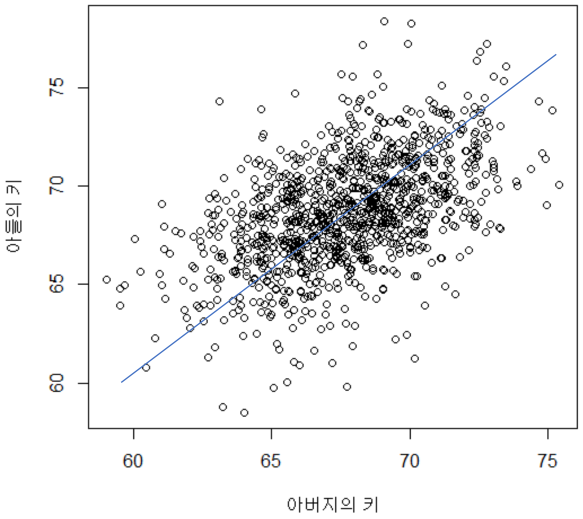
컴퓨터응용통계

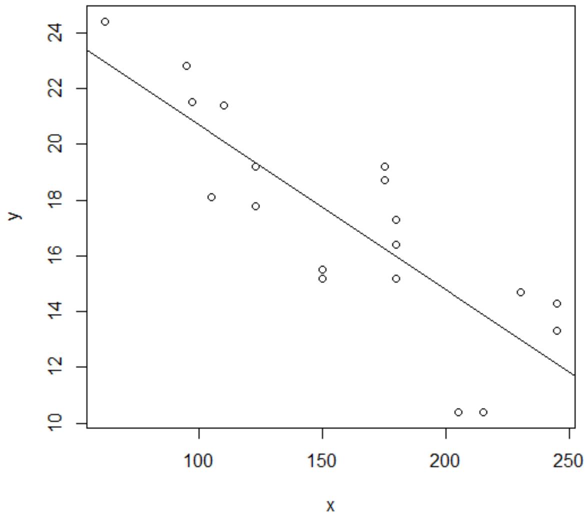
9장 회귀분석

최경미

단순회귀분석 (simple linear regression)

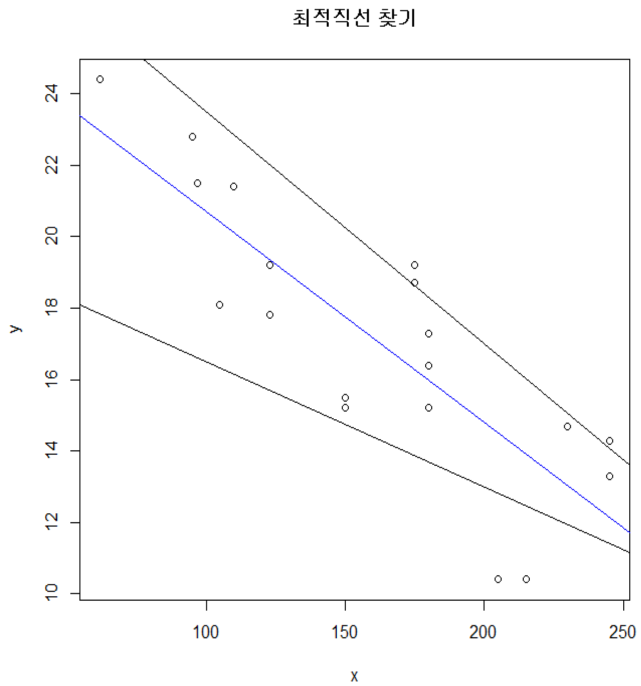
- 두 변수 X, Y가 모두 연속형일 때, 두 변수 사이의 인과관계를 직선으로 표현한다.
- 아버지의 키가 아들의 키에 어느 정도 영향을 미치는가?
- 자동차의 마력이 연비에 어느 정도 영향을 미치는가?





예제 8.1 마력과 연비

마력 x	110	175	105	245	62	95	123	123	180	180
연비 y	21.4	18.7	18.1	14.3	24.4	22.8	19.2	17.8	16.4	17.3
마력 x	180	205	215	230	97	150	150	245	175	
연비 y	15.2	10.4	10.4	14.7	21.5	15.5	15.2	13.3	19.2	



$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_n, y_n)$$

$$Y_i = \alpha + \beta x_i + \varepsilon_i,\ \mathrm{i} = 1, \cdots, \mathrm{n}$$

단순회귀모형

(x_1, y_1), (x_2, y_2), ⋯, (x_n, y_n)

$Y_i = \alpha + \beta x_i + \varepsilon_i$, $\mathrm{i} = 1, \cdots, \mathrm{n}$

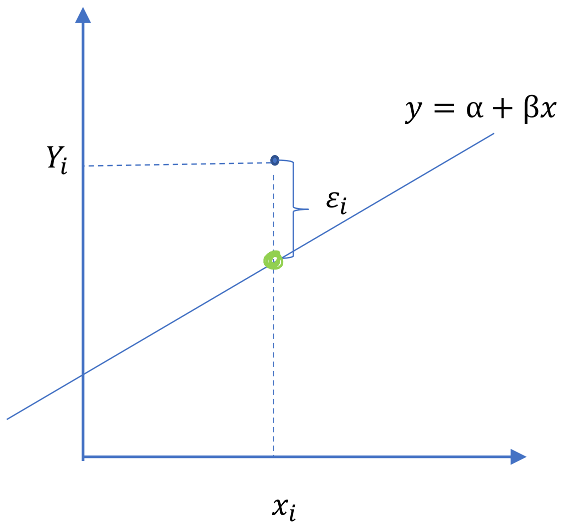
x_i : 설명변수(explanatory variable), 독립변수(independent variable)

Y_i : 반응변수(response variable), 종속변수(dependent variable)

