- F-test for general regression models
  - Hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_1: \text{not all } \beta_i (i = 1,2,\cdots,p) \text{ equal zero}$$

  - Test statistic

$$F^* = MSR/MSE$$

    - F follows F-distribution with $(p, n-p-1)$ degree of freedom
  - Decision rule

$$\text{If } F^* \leq F(1-\alpha; p, n-p-1), \text{ conclude } H_0$$
$$\text{If } F^* > F(1-\alpha; p, n-p-1), \text{ conclude } H_1$$

    - α: significance level

※ Statistical Test

- A statistical test provides a mechanism for making quantitative decisions about a process or processes
  - The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process
  - The procedure is based on how likely it would be for a set of observations to occur if the null hypothesis were true

- Null hypothesis
  - A general statement or default position that there is no relationship between two measured phenomena, or no association among groups

- Alternative hypothesis
  - It is the hypothesis used in hypothesis testing that is contrary to the null hypothesis

※ Statistical Test

- Steps in testing for statistical significance

State the null hypothesis and alternative hypothesis

Select a probability of error level (alpha level)

Collect data

Select and compute the test statistics

Interpret the results

※ Statistical Test

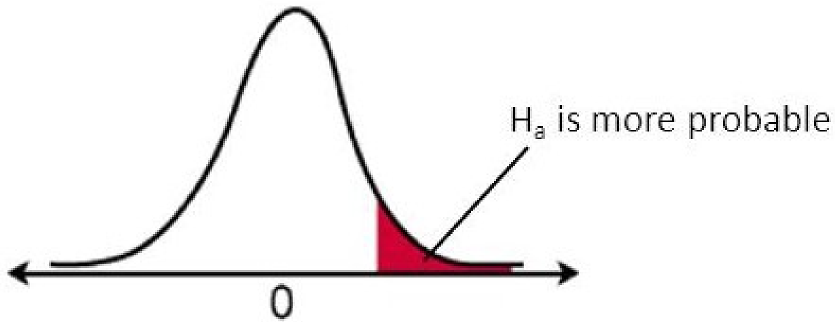
- Consider 20 first year resident female doctors drawn at random from one area
  - resting systolic blood pressures measured using an electronic sphygmomanometer
    - Sample mean = 130.05

  - Research hypothesis is that a resting systolic blood pressure of 120 mm Hg is predicted as the population mean
  - Null hypothesis and alternative hypothesis

$$H_0: \mu = 120$$
$$H_1: \mu \neq 120$$

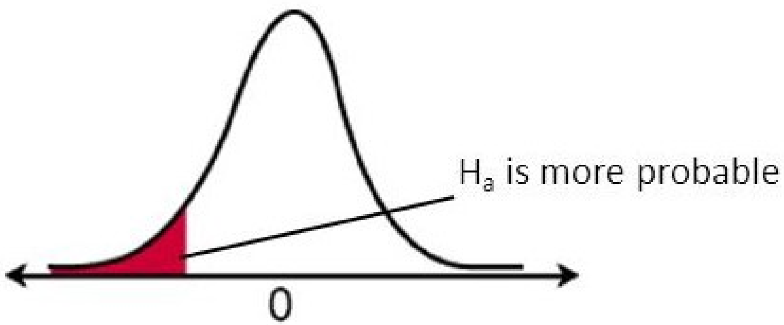
  - Set significance level as 0.05
  - Determine test statistics and underlying distribution
    - $t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}}$
    - t follows t -distribution with the degree of freedom as $n-1$

