회귀분석 1

과제 3

학번	202014107
학과	경제학부
이름	강신성

수신자 | 통계학과 이영미 교수님

전북대학교
JEONBUK NATIONAL UNIVERSITY

Regression Analysis : HW03

CH 04

1. (1) β_0의 90% 신뢰구간을 구하시오

$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$ 에서 $MSE = 0.11$, $S_{xx} = 3.26$, $\bar{x} = 1.6$, $n = 9$, $\hat{\beta}_0 = -0.11$ 이므로,

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{0.11 \times \left(\frac{1}{9} + \frac{1.6^2}{3.26}\right)} = 0.314.$$

즉, 귀무가설 $H_0 : \beta_0 = 0$ 하에서 $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(7)$ 이므로,

$$P\left(-t_{0.05}(7) < \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} < t_{0.05}(7)\right) = P\left(\hat{\beta}_0 - t_{0.05}(7)\cdot\hat{\sigma}_{\hat{\beta}_0} < \beta_0 < \hat{\beta}_0 + t_{0.05}(7)\cdot\hat{\sigma}_{\hat{\beta}_0}\right) = 0.9$$

$\therefore$ β_0의 90% CI : $(-0.705, 0.485)$.

(2) 다음의 가설 검정을 수행하시오.

$$H_0 : \beta_1 = 1 \quad vs. \quad H_1 : \beta_1 > 1.$$

i) $\alpha = 0.05$로 설정

ii) $T_0 = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(7)$ as H_0

iii) $t_0 > t_{0.05}(7) = 1.895$ 이면 귀무가설은 기각 가능 (단측 검정)

iv) $t_0 = \frac{2.16 - 1}{0.18} = 6.44 > t_{0.05}(7)$ 이므로 유의수준 $\alpha = 0.05$ 에서 귀무가설은 기각하고 대립가설을 수용한다 즉, β_1은 1보다 크다.

(3) 무게가 3,000 kg이 되는 차량의 평균 에너지 소모량을 예측하시오. 이것은 무게가 1,000 kg이 되는 차량의 에너지 소모량의 몇배인가?

i) $\widehat{E(Y|X=3)} = -0.11 + 2.16 \times 3 = 6.37.$

ii) $\widehat{E(Y|X=1)} = -0.11 + 2.16 \times 1 = 2.05.$

$\therefore$ 대략 3.1배를 소모한다고 예측할 수 있다.

1. (4) 무게가 3.000kg이 되는 차량의 평균 에너지 소모량과 하나의 개별 y값의 90% 신뢰구간을 각각 구하시오.

i) $\mu_0 = E(Y|x_0)$, $\hat{\mu}_0 = \widehat{E(Y|x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, $\widehat{Var}(\hat{\mu}_0) = MSE\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{Sxx}\right)$.

$x_0 = 3$일 때, $\hat{\mu}_0 = 6.37$, $\widehat{Var}(\hat{\mu}_0) = 0.11 \times \left(\frac{1}{9} + \frac{(3-1.6)^2}{3.26}\right)$, $\widehat{s.e.}(\hat{\mu}_0) = \sqrt{\widehat{Var}(\hat{\mu}_0)} \doteq 0.28$

$\therefore$ μ_0의 90% CI : $6.37 \pm t_{0.05}(7) \cdot 0.28 = (5.839,\ 6.901)$ (Since $t_{0.05}(7) = 1.895$).

ii) $y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$, $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, $\widehat{Var}(\hat{y}_0) = MSE\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{Sxx}\right)$.

$x_0 = 3$일 때, $\hat{y}_0 = 6.37$, $\widehat{Var}(\hat{y}_0) = 0.11 \times \left(1 + \frac{1}{9} + \frac{(3-1.6)^2}{3.26}\right)$, $\widehat{s.e.}(\hat{y}_0) = \sqrt{\widehat{Var}(\hat{y}_0)} \doteq 0.434$.

$\therefore$ y_0의 90% CI : $6.37 \pm t_{0.05}(7) \cdot 0.434 = (5.548,\ 7.192)$.

(5) 원점을 지나는 회귀직선에서 회귀계수(기울기)에 대한 90% 신뢰구간을 구하시오.

적합된 원점을 지나는 회귀직선 : $\hat{y} = 2.10x$.

$\hat{\beta}_1 = 2.10$, $\widehat{Var}(\hat{\beta}_1) = \frac{MSE}{Sxx} = \frac{0.096}{3.26}$, $\widehat{s.e.}(\hat{\beta}_1) = 0.172$. (Since SST = 116.67, SSR = 115.9, df = 8).

$\therefore$ 원점을 지나는 회귀직선의 회귀계수 β_1의 90% CI : $2.10 \pm t_{0.05}(8) \cdot 0.172 = (1.78,\ 2.42)$.