ε_i : 오차(errors)

α : y-절편(intercept)

β : 기울기(slope)



- 오차에 대한 가정:

"오차는 독립이고, 동일한 $\mathrm{N}(0, \sigma^2)$를 따른다."

$$E[\varepsilon_i] = 0, Var(\varepsilon_i) = \sigma^2$$

① 독립성 ② 정규성 ③ 등분산성

- 모형 $\mathrm{E}[\mathrm{Y_i}|x_i]$의 선형성

ε_i와 Y_i는 확률변수

x_i는 확률변수가 아닌 주어진 값

$$\mathrm{E}[\mathrm{Y_i}|x_i] = E[\alpha + \beta x_i + \varepsilon_i] = \alpha + \beta x_i + E[\varepsilon_i] = \alpha + \beta x_i$$

회귀분석 순서

①회귀계수 α, β를 추정하자.

②최적 직선식을 추정하자.

③선형 회귀모형의 적합성 검정하자.

④계수 추정값의 유의성 검정하자.

⑤주어진 점 x에서 Y의 평균반응에 대한 신뢰구간과 예측구간을 구하자.

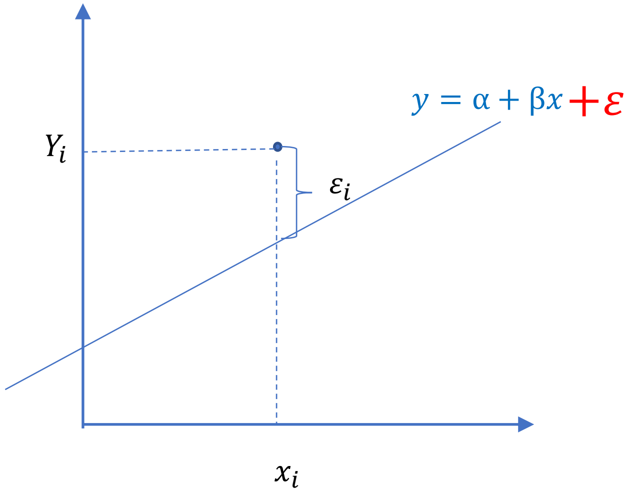
⑥잔차도를 이용하여 모형의 적합도를 살펴보자.

⑦이상점 및 영향점의 존재를 파악해보자.

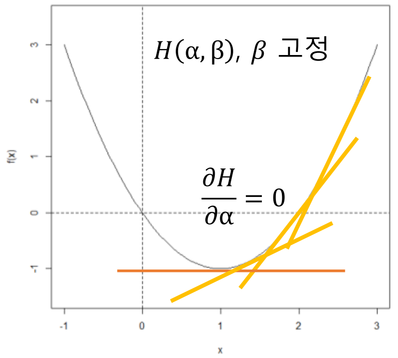
⑧여러 개의 설명변수를 사용하는 다중회귀분석을 사용하자.

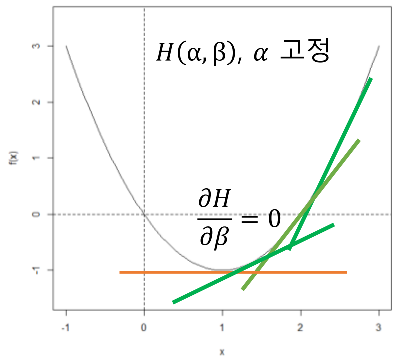
⑨변수 또는 모형을 선택하자.

9.2 최소제곱법 (Least Squares Method)



미분계수=접선의 기울기





오차 제곱합

$$H(\alpha,\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

H가 최소가 되는 α와 β를 찾아보자.

$$\frac{\partial H}{\partial \alpha} = 0 \ (\beta \text{ 고정})$$

$$\frac{\partial H}{\partial \beta} = 0 \ (\alpha \text{ 고정})$$

$$\frac{\partial H}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial H}{\partial \beta} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \alpha \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

최소제곱추정치 (Least Squares Estimators)

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

추정값 (predictors)

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

잔차 (residuals) 잔차 = 관측값 – 추정값

$$e_i = y_i - \hat{y}_i$$

- 통계와 머신러닝의 차이... 통계는 수학적인 최적값을 찾고, 머신러닝은 근사적인 최적값을 찾는다.

R을 이용한 계산

```
# 표 9.1
# 방법 1
 x<- mtcars[mtcars$am==0,"hp"]
 y<- mtcars[mtcars$am==0,"mpg"]

# 방법 2
 auto <- subset(mtcars, am==0)
 y<-auto$mpg
 x<-auto$hp

# 방법 3
 y<-c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 21.5, 15.5, 15.2, 13.3, 19.2)
 x<-c(110, 175, 105, 245,   62,   95, 123, 123, 180, 180, 180, 205, 215, 230,   97, 150, 150, 245, 175)
```

예제 9.2 계산 (R 계산)

```
x<- mtcars[mtcars$am==0,"hp"]
y<- mtcars[mtcars$am==0,"mpg"]

> fit <- lm(y~x)
> fit
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
  26.62485     -0.05914
```

