Machine Learning

-Linear Regression-

SCH Univ.
Dept. of AI and Bigdata
Kim JinSeong

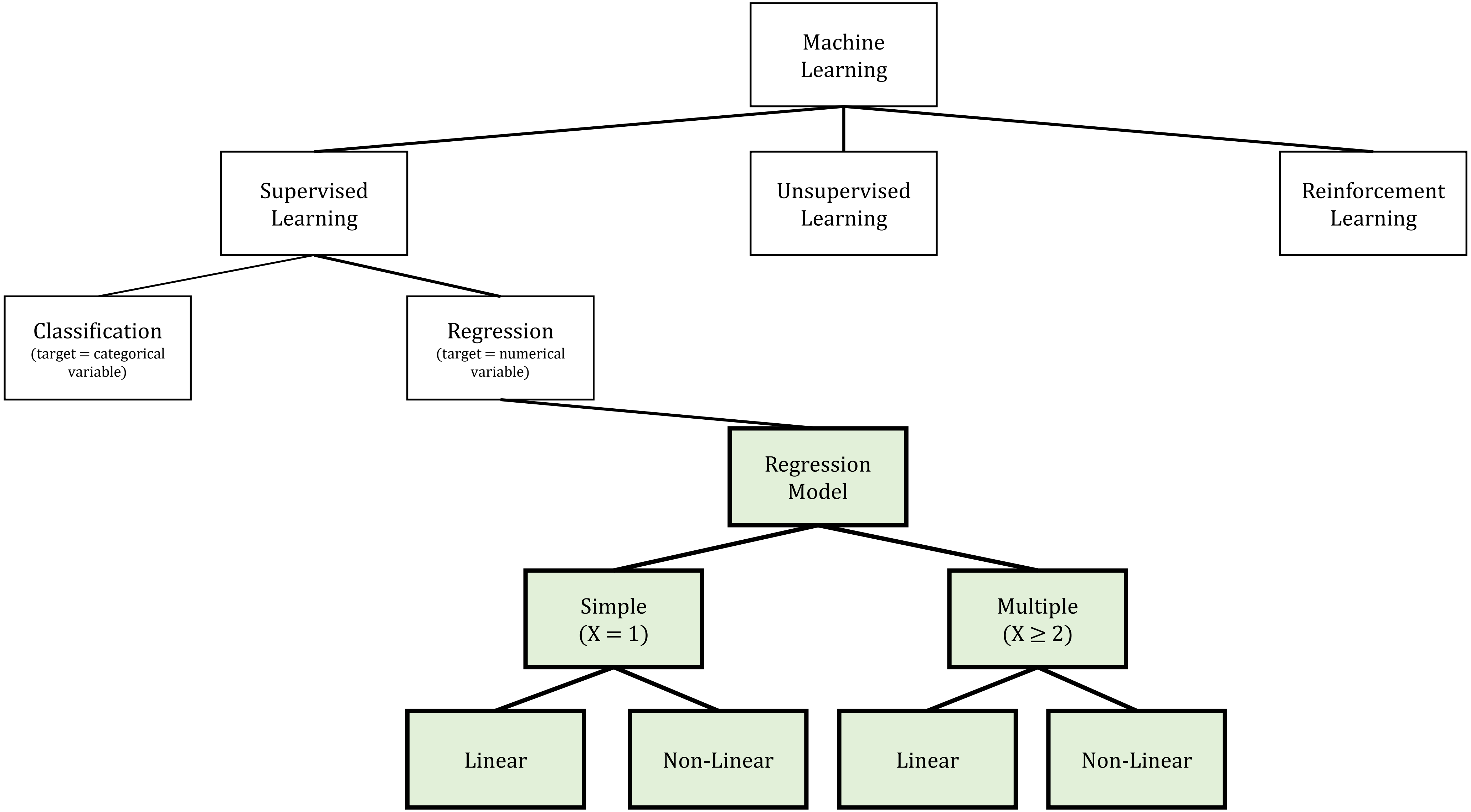
SOON CHUN HYANG UNIVERSITY

Contents

Chapter Ⅰ

-Definition of Linear Regression-

Types of Machine Learning

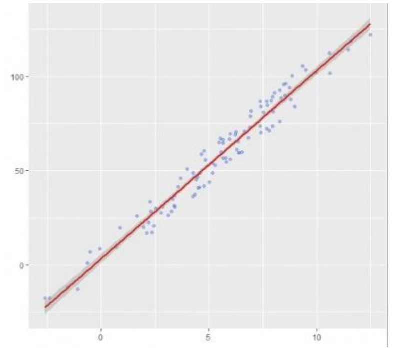


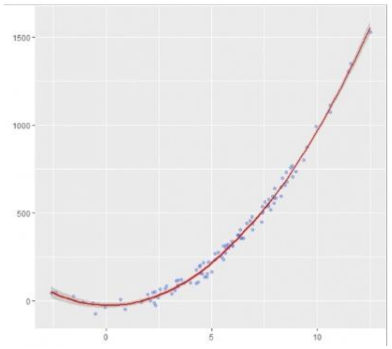
Types of Regression Model

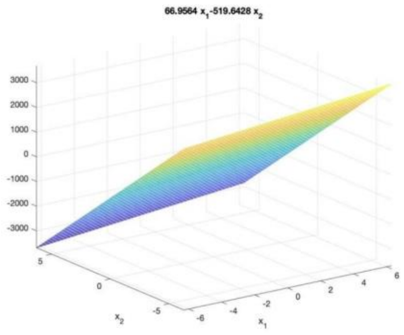
Regression Model

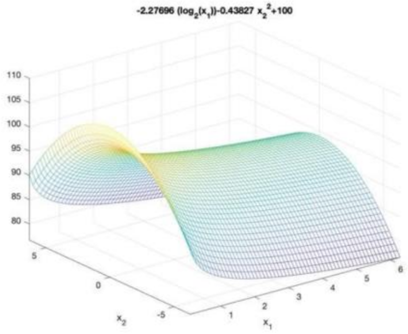
Simple (X = 1) | Multiple (X ≥ 2)

Linear | Non-Linear | Linear | Non-Linear







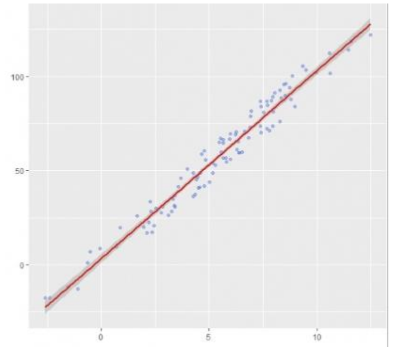


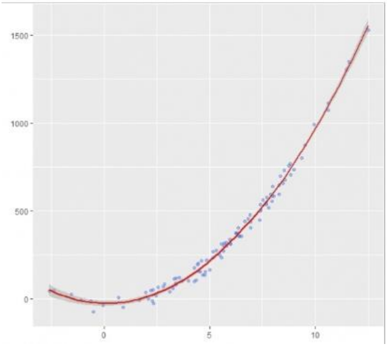
https://www.researchgate.net/figure/Multivariate-non-linear-regression-model-2_fig2_351938618
https://wikidocs.net/34065

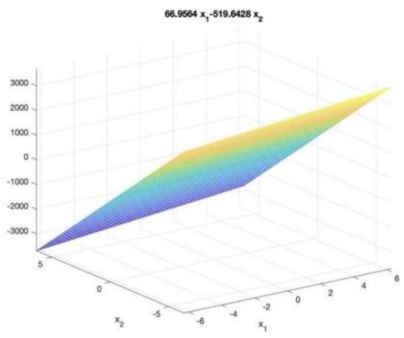
Types of Regression Model

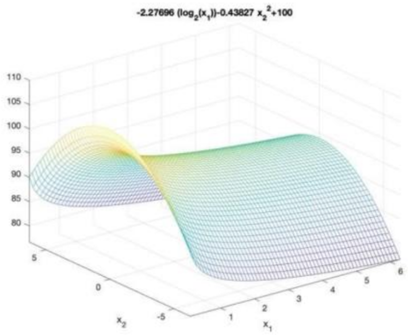
Regression Model

- Simple ($X = 1$)
  - Linear
  - Non-Linear
- Multiple ($X \geq 2$)
  - Linear
  - Non-Linear







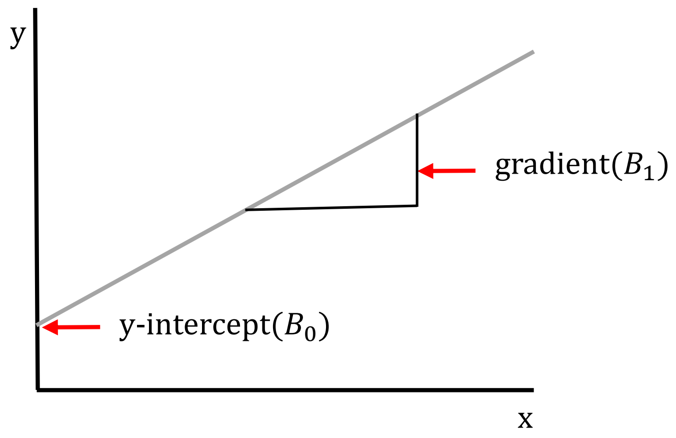


https://www.researchgate.net/figure/Multivariate-non-linear-regression-model-2_fig2_351938618
https://wikidocs.net/34065

