EMS702P Statistical Thinking and Applied Machine Learning

Week 6 .1 – Linear regression

Yunpeng Zhu

Linear regression

Edition: v1.1

Table of Contents

1 Matrix calculation

1) Transpose: Flips a matrix over its diagonal

$$\underset{m\times n}{\mathbf{A}} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$$

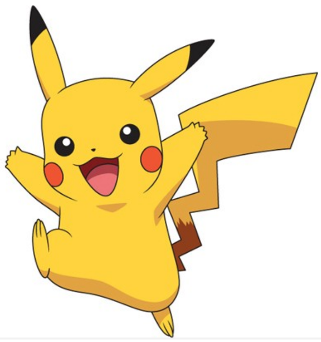
Switches the row and column indices of the matrix

$$\underset{n\times m}{\mathbf{A}^{\mathrm{T}}} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Quiz 1.1:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 2 & 3 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} 0.7 & -2 & 1 \end{bmatrix},$$



2) Addition: Add matrices together

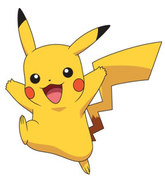
$$\underset{m\times n}{\mathbf{A}} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \text{ and } \underset{m\times n}{\mathbf{B}} = \begin{bmatrix} b_{1,1} & \cdots & b_{1,n} \\ \vdots & \ddots & \vdots \\ b_{m,1} & \cdots & b_{m,n} \end{bmatrix}$$

$$\underset{m\times n}{\mathbf{A}} + \underset{m\times n}{\mathbf{B}} = \begin{bmatrix} a_{1,1}+b_{1,1} & \cdots & a_{1,m}+b_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1}+b_{n,1} & \cdots & a_{n,m}+b_{n,m} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \end{bmatrix}$$

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} 4 & 4 & 4 \\ 10 & 10 & 10 \end{bmatrix}$$

Quiz 1.2:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 0 & 3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix},$$



3) Multiplication: Time matrices together

$$\underset{m\times n}{\mathbf{A}} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \text{ and } \underset{n\times p}{\mathbf{B}} = \begin{bmatrix} b_{1,1} & \cdots & b_{1,p} \\ \vdots & \ddots & \vdots \\ b_{n,1} & \cdots & b_{n,p} \end{bmatrix}$$

$$\underset{m\times n}{\mathbf{A}}\ \underset{n\times p}{\mathbf{B}} = \underset{m\times p}{\mathbf{C}}$$

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$$

$$i = 1,\ldots,m;\ j = 1,\ldots,p$$

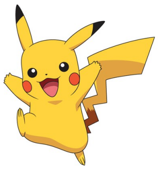
$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 1\times1+2\times0+3\times1 & 1\times2+2\times1+3\times0 \\ 4\times1+5\times0+6\times1 & 4\times2+5\times1+6\times0 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 4 \\ 10 & 13 \end{bmatrix}$$

Quiz 1.3:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 0 & 3 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix},$$



4) Inverse matrix (2x2)

$$\underset{2\times2}{\mathbf{A}} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}, \underset{2\times2}{\mathbf{A}^{-1}} = \frac{1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}, \mathbf{A}^{-1} = \frac{1}{1\times4 - 2\times3} \begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1.5 \\ 1 & -0.5 \end{bmatrix}$$

Quiz 1.4:

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix},$$

Matrix calculation properties [1]:

- *Non-commutativity:* $\mathbf{AB} \neq \mathbf{BA}$

$$\begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 4 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} -1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -2 & 4 \end{bmatrix}$$

➢ *Distributivity:* $\begin{cases}\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}\\(\mathbf{A}+\mathbf{B})\mathbf{C}=\mathbf{AC}+\mathbf{BC}\end{cases}$

$$\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times(\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}+\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})=\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}+\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}=\begin{bmatrix}0 & 2\\4 & 4\end{bmatrix}$$

$$(\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}+\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix})\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}=\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}+\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}=\begin{bmatrix}0 & 1\\2 & 3\end{bmatrix}$$

➢ *Product with a scalar:* $\alpha\mathbf{A}=\mathbf{A}\alpha$

$$3\times\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}=\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times 3=\begin{bmatrix}-3 & 3\\0 & 6\end{bmatrix}$$

➢ *Transpose:* $(\mathbf{AB})^{\mathrm{T}}=\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$

$$(\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix})^{\mathrm{T}}=\begin{bmatrix}1 & 1\\4 & 2\end{bmatrix}^{\mathrm{T}}=\begin{bmatrix}1 & 4\\1 & 2\end{bmatrix}$$

$$\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}^{\mathrm{T}}\times\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}^{\mathrm{T}}=\begin{bmatrix}1 & 2\\0 & 1\end{bmatrix}\times\begin{bmatrix}-1 & 0\\1 & 2\end{bmatrix}=\begin{bmatrix}1 & 4\\1 & 2\end{bmatrix}$$

➢ *Associativity:* $(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$

$$(\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix})\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}=\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times(\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}\times\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix})=\begin{bmatrix}1 & 1\\4 & 2\end{bmatrix}$$

➢ *Inverse:* $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$

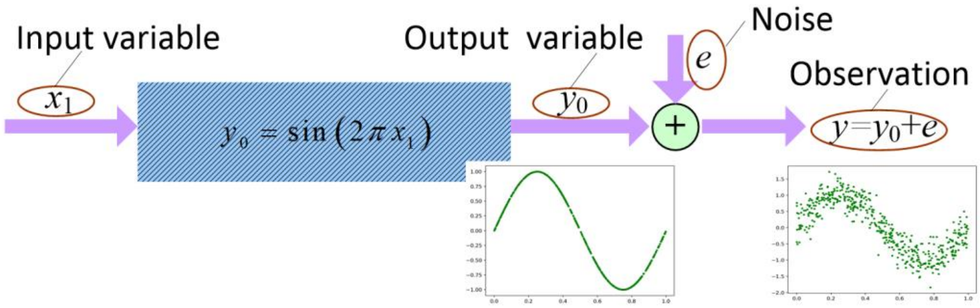
$$(\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}\times\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix})^{-1}=\begin{bmatrix}1 & 1\\4 & 2\end{bmatrix}^{-1}=\frac{1}{1\times 2-1\times 4}\begin{bmatrix}2 & -1\\-4 & 1\end{bmatrix}=\frac{1}{-2}\begin{bmatrix}2 & -1\\-4 & 1\end{bmatrix}$$

$$\begin{bmatrix}1 & 0\\2 & 1\end{bmatrix}^{-1}\times\begin{bmatrix}-1 & 1\\0 & 2\end{bmatrix}^{-1}=\frac{1}{1\times1-0\times 2}\begin{bmatrix}1 & 0\\-2 & 1\end{bmatrix}\times\frac{1}{(-1\times 2)-1\times 0}\begin{bmatrix}2 & -1\\0 & -1\end{bmatrix}$$

$$=\begin{bmatrix}1 & 0\\-2 & 1\end{bmatrix}\times\frac{1}{-2}\begin{bmatrix}2 & -1\\0 & -1\end{bmatrix}=\frac{1}{-2}\begin{bmatrix}2 & -1\\-4 & 1\end{bmatrix}$$

