Signal & Noise I
Regression

D:plus

목차

D:plus

1. Signal & Noise

1. Signal & Noise

복습할 겸! 다시 고민하자!

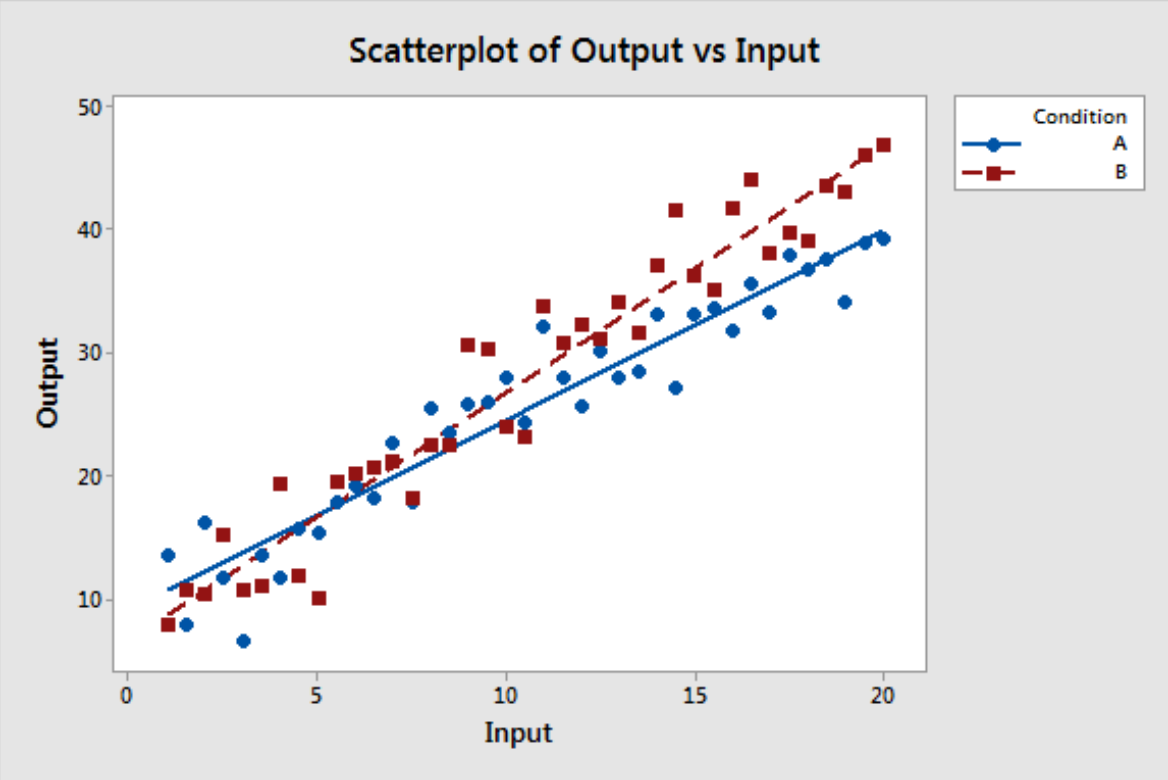
	A	B	C	D
1	Quantity Sold	Price	Advertising	
2	8500	$2	$2,800	
3	4700	$5	$200	
4	5800	$3	$400	
5	7400	$2	$500	
6	6200	$5	$3,200	
7	7300	$3	$1,800	
8	5600	$4	$900	
9				

끌어낼 수 있는 이야기는 뭐가 있을까?

D:plus

1. Signal & Noise

복습할 겸! 다시 고민하자!



끌어낼 수 있는 이야기는 뭐가 있을까?

D:plus

1. Signal & Noise

복습할 겸! 다시 고민하자!

Contains "Money"	Domain type	Has attach.	Time received	spam
yes	com	yes	night	yes
yes	edu	no	night	yes
no	com	yes	night	yes
no	edu	no	day	no
no	com	no	day	no
yes	cat	no	day	yes

끌어낼 수 있는 이야기는 뭐가 있을까?

1. Signal & Noise

‘무언가’의 결과에게 기대하는 것.

