3강 단순회귀모형 (2)

정보통계학과 김성수교수

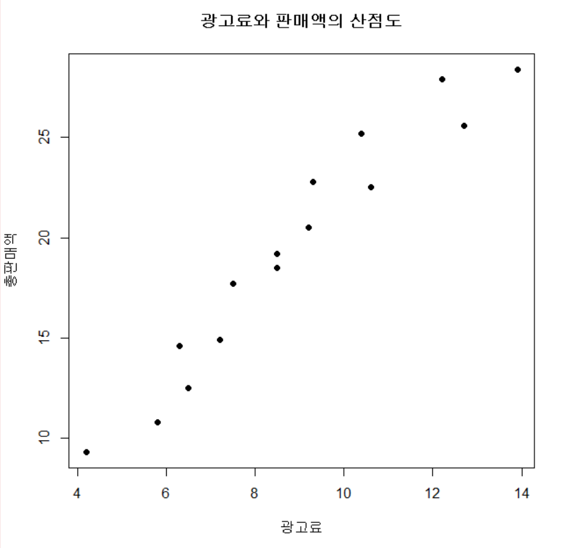
학습목차

1. R을 이용한 회귀분석 (복습)
2. 1.5 단순회귀의 추정과 검정
3. 1.6 가중회귀
4. 1.7 분석 사례

1 R을 이용한 회귀분석 (복습)

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.

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
```

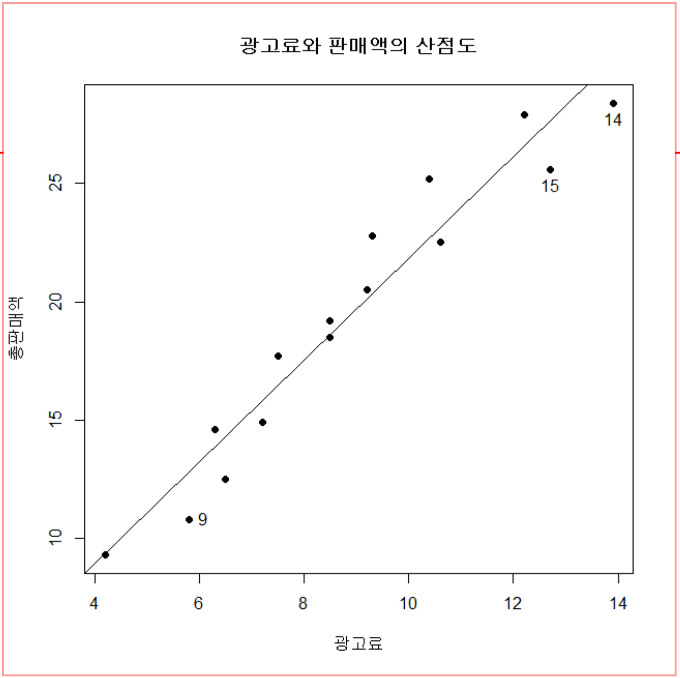
추정된 회귀식

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

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
```

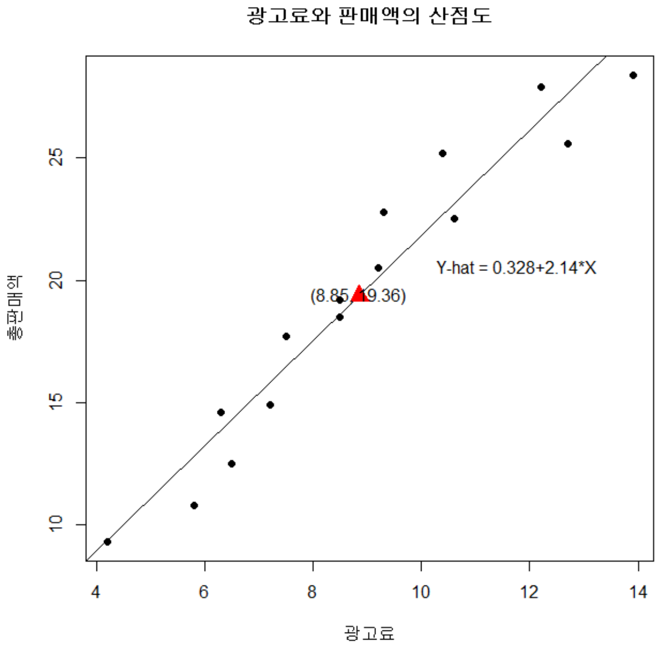


잔차(residual)

