2강

단순회귀모형 (1)

정보통계학과 김성수교수

학습목차

1 회귀분석이란?

회귀분석이란?

✓ 회귀분석 관련 변수(variable)의 분류

- 독립변수(independent variable) : 다른 변수에 영향을 주는 변수.
- 설명변수(explanatory variable)이라고도 함. 보통 X 로 표시.

- 종속변수(dependent variable) : 다른 변수에 의하여 영향을 받는 변수.
- 반응변수(response variable)이라고도 함. 보통 Y 로 표시.

➢ 한 나라에서 국민소득이 증가하면 자동차 보유대수가 증가한다.

이 경우, 국민소득은 독립변수, 자동차 보유대수는 종속변수가 됨.

회귀분석(regression analysis)

- 독립변수(들)과 종속변수 간의 함수관계를 규명하는 통계적인 분석방법
- 회귀(regression)란 뜻은 “본래의 자기 자리로 돌아 온다”라는 뜻.
- 회귀란 용어는 영국의 우생학자 갈튼(Galton;1822~1911)이 처음 사용.

완두콩 시험에서 부모콩(mother seeds)의 무게를 X축으로, 자식콩(daughter seeds)의 무게를 Y축으로 산점도를 그려 두 세대간의 관계를 살펴봄. 이들의 관계식은 양의 직선관계이나 기울기는 1보다 작아서 자식의 무게는 평균 무게로 회귀하려는 경향이 있다는 사실을 발견하고 이를 회귀(regression)로 명명함.

Pearson 이 계량적으로 처음으로 분석하여 발표함.

(Stanton, 2001; “Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors”, *Journal of Statistics Education,* Volume 9, Number 3, 2001)

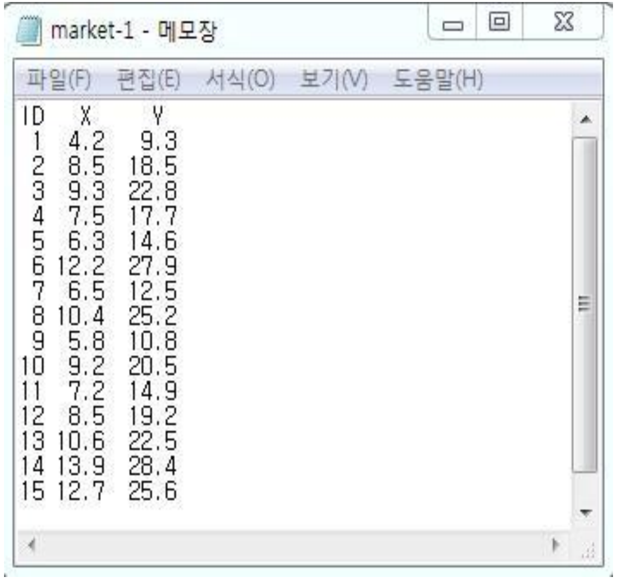
2 단순회귀모형

산점도(scatterplot)

- 한 변수를 x축으로 놓고, 다른 한 변수를 y축으로 그린 그림으로서, 두 연속인 변수들 간의 관계를 밝히고자 할 때 가장 널리 이용되는 그래프임.