Descriptive

VS

Predictive

D:plus

1. Signal & Noise

모델링을 하기 전, 고민 / 점검 해둬야 하는 부분

데이터를 먼저 확인해야 한다!

$$Y = f(X) + \epsilon$$

반드시 알아야 하는 / 극복해야 하는 지점

우리 데이터(X)는 우리 관심사(Y)를 설명하기에 충분한가?

충분하다 : 1. 많다. 2. 다양하다. **3. 적절하다.**

D:plus

1. Signal & Noise

모델링을 하면서 고민해야 하는 부분!

$$Y = f(X) + \epsilon \qquad \hat{Y} = \hat{f}(X)$$

반드시 알아야 하는 / 극복해야 하는 지점

우리의 모델은 실제 Signal을 잡아내기에 충분한가?

충분하다 : 1. 오차가 작다. 2. 정확하다. 3. 설명하기 편하다.

D:plus

1. Signal & Noise

모델링의 기본 생각.

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

실제 신호와의 오차를 최대한 줄이고자 하는 것!

D:plus

1. Signal & Noise

모델링의 결과를 바라보는 법

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

잡아낸 신호와 오차의 크기는?

오차를 바라보는 여러 방식들이 필요!

D:plus

1. Signal & Noise

머신러닝에서 러닝이란?

사실상, 수학에서 말하는 최적화 문제!

1. 모델구조를 설계한다.
2. 모델을 평가할 지표를 설계한다. (에러에 관련해서)
3. 데이터를 모델에 넣어 예측 값을 만들어 낸다.
4. 예측값과 실제값을 이용해 지표를 계산한다.
5. 지표를 보고 모델에 피드백한다.

D:plus

2. Regression [Analysis]

2. Regression [Analysis]

엄격하게 시작해봅시다.

$$Y = f(X) + \epsilon \qquad \hat{Y} = \hat{f}(X)$$

가정/가설로 시작 : X와 Y는 간단한 선형적인 관계에 있지 않을까.
-> 관계(f)가 Y=aX+b 같은 형태를 띄고 있을 거야!

Q1. 관계가 선형적이라면 무슨 장점이 있을까?

Q2. 선형적인 관계를 자연스럽게 가정하려면 어떤 상황일까?

Q3. 자연스러운 상황이 없다면, 가정해볼 수 없는 걸까?

D:plus

2. Regression [Analysis]

회귀 분석! : 손으로 적어보자.

선형 회귀 (수식)

$$Y = f(X) + \epsilon$$

$$y_i = w_1 X_{i1} + w_2 X_{i2} + \cdots + w_p X_{ip} + w_0 + e_i$$

$$\hat{Y} = \hat{f}(X)$$

선형 회귀 (그래프)

2. Regression [Analysis]

다 끝나고 와서, 아래의 내용들 중, 굵직한 것들을 이해할 수 있어야 함!

```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

D:plus

2. Regression [Analysis]

다 끝나고 와서, 아래의 내용들 중, 굵직한 것들을 이해할 수 있어야 함!

```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

Coefficient 하나 하나에 대해서, 가설 검정을 한다.

무슨 의미가 될까?

$$H_0: \beta_0 = \beta_{0,0}$$
$$H_1: \beta_0 \neq \beta_{0,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

