9강

자료진단

정보통계학과 김성수교수

학습목차

1. 회귀진단
2. 잔차 분석
3. 특이점
4. 영향력 관측값

1 회귀진단

회귀진단(Regression Diagnostics)

- 총괄분석(aggregate analysis) : 회귀식을 구하고 추정 및 검정을 하는 과정

⇒ 이 과정의 목적은 전체자료를 결합하여 적합모형(fitted model)이라고 하는 하나의 요약된 결과를 구하는 것임.

⇒ 이러한 총괄분석은 설정한 회귀모형과 가정들이 정확하다는 전제하에 이루어 진 것임.

⇒ 어떠한 문제에서도 그렇듯이 제시된 회귀모형 및 가정은 얼마든지 틀릴 수가 있음.

- 총괄분석에 이어 회귀분석의 두 번째 중요한 과정은 사용된 회귀모형 및 가정이 타당한 지, 그리고 각각의 관측값이 모형 및 가정에 어떠한 영향을 미치는 지를 진단하는 것임. 이러한 과정을 회귀진단이라고 함.

  총괄분석에 대비하여 이 과정을 개별분석(case analysis) 이라고도 함.

회귀진단(Regression Diagnostics)

- 회귀진단은 모형진단(model diagnostics)과 자료진단(data diagnostics)의 두 가지로 구분

⇒ 모형진단에서는 모형이나 가정에 어떤 문제점이 있나를 알아봄.

⇒ 자료진단에서는 자료의 조그만 변화가 모형의 추정에 어떠한 영향을 미치는 가를 알아봄.

2 잔차 분석

잔차(Residual)

- 잔차 e_i :

  반응변수의 측정값 Y_i 와 추정값 $\hat{Y}_i$ 의 차이, 즉 $e_i = Y_i - \hat{Y}_i$

- 행렬을 이용한 잔차벡터 표현

$$Y = X\beta + \epsilon, \quad E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2 I$$

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'Y \\ &= HY\end{aligned}$$

$H = X(X'X)^{-1}X'$: 햇행렬(hat matrix), $n \times n$ 행렬

- 잔차벡터

$$\begin{aligned}e &= Y - \hat{Y} \\ &= Y - HY \\ &= (I - H)Y\end{aligned}$$

잔차의 표준화

- 오차 ϵ : 관찰할 수 없는 변수로서 기대값 0 이며 서로 상관관계가 없고 등분산 σ^2이라고 가정.

- 잔차 e 는 계산되어지는 값으로서 기대값벡터와 분산-공분산 행렬은 다음과 같음.

$$E(\boldsymbol{e}) = \mathbf{0}, \quad Var(\boldsymbol{e}) = \sigma^2 (\boldsymbol{I} - \boldsymbol{H})$$

⇒ 오차와 같이 각각의 잔차는 기대값이 0 이지만 그 분산은 같지 않을 뿐 아니라 서로 상관관계가 있음.

잔차의 표준화

- $Var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$ 이므로 $Var(e_i) = \sigma^2(1-h_{ii})$ 이 됨.

⇒ $Var(e_i)$는 h_{ii}가 클 때 작아짐. 즉, X_i가 $\overline{X}$에서 멀리 떨어져있는 경우 더 작은 잔차를 가질 가능성이 높음.

⇒ 그러나 $\overline{X}$에서 멀리 떨어져있는 곳에서는 모형이 다를 가능성이 많은데 이 곳에서 잔차가 작아지는 경향이 있다는 사실은 바람직하지 못함.

⇒ 이를 개선하기 위해 e_i를 잔차의 표준오차인 $\sigma\sqrt{1-h_{ii}}$로 나누어 표준화 시킬 필요가 있음. 이것을 **잔차의 표준화**라고 함.

표준화 잔차(standardized residual)

- 표준화 잔차(standardized residual)

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

⇒ σ^2을 추정하는데 있어서 i번째 관측값을 포함하는 모든 자료를 사용.

⇒ 내면스튜던트화 잔차(internally studentized residuals) 라고도 함.