Definition of Linear Regression Model

Linear Regression Model: Model that expresses Y(output variable) as a linear combination of X(input variable)

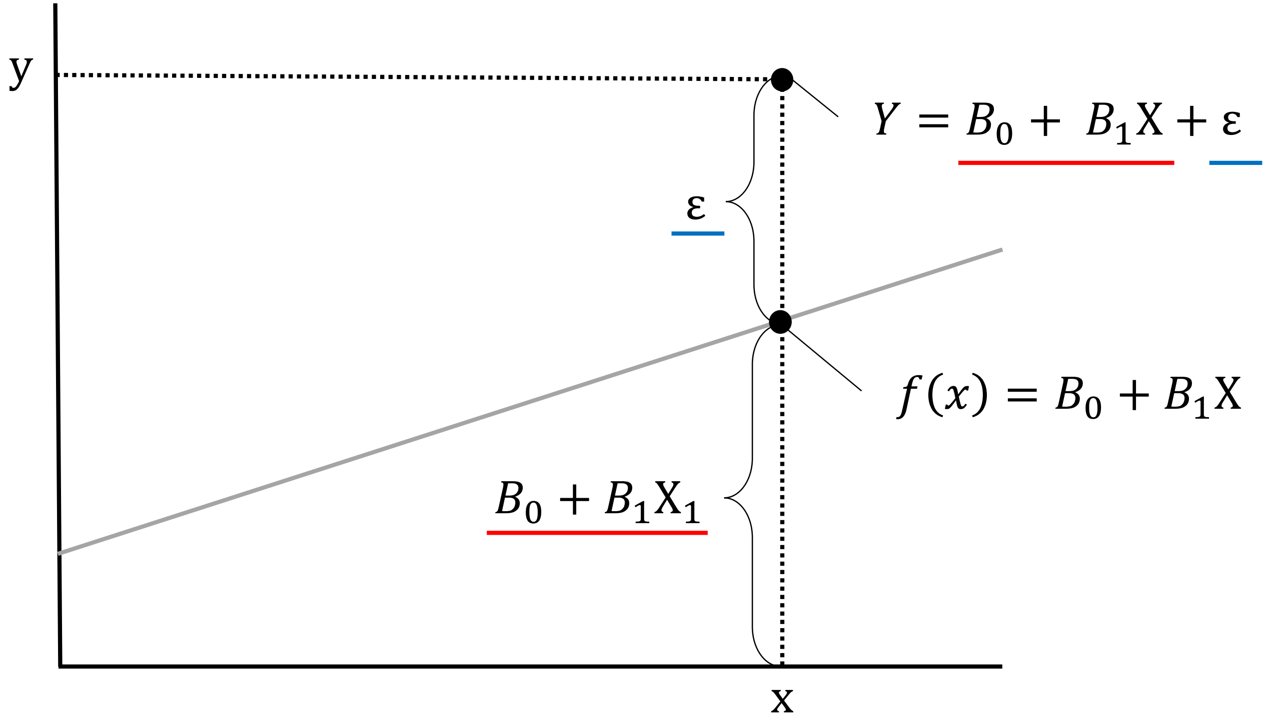
* Linear combination : Combine variables by adding/subtracting (constant multiplication)
ex) $Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$



$$Y = B_0 + B_1 X_1$$

Purpose 1. Explain the relationship between X variables and Y variables
2. Predict future Y(output variables)

Definition of Linear Regression Model



$$Y = \underline{can\ be\ explained\ by\ X(f(x))} + \underline{can't\ be\ explained\ by\ X(\varepsilon)}$$

ε = random error

Assumption of Linear Regression Model

- Assumption of random error

➡ $\varepsilon_i \sim N(0, \sigma^2) \quad i = 1,2,3,..,n$

ε_i conforms to a normal distribution $\rightarrow E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2$ for all i

In $Y = B_0 + B_1 X + \varepsilon$, ε Follows probability distribution(normal distribution)
So, Y also follows any probability distribution

1. $E(Y_i) = E(B_0 + B_1 X_i) + E(\varepsilon) = B_0 + B_1 X_i$

$B_0 + B_1 X_i$ is constant $\rightarrow E(B_0 + B_1 X_i) = B_0 + B_1 X_i$

$E(\varepsilon_i) = 0$

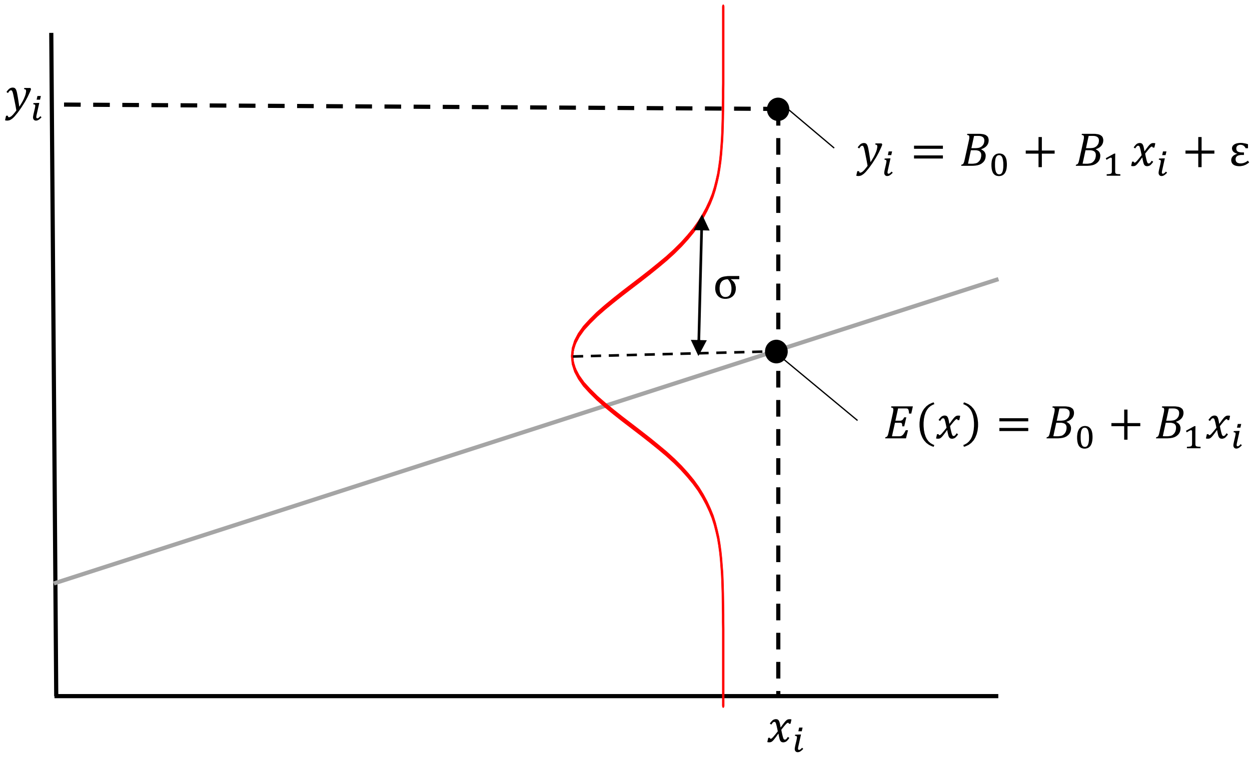
2. $V(Y_i) = V(B_0 + B_1 X_i) + V(\varepsilon) = \sigma^2$

$B_0 + B_1 X_i$ is constant $\rightarrow V(B_0 + B_1 X_i) = 0$

$V(\varepsilon_i) = \sigma^2$

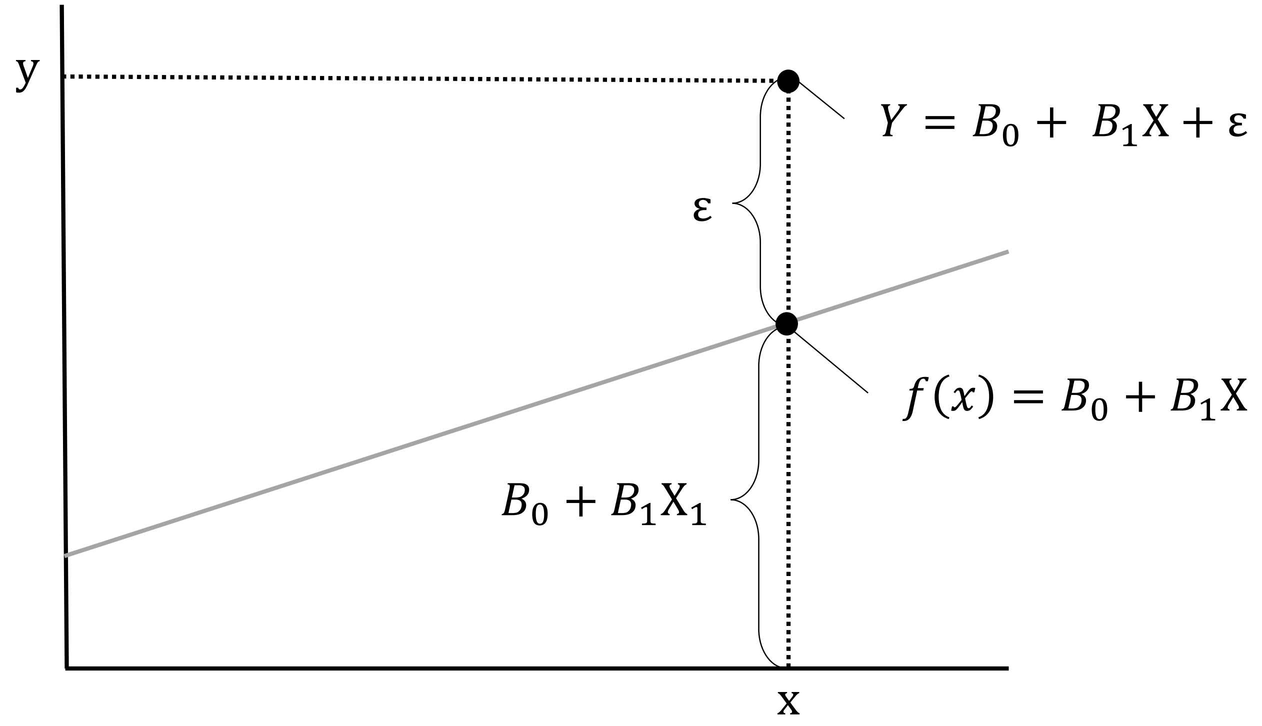
i.e., $\boldsymbol{Y_i \sim N\left(B_0 + B_1 X_i, \sigma^2\right) \quad i = 1, 2, \cdots, n}$