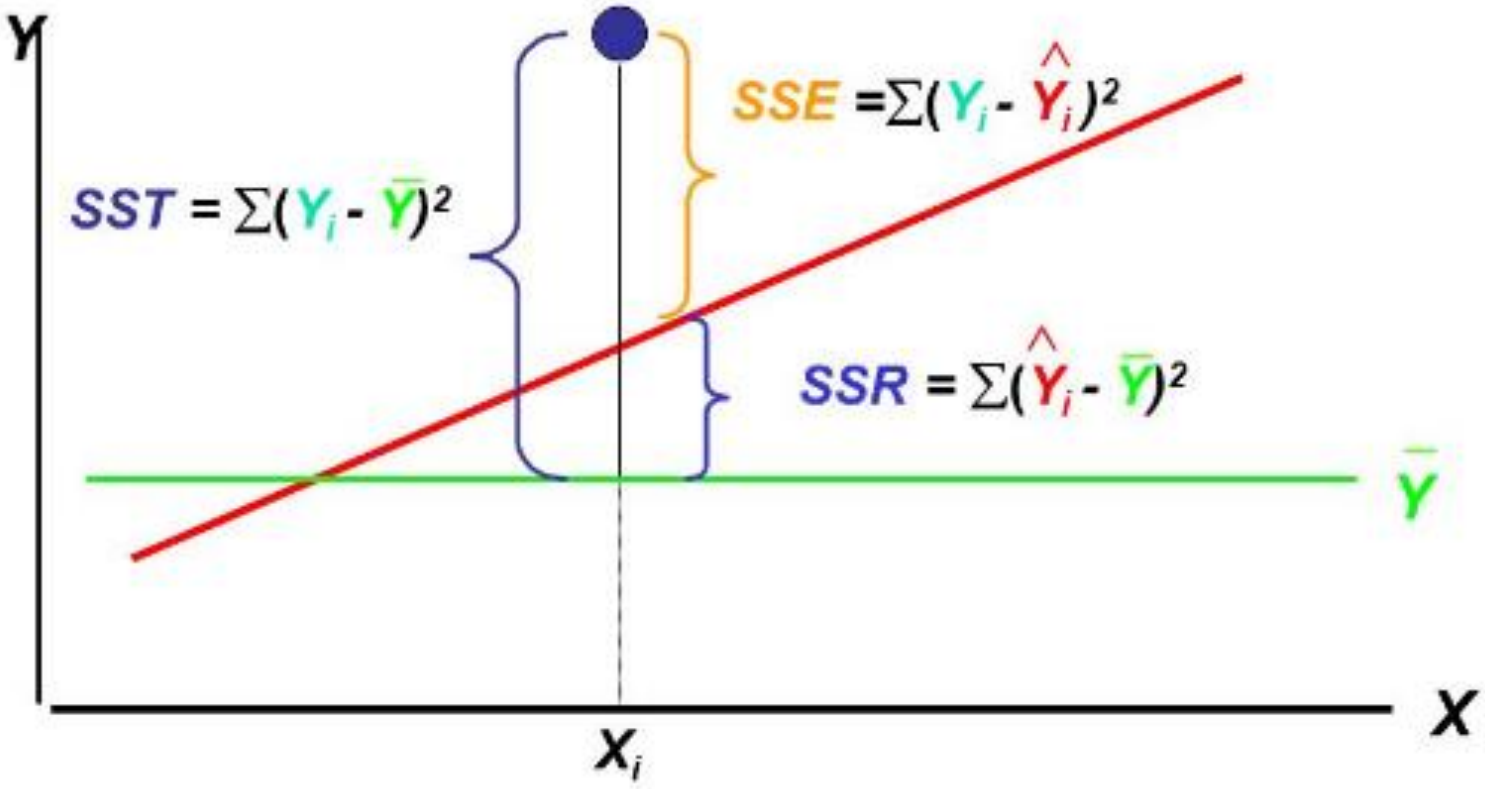
D:plus

2. Regression [Analysis]

회귀분석에서 에러를 바라보는 첫 번째 관점

$$Y = f(X) + \epsilon$$

$$\hat{Y} = \hat{f}(X)$$



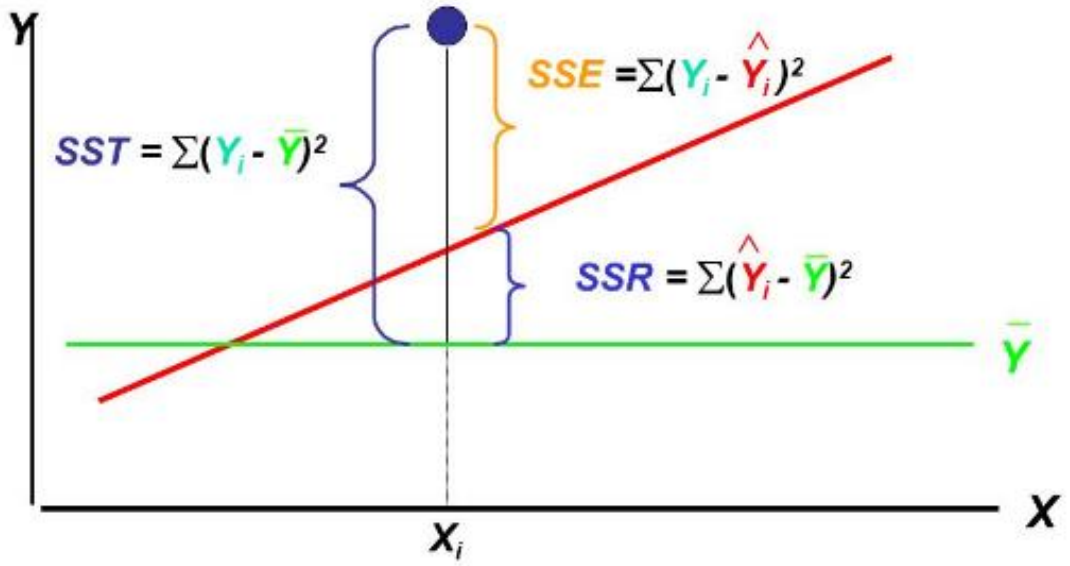
D:plus

2. Regression [Analysis]

회귀분석에서 에러를 바라보는 첫 번째 관점

$$Y = f(X) + \epsilon$$

$$\hat{Y} = \hat{f}(X)$$



