Chapter 6 다중공선성

다중회귀($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$)는 설명변수와 종속변수간의 관계에 대한 유의성 검정(F-검정)과 각 설명변수의 유의성(t-검정)에 가장 큰 관심을 갖지만 다음 사항에 대해 의구심을 갖게 된다. (1)서로 다른 설명변수의 상대적 중요 정도는 무엇인가?(표준화 회귀계수) (2)종속변수에 대한 설명변수의 설명력의 크기는 얼마인가? (OLS 추정 회귀계수)

만약 설명변수가 서로 독립이고 모형 설정 시 고려되지 않은 다른 설명변수와 독립이라면 설명변수의 회귀계수(β_k , 상대적 효과, 설명력, marginal effect)에 의해 위의 질문에 답이 가능하다. 그러나 현실에서 설명변수가 완벽하게 독립인 경우는 없다. 설명변수들 간 상관계수가 낮으면(유의하지 않으면) 여전히 회귀계수 해석이 적절하다. 그러나 상관관계가 높으면(유의하면) OLS 회귀계수 추정과 검정(추정 분산이 변함)이 쓸모가 없게 되는데 이를 다중공선성(Multicollinearity) 문제라 한다.

6.1 다중공선성 문제 맛보기

EXAMPLE 6-1 상관계수 구하기

종속변수 FSIQ, 설명변수 5개(VIQ, PIQ, WEIGHT, HEIGHT, MRI)를 고려한 다중모형을 생각해 보자. 🖫MRI_IQ.txt

앞의 산점도 행렬에서 설명변수 간의 상관 관계 정도를 시각적으로 판단할 수 있으나 관계 정도가 유의한지 알아보기 위하여 상관계수 유의성 검정을 해 보자. 다중공선성 진단을 위한 통계량이 있으나 산점도나 상관계수만으로 1차 진단을 한다. 산점도만으로는 정확한 판단이 어려워 상관분석(Pearson 상관계수)을 실시한다. 상관계수는 두 변수 간의 상관관계만 측정하므로 다중공선성 진단에 VIF, Condition Index 사용한다.

```
proc corr data=mri nosimple;
    var VIQ PIQ Weight Height MRI;
run;
```

	VIQ	PIQ	Weight	Height	MRI
VIQ	1.00000	0.77602 <.0001	-0.07609 0.6498	-0.10706 0.5223	0.30028 0.0670
PIQ	0.77602 <.0001	1.00000	0.00251 0.9881	-0.08154 0.6265	0.37778 0.0194
Weight	-0.07609 0.6498	0.00251 0.9881	1.00000	0.70000 <.0001	0.51338 0.0010
Height	-0.10706 0.5223	-0.08154 0.6265	0.70000 <.0001	1.00000	0.58073 0.0001
MRI	0.30028 0.0670	0.37778 0.0194	0.51338 0.0010	0.58073 0.0001	1.00000

유의수준 0.05 하에서 설명변수 (VIQ, PIQ), (PIQ, MRI), (WEIGHT, HEIGHT), (HEIGHT, MRI)가 유의하였다.

6.1.1 상관 관계가 낮은 설명변수

우선 상관계수가 낮은 두 변수($PIQ, Weight$)만을 설명변수로 사용한 회귀 모형을 생각해 보자.

```
PROC REG DATA=MRI;
     MODEL FSIQ=PIQ;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.90611	7.10359	0.55	0.5858
PIQ	PIQ	1	0.98477	0.06256	15.74	<.0001