스튜던트화 잔차(studentized residual)

- 스튜던트화 잔차(studentized residuals)

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

⇒ σ^2을 추정하는데 있어서 i번째 관측값을 제외하고 구한 회귀모형에서 얻어진 σ^2의 추정량인 $MSE_{(i)}$를 사용. 아래첨자 (i)는 i번째 값이 제외되었음을 의미.

⇒ r_i 와는 달리 t_i에서의 분모와 분자는 서로 독립적임. t_i는 자유도가 $n-k-2$인 t-분포를 따르며 특이점(outlier)을 찾기 위한 검정통계량으로 쓰임.

표준화잔차와 스튜던트화 잔차의 관계

- 표준화 잔차 r_i 와 스튜던트화잔차 t_i 와의 관계

$$t_i = r_i \left(\frac{n-k-2}{n-k-1-r_i^2} \right)^{1/2}$$

3 특이점

표준화잔차와 스튜던트화 잔차의 관계

✓ 특이점(outlier)

회귀분석에서 중요한 가정 중의 하나는 자료에 포함된 모든 관측값에 대해 사용된 모형이 적절하다는 것임. 그러나, 실제 문제에서는 1-2개의 관측값이 대부분의 자료가 적합되는 모형을 따르지 않는 경우를 많이 봄. 이와 같이 나머지 관측값들과는 달리 주어진 모형을 따르지 않는 관측값을 특이점(outlier) 또는 이상점이라 부름

=〉 자료진단의 중요한 기능 중의 하나는 이러한 특이점을 찾아내는 것

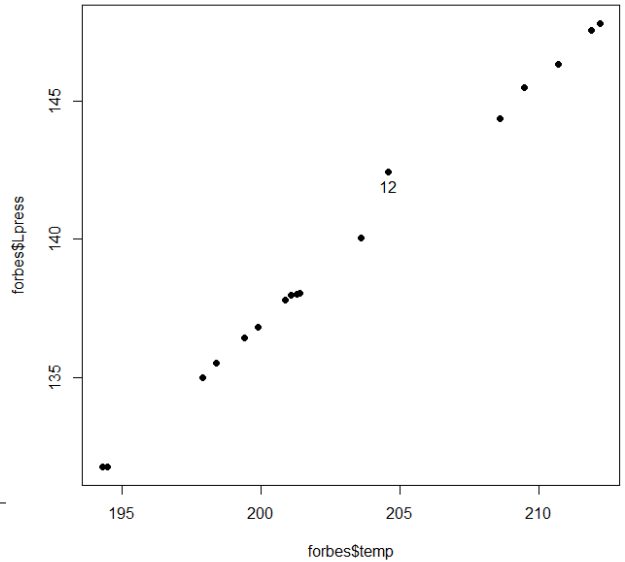
✓ 특이점의 검출 : 스튜던트화 잔차를 이용

: 특이점 검정 : Bonferroni t-검정 이용

R 활용 예 : 자료 읽기

◆ **Forbes 자료**

번호	물끓는온도 (°F)	대기압력 (in. Hg)
1	194.5	20.79
2	194.3	20.79
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		



Forbes는 물 끓는 온도와 대기압력에 로그를 취한 값 사이에는 선형관계가 존재한다고 함. 따라서 대기압력에 상용로그를 취한 값에 100을 곱한 값을 반응변수로 함.

```
> forbes = read.table("c:/data/reg/forbes.txt",
          header=T)
> forbes$Lpress = 100*log10(forbes$press)
> head(forbes, 3)
  num  temp press   Lpress
1   1 194.5 20.79 131.7854
2   2 194.3 20.79 131.7854
3   3 197.9 22.40 135.0248
> plot(forbes$temp, forbes$Lpress, pch=19)
> identify(forbes$temp, forbes$Lpress)
[1] 12
```

R 활용 예 : 모형적합

```
> forbes.lm = lm(Lpress ~ temp, data=forbes)
> summary(forbes.lm)

Call:
lm(formula = Lpress ~ temp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418    3.34136  -12.62 2.17e-09 ***
temp          0.89562    0.01646   54.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

R 활용 예 : 분산분석표 작성

```
> anova(forbes.lm)
Analysis of Variance Table

Response: Lpress
          Df Sum Sq Mean Sq F value    Pr(>F)
temp       1 425.76  425.76  2961.5 < 2.2e-16 ***
Residuals 15   2.16    0.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

