회귀분석 1

과제 4

학번 202014107
학과 경제학부
이름 강신성

수신자 | 통계학과 이영미 교수님

전북대학교
JEONBUK NATIONAL UNIVERSITY

Regression Analysis : HW04

CH 05.

1. 생선을 잡아서 얼음창고에 일주일 동안 보관한 후에 신선도가 어느 정도 변하는지 실험하였다. 신선도를 y로 놓고 10점 만점으로 하여 0점이 신선도가 전혀 없는 것이고 10점이 가장 좋은 경우이다. 설명변수 x는 생선을 잡은 지 x시간이 경과한 후에 얼음창고에 넣는 것을 가리킨다. 실험으로 10개의 데이터를 얻었다.

y(신선도)	8.5	8.4	7.9	8.1	7.8	7.6	7.3	7.0	6.8	6.7
x(경과시간)	0	0	3	3	6	6	9	9	12	12

(1) 선형회귀모형 ($y=\beta_0+\beta_1 x+\epsilon$)이 타당한가를 유의수준 $\alpha=0.05$를 사용하여 적합결여검정을 행하라.

(a) 먼저 해당 데이터를 기반으로 최소제곱법을 통해 회귀직선을 적합하면,

$$\sum_{i=1}^{10} x_i^2 = 2(3^2+6^2+9^2+12^2) = 540.$$

$$\sum_{i=1}^{10} x_i y_i = 3\times(7.9+8.1) + 6\times(7.8+7.6) + 9\times(7.3+7.0) + 12\times(6.8+6.7) = 431.1$$

$$\sum_{i=1}^{10} y_i^2 = 582.85$$

$$\bar{x}=6,\ \bar{y}=7.61.$$

$$\therefore\ \hat{\beta}_1 = \frac{431.1 - 10\times 6\times 7.61}{540 - 10\times 6^2} = -\frac{25.5}{180} = -\frac{1.7}{12} = -0.14167.$$

$$\hat{\beta}_0 = 7.61 - \left(-\frac{1.7}{12}\right)\times 6 = 8.46.$$

$\therefore$ 적합된 회귀직선 $\hat{y} = 8.46 - 0.14167x$

$$SST = \sum_{i=1}^{10}(y_i-\bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 582.85 - 7.61^2\times 10 = 3.729.$$

$$SSR = \sum_{i=1}^{10}(\hat{y}_i-\bar{y})^2.$$

i	1	2	3	4	5	6	7	8	9	10
$\hat{y}_i$	8.46	8.46	8.035	8.035	7.61	7.61	7.185	7.185	6.76	6.76
$\hat{y}_i-\bar{y}$	0.85	0.85	0.425	0.425	0	0	-0.425	-0.425	-0.85	-0.85

$\therefore SSR = 4\times 0.85^2 + 4\times 0.425^2 = 3.6125$, $SSE = 3.729 - 3.6125 = 0.1165$

$$SSLF = \sum_{i=1}^{5}\sum_{j=1}^{2}(\hat{y}_i-\bar{y}_i)^2 = 2\sum_{i=1}^{5}(\hat{y}_i-\bar{y}_i)^2 = 2\cdot\{(8.46-8.45)^2 + (8.035-8.0)^2 + (7.61-7.7)^2 + (7.185-7.15)^2 + (6.76-6.75)^2\}$$

$$= 0.0215.$$

$$SSPE = 0.1165 - 0.0215 = 0.095$$

1. (1)

(b) 위 내용을 기반으로 적합결여검정을 행하면,

i) $H_0 : E(Y|X=x) = \beta_0 + \beta_1 x$

$H_1 : E(Y|X=x) \neq \beta_0 + \beta_1 x$.

ii) $\alpha = 0.05$

iii) $F_0 = \dfrac{MSLF}{MSPE} = \dfrac{\frac{SSLF}{3}}{\frac{SSPE}{5}} \sim F(3,5)$ as H_0.

iv) $f_0 > F_{0.05}(3,5) = 3.62$이면 귀무가설은 기각 가능.

v) $f_0 = \dfrac{\frac{0.0215}{3}}{\frac{0.095}{5}} = 0.3772 < F_{0.05}(3,5)$ 이므로 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 없다. 따라서 선형 회귀모형은 적합하다.

(2) 선형모형이 타당한 경우, 신선도의 점수가 시간당 얼마만큼이나 떨어지는가를 95% 신뢰계수를 가지고 구간추정하라.

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, $\dfrac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$, $\hat{\sigma}^2 = MSE \Rightarrow \dfrac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim t(8)$.

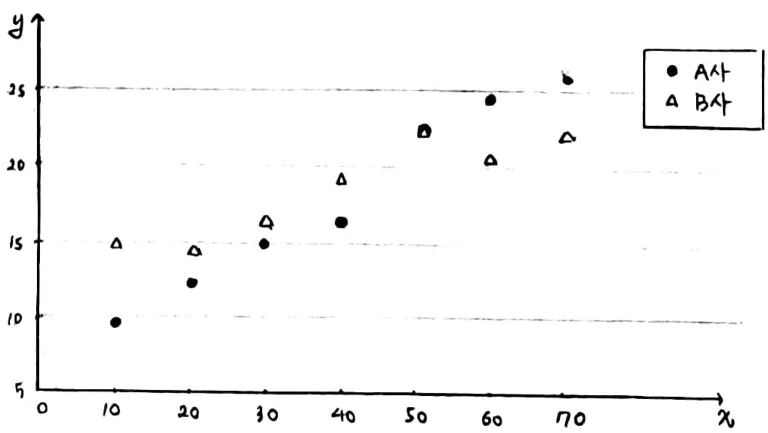
$\hat{\sigma}^2 = MSE = \dfrac{0.1165}{8}$, $\widehat{s.e.}(\hat{\beta}_1) = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\dfrac{\frac{0.1165}{8}}{540-360}} = 0.009$. $t_{0.025}(8) = 2.306$

∴ 시간 당 신선도의 점수 변화 β_1에 대한 95% 신뢰구간 : $0.14167 \pm 2.306 \cdot 0.009 = (0.1209, 0.1624)$.