```
PROC REG DATA=MRI;
     MODEL FSIQ=WEIGHT;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	121.44883	25.80102	4.71	<.0001
WEIGHT	Weight	1	-0.05222	0.16883	-0.31	0.7589

```
PROC REG DATA=MRI;
     MODEL FSIQ=PIQ WEIGHT;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.13817	11.55584	1.05	0.3007
PIQ	PIQ	1	0.98491	0.06272	15.70	<.0001
WEIGHT	Weight	1	-0.05460	0.06036	-0.90	0.3719

설명변수 간에 상관 관계가 낮은 경우 설명변수의 회귀계수 OLS 추정치와 추정 오차(추정 표준편차 STD error)의 변화가 거의 없다. 물론 유의하지 않은 설명변수의 추정 분산은 크기가 달라진다. 다중회귀에서 설명변수들간의 상관 관계가 낮은 경우(다중공선성 문제가 발생하지 않음) 회귀계수(편 미분 계수($\hat{\beta}_k$)를 이용하여 각 설명변수가 종속변수에 미치는 영향 정도 설명해도(물론 표준화 회귀계수이지만) 무방하다.

6.1.2 상관 관계가 유의한 경우

이제 상관 관계가 유의한 두 변수(PIQ, VIQ)가 있는 회귀 모형을 생각해 보자.