$$SST = SSR + SSE$$

Total Sample Variability = **Explained Variability** + **Unexplained Variability**

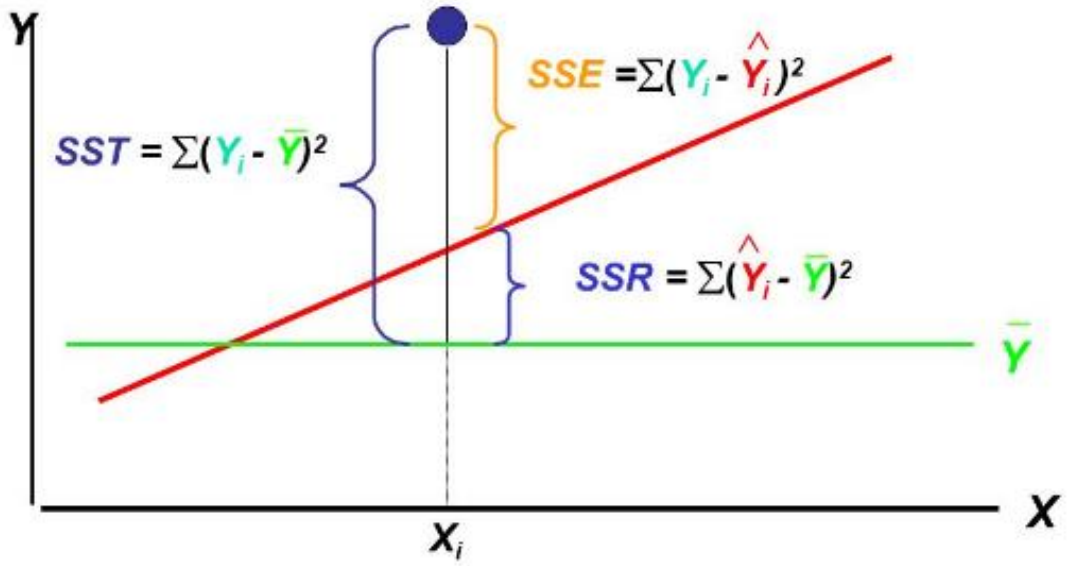
D:plus

2. Regression [Analysis]

회귀분석에서 에러를 바라보는 첫 번째 관점

$$Y = f(X) + \epsilon$$

$$\hat{Y} = \hat{f}(X)$$



SST :

SSE :

SSR :

D:plus

2. Regression [Analysis]

참고만!