Assumption of Linear Regression Model



i.e., $\boldsymbol{Y_i \sim N\left(B_0 + B_1 X_i, \sigma^2\right) \quad i = 1, 2, \cdots, n}$

Linear Regression Model



$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p + \varepsilon$$

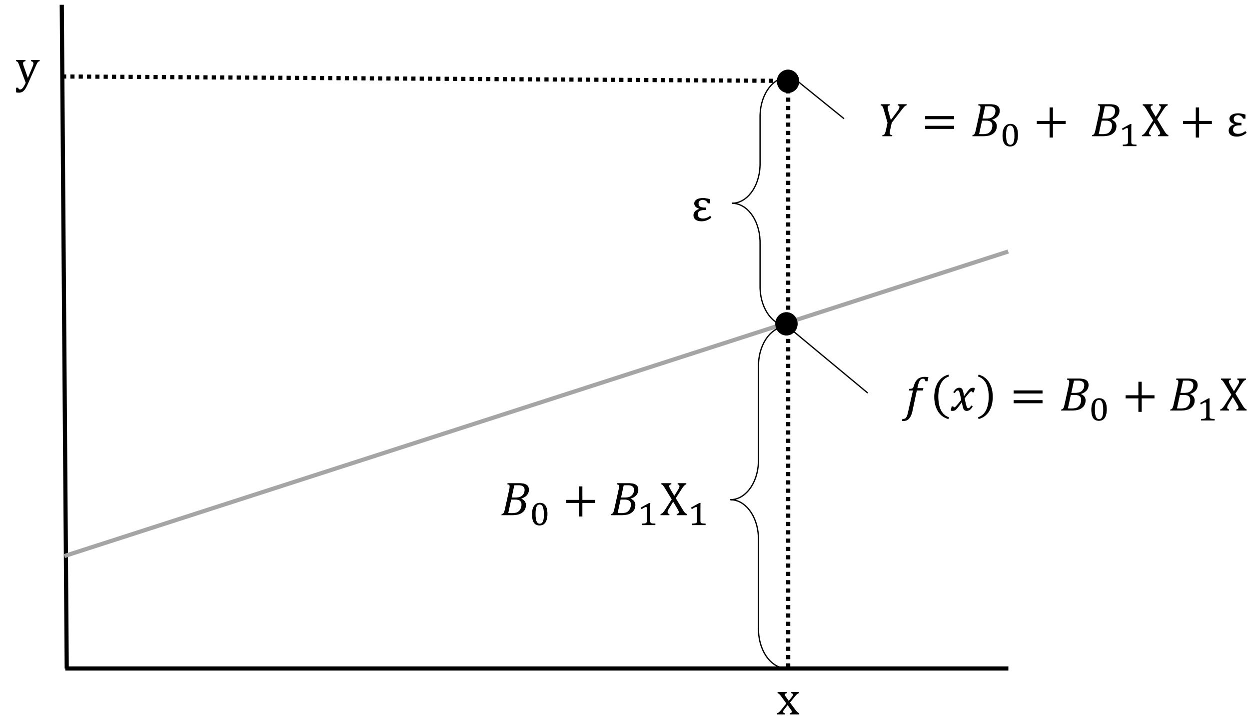
$$\mathrm{E(Y)} = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$$

View Point.

Find a linear regression line that describes **the relationship** between the input variable(X) and the mean of output variable(Y)

i.e., **Find Parameter**$(B_0,\ B_1, .., \ B_p)$ using the function of data

Linear Regression Model



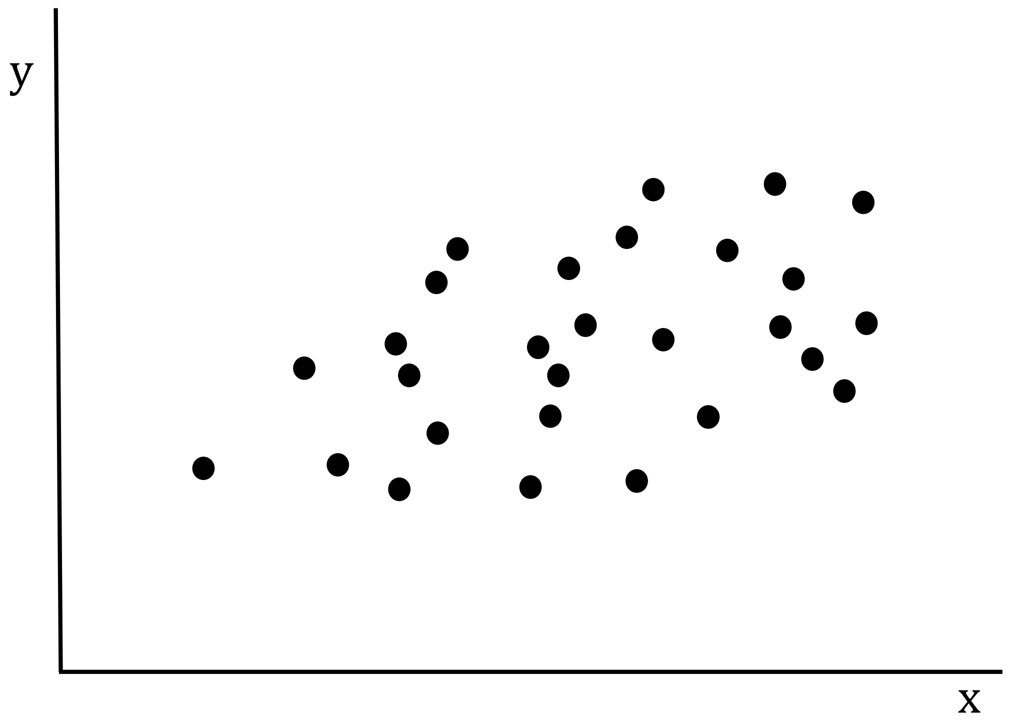
$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p + \varepsilon$$

$$E(Y) = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$$

View Point.

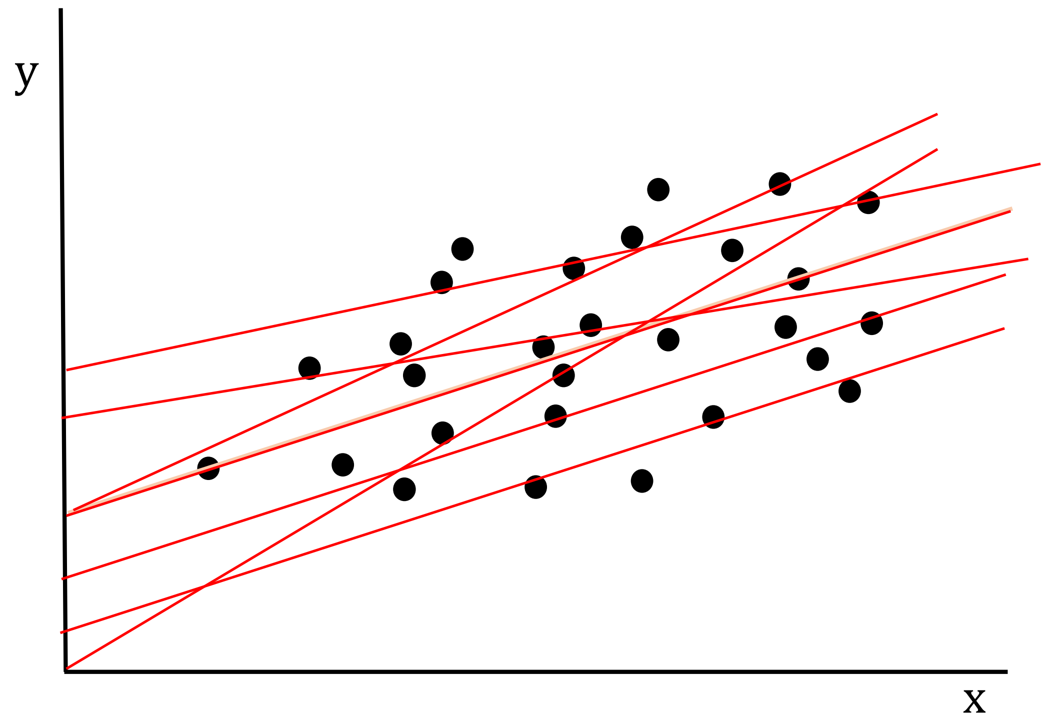
Find a linear regression line that describes **the relationship** between the input variable(X) and the mean of output variable(Y)