(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)
```



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.

참고 1 : 유의수준 0.05에서 F-기각역

```
> qf(0.95, 1, 13)
[1] 4.667193
```

F-value = 192.9 > F(1,13,0.05) 이므로 귀무가설을 기각함.

참고 2 : p-값 구하기

```
> 1-pf(192.9, 1,13)
[1] 3.554018e-09
```

결정계수, 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
```

```
> 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
```

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

결정계수 0.9369

2 단순회귀의 추정과 검정

기본 가정

- n개의 관찰점 (X_i, Y_i)에서 적합된 회귀직선

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

에서 $\hat{Y}$는 $\mu_{Y \cdot X}$, b_0는 β_0, b_1은 β_1의 추정량들로서 이들을 이용하여 각 모수들에 대한 구간추정과 가설검정을 하게 됨.

- 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ 에서 오차항 ϵ_i 에 대해서 $\epsilon \sim N(0, \sigma^2)$ 의 분포를 따른다고 가정하며, 따라서 반응변수 Y_i도 $Y_i \sim N(\mu_{Y \cdot X}, \sigma^2)$ 분포를 따른다고 가정.

β_1 의 신뢰구간

- 회귀계수 기울기 β_1 에 대한 추정량

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

- 기댓값, 분산

$$E(b_1) = \beta_1$$

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

- σ^2의 추정값은 MSE 에 의하여 구해짐

- b_1의 분산의 추정값

$$\widehat{Var}(b_1) = \frac{MSE}{S_{XX}}$$

- β_1 의 신뢰계수$100(1-\alpha)\%$ 신뢰구간

$$b_1 \pm t(n-2;\alpha/2)\sqrt{\frac{MSE}{S_{XX}}}$$

β_0의 신뢰구간

- 절편 β_0 의 추정량

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

- 기댓값 및 분산

$$E(b_0) = \beta_0$$

$$Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right)$$

- β_0 의 $100(1-\alpha)\%$ 신뢰구간

$$b_0 \pm t(n-2;\alpha/2)\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right)}$$

R 결과에서 β_1 , β_0 신뢰구간 구하기

```
> 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
```

β_1 의 95% 신뢰구간

```
> q.val = qt(0.975,13)
> 2.1497 - q.val*0.1548
[1] 1.815275
> 2.1497 + q.val*0.1548
[1] 2.484125
```

β_0 의 95% 신뢰구간

```
> q.val = qt(0.975,13)
> 0.3282 - q.val*1.4302
[1] -2.761559
> 0.3282 + q.val*1.4302
[1] 3.417959
```

추정값의 신뢰구간

- 어떤 주어진 값 X에서 Y의 기대값을 $E(Y)=\mu_{Y.X}=\beta_0+\beta_1 X$ 라고 하면 이는 다음과 같이 추정됨.

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

- 추정량 $\hat{Y}$ 의 기댓값 및 분산

$$E(\hat{Y})=E(b_0+b_1 X)=\beta_0+\beta_1 X=\mu_{Y.X}$$

$$Var(\hat{Y})=Var(\overline{Y})+(X-\overline{X})^2 Var(b_1)$$

$$=\sigma^2\left(\frac{1}{n}+\frac{(X-\overline{X})^2}{S_{XX}}\right)$$

추정값의 신뢰구간

⇒ $\hat{Y}$의 분산은 X의 함수로서 $X=\overline{X}$ 일 경우에 최소가 되며, $X=\overline{X}$를 대칭으로 X의 값이 $\overline{X}$에서 멀어질수록 커짐. 또한 표본의 크기 n이 커져도 $Var(\hat{Y})$이 작아짐을 알 수 있음.