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

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

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

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 + \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \boxed{\sum_{i=1}^{n}(y_i - \hat{y})(\hat{y} - \bar{y})}$$

$$\underbrace{\boxed{\sum_{i=1}^{n}(y_i - \hat{y})^2}}_{\text{SSE}} + \underbrace{\boxed{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}}_{\text{SSR}}$$

=0, as imposed in the estimation, E(εx)=0.

D:plus

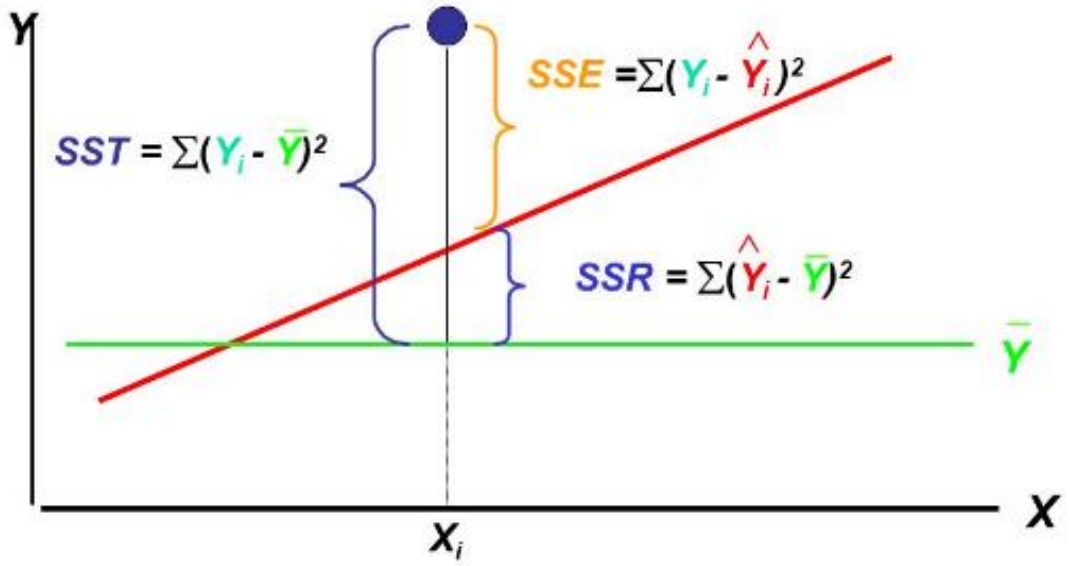
2. Regression [Analysis]

그래서 MSE!

$$Y = f(X) + \epsilon$$

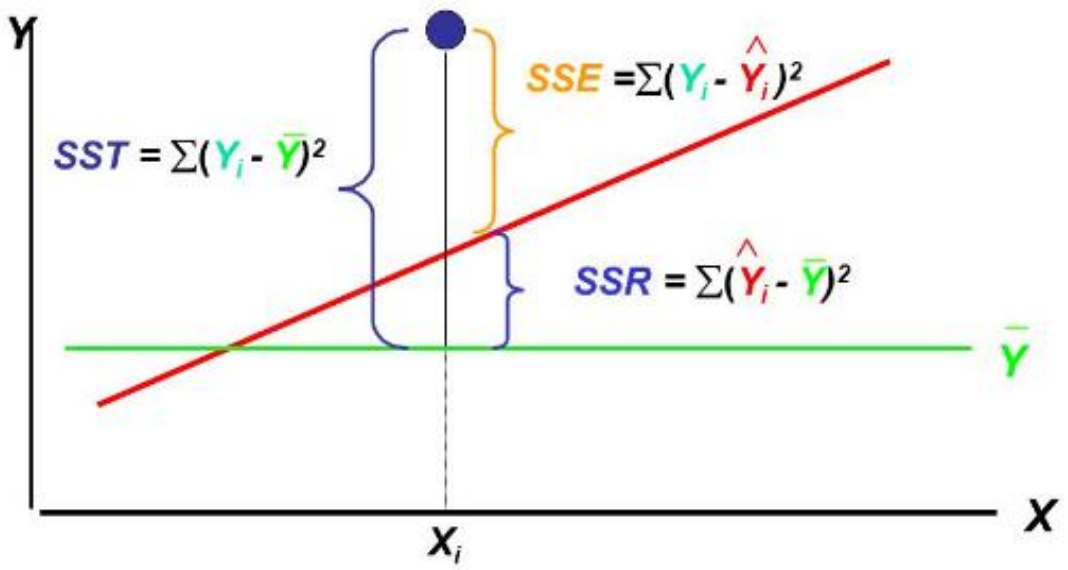
$$\hat{Y} = \hat{f}(X)$$

Mean Squared ERROR!



