



변수선택

정보통계학과 김성수교수

✓ 학습목차

1

변수선택과 다중공선성

2

변수선택 기준

3

변수선택 방법

1

변수선택과 다중공선성

변수선택과 다중공선성

✓ 변수선택

- 일반적으로 많은 사람들은 복잡한 모형보다는 간편한 모형을 선호.
반응변수에 영향을 미치리라고 예상되는 많은 설명변수 중에서 모형에 포함시킬 변수를 결정. 이를 **변수선택의 문제**(variable selection problem)라 함.

✓ 다중공선성

- 포함되는 설명변수들 사이에 연관성이 있는 경우에는 적합된 모형의 안정성과 신뢰성을 떨어뜨림. 이를 **공선성**, 혹은 **다중공선성**이라고 함.

다중공선성

✓ 다중공선성

- 다중공선성 : 두 설명변수 X_1 과 X_2 가 임의의 상수 c_0, c_1, c_2 에 대하여

$$c_1 X_1 + c_2 X_2 = c_0$$

인 경우, 두 변수 사이에 완벽한 공선성(exact collinearity)이 있다고 함.

(예) $X_1 + X_2 = 100$ 인 경우, X_2 는 X_1 으로 완전히 결정됨.

- 두 설명변수의 표본상관계수의 제곱 r_{12}^2 이 1 인 경우 두 변수 사이에 완벽한 공선성이 존재함.

다중공선성

- 설명변수의 수가 2가 넘는 경우, 설명변수 X_1, \dots, X_k 들이 임의의 상수 c_0, c_1, \dots, c_k 에 대하여 다음과 같은 관계가 성립하거나 또는 근사적으로 성립할 때

$$c_1X_1 + c_2X_2 + \dots + c_kX_k = c_0$$

이 설명변수들 사이에 다중공선성(multicollinearity)이 존재

⇒ 설명변수 X_h 와 나머지 설명변수간의 결정계수(다중상관계수의 제곱)

R_h^2 이 다중공선성의 정도를 나타낸다고 할 수 있음.

다중공선성

✓일반적으로 회귀모형을 분석하는 과정에서 다음 중 하나 이상의 현상이 발생하면 설명변수들 사이의 다중공선성에 대한 의심

- (1) 설명변수들의 표본상관행렬에서 상관계수가 크게 (+1 또는 -1에 가까운 경우)나타날 때
- (2) 어떤 설명변수를 모형에 추가하거나 제거시 추정된 회귀계수의 크기나 부호에 큰 변화를 줄 때
- (3) 새로운 자료를 추가하거나 기존의 자료를 제거시 추정된 회귀계수의 크기나 부호에 큰 변화를 줄 때
- (4) 중요하다고 생각되어지는 설명변수에 대한 검정결과가 유의하지 않게 나타나거나, 이 설명변수에 대한 회귀계수의 신뢰구간이 상당히 넓을 때
- (5) 추정된 회귀계수의 부호가 과거의 경험이나 이론적인 면에서 기대되는 부호와 상반될 때

다중공선성

✓ 분산팽창인자 (VIF; variance inflation factor)

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

, R_j^2 : X_j 를 반응변수로 보고 나머지 설명변수들에 대한 결정계수

⇒ 일반적으로 k 개의 VIF_j 중 가장 큰 값이 5~10을 넘으면
다중공선성이 있다고 판정.

R 활용

<해군병원의 인력 자료>

X_1	X_2	X_3	X_4	X_5	Y
15.57	2463	472.92	18.0	4.45	566.52
44.02	2048	1339.75	9.5	6.92	696.82
20.42	3940	620.25	12.8	4.28	1033.15
18.74	6505	568.33	36.7	3.90	1603.62
49.20	5723	1497.60	35.7	5.50	1611.37
44.92	11520	1365.83	24.0	4.60	1613.27
55.48	5779	1687.00	43.3	5.62	1854.17
59.28	5969	1639.92	46.7	5.15	2160.55
94.39	8461	2872.33	78.7	6.18	2305.58
128.02	20106	3655.08	180.5	6.15	3503.93
96.00	13313	2912.00	60.9	5.88	3571.89
131.42	10771	3921.00	103.7	4.88	3741.40
127.21	15543	3865.67	126.8	5.50	4026.52
252.90	36194	7684.10	157.7	7.00	10343.81
409.20	34703	12446.33	169.4	10.78	11732.17
463.70	39204	14098.40	331.4	7.05	15414.94
510.22	86533	15524.00	371.6	6.35	18854.45