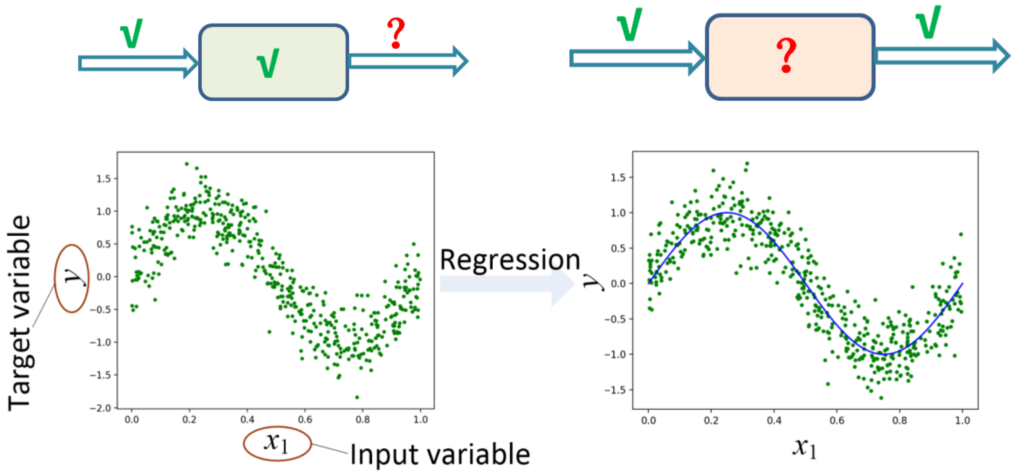
2 Introduction to linear regression

Consider a system governed by an equation: $y_0 = \sin(2\pi x_1)$



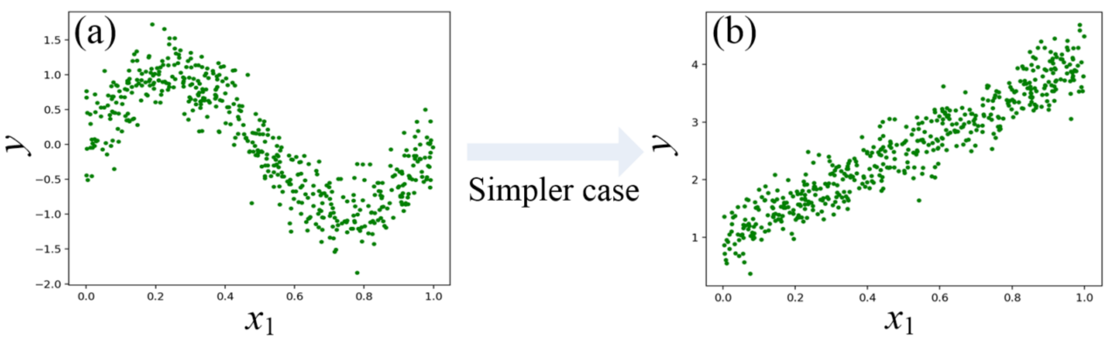
- In practice, given one x_1 gets one observation value of y ;

How to predict the value of y_0 under a x_1 **without** knowing the sine equation?



The goal of regression is to **predict** the value of one or more continuous **target** variables given the value of one/multiple dimensional **input** variables.

2.1 Linear model and linearity in parameters

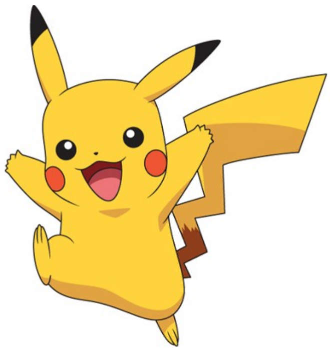


Consider a simpler case (b). The observation y is basically **a straight line**, so that the system model can be represented by a linear equation: $y = w_0 + w_1 x_1$, where w_1 and w_0 are parameters to be determined from the observation data.

- The equation $y = w_0 + w_1 x_1$ is a linear model (Straight line)

Quiz 2.1:

Plot the linear model $y = 1 + 2x$ (*How many points do you need?*)



Expansion 1: Linear model of *n*-dimensional space ($n \geq 2$, plane surface):

$$y = w_0 + w_1 x_1 + \cdots + w_n x_n$$

where $x_1, \ldots, x_n$ are n input variables, w_0 and $w_1, \ldots, w_n$ are the model parameters.

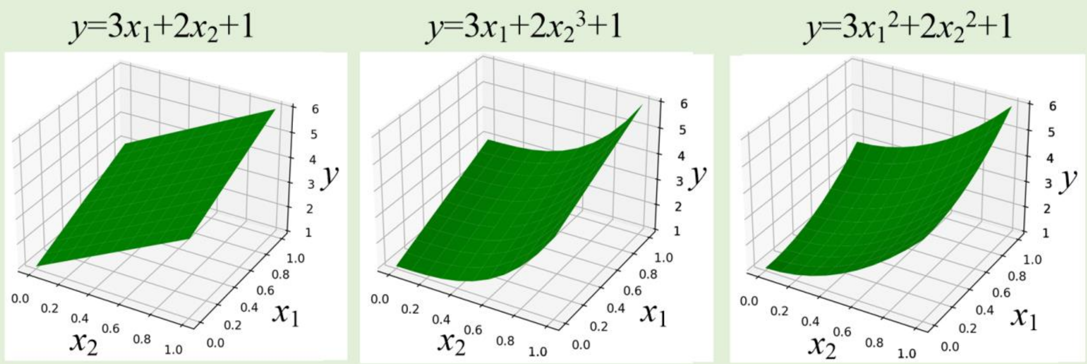
Expansion 2: By replacing $x_1, \ldots, x_n$ above with some **basis functions**:

$$y = w_0 + w_1 \varphi_1(\bar{\mathbf{x}}) + \cdots + w_n \varphi_n(\bar{\mathbf{x}})$$

where $\bar{\mathbf{x}} = [x_1, \ldots, x_n]^{\mathrm{T}}$, $\varphi_1(\bar{\mathbf{x}}), \ldots, \varphi_n(\bar{\mathbf{x}})$ are basis functions. This model is known as the **linear in parameters model**.