2. 기울기 β_1의 $100(1-\alpha)\%$ 신뢰구간이 0을 그 구간 속에 포함하고 있으면, 가설검정 $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ 에서 유의수준 α로 귀무가설이 채택되고, 만약 신뢰구간이 0을 포함하고 있지 않으면, 대립가설이 채택된다. 이것은 옳은 주장인가?

위 주장은 옳은 주장이다. 신뢰구간에 0이 포함되었다는 것은 아래와 같다.

$\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{s.e.}(\hat{\beta}_1) < 0 < \hat{\beta}_1 + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{s.e.}(\hat{\beta}_1)$. 해당 부등식을 다르게 표현하면,

$-t_{\frac{\alpha}{2}}(n-2) < \frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)} < t_{\frac{\alpha}{2}}(n-2) \Rightarrow \left|\frac{\hat{\beta}_1 - \beta_1}{\widehat{s.e.}(\hat{\beta}_1)}\right| < t_{\frac{\alpha}{2}}(n-2)$ ($\because \beta_1 = 0$ as H_0) 으로, 가설검정에서 귀무가설을 기각하지 못하는 조건과 동일하다.

반대로, 신뢰구간에 0이 포함되어 있지 않다는 것도 비슷하게,

$\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{s.e.}(\hat{\beta}_1) > 0$ or $\hat{\beta}_1 + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{s.e.}(\hat{\beta}_1) < 0 \Rightarrow \frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)} > t_{\frac{\alpha}{2}}(n-2)$ or $-\frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)} > t_{\frac{\alpha}{2}}(n-2)$

$\Rightarrow \left|\frac{\hat{\beta}_1}{\widehat{s.e.}(\hat{\beta}_1)}\right| > t_{\frac{\alpha}{2}}(n-2)$ 로서, 가설검정에서 귀무가설 기각 및 대립가설 수용의 조건과 동일하다.

따라서 문제의 주장은 옳다.

3. 단순 선형회귀모형에서 기울기 $\beta_1 = 0$ 에 대한 검정통계량은 다음과 같이 표본크기 (n)와 표본 상관계수(r)의 함수임을 보이시오

$$t = \frac{\hat{\beta}_1}{\sqrt{MSE / S_{(xx)}}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}.$$

i) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{r \cdot \sqrt{S_{yy}}}{\sqrt{S_{xx}}}.$

ii) $SSE = SST - SSR. = SST\left(\frac{SST}{SST} - \frac{SSR}{SST}\right) = SST(1-R^2)$

$\therefore MSE = \frac{SST(1-R^2)}{n-2.} = \frac{S_{yy}(1-R^2)}{n-2}.$

iii) $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}.$

$r^2 = \frac{\{S_{xy}\}^2}{S_{xx}S_{yy}} = \frac{\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 \times \sum_{i=1}^{n}(y_i-\bar{y})^2}.$ $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i-\bar{x})^2.$

$\therefore r^2 = \frac{\{\hat{\beta}_1 \sum_{i=1}^{n}(x_i-\bar{x})^2\}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 \times \sum_{i=1}^{n}(y_i-\bar{y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}$ 이므로, $R^2 = r^2.$

i) + ii) + iii) 에 따라 윗식을 다시 쓰면,

$$t = \frac{\frac{r \cdot \sqrt{S_{yy}}}{\sqrt{S_{xx}}}}{\sqrt{\frac{\frac{S_{yy}(1-r^2)}{n-2}}{S_{xx}}}} = \sqrt{n-2} \cdot \frac{r}{\sqrt{1-r^2}}$$ 이다.

`4.` **R 실습**. 보스턴 집값 데이터 (데이터 출처 : MASS 패키지). 이 데이터는 Boston 근처 지역의 지역적 특징과 주택 가격의 중앙값 등을 포함하고 있다. 데이터는 MASS 패키지 설치를 통해 Boston 데이터를 사용할 수 있다. 아래와 같이 사용가능하며 자세한 내용을 살펴볼 수 있다.

```
library(MASS)
head(Boston)
?Boston # Boston 데이터의 자세한 설명을 볼 수 있음
```

1인당 범죄율 `crim` 을 설명변수 x로 하고, 주택가격의 중앙값 `medv` 을 반응변수 y 로 할 때 다음에 대하여 답하시오.

In [3]:

```
### 라이브러리 불러오기
##library(MASS)
##library(lmtest)
```

(1) 선형 회귀모형 $y = \beta_0 + \beta_1 x + \epsilon$을 적합시켜라.

