다중공선성 multicollinearity

개념

- 다중 회귀모형($y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i$)는 설명변수와 종속변수간의 관계에 대한 유의성 검정(F-검정)과 각 설명변수의 유의성(t-검정)에 가장 큰 관심
- 의구심- (1)서로 다른 설명변수의 상대적 중요 정도는 무엇인가?(표준화 회귀계수) (2)종속변수에 대한 설명변수의 설명력의 크기는 얼마인가? (OLS 추정 회귀계수)
- 만약 설명변수가 서로 독립(상관계수 0)이라면 설명변수의 회귀계수(β_k , 상대적 효과, 설명력, marginal effect)에 의해 위의 질문에 답이 가능하다.
- 현실에서 설명변수가 완벽하게 독립인 경우는 없다. 설명변수들 간 상관계수가 낮으면(유의하지 않으면) 여전히 회귀계수 해석이 적절하다.
- 상관관계가 높으면(유의하면) OLS 회귀계수 추정과 검정(추정 분산이 변함)이 쓸모가 없게 되는데 이를 다중공선성(Multicollinearity) 문제

문제

1) OLS 추정치 $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$ 단문제 발생 - 두 설명변수의 상관관계가 높으면 (데이터 행렬의 한 열(X_i 변수)이 다른 열(X_i 변수와 상관계수가 큰 변수)로 표현되므로 $\det(X'X)$ 가 거의 0에 가깝다. (상관계수가 ± 1 이면 $\det(X'X) = 0$ $(X'X)$ 〈=〉 역행렬이 불안정 (상관계수 ± 1 이면) 구할 수 없음

2) OLS 추정치의 추정오차 $s^2_{\hat{\beta}} = MSE(X'X)^{-1}$ 이므로 추정치가 불안정해지므로 추정치를 신뢰할 수 없음

진단

① 산점도나 상관계수 이용

산점도 행렬이나 상관계수를 계산하여 상관 관계가 높은 설명변수들을 판단하고 다중공선성 문제가 일어날 것이라는 예상을 한다. 회귀분석의 시작은 산점도 행렬과 상관계수 행렬이다. 상관계수 행렬은 종속변수에 영향을 주는 설명변수를 사전에 판단하고 설명변수 간 높은 상관관계로 인해 발생하는 다중공선성 사전 진단.

② 분산팽창지수(Variation Index Factor)

$$VIF_k = \frac{1}{(1 - R^2_k)}$$

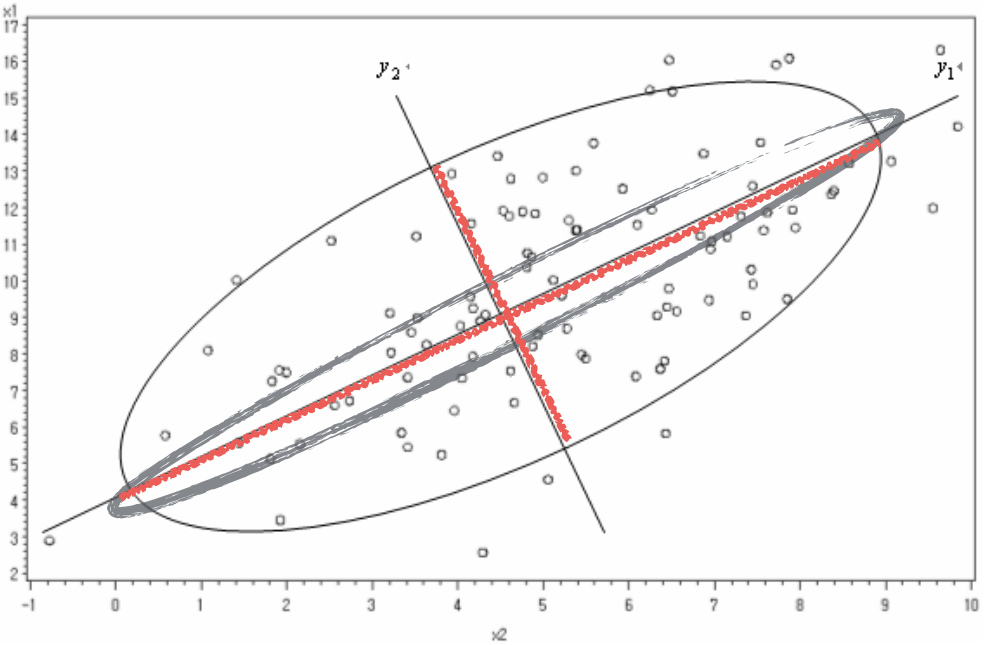
R^2_k 는 종속변수 X_k , 나머지 설명변수를 설명변수로 하여 추정한 회귀모형의 결정계수이다. 그러므로 R^2_k 가 1에 가깝다는 것은 X_k 가 다른 설명변수로 표현될 수 있다는 것이다. 종속변

수 분산팽창지수(VIF) VIF_k 가 크다는 것은 X_k 가 다른 설명변수들에 의해 선형 함수(모형)로 표현될 수 있다는 것이고 다중공선성 문제가 발생한다고 한다. 일반적으로 3이상인 설명변수가 다중공선성 문제를 발생시킨다고 판단한다.