- **Basis function:** Every function in the function space can be represented as a linear combination of basis functions.

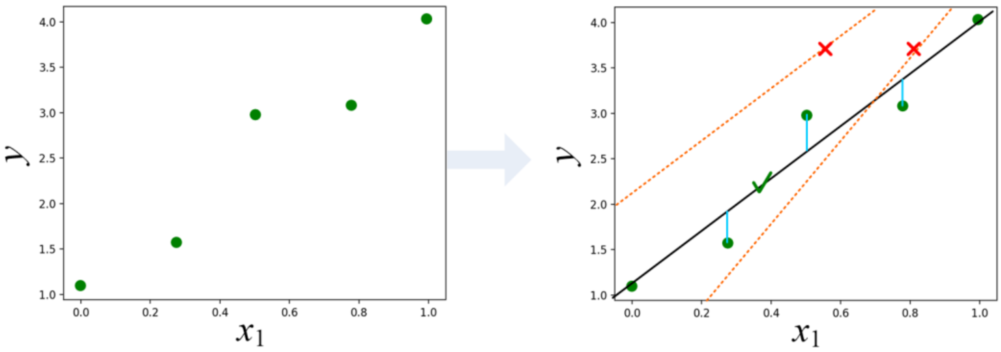
Example: 2-dimensional polynomial basis functions



- Any continuous curves/surfaces can be represented by a polynomial model with up to a sufficiently high order.

2.2 Regression and data fitting

Consider there are 5 observed values. Assuming the regression model is a straight line: $y = w_0 + w_1 x_1$. In the following plots, there are 3 lines. Which one is correct?



The black line is correct because the **distance** between each of the 5 points and the line is small (Note the distance is the **vertical distance**, representing the difference between two observations under the same variable). We say the black line **fits the data** well.

The criterion of regression and data fitting: Minimizes the vertical distance from the data points to the regression line.

Denote $\hat{y} = f\left[x_1\right]$ is the predicted value of the regression line:

- **Index 1 (Least absolute deviations):**

$$\min \sum_{i=1}^{5} \left| y(i) - \hat{y}(i) \right| = \min \sum_{i=1}^{5} \left| y(i) - f\left[x_1(i)\right] \right|$$

- **Index 2 (Least squares):**

$$\min \sum_{i=1}^{5} \left[y(i) - \hat{y}(i) \right]^2 = \min \sum_{i=1}^{5} \left[y(i) - f\left[x_1(i)\right] \right]^2$$

3 Least squares and maximum likelihood

3.1 The least squares algorithm

Consider a one-dimensional linear regression, we have:

$$\begin{aligned} y(1) &= \underbrace{w_0 \times 1 + w_1 x_1(1)}_{} + e(1) \\ y(2) &= w_0 \times 1 + w_1 x_1(2) + e(2) \\ y(3) &= w_0 \times 1 + w_1 x_1(3) + e(3) \\ y(4) &= w_0 \times 1 + w_1 x_1(4) + e(4) \\ y(5) &= \underbrace{w_0 \times 1 + w_1 x_1(5)}_{\hat{y}} + e(5) \end{aligned} \Rightarrow \underbrace{\begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ y(4) \\ y(5) \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & x_1(1) \\ 1 & x_1(2) \\ 1 & x_1(3) \\ 1 & x_1(4) \\ 1 & x_1(5) \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}_{\mathbf{W}} + \underbrace{\begin{bmatrix} e(1) \\ e(2) \\ e(3) \\ e(4) \\ e(5) \end{bmatrix}}_{\mathbf{e}}$$

where $e(i)$ represent the errors.

According to the least squares criterion, we need to find the values of w_0 and w_1 to achieve

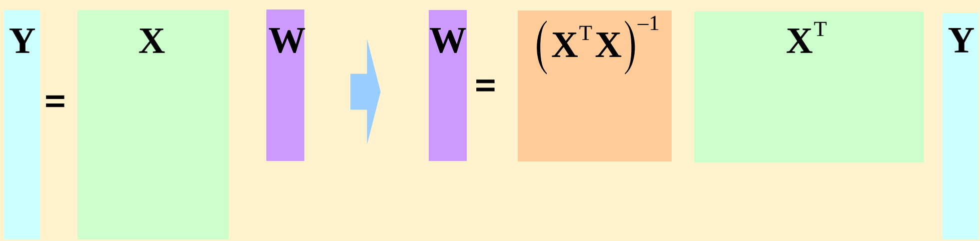
$$\begin{aligned} \min \sum_{i=1}^{5} [y(i) - \hat{y}(i)]^2 &= \min \sum_{i=1}^{5} [e(i)]^2 = \min \left[\mathbf{e}^{\mathrm{T}} \mathbf{e} \right] \\ &= \min \left[(\mathbf{Y} - \mathbf{XW})^{\mathrm{T}} (\mathbf{Y} - \mathbf{XW}) \right] = \min \|\mathbf{Y} - \mathbf{XW}\|_2^{\ 2} \end{aligned}$$

* $\|\bar{\mathbf{x}}\|_2 = \sqrt{\bar{\mathbf{x}}^{\mathrm{T}} \bar{\mathbf{x}}} = \sqrt{x_1^{\ 2} + \cdots + x_n^{\ 2}}$ called L2-norm [2].

Solve the minimization problem, we have:

The least squares function:

$$\underset{M\times 1}{\mathbf{W}} = \Big(\underset{M\times N}{\mathbf{X}^{\mathrm{T}}} \ \underset{N\times M}{\mathbf{X}} \Big)^{-1} \underset{M\times N}{\mathbf{X}^{\mathrm{T}}} \ \underset{N\times 1}{\mathbf{Y}}$$



Derivation: *See supplementary material 'Derivation of LS algorithm' on QM+.*

Example: Least squares regression of a linear model

$$y_0 = 1 + 3x_1 \rightarrow \begin{matrix} [x_1(1), y(1)] = [0, 1.1] \\ [x_1(2), y(2)] = [0.5, 2.3] \\ [x_1(3), y(3)] = [1, 4.2] \end{matrix} \rightarrow \begin{matrix} 1.1 = w_0 + e(1) \\ 2.3 = w_0 + 0.5w_1 + e(2) \\ 4.2 = w_0 + w_1 + e(3) \end{matrix}$$

$$\mathbf{Y} = \mathbf{XW} + \mathbf{e}$$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 0.96 \\ 3.1 \end{bmatrix} \leftarrow \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \begin{bmatrix} 1.1 \\ 2.3 \\ 4.2 \end{bmatrix} \leftarrow \begin{bmatrix} 1.1 \\ 2.3 \\ 4.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} + \begin{bmatrix} e(1) \\ e(2) \\ e(3) \end{bmatrix}$$