(예) 표본상점의 광고료와 총판매액

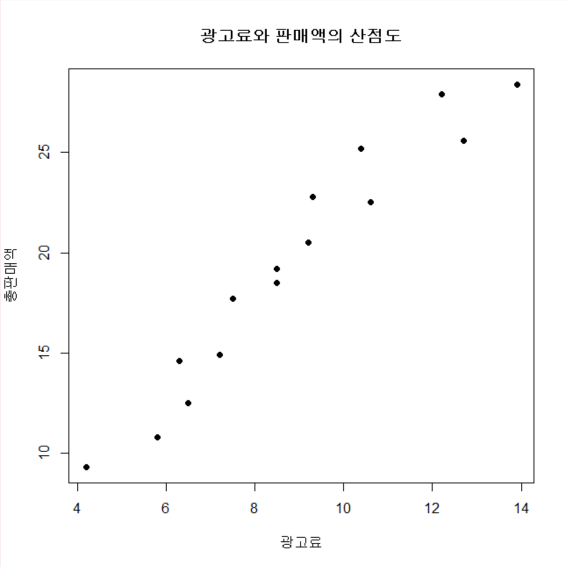
상점번호	광고료(단위;100만원)	총판매액(단위:1000만원)
1	4.2	9.3
2	8.5	18.5
3	9.3	22.8
4	7.5	17.7
5	6.3	14.6
6	12.2	27.9
7	6.5	12.5
8	10.4	25.2
9	5.8	10.8
10	9.2	20.5
11	7.2	14.9
12	8.5	19.2
13	10.6	22.5
14	13.9	28.4
15	12.7	25.6




```
ID  X    Y
1   4.2  9.3
2   8.5  18.5
3   9.3  22.8
4   7.5  17.7
5   6.3  14.6
6  12.2  27.9
7   6.5  12.5
8  10.4  25.2
9   5.8  10.8
10  9.2  20.5
11  7.2  14.9
12  8.5  19.2
13 10.6  22.5
14 13.9  28.4
15 12.7  25.6
```

산점도(scatterplot)

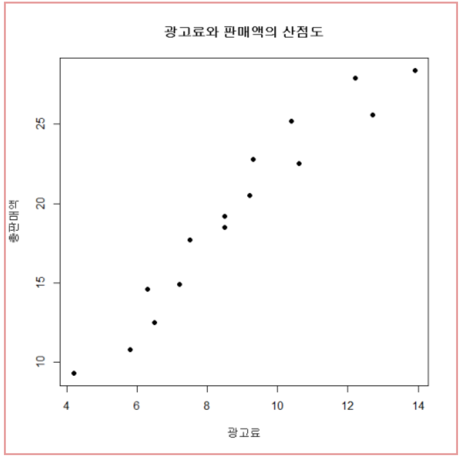
```
> market = read.table("c:/data/reg/market-1.txt", header=T)
> head(market)
  ID    X    Y
1  1  4.2  9.3
2  2  8.5 18.5
3  3  9.3 22.8
4  4  7.5 17.7
5  5  6.3 14.6
6  6 12.2 27.9
> plot(market$X, market$Y, xlab="광고료", ylab="총판매액", pch=19)
> title("광고료와 판매액의 산점도")
```



산점도해석 : 광고료가 증가하면 총판매액도 증가한다는 사실을 쉽게 알 수 있고, 또한 그 관계가 직선인 것도 알 수 있음.

단순회귀모형

단순회귀모형



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

여기서

Y_i = i번째 측정된 반응변수 Y 의 값

β_0 = 절편 회귀계수

β_1 = 기울기 회귀계수

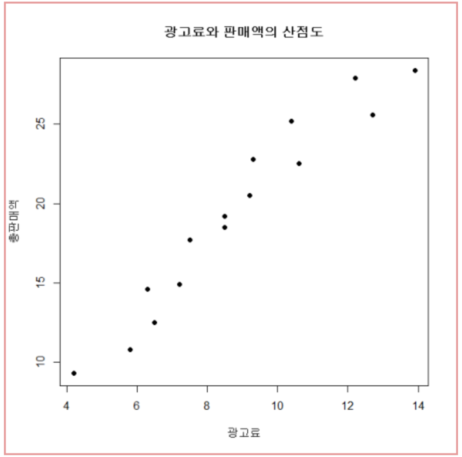
X_i = i번째 주어진 상수 X 값

ϵ_i = i번째 측정된 Y 의 오차항으로 평균 $E(\epsilon_i) = 0$, 분산

$Var(\epsilon_i) = \sigma^2$ 이면서, 다른 오차항과는 상관관계가 없는 것으로 가정.

단순회귀모형

단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



(1) 반응변수 Y_i는 상수항 $\beta_0 + \beta_1 X_i$ 와 오차항 ϵ_i로 이루어져 있으며,
따라서 Y_i는 확률변수임.