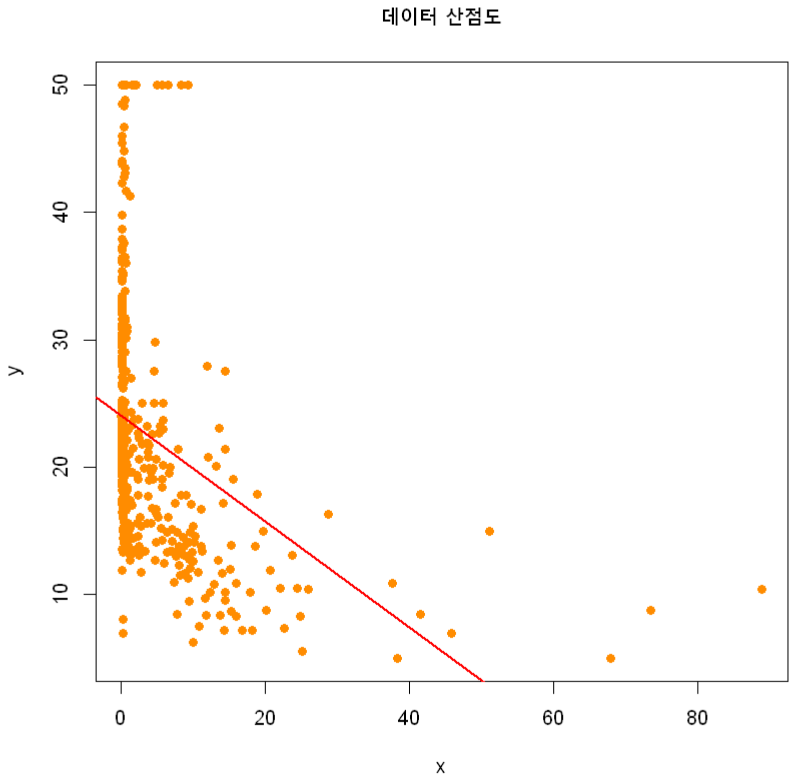
In [4]:

```
x <- Boston$crim
y <- Boston$medv

model <- lm(y~x)

summary(model)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.957  -5.449  -2.007   2.512  29.800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.03311    0.40914   58.74   <2e-16 ***
x           -0.41519    0.04389   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.484 on 504 degrees of freedom
Multiple R-squared:  0.1508,	Adjusted R-squared:  0.1491 
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

In [5]:

```
plot(y~x, pch = 16, col = 'darkorange', main = '데이터 산점도')
abline(model, lwd = 2, col = 'red')
```

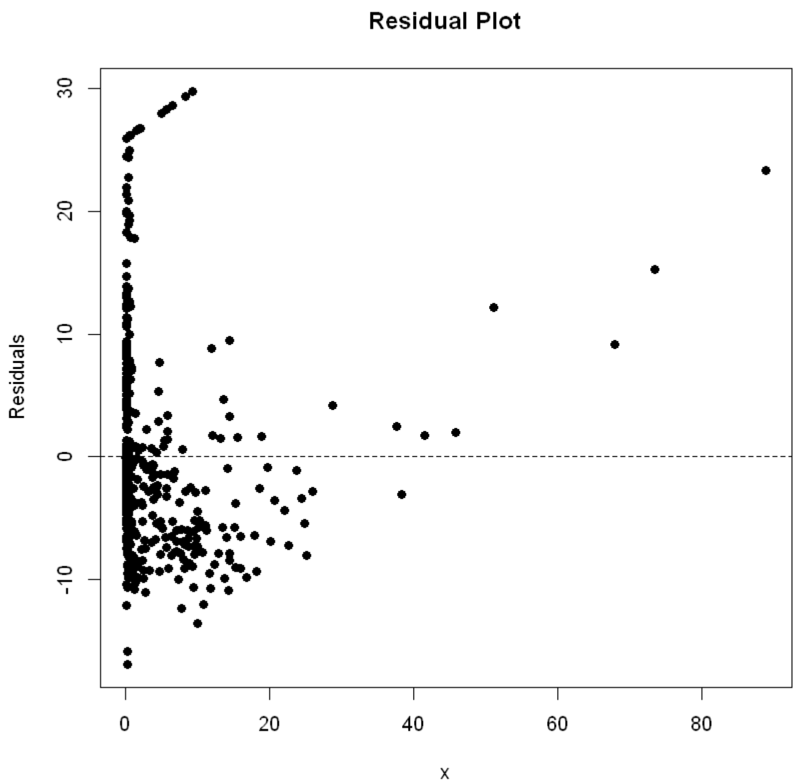


> `p-value` 로 보았을 때 모형은 통계적으로 유의하나, 잘 적합된 것 같아 보이진 않는다.

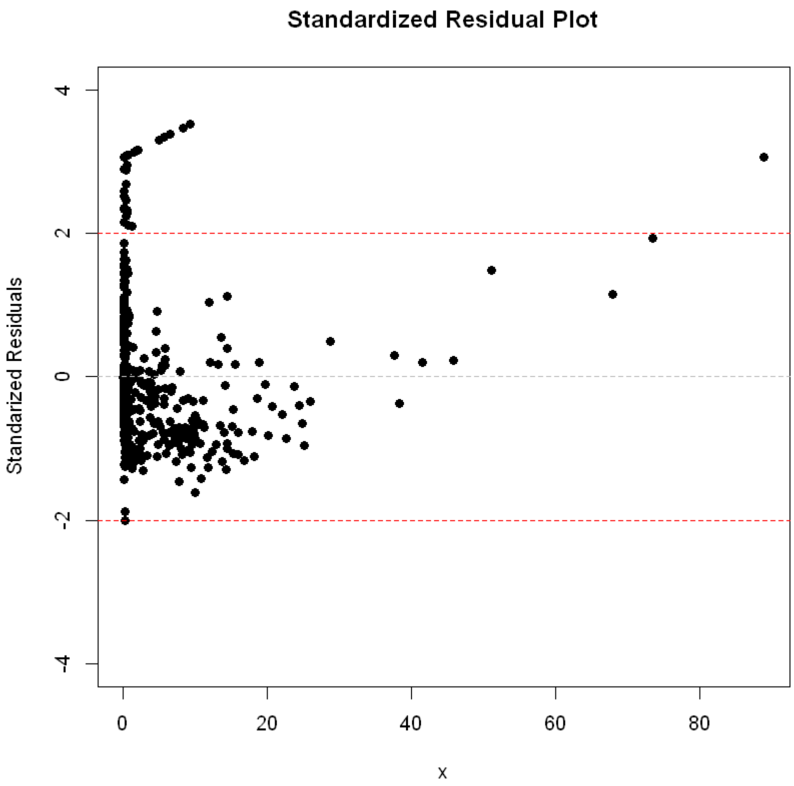
---

(2) 잔차의 산점도를 그려보고 모형의 타당성과 등분산성에 대하여 설명하시오.

In [6]:
```
plot(model$residuals ~ x, pch = 16, ylab = 'Residuals', main = 'Residual Plot')
abline(h = 0, lty = 2)
```



In [7]:
```
plot(rstandard(model) ~ x, pch = 16, ylab = 'Standarized Residuals', ylim = c(-4
abline(h = c(-2, 0, 2), col = c('red', 'grey', 'red'), lty = 2)
```



i) 선형성 : 모형이 0을 중심으로 대칭인 것처럼 보이진 않는다.

ii) 등분산성 : 설명변수 x값에 관계없이 잔차의 분산이 일정해야 하는데, 이 경우 x값이 작을수록 잔차의 분산이 커지고 있다. 따라서 모형의 등분산성 가정은 옳지 못하다.

iii) 정규성 : 표준화된 잔차를 산점도로 나타냈을 때 2를 넘는 값이 많이 관측되며, 3을 넘는 값들도 꽤 많이 보인다. 이에 따라 정규성을 가정하기 어려울 것으로 예상된다.

iv) 독립성 : 잔차가 x의 값에 따라 비선형의 패턴이 있어보인다. 이에 따라 독립성을 가정하기 어려울 것으로 예상된다.

> 따라서 선형회귀의 기본가정에 위배되므로, 단순선형회귀로 적합한 모형은 타당하지 않은 것 같다.

---

(3) 유의수준 $\alpha = 0.1$에서 $H_0 : \beta_1 = 0.3$, $H_1 : \beta_1 \neq 0.3$을 검정하시오

In [8]:
```
## T-통계량
t_0 <- (as.numeric(model$coefficients[2]) - 0.3)/summary(model)$coefficients[2,2
print(paste('T-statiscic =', t_0))

## p-value
p_value <- pt(q = t_0, df = length(x-2)) + pt(q = -t_0, df = length(x-2), lower.
print(paste('p_value =', p_value))

## 검정
print(paste('Result of \'p_value < 0.1\' :', p_value < 0.1,'// So reject null hy
```

```
[1] "T-statiscic = -16.2949207615693"
[1] "p_value = 2.69319568160759e-48"
[1] "Result of 'p_value < 0.1' : TRUE // So reject null hypothesis"
```

> 유의수준 $\alpha = 0.1$에서 `p_value`가 α보다 작으므로, 귀무가설을 기각하고, 대립가설을 수용한다. 즉, β_1은 0.3이 아니다.

---

(4) $Durbin - Watson\ d$통계량을 사용하여 $H_0 : \rho = 0, H_0 : \rho > 0$을 유의수준 $\alpha = 0.05$에서 검정하시오

In [10]:
```
dwtest(model, alternative = 'greater')
```

```
	Durbin-Watson test

data:  model
DW = 0.71342, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

> `p-value`가 $\alpha = 0.05$보다 작기 때문에 귀무가설을 기각한다. 따라서 잔차는 1차 양의 자기상관을 지닌다.