※ Statistical Test



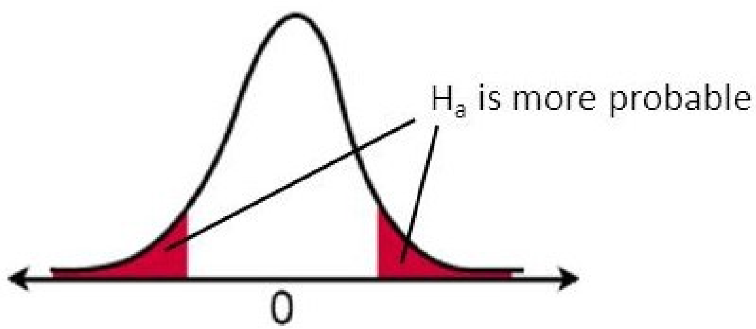
Right-tail test

$$H_a: \mu > \text{value}$$



Left-tail test

$$H_a: \mu < \text{value}$$

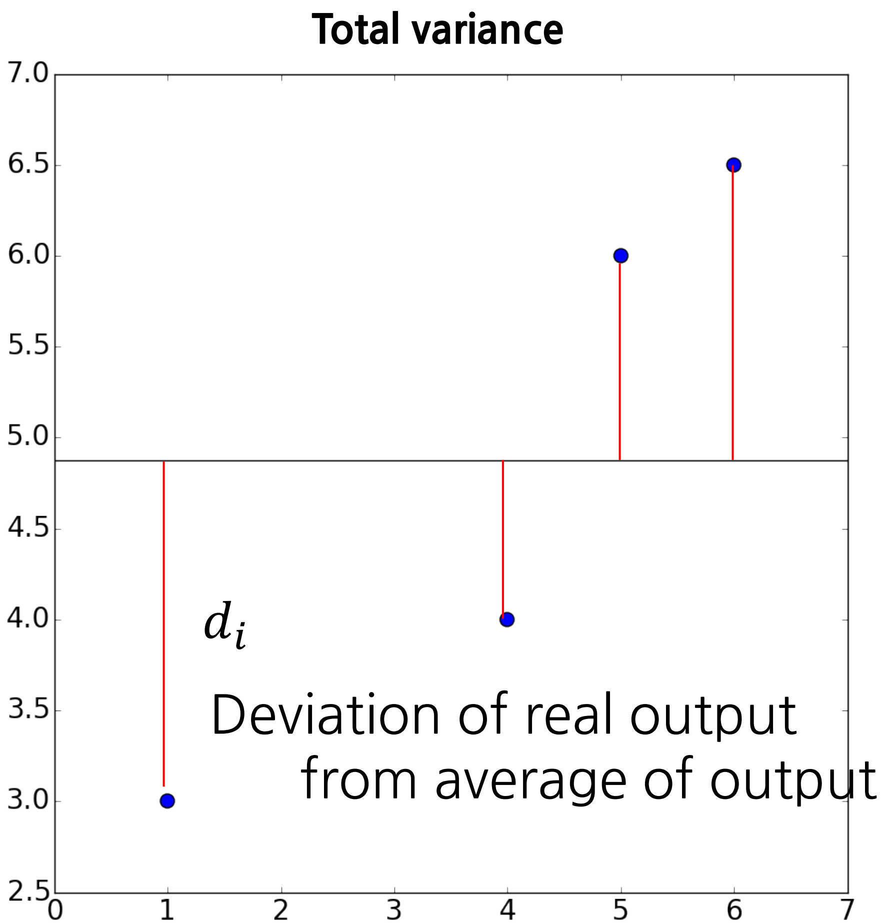


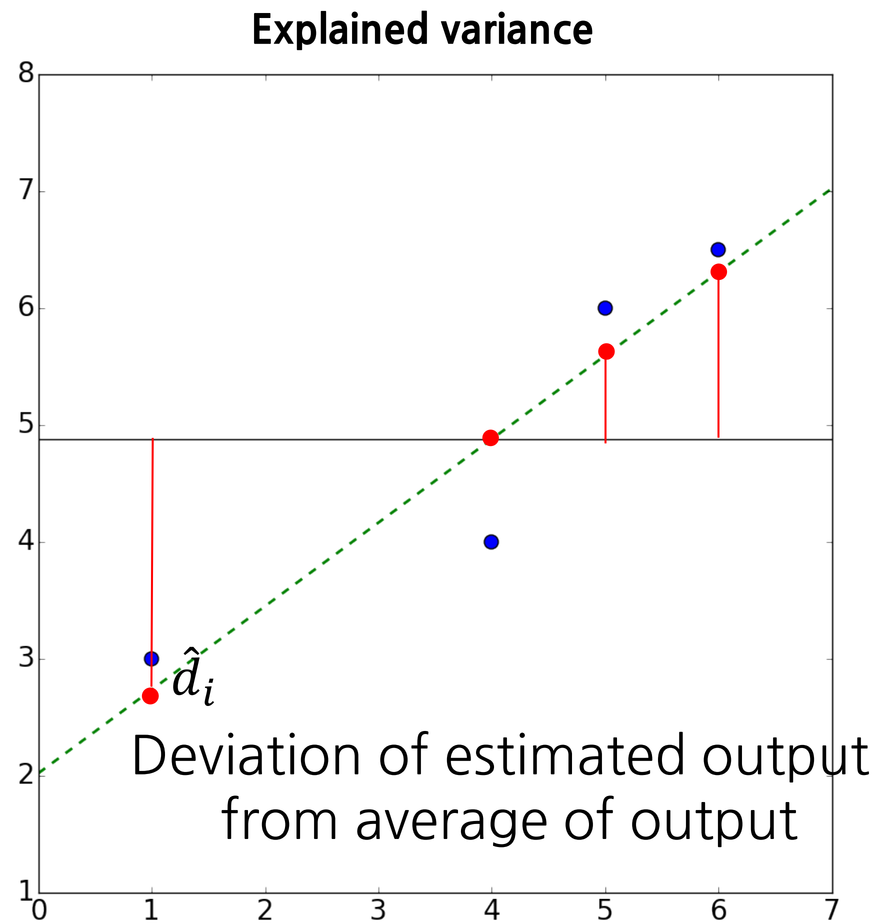
Two-tail test

$$H_a: \mu \neq \text{value}$$

Sum of Squares

- Total variance(SST) and Explained variance(SSR)





$\sum_i d_i^2$

$\sum_i \hat{d}_i^2$

Sum of Square

- Total variance: the total sum of squares

$$SST = \sum_i (y_i - \bar{y})^2$$

- Explained variance: the regression sum of squares, also called the explained sum of squares

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

- Residual variance: the sum of squares errors, also called the residual sum of squares

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- Relationship among three values

$$SST = SSR + SSE$$

Test of Model Significance

- ANOVA table for multiple regression model with p input variables

Factor	Sum of square	Degree of freedom	Mean square	F-value	p-value
Model	SSR	p	$MSR = SSR/p$	$F_0 = \mathrm{MSR}/\mathrm{MSE}$	$P\{F_{p,n-p-1} > F_0\}$
Residual	SSE	$n-p-1$	$MSE = SSE/(n-p-1)$		
Total	SST	$n-1$			

  - Analysis of Variance (ANOVA)

Degree of Freedom

- The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary
  - The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it
- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