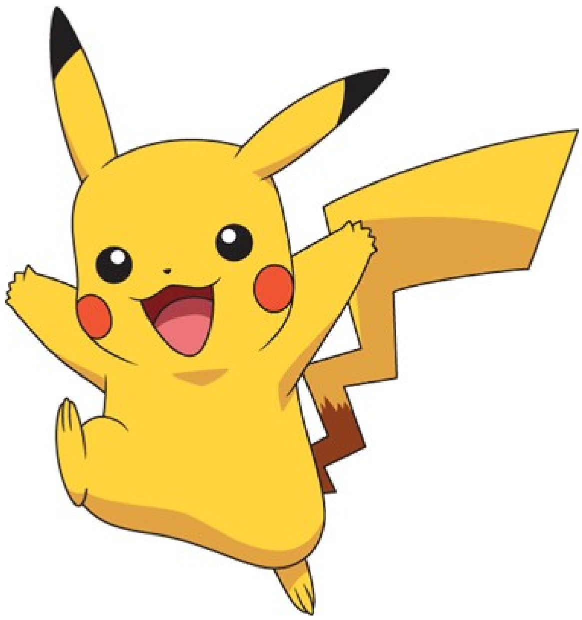
$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \\ 1 & 1 \end{bmatrix}, \mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1.5 \\ 1.5 & 1.25 \end{bmatrix},$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \frac{1}{3\times1.25 - 1.5\times1.5} \begin{bmatrix} 1.25 & -1.5 \\ -1.5 & 3 \end{bmatrix} = \begin{bmatrix} 0.83 & -1 \\ -1 & 2 \end{bmatrix},$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} 0.83 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.83 & 0.33 & -0.17 \\ -1 & 0 & 1 \end{bmatrix}$$

Quiz 3.1:

Find out the linear model using the following data: $(x, y): (0,2), (1,4)$



In the above discussions, $\mathbf{X}$ and $\mathbf{W}$ can be expanded to more complex and general cases:

Expansion 1: Linear model of *n*-dimensional space ($n \geq 2$, plane surface):

$$\underset{N\times1}{\mathbf{Y}} = \underset{N\times M}{\mathbf{X}}\ \underset{M\times1}{\mathbf{W}} + \underset{N\times1}{\mathbf{e}}$$

with

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1{}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N{}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(N) & \cdots & x_n(N) \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

K is the total number of the observation.

Expansion 2: By replacing $x_1, \ldots, x_n$ above with some basis functions:

$$\underset{N\times1}{\mathbf{Y}} = \underset{N\times M}{\mathbf{X}}\ \underset{M\times1}{\mathbf{W}} + \underset{N\times1}{\mathbf{e}}$$

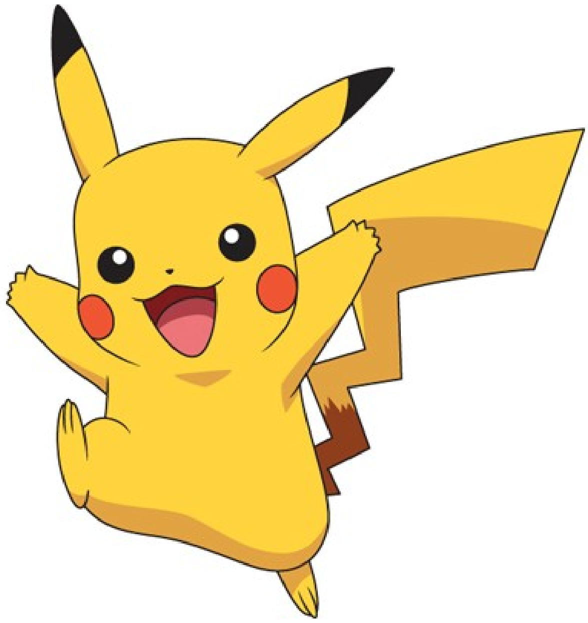
with

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1{}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N{}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & \varphi_1(\bar{\mathbf{x}}\{1\}) & \cdots & \varphi_n(\bar{\mathbf{x}}\{1\}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \varphi_1(\bar{\mathbf{x}}\{N\}) & \cdots & \varphi_n(\bar{\mathbf{x}}\{N\}) \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

N is the total number of the observation.

Quiz 3.2:

We know a model is $y = ax_1 + bx_2{}^2$. Determine the model using the following data: $(x_1, x_2, y): (1,1,3), (1,2,6)$



IT class (Python code):

```
from scipy.optimize import leastsq
def residuals_func(weights_vab, y, x):
   ret = fit_func(weights_vab, x) - y
   return ret
Weights = leastsq( residuals_func, Weights_init, args=(y, x) )
```

3.2 The maximum likelihood method

The **likelihood** of something happening is how likely it is to happen.

In $\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{e}$, the regression is to find $\mathbf{W}$ that are the highest likely to achieve $\mathbf{Y}$ under given $\mathbf{X}$. The function $P(\mathbf{Y}|\mathbf{W})$

is known as the **probability function** if $\mathbf{W}$ are known. It computes the probability of achieving $\mathbf{Y}$ under $\mathbf{W}$.

is known as the **likelihood function** if $\mathbf{Y}$ are known. It describes the probabilities of achieving $\mathbf{Y}$ under different values of $\mathbf{W}$.

Given different values of $\mathbf{W}$, the **maximum likelihood method** is to maximize the probability $P(\mathbf{Y}|\mathbf{W})$ by determining certain $\mathbf{W}$:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[P(\mathbf{Y}|\mathbf{W}) \right] = \arg\max_{\mathbf{W}} \left[\prod_{i=1}^{N} P(y(i)|\mathbf{W}) \right]$$

$$\prod_{i=1}^{N} x_i = x_1 x_2 x_3 \cdots x_N$$

Example: Coin flipping

Front: +; Back: -

11 test results: **Y**=[+ - - + - + + - + - +]

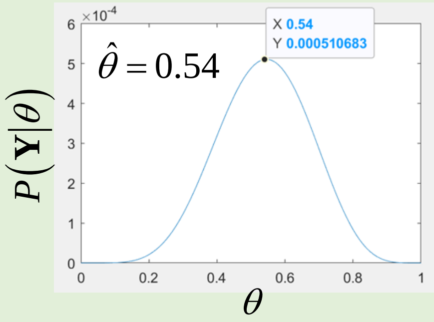
Estimate the front's probability: θ

$$
\begin{aligned}
P(\mathbf{Y}|\theta) &= \theta\times(1-\theta)\times(1-\theta)\times\theta\times(1-\theta)\times\theta\times\theta\times(1-\theta)\times\theta\times(1-\theta)\times\theta \\
&= \theta^6(1-\theta)^5
\end{aligned}
$$

representing the probability of achieving $\mathbf{Y}$ under a front's probability θ.