2. 두 타이어회사 A, B에서 생산되는 타이어를 비교하기 위해서 고속도로에서 트럭이 달리는 상황을 모의실험 (simulated experiment) 하여 다음의 데이터를 얻었다. x는 트럭이 달리는 속도이고, y는 타이어가 마모되기까지의 총 주행거리이다.

$x_{1j} = x_{2j}$	10	20	30	40	50	60	70
y_{1j} (A)	9.8	12.5	14.9	16.5	22.4	24.1	25.8
y_{2j} (B)	15.0	14.5	16.5	19.1	22.3	20.8	22.4

(1) 산점도를 그리시오.



2. (2) 각 회사별로 속도와 총주행거리 간의 회귀모형을 구한다면, 두 개의 직선이 동일하다고 볼 수 있는가? 유의수준 $\alpha = 0.05$에서 가설검정하시오.

(a) Full model 적합.

i) $\sum_{j=1}^{7} x_{1j}^2 = \sum_{j=1}^{7} x_{2j}^2 = 14000$, $\bar{x}_1 = \bar{x}_2 = 40$.

$\sum_{j=1}^{7} x_{1j} y_{1j} = 5827$, $\bar{y}_1 = 18$

$$\therefore \hat{\beta}_{11} = \frac{\sum_{j=1}^{7} x_{1j} y_{1j} - 7 \times \bar{x}_1 \times \bar{y}_1}{\sum_{j=1}^{7} x_{1j}^2 - 7 \times \bar{x}_1^2} = \frac{5827 - 7 \times 40 \times 18}{14000 - 7 \times 40^2} = 0.2811$$

$\hat{\beta}_{01} = \bar{y}_1 - \hat{\beta}_{11} \bar{x}_1 = 6.756 \Rightarrow \hat{y}_1 = 6.756 + 0.2811 x_1$

ii) $\sum_{j=1}^{7} x_{2j} y_{2j} = 5654$, $\bar{y}_2 = \frac{130.6}{7} = 18.657$

$$\therefore \hat{\beta}_{12} = \frac{\sum_{j=1}^{7} x_{2j} y_{2j} - 7 \times \bar{x}_2 \times \bar{y}_2}{\sum_{j=1}^{7} x_{2j}^2 - 7 \times \bar{x}_2^2} = \frac{5654 - 7 \times 40 \times \frac{130.6}{7}}{14000 - 7 \times 40^2} = 0.1536$$

$\hat{\beta}_{02} = \bar{y}_2 - \hat{\beta}_{12} \bar{x}_2 = 12.513 \Rightarrow \hat{y}_2 = 12.513 + 0.1536 x_2$.

j	1	2	3	4	5	6	7
$\hat{y}_{1j}$	9.567	12.378	15.189	18	20.811	23.622	26.433
$\hat{y}_{2j}$	14.049	15.585	17.121	18.657	20.198	21.729	23.265

$\Rightarrow SSE_1 = \sum_{j=1}^{7} (y_{1j} - \hat{y}_{1j})^2 = 5.5568$, $SSE_2 = \sum_{j=1}^{7} (y_{2j} - \hat{y}_{2j})^2 = 8.7142$.

$\therefore SSE(F) = SSE_1 + SSE_2 = 14.27102$, $df_F = 7-2+7-2 = 10$.

(b) Reduced model 적합.

$\sum_{i=1}^{2}\sum_{j=1}^{7} x_{ij}^2 = 28000$, $\sum_{i=1}^{2}\sum_{j=1}^{7} x_{ij} y_{ij} = 5827 + 5654 = 11481$. $\bar{x} = 40$, $\bar{y} = \frac{\frac{130.6}{7} + 18}{2} = 18.3286$.

$$\therefore \hat{\beta}_1 = \frac{11481 - 14 \times 40 \times (\frac{130.6}{14} + 9)}{28000 - 14 \times 40^2} = 0.2173$$

$\hat{\beta}_0 = 18.3286 - 0.2173 \times 40 = 9.6366 \Rightarrow \hat{y} = 9.6366 + 0.2173 x$.

x	10	20	30	40	50	60	70
$\hat{y}$	11.8096	13.9826	16.1556	18.3286	20.5016	22.6746	24.8476

$\therefore SSE(R) = 41.5924$, $df_R = 14-2 = 12$.

2. (2)

(c) 두 직선의 동일성 여부를 유의수준 $\alpha = 0.05$에서 검정

i) $H_0 : \beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$

$H_1 : \beta_{01} \neq \beta_{02}$ or $\beta_{11} \neq \beta_{12}$.

ii) $\alpha = 0.05$

iii) 검정통계량 $F_0 = \dfrac{SSE(R) - SSE(F)}{df_R - df_F} \div \dfrac{SSE(F)}{df_F} \sim F(2, 10)$ as H_0.

iv) $f_0 > F_{0.05}(2, 10) = 2.92$이면 귀무가설은 기각 가능.

v) $f_0 = \dfrac{41.5924 - 14.27102}{2} \div \dfrac{14.27102}{10} = 9.5723 > F_{0.05}(2,10)$ 이므로,

유의수준 $\alpha = 0.05$에서 귀무가설은 기각하고, 대립가설은 수용한다. 즉, 두 개의 직선은 동일하지 않다.

