Regression Diagnostic

회귀진단

(1) 오차항의 가정 검토

(2) 적절한 모형의 선택

(3) 독립변수들간의 상관관계 검토

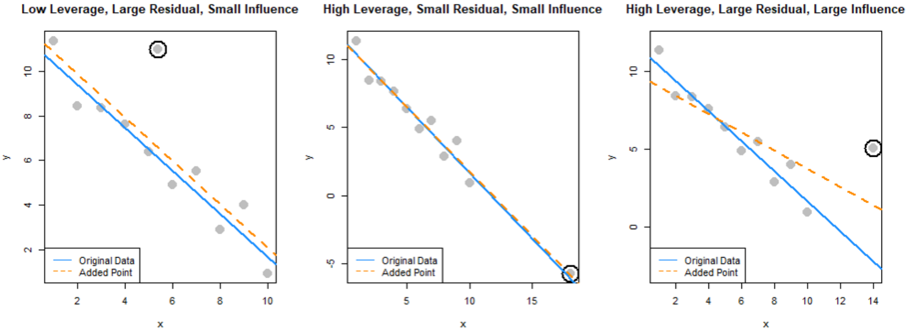
(4) 지렛대점(leverage point)의 검출

(5) 이상치(outlier) 확인

(6) 영향점(influential observation)의 검출

회귀진단

- leverge vs. outlier vs. influence point



Hat matrix

- 추정된 회귀직선

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = H\boldsymbol{y}$$

- H : hat matrix, $n \times n$ matrix
- $\mathrm{Var}(\hat{\boldsymbol{y}}) = \sigma^2 H, \quad \mathrm{Var}(\mathbf{e}) = (\mathbf{I}_n - H)\sigma^2$
- for $i, j = 1, ..., n$

$$h_{ij} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_j$$

  where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$
- h_{ii} : H의 대각 원소 (diagonal elements)

Hat matrix

$$H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

- $HH = H$, idempotent matrix, $rank(H) = p + 1$,

$$\mathsf{tr}(H) = \sum_{i=1}^{n} h_{ii} = p + 1$$

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

- H : positive definite (양반정치행렬)

$$0 \leq h_{ii} < 1, \ \ -1/2 \leq h_{ij} \leq 1/2$$

Hat matrix

- $p = 1$

$$h_{ij} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_j$$

$$= \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \dfrac{\sum x_i^2}{nS_{(xx)}} & \dfrac{-\bar{x}}{S_{(xx)}} \\[2ex] \dfrac{-\bar{x}}{S_{(xx)}} & \dfrac{1}{S_{(xx)}} \end{pmatrix} \begin{pmatrix} 1 \\[2ex] x_j \end{pmatrix} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{(xx)}}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{(xx)}}$$

Hat matrix

- $p > 1$

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathcal{X}^T \mathcal{X})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

where $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $\bar{\mathbf{x}}^\top = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$,

$\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$,

$$\mathcal{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

Leverage point

- $h_{ii} > 2\bar{h}$ 이면, i번째 관측치가 leverage point로 고려 가능

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$$

Residual

$$
\begin{aligned}
\mathbf{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= [\mathbf{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} = (\mathbf{I}_n - H)\boldsymbol{y}
\end{aligned}
$$

- $\mathrm{E}(\mathbf{e}) = \mathbf{0}_n, \ \ \mathrm{Var}(\mathbf{e}) = (\mathbf{I}_n - H)\sigma^2$
- $\mathrm{E}(e_i) = 0, \ \ \mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2, \ \ \mathrm{Cov}(e_i, e_j) = -h_{ij}\sigma^2$

$$
\mathrm{Corr}(e_i, e_j) = \rho_{ij} = \frac{-h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}}
$$

Standardized residual

- Since $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \mathbf{I}_n\sigma^2)$, $\mathbf{e} \sim N(\mathbf{0}_n, (\mathbf{I}_n - H)\sigma^2)$ and

$$e_i \sim N(0, (1 - h_{ii})\sigma^2)$$

- (내적) 표준화잔차 ((internally) standardized residual)

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \;\; \hat{\sigma}^2 = MSE$$

Studentized residual

- (외적) 스튜던트화 잔차 ((externally) studentized residual)