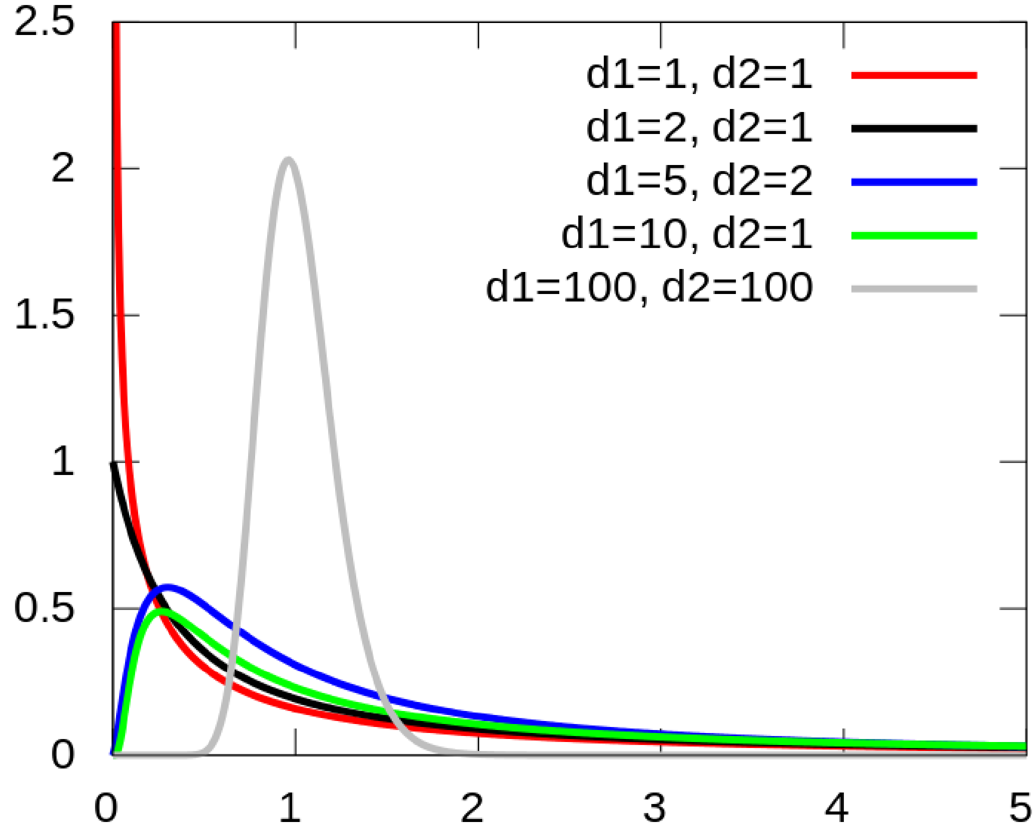
  - The reason that denominator is $n-1$ is that degree of freedom of sample mean, $\bar{y}$ is $n-1$
  - Another reason is that in the case of that denominator is $n-1$, S^2 is unbiased estimator of variance of population
- Mean squared error for simple linear regression

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - The reason that denominator is $n-2$ is that $\hat{y}_i$ is calculate from $\hat{\beta}_0 + \hat{\beta}_1 x_i$ and it depends on two estimators $\hat{\beta}_0, \hat{\beta}_1$ $\rightarrow$ Decrease two degrees of freedom

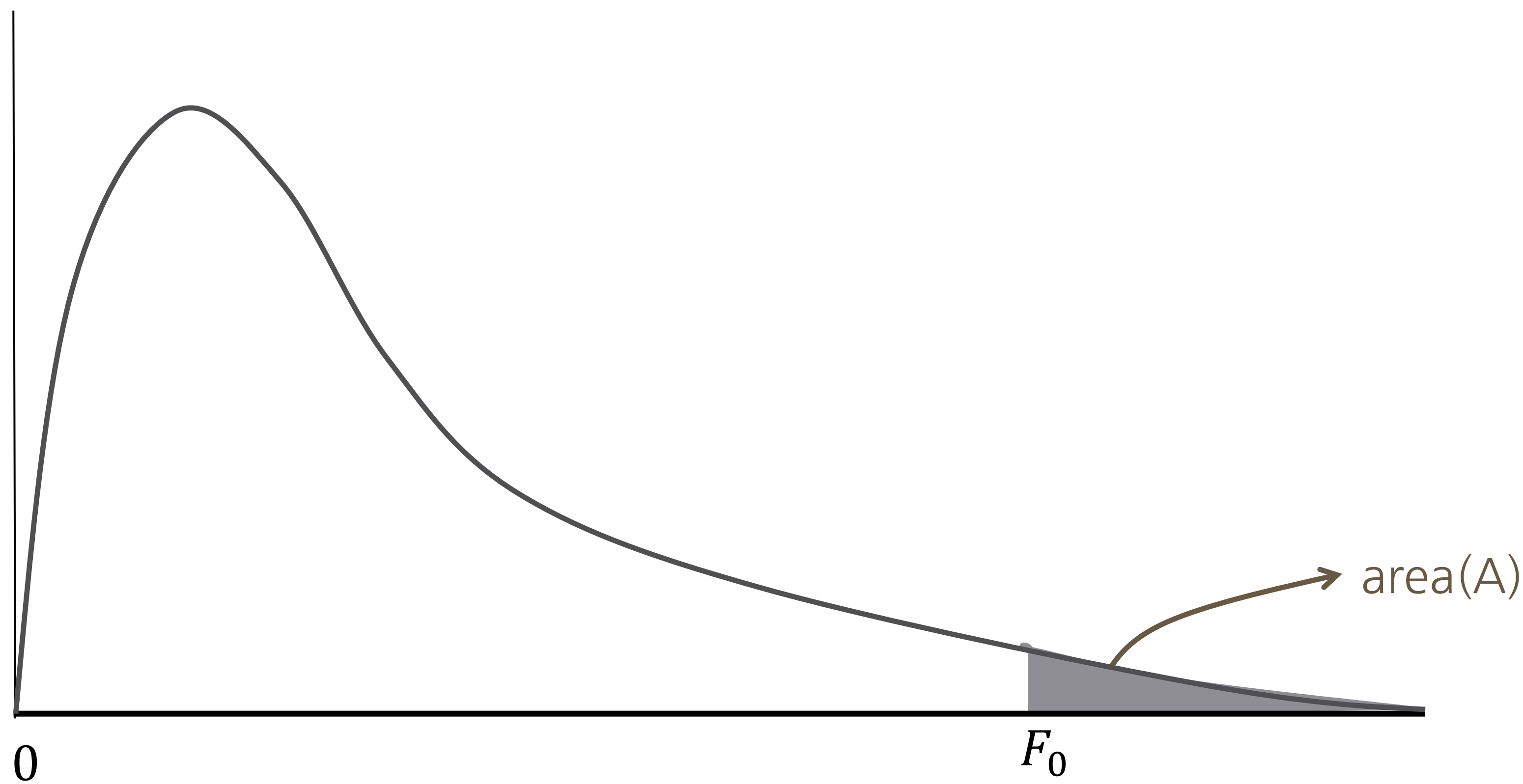
※ F distribution

- F statistics follows F distribution with $(p, n-p-1)$ degree of freedom
  - Probability density function of F distribution with different parameters
    - F distribution is determined by two parameters



Test of Model Significance

- If (area under density function from F_0 to ∞) 〈 α
  → Reject null hypothesis → not all $\beta_i (i = 1,2,\cdots,p)$ equal zero
  - α is significance level
  - significance level is usually set to 0.1, 0.05, or 0.01
    - The higher significance level, the level probability to reject null hypothesis