(3) 관심의 대상이 x가 증가함에 따라 y가 얼마나 증가하는가에 있다. 두 회사의 타이어에 대하여 각각 회귀모형을 적합했을 때, 기울기가 같은지 유의수준 5%로 검정하시오.

i) $H_0 : \beta_{11} = \beta_{12}$

$H_1 : \beta_{11} \neq \beta_{12}$.

ii) $\alpha = 0.05$

iii) $\dfrac{(\hat{\beta}_{11} - \hat{\beta}_{12}) - (\beta_{11} - \beta_{12})}{s.e.(\hat{\beta}_{11} - \hat{\beta}_{12})} \sim N(0,1)$ as H_0.

$s.e.(\hat{\beta}_{11} - \hat{\beta}_{12}) = \sqrt{Var(\hat{\beta}_{11} - \hat{\beta}_{12})}$. 각 표본은 독립이므로 $Var(\hat{\beta}_{11} - \hat{\beta}_{12}) = Var(\hat{\beta}_{11}) + Var(\hat{\beta}_{12}) = \dfrac{\sigma^2}{S_{xx_1}} + \dfrac{\sigma^2}{S_{xx_2}}$.

$\Rightarrow s.e.(\hat{\beta}_{11} - \hat{\beta}_{12}) = \sigma\sqrt{\dfrac{1}{S_{xx_1}} + \dfrac{1}{S_{xx_2}}}$. σ^2을 알지 못하므로 $\hat{\sigma}^2 = MSE = \dfrac{SSE_1 + SSE_2}{n_1 - 2 + n_2 - 2}$로 추정하면,

검정통계량 $T_0 = \dfrac{(\hat{\beta}_{11} - \hat{\beta}_{12}) - (\beta_{11} - \beta_{12})}{\sqrt{MSE}\sqrt{\dfrac{1}{S_{xx_1}} + \dfrac{1}{S_{xx_2}}}} \sim t(10)$. as H_0.

iv) $|t_0| > t_{0.025}(10) = 2.228$이면 귀무가설은 기각 가능.

v) $t_0 = \dfrac{0.2811 - 0.1536}{\sqrt{MSE}\sqrt{\dfrac{1}{S_{xx_1}} + \dfrac{1}{S_{xx_2}}}} = \dfrac{0.1275}{\sqrt{\dfrac{14.27102}{10}} \times \sqrt{\dfrac{2}{14000 - 7\times 40^2}}} = 3.99 > t_{0.025}(10)$ 이므로

유의수준 $\alpha = 0.05$에서 귀무가설은 기각하고 대립가설은 수용한다. 즉, 두 회사의 타이어에 대하여 속도에 따른 총 주행거리의 변화 정도는 다르다.

3. 단순회귀에서 회귀제곱합,

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

을 이차형식 $y^T B y$로 표현하시오. 이 이차형식의 분포를 구하고, 또한 기대치를 <정리 5.1>에 의하여 구하시오.

단순선형회귀식 $y_i = \beta_0 + \beta_1 x + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0,1)$, $i=1,2,\cdots,n$ 에서

$$\underset{\sim}{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \underset{\sim}{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \underset{\sim}{\mu} = X\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix}$$ 이라 하면,

$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$. $\underset{\sim}{y} \sim N(\underset{\sim}{\mu}, \sigma^2 I_n)$ 이고, 적합된 회귀모형에서 $\hat{\underset{\sim}{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$, $\hat{\underset{\sim}{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$, $\underset{\sim}{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ 이라 할 때,

$$\hat{\underset{\sim}{\beta}} = \arg\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\sim}{\epsilon}^T \underset{\sim}{\epsilon},$$

$$\underset{\sim}{\epsilon}^T\underset{\sim}{\epsilon} = (\underset{\sim}{y} - X\underset{\sim}{\beta})^T(\underset{\sim}{y} - X\underset{\sim}{\beta}) = \underset{\sim}{y}^T\underset{\sim}{y} - \underset{\sim}{\beta}^T X^T \underset{\sim}{y} - \underset{\sim}{y}^T X \underset{\sim}{\beta} + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta} = \underset{\sim}{y}^T\underset{\sim}{y} - 2\underset{\sim}{\beta}^T X^T \underset{\sim}{y} + \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta} \quad (\because \underset{\sim}{\beta}^T X^T \underset{\sim}{y} = \underset{\sim}{y}^T X \underset{\sim}{\beta})$$

$$\Rightarrow \frac{\partial \underset{\sim}{\epsilon}^T\underset{\sim}{\epsilon}}{\partial \underset{\sim}{\beta}} = -2X^T\underset{\sim}{y} + 2X^TX\underset{\sim}{\beta} = 0. \Rightarrow X^TX\underset{\sim}{\beta} = X^T\underset{\sim}{y}.$$

$$\therefore \hat{\underset{\sim}{\beta}} = (X^TX)^{-1}X^T\underset{\sim}{y}, \quad \hat{\underset{\sim}{y}} = X\hat{\underset{\sim}{\beta}} = X(X^TX)^{-1}X^T\underset{\sim}{y}.$$

이에 따라 $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i\hat{y}_i + \sum_{i=1}^{n}\hat{y}_i^2$

$$= \underset{\sim}{y}^T\underset{\sim}{y} - 2\underset{\sim}{y}^T\hat{\underset{\sim}{y}} + \hat{\underset{\sim}{y}}^T\hat{\underset{\sim}{y}}$$

$$= \underset{\sim}{y}^T I_n \underset{\sim}{y} - 2\underset{\sim}{y}^T X(X^TX)^{-1}X^T\underset{\sim}{y} + \underset{\sim}{y}^T X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T\underset{\sim}{y}$$

$$= \underset{\sim}{y}^T I_n \underset{\sim}{y} - 2\underset{\sim}{y}^T X(X^TX)^{-1}X^T\underset{\sim}{y} + \underset{\sim}{y}^T X(X^TX)^{-1}X^T\underset{\sim}{y}$$

$= \underset{\sim}{y}^T(I_n - X(X^TX)^{-1}X^T)\underset{\sim}{y}$ 의 이차형식으로 나타낼 수 있다.