Maximum likelihood estimation:

$$
\hat{\theta} = \arg\max_{\theta}\left[\theta^6(1-\theta)^5\right]
$$



3.3 The maximum likelihood based linear regression

Consider the linear regression problem

$$
y = \begin{cases} w_0 + w_1 x_1 + \cdots + w_n x_n + e \\ w_0 + w_1\varphi_1(\bar{\mathbf{x}}) + \cdots + w_n\varphi_n(\bar{\mathbf{x}}) + e \end{cases} = \underset{1\times M}{\mathbf{x}^{\mathrm{T}}}\ \underset{M\times 1}{\mathbf{W}} + e
$$

for example,

$$
y = 2 + x_1 + 3x_2 - 2x_3 + e = \begin{bmatrix} 1 & x_1 & x_2 & x_3 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \\ 3 \\ -2 \end{bmatrix} + e
$$

Assuming the regression error (residual) is normally distributed as

$$
e \sim \mathrm{N}(0,\sigma^2) \Rightarrow y \sim \mathrm{N}(\mathbf{x}^{\mathrm{T}}\mathbf{W},\sigma^2)
$$

$$
P(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(y-\mathbf{x}^{\mathrm{T}}\mathbf{W})^2}{2\sigma^2}\right)
$$

The maximum likelihood based linear regression is evaluated as

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[P\left(\mathbf{Y}|\mathbf{W}\right) \right] = \arg\max_{\mathbf{W}} \left[\prod_{i=1}^{N} P\left(y(i)|\mathbf{W}\right) \right]$$

$$= \arg\max_{\mathbf{W}} \left[\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\left(y(i) - \mathbf{x}_i^{\mathrm{T}}\mathbf{W}\right)^2}{2\sigma^2} \right) \right]$$

$$= \arg\max_{\mathbf{W}} \left[\frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} \exp\left(\sum_{i=1}^{N} \frac{-\left(y(i) - \mathbf{x}_i^{\mathrm{T}}\mathbf{W}\right)^2}{2\sigma^2} \right) \right]$$

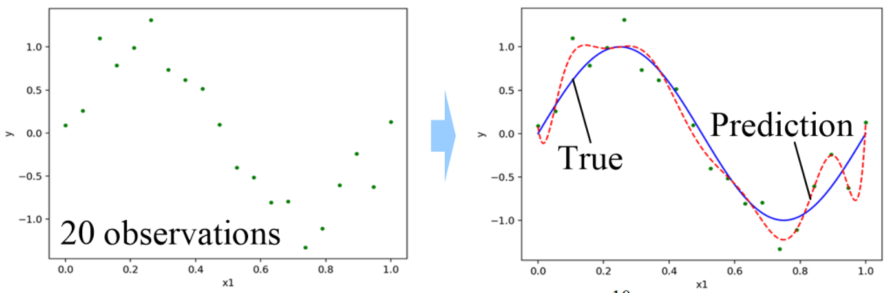
where $\mathbf{x}_i$ represents the i th values of model terms, so that

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left(y(i) - \mathbf{x}_i^{\mathrm{T}}\mathbf{W}\right)^2$$

which is exactly the same as the least square criterion.

4 Regularized least squares for linear regression

Consider the following case:



$$y = w_0 + w_1 x_1 + \cdots + w_{10} x_1^{10}$$

The data is **over-fitted** by the curve: The curve tends to approach the observed points with small error [3].

Regularization: Relax the errors to make the curve smooth.

If the evaluated $\mathbf{W}$ is large, when the value of $\mathbf{x}$ changes slightly, the value of y changes significantly. To make the curve smooth, constraints of the evaluated values of $\mathbf{W}$ are considered.

$$\min\|\mathbf{Y}-\mathbf{XW}\|_2 \quad \xrightarrow{\text{Regularization}} \quad \min\left(\|\mathbf{Y}-\mathbf{XW}\|_2 + \boxed{\lambda\|\mathbf{W}\|_2}\right)$$

Penalty

$$\underbrace{\mathbf{W}}_{M\times 1} = \left(\underbrace{\mathbf{X}^{\mathrm{T}}}_{M\times N}\underbrace{\mathbf{X}}_{N\times M}\right)^{-1}\underbrace{\mathbf{X}^{\mathrm{T}}}_{M\times N}\underbrace{\mathbf{Y}}_{N\times 1}$$

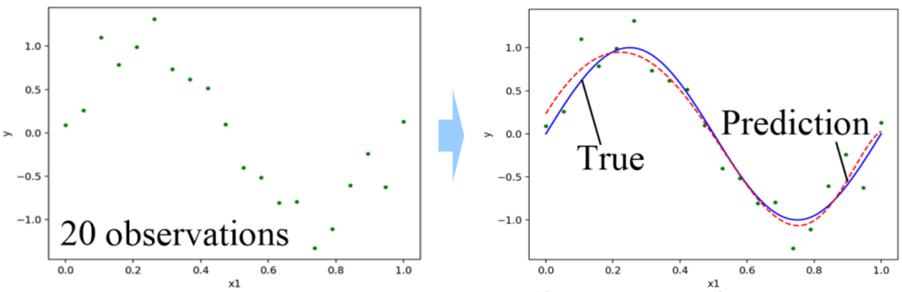
Lese squares

$$\underbrace{\mathbf{W}}_{M\times 1} = \left(\underbrace{\mathbf{X}^{\mathrm{T}}}_{M\times N}\underbrace{\mathbf{X}}_{N\times M} + \underbrace{\boxed{\lambda\mathbf{I}}}_{M\times M}\right)^{-1}\underbrace{\mathbf{X}^{\mathrm{T}}}_{M\times N}\underbrace{\mathbf{Y}}_{N\times 1}$$

Regularized lese squares

$\lambda > 0$ is a constant, $\mathbf{I}$ is an unit matrix.