2. Regression [Analysis]

R squared! 그야 말로 Signal & Noise

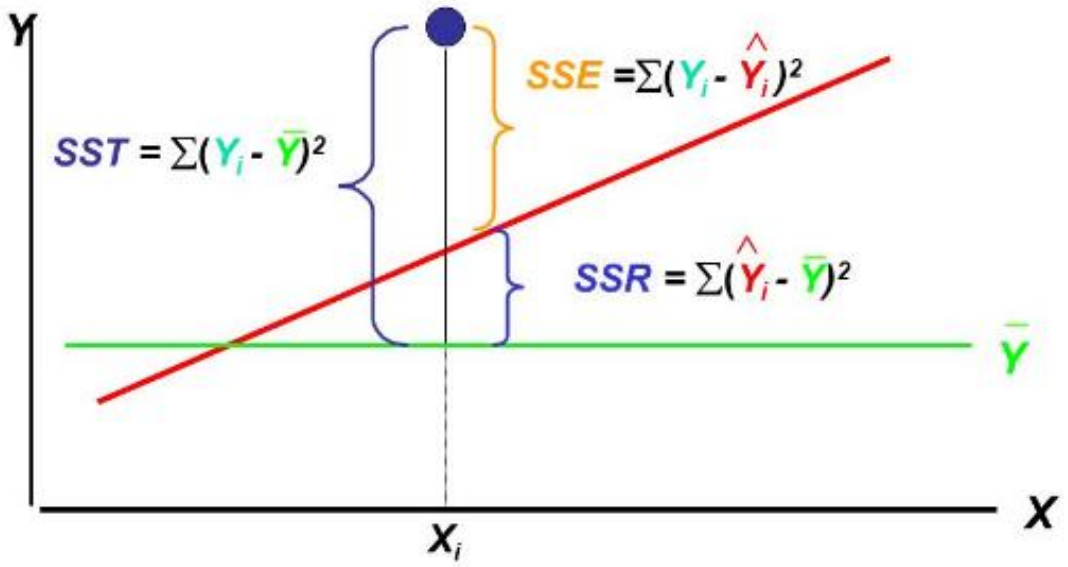


$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

D:plus

2. Regression [Analysis]

F-statistics!　이것도 결국 Signal & Noise



$$F = \frac{MSR}{MSE}$$

2. Regression [Analysis]

의아함이 없어야 한다!

```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

관심사가, u, v, w 세 변수를 사용해 선형적으로 설명이 될 것이라 가정했음.

잔차(오차)의 50%는 -1~ 1.4 사이에 몰려있다, 오차의 최대값은 3.4 정도

Intercept 는 Regressor와 상관없이 설명되는 bias인데, 그 값은 크지만, 실제 Parameter가 0이라고 가정 해도, 샘플데이터로 1.422 정도의 추정 값을 얻을 확률이 32%는 된다.

U와 v는 각각 1unit이 커지면 y에 1unit정도의 변화를 준다고 추정되고 있다. 실제로 아무런 관련이 없다고(parameter 0) 가정하면, 샘플데이터로 저 정도 추정 값을 얻을 확률은 각각 0.1%, 2% 정도이다. 우연으로 치부하기에는, 통계적으로 유의미하다.

잔차의 표준 오차는 1.62 정도, 변동폭이 조금 큰 것은 아쉽다. 평균선으로 못 잡아낸 에러 중, 약 50%는 설명을 해내며, 적어도, u,v,w중 하나는 y를 설명하는데 유의미하다.

D:plus

2. Regression [Analysis]

Regression ‘Analysis’

Y는 아이스크림 판매량. U는 온도 V는 체감온도, W는 습도 라고 해보자.

모두가 알아들을 수 있도록 쉽게 설명해보자.