이때 $B = I_n - X(X^TX)^{-1}X^T$ 는,

$$BB = (I_n - X(X^TX)^{-1}X^T)(I_n - X(X^TX)^{-1}X^T)$$

$$= I_nI_n - 2X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$

$$= I_n - X(X^TX)^{-1}X^T$$

$= B$ 이므로, 멱등행렬이다.

$$\therefore rank(B) = tr(B) = tr(I_n) - tr(X(X^TX)^{-1}X^T) = n - tr(X^TX(X^TX)^{-1}) = n - tr(I_2)$$

$$= n-2.$$

3.

$$\underset{\sim}{\mu}^T B \underset{\sim}{\mu} = \underset{\sim}{\beta}^T X^T B X \underset{\sim}{\beta} = \underset{\sim}{\beta}^T X^T (I_n - X(X^TX)^{-1}X^T) X \underset{\sim}{\beta}$$

$$= \underset{\sim}{\beta}^T X^T I_n X \underset{\sim}{\beta} - \underset{\sim}{\beta}^T X^T X (X^TX)^{-1} X^T X \underset{\sim}{\beta}$$

$$= \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta} - \underset{\sim}{\beta}^T X^T X \underset{\sim}{\beta}$$

$$= 0.$$

$\therefore$ $\underset{\sim}{y} \sim N(\underset{\sim}{\mu}, \sigma^2 I_n)$, $\text{rank}(B) = n-2$, $\frac{1}{2}\underset{\sim}{\mu}^T B \underset{\sim}{\mu} = 0$, B는 멱등행렬 이므로,

$\underset{\sim}{y}^T B \underset{\sim}{y} \sim \chi^2(n-2, 0) = \chi^2(n-2)$. 이다.

또한, $\underset{\sim}{y}^T B \underset{\sim}{y}$의 기대치를 구하면,

$$E(\underset{\sim}{y}^T B \underset{\sim}{y}) = tr(B \cdot \sigma^2 I_n) + \underset{\sim}{\mu}^T B \underset{\sim}{\mu}$$

$$= tr(\sigma^2 B) + 0$$

$$= \sigma^2 tr(B)$$

$$= (n-2)\sigma^2 .$$

4. **R 실습**. 어떤 슈퍼마켓에서 고객이 구입하는 상품의 금액과 카운터에서 값을 치르는 데 걸리는 시간 사이에 회귀함수관계가 있는가를 알아보기 위하여 10명의 고객을 임의로 추출하여 다음의 데이터를 얻었다.

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

(1) 데이터의 산점도를 그려라.

- 먼저 데이터프레임을 정의하면 아래와 같다.

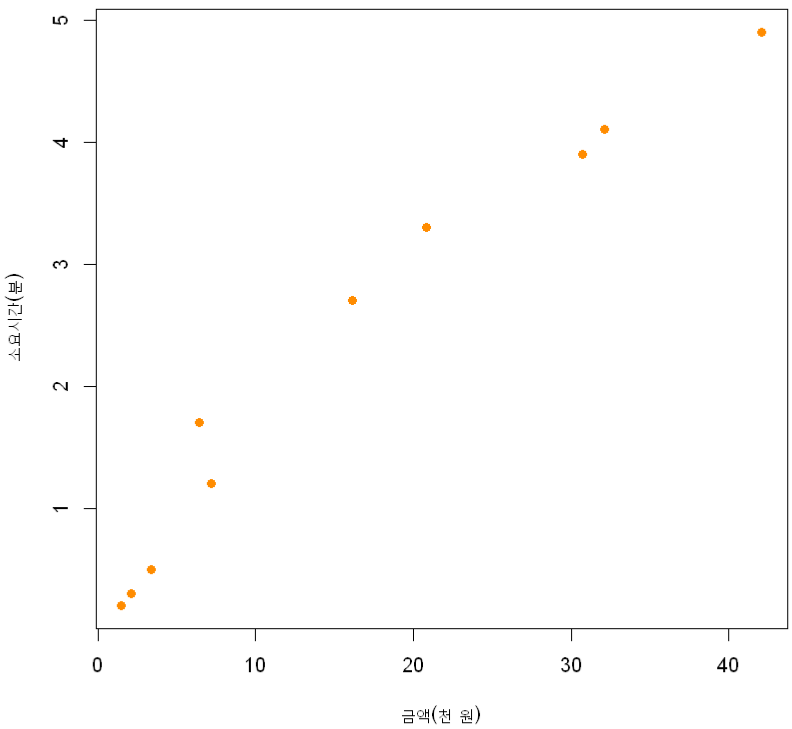
In [1]:
```
x <- c(6.4, 16.1, 42.1, 2.1, 30.7, 32.1, 7.2, 3.4, 20.8, 1.5)
y <- c(1.7, 2.7, 4.9, 0.3, 3.9, 4.1, 1.2, 0.5, 3.3, 0.2)

df <- data.frame(x, y)
df
```

A data.frame: 10 × 2

x	y
<dbl>	<dbl>
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

- 데이터의 산점도

In [2]:
```
plot(y~x, df, pch = 16, col = 'darkorange', xlab = '금액(천 원)', ylab = '소요시
```



> 데이터가 약간의 비선형성이 존재하는 것으로 보인다.