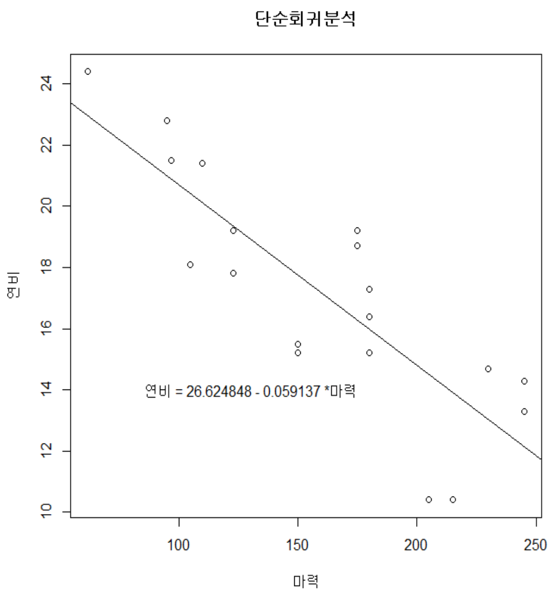
Y절편 $\hat{\alpha}$

기울기 $\hat{\beta}$

- 그래프 그리기

```
plot(x,y)

abline(fit)
```



$\hat{\alpha} = 26.625$

$\hat{\beta} = -0.059$

$\therefore$ 연비 $= 26.625 - 0.059 \times$ 마력

9.3 분산분석을 이용한 회귀모형의 적합도 검정 (Goodness-of-fit)

H_0: 회귀모형이 유의하지 않다.

H_1: 회귀모형이 유의하다.

H_0: $Y_i = \alpha + \varepsilon_i$

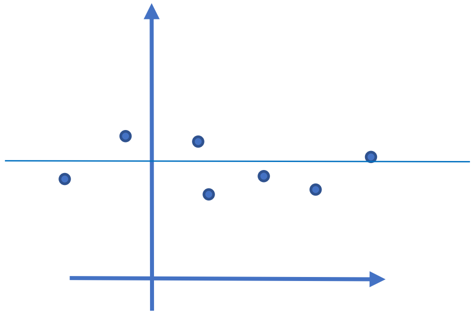
H_1: $Y_i = \alpha + \beta x_i + \varepsilon_i$

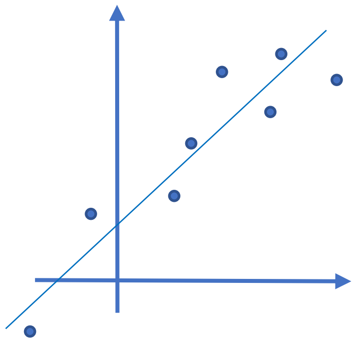
H_0: $\beta = 0$

H_1: $\beta \neq 0$

H_0: 모든 회귀 계수(β)가 0이다.

H_1: 0이 아닌 회귀계수(β)가 존재한다.





자료의 변동

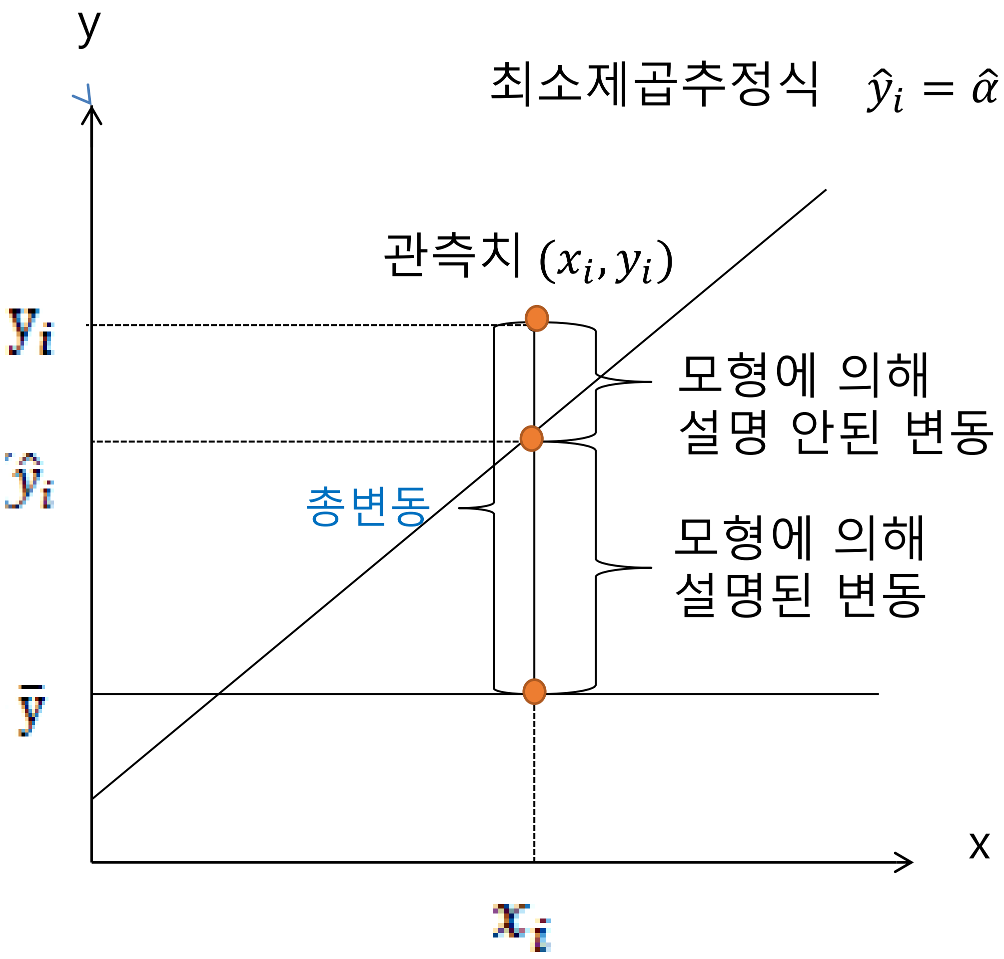
회귀모형 $Y_i = \alpha + \beta x_i + \varepsilon_i,\ \varepsilon_i \sim \text{iid N}(0, \sigma^2)$

최소제곱추정식 $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

관측값
(observed value)

추정값
(fitted value,
predicted value)