Python: Draw Figure

Matplotlib

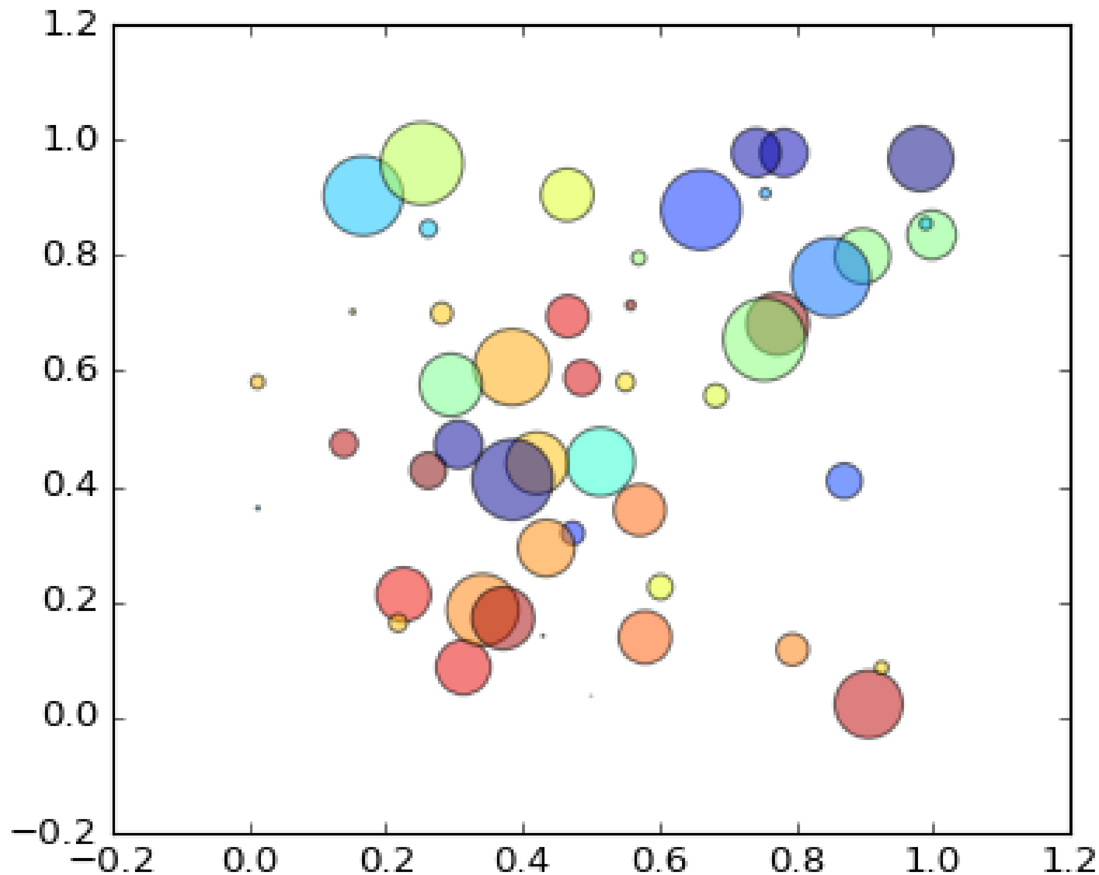
- matplotlib is a python 2D plotting library
  - url: http://matplotlib.org/index.html
  - import matplotlib

```
import matplotlib.pyplot as plt
```

- Scatter plot

```
import numpy as np
import matplotlib.pyplot as plt
N = 50
x = np.random.rand(N)
y = np.random.rand(N)
colors = np.random.rand(N)
area = np.pi * (15 * np.random.rand(N))**2  # 0 to 15 point radiuses

plt.scatter(x, y, s=area, c=colors, alpha=0.5)
plt.show()
```