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

단, $\hat{\sigma}_{(i)}$: i 번째 측정값 y_i 를 제외하고 얻어진 $\hat{\sigma}$

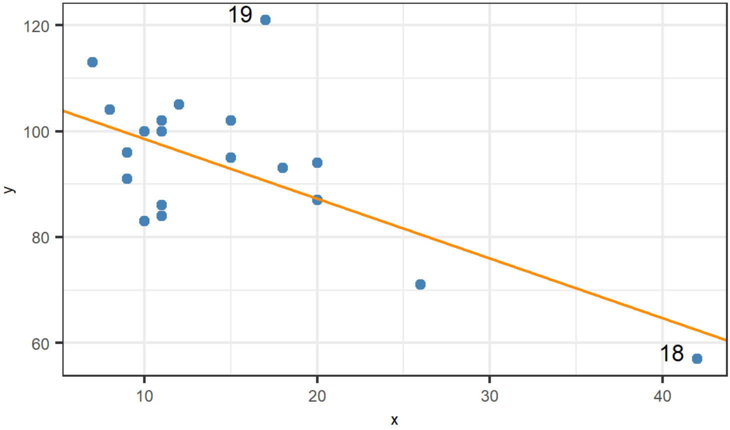
$$\hat{\sigma}_{(i)}^2 = \left[(n-p-1)\hat{\sigma}^2 - \frac{e_i^2}{1-h_{ii}}\right] / (n-p-2)$$

- $|r_i^*| \geq t_{\alpha/2}(n-p-2)$ 이면, 유의수준 α 에서, y_i 를 이상점이라고 판정

Example

실험번호	x	y	실험번호	x	y	실험번호	x	y
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

Example



Example

- 회귀직선 $\hat{y} = 109.874 - 1.127x$
- 분산분석표

요인	제곱합	자유도	평균제곱	F_0
회귀	1,604.08	1	1,604.08	12.20
잔차	2,308.59	19	121.505	
계	3,912.67	20		

- $F_0 = 13.20 > F_{0.05}(1, 19) = 4.38$ 이므로 회귀직선은 유의
- $\hat{\sigma} = \sqrt{121.505} = 11.0229$

Example

Table: 지렛대점과 이상점을 찾는 측도(h_{ii}, r_i와 $r_i{}^*$)

실험번호	e_i	h_{ii}	r_i	$r_i{}^*$
⋮	⋮	⋮	⋮	⋮
8	2.5230	0.0567	0.2357	0.2297
9	3.1421	0.0799	0.2972	0.2899
10	6.6666	0.0726	0.6280	0.6177
11	11.0151	0.0908	1.0480	1.0508
12	-3.7309	0.0705	-0.3511	-0.3428
13	-15.6040	0.0628	-1.4623	-1.5108
14	-13.4770	0.0567	-1.2588	-1.2798
15	4.5230	0.0567	0.4225	0.4132
16	1.3961	0.0628	0.1308	0.1274
17	8.6500	0.0521	0.8060	0.7983
18	-5.5403	$\boxed{0.6516}$	-0.8515	-0.8451
19	$\boxed{30.2850}$	0.0531	$\boxed{2.8234}$	$\boxed{3.6070}$
20	-11.4770	0.0567	-1.0720	-1.0765
21	1.3961	0.0628	0.1308	0.1274

Example

- $\bar{h} = \frac{p+1}{n} = \frac{2}{21} = 0.0952$
- $2\bar{h} = 0.1905, \;\; 3\bar{h} = 0.2857$
- $t_{0.025}(n-p-2) = t_{0.025}(18) = 2.101$
- 18번째 데이터 : leverage point
- 19번째 데이터 : outlier

영향점을 검출하는 방법

(1) DFFITS (Difference if Fits)

$$\mathrm{DFFITS}(i) = \frac{\hat{y}_i - \tilde{y}_i(i)}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

- $\tilde{y}_i(i)$: i번째 데이터를 제외시키고 $n-1$개 데이터에서 얻은 예측값
- $|\mathrm{DFFITS}(i)| \geq 2\sqrt{\frac{p+1}{n-p-1}} \Rightarrow$ 영향점

영향점을 검출하는 방법

(2) Cook's Distance

$$\mathrm{D}(i) = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_j(i)\right)^2}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^\top \boldsymbol{X}^\top \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))}{(p+1)\hat{\sigma}^2}$$

- $\hat{\boldsymbol{\beta}}(i)$: i 번째 관측치를 제외시키고 $n-1$ 개의 관측값에서 구한 $\boldsymbol{\beta}$ 의 최소제곱추정량

$$\mathrm{D}(i) = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{p+1} = \frac{r_i^2}{p+1} \cdot \frac{h_{ii} e_i^2}{\hat{\sigma}^2 (1-h_{ii})^2}$$

- $\mathrm{D}(i) \geq F_{0.5}(p+1, n-p-1)$ 이면 영향점으로 의심

영향점을 검출하는 방법

(3) Andrews-Pregibon의 통계량

- 행렬 $\boldsymbol{X}$와 벡터 $\boldsymbol{y}$를 같이 고려

$$\mathrm{AP}(i) = \frac{|\boldsymbol{X}^*(i)^\top \boldsymbol{X}^*(i)|}{|\boldsymbol{X}^{*\top} \boldsymbol{X}^*|}$$

