통계 분석

2019년 2학기

강봉주

다중 회귀 분석

회귀 분석

[개요]

- 독립변수가 2개 이상인 다중 선형회귀(multiple linear regression)
- 모형식: $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,\ i = 1, \dots, n$
- 오차항에 대한 가정: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

회귀 분석

[모수의 추정]

- $\{(x_i, y_i) | i = 1, \ldots, n\}$: 표본
- $$\begin{gathered} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ \ldots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \epsilon_n \end{gathered}$$
- 행렬 표현식: $y = \mathrm{X}\beta + \epsilon$
- $y = (y_1, \ldots, y_n)^T, \mathrm{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ . & . & . & . \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta = (\beta_0, \beta_1, \ldots, \beta_k)^T, \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$

회귀 분석

[모수의 추정]

- $\underset{\beta}{\text{minimize}}\left((y - \mathrm{X}\beta)^T (y - \mathrm{X}\beta)\right)$
- β에 대하여 미분하여 정리
- $\mathrm{X}^{\mathrm{T}}\mathrm{X}\beta = \mathrm{X}^{\mathrm{T}}y$: 정규 방정식
- $\hat{\beta} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}y$: 역행렬이 존재하면.
- 역행렬을 이용하지 않은 경우: QR 분해 기법 이용

회귀 분석

[모수의 추정]

- QR 분해 기법에 의한 추정

```
# 회귀계수값 구하기
betahat <- qr.solve(t(X) %*% X,
t(X) %*% y)
# betahat은 열 벡터
# dim(betahat)
# is.matrix(betahat)

#  예측값
yhat <- X %*% betahat

# 잔차
r <- y - yhat
```

회귀 분석

[모수의 추정]

- QR 분해 기법에 의한 추정

```
# 오차 분산 추정값 구하기
n <- nrow(sdf)
p <- length(betahat)
SSE <- crossprod(r, r)
MSE <- SSE/(n-p) # 평균제곱오차
RSE <- sqrt(MSE) # Residual standard
error
print(RSE)
```

회귀 분석

[모수의 추론]

- $Y \sim \mathrm{N}(\mathrm{X}\beta, \mathrm{I}\sigma^2)$
- $\mathrm{E}(\hat{\beta}) = \mathrm{E}\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}Y\right) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{E}(Y) = \beta$
- $\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}Y\right) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Cov}(y)\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\right)^{T} =$

  $\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\{\mathrm{I}\sigma^2\}\mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\sigma^2$
- $\hat{\beta} \sim \mathrm{N}\left(\beta, \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\sigma^2\right), \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p), \hat{\beta}, \hat{\sigma}^2$ 이 서로 독립

회귀 분석

[모수의 추론]

- $\sqrt{(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}} \overset{let}{=} \mathrm{D}$
- $t_k = \frac{\hat{\beta}_k - \beta_k}{\mathrm{D}_{kk}\hat{\sigma}} \sim t(n-p)$
- $\mathrm{H}_0: \beta_k = 0$
- $|t_{k0}| \geq t_{\frac{\alpha}{2}}(n-p)$, 영가설 기각, α 는 유의 수준

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

- $F = \frac{SSR/k\sigma^2}{SSE/(n-k-1)\sigma^2} \sim F(k, n-k-1)$, k는 절편을 제외한 모수의 개수

요인	제곱합	자유도	평균제곱	F값	P값
회귀	SSR	k	$MSR = \frac{SSR}{k}$	$f_0 = \frac{MSR}{MSE}$	$\Pr(F \geq f_0)$
잔차	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$		
계	SST	$n-1$			

회귀 분석

[회귀직선의 유의성 검증: 분산 분석표]