평균값
mean



분산과 총변동

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

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

변동의 분해

$$y_i - \bar{y} \quad = \quad (\hat{y}_i - \bar{y}) \quad + \quad (y_i - \hat{y}_i)$$

$$\text{총변동} = \begin{pmatrix}\text{회귀모형에 의해서}\\ \text{설명된 변동}\end{pmatrix} + \begin{pmatrix}\text{회귀모형에 의해서}\\ \text{설명이 안된 변동}\end{pmatrix}$$

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

$$SST \quad = \quad SSR \quad + \quad SSE(RSS)$$

분산 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

분산의 분포 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

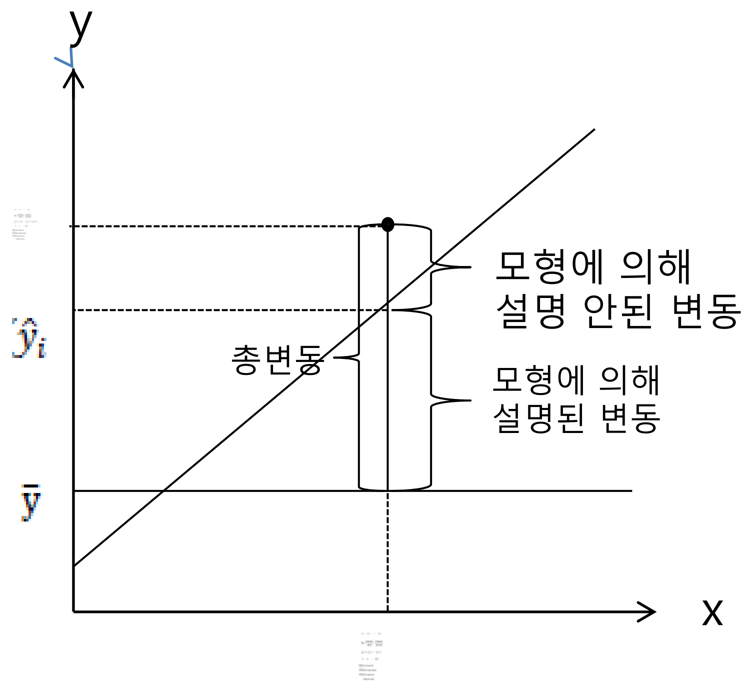
∴ **총변동의 분포** $\frac{SST}{\sigma^2} \sim \chi^2(n-1)$

총제곱합 SST(Sum of squares for total)

회귀제곱합 SSR(Sum of squares for regression)

오차제곱합 SSE(Sum of squares for error)

RSS(Residual sum of squares)



결정계수

- $SST \quad = \quad SSR \quad + \quad SSE$
- R^2은 총변동 중 회귀직선에 의해 설명된 비율이다.
- $R^2 = \frac{SSR}{SST}$
- R^2이 클수록 회귀모형이 자료의 변동을 잘 설명한다.
- $0 \le R^2 \le 1$
- 독립변수의 수가 증가하면 R^2이 커진다.
- R^2의 증가가 둔화되는 지점에서 적절한 독립변수의 수를 대략적으로 짐작할 수 있다.
- SSR과 SST가 독립이 아니어서 R^2이 확률분포를 갖지 않는다.
- R^2을 이용하여 모형에 대한 검정을 실시할 수 없고, 모형을 결정하기 어렵다.

제곱합의 분포와 자유도

- 귀무가설이 참일 때 제곱합의 분포는 다음과 같다

$$\left(\frac{SST}{\sigma^2} \sim \chi^2(n-1)\right) = \left(\frac{SSR}{\sigma^2} \sim \chi^2(1)\right) + \left(\frac{SSE}{\sigma^2} \sim \chi^2(n-2)\right)$$

- 회귀모형의 자유도 df_R은 y-절편을 제외한 설명변수의 개수이다. 단순회귀분석에서는 설명변수가 1개이므로, 단순회귀모형에서 회귀모형의 자유도는 항상 1이다.

- 회귀모형의 자유도 df_R과 오차의 자유도 df_E를 더하면, 총 자유도 df_T를 얻을 수 있다.

$$df_T = df_R + df_E$$

- 총자유도는 $df_T = (\text{표본의 크기} - 1) = n - 1$이다.

- 오차의 자유도df_E는 SSE의 자유도는 SST의 자유도와 SSR의 자유도의 차이다.

$$df_E = df_T - df_R = \left(\text{표본의 크기} - 1\right) - \text{설명변수의 개수} = \left(\text{단순회귀모형}\right) n - 2$$

평균제곱합

- 평균제곱합(Mean Squares; MS)은 제곱합을 자유도로 나눈 것으로 정의된다 ($MS = SS/df$). 오차 분산 σ^2의 추정값으로 MSE를 사용하며, 다음과 같다.
- $\hat{\sigma}^2 = MSE$
- $\mathrm{E}[MSE] = \sigma^2$
- $\hat{\sigma} = s = \sqrt{MSE} = \text{residual standard error}$

검정통계량

- $\left(\frac{SST}{\sigma^2} \sim \chi^2(n-1)\right) = \left(\frac{SSR}{\sigma^2} \sim \chi^2(1)\right) + \left(\frac{SSE}{\sigma^2} \sim \chi^2(n-2)\right)$
- 선형회귀모형의 적합도 검정을 위한 검정통계량 F은 귀무가설이 참일 때 다음과 같다.

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2)$$

- $F \geq F_\alpha(1, n-2)$, 또는 "$p-$값 $\leq$ 유의수준 $\alpha = 0.05$"이면, 귀무가설 H_0: $\beta = 0$을 기각하고, 단순회귀모형이 유의하다고 결론짓는다.