Linear Regression Model



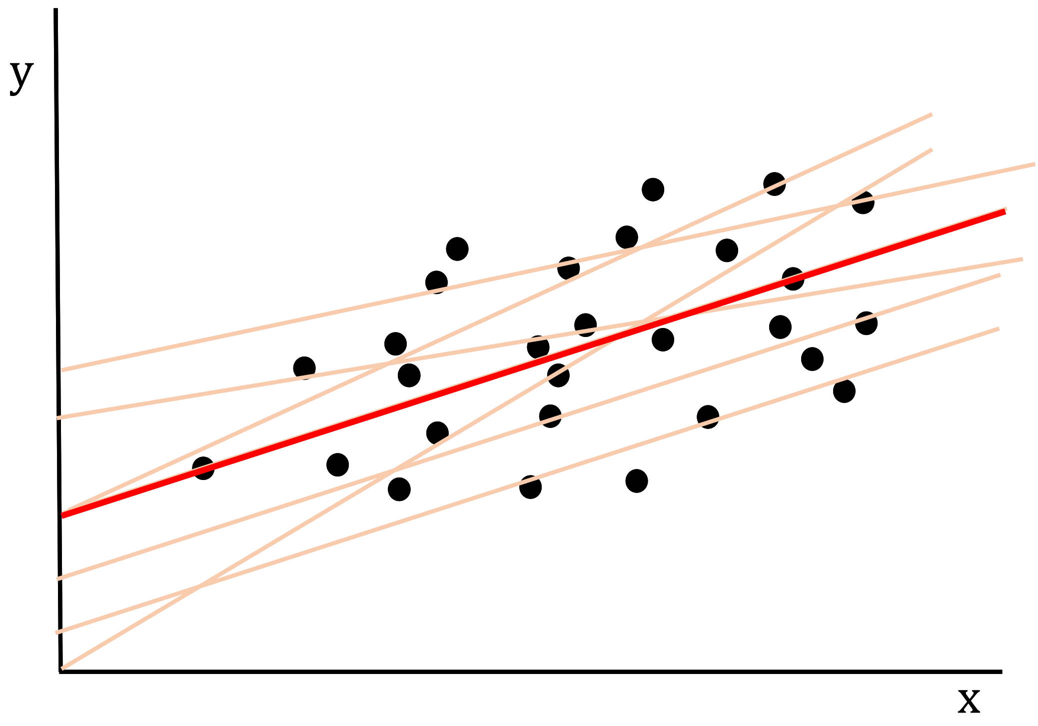
$$E(Y) = f(X) = B_0 + B_1 X$$

Linear Regression Model



$$E(Y) = f(X) = B_0 + B_1 X$$

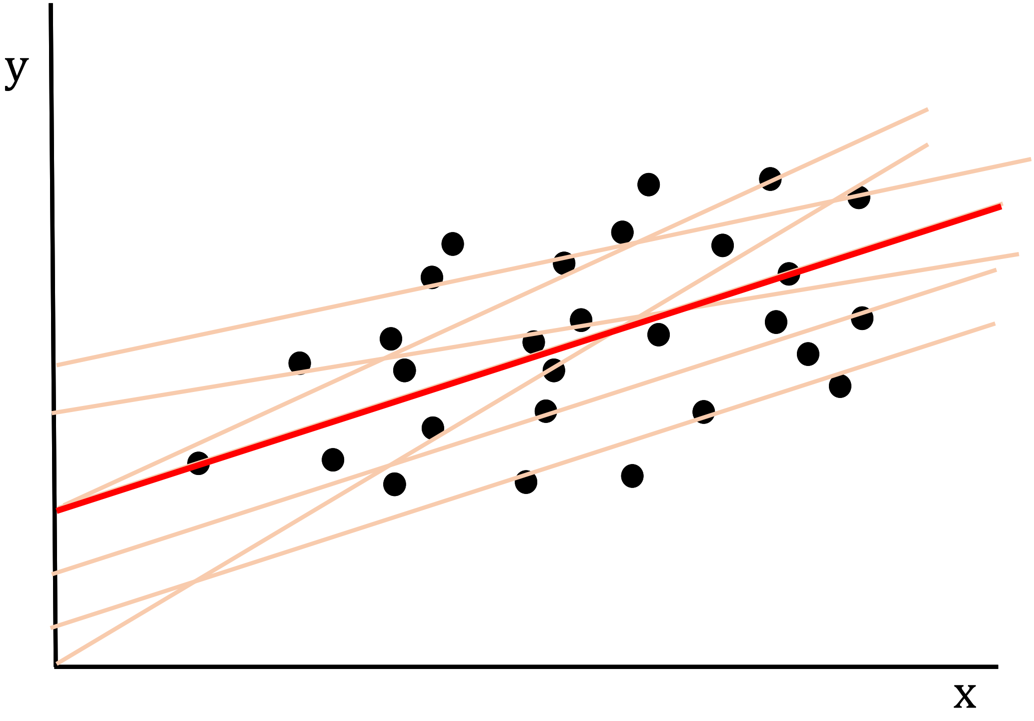
Linear Regression Model



$$E(Y) = f(X) = B_0 + B_1 X$$

Find Best Parameter $(B_0, B_1, .., B_p)$ using data

Linear Regression Model



$E(Y) = f(X) = B_0 + B_1 X$

Find Best Parameter $(B_0,\ B_1, .., \ B_p)$ using data

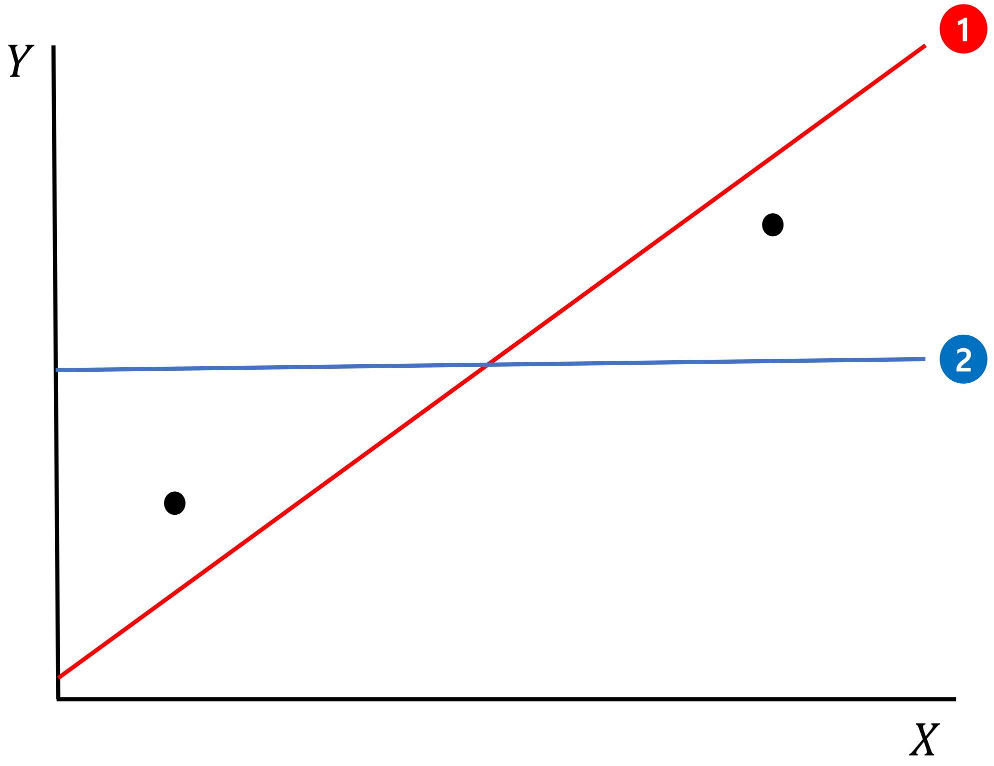
How to find good parameter?