Scatter plot

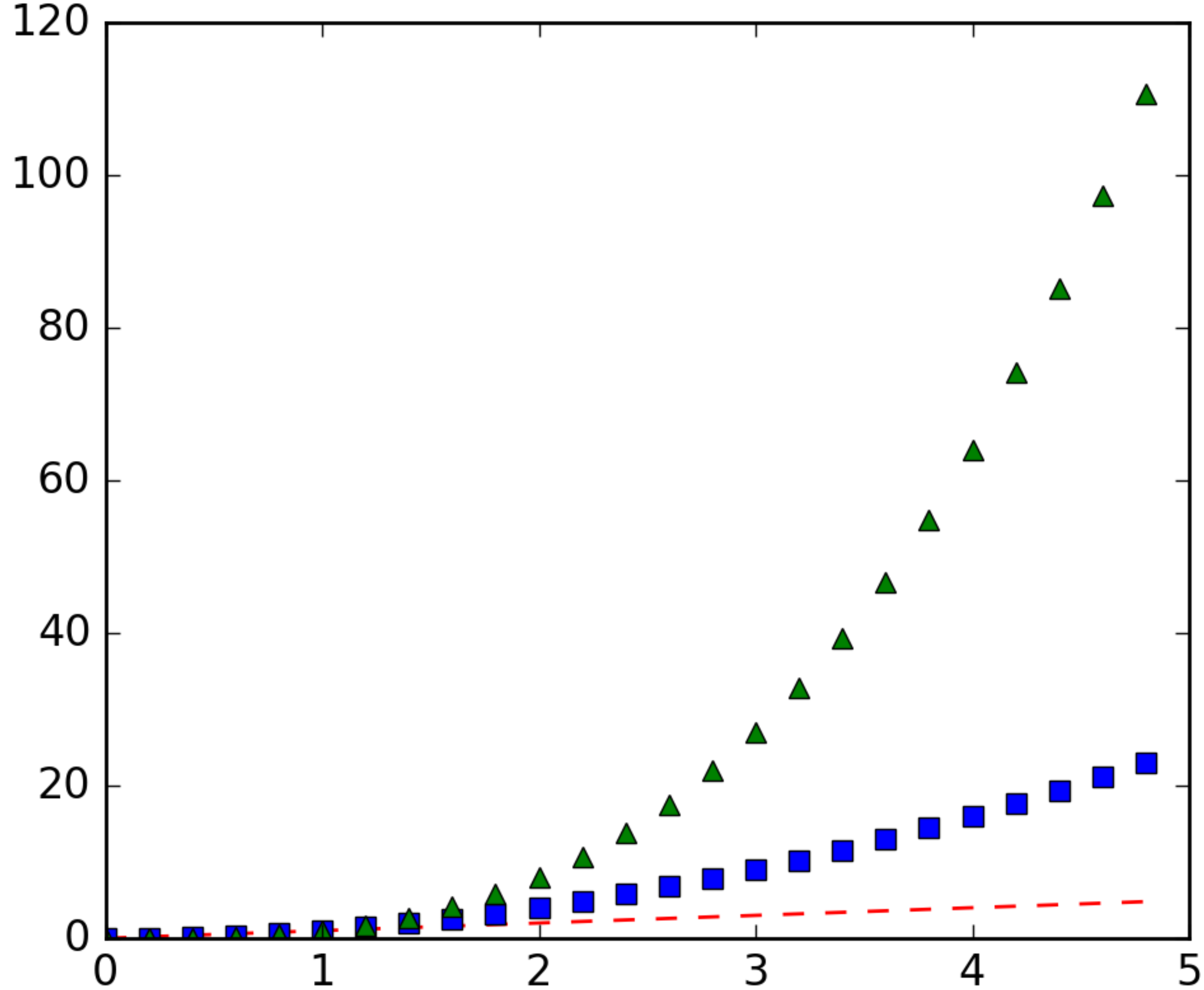


Line Plot

- Line plot

```
import numpy as np
import matplotlib.pyplot as plt

# evenly sampled time at 200ms intervals
t = np.arange(0., 5., 0.2)

# red dashes, blue squares and green triangles
plt.plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^')
plt.show()
```

Line Plot



Line Properties

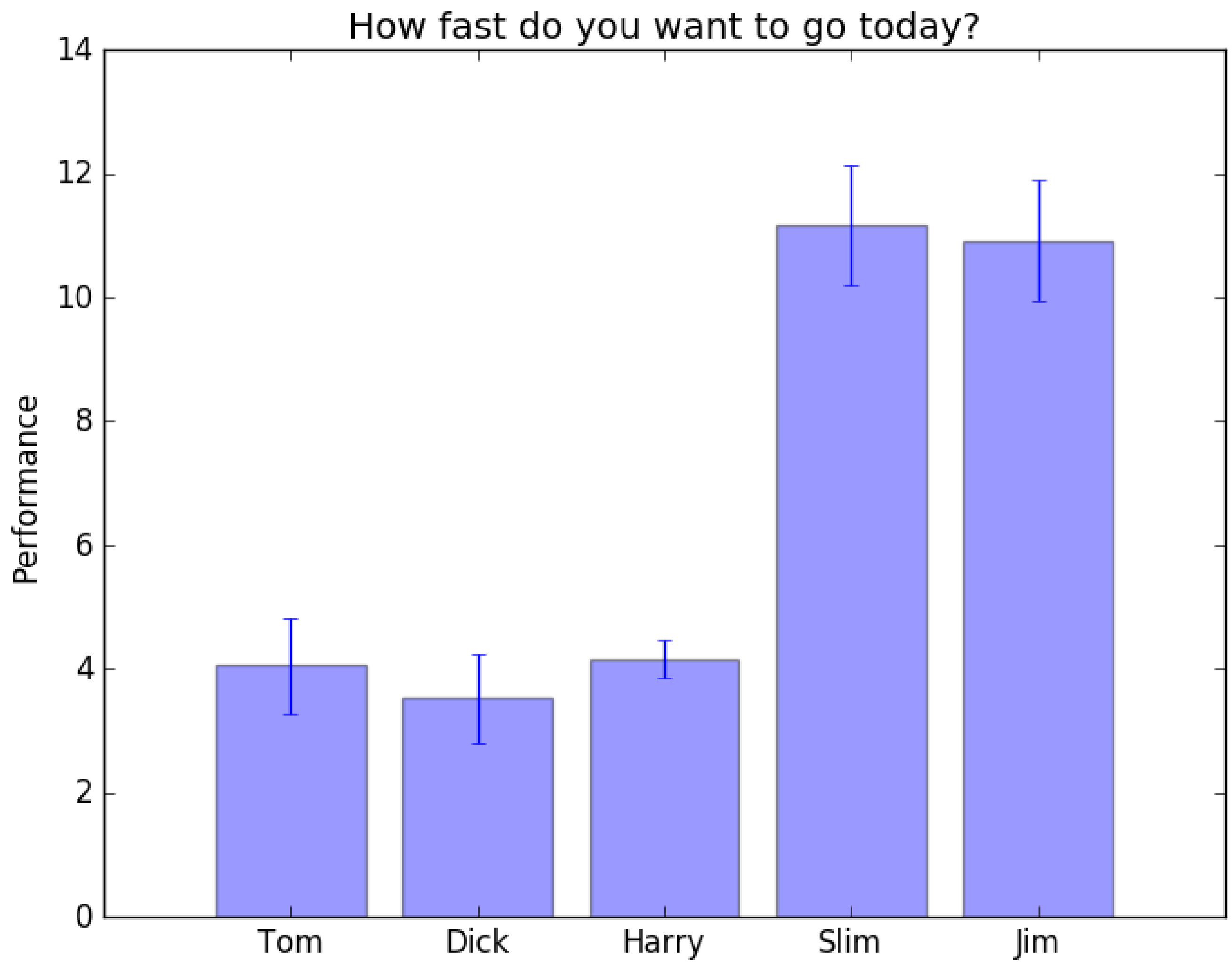
<table>
<tr><th>Property</th><th>Value Type</th></tr>
<tr><td>linestyle or ls</td><td>['-' | '--' | '-.' | ':' | 'steps' | ...]</td></tr>
<tr><td>linewidth or lw</td><td>float value in points</td></tr>
<tr><td>marker</td><td>['+' | ',' | '.' | '1' | '2' | '3' | '4']</td></tr>
<tr><td>markersize or ms</td><td>float</td></tr>
<tr><td>markeredgecolor or mec</td><td>any matplotlib color</td></tr>
<tr><td>markeredgewidth or mew</td><td>float value in points</td></tr>
<tr><td>markerfacecolor or mfc</td><td>any matplotlib color</td></tr>
<tr><td>alpha</td><td>float</td></tr>
</table>