---

(2) 단순회귀모형, $y = \beta_0 + \beta_1 x + \epsilon$을 가정하고, 이를 적합한 경우에 결정계수 R^2의 값은 얼마인가? 만족할 만큼 충분히 큰가?

In [3]:
```
model <- lm(y~x, data = df)
print(paste('R-squared = ',summary(model)$r.squared))
```

```
[1] "R-squared =  0.954247030991951"
```

> 결정계수의 값을 더 높일 수 있을 것 같다.

---

(3) 다음의 비선형모형을 고려하자.

(a) $y = e^{\beta_0 + \beta_1 x + \epsilon}$

(b) $y = \beta_0 + \beta_1 \sqrt{x} + \epsilon$

(c) $y = \beta_0 x^{\beta_1} \epsilon$

(d) $y = \beta_0 \cdot {\beta_1}^x \epsilon$

(e) $y = \beta_0 + \beta_1 (\frac{1}{x}) + \epsilon$

위의 모형을 적절한 모형변환을 통하여 선형모형으로 만든 후 회귀모형을 적합하고 각각 R^2을 구하시오. 어떤 모형이 가장 큰 R^2의 값을 가지는가?

- 각각 변환 후 적합

In [22]:
```
## (a)
log_y <- log(y)
model_a <- lm(log_y~x)

## (b)
sqrt_x <- sqrt(x)
model_b <- lm(y~sqrt_x)

## (c)
log_x <- log(x)
model_c <- lm(log_y~log_x)  ## _beta0 = log(beta0), _epsilon = log(epsilon)

## (d)
model_d <- lm(log_y~x)  ## _beta0 = log(beta0), _beta1 = log(beta1), _epsilon =

## (e)
inv_x <- 1/x
model_e <- lm(y~inv_x)
```

- 각각의 R-squared 산출

In [34]:
```
data.frame(
    'Model' = c('a', 'b', 'c', 'd', 'e'),
    'R-squared' = c(summary(model_a)$r.squared, summary(model_b)$r.squared,
                    summary(model_c)$r.squared, summary(model_d)$r.squared,
                    summary(model_e)$r.squared)
)
```

A data.frame: 5 × 2

Model	R.squared
<chr>	<dbl>
a	0.7384658
b	0.9888110
c	0.9624502
d	0.7384658
e	0.6873067

> (b)의 경우가 가장 큰 R^2값을 가진다.

---

(4) 위의 (3)에서 R^2의 값이 가장 큰 모형이 선택되었을 경우, 이 모형의 추정식을 사용하여 구매상품의 총 금액이 $10,000$원인 경우에, 카운터에서 값을 치르는 데 평균 몇 분이 소요되리라고 예측하는가?

- `model_b` 를 사용하여 예측하면...

```
In [49]: predict(model_b, newdata = data.frame(sqrt_x = sqrt(10)))  ## x값에 10을 넣어줘0
```

1: 1.88569333051879

> 따라서 카운터에서 값을 치르는 데 평균 1.89분이 필요하다고 예측된다.

---

5. **R 실습**. 아마존 강 수위 문제 (문제의 출처 : 참고문헌(3.8)) 아마존 강 유역은 지구상의 가장 큰 열대우림 지역이지만, 대부분의 다른 자연자원과 마찬가지로 개발의 손길이 미치면서 열대림이 급속히 파괴됐다. 1970년대 이후 아마존 상류 지역에 도로가 건설되면서 인구가 빠르게 증가되었고 대규모의 삼림파괴가 이뤄졌다. 강수량과 유수량이 모두 영향을 받을 수 있기 때문에 이것은 결국 아마존 강 전체에 영향을 미치는 심각한 기후학적 및 수문학적 변화를 가져왔다. 다음의 표는 페루 이키토스(Iquitos)에서 1962년부터 1978년까지 기록한 아마존 강 최고수위(High)와, 최저수위(Low)를 기록한 것이다. (단위 : 미터)

Table 1 : 아마존 강 데이터 (Amazom River data)

Year	High (m)	Low (m)	Year	High (m)	Low (m)
1962	25.82	18.24	1971	27.36	21.91
1963	25.35	16.50	1972	26.65	22.51
1964	24.29	20.26	1973	27.13	18.81
1965	24.05	20.97	1974	27.49	19.42
1966	24.89	19.43	1975	27.08	19.10
1967	25.35	19.31	1976	27.51	18.80
1968	25.23	20.85	1977	27.54	18.80
1969	25.06	19.54	1978	26.21	17.57
1970	27.13	20.49			

1962년부터 1969년까지의 데이터는 개발 이전에 수집된 데이터이고, 1970년부터 1978년까지의 데이터는 개발 이후에 관측된 데이터를 나타낸다. 이 데이터는 아마존 상류지역의 삼림파괴가 아마존 유역의 강 수위에 변화를 일으켰는지 분석하고자 한다. 우리의 관심은 시간에 따른 아마존 강 수위 변화여부이다. 예를 들어 우리가 다음을 적합한다면,

$$High = \beta_0 + \beta_1 \times Year + \epsilon$$

(1) $\beta_1 = 0$은 시간에 따른 아마존 강의 최고수위에 아무런 (선형)변화가 없다는 것을 의미하고, (2) $\beta_1 > 0$은 아마존 강의 최고수위가 증가된 것을 의미하는데, 이것은 해마다 아마존 강에 흐르는 물이 늘어난 것을 나타낼 수 있다. (3) $\beta_1 < 0$는 시간에 따라 아마존 강의 최고수위가 낮아 진 것을 의미하는데, 이것은 해마나 아마존 강의 흐르는 물이 줄어든 것을 의미한다.