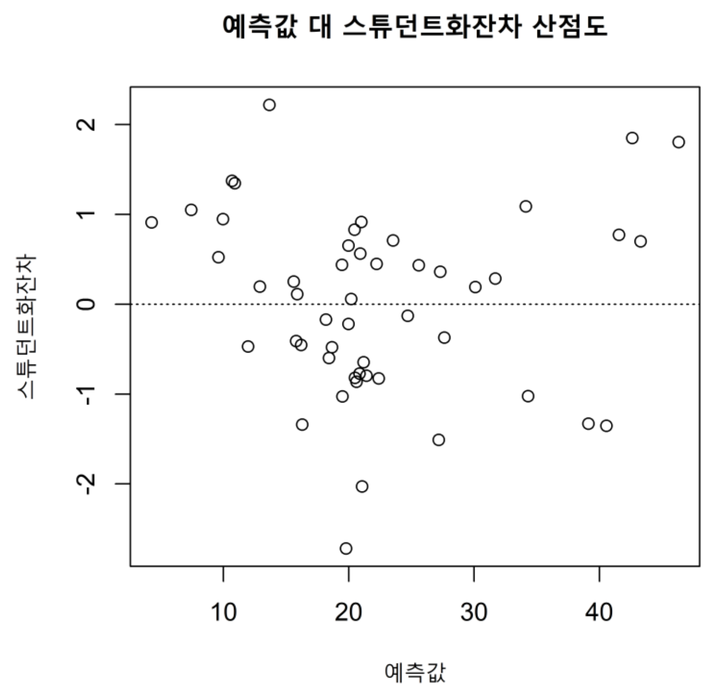
- $F = \frac{SSR/k\sigma^2}{SSE/(n-k-1)\sigma^2} \sim F(k, n-k-1)$, k는 절편을 제외한 모수의 개수

```
n <- nrow(sdf)
p <- length(fit$coefficients)
SSE <- sum(fit$residuals^2) # 잔차제곱합
MSE <- SSE/(n-p) # 평균제곱오차

SST <- (n-1)*var(y)
SSR <- SST-SSE
f0 <- (SSR/(p-1))/(SSE/(n-p))
pvalue <- 1-pf(f0, df1=(p-1), df2=n-p)
print(paste('F-statistic: ',f0, 'on ', p-1, ' and ', n-p,' DF, ', 'p-
value: ', pvalue))

#함수 이용
summary(fit)
```

회귀 분석

[오차 가정에 대한 검증]

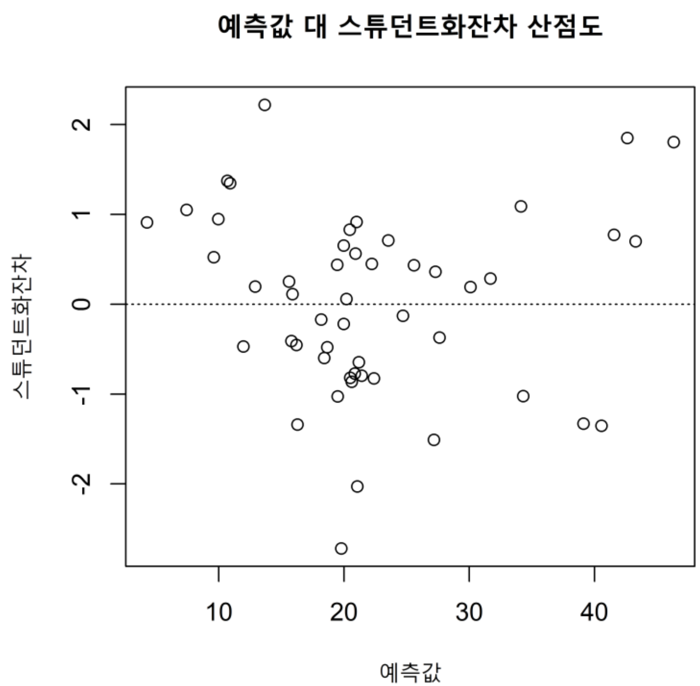
- 독립성 확인



회귀 분석

[오차 가정에 대한 검증]

- 등분산성 확인



회귀 분석

[오차 가정에 대한 검증]

- 정규성 확인