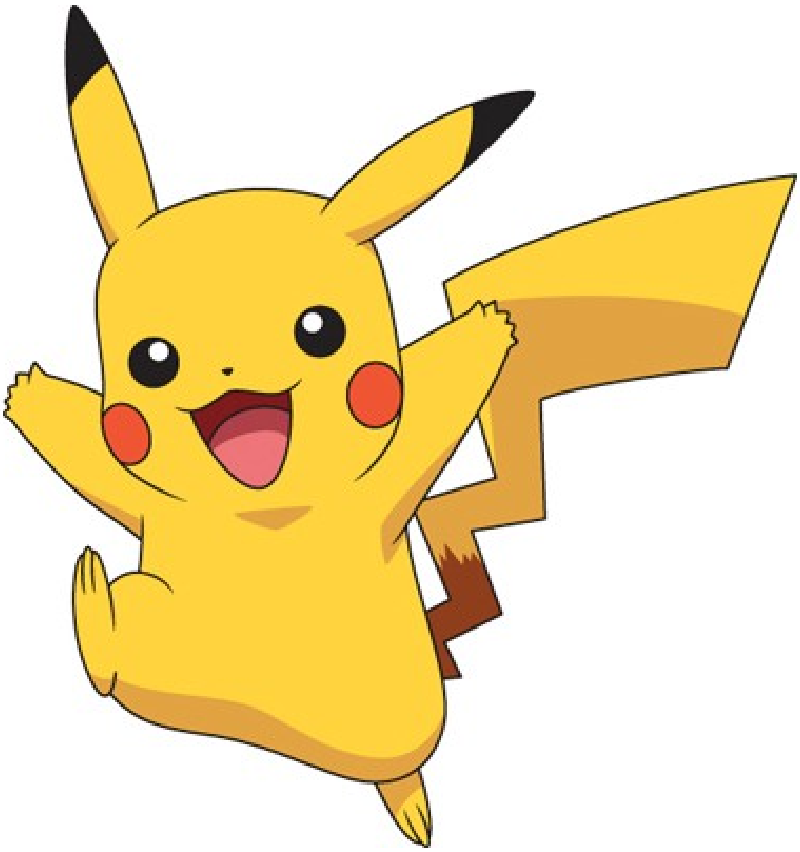
Derivation: *See supplementary material 'Derivation of RLS algorithm' on QM+.*



$$y = w_0 + w_1x_1 + \cdots + w_{10}x_1^{10},\ \lambda = 1\times10^{-3}$$

Quiz 4.1:

We know a model is $y = ax_1 + bx_2^{\,2}$. Determine the model using regularized least squares with the following data: $(x_1, x_2, y): (1,1,3), (1,2,6)$



IT class (Python code):

```
from scipy.optimize import leastsq
def residuals_func(weights_vab, y, x):
    ret = fit_func(weights_vab, x) - y
    ret = np.append( ret, np.sqrt(lamda * np.square(weights_vab)) )
    return ret
weights = leastsq( residuals_func, Weights_init, args=(y, x) )
```

5 Model validation

5.1 Prediction errors

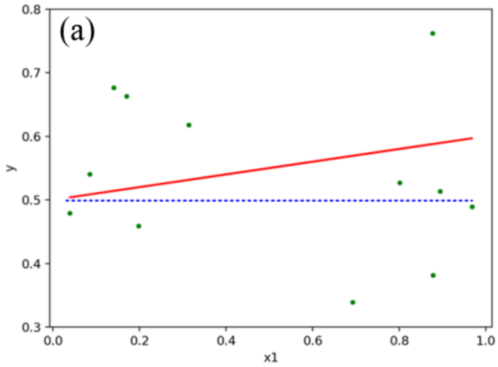
Root Mean Square Error (RMSE): $\eta_{\text{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[y(i)-\hat{y}(i)\right]^2}$

Mean Square Error (MSE): $\eta_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left[y(i)-\hat{y}(i)\right]^2$

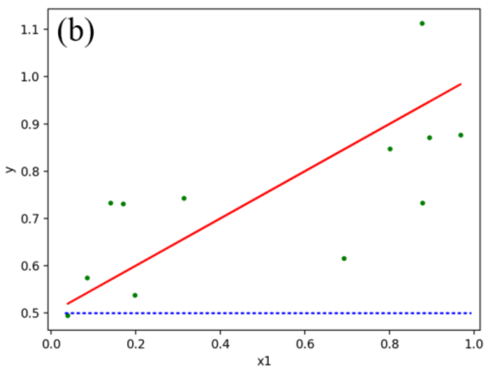
Mean Absolute Error (MAE): $\eta_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}\left|y(i)-\hat{y}(i)\right|$

Mean Relative Error (MAE): $\eta_{\text{MRE}} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y(i)-\hat{y}(i)}{y(i)}\right|$

5.2 Hypothesis test for linear regression



$y = 0.5 + 0.1x_1$ ➡ Is 0.1 significant?



$y = 0.5 + 0.51x_1$ ➡ Is 0.51 significant?

(1) T-test

T-distribution (Student's T-distribution): Estimating the mean of a **normally distributed population** in situations where the **sample size is small** and the **population's standard deviation is unknown**.

$\bar{s}, s$: Stansard deviation of sample

$\bar{\sigma}, \sigma$: Stansard deviation of population

$$x_n \sim \mathrm{N}\left(\mu, \sigma^2\right) \quad \xrightarrow{Z = \frac{x_n - \mu}{\sigma}} \quad Z \sim \mathrm{N}(0,1)$$

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_i \sim \mathrm{N}\left(\mu, \bar{\sigma}^2 = \sigma^2 / N\right) \quad \xrightarrow{Z = \frac{\bar{x} - \mu}{\bar{\sigma}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}} \quad Z \sim \mathrm{N}(0,1)$$

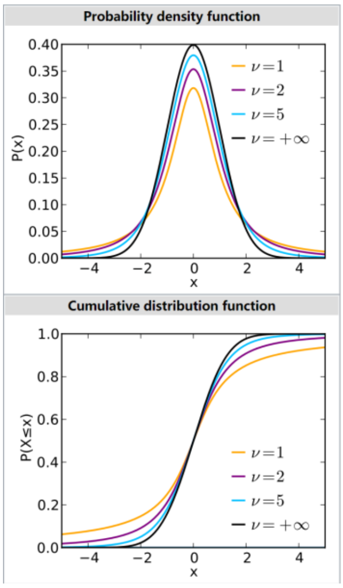
$$T = \frac{\bar{x} - \mu}{\bar{s}} = \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

$$T \sim \mathrm{t}(v)$$

$$v = N - p,\ p = 1$$

v : Degree of freedom

p: Number of variables



Consider a linear regression model:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + e, e \sim \mathrm{N}\left(0, \sigma^2\right)$$

and in matrix form

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{e}$$

The aim of the T-test is to check the linearities of the relationship between the response variable y and different model coefficients $w_1, \ldots, w_n$.