(2) 오차항의 평균 $E(\epsilon_i) = 0$ 이므로,

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) = \beta_0 + \beta_1 X_i$$

주어진 X에서 Y의 기대값을 $\mu_{Y \cdot X} = \beta_0 + \beta_1 X$ 라고 하면

$$Y = \mu_{Y \cdot X} + \epsilon$$

(3) 오차항 ε_i의 분산은 등분산(homoscedastic) σ^2 으로 가정. 따라서
반응변수 Y_i의 분산은

$$Var(Y_i) = Var(\beta_0 + \beta_1 X_i + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

이므로, 반응변수 Y_i 도 등분산 σ^2 임.

(4) 반응변수 Y 의 오차항들은 서로 독립이라고 가정. 두 변수간의
공분산(covariance)

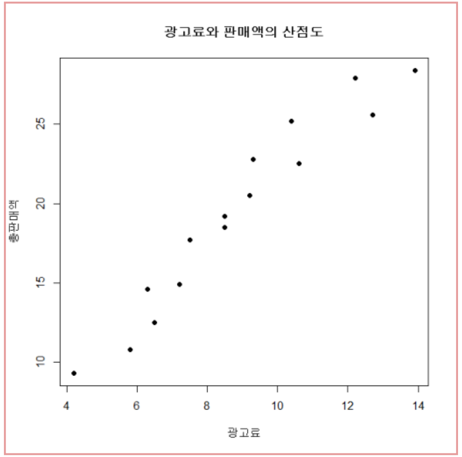
$$Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$$

이 성립되는 가정임.

오차항 ϵ_i 와 ϵ_j 가 서로 독립이므로, 반응변수 Y_i 와 Y_j 도 서로 독립.

단순회귀모형

단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



- 대체모형(alternative model)

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
&= (\beta_0 + \beta_1 \overline{X}) + \beta_1 (X_i - \overline{X}) + \epsilon_i \\
&= \beta_0^* + \beta_1 (X_i - \overline{X}) + \epsilon_i
\end{aligned}
$$

여기서 $\beta_0 + \beta_1 \overline{X}$ 를 β_0^* 로 대치시킨 것임.

대체모형은 설명변수로서 X_i 대신에 $(X_i - \overline{X})$를 사용하는 경우임.

3 회귀선의 추정

회귀선

단순회귀모형 $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

회귀선 : 표본자료(sample data)로부터 모형식을 추정하여 얻은 직선

$$\hat{Y} = b_0 + b_1 X$$

이를 추정된 회귀직선, 또는 간단히 회귀선이라고 함.

- b_0, b_1은 각각 β_0, β_1의 추정값
- $\hat{Y}$(Y hat 이라고 읽음)은 주어진 X에서의 기대값 $E(Y)$의 추정값임.
- b_0는 $X = 0$일 때, $\hat{Y}$의 값으로 추정된 회귀절편(intercept)이라고 함.
- b_1는 X가 한 단위 증가할 때에 $\hat{Y}$의 증가량을 나타내주며, 이를 기울기 (slope)라고 함.

최소제곱법

- 회귀계수 b_0, b_1을 구하는 방법

\- 최소제곱법(method of least squares)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

에서 오차제곱들의 합

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

을 최소로 하는 β_0와 β_1의 값들을 이들의 추정값 b_0와 b_1으로 하는 방법임.

\- 오차제곱합 S를 최소화시키는 β_0와 β_1의 값을 구하기 위하여 S를 β_0와 β_1으로 각각 편미분

$$\frac{\partial S}{\partial \beta_0} = -2\sum (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

최소제곱법

- 편미분값을 0으로 만드는 β_0와 β_1을 각각 b_0와 b_1 으로 대체하여 정리

$$b_0 n + b_1 \sum X_i = \sum Y_i$$

$$b_0 \sum X_i + b_1 \sum X_i^2 = \sum X_i Y_i$$

이 식을 정규방정식(normal equations)이라고 부름.