---

(1) $High$**와** $Year$**,** Low**와** $Year$**, 그리고** $High$**와** Low**에 대해 산점도를 그리시오.**

- 데이터를 데이터프레임으로 입력

In [28]:
```
df <- data.frame(
    Year = seq(1962, 1978),
    High = c(25.82, 25.35, 24.29, 24.05, 24.89, 25.35, 25.23, 25.06, 27.13,
             27.36, 26.65, 27.13, 27.49, 27.08, 27.51, 27.54, 26.21),
    Low = c(18.24, 16.50, 20.26, 20.97, 19.43, 19.31, 20.85, 19.54, 20.49,
           21.91, 22.51, 18.81, 19.42, 19.10, 18.80, 18.80, 17.57)
)

head(df)
```

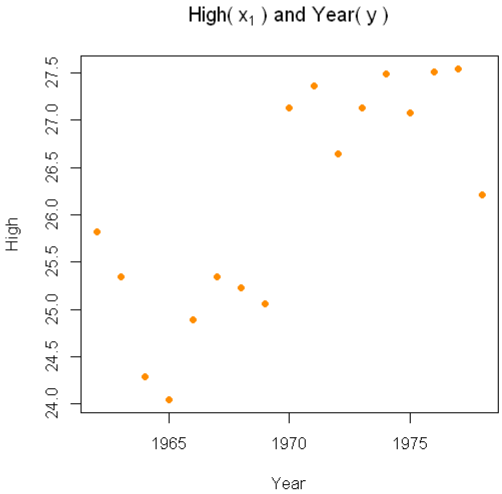
A data.frame: 6 × 3

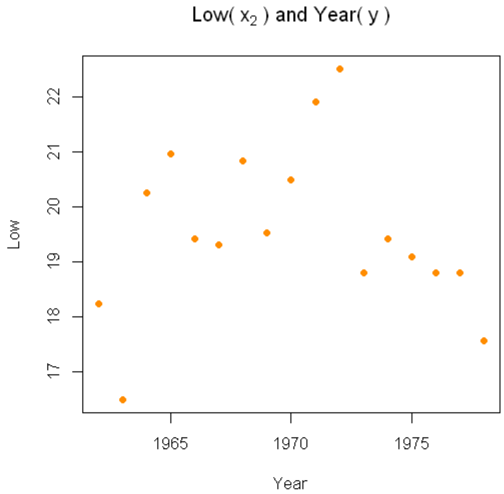
	Year	High	Low
	<int>	<dbl>	<dbl>
1	1962	25.82	18.24
2	1963	25.35	16.50
3	1964	24.29	20.26
4	1965	24.05	20.97
5	1966	24.89	19.43
6	1967	25.35	19.31

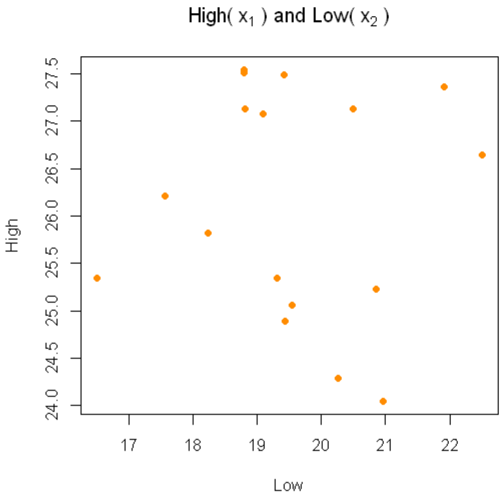
- 개별 산점도 산출

In [29]:
```
options(repr.plot.width = 9, repr.plot.height = 9)
```

In [30]:
```
par(mfrow = c(2,2))
plot(High ~ Year, data = df, pch = 16, col = 'darkorange', main = bquote('High('
plot(Low ~ Year, data = df, pch = 16, col = 'darkorange', main = bquote('Low('~x
plot(High ~ Low, data = df, pch = 16, col = 'darkorange', main = bquote('High('~
```







---

(2) $Year$에 대한 $High$, $Year$에 대한 Low, 그리고 Low에 대한 $High$의 회귀모형을 구하시오. 3개 모형의 결과를 요약하고, 각 모형별로 회귀계수의 의미를 설명하시오.

- `Year`에 대한 `High`

In [31]:
```
model_1 <- lm(High ~ Year, data = df)
summary(model_1)
```

```
Call:
lm(formula = High ~ Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3629 -0.5341  0.1479  0.4903  1.1412 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -330.21235   78.03319  -4.232 0.000725 ***
Year           0.18088    0.03961   4.567 0.000371 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8001 on 15 degrees of freedom
Multiple R-squared:  0.5816,	Adjusted R-squared:  0.5537 
F-statistic: 20.85 on 1 and 15 DF,  p-value: 0.0003708
```

In [32]:
```
model_1$coefficients[2]
```

Year: 0.180882352941173

- 모형의 F값에 대한 `p-value` 가 0.01보다 작으므로 해당 모형은 유의수준 0.01에서 통계적으로 유의하다.
- 회귀계수의 경우 `p-value` 가 0.01보다 작으므로 통계적으로 유의하고, 해당 값이 양수이므로, 시간이 지남에 따라 아마존 강의 최고수위가 증가함을 의미한다.