③ 상태지수(condition index) $Condition_k = \sqrt{\frac{\lambda_{\max}}{\lambda_k}}$

$(X'X)$ 의 대각행렬이 1이 되게 변환한 후(상관변환) 고유치(eigen value)를 구하고 가장 큰 고유치 값으로 나눈 후 제곱근을 구한 값을 상태지수라 한다. 고유치는 원변수(설명변수)의 선형결합에 의해 만들어진 주성분 변수의 원변수 변동에 대한 설명력이다. 그러므로 고유치가 크다는 것(상태지수 값이 큰 값) 주성분의 원 변수 변동에 설명력이 크다는 것을 의미하므로 주성분에 의해 원변수의 차수를 줄일 수 있음을 의미한다. 원변수가 상관 관계가 높음을 의미한다. 상태지수가 10이면 원변수(설명변수)들간 약한 상관 관계가 존재하고 100 이상인 값이 있으면 상관 관계가 매우 유의한 설명변수가 존재한다. 즉 다중공선성 문제가 발생한다고 할 수 있다.

(고유치 의미) 상관계수 크기는 타원형의 길이와 폭의 넓이다. 길고 폭이 좁으면 (거의 선에 가까운 형태이면, 안의 굵은 타원) 상관관계 높음, 두 축 (y1, y2)이 길이(붉은 선 길이)는 고유치의 값이다. 가장 큰 고유치 값이 다른 고유치 값에 비해 크다는 것은 변수 간의 상관관계가 높다는 것을 의미 〈=〉 다중공선성 문제 발생



해결

① 변수 제거

다중공선성 문제를 일으키는 변수를 제외한다. 일반적으로 다중공선성 문제를 일으키는 변수 중 종속변수와의 상관 관계가 높은 것을 남겨둔다. 상관 관계의 차이가 거의 없다면 해석이 용이한 설명변수를 남겨 둔다. 모형에 고려된 설명변수의 수가 적으면 제거하는 방법보다는 다른 방법을 사용하는 것을 권한다.

*) 종속변수와 설명변수의 부호가 바뀌지 않으면 그냥 가는 경우도 많음 – 손실이 없음

② 주성분 분석 이용하기

주성분 분석(PCA, Principal Component Analysis)은 다음 원칙에 의해 원 변수의 선형 결합인 주성분(principal components)을 얻는다.

○주성분 변수 간에는 서로 상관 관계가 전혀 없다. (독립이다)

○첫번째 주성분은 변수들의 변동을(분산, 이를 변수가 가진 정보로 표현) 가장 많이 설명하고 계속 구해지는 2, 3, …번째 주성분은 자료의 나머지 정보들을 설명하고 크기는 점점 줄어든다.

주성분 변수는 원 변수(회귀분석에서는 설명변수 $X_1, X_2, \ldots, X_p$)의 선형 결합이며 서로는 독립이다. 주성분 변수는 서로 독립이므로 주성분 변수를 설명변수로 사용한다면 다중공선성 문제가 발생하지 않을 것이다.

③ 능형 회귀분석 Ridge Regression

다중공선성은 회귀계수의 분산을 증가시키므로 불편성(OLS는 불편 추정량이다)을 포기하는 대신 MSE(Mean Square of Error)를 최소화 하는 편기(biased) 추정량을 구하는 계수 추정 방법을 사용함으로써 다중공성선 문제를 해결하는데 이를 능형 회귀분석((Ridge Regression)이라 한다.

$$MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = V(\hat{\beta}) + (E(\hat{\beta}) - \beta))^2 = V(\hat{\beta}) + Bias^2$$

OLS 추정치의 편기(Bias)=0이므로 OLS 추정치 분산은 MSE 이다. 다중 회귀모형의 회귀계수에 대한 추정치로 다음을 생각해 보자. $(X'X + cI)\hat{\beta} = X'y$: $c = 0$ 인 경우 OLS 추정치이고 불편 추정량이다. $c \neq 0$ 이면 $\hat{\beta}$ 는 불편 추정량이고 $MSE(\hat{\beta})$ 최소화 하는 c 구하면 능형 추정량 $\hat{\beta}_R = (X'X + cI)^{-1}X'y$ 을 얻는다. c 을 어떻게 구하겠는가? Ridge trace(c 에 대한 추정 회귀계수 $\hat{\beta}_1^R, \hat{\beta}_2^R, \cdots, \hat{\beta}_p^R$ 의 산점도)와 VIF_k 이용한다. 각 값들이 안정화 되는 가장 작은 c 값을 선택하면 된다.