- $\boldsymbol{X}^* = (\boldsymbol{X}, \boldsymbol{y})$, $\boldsymbol{X}^*(i) = (\boldsymbol{X}(i), \boldsymbol{y}(i))$
- $\boldsymbol{X}^*(i)$: $\boldsymbol{X}^*$ 행렬에서 i번째 행을 제거한 것
- 가장 작은 $\mathrm{AP}(i)$의 값을 영향점으로 간주

영향점을 검출하는 방법

(4) COVRATIO

$$\text{COVRATIO}(i) = \frac{1}{\left[1 + \frac{(r_i^*)^2 - 1}{n-p-1}\right]^{p+1} (1 - h_{ii})}$$

- COVRATIO의 값이 1에 가까우면 y_i는 별로 영향을 주지 못함
- 1에서 멀어질수록 영향을 크게 주고 있음
- $|\text{COVRATIO}(i) - 1| \geq 3(p+1)/n$ 이면 i번째 관측치를 영향을 크게 주는 측정값으로 볼 수 있음

영향점을 검출하는 방법

(5) FVARATIO

$$\begin{aligned}\mathrm{FVARATIO}(i) =& \frac{h_{ii}\hat{\sigma}^2(i)/(1-h_{ii})}{h_{ii}\hat{\sigma}^2} = \frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2(1-h_{ii})} \\ =& \frac{e_i^2}{(r_i^*)^2(1-h_{ii})^2\hat{\sigma}^2}\end{aligned}$$

▷ $\mathrm{FVARATIO}(i) \leq 1 - \dfrac{3}{n}$ 이거나 $\mathrm{FVARATIO}(i) \geq 1 + \dfrac{2p+3}{n}$ 이면 i 번째 관측치를 영향을 크게 주는 측정값으로 제안

지렛대점, 이상점과 영향점을 찾는 데 사용되는 측도

지렛대점	이상점	영향점
1. 행렬 H의 대각선원소 h_{ii} 2. Mahalanobis의 거리 $\mathrm{M}(i) = (n-1)\left(h_{ii} - \frac{1}{n}\right)$	1. 표준화잔차 $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$ 2. 스튜던트화 잔차 $r_i^* = \frac{e_i}{s(i)\sqrt{1-h_{ii}}} = r_i\left(\frac{n-p-2}{n-p-1-r_i^2}\right)^{1/2}$	1. $\mathrm{DFFITS}(i) = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} r_i^*$ 2. Cook의 통계량 $\mathrm{D}(i) = \frac{h_{ii}}{(p+1)(1-h_{ii})} \cdot r_i^2$ 3. Andrews-Pregibon의 통계량 $\mathrm{AP}(i) = 1 - h_{ii} - \frac{e_i^2}{(n-p-1)s^2}$ 4.$\mathrm{COVRATIO}(i)$ $= \frac{1}{\left[1+\frac{(r_i^*)^2-1}{n-p-1}\right]^{p+1} \cdot (1-h_{ii})}$ 5. $\mathrm{FVARATIO}(i)$ $= \frac{e_i^2}{(r_i^*)^2(1-h_{ii})^2 s^2}$

Example

실험번호	DFFITS(i)	D(i)	M(i)	AP(i)	COVRATIO(i)	FVARATIO(i)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
7	0.0772	0.0031	0.2074	0.9370	1.1702	1.1145
8	0.0563	0.0017	0.1810	0.9406	1.1742	1.1157
9	0.0854	0.0038	0.6448	0.9159	1.1997	1.1418
10	0.1728	0.0154	0.5000	0.9081	1.1521	1.1146
11	0.3320	0.0548	0.8027	0.8567	1.0878	1.0938
12	-0.0944	0.0047	0.4585	0.9234	1.1833	1.1283
13	-0.3911	0.0717	0.3039	0.8317	0.9363	0.9996
14	-0.3137	0.0416	0.1810	0.8647	0.9923	1.0256
15	0.1013	0.0054	0.1810	0.9345	1.1590	1.1085
16	0.0330	0.0006	0.3039	0.9363	1.1867	1.1253
17	0.1972	0.0179	0.0898	0.9155	1.0964	1.0755
18	-1.1558	0.6781	12.0498	0.3351	2.9587	2.9142
19	0.8537	0.2233	0.1086	0.5497	0.3964	0.6470
20	0.2638	0.0345	0.1810	0.8863	1.0426	1.0513
21	0.0330	0.0006	0.3039	0.9363	1.1867	1.1253