R 활용 예 : 잔차분석

```
> forbes.res = ls.diag(forbes.lm)
> names(forbes.res)
 [1] "std.dev"      "hat"          "std.res"      "stud.res"
 [5] "cooks"        "dfits"        "correlation"  "std.err"
 [9] "cov.scaled"   "cov.unscaled"
> resid.result = cbind(forbes.res$std.res, forbes.res$stud.res, forbes.res$hat)
> colnames(resid.result) = c("standardized resid", "studentized resid", "Hat")
> resid.result = round(resid.result,3)
> print(resid.result)
      standardized resid studentized resid   Hat
 [1,]             -0.728             -0.716 0.193
 [2,]             -0.203             -0.197 0.200
 [3,]             -0.150             -0.145 0.107
 [4,]              0.052              0.050 0.098
 [5,]              0.091              0.088 0.083
[10,]             -0.230             -0.223 0.064
[11,]             -0.400             -0.388 0.060
[12,]              3.707             12.374 0.064
[13,]              0.004              0.004 0.140
[16,]             -0.235             -0.227 0.210
[17,]             -0.260             -0.252 0.220
```

```
> rstudent(forbes.lm) #스튜던트화 잔차
          1           2           3           4           5
-0.716454916 -0.196531386 -0.145084092  0.050361279  0.088080643
```

```
> Bonferroni 유의수준 0.01에서 기각치
> qt(0.01/(2*17), 14)
[1] -4.414447
```

```
> Bonferroni p-value for obs.12
> 2*17*(1-pt(12.374,14))
[1] 1.071262e-07
```

R 활용 예 : 특이값 검정

```
> library(car)
> outlierTest(forbes.lm)
   rstudent unadjusted p-value Bonferonni p
12 12.37386         6.3025e-09   1.0714e-07
```

특이점 식별 후 조치

특이점으로 판정되면 먼저 그 원인이 어디에 있는가를 규명

- 자료 입력시 오류가 발생하였거나
- 실험이 잘못 되었거나
- 잘못된 원료가 사용되었거나
- 자료를 측정하는 기계가 고장이 났다거나 등의 여러 가지 원인이 있을 수 있음.

이와 같이 원인이 밝혀지면 가능한 경우 다시 실험을 하여 새로운 자료로 대체시키고,
이것이 불가능한 경우 특이점을 제거시키고 분석.

4 영향력 관측값

영향력이 큰 관측값(influential observation)

• 영향력이 큰 관측값(influential observation) : 자료에서 관측값을 제거하고 얻은 회귀분석 결과가 이 데이터를 포함시키고 얻은 결과와 판이하게 다를 때 이를 영향력있는 관측값이라고 함.

• $b_{(i)} = (b_{0(i)}, b_{1(i)}, \cdots, b_{k(i)})'$: i번째 관측값이 제거되었을 때 추정량

$X_{(i)}$: X 에서 i번째 행이 제거된 $(n-1)\times k$ 행렬

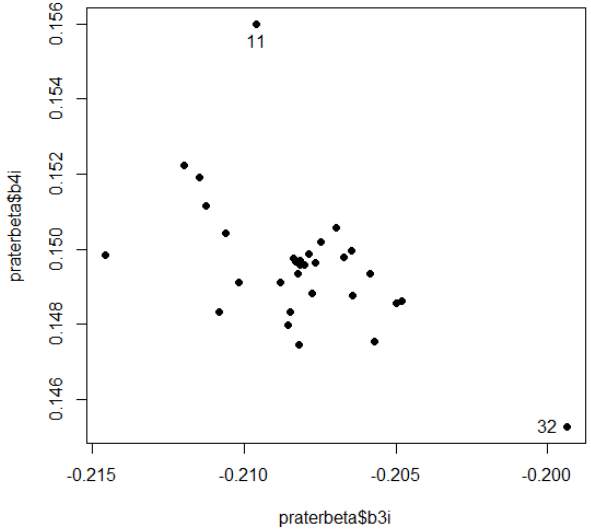
$Y_{(i)}$: Y 에서 Y_i 가 제거된 $(n-1)\times 1$ 벡터

$\Rightarrow b_{(i)} = (X_{(i)}^{'} X_{(i)})^{-1} X_{(i)}^{'} Y_{(i)}$

- 만일, 관측값 i가 영향력있는 관측값이라면 $b_{(i)}$의 값은 다른 관측값 j를 제거하고 구한 $b_{(j)}$과 비교하여 많이 다를 것임.