```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

D:plus

3. Design a Cost function

3. Design a Cost function

Cost Function, 그리고 Gradient Descent // 손으로 해보자!

모델 구조 잡기 (가정하기)

$Y = f(X) + \epsilon$ $\quad y = ax + \epsilon$ 라고 가정.

$\hat{Y} = \hat{f}(X)$ $\quad \hat{y} =$ ← 채워 넣자.

⇨

Cost Function

$$C(a) = SSE = \sum(\hat{y} - y)^2 = \sum(\hat{a}x - y)^2 = 5(\hat{a} - 2)^2$$

Gradient

$$\frac{dC(\hat{a})}{d\hat{a}} = 2\sum x(\hat{a}x - y) = 10(\hat{a} - 2)$$

⇦

$$\hat{a}_{new} := \hat{a}_{old} - 0.01\frac{dC(\hat{a})}{d\hat{a}}$$

Y	X
2	1
4	2

	8

(weight, Cost Function) Graph

D:plus

3. Design a Cost function

Cost Function, 그리고 Gradient Descent // 손으로 해보자!

모델 구조 잡기 (가정하기)

$Y = f(X) + \epsilon$ $\quad y = ax + b + \epsilon$ 가정.

$\hat{Y} = \hat{f}(X)$ $\quad \hat{y} =$ ← 채워 넣자.

Cost Function

$$C(a,b) = SSE = \sum(\hat{y} - y)^2 = \sum(\hat{a}x + \hat{b} - y)^2$$

Gradient

$$\frac{\partial C}{\partial \hat{a}} = 2\sum x(\hat{a}x + \hat{b} - y) = 10\hat{a} + 6\hat{b} - 20$$

$$\frac{\partial C}{\partial \hat{b}} = 2\sum(\hat{a}x + \hat{b} - y) = 6\hat{a} + 4\hat{b} - 12$$

$$\hat{a}_{new} := \hat{a}_{old} - 0.01\frac{\partial C}{\partial \hat{a}}$$

	0

$$\hat{b}_{new} := \hat{b}_{old} - 0.01\frac{\partial C}{\partial \hat{b}}$$

	3

Y	X
2	1
4	2

D:plus

3. Design a Cost function

(1, 2)보다 (2, 4)가 더 중요하다고 해보자. 현실이라면, 큰 값을 더 잘 맞춰야 할 수도 있다

모델 구조 잡기 (가정하기)

$Y = f(X) + \epsilon$ $y = ax + b + \epsilon$ 가정.

$\hat{Y} = \hat{f}(X)$ $\hat{y} =$ ← 채워 넣자.

Cost Function

$$C(a,b) = [\quad]SSE = \sum[\quad](\hat{y} - y)^2 = \sum[\quad](\hat{a}x + \hat{b} - y)^2$$

Gradient

$$\frac{\partial C}{\partial \hat{a}} = 2\sum x[\quad](\hat{a}x + \hat{b} - y) = 36\hat{a} + 20\hat{b} - 72$$

$$\frac{\partial C}{\partial \hat{b}} = 2\sum[\quad](\hat{a}x + \hat{b} - y) = 20\hat{a} + 12\hat{b} - 40$$

$\hat{a}_{new} := \hat{a}_{old} - 0.01\frac{\partial C}{\partial \hat{a}}$

	0

$\hat{b}_{new} := \hat{b}_{old} - 0.01\frac{\partial C}{\partial \hat{b}}$

	3

Y	X
2	1
4	2

D:plus

3. Design a Cost function

Summary & Challenge

Cost Function은 있는 것 그대로 가져다 쓰는 것도 좋지만, 상황에 따라 새로 디자인을 해야 할 수도 있다.

Q1. Linear Regression에서, Cost Function 에 L2 Regularization을 가할 경우 Ridge Regression이라고 한다.
혹은, L1 Regularization을 가할 경우 Lasso Regression이라고 한다. 무슨 효과가 있을까?

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

D:plus