- 정규방정식을 b_0와 b_1에 대하여 풀면,

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

최소제곱법

- 간편한 표현 :

$$S_{XX} = \sum (X_i - \overline{X})^2$$
$$S_{YY} = \sum (Y_i - \overline{Y})^2$$
$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

라 하면, 기울기 b_1은

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

이 됨.

- $$\begin{aligned}\widehat{Y}_i &= b_0 + b_1 X_i \\ &= (\overline{Y} - b_1 \overline{X}) + b_1 X_i\end{aligned}$$

$$\Rightarrow \widehat{Y}_i - \overline{Y} = b_1 (X_i - \overline{X})$$

R 활용

(예제) 표본상점의 광고료와 총판매액 자료에 대하여 회귀직선을 구하고, 산점도 위에 회귀직선을 그려보아라.

```
> market.lm = lm(Y ~ X, data=market)
> summary(market.lm)
Call:
lm(formula = Y ~ X, data = market)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02908 -1.35349 -0.05685  0.98903  2.51517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3282     1.4302   0.229    0.822
X             2.1497     0.1548  13.889 3.55e-09 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 1.587 on 13 degrees of freedom
Multiple R-squared:  0.9369,    Adjusted R-squared:  0.932
F-statistic: 192.9 on 1 and 13 DF,  p-value: 3.554e-09
```

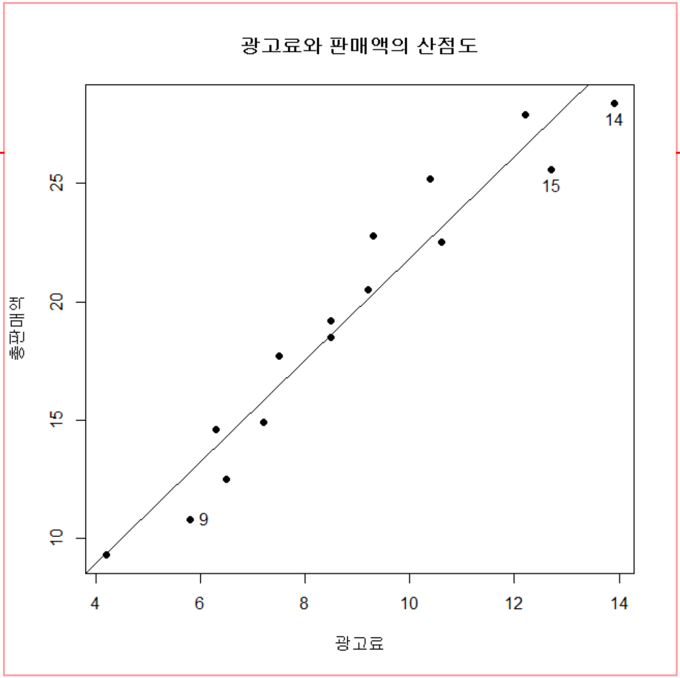
추정된 회귀식

$$\hat{Y} = 0.3282 + 2.1497X$$

R 활용

(예제) 표본상점의 광고료와 총판매액 자료에 대하여 회귀직선을 구하고, 산점도 위에 회귀직선을 그려보아라.

```
> plot(market$X, market$Y, xlab="광고료", ylab="총판매액", pch=19)
> title("광고료와 판매액의 산점도")
> abline(market.lm)
> identify(market$X, market$Y)
[1]  9 14 15
```



잔차(residual)

- 잔차(residual) X_i에서 측정된 값 Y_i 와 추정된 $\widehat{Y}_i$과의 차이

$$e_i = Y_i - \widehat{Y}_i$$

(1) 잔차들의 합은 0. 즉,

$$\sum e_i = 0$$

(2) 잔차제곱의 합, $\sum e_i^2$ 은 최소가 됨.