```
PROC REG DATA=MRI;
     MODEL FSIQ=VIQ;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.52949	6.46731	0.55	0.5886
VIQ	VIQ	1	0.98120	0.05654	17.36	<.0001

```
PROC REG DATA=MRI;
     MODEL FSIQ=PIQ VIQ;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-10.11448	1.53835	-6.57	<.0001
PIQ	PIQ	1	0.53251	0.02036	26.15	<.0001
VIQ	VIQ	1	0.57412	0.02006	28.62	<.0001

PIQ, VIQ의 회귀계수 추정치가 0.98에서 0.57로 바뀌었다. 다른 설명변수에 따라 추정 회귀계수의 변화가 심하므로 회귀계수가 더 이상 종속변수에 대한 상대 설명력으로 해석될 수 없다.

다른 측면에서 다중공선성 문제를 살펴보자. 다음 추정 결과를 보면 (MRI, VIQ) 설명변수를 사용하면 MRI 회귀계수의 추정치가 0.00001801이지만 (MRI, VIQ, PIQ) 설명변수 사용하면(VIQ, PIQ 상관 관계 매우 높음) MRI 회귀계수 추정치는 -0.00000899로 부호가 변한다.(두뇌 세포 수가 많으면 FSIQ 점수가 낮다?) 또한 (MRI, VIQ, PIQ)가 모두 유의하다고 검정되고 잔차분석만으로는 다중공선성 문제점을 발견하지 못한다. 즉, 잘못된 회귀 모형 추정 결과를 발표하게 될 것이다.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-10.88005	16.34802	-0.67	0.5101
MRI	MRI	1	0.00001801	0.00001876	0.96	0.3437
VIQ	VIQ	1	0.96409	0.05934	16.25	<.0001

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-3.26299	3.48702	-0.94	0.3560
PIQ	PIQ	1	0.54290	0.01995	27.21	<.0001
MRI	MRI	1	-0.00000889	0.00000411	-2.16	0.0375
VIQ	VIQ	1	0.57462	0.01908	30.12	<.0001

6.2 다중공선성 문제 및 발견 방법

6.2.1 다중공선성 문제

관측치	X_1	X_2	Y
1	2	6	23
2	8	9	83
3	6	8	63
4	10	10	103

위의 데이터에서 설명변수 X_1, X_2와 상관계수는 1이다.(완벽한 상관관계) 데이터 행렬 X의 한 열이 다른 열의 선형 함수 형태로 표현되므로 $|X'X|=0$ ($\hat{\beta}=(X'X)^{-1}X'\underline{y}$)이다. 이로 인하여 회귀계수 추정치는 무한히 많이 존재한다.

$$\hat{Y}=-87+X_1+18X_2,\ \hat{Y}=-7+9X_1+2X_2$$

이처럼 설명변수간의 상관 관계가 높으면 추정 회귀계수를 믿을 수 없게 된다. 다중회귀모형에서 회귀계수의 의미는 설명변수의 영향력(한 단위 변화할 때 종속변수의 변화량)이므로 이제 더 이상 이런 해석이 불가능해진다.

설명변수들간의 상관 관계가 유의하면(높으면) $|X'X|\approx 0$이 된다. ($|X'X|\approx 0$에 대하여: 자료행렬 X의 열은 각 설명변수가 된다. 설명변수가 상관 관계가 높다는 것은 한 설명변수를 다른 설명변수의 선형 함수로 표시할 수 있다는 것이다. ($X_k\approx aX_j$) 행렬의 성질에 의하면 한 열이 다른 열의 선형 함수로 표현되면 행렬식의 값은 0이다.) $(X'X)^{-1}=\frac{1}{|X'X|}adj(X'X)$ 이므로 $(X'X)^{-1}$가 매우 커지게 된다. 회귀계수 OLS 추정치 $\hat{\underline{\beta}}=(X'X)^{-1}X'\underline{y}$, OLS 추정치의 분산 $s^2_{\hat{\beta}}=MSE(X'X)^{-1}$ 이므로 다중공선성 문제가 발생하면 추정치가 불안해진다. (계수 부호까지도 반대가 되는 경우 발생) 이는 모형의 유의성 검정 F-검정이나 각 설명변수 유의성 검정 t-검정, 그리고 잔차분석에 의해서도 발견되지 않는다.

6.2.2 문제 발견

①산점도나 상관계수 이용

산점도 행렬이나 상관계수를 계산하여 상관 관계가 높은 설명변수들을 판단하고 다중공선성 문제가 일어날 것이라는 예상을 한다. 다중공선성 문제를 일으킬 것이라 판단되는 설명변수를 처음부터 제거하여도 무방하다. 제외할 때는 (1)종속변수와 상관 관계가 낮은 설

명변수(VIQ, PIQ 중에는 PIQ를 제거한다. (2)상관계수의 값의 차이가 크지 않으면 해석하기 하기 용이한 변수를 남기면 된다.

산점도나 상관계수는 두 변수 간의 상관 관계만 파악할 수 있으므로 하나의 설명변수가 2개 이상의 설명변수의 선형 결합($X_k = a_1 X_j + a_2 X_l + a_3 X_m + \ldots$)으로 표현되어 발생하는 다중공선성 문제는 발견할 수 없다. 이에 이용되는 통계량이 분산팽창지수(VIF)나 상태지수(condition index)이다. 산점도나 상관계수에 의해 다중공선성 문제를 예상하고 실제 발생 여부는 분산팽창지수나 상태지수를 이용하여 진단하는 방법을 권한다.