hospital - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

X1	X2	X3	X4	X5	Y
15.57	2463	472.92	18.0	4.45	566.52
44.02	2048	1339.75	9.5	6.92	696.82
20.42	3940	620.25	12.8	4.28	1033.15
18.74	6505	568.33	36.7	3.90	1603.62
49.20	5723	1497.60	35.7	5.50	1611.37
44.92	11520	1365.83	24.0	4.60	1613.27
55.48	5779	1687.00	43.3	5.62	1854.17
59.28	5969	1639.92	46.7	5.15	2160.55
94.39	8461	2872.33	78.7	6.18	2305.58
128.02	20106	3655.08	180.5	6.15	3503.93
96.00	13313	2912.00	60.9	5.88	3571.89
131.42	10771	3921.00	103.7	4.88	3741.40
127.21	15543	3865.67	126.8	5.50	4026.52
252.90	36194	7684.10	157.7	7.00	10343.81
409.20	34703	12446.33	169.4	10.78	11732.17
463.70	39204	14098.40	331.4	7.05	15414.94
510.22	86533	15524.00	371.6	6.35	18854.45

Y : 월간 의사 연 근무시간

X_1 : 일 평균 환자수

X_2 : 월간 X-ray 촬영 횟수

X_3 : 월간 이용 병석수

X_4 : 해당지역의 병원 이용가능 인구 / 1000

X_5 : 평균입원일

```
> hospital <- read.table("c:/data/reg/hospital.txt", header=T)
```

```
> head(hospital, 3)
```

	X1	X2	X3	X4	X5	Y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15

회귀적합

```
> hospital.lm <- lm(Y~ ., data=hospital)
```

```
> summary(hospital.lm)
```

```
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1962.94816	1071.36170	1.832	0.0941 .
X1	-15.85167	97.65299	-0.162	0.8740
X2	0.05593	0.02126	2.631	0.0234 *
X3	1.58962	3.09208	0.514	0.6174
X4	-4.21867	7.17656	-0.588	0.5685
X5	-394.31412	209.63954	-1.881	0.0867 .

Residual standard error: 642.1 on 11 degrees of freedom

Multiple R-squared: 0.9908, Adjusted R-squared: 0.9867

F-statistic: 237.8 on 5 and 11 DF, p-value: 8.068e-11

$$\hat{Y} = 1962.948 - 15.852X_1 + 0.056X_2 + 1.590X_3 - 4.219X_4 - 394.314X_5$$

반응변수인 월간 의사 연 근무시간(Y)은 일 평균 환자(X_1), 병원가능인구(X_4), 그리고 평균입원일(X_5)에 비례할 것으로 예상되나 이들 세 설명변수에 대한 회귀계수, $\beta_1, \beta_4, \beta_5$ 의 추정값의 부호가 모두 음으로 나타남.

⇒ 설명변수들 간에 다중공선성의 존재가 예상되며, 따라서 분산팽창인자 등에 의한 진단이 요구.

분산팽창인자 계산

```
> library(fmsb)
> VIF(lm(X1~X2+X3+X4+X5, data=hospital))
[1] 9597.571
> VIF(lm(X2~X1+X3+X4+X5, data=hospital))
[1] 7.940593
> VIF(lm(X3~X1+X2+X4+X5, data=hospital))
[1] 8933.087
> VIF(lm(X4~X1+X2+X3+X5, data=hospital))
[1] 23.29386
> VIF(lm(X5~X1+X2+X3+X4, data=hospital))
[1] 4.279835
```

```
> cor(hospital[, -6])
      X1      X2      X3      X4      X5
X1 1.0000000 0.9073795 0.9999040 0.9356913 0.6711974
X2 0.9073795 1.0000000 0.9071493 0.9104688 0.4466496
X3 0.9999040 0.9071493 1.0000000 0.9331680 0.6711095
X4 0.9356913 0.9104688 0.9331680 1.0000000 0.4628609
X5 0.6711974 0.4466496 0.6711095 0.4628609 1.0000000
```



〈분산팽창인자〉

변수	분산팽창인자
X1	9597.571
X2	7.9406
X3	8933.087
X4	23.29386
X5	4.2798

: 다중공선성 문제가 존재

설명변수들 사이, 특히 X_1, X_2, X_3, X_4 간에

강한 선형종속관계가 있음.

설명변수 X1을 제거하는 경우