SAS 활용

```
proc reg data=ds.smsa;
     model mortality=income;
run;
proc reg data=ds.smsa;
     model mortality=income education; /*상관계수 높은 edu. 삽입*/
run;
proc reg data=ds.smsa;
     model mortality=income S02Pot; /*상관계수 낮은 S02 삽입*/
run;
```

설명변수 income과 상관관계가 낮은 변수가 들어가면 회귀계수 크기의 변동은 적으나 (-0.00395->-0.00437) 높은 변수 education이 들어가면 -0.000478로 변동이 크다. 회귀계수는 설명변수가 종속변수에 미치는 영향정도이다. 값의 크기가 이렇게 바뀌면... 그리고 income 유의성이 없어짐

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1072.60937	59.45643	18.04	<.0001
income	income	1	-0.00395	0.00177	-2.23	0.0297

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1062.77655	53.38366	19.91	<.0001
income	income	1	-0.00437	0.00159	-2.74	0.0082
S02Pot	S02Pot	1	0.43244	0.11215	3.86	0.0003

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1351.92179	93.13023	14.52	<.0001
income	income	1	-0.00047841	0.00186	-0.26	0.7983
Education	Education	1	-36.00580	9.79572	-3.68	0.0005

```
proc reg data=ds.smsa;
     model mortality=income--S02Pot/vif collin;
run;
```

VIF -> HCPot, NOxPot 문제

상태지수 -> HCPot, NOxPot 문제, 동일한 결과 - 상태지수가 10이상인 행에서 변동비율이 큰(일반적으로 0.7 이상) 설명변수들이 (절편 제외) 2개 이상 존재하면 그 변수들이 다중공선성을 일으킴

NOxPot와 종속변수 상관계수는 음이었는데 다중공선성으로 양의 부호로 바뀌었다. 그리고 개별적으로 종속변수와 상관관계 유의하지 않았으나 HCPOT이런 경우 반드시 해결해야 한다. 다중공선성 문제가 진단되었지만 종속변수와 유의한 설명변수의 부호가 바뀌지 않았다면 굳이 문제 변수를 제거할 필요 없음.

여기서는 부호가 바뀌었으므로 (HCPot, NOxPot) 문제 변수 둘 중 하나 제거 – 종속변수와 상관계수가 낮은 NOxPot을 제외하고 재분석 실시한다

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1025.42773	52.42944	19.56	<.0001	0
income	income	1	-0.00306	0.00158	-1.93	0.0583	1.12571
HCPot	HCPot	1	-1.52400	0.59741	-2.55	0.0136	68.90525
NOxPot	NOxPot	1	2.84599	1.24198	2.29	0.0259	75.56807
SO2Pot	SO2Pot	1	0.19607	0.17024	1.15	0.2545	2.63313

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	income	HCPot	NOxPot	SO2Pot
1	3.41732	1.00000	0.00098527	0.00100	0.00063583	0.00062959	0.01201
2	1.14778	1.72549	0.00283	0.00238	0.00355	0.00245	0.00460
3	0.42128	2.84810	0.00334	0.00367	0.00118	0.00002053	0.39645
4	0.00804	20.61565	0.94585	0.93273	0.01955	0.03259	0.03487
5	0.00558	24.74516	0.04698	0.06022	0.97508	0.96431	0.55207

R 활용

```
fit=lm(Mortality~income+HCPot+NOxPot+SO2Pot)
library(car)
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?
```

```
> fit=lm(Mortality~income+HCPot+NOxPot+SO2Pot)
> vif(fit) # variance inflation factors

   income     HCPot    NOxPot    SO2Pot
 1.125708 68.905252 75.568075  2.633135
> sqrt(vif(fit)) > 2 # problem?
income  HCPot NOxPot SO2Pot
 FALSE   TRUE   TRUE  FALSE
```

VIF에 의해 문제가 되는 변수가 true가 진단되어 HCPot, NOXPot 두 변수