Chapter Ⅱ

- Parameter Estimation -

Parameter Estimation

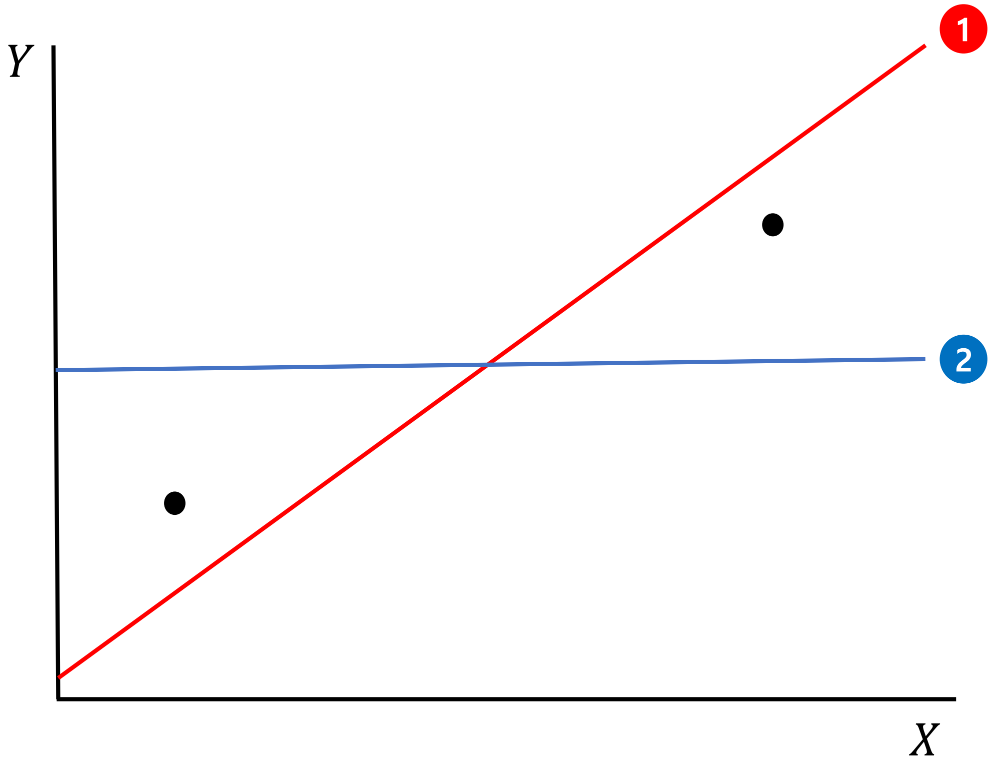
Question. Let's compare with red and blue. Which one is correct prediction line?



Parameter Estimation

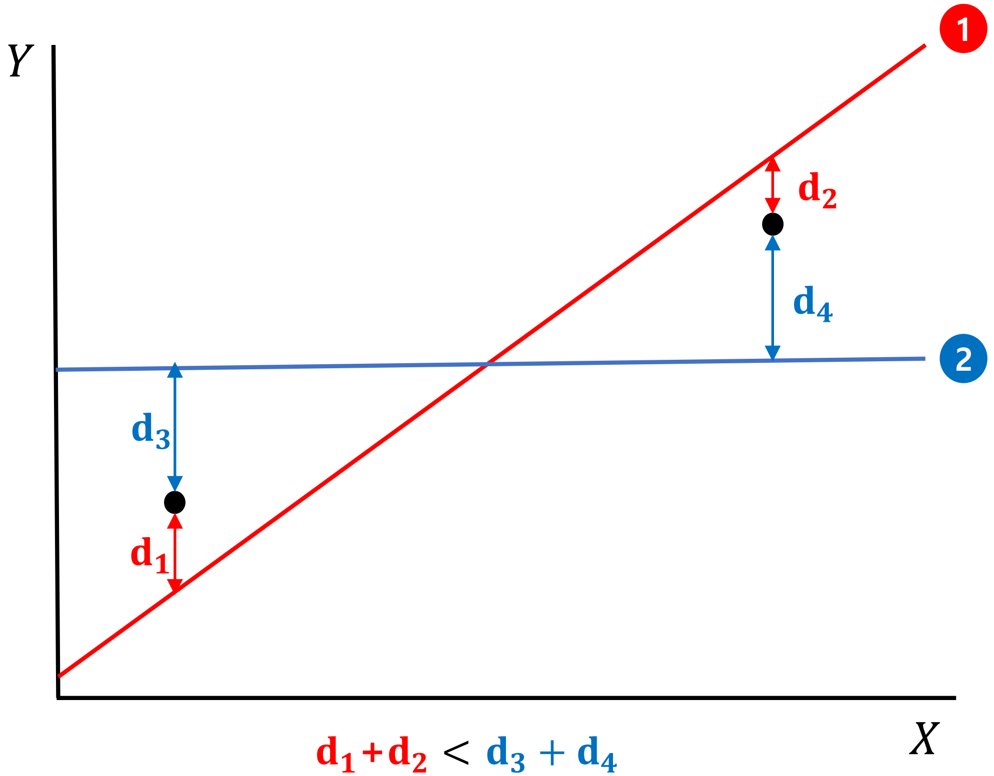
Question. Let's compare with red and blue. Which one is correct prediction line?

red !!!



Parameter Estimation

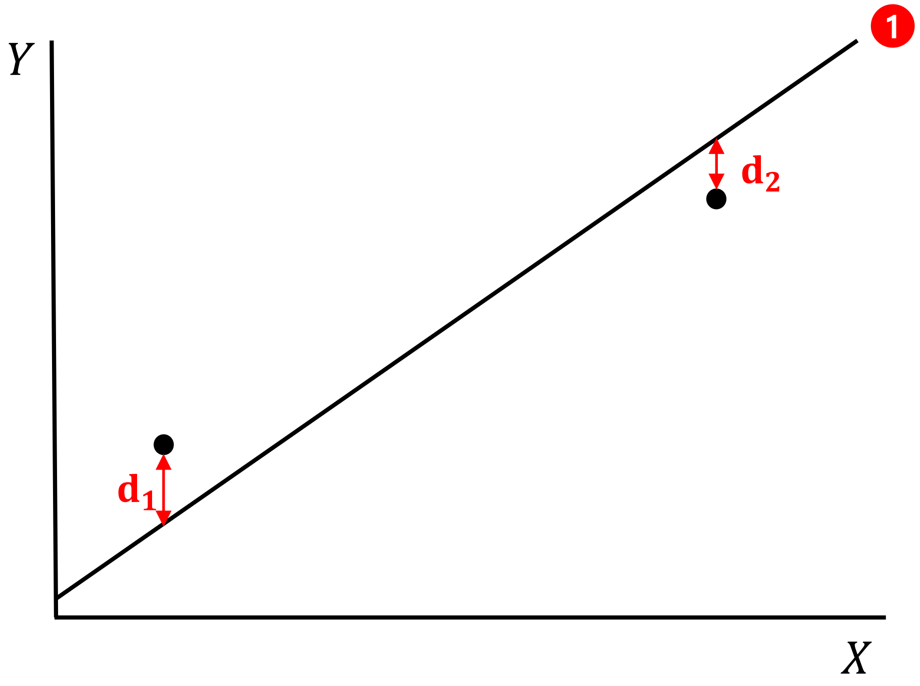
Question. Let's compare with red and blue. Which one is correct prediction line?



$d_1 + d_2 < d_3 + d_4$

Answer. Red is a better regression line than blue

Parameter Estimation



$$d_1 + d_2 + \cdots + d_n = 0$$
$$d_1^2 + d_2^2 + \cdots + d_n^2 \geq 0$$

$$\begin{aligned} d_1 &= Y_1 - E(Y_1) \\ &= Y_1 - (B_0 + B_1 X_1) \end{aligned}$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{Y_i - (B_0 + B_1 X_i)\}^2 \longleftarrow \text{Cost Function}$$

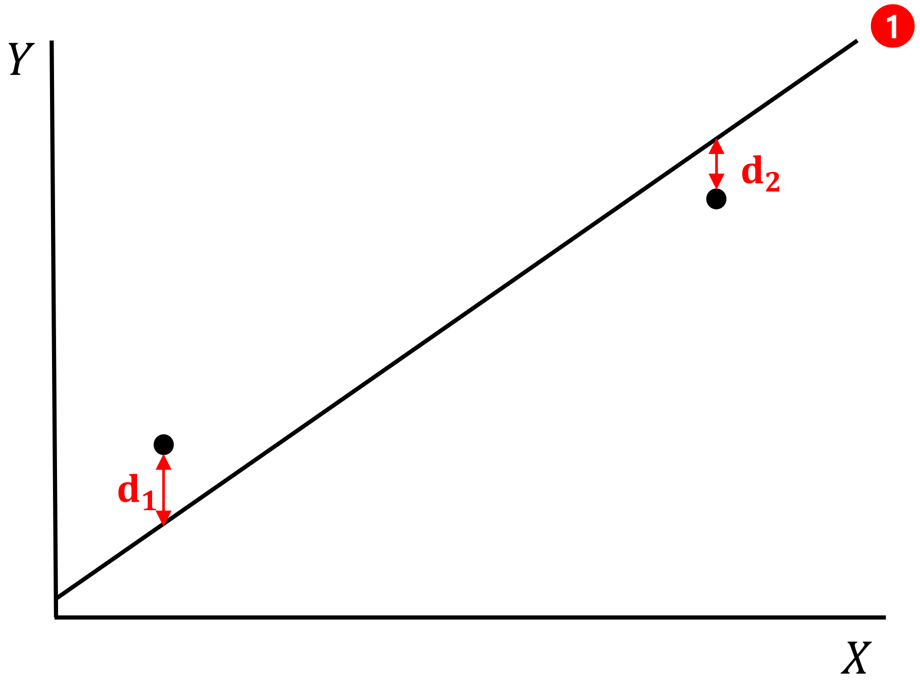
i.e., Finding the smallest Cost function is finding the best parameters !!!

$$\min_{B_0, B_1} \sum_{i=1}^{n} \{Y_i - (B_0 + B_1 X_i)\}^2$$

※ Cost Function vs Lost Function

Lost Function: loss of a data vs Cost Function: Sum of loss functions for all data