```
> summary(lm(Y~X2+X3+X4+X5, data=hospital))
```

```
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2032.18806	942.07483	2.157	0.0520 .
X2	0.05608	0.02036	2.755	0.0175 *
X3	1.08837	0.15340	7.095	1.26e-05 ***
X4	-5.00407	5.08071	-0.985	0.3441
X5	-410.08296	178.07810	-2.303	0.0400 *

—
Residual standard error: 615.5 on 12 degrees of freedom
Multiple R-squared: 0.9908, Adjusted R-squared: 0.9877
F-statistic: 323.5 on 4 and 12 DF, p-value: 4.181e-12

```
> VIF(lm(X2~X3+X4+X5, data=hospital))
```

```
[1] 7.925796
```

```
> VIF(lm(X3~X2+X4+X5, data=hospital))
```

```
[1] 23.92684
```

```
> VIF(lm(X4~X2+X3+X5, data=hospital))
```

```
[1] 12.70597
```

```
> VIF(lm(X5~X2+X3+X4, data=hospital))
```

```
[1] 3.360866
```

결정계수 $R^2 = 0.99$ 도 거의 변하지 않았으며

각각의 추정된 회귀계수의 표준오차는 모두 줄어들었음을 알 수 있음.

또한 X_1 을 제거한 경우의 분산팽창인자도

X_1 을 포함하는 경우에 비하여 모두 작은 값을 가지는 것을 알 수 있음.

2 변수 선택의 기준

모형선택의 기준 : 결정계수

R_p^2 는 k 개의 설명변수 중에서 p 개의 설명변수로 구성되는 모형에서 다음과 같이 정의.

$$R_p^2 = 1 - \frac{SSE_p}{SST}$$

, SSE_p : p 개의 설명변수를 포함한 모형에서의 잔차제곱합

⇒ R_p^2 는 설명변수가 추가되어 p 가 증가함에 따라 증가하여 모든 설명변수가 다 포함이 된 모형($p=k$)일 때 최대가 됨. 이러한 이유로 최대의 결정계수의 값을 가지는 모형을 선택하는 것은 의미가 없고 변수를 하나씩 증가시키면서 R_p^2 의 증가가 둔화되는 지점에서 p 개의 설명변수를 선택.

⇒ 그러나 이 지점을 선택하는 과정은 다분히 분석자의 주관에 의존.

모형선택의 기준 : 수정결정계수

- 결정계수 R_p^2 가 가지고 있는 결점을 보완하기 위하여 제시된 방법이

수정결정계수(adjusted coefficient of determination), \bar{R}_p^2

$$\bar{R}_p^2 = 1 - \frac{SSE_p / (n - p - 1)}{SST / (n - 1)} = 1 - \left(\frac{n - 1}{n - k - 1} \right) (1 - R_p^2)$$

⇒ 제곱합 SS 를 그대로 쓰기보다는 해당되는 자유도로 나누는 조정 과정을 거침

⇒ \bar{R}_p^2 는 설명변수의 수가 증가하여도 항상 증가하는 값은 아님. 따라서 R_p^2 와는

달리 \bar{R}_p^2 를 기준으로 하는 경우에는 \bar{R}_p^2 을 최대로 하는 p 개의 설명변수를 선택.

- \bar{R}_p^2 를 다시 쓰면

$$\bar{R}_p^2 = 1 - \frac{MSE_p}{Syy / n - 1}$$

⇒ 모형선택의 기준으로서의 최소 MSE_p 와 최대 \bar{R}_p^2 기준은 동일함.

모형선택의 기준 : Mallows Cp, AIC

Mallows의 C_p 통계량

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p - 2)$$

⇒ C_p 를 기준으로 k 개의 변수 중에서 p 개의 변수를 선택할 때
 $C_p \approx p+1$ 에 가까운 모형 중에서 C_p 값이 최소가 되는 모형을
선택

AIC(Akaike Information Criteria)

$$AIC = n \log(SSE_p/n) + 2(p+1)$$

⇒ 작은 값을 갖는 모형을 선택

3 변수선택 방법

변수선택의 방법

- (1) 모든 가능한 회귀 (all possible regression)
- (2) 앞으로부터 선택법 (forward selection)
- (3) 뒤로부터 제거법 (backward elimination)
- (4) 단계별 회귀방법 (stepwise regression)