- 설명변수가 두 개인 경우 각 관측값이 제거되어 구해질 $b_{1(i)}$와 $b_{2(i)}$에 대한 산점도에서 다른 점들과 유난히 구별되는 점이 영향력있는 관측값이 될 것임.

예) Prater 자료 (교재 152 페이지)



Cook의 D_i 통계량

Cook의 통계량 D_i

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \beta)}{(k+1)MSE}$$

여기서 $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$로 정의하면 D_i는 다음과 같이 표현됨.

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T(\hat{Y}_{(i)} - \hat{Y})}{(k+1)MSE}$$

⇨ 따라서 D_i는 $\hat{Y}$와 $\hat{Y}_{(i)}$사이의 거리 개념으로 이해

⇨ D_i가 큰 관측값은 회귀계수의 추정값 $\hat{\beta}$와 예측값 $\hat{Y}$에 상당한 영향을 미침. 그리고 그 관측값의 제거는 결론에 큰 변화를 초래할 수도 있음.

⇨ 가장 큰 D_i를 가진 관측값이나 관측값의 수가 매우 많은 자료에서 몇 개의 큰 D_i를 가진 관측값에 대해서는 주의를 기울여야 함.

Cook의 D_i 통계량

Cook의 D_i 통계량의 간편식

$$D_i = \frac{1}{(k+1)} r_i^2 \left(\frac{h_{ii}}{1-h_{ii}} \right)$$

즉, D_i는 i 번째 표준화 잔차 r_i의 제곱과 h_{ii}의 단조함수와의 곱.

⇨ D_i의 크기는 i번째 관측값에서 모형적합의 부족을 반영하는 r_i와 x_i의 $\bar{x}$로 부터 떨어져 있는 위치를 반영하는 h_{ii}에 의해 D_i의 값이 결정됨.

⇨ 즉, r_i가 크거나 h_{ii}의 값이 크면 D_i의 값은 커짐.

R 활용 예 : 자료 읽기

◆ 토양침식자료

번호	SL	SG	LOBS	PGC
1	27.1	0.43	1.95	0.34
2	35.6	0.47	5.13	0.32
3	31.4	0.44	3.98	0.29
4	37.8	0.48	6.25	0.30
5	40.2	0.48	7.12	0.25
6	39.8	0.49	6.50	0.26
7	55.5	0.53	10.67	0.10
8	43.6	0.50	7.08	0.16
9	52.1	0.55	9.88	0.19
10	43.8	0.51	8.72	0.18
11	35.7	0.48	4.96	0.28

```
> soil = read.table("c:/data/reg/soil.txt", header=T)
> head(soil,3)
  obs   SL   SG LOBS  PGC
1   1 27.1 0.43 1.95 0.34
2   2 35.6 0.47 5.13 0.32
3   3 31.4 0.44 3.98 0.29
```

R 활용 예 : 모형적합

```
> soil.lm = lm(SL ~ SG+LOBS+PGC, data=soil)
> summary(soil.lm)
Call:
lm(formula = SL ~ SG + LOBS + PGC, data = soil)

Residuals:
     Min       1Q   Median       3Q      Max
-3.04950 -0.27965 -0.01715  0.68759  2.15143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.8793    18.1342  -0.104   0.9204
SG           77.3258    44.5055   1.737   0.1259
LOBS          1.5591     0.7345   2.123   0.0714 .
PGC         -23.9038    13.4294  -1.780   0.1183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.516 on 7 degrees of freedom
Multiple R-squared:  0.9769,    Adjusted R-squared:  0.967
F-statistic: 98.66 on 3 and 7 DF,  p-value: 4.325e-06
```