(3) 관찰값 Y_i 의 합과 추정값 $\widehat{Y}_i$ 의 합은 같음.

$$\sum Y_i = \sum \widehat{Y}_i$$

```
> names(market.lm)
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
> resid = market.lm$residuals
> fitted = market.lm$fitted
> sum(resid)
[1] 4.718448e-16
> sum(fitted)
[1] 290.4
> sum(market$Y)
[1] 290.4
```

잔차(residual)

- 잔차(residual) X_i에서 측정된 값 Y_i 와 추정된 $\hat{Y}_i$과의 차이

$$e_i = Y_i - \hat{Y}_i$$

(4) 잔차들의 X_i에 의한 가중합은 0. 즉,

$$\sum X_i e_i = 0$$

(5) 잔차들의 $\hat{Y}_i$에 의한 가중합은 0. 즉,

$$\sum \hat{Y}_i e_i = 0$$

(6) 점 $(\overline{X}, \overline{Y})$ 는 적합된 회귀선상에 있음.

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X})$$

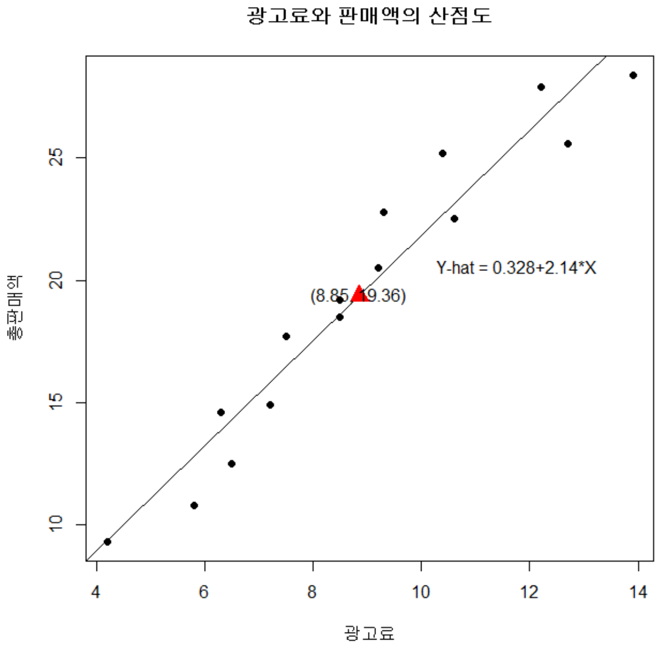
```
> names(market.lm)
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
> sum(market$X*resid)
[1] 9.547918e-15
> sum(fitted*resid)
[1] 5.107026e-15
```

잔차(residual)

(6) 점 $(\overline{X}, \overline{Y})$ 는 적합된 회귀선상에 있음.

$$\hat{Y}_i = \overline{Y} + b_1(X_i - \overline{X})$$

```
> plot(market$X, market$Y, xlab="광고료", ylab="총판매액",
        pch=19)
> title("광고료와 판매액의 산점도")
> abline(market.lm)
> xbar = mean(market$X)
> ybar = mean(market$Y)
> xbar
[1] 8.853333
> ybar
[1] 19.36
> points(xbar, ybar, pch=17, cex=2.0, col="RED")
> text(xbar, ybar, "(8.85, 19.36)")
> fx <- "Y-hat = 0.328+2.14*X"
> text(locator(1), fx)
```



4 회귀모형의 정도

분산분석표에 의한 F-검정

- 변동의 분해

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \overline{Y})^2$$

$$SST \quad = \quad SSE \quad + \quad SSR$$

총제곱합   잔차제곱합   회귀제곱합

자유도 $(n-1) \quad = \quad (n-2) \quad + \quad 1$

분산분석표에 의한 F-검정

<단순회귀의 분산분석표>

요인	자유도	제곱합	평균제곱	F_0
회귀	1	SSR	$MSR = SSR$	$\frac{MSR}{MSE}$
잔차	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
계	$n-1$	SST		

- 가설검정 $H_0 : \beta_1 = 0$

  $H_1 : \beta_1 \neq 0$

- 검정통계량 $F_0 = \frac{MSR}{MSE}$

- 검정방법 : "$F_0 > F(1, n-2; \alpha)$" 이면 귀무가설을 기각하고, 회귀직선이 유의하다 라고 말함.