⇒ 주어진 X에서 $\mu_{Y.X}$의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{Y} \pm t(n-2;\alpha/2)\sqrt{MSE\left[\frac{1}{n}+\frac{(X-\overline{X})^2}{S_{XX}}\right]}$$

추정값의 신뢰구간

- 주어진 X의 값에서 새로운 예측값 Y_{new} 의 신뢰구간

$$Var(\hat{Y}_{\text{new}}) = Var(\epsilon) + Var(\hat{Y})$$
$$= \sigma^2 + \sigma^2 [\frac{1}{n} + \frac{(X - \overline{X})^2}{S_{XX}}]$$
$$= \sigma^2 [1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{S_{XX}}]$$

- $100(1-\alpha)\%$ 신뢰구간:

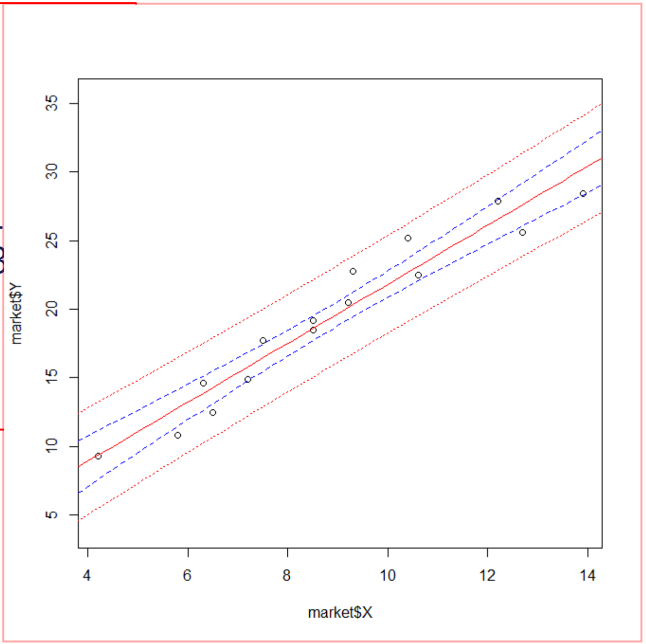
$$\hat{Y}_{\text{new}} \pm t(n-2;\alpha/2) \sqrt{MSE[1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{S_{XX}}]}$$

X의 주어진 값에서 신뢰대 그리기

```
> pred.frame = data.frame(X=seq(3.5, 14.5, 0.2))
> pc = predict(market.lm, int="c", newdata=pred.frame)     #기댓값 신뢰구간
> pp = predict(market.lm, int="p", newdata=pred.frame)     #새로운 값 신뢰구간
> head(pc, 3)
       fit      lwr       upr
1 7.852079 5.855247  9.848911
2 8.282014 6.344903 10.219125
3 8.711949 6.834076 10.589821
> head(pp, 3)
       fit      lwr      upr
1 7.852079 3.885278 11.81888
2 8.282014 4.344937 12.21909
3 8.711949 4.803678 12.62022
```

X의 주어진 값에서 신뢰대 그리기

```
> pred.X = pred.frame$X
> pred.X
 [1]  3.5  3.7  3.9  4.1  4.3  4.5  4.7  4.9  5.1  5.3  5.5  5.7  5.9  6.1
[15]  6.3  6.5  6.7  6.9  7.1  7.3  7.5  7.7  7.9  8.1  8.3  8.5  8.7  8.9
[29]  9.1  9.3  9.5  9.7  9.9 10.1 10.3 10.5 10.7 10.9 11.1 11.3 11.5 1
[43] 11.9 12.1 12.3 12.5 12.7 12.9 13.1 13.3 13.5 13.7 13.9 14.1 14.3
> plot(market$X, market$Y, ylim=range(market$Y, pp))
> matlines(pred.X, pc, lty=c(1,2,2), col="BLUE")
> matlines(pred.X, pp, lty=c(1,3,3), col="RED")
```