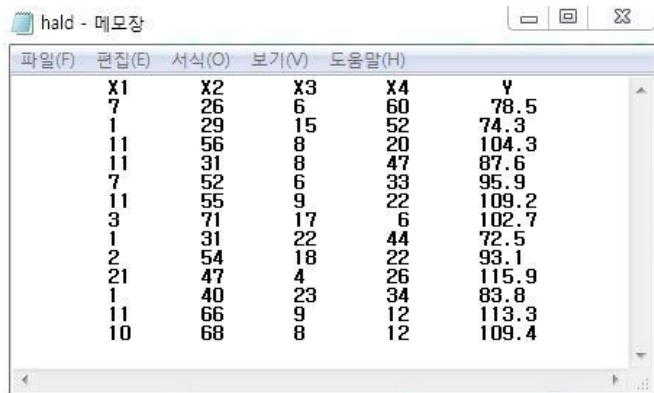
모든 가능한 회귀

➤ 이 방법은 모든 가능한 변수들의 조합을 회귀분석하여 보는 것임.

만약 k 개의 설명변수가 있는 경우 $2^k - 1$ 개의 회귀모형을 적합

⇒ k 가 작은 경우는 문제가 없으나 k 가 커지면 방대한 계산이 요구되고
효율적인 알고리즘을 이용한 컴퓨터의 도움이 없이는 불가능

R 활용 : Hald 자료 읽기



X1	X2	X3	X4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

```
> hald = read.table("c:/data/reg/hald.txt", header=T)
```

```
> head(hald, 3)
```

	X1	X2	X3	X4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3

R 활용 : 모든 가능한 회귀

```
> install.packages("leaps")
> library(leaps)
> all.lm = regsubsets(Y ~ ., data=hald)
> (rs=summary(all.lm))
```

Subset selection object
 Call: regsubsets.formula(Y ~ ., data = hald)
 4 Variables (and intercept)

	Forced in	Forced out
X1	FALSE	FALSE
X2	FALSE	FALSE
X3	FALSE	FALSE
X4	FALSE	FALSE

1 subsets of each size up to 4
 Selection Algorithm: exhaustive

		X1	X2	X3	X4
1	(1)	" "	" "	" "	"*"
2	(1)	"*"	"*"	" "	" "
3	(1)	"*"	"*"	" "	"*"
4	(1)	"*"	"*"	"*"	"*"

```
> names(rs)
[1] "which" "rsq" "rss" "adjr2" "cp" "bic"
"outmat" "obj"
> rs$rsq
[1] 0.6745420 0.9786784 0.9823355 0.9823756
> rs$adjr2
[1] 0.6449549 0.9744140 0.9764473 0.9735634
> rs$cp
[1] 138.730833 2.678242 3.018233 5.000000
```

p	선택된 변수	R_p^2	\bar{R}_p^2	C_p
1	X_4	0.6745	0.6450	138.731
2	$X_1 X_2$	0.9787	0.9744	2.678
3	$X_1 X_2 X_4$	0.9823	0.9764	3.018
4	$X_1 X_2 X_3 X_4$	0.9824	0.9736	5.00

앞으로부터의 선택(forward selection method)

- ✓ 반응변수에 영향을 줄 것으로 생각되는 k 개의 설명변수들 중에서 가장 크게 영향을 줄 것으로 판단되는 변수부터 하나씩 선택하여 더 이상 중요한 변수가 없다고 판단될 때 변수의 선택을 중단하는 방법.

R 활용 : 앞으로부터의 선택

```
> start.lm = lm(Y~1, data=hald)
> full.lm = lm(Y~ ., data=hald)
> step(start.lm, scope=list(lower=start.lm,
  upper=full.lm), direction="forward")
```

Start: AIC=71.44

Y ~ 1

	Df	Sum of Sq	RSS	AIC
+ X4	1	1831.90	883.87	58.852
+ X2	1	1809.43	906.34	59.178
+ X1	1	1450.08	1265.69	63.519
+ X3	1	776.36	1939.40	69.067
<none>			2715.76	71.444

Step: AIC=58.85

Y ~ X4

	Df	Sum of Sq	RSS	AIC
+ X1	1	809.10	74.76	28.742
+ X3	1	708.13	175.74	39.853
<none>			883.87	58.852
+ X2	1	14.99	868.88	60.629