The T-Test of the model coefficients is conducted one by one. The null and alternative hypotheses for the T-test are

$$\begin{matrix} H_0 : w_j = 0 \\ H_A : w_j \neq 0 \end{matrix} \quad \text{for } j = 1, \ldots, n$$

Assuming the model coefficients are normally distributed:

$$\hat{w}_j \sim \mathrm{N}\left(w_j, \sigma_j^{\ 2}\right)$$

where $\sigma_j^{\ 2} = \underbrace{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}_{j+1,\, j+1} \sigma^2,\ j = 1, \ldots, n$.

Derivation: *See supplementary material 'T-test for linear regression' on QM+.*

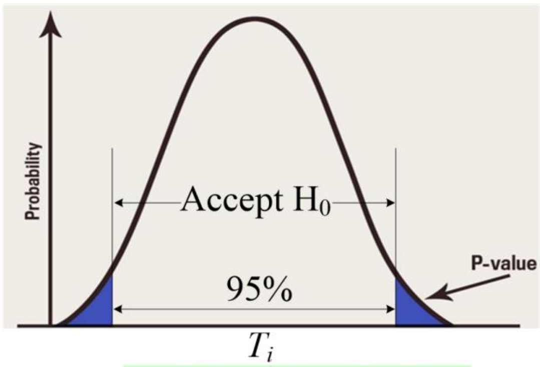
Denote $\mathbf{C} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$, consider the variable

$$T_j = \frac{\hat{w}_j - 0}{s_j} = \frac{\hat{w}_j}{\sqrt{\mathbf{C}(j+1, j+1)}s} \sim \mathrm{t}\left(N - n - 1\right),\ j = 1, \ldots, n$$

where s_j, s are the standard deviation of samples corresponding to σ_j, σ for population, respectively.

$$s = \sqrt{\frac{\sum_{i=1}^{N} e(i)^2}{N-n-1}} = \sqrt{\frac{\sum_{i=1}^{N} \left[y(i) - \hat{y}(i)\right]^2}{N-n-1}}$$

where $\hat{y}(i)$ is the prediction value under the **estimated parameters**.



$v = N - n - 1 = 12 - 1 - 1$

$T_1 \sim \mathrm{t}(\boxed{10});\ P = 0.05$

(a) $T_1 = 0.862 < 2.228 \rightarrow$ Accept $\mathrm{H_0}$

(b) $T_1 = 4.311 > 2.228 \rightarrow$ Reject $\mathrm{H_0}$

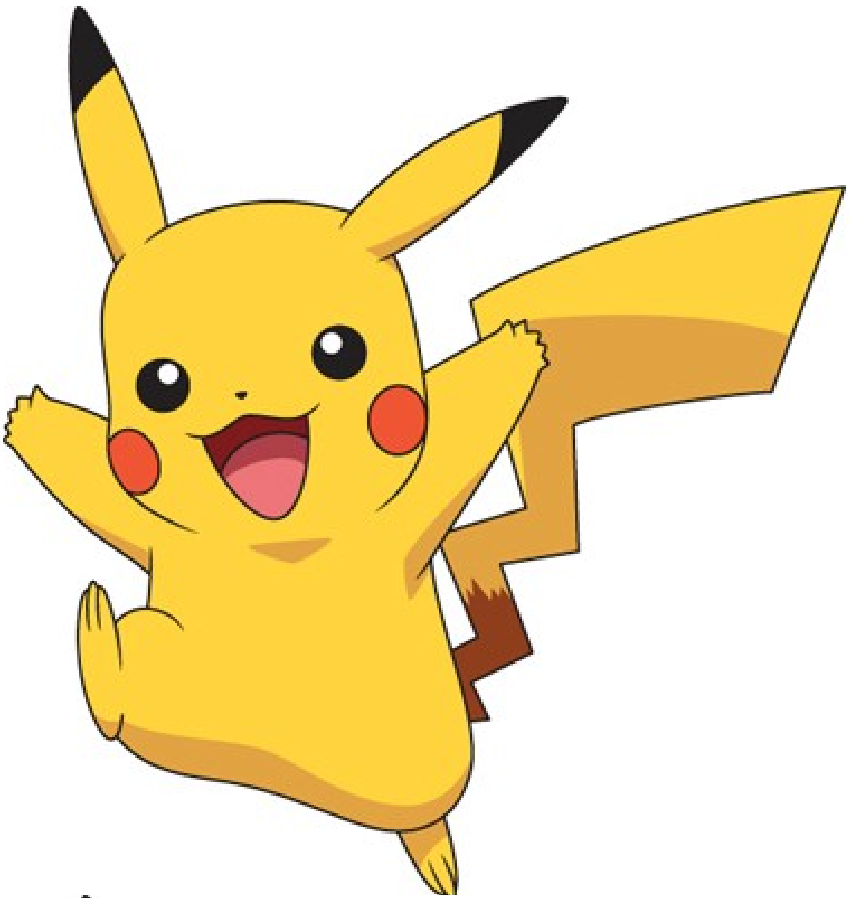
Significance level (α) (2-tail)