Parameter Estimation



$$\mathrm{d}_1 + \mathrm{d}_2 + \cdots + \mathrm{d}_n = 0$$
$$d_1^2 + d_2^2 + \cdots + d_n^2 \geq 0$$

$$\mathrm{d}_1 = \mathrm{Y}_1 - E(\mathrm{Y}_1)$$
$$= \mathrm{Y}_1 - (\mathrm{B}_0 + \mathrm{B}_1\mathrm{X}_1)$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{\mathrm{Y}_i - (\mathrm{B}_0 + \mathrm{B}_1\mathrm{X}_i)\}^2 \longleftarrow \text{Cost Function}$$

i.e., Finding the smallest Cost function is finding the best parameters !!!

$$\min_{\mathrm{B}_0,\mathrm{B}_1} \sum_{i=1}^{n} \{\mathrm{Y}_i - (\mathrm{B}_0 + \mathrm{B}_1\mathrm{X}_i)\}^2$$

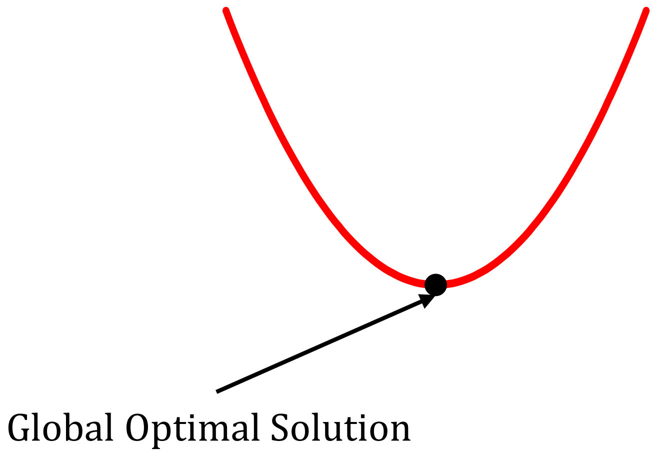
※Cost Function vs Lost Function

How to find the smallest Cost function?

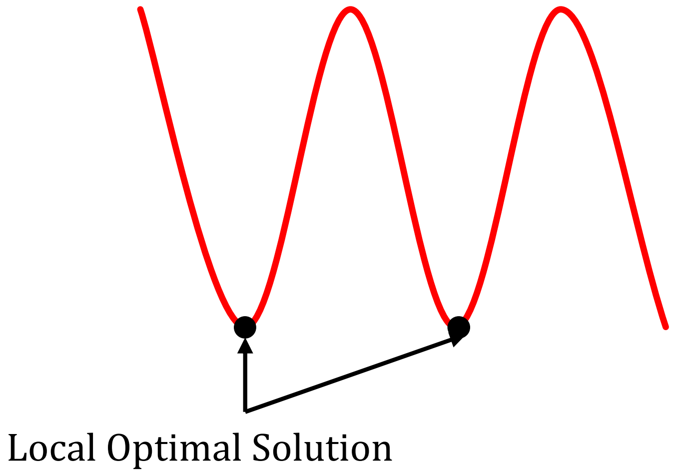
Lost Function: loss of a data vs Cost Function: Sum of loss functions for all data

Parameter Estimation

In linear regression, Cost Function is always **convex** = globally optional solution exists



Convex Function



Non-Convex Function

i.e., The way that finds the smallest cost function(estimates best parameter) is
Finding a point where the derivative is 0

Parameter Estimation

- Partial derivative based on Parameter(B_1, B_0)

(B_1: gradient, , B_0: y-intercept)

C*ost Function:* $\sum_{i=1}^{n}\{Y_i - (B_0 + B_1X_i)\}^2$

B_0 partial derivative $\rightarrow \dfrac{\partial C(B_0, B_1)}{\partial B_0} = -2\sum_{i=1}^{n} Y_i - (B_0 + B_1X_i) = 0$

B_1 partial derivative $\rightarrow \dfrac{\partial C(B_0, B_1)}{\partial B_1} = -2\sum_{i=1}^{n} Y_i - (B_0 + B_1X_i)X_i = 0$

The result of partial derivative

$$\hat{B}_0 = \bar{Y} - \hat{B}_0\,\bar{X}$$

$$\hat{B}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The linear regression function that has best parameter

$$\boldsymbol{f(x) = \hat{Y} = \hat{B}_0 + \hat{B}_1 X}$$

Least Squares Estimation Algorithm

Goal. Find estimator of B_0 and B_1 (i.e., $\hat{B}_0$ and $\hat{B}_1$)

Step1. Cost Function(Squared the sum of the difference between the actual y value and y value on the regression line)

$$\sum_{i=1}^{n}\{Y_i - (B_0 + B_1X_i)\}^2$$

Step2. Find B_0, B_1 to minimize Cost Function

$$\min_{B_0,B_1}\sum_{i=1}^{n}\{Y_i - (B_0 + B_1X_i)\}^2$$

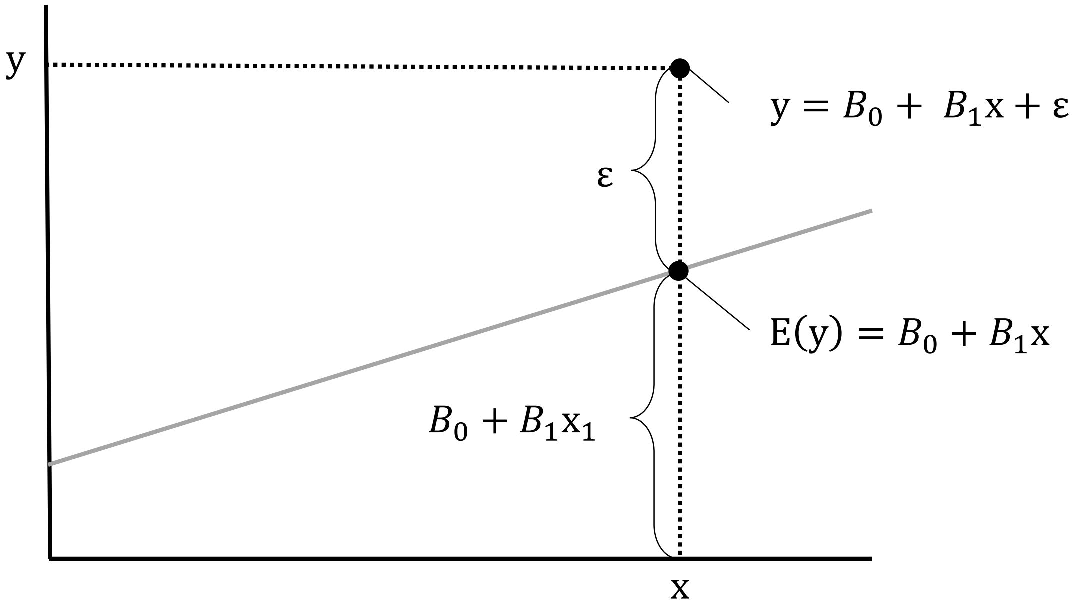
Step3. Find the point where the derivative(gradient) is 0

$$\frac{\partial C(B_0, B_1)}{\partial B_0} = -2\sum_{i=1}^{n} Y_i - (B_0 + B_1X_i) = 0$$

$$\frac{\partial C(B_0, B_1)}{\partial B_1} = -2\sum_{i=1}^{n} Y_i - (B_0 + B_1X_i)X_i = 0$$

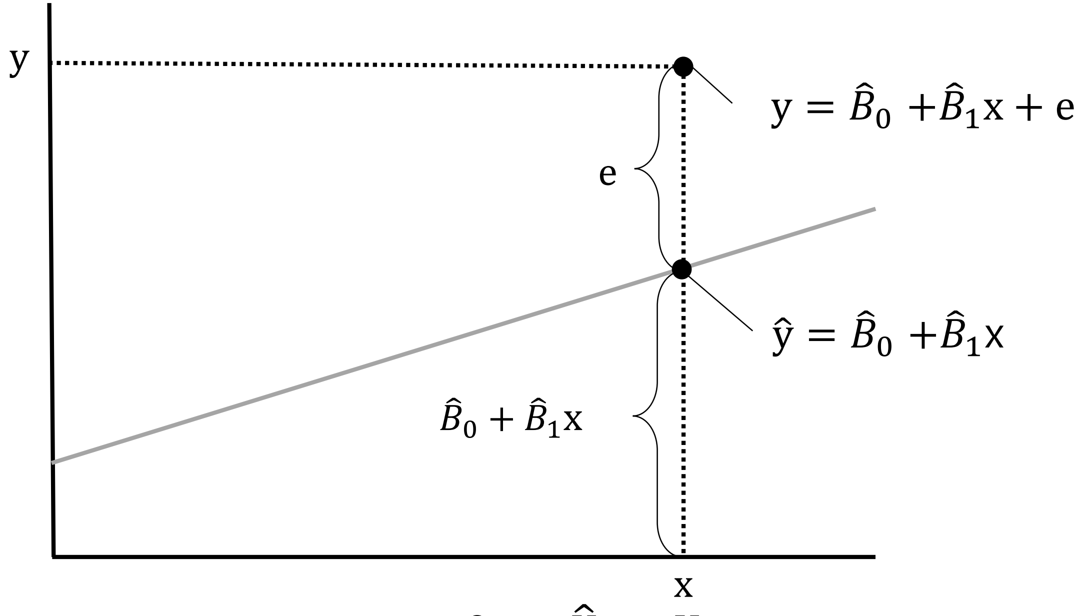
Solutions. $\hat{B}_0 = \bar{Y} - \hat{B}_1\,\bar{X},\ \ \hat{B}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