Bar Chart

- Bar chart

```
import numpy as np
import matplotlib.pyplot as plt

# Example data
people = ('Tom', 'Dick', 'Harry', 'Slim', 'Jim')
y_pos = np.arange(len(people))
performance = 3 + 10 * np.random.rand(len(people))
error = np.random.rand(len(people))

plt.bar(y_pos, performance, yerr=error, align='center', alpha=0.4)
plt.xticks(y_pos, people)
plt.ylabel('Performance')
plt.title('How fast do you want to go today?')

plt.show()
```

Bar Chart

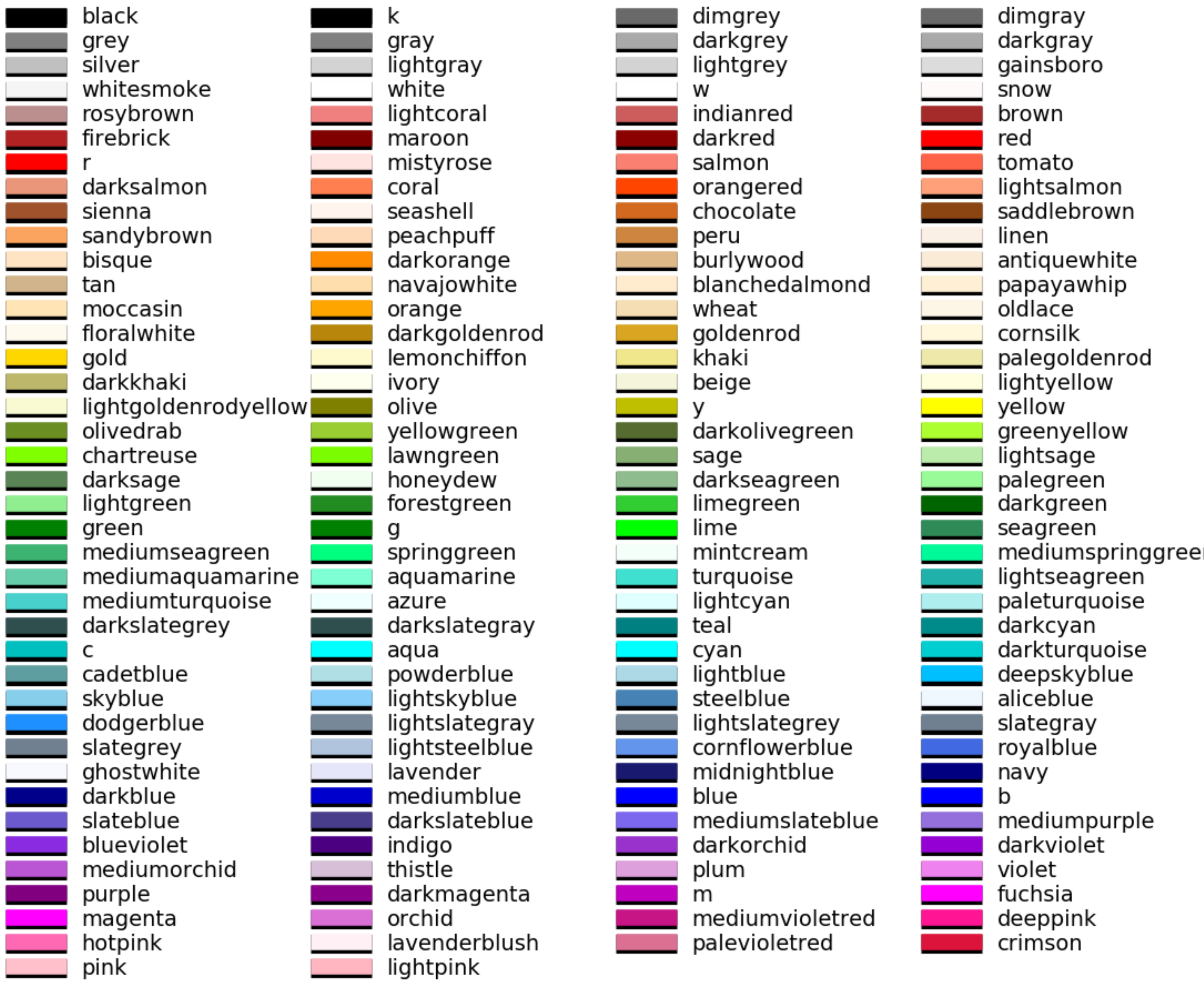

How fast do you want to go today?
Performance
14
12
10
8
6
4
2
0
Tom
Dick
Harry
Slim
Jim