분산분석표 (ANOVA table, Analysis of Variance table)

요인	제곱합(SS)	자유도 (df)	평균제곱합 $MS = SS/df$	검정통계량 F	유의확률 $p-$값
회귀모형	SSR	df_R=X 변수 개수=1	$MSR = SSR/1$	$F = MSR/MSE$	$p \leq \alpha$ 이면
잔차	SSE	$df_E = df_T - df_R$ $= \mathrm{n} - 2$	MSE $= SSE/(n-2)$	$\sim F(1, n-2)$ (H_0이 참일 때)	"H_0: 회귀계수 = 0" 을 기각하고 회귀모형이 유의하다 결론지음
총합	SST	df_T = 표본수 $- 1$ $= \mathrm{n} - 1$			

오차의 분산 추정치 $\hat{\sigma}^2 = \mathrm{MSE}$ 결정계수 $R^2 = SSR/SST$= 회귀모형에 의해서 설명된 변동의 비율

R의 분산분석표에는 총합이 출력되지 않는다.

R 추정

```r
x<- mtcars[mtcars$am==0,"hp"]
y<- mtcars[mtcars$am==0,"mpg"]
fit <- lm(y~x)
```

```
> anova(fit)
Analysis of Variance Table

Response: y
          Df  Sum Sq  Mean Sq  F value   Pr(>F)
x          1 182.937  182.937   38.088  1.025e-05 ***
Residuals 17  81.651    4.803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p=1.025× $10^{-5} < 0.05$이므로, 회귀모형이 유의하다.

H_0: 회귀모형이 유의하지 않다. H_1: 회귀모형이 유의하다.

$H_0: Y_i = \alpha + \varepsilon_i$ $H_1: Y_i = \alpha + \beta x_i + \varepsilon_i$

$H_0: \beta = 0$ $H_1: \beta \neq 0$

H_0: 모든 회귀 계수(β)가 0이다. H_1: 0이 아닌 회귀계수(β)가 존재한다.

요인	제곱합 SS	자유도 df	평균제곱합 MS = SS/df	검정통계량 F	유의확률 p
회귀 (마력hp)	182.937	1	182.937	F=38.088	1.025e-05
잔차 (residual)	81.651	17	4.803		
총합	264.588	18			

$\hat{\sigma} = \sqrt{MSE}$ =Residual standard error: 2.192 on 17 degrees of freedom

R^2 =Multiple R-squared: 0.6914, Adjusted R-squared: 0.6733

요인	제곱합(SS)	자유도 (df)	평균제곱합 $MS = SS/df$	검정통계량 F	유의확률 $p-$값
회귀모형	SSR	df_R=X 변수 개수=1	$MSR = SSR/1$	$F = MSR/MSE$	$p \le \alpha$ 이면
오차	SSE	$df_E = df_T - df_R = n-2$	$MSE = SSE/(n-2)$	$\sim F(1, n-2)$ (H_0이 참일 때)	"H_0: 회귀계수 = 0"을 기각하고 회귀모형이 유의하다 결론 지음
총합	SST	df_T = 표본수 − 1 = n − 1			

오차의 분산 추정치 $\hat{\sigma}^2$ = MSE 결정계수 $R^2 = SSR/SST$= 회귀모형에 의해서 설명된 변동의 비율

R의 분산분석표에는 총합이 출력되지 않는다.

9.4 추정된 계수의 유의성 검정

$Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \cdots, n$, ε_i~iid N(0, σ^2)

(절편의 유의성) 가설 H_0: $\alpha = 0$ H_1: $\alpha \neq 0$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\mathrm{E}[\hat{\alpha}] = \alpha$$

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$se(\hat{\alpha}) = \sqrt{\widehat{\mathrm{Var}(\hat{\alpha})}} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

$t = \frac{\hat{\alpha}}{se(\hat{\alpha})} \sim t(df_{Error})$ under H_0

(기울기의 유의성) 가설 H_0: $\beta = 0$ H_1: $\beta \neq 0$

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

$$\mathrm{E}[\hat{\beta}] = \beta$$

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$$

$$se(\hat{\beta}) = \sqrt{\widehat{Var}(\hat{\beta})} = \frac{\sqrt{MSE}}{\sqrt{S_{xx}}}$$

$t = \frac{\hat{\beta}}{se(\hat{\beta})} \sim t(df_{Error})$ under H_0

계수의 유의성 검정

(절편의 유의성) 가설 H_0: $\alpha = 0 \quad H_1: \alpha \neq 0$

(기울기의 유의성) 가설 H_0: $\beta = 0 \quad H_1: \beta \neq 0$

	추정값 (Estimate)	표준오차 (Std. Error)	검정통계량 t	유의확률 $p = \Pr(> \lvert t \rvert)$
절편(Intercept	$\hat{\alpha}$	$se(\hat{\alpha})$	$t = \dfrac{\hat{\alpha} - \alpha}{se(\hat{\alpha})}$	$p \leq \alpha$이면 $H_0: \alpha = 0$ 을 기각한다.
x	$\hat{\beta}$	$se(\hat{\beta})$	$t = \dfrac{\hat{\beta} - \beta}{se(\hat{\beta})}$	$p \leq \alpha$이면 $H_0: \beta = 0$ 을 기각한다.

$\hat{\sigma} = \sqrt{MSE}$ = 잔차에 대한 표준오차 (Residual standard error). 자유도=(n-1)-x의 갯수

R^2 =결정계수 (Multiple R-squared) =SSR/SST, 조정된 결정계수 (Adjusted R-squared)

se: 모수에 대한 표준오차 (standard error)

예제 마력과 연비.