Residual



$$\varepsilon = y \ - \ E(y)$$

B_0, B_1 is not fixed value, just status of parameter
ε follows normal distribution



$$e \ = \ \hat{y} \ - \ y$$

$\hat{B}_0, \ \hat{B}_1$ is fixed value
e is error of fixed values (constant)

e(residual) = the value that ε(random error) is actually implemented

Chapter Ⅲ

- Parameter Inference -

Parameter inference

- There are two ways of infer parameters

1. Estimator

2. Hypothesis test

Estimator of parameter

● Estimators($\hat{B}_0, \hat{B}_1$) that calculated by using Least Squared Estimation Algorithm

$$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X}, \qquad \hat{B}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

▶ Estimator : a function of the sample(data)
$\hat{B}_0, \hat{B}_1$

▶ Usage of Estimator: estimate unknown parameter(B_0, B_1)

▶ Types of Estimator – Point Estimator
– Interval Estimator

Point estimator of parameter

$Y_i = B_0 + B_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \quad i = 1,2,\cdots,n$

1) Point Estimator of B_0: $\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X}$
2) Point Estimator of B_1: $\hat{B}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
3) Point Estimator of σ^2: $\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} {e_i}^2 \quad (n = number\ of\ samples,\ e = residual)$

Gauss-Markov Theorem: Least Square Estimator is the Best Linear Unbiased Estimator (BLUE)

BLUE : The BLUE is (1)unbiased estimator and (2)has the smallest average squared error(variance) compared to any unbiased estimators.

(1) unbiased estimator : $E(\hat{B}_0) = B_0,\ E(\hat{B}_1) = B_1$
(2) smallest variance estimator : $V(a\hat{B}_0) \leq V(b\hat{\theta})$, $V(a\hat{B}_1) \leq V(b\hat{\theta})$ $\quad \hat{\theta}$: any other unbiased estimate

Interval estimator of parameter

- Con(s) of interval estimate
  → Estimate intervals to provide more flexible information

- Basic form that interval estimator of $\theta(parameter)$

$$\hat{\theta} - C * \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + C * \sigma(\hat{\theta}) \qquad \hat{\theta} : point\ estimator\ of\ \theta$$

i.e., have to know parameter(**point estimator**, **constance**, **standard deviation**)

1) Confidence interval for gradient(B_1) → $(100(1-a)\%)$

$$\hat{B}_1 - t_{\frac{a}{2},n-2} sd(\hat{B}_1) \leq \hat{B}_1 \leq \hat{B}_1 + t_{\frac{a}{2},n-2} sd(\hat{B}_1)$$

2) Confidence interval for y-interval(B_0)

Same form as confidence interval for B_1

1) $\hat{B}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$: $point\ estimator\ of\ B_1$

2) $t_{\frac{a}{2},n-2}$: The value of the t-distribution with a degree of freedom of n-2 under the significance level (1-a)

3) $sd(\hat{B}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X-\bar{X})^2}}$: $standard\ deviation\ of\ \hat{B}_1$

Hypothesis test for gradient($\boldsymbol{B_1}$)

What is hypothesis test? Hypothesis and test for unknown parameters

● hypothesis test

$H_0: B_1 = 0 \quad vs \quad H_1: B_1 \neq 0$ (H_0: Null Hypothesis, H_1: Alternative Hypothesis)

* If B_1(gradient)=0, There is no relationship between X and Y

$$t^* = \frac{\widehat{B}_0 - 0}{sd(\hat{B}_1)} \leftarrow \ test\ statistic\ for\ null\ hypothesis$$

$(\widehat{B}_0: made\ of\ data,\ \ 0: made\ of\ hypothesis,\ sd(\hat{B}_1): use\ for\ scaling)$

Prove hypothesis test by one of the two methods

1) $IF\ |t^*| > t_{\frac{\alpha}{2}, n-2} \rightarrow we\ reject\ H_0$
2) $\text{p}-\text{value} = 2P(T > |t^*|)\ where\ T \sim t(n-2)$

Generally, if p-value is less than 0.05 or 0.01, the null hypothesis is rejected

Example (Regression analysis)

The regression equation ➡ Y(Appraised value) = -29.6 + 0.0779X(Area)

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.00

S = 16.9065

Q1. What are point estimates of the parameters?

=> $\hat{B}_0$ = -29.56, $\hat{B}_1 = 0.077939$

Q2. What is the standard deviation(standard error) of the parameter?

$$=> sd(\hat{B}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} = 10.66$$

$$sd(\hat{B}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = 0.004370$$

Example (Regression analysis)

The regression equation ➡ Y(Appraised value) = -29.6 + 0.0779X(Area)

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.00

S = 16.9065

Q3. What is the T in the above table?

=> $H_0: B_1 = 0 \quad vs \quad H_1: B_1 \neq 0$

$$\text{T} = t^* = \frac{\widehat{B}_0 - 0}{sd(\hat{B}_1)} = \frac{0.077939 - 0}{0.004370} = 17.83$$

Q4. What is the P in the above table?

=> $\text{p-value} = 2P(T > |t^*|) = 2P(T > |17.83|)\ where\ T \sim t(13)_{(n\,=\,15\ \rightarrow\ n-2\,=\,13)} = 0.00$

➡ H_0 is rejected , $H_1 \neq 0$ i.e., X(Area) has significant effect on Y(Appraised value)

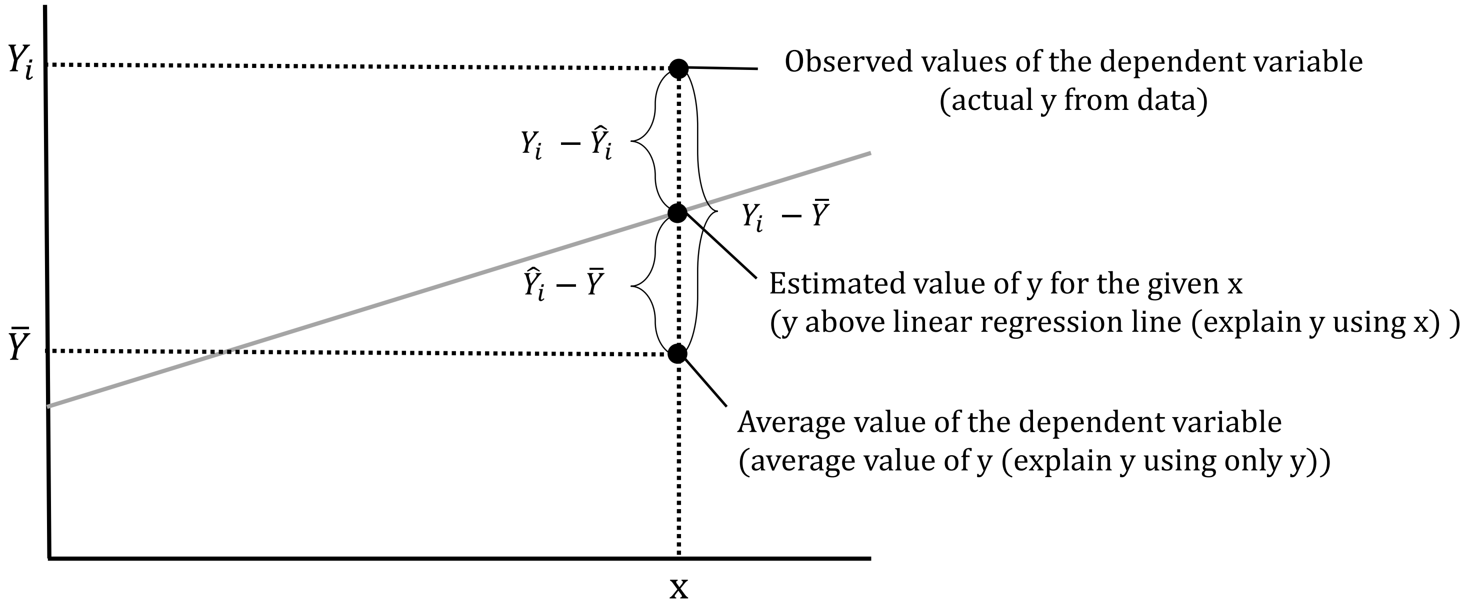
Q5. What is the S in the above table?