β_1 의 검정

- 추정된 회귀직선의 기울기 b_1의 분포는 $b_1 \sim N(\beta_1, \frac{\sigma^2}{S_{XX}})$ 이므로,

$$\frac{b_1 - \beta_1}{\sqrt{\frac{MSE}{S_{XX}}}} \sim t(n-2)$$

- β_1 에 대한 가설검정

귀무가설 $H_0 : \beta_1 = 0$

대립가설 $H_1 : \beta_1 \neq 0$

- 검정통계량

$$t_0 = \frac{b_1}{\sqrt{\widehat{Var}(b_1)}} = \frac{b_1}{\sqrt{\frac{MSE}{S_{XX}}}}$$

- 검정방법

양측검정이므로 $t(n-2;\alpha/2)$인 기각값을 구한 후,

만약 $|t_0| > t(n-2;\alpha/2)$ 이면 귀무가설을 기각.

R 결과 : β_1 검정

```
> 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
```

기각역 및 p-값 구하기

```
> # 유의수준 0.05 기각역
> qt(0.975, 13)
[1] 2.160369
> # 유의확률 p-값
> 2*(1-pt(13.889, 13))
[1] 3.553531e-09
```

⇨ 이 결과에서 기울기 β_1의 추정값 $b_1 = 2.1497$ 이고, t-값

$$t_0 = \frac{2.1497}{0.1548} = 13.889$$

3 가중회귀

가중회귀

- 오차항마다 분산이 다른 경우

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

$$Var(\epsilon_i) = \sigma_i^2 = \frac{\sigma^2}{w_i}$$

- **가중최소제곱법** : 이와 같이 오차항마다 분산이 다른 경우 가중최소제곱법(method of weighted least squares)을 사용하여 회귀분석하는 것을 가중회귀(weighted regression)분석이라고 부름. 이 때 w_i를 가중값(weights)이라고 함.

R 활용 예 : 가중회귀

(예제) 두 변수 X, Y 에 대하여 다음의 데이터가 얻어졌다.

X_i	1	2	3	4	5
Y_i	2	3	5	8	7

단순회귀모형이 $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\epsilon \sim N(0, X_i\sigma^2)$ 인 경우에 가중회귀직선을 구하라.

```
> x = c(1,2,3,4,5)
> y = c(2,3,5,8,7)
> w = 1/x
> w.lm = lm(y ~ x, weights=w)
> summary(w.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3784     0.6891   0.549   0.6212
x             1.5405     0.2688   5.730   0.0106 *
---
Residual standard error: 0.5411 on 3 degrees of freedom
Multiple R-squared:  0.9163,    Adjusted R-squared:  0.8884
F-statistic: 32.84 on 1 and 3 DF,  p-value: 0.01055
```

⇒ 가중회귀직선 $\hat{Y} = 0.3784 + 1.5405X$

4 분석사례

분석사례

어떤 슈퍼마켓에서 고객이 구입하는 상품의 금액과 카운터에서 값을 치르는데 걸리는 시간사이에 회귀함수 관계가 있는가를 알아보기 위하여 10명의 고객을 임의로 추출하여 다음의 데이터를 얻었다. R을 이용하여 회귀모형을 적합해보자.