```
# 예제 9.4 계수추정 (parameter estimates)과 적합도 검정
> summary(fit)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q   Median     3Q     Max
-4.1018 -1.9026  0.6114  1.5592  2.9241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.624848   1.615883  16.477 6.92e-12 ***
x           -0.059137   0.009582  -6.172 1.02e-05 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.192 on 17 degrees of freedom
Multiple R-squared:  0.6914,   Adjusted R-squared:  0.6733
F-statistic: 38.09 on 1 and 17 DF,  p-value: 1.025e-05
```

모형

잔차에 대한 기술통계량

절편

추정값, 표본오차, 검정통계량, 유의확률

$\hat{\sigma} = \sqrt{MSE}$와 자유도

설명변수

R^2

모형에 대한 검정통계량과 유의확률

$p < \alpha = 0.05$이면, 모형이 유의하다.

```
# 그림 9.3
> plot(x,y, main="단순회귀분석",
       xlab="마력", ylab="연비",
       text(130,14,"연비 = 26.624848 - 0.059137 *마력") )
> abline(fit)
```

	추정값(Estimate)	표준오차(Std. Error)	검정통계량 (t value)	유의확률 Pr(>|t|)
절편(Intercept)	26.624848	1.615883	16.477	6.92e-12 ***
x	-0.059137	0.009582	-6.172	1.02e-05 ***

$\hat{\sigma} = \sqrt{MSE}$ =Residual standard error: 2.192 on 17 degrees of freedom

R^2 =Multiple R-squared: 0.6914, Adjusted R-squared: 0.6733

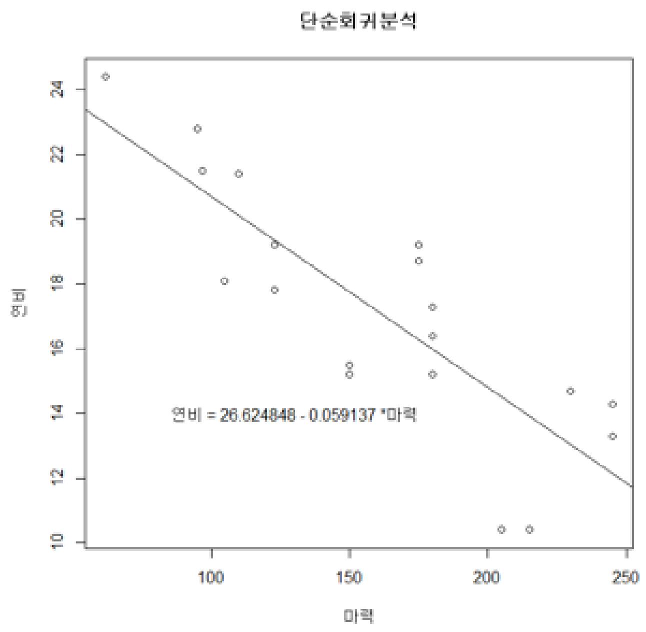


그림 9.3 마력과 연비에 대하여 추정된 단순회귀모형

9.5 잔차도를 이용한 모형 진단

- 오차의 독립성, 정규성, 등분산성
- 잔차 $e_i = Y_i - \hat{Y}_i$
- 표준화된 잔차(standardized residual)

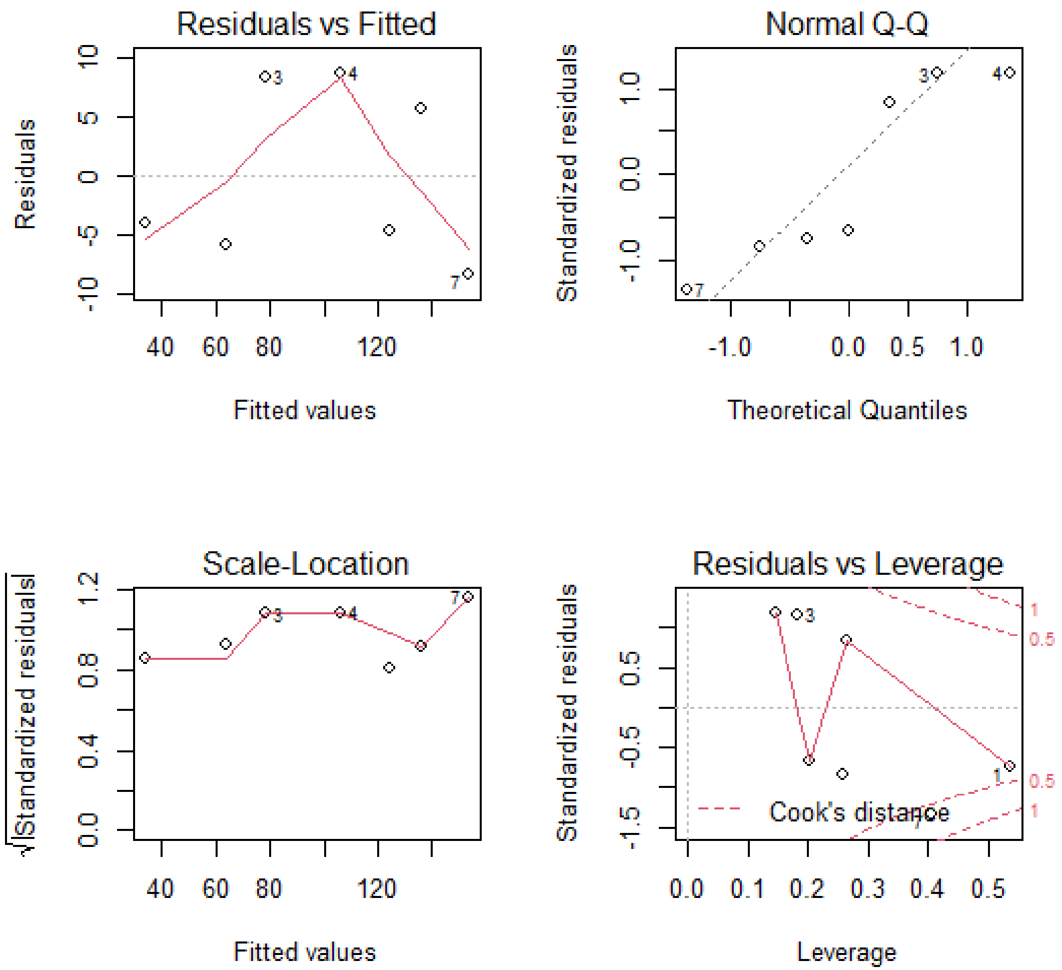
$$r_i = \frac{e_i}{\widehat{sd}(e_i)} \sim \text{근사적으로 } N(0,1)$$

① 잔차는 0 주변에 대칭적으로 놓여있다.

② 대부분의 잔차는 ±2 근방 이내로 놓여있다.

③ 규칙적인 패턴이 없다.