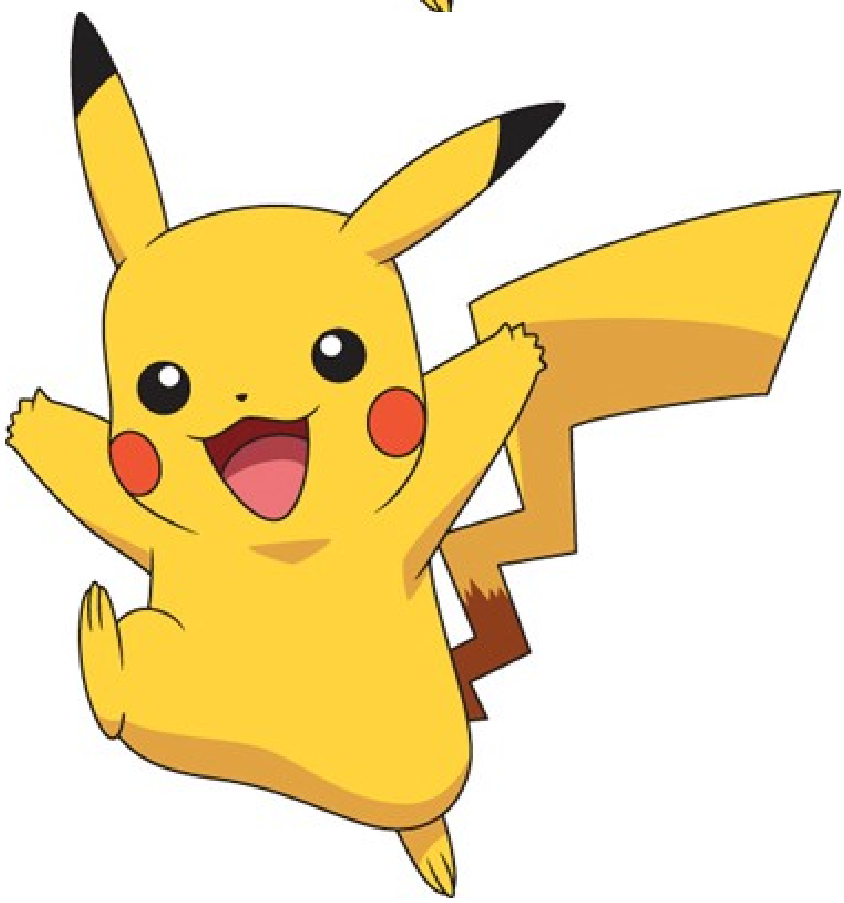
Degrees of freedom (df)	.2	.15	.1	.05	.025
1	3.078	4.165	6.314	12.706	25.452
2	1.886	2.282	2.920	4.303	6.205
3	1.638	1.924	2.353	3.182	4.177
4	1.533	1.778	2.132	2.776	3.495
5	1.476	1.699	2.015	2.571	3.163
6	1.440	1.650	1.943	2.447	2.969
7	1.415	1.617	1.895	2.365	2.841
8	1.397	1.592	1.860	2.306	2.752
9	1.383	1.574	1.833	2.262	2.685
10	1.372	1.559	1.812	2.228	2.634
11	1.363	1.548	1.796	2.201	2.593
12	1.356	1.538	1.782	2.179	2.560
13	1.350	1.530	1.771	2.160	2.533
14	1.345	1.523	1.761	2.145	2.510
15	1.341	1.517	1.753	2.131	2.490
16	1.337	1.512	1.746	2.120	2.473
17	1.333	1.508	1.740	2.110	2.458
18	1.330	1.504	1.734	2.101	2.445
19	1.328	1.500	1.729	2.093	2.433
20	1.325	1.497	1.725	2.086	2.423

Quiz 5.1:

I identified a model $y = 1 + 0.2x$ from the data set: $(x, y):(0,1),(1,0.9),(2,1.1)$.
Is this model correct?

The linear model under all data can be written as





(2) F- test

The T-test can only check one parameter each time. If we want to check multiple parameters at the same time, F-test can be applied.

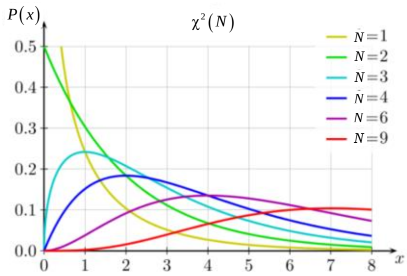
$$X_i \sim \mathrm{N}\left(\mu, \sigma^2\right), i = 1, \ldots, N$$

$$Z_i = \frac{X_i - \mu}{\sigma}$$

$$Z_i \sim \mathrm{N}(0,1)$$

$$P = \sum_{i=1}^{N} Z_i^{\,2}$$

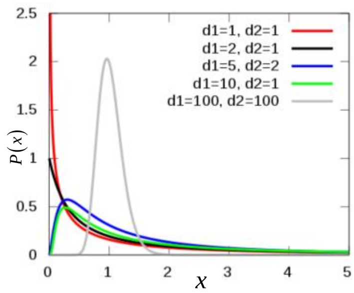
$$P \sim \chi^2(N)$$



$$A \sim \chi^2(d_1),\ B \sim \chi^2(d_2)$$

$$F = \frac{A/d_1}{B/d_2}$$

$$F \sim \mathrm{f}\left(d_1, d_2\right)$$



Denote the full linear regression model as **the unrestricted model**, i.e.

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$$

Under the hypothesis

$$H_0 : w_4 = w_5 = 0$$
$$H_A : w_4, w_5 \neq 0$$

the **restricted model** is defined as, i.e.

$$y = w_0' + w_1'x_1 + w_2'x_2 + w_3'x_3$$

where w_0', w_1', w_2', w_3' are coefficients of the restricted model.

Denote **sum of squares of residuals (SSR)** is:

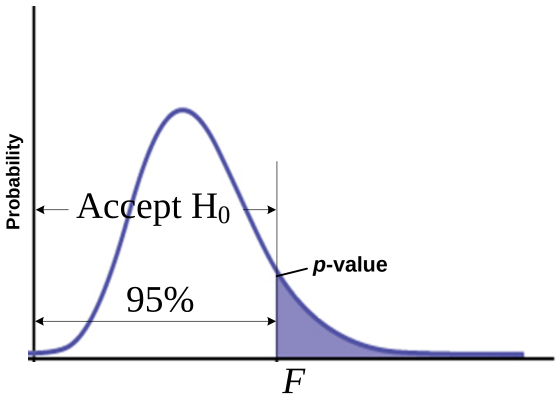
$$SSR = \sum_{i=1}^{N} e(i)^2 = \sum_{i=1}^{N} \left[y(i) - \hat{y}(i) \right]^2$$

Then the variable F is defined as

$$F = \frac{\left(SSR_{\mathrm{r}} - SSR_{\mathrm{ur}}\right) / \left(n_{\mathrm{r}} - n_{\mathrm{ur}}\right)}{SSR_{\mathrm{ur}} / \left(N - n_{\mathrm{ur}} - 1\right)} \sim \mathrm{f}\left(n_{\mathrm{r}} - n_{\mathrm{ur}}, N - n_{\mathrm{ur}} - 1\right)$$

where n_r and n_{ur} are the numbers of the restricted and unrestricted model parameters, respectively. For example, $n_r = 3$ and $n_{ur} = 5$ for the above case.

Derivation: *See supplementary material 'F-test for linear regression' on QM+.*



$$F \sim f(1,10);\ P = 0.05$$

(a) $F = 0.44 < 4.96$ ➡ Accept H_0

(b) $F = 10.50 > 4.96$ ➡ Reject H_0

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	$\boxed{4.96}$	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35

6 Further Readings

[1] Matrix multiplication on Wikipedia.

https://en.wikipedia.org/wiki/Matrix_multiplication

[2] Norm on Wikipedia.

https://en.wikipedia.org/wiki/Norm_(mathematics)

[3] Overfitting and Underfitting problems

https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765