`-` `Year` 에 대한 `Low`

In [33]:
```
model_2 <- lm(Low ~ Year, data = df)
summary(model_2)
```

```
Call:
lm(formula = Low ~ Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1147 -0.7121 -0.1610  0.9306  2.9664 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.106961 151.723912   0.231     0.82
Year        -0.007892   0.077017  -0.102     0.92

Residual standard error: 1.556 on 15 degrees of freedom
Multiple R-squared:  0.0006996,	Adjusted R-squared:  -0.06592 
F-statistic: 0.0105 on 1 and 15 DF,  p-value: 0.9197
```

- 모형의 F값에 대한 `p-value` 가 0.9197로 해당 모형은 통계적으로 유의미하지 않다.
- 회귀계수는 시간이 지남에 따라 아마존 강의 최저수위가 얼마나 변하는지를 의미한다.

`-` `Low` 에 대한 `High`

In [34]:
```
model_3 <- lm(High ~ Low, data = df)
summary(model_3)
```

```
Call:
lm(formula = High ~ Low, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05605 -0.87774  0.05615  1.01720  1.40344 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.40088    4.02478   6.560 9.05e-06 ***
Low         -0.01406    0.20520  -0.069    0.946    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 15 degrees of freedom
Multiple R-squared:  0.0003129,	Adjusted R-squared:  -0.06633 
F-statistic: 0.004695 on 1 and 15 DF,  p-value: 0.9463
```

- 모형의 F값에 대한 `p-value` 가 0.9463으로, 해당 모형은 통계적으로 유의미하지 않다.
- 회귀계수는 아마존 강의 최저수위가 증가함에 따라 최고수위가 얼마나 변하는지를 의미한다.

---

(3) 이 자료를 근거로 우리는 삼림파괴가 아마존 강 수위의 변화를 일으킨다고 할 수 있는가? 이용가능하다면 이러한 인과관계를 추론하는 데 사용될 수 있는 추가 정보는 무엇이 있겠는가?

`-` 풀이

 시간에 따라 아마존 강 최고수위가 올라간다는 것은 통계적으로 유의미했다. 다만, 이를 삼림파괴와 직접적으로 연결할 수는 없다. 삼림파괴의 정도가 시간에 따라 얼마나 많이 변했느냐에 따른 정보도 없으며, 삼림파괴가 아닌 이외의 요인이 아마존 강 수위의 변화를 일으켰을 수도 있다.
 이에 따라 아마존 삼림지역의 면적을 연도에 따라 추가적으로 기재하여 이러한 인과관계를 추론하는 데 사용할 수 있을 것이다.

(4) 아마존 강의 최저수위와 최고수위와의 산점도를 1960년대, 1970년대 자료별로 다르게 그리고, 각각의 회귀선을 적합하시오.

`-` 데이터 분할

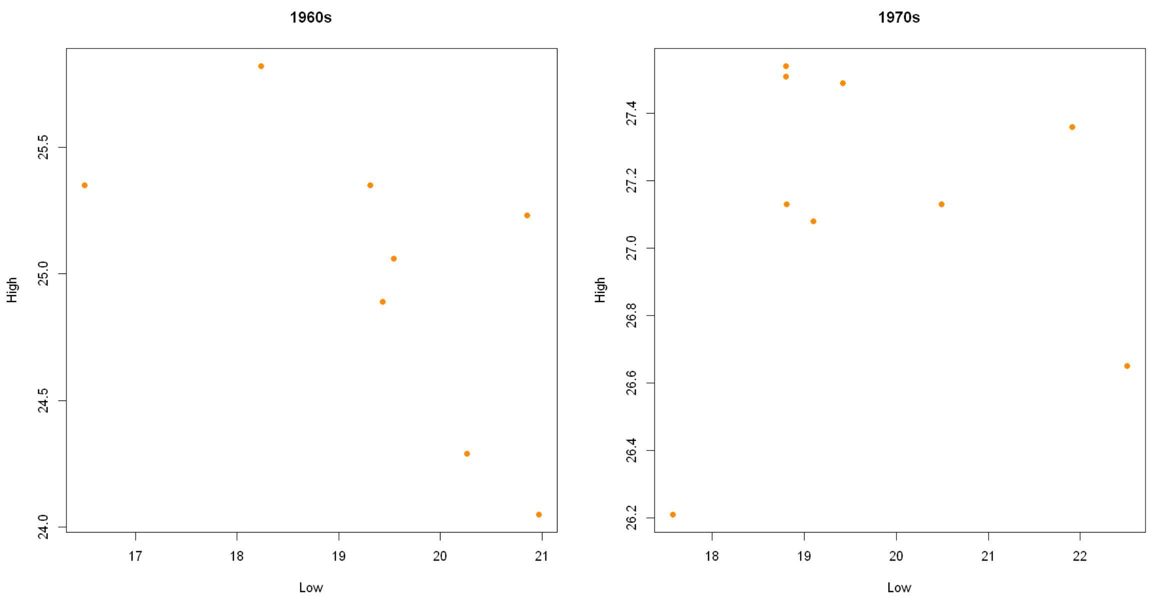
In [35]:
```
df_60 <- df[which(substr(df$Year, 1, 3) == 196),]
df_70 <- df[which(substr(df$Year, 1, 3) == 197),]
```