Pie Chart

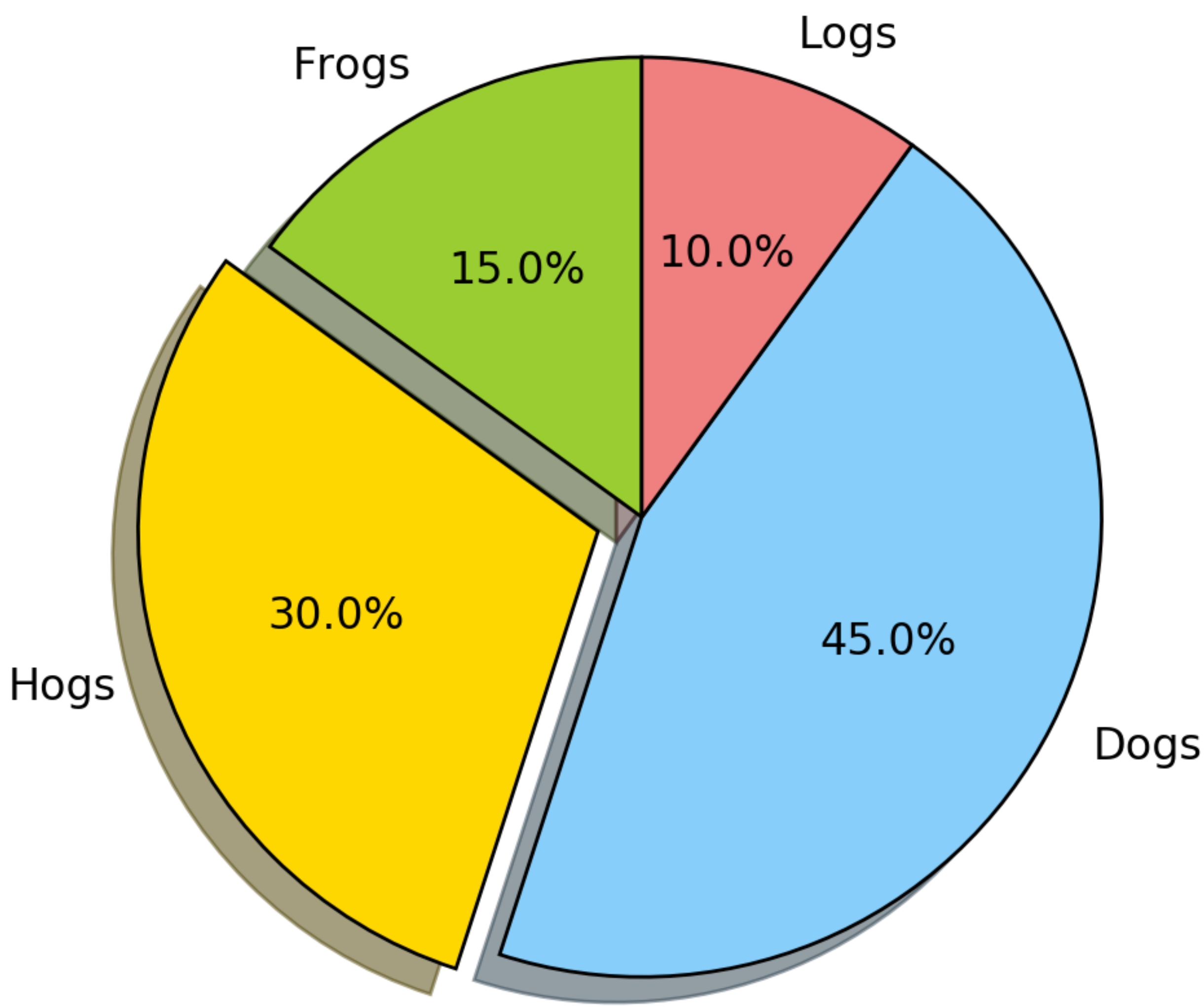
- Pie Chart

```
# The slices will be ordered and plotted counter-clockwise.
labels = 'Frogs', 'Hogs', 'Dogs', 'Logs'
sizes = [15, 30, 45, 10]
colors = ['yellowgreen', 'gold', 'lightskyblue', 'lightcoral']
explode = (0, 0.1, 0, 0)  # only "explode" the 2nd slice (i.e. 'Hogs')

plt.pie(sizes, explode=explode, labels=labels, colors=colors,
        autopct='%1.1f%%', shadow=True, startangle=90)
# Set aspect ratio to be equal so that pie is drawn as a circle.
plt.axis('equal')
plt.show()
```