=> $\text{S} = \hat{\sigma} = \sqrt{\left(\frac{1}{n-2}\right) \sum_{i=1}^{n} {e_i}^2} = 16.9065$

Chapter IV

- Coefficient of Determination & ANOVA-

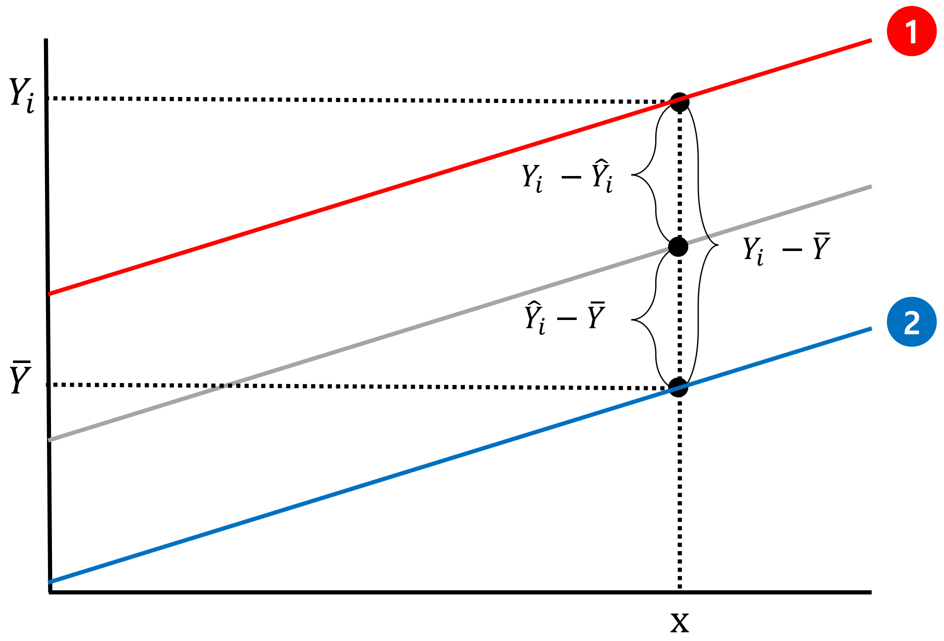
Coefficient of Determination: $\boldsymbol{R^2}$



SSE(Sum of Square Error) $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
SSR(Sum of Square Regression) $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
SST(Sum of Square Total) $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$\rightarrow \mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$

Coefficient of Determination: R^2



$$\text{Coefficient of Determination}(R^2) = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (SST = SSE + SSR)$$

1 $\frac{SSR}{SST} = 1 \rightarrow \begin{matrix} SSE = 0 \\ SSR = SST \end{matrix}$ ➡ There is no error, Completely same

2 $\frac{SSR}{SST} = 0 \rightarrow \begin{matrix} SSR = 0 \\ SSE = SST \end{matrix}$ ➡ Average of y = use x (above linear regression line)

Coefficient of Determination: R^2

- Property of R^2

1. $0 \leq R^2 \leq 1$
2. $R^2=1$: X variable can explain 100% of Y.
   i.e., all data are above the regression line
3. $R^2=0$: X variable can't explain Y
   i.e., X variable does not help description(prediction) of Y at all
4. How much the X variable in use reduced the variance of the Y variable
5. The degree of performance improvement gained by using X information compared to simply using Y average value
6. Quality of X Variables in Use

But, R^2 always increases even if non-significant variable is added

➡ (Adding non-significant variable to y $\rightarrow$ SSE value decreases $\rightarrow$ R^2 increases)

Adjusted Coefficient of Determination ($\boldsymbol{R^2}_{\boldsymbol{adj}}$)

- Adjusted R^2

$$\mathrm{R^2}_{\mathrm{adj}} = 1 - \left[\frac{\mathrm{n}-1}{\mathrm{n}-(\mathrm{p}+1)}\right]\frac{\mathrm{SSE}}{\mathrm{SST}} \quad (\mathrm{n} = \text{number of data}, \mathrm{p} = \text{ number of variable})$$

- Property of Adjusted R^2

1. Adjusted R^2 is multiplied by a particular coefficient,
so that when a non-significant variable is added, it does not increase

➡ Adding a non-significant variable to y $\rightarrow$ value of p increases
$\rightarrow$ the denominator of a particular constant increases $\rightarrow$ Adjusted R^2 decreases

➡ Adding a significant variable to y $\rightarrow$ SSE decreases

2. Use to compare explanatory power of regression models with different explanatory variables

Example (R^2)

Q. How does the number of salespeople and advertising costs of each store affect sales?

Variable	Estimate	T	P-Value
Constant	141.516	0.706	0.472
The number of salespeople (X_1)	13.035	1.854	0.106
Advertising costs (X_2)	14.469	3.025	0.019

$$\mathrm{SSR} = 54809.18, \quad \mathrm{SSE} = 25440.82, \quad \mathrm{SST} = 80250.00$$

A. $\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{54809.18}{25440.82} = 0.683$

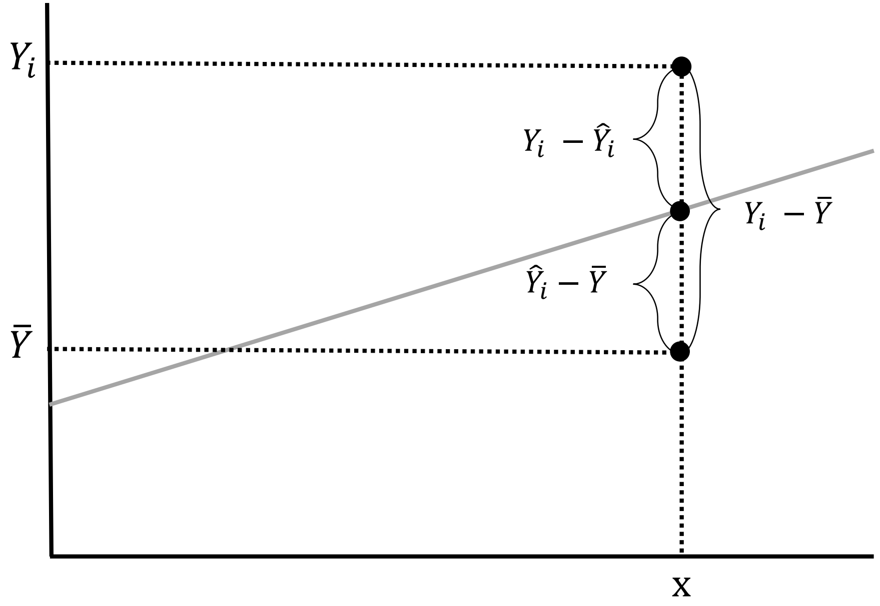
1. The number of salespeople and advertising cost variables reduced the volatility of the sales variable by 68.3%
2. Using the number of salespeople and advertising costs compared to the (simple) average of sales increases explanatory power by 68.3%
3. The degree of "variable quality" of the number of salespeople and advertising costs is 68.3 (based on 100)

Analysis of Variance(ANOVA) in Linear Regression Model

● Analysis of Variance(ANOVA) in Linear Regression Model

1. analysis by using variance

2. Ultimately used for hypothesis testing



variation

- SST $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$: total amount of variation in Y
- SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$: amount described by the X
- SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$: amount described by the Error

Analysis of Variance(ANOVA) in Linear Regression Model

$\frac{SSR}{SSE}$: Fractions to see how large the SSR is compared to SSE

$\frac{SSR}{SSE} > 1$

- amount described by the X $>$ amount described by the Error
- X variable has significant effect on description(prediction) of Y variable
- The coefficient of the X variable(gradient) is not 0

$0 \leq \frac{SSR}{SSE} \leq 1$

- amount described by the X $<$ amount described by the Error
- X variable has non-significant effect on Y variable
- Statistically, the coefficient of the X variable(gradient) is 0

Analysis of Variance(ANOVA) in Linear Regression Model

Question. $In\ \frac{SSR}{SSE} > 1\ case,$ how can judge it is big ?

Answer. If we know the distribution, we can judge statistically. However, the distribution cannot be defined directly
But, SSE, SSR follows Chi-Square Distribution(Parameter : degree of freedom)

Let $Y_1\ be\ \chi^2(v_1)\ and\ Y_2\ be\ \chi^2(v_2),\ \ define\ F = \frac{Y_1/v_1}{Y_2/v_2}$

F has an F-distribution with v_1 degrees of freedom in the numerator and v_2 degrees of freedom in the denominator, denoted as F(v_1, v_2)

In case of simple linear regression,

SSR ~ $\chi^2(v_1 = 1)$, SSE ~ $\chi^2(v_1 = n - 2)$

$$F^* = \frac{SSR/1}{SSE/n-2} \sim F(1, n-2)$$

ANOVA Table

Source	DF	SS	MS	F	P
Model Error	1 n-2	SSR SSE	MSR MSE	F^*	P-Value
Total	n-1	SST			

$$H_0: B_1 = 0 \quad vs \quad H_1: B_1 \neq 0$$

$$F^* = \frac{SSR/_1}{SSE/_{n-2}} = \frac{MSR}{MSE} \sim F(1, n-2)$$

$$\text{p−value} = P(\mathrm{Y} > F^*) \; where \; Y \sim F(1, n-2)$$

If F^*value is large(MSR is relatively enough large than MSE), H_0 is rejected

F^*value(test statistic) is large → The probability that the T value is greater than the F^*value is less → p-value is small → Reject the null hypothesis(H_0)

Example (ANOVA)

Source	DF	SS	MS	F	P
Model Error	2 7	54809.18 25440.82	27404.59 3634.40	$F7.540^*$	0
Total	9	80250.00			

$$H_0: B_1 = B_2 = 0 \quad vs \quad H_1: At\ least\ one\ B \neq 0$$

$$F^* = \frac{MSR}{MSE} = \frac{54809.18/_2}{25440.82/_7} = \frac{27404.59}{3634.40} = 7.540$$

$$\text{p}-\text{value} = P(Y > 7.540) \approx 0, where\ Y \sim F(2,7)$$

$At\ least\ one\ B \neq 0$ (The number of salespeople or advertising costs or both are significant)

Thank you