```
> par(mfrow=c(2,2))
> plot(fit)
```



QQ-plot

- QQ-plot은 이론적인 정규분포로부터 생성된 값과 회귀모형으로 추정된 잔차를 크기 순서대로 짝을 지워서 산점도로 나타낸 그래프이다.
- QQ-plot이 일직선이면 잔차가 정규분포를 따른다고 볼 수 있다.

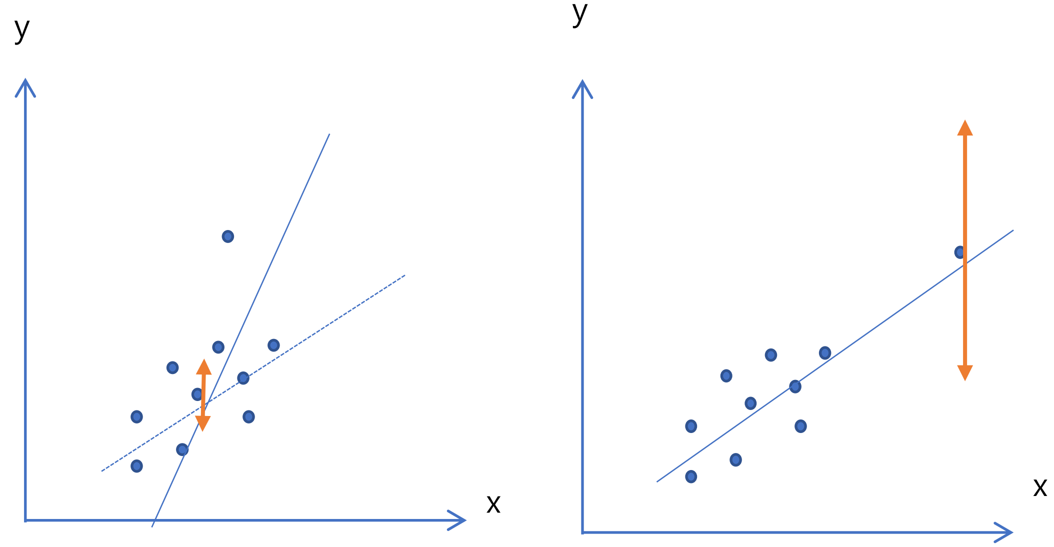
이상점과 영향점

- ① 쿡스 거리 D_i는 i번째 관측치가 있을 때와 없을 때 전체 추정값이 얼마나 달라지는지를 계산한 값이다.

  $D_i > \frac{4}{n}$ 또는 $D_i > \frac{4}{n-p-1}$

- ② 한 점에서 y_i를 아래 위로 움직일 때, $\hat{y}_i$의 변화율을 레버리지라고 부른다. x_i가 $\bar{x}$로부터 멀리 떨어질수록 큰 지렛대 효과가 발생하여, 레버리지 점수가 커진다.

  $h_i > \frac{2(p+1)}{n}$



보고서

마력과 연비에 대한 회귀분석

R의 mtcars 중 자동 트랜스미션 차 19대를 이용하여, 단순회귀모형을 이용하여, 마력(x)이 연비(y)를 어떻게 설명할 수 있는지 살펴보자. 이때 유의수준 0.05를 사용한다.

표 1의 분산분석표에서 검정통계량 F=38.088에 대한 유의확률 p=1.025 $\times 10^{-5}$이 유의수준 0.05보다 작으므로, $H_0: \beta = 0$ 또는 H_0: 단순회귀모형이 유의하지 않다를 기각한다. 즉, 추정된 단순회귀모형이 적합하며 유의하다. 모형의 결정계수가 $R^2 = \frac{182.937}{264.588} = 0.6914$ 이므로, 마력은 연비의 총변동 중 69.14 %를 설명한다.

표 1. 분산분석표

요인	제곱합 SS	자유도 df	평균제곱합 MS = SS/df	검정통계량 F	유의확률 p
회귀 (마력hp)	182.937	1	182.937	F=38.088	1.025e-05
잔차 (residual)	81.651	17	4.803		
총합	264.588	18			

표 2의 계수추정표로부터 구한 추정된 회귀직선식은 다음과 같다 (그림1).

$$\text{연비} = 26.62485 - 0.05914 \times \text{마력}$$

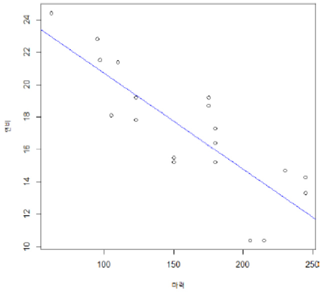


그림 1. 마력과 연비의 회귀분석모형

여기서, 절편에 대한 유의확률 $p = 6.92 \times 10^{-12}$이 유의수준 0.05보다 작으므로, 절편은 유의하다.

기울기에 대한 유의확률 $p = 1.02 \times 10^{-5}$이 유의수준 0.05보다 작으므로, 기울기가 유의하다. 즉, 마력이 연비에 유의하게 영향을 미치며, 마력이 1 증가하면, 연비가 0.05914 (마일/갤런) 감소한다. 잔차의 표준오차는 $\hat{\sigma} = \sqrt{MSE} = 2.192$이다.

표 2 계수추정표

	추정값 (Estimate)	표준오차 (Std. Error)	검정통계량 (t value)	유의확률 Pr(>|t|)
절편(Intercept)	26.624848	1.615883	16.477	6.92e-12 ***
마력	-0.059137	0.009582	-6.172	1.02e-05 ***

$\hat{\sigma} = \sqrt{MSE}$ =Residual standard error: 2.192 on 17 degrees of freedom

R^2 =Multiple R-squared: 0.6914, Adjusted R-squared: 0.6733

그림2에서 잔차의 QQ-plot을 보면, 잔차가 대체로 정규분포를 벗어나지 않음을 볼 수 있다. 잔차에 대한 샤피로의 정규성검정에서 유의확률 p=0.129이므로, 단순회귀모형의 정규성 가정이 성립함을 알 수 있다.

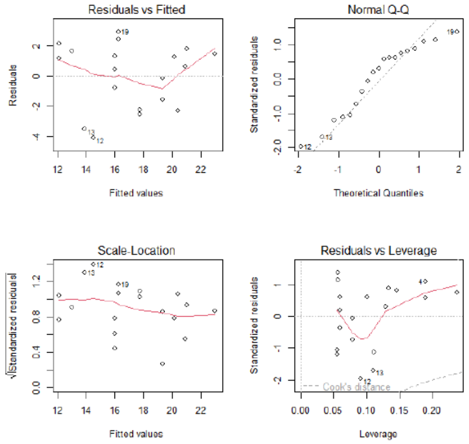


그림 2. 잔차도

부록