Colors of Matplotlib



Pie Chart

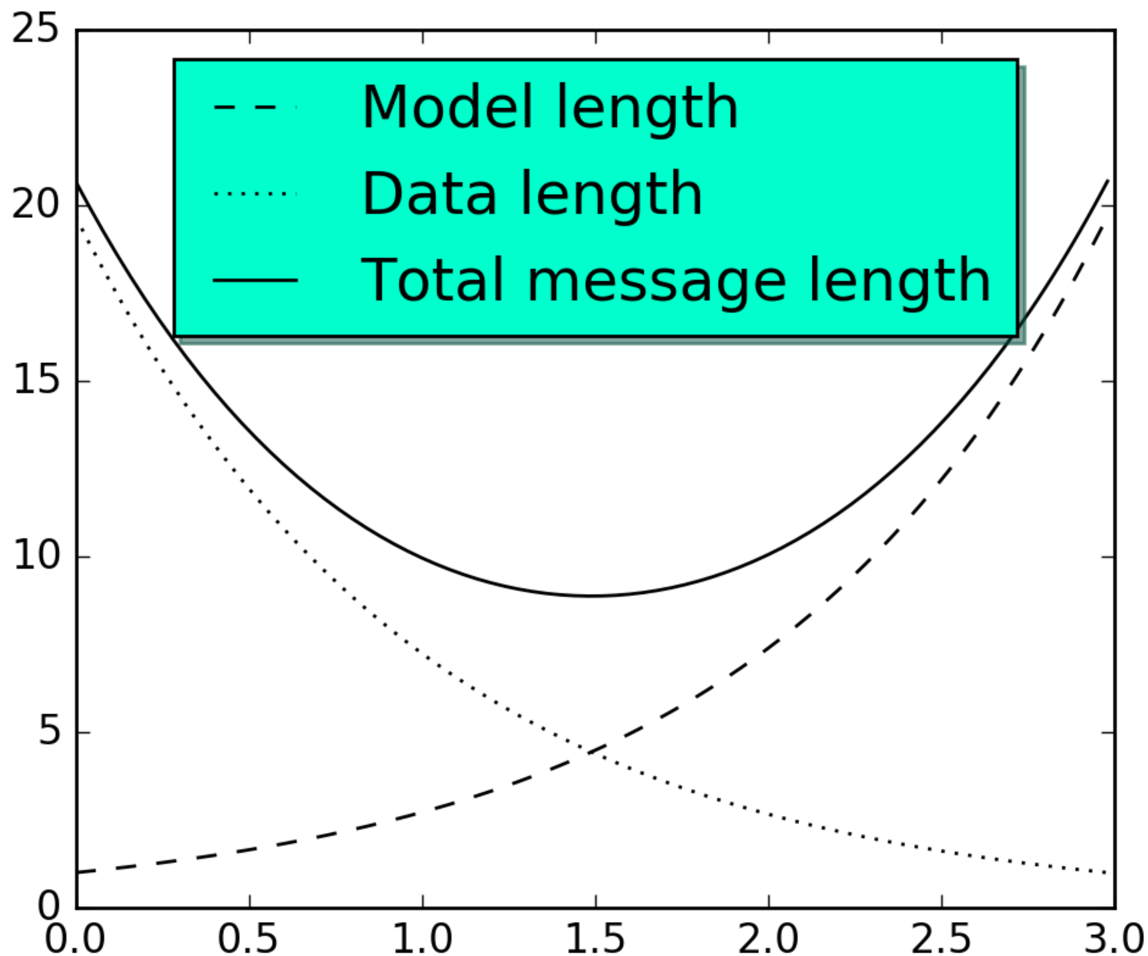


Legend

- Add legend

```
import numpy as np
import matplotlib.pyplot as plt

# Make some fake data.
a = b = np.arange(0, 3, .02)
c = np.exp(a)
d = c[::-1]
# Create plots with pre-defined labels.
plt.plot(a, c, 'k--', label='Model length')
plt.plot(a, d, 'k:', label='Data length')
plt.plot(a, c + d, 'k', label='Total message length')
legend = plt.legend(loc='upper center', shadow=True, fontsize='x-large')

# Put a nicer background color on the legend.
legend.get_frame().set_facecolor('#00FFCC')
plt.show()
```

Legend



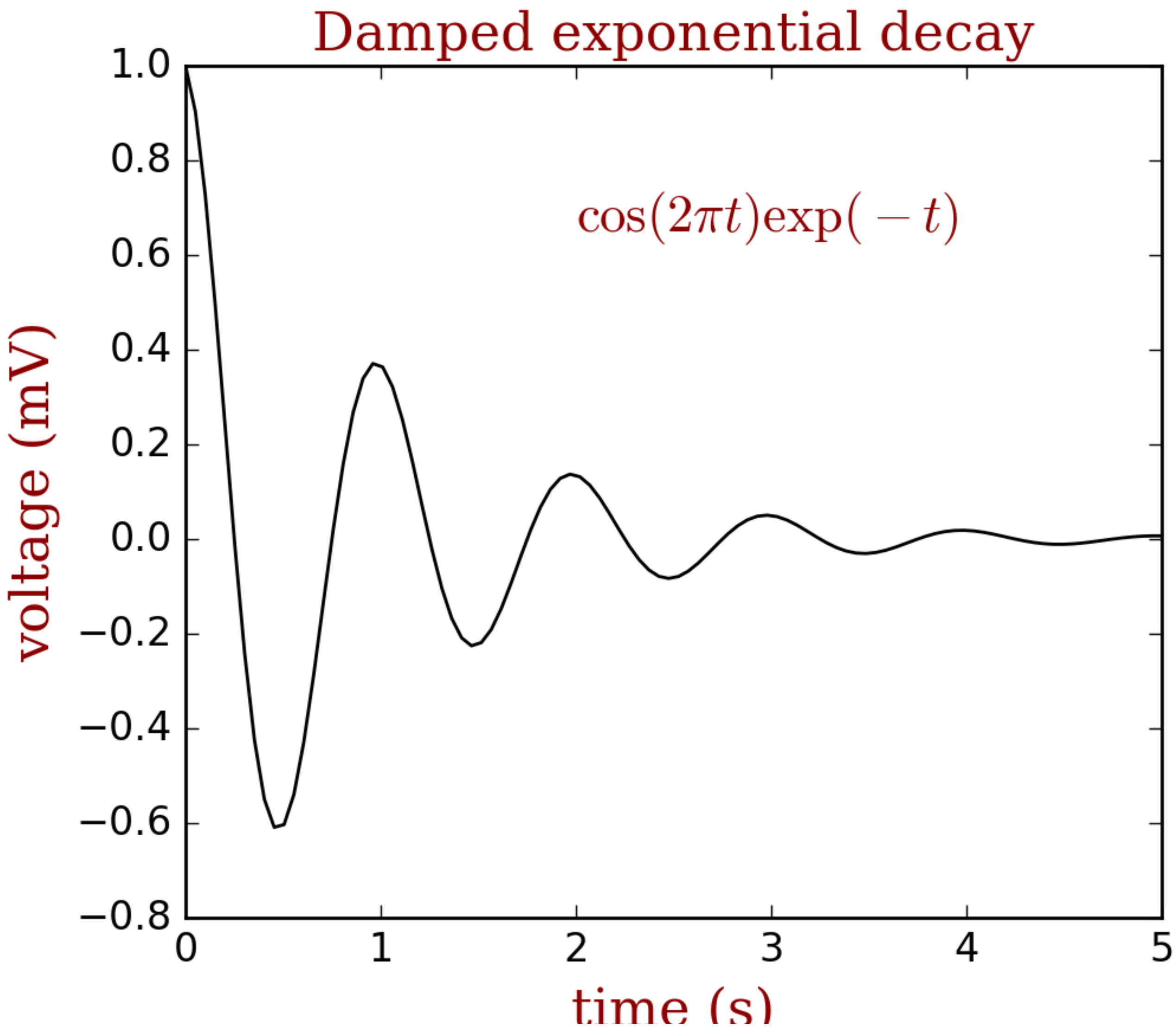
Model length
Data length
Total message length
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Text

- Put some texts

```
import numpy as np
import matplotlib.pyplot as plt

font = {'family': 'serif','color':  'darkred','weight': 'normal','size': 16,}

x = np.linspace(0.0, 5.0, 100)
y = np.cos(2*np.pi*x) * np.exp(-x)

plt.plot(x, y, 'k')
plt.title('Damped exponential decay', fontdict=font)
plt.text(2, 0.65, r'$₩cos(2 ₩pi t) ₩exp(-t)$', fontdict=font)
plt.xlabel('time (s)', fontdict=font)
plt.ylabel('voltage (mV)', fontdict=font)

# Tweak spacing to prevent clipping of ylabel
plt.subplots_adjust(left=0.15)
plt.show()
```

Text



Damped exponential decay
$\cos(2\pi t)\exp(-t)$
voltage (mV)
time (s)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
1
2
3
4
5