〈 슈퍼마켓 자료 〉

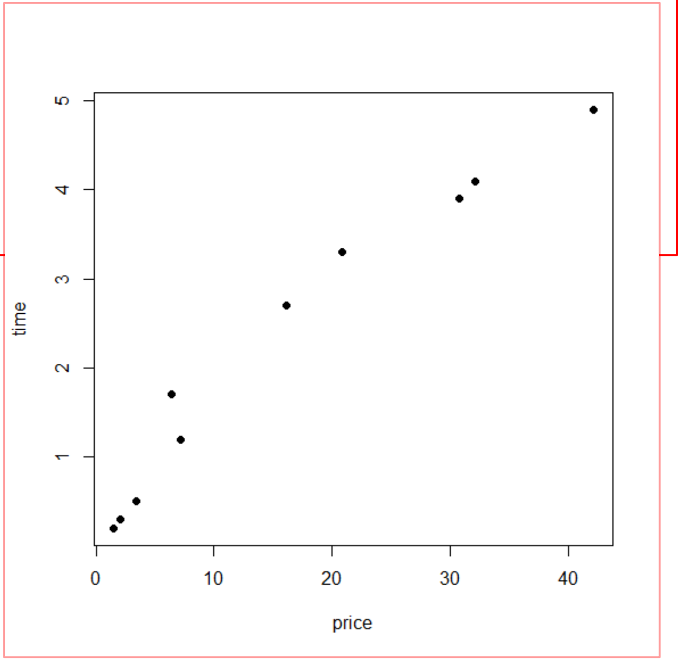
구매상품의 금액 (단위: 천원)	소요되는 시간 (단위: 분)
6.4	1.7
16.1	2.7
42.1	4.9
2.1	0.3
30.7	3.9
32.1	4.1
7.2	1.2
3.4	0.5
20.8	3.3
1.5	0.2

1) 자료파일 만들기

다음과 같이 메모판을 이용하여 supermarket.txt 파일을 만든다.

2) 자료를 읽어 산점도 그리기

```
> super = read.table("c:/data/reg/supermarket.txt", header=T)
> head(super, 3)
  price time
1  6.4   1.7
2 16.1   2.7
3 42.1   4.9
> attach(super)
> plot(price, time, pch=19)
```



3) 회귀모형 적합하기

```
> super.lm = lm(time ~ price, data=super)
> summary(super.lm)
…
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.396460   0.191488    2.07   0.0722 .
price       0.115982   0.008979   12.92 1.22e-06 ***
---
Residual standard error: 0.3925 on 8 degrees of freedom
Multiple R-squared:  0.9542,    Adjusted R-squared:  0.9485
F-statistic: 166.9 on 1 and 8 DF,  p-value: 1.221e-06
```

- 단순회귀방정식 : $\widehat{time} = 0.396 + 0.116 \times price$
- 기울기 검정 : t-값=12.92 이고 p-값=1.22×10^{-6} 이 매우 작으므로 $H_0 : \beta_1 = 0$ 이라는 귀무가설을 기각.
- 결정계수 $R^2 = 0.9542$ 로서, 총변동 중에서 95.42%가 회귀방정식으로 설명되는 회귀변동이 차지하고 있다는 것을 나타냄.
- F-값=166.9 이고, 이에 대한 p-값=1.221×10^{-6} 으로서 적합된 회귀직선이 유의하다는 것을 알 수 있음.

4) 분산분석표 구하기

```
> anova(super.lm)
Analysis of Variance Table

Response: time
          Df  Sum Sq Mean Sq F value    Pr(>F)
price      1 25.7036 25.7036  166.85 1.221e-06 ***
Residuals  8  1.2324  0.1541
```

분산분석표에서 보면 검정통계량 F_0=166.85 이고, 이에 대한 유의확률 p-값=1.221×10^{-6} 이 매우 작으므로 적합된 회귀선이 유의하다는 것을 알 수 있음.

5) 잔차 및 추정값 보기

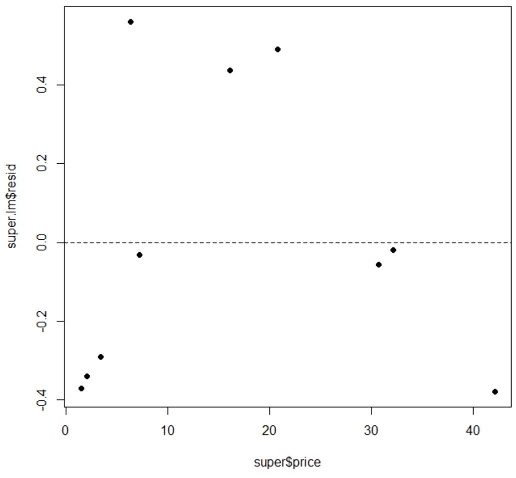
```
> names(super.lm)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> cbind(super, super.lm$resid, super.lm$fitted)
   price time super.lm$resid super.lm$fitted
1    6.4  1.7     0.56125840        1.138742
2   16.1  2.7     0.43623742        2.263763
3   42.1  4.9    -0.37928275        5.279283
4    2.1  0.3    -0.34002095        0.640021
5   30.7  3.9    -0.05709314        3.957093
6   32.1  4.1    -0.01946730        4.119467
7    7.2  1.2    -0.03152683        1.231527
8    3.4  0.5    -0.29079696        0.790797
9   20.8  3.3     0.49112416        2.808876
10   1.5  0.2    -0.37043203        0.570432
```