`-` 산점도

In [36]:
```
options(repr.plot.width = 15, repr.plot.height = 8)
```

In [37]:

```
par(mfcol = c(1,2))
plot(High ~ Low, data = df_60, pch = 16, col = 'darkorange', main = '1960s')
plot(High ~ Low, data = df_70, pch = 16, col = 'darkorange', main = '1970s')
```



`-` 각각의 회귀선 적합

In [38]:

```
## 1960s
model_60 <- lm(High ~ Low, data = df_60)  ## high를 y로

## 1970s
model_70 <- lm(High ~ Low, data = df_70)
```

In [39]:

```
summary(model_60)
```

```
Call:
lm(formula = High ~ Low, data = df_60)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5606 -0.4053 -0.0057  0.3765  0.5895 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  29.8367     2.4640  12.109 1.93e-05 ***
Low          -0.2492     0.1268  -1.966   0.0969 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4925 on 6 degrees of freedom
Multiple R-squared:  0.3918,	Adjusted R-squared:  0.2904 
F-statistic: 3.864 on 1 and 6 DF,  p-value: 0.09691
```

> 1960년대 자료의 경우 회귀계수는 음수이다.

In [40]:

```
summary(model_70)
```

```
Call:
lm(formula = High ~ Low, data = df_70)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.87738 -0.03226  0.02245  0.37253  0.43262 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.80160    2.05235  13.059  3.6e-06 ***
Low          0.01627    0.10381   0.157     0.88    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4733 on 7 degrees of freedom
Multiple R-squared:  0.003495,	Adjusted R-squared:  -0.1389 
F-statistic: 0.02455 on 1 and 7 DF,  p-value: 0.8799
```

> 1970년대 자료의 경우 회귀계수는 양수이며, 회귀직선은 통계적으로 유의미하지 않다.

---

(5) 아마존 강의 최저수위와 최고수위와의 관계가 1960년대와 1970년대에 따라 차이가 있는가? 두 회귀모형의 동일성 여부를 유의수준 $\alpha = 0.01$에서 검정하시오.

ⅰ) 가설 설정

$$H_0 : \beta_{01} = \beta_{02} \ and \ \beta_{11} = \beta_{12}, \quad vs. \quad H_1 : \beta_{01} \neq \beta_{02} \ or \ \beta_{11} \neq \beta_{12}$$

ⅱ) 검정통계량

In [41]:
```
SSE1 = anova(model_60)[2,2]
SSE2 = anova(model_70)[2,2]

SSE_full = SSE1 + SSE2
SSE_reduced = anova(model_3)[2,2]  ## Low에 대한 High
```

In [55]:
```
F0 = ((SSE_reduced - SSE_full)/2)/(SSE_full/(nrow(df)-4))
F0
```

42.827448434212

In [26]:
```
library(gap)

y1 <- df_60[,'High']
x1 <- df_60[,'Low']
y2 <- df_70[,'High']
x2 <- df_70[,'Low']

gap::chow.test(y1, x1, y2, x2)
```

F value: 42.8274484342117 **d.f.1:** 2 **d.f.2:** 13 **P value:** 1.90046810365518e-06 iii) 기각역

In [63]:
```
c <- qf(p = 0.01, df1 = 2, df2 = nrow(df)-4, lower.tail = FALSE)
c  ## critical value
```

6.70096453588078

iv) 검정통계량의 관측치와의 비교

In [61]:
```
F0 > c
```

TRUE

> 검정통계량의 관측값이 임계치보다 크므로(`p-value` 가 0.01보다 작으므로) 유의수준 $\alpha = 0.01$에서 귀무가설을 기각하고, 대립가설을 수용한다. 즉, 아마존 강의 최저수위와 최고수위와의 관계는 1960년대와 1970년대에 따라 차이가 있다.

---

(6) (4)에서 구한 두 회귀모형의 기울기가 같은지 유의수준 $\alpha = 0.01$에서 검정하시오.

i) 가설 설정

$$H_0 : \beta_{11} = \beta_{12}, \quad vs. \quad H_1 : \beta_{11} \neq \beta_{12}$$

ii) 검정통계량

`-` $v\hat{a}r(\hat{\beta_{11}} - \hat{\beta_{12}})$의 값 계산

In [67]:
```
MSE_full = SSE_full/(nrow(df_60)-2+nrow(df_70)-2)  ## 분산 추정치
Sxx1 = sum((x1 - mean(x1))**2)
Sxx2 = sum((x2 - mean(x2))**2)
```

In [68]:
```
var_hat = MSE_full * (1/Sxx1 + 1/Sxx2)
var_hat
```

0.0265987528460711

`-` 검정통계량 산출

In [70]:
```
t0 = (model_60$coefficients[2] - model_70$coefficients[2])/sqrt(var_hat)
t0
```

Low: -1.62782421589498

iii) 기각역

`-` 기각역 산출

In [71]:
```
df = nrow(df_60)-2+nrow(df_70)-2
df
```

13

In [74]:
```
c = qt(0.005, df, lower.tail = FALSE)  ## 양측검정
c
```

3.01227583871658

`-` 또는 `p-value` 산출

In [79]:
```
p_value = 2*pt(t0, df)
p_value
```

Low: 0.127543965251389

iv) 검정통계량의 관측치 비교

In [82]:
```
abs(t0) > c
```

Low: FALSE

> 검정통계량의 관측값이 임계치보다 작으므로, 유의수준 $\alpha = 0.01$에서 귀무가설을 기각할 수 없다. 즉, 두 회귀모형의 기울기가 다르다고 할 수 없다.