```
> x<- mtcars[mtcars$am==0,"hp"]
> y<- mtcars[mtcars$am==0,"mpg"]
> fit <- lm(y~x)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 182.937 182.937  38.088 1.025e-05 ***
Residuals 17  81.651   4.803
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1018 -1.9026  0.6114  1.5592  2.9241

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.624848   1.615883  16.477 6.92e-12 ***
x           -0.059137   0.009582  -6.172 1.02e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.192 on 17 degrees of freedom
Multiple R-squared:  0.6914,	Adjusted R-squared:  0.6733
F-statistic: 38.09 on 1 and 17 DF,  p-value: 1.025e-05

>   plot(x,y, main="자동 트랜스미션 차들의 마력과 연비", xlab="마력", ylab="연비")
> abline(fit)
> par(mfrow=c(2,2))
> plot(fit)
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.92307, p-value = 0.129
```

과제. 보고서와 문제 풀기

- 2. (1)-(2) 문맹이 수명에 미치는 영향을 알아보기 위하여, 미국 50개 주의 문맹률 % (1970년)과 기대수명 (세)(1969–71년)을 조사하였다. (자료: state.x77 in R, U.S. Department of Commerce, Bureau of the Census (1977)) 단순회귀분석을 실시하여 아래의 표2과 표3를 얻었다.

```
mydata<-data.frame(state.x77)

fit <-lm(Life.Exp ~ Illiteracy, data=mydata)

anova(fit)

summary(fit)

plot(mydata$Illiteracy, mydata$Life.Exp, xlab="문맹률 %", ylab="기대수명")

abline(fit)
```

- 표2. 분산분석: 종속변수 y는 기대수명 (세) , 독립변수 x는 문맹률 (%)

요인	제곱합	자유도	평균 제곱	F	유의확률
회귀 모형	30.578				6.969e-06
잔차					
합계	88.299	49			

표3. 계수추정표

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	72.3949	0.3383	213.973	< 2e-16 ***
x	-1.2960	0.2570	-5.043	6.97e-06 ***

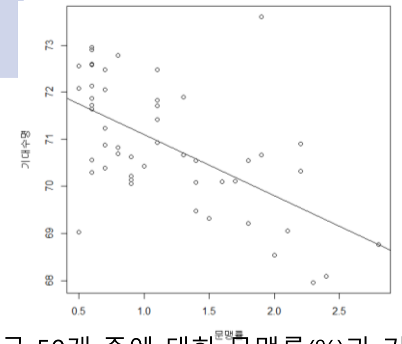


그림2. 미국 50개 주에 대한 문맹률(%)과 기대수명((세)의 산점도와 회귀직선

(1) 표2에 대한 설명으로 틀린 것은 어느 것인가?

① 결정계수 R^2= 0.346 이므로 총변동 중 회귀모형이 설명하는 변동은 34.6%이다.

② H_0: $\beta = 0$에 대한 검정통계량은 F=25.4이다.

③ 잔차의 평균제곱합은 57.721이다.

④ 유의수준 0.05에서 귀무가설을 기각하므로 <그림1>의 직선 모형이 유의하다.

⑤ 위 보기 중 답 없음

(2) 표3에 대한 설명으로 틀린 것은 어느 것인가?

① H_0: $\alpha = 0$에 대한 검정통계량은 t=213.973이고, 유의수준 0.05에서 y절편이 0이 아니다.

② 유의수준 0.05에서 H_0: $\beta = 0$ 에 대한 p 값이 0.05보다 작으므로, 문맹률이 수명에 유의하게 영향을 미친다고 볼 수 있다.

③ 표2의 F-검정통계량과 표3에서 H_0: $\beta = 0$에 대한 t-검정통계량의 제곱은 동일하다.

④ 유의수준 0.05에서 문맹률이 1%감소할 때 수명이 1.296세 감소한다고 할 수 있다.

⑤ 위 보기 중 답 없음