변수	추정값	표준오차	t_0	유의확률
Intercept	-1.8793	18.1342	-0.104	0.9204
SG	77.3258	44.5055	1.737	0.1259
LOBS	1.5591	0.7345	2.123	0.0714
PGC	-23.9038	13.4294	-1.780	0.1183

R 활용 예 : 분산분석표 작성

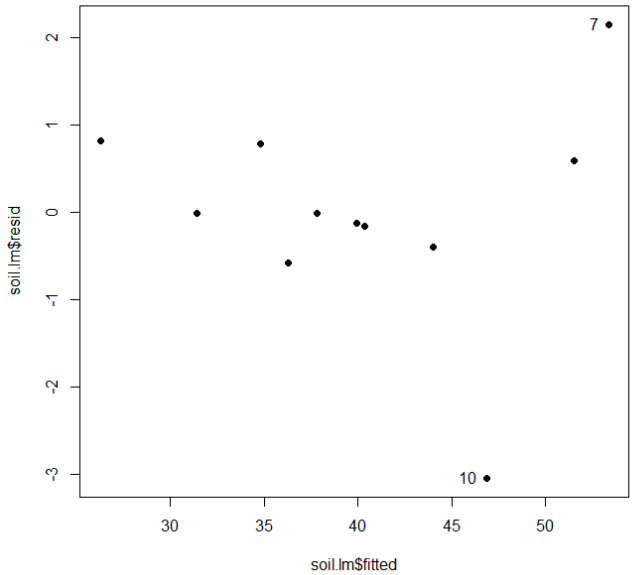
```
> anova(soil.lm)
Analysis of Variance Table

Response: SL
          Df Sum Sq Mean Sq  F value    Pr(>F)
SG         1 640.42  640.42 278.5454 6.778e-07 ***
LOBS       1  32.78   32.78  14.2581  0.006928 **
PGC        1   7.28    7.28   3.1683  0.118301
Residuals  7  16.09    2.30
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

요인	자유도	제곱합	평균제곱	F_0	유의확률
회귀	3	680.48	226.83	98.66	0.0000
잔차	7	16.09	2.30		
계	10	696.57			

R 활용 예 : (추정값, 잔차) 산점도

```
> plot(soil.lm$fitted, soil.lm$resid, pch=19)
> identify(soil.lm$fitted, soil.lm$resid)
[1]  7 10
```



예측값이 가장 큰 7번이 영향력 있는 관측값으로, 7번보다는 조금 더 중간에 가깝고 잔차가 큰 10번이 특이점이 될 가능성이 높은 것으로 보임.

R 활용 예 : 잔차분석 및 cook 통계량

```
> soil.diag = ls.diag(soil.lm)
> names(soil.diag)
 [1] "std.dev"      "hat"          "std.res"      "stud.res"
 [5] "cooks"        "dfits"        "correlation"  "std.err"
 [9] "cov.scaled"   "cov.unscaled"
> diag.st = cbind(soil.diag$hat, soil.diag$std.res, soil.diag$stud.res, soil.diag$cooks)
> colnames(diag.st) = c("Hii", "ri" , "ti", "Di")
> round(diag.st, 3)
       Hii     ri     ti    Di
 [1,] 0.464  0.736  0.709 0.117
 [2,] 0.248  0.599  0.569 0.029
 [3,] 0.363 -0.014 -0.013 0.000
 [4,] 0.299 -0.008 -0.008 0.000
 [5,] 0.332 -0.131 -0.121 0.002
 [6,] 0.118 -0.091 -0.084 0.000
 [7,] 0.533  2.075  3.098 1.227
 [8,] 0.530 -0.382 -0.358 0.041
 [9,] 0.629  0.636  0.607 0.171
[10,] 0.188 -2.232 -3.851 0.289
[11,] 0.298 -0.454 -0.427 0.022
```

7번 : 영향력 있는 관측값

```
> Di = cooks.distance(soil.lm)
> round(Di, 3)
    1     2     3     4     5     6     7     8     9    10    11
0.117 0.029 0.000 0.000 0.002 0.000 1.227 0.041 0.171 0.289 0.022
> library(car)
> outlierTest(soil.lm)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
10 -3.850967          0.0084505      0.092955
```

다음시간 안내

10강. 모형진단