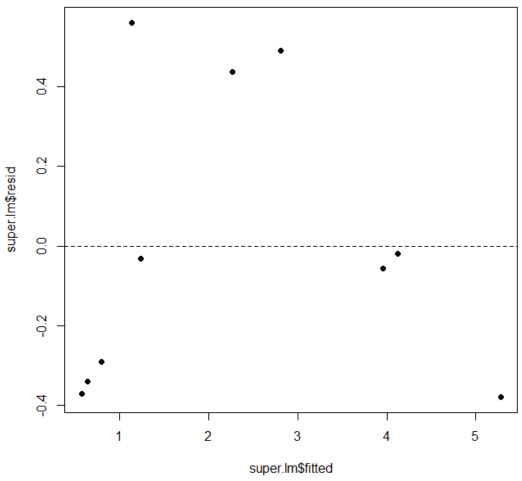
6) 잔차 그림 그리기

```
> plot(super$price, super.lm$resid, pch=19)
> abline(h=0, lty=2)
> plot(super.lm$fitted, super.lm$resid, pch=19)
> abline(h=0, lty=2)
```

이 그림에서 보면 잔차는 0을 중심으로 일정한 범위내에 있으므로 회귀에 대한 기본 가정을 만족한다고 할 수 있으나, X가 증가함에 따라 곡선관계를 보여주고 있음.

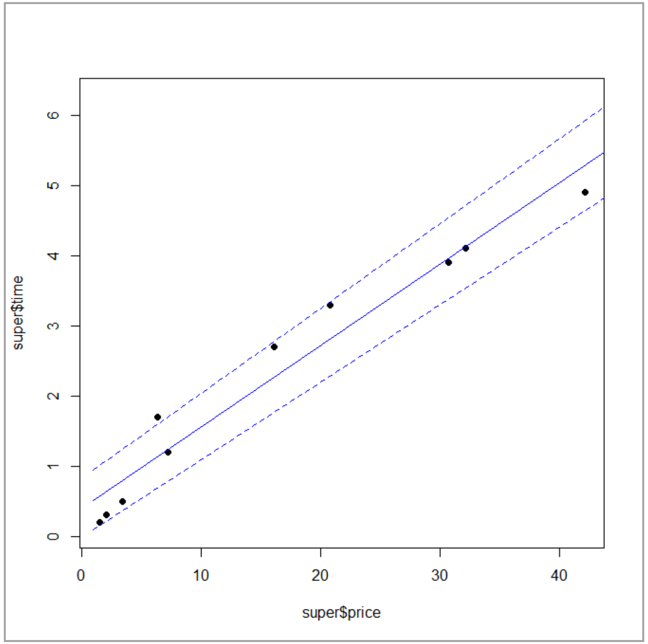
따라서 2차곡선 회귀식 $\hat{Y} = b_0 + b_1 X + b_2 X^2$ 을 구해보는 것도 의미가 있으리라고 생각됨.





7) 추정값의 신뢰대 그리기

```
> p.x = data.frame(price=c(1,45))
> pc = predict(super.lm, int="c", newdata=p.x)
> pred.x = p.x$price
> plot(super$price, super$time, ylim=range(super$time, pc), pch=19)
> matlines(pred.x, pc, lty=c(1,2,2), col="BLUE")
```



다음시간 안내

4강. 중회귀모형 (1)