- R 에서는 검정통계량 F_0에 대한 유의확률 p-값이 제공됨.

  "p-값 < 유의확률 α" 이면 귀무가설을 기각함.

R 활용 : 분산분석표

```
> market.lm = lm(Y ~ X, data=market)
> anova(market.lm)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq  F value    Pr(>F)
X          1 485.57  485.57    192.9   3.554e-09 ***
Residuals 13 32.72     2.52
```

분산분석 결과 해석 : p-값$=3.554\times10^{-9}$ **로 매우 작은 값이므로**
$H_0:\beta_1=0$ **을 기각.**

결정계수

- 결정계수(coefficient of determination)

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

: 총변동중에서 회귀선에 의하여 설명되는 비율이며,
R^2 의 범위는 $0 \leqq R^2 \leqq 1$ 임.

- 결정계수 R^2의 값은 0에서 1사이에 있으며, X 와 Y 사이에 높은 상관관계가 있을수록 R^2의 값은 1에 가까워짐. 즉, R^2의 값이 0에 가까운 값을 가지는 회귀선은 쓸모가 없는 회귀선이고, 회귀분석이 의미가 없으며, R^2의 값이 큰 값을 가질수록 회귀선의 유용성이 높아짐.

- 결정계수는 총변동을 설명하는 데 있어서 회귀선에 의하여 설명되는 변동이 기여하는 비율을 의미하므로 결정계수를 회귀선의 기여율이라고 부르기도 함.

R 활용 : 결정계수

```
> market.lm = lm(Y ~ X, data=market)
> anova(market.lm)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq  F value    Pr(>F)
X          1 485.57  485.57    192.9  3.554e-09 ***
Residuals 13 32.72     2.52
```

$\Rightarrow R^2 = 485.57/(485.57+32.72) = 0.9369$

이는 총변동 중에서 회귀직선에 의하여 설명되는 부분이 94%라는 의미로서, 추정된 회귀선의 정도가 높다는 것을 알 수 있음.

추정값의 표준오차

- 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$
- 분산분석표에서 잔차평균제곱 MSE 는 오차분산 σ^2의 불편추정량이 됨. 따라서 MSE 의 제곱근을 추정값의 표준오차(standard error of estimate)라고 부름.

$$S_{Y \cdot X} = \sqrt{MSE}$$

- 추정값의 표준오차는 두 모형의 비교에서 이 값이 작은 모형이 주어진 자료에 더 잘 적합한다는 의미로 이용됨.

R 활용 : 추정값 표준오차

```
> market.lm = lm(Y ~ X, data=market)
> summary(market.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3282     1.4302   0.229    0.822
X             2.1497     0.1548  13.889 3.55e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.587 on 13 degrees of freedom
Multiple R-squared:  0.9369,    Adjusted R-squared:  0.932
F-statistic: 192.9 on 1 and 13 DF,  p-value: 3.554e-09
```

$$\Rightarrow S_{Y \cdot X} = \sqrt{MSE} = \sqrt{2.52} = 1.587$$

```
> anova(market.lm)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq  F value   Pr(>F)
X          1 485.57  485.57    192.9  3.554e-09 ***
Residuals 13  32.72    2.52
```

상관계수와 결정계수 관계

- 상관계수 : 상관계수는 두 연속인 변수간의 선형관계(linear relationship)가 어느 정도인가를 재는 척도

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

- 단순회귀분석에서는 상관계수 r 을 다음과 같이 구할 수 있음. $r = \pm\sqrt{R^2}$
즉, 상관계수 는 결정계수 R^2의 제곱근이며, 만약 추정된 회귀선의 기울기 b_1이 양이면 $r = \sqrt{R^2}$ 으로 양의 상관계수를 갖고, 기울기 b_1이 음이면 $r = -\sqrt{R^2}$ 으로 음의 상관계수를 가짐.

다음시간 안내

3강. 단순회귀모형 (2)