Step: AIC=28.74

Y ~ X4 + X1

	Df	Sum of Sq	RSS	AIC
+ X2	1	26.789	47.973	24.974
+ X3	1	23.926	50.836	25.728
<none>			74.762	28.742

Step: AIC=24.97

Y ~ X4 + X1 + X2

	Df	Sum of Sq	RSS	AIC
<none>			47.973	24.974
+ X3	1	0.10909	47.864	26.944

Call:

lm(formula = Y ~ X4 + X1 + X2, data = hald)

Coefficients:

(Intercept)	X4	X1	X2
71.6483	-0.2365	1.4519	0.4161

뒤로부터 제거 (backward elimination method)

✓ 앞으로부터 선택법은 설명변수를 하나씩 선택하는 방법이나
뒤로부터 제거법은 이와 반대로 반응변수에 영향을 주리라고 생각되는
 k 개의 설명변수들 중에서 가장 작게 영향을 주리라고 여겨지는
변수부터 하나씩 제거하여 나가면서 더 이상 제거할 변수가 없다고
판단될 때 변수의 제거를 중단하는 방법

R 활용 : 뒤로부터 제거

```
> full.lm = lm(Y~ ., data=hald)
> step(full.lm, data=hald, direction="backward")
```

Start: AIC=26.94
Y ~ X1 + X2 + X3 + X4

	Df	Sum of Sq	RSS	AIC
- X3	1	0.1091	47.973	24.974
- X4	1	0.2470	48.111	25.011
- X2	1	2.9725	50.836	25.728
<none>			47.864	26.944
- X1	1	25.9509	73.815	30.576

Step: AIC=24.97
Y ~ X1 + X2 + X4

	Df	Sum of Sq	RSS	AIC
<none>			47.97	24.974
- X4	1	9.93	57.90	25.420
- X2	1	26.79	74.76	28.742
- X1	1	820.91	868.88	60.629

Call:
lm(formula = Y ~ X1 + X2 + X4, data = hald)

Coefficients:
(Intercept) X1 X2 X4
71.6483 1.4519 0.4161 -0.2365

단계별 회귀(stepwise selection)

- ✓ 앞으로부터 선택법에 뒤로부터 제거법을 가미한 방법.
- 이 방법은 중요한 변수를 하나씩 선택하여 나가면서 이미 선택된 변수가 의미가 있는 지를 체크
- 즉, 새로운 변수가 추가될 때마다, 기존의 변수가 제거될 필요가 있는지를 매 단계별로 검토하여 선택하는 방법임

R 활용 : 단계별 선택

```
> start.lm = lm(Y~1, data=hald)
> full.lm = lm(Y~ ., data=hald)
> step(start.lm, scope=list(upper=full.lm), data=hald,
      direction="both")
```

Start: AIC=71.44

Y ~ 1

	Df	Sum of Sq	RSS	AIC
+ X4	1	1831.90	883.87	58.852
+ X2	1	1809.43	906.34	59.178
+ X1	1	1450.08	1265.69	63.519
+ X3	1	776.36	1939.40	69.067
<none>			2715.76	71.444

Step: AIC=58.85

Y ~ X4

	Df	Sum of Sq	RSS	AIC
+ X1	1	809.10	74.76	28.742
+ X3	1	708.13	175.74	39.853
<none>			883.87	58.852
+ X2	1	14.99	868.88	60.629
- X4	1	1831.90	2715.76	71.444

Step: AIC=28.74

Y ~ X4 + X1

	Df	Sum of Sq	RSS	AIC
+ X2	1	26.79	47.97	24.974
+ X3	1	23.93	50.84	25.728
<none>			74.76	28.742
- X1	1	809.10	883.87	58.852
- X4	1	1190.92	1265.69	63.519

Step: AIC=24.97

Y ~ X4 + X1 + X2

	Df	Sum of Sq	RSS	AIC
<none>			47.97	24.974
- X4	1	9.93	57.90	25.420
+ X3	1	0.11	47.86	26.944
- X2	1	26.79	74.76	28.742
- X1	1	820.91	868.88	60.629

Call:

lm(formula = Y ~ X4 + X1 + X2, data = hald)

Coefficients:

(Intercept)	X4	X1	X2
71.6483	-0.2365	1.4519	0.4161



다음시간 안내

8강. 모형개발