②분산팽창지수(Variation Index Factor) $VIF_k = \dfrac{1}{(1-R_k^2)}$

R_k^2 는 회귀 모형 $X_k = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{k-1} X_{k-1} + \cdots + \beta_{k+1} X_{k+1} + \beta_p X_p + e$ (종속변수는 X_k , 설명변수는 나머지 설명변수인 회귀 모형)의 결정 계수이다. 그러므로 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 10이상인 설명변수가 다중공선성 문제를 발생시킨다고 판단한다.

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

$(X'X)$ 의 대각행렬이 1이 되게 변환한 후(상관변환: correlation transformation) 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)들 간 약한 상관 관계가 존재하고 100 이상인 값이 있으면 상관 관계가 매우 유의한 설명변수가 존재한다. 즉 다중공선성 문제가 발생한다고 할 수 있다.

참고 고유치와 고유벡터

행렬 A 의 고유 방정식(characteristic equation) $|A_{n\times n} - \lambda I_n| = \underline{0}$ 를 만족하는 $\lambda_1, \lambda_2, \ldots, \lambda_n$ 들을 고유치(eigen value, characteristic value, latent value)라 하고 각 고유치에 대해 $A\underline{e}_i = \lambda_i \underline{e}_i$ 를 만족하는 벡터를 고유 벡터(eigen vector)라 한다. 또한 고유 벡터를 구하는 식에서 $A\underline{e}_i = \lambda_i \underline{e}_i$ 을 살펴보면 행렬 A의 크기 고유치에 나타나 있다. 대칭행렬에 대해서는 다음이 성립한다. 회귀분석에 다루는 $X'X$, 공분산 행렬이나 상관 행렬은 모두 대칭 행렬이므로 알아 두면 유용한 성질이다.

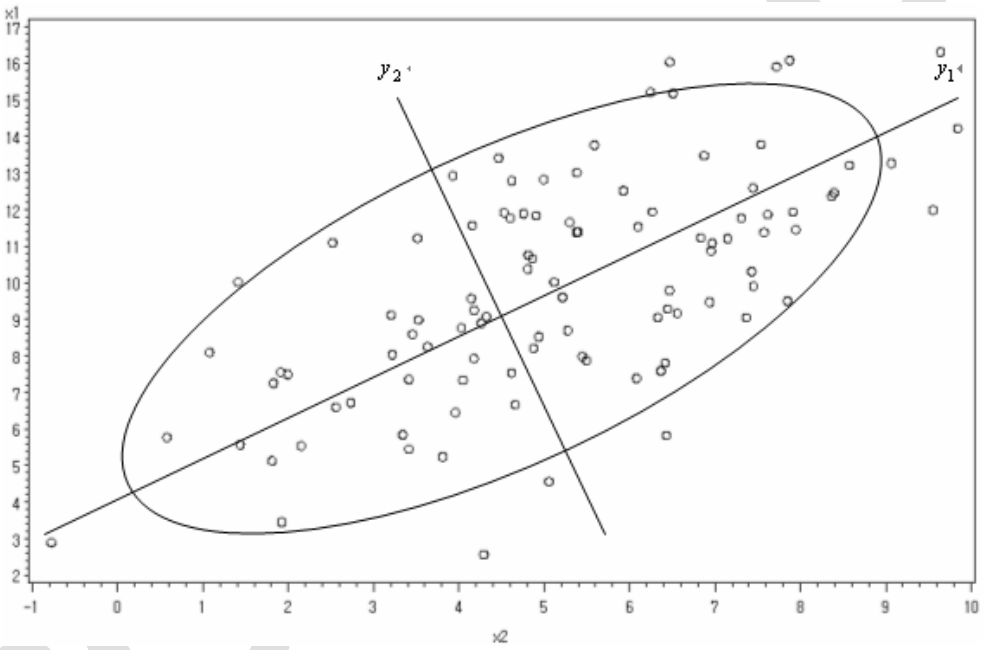
①고유치는 실수이다.

②대칭 행렬은 대각화가 가능하다(Diagnosable). $A = U^{-1}DU$ D는 대각원소가 A의 고유치인 대각 행렬이고 U는 직교 행렬이다.

③고유 벡터는 orthogonal하다. 즉 $e_i' e_j = 0 \; for \; i \neq j$

④행렬의 계수와 0이 아닌 고유치의 수는 같다. 즉 0인 고유치가 존재하는 행렬은 full-rank가 아니며 역 행렬이 존재하지 않는다.

평균 $\underline{\mu} = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$, 공분산 행렬 $\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$인 이변량 정규분포를 고려하자. $\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$의 고유치는 $\lambda_1 = 9.7$, $\lambda_2 = 3.2$이고 각 고유치에 대응하는 고유벡터는 $\underline{a}_1 = \begin{bmatrix} 0.94 \\ 0.33 \end{bmatrix}$, $\underline{a}_2 = \begin{bmatrix} -0.33 \\ 0.94 \end{bmatrix}$이다.



EXAMPLE 6-1 상관계수 구하기

지적 능력에 영향을 준다고 생각된 변수들 중 (VIQ, PIQ, MRI) 만을 고려한 모형을 생각하자. MRI_IQ.txt

```
PROC REG DATA=MRI;
    MODEL FSIQ=PIQ VIQ MRI/VIF COLLIN;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-3.26299	3.48702	-0.94	0.3560	0
PIQ	PIQ	1	0.54290	0.01995	27.21	<.0001	2.66837
VIQ	VIQ	1	0.57462	0.01908	30.12	<.0001	2.51425
MRI	MRI	1	-0.00000889	0.00000411	-2.16	0.0375	1.16665

VIF 통계량으로는 다중공선성 문제를 일으키는 설명변수를 찾지 못했다.

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	PIQ	VIQ	MRI
1	3.95707	1.00000	0.00038561	0.00091212	0.00098204	0.00033642
2	0.03122	11.25746	0.05650	0.10702	0.13381	0.03595
3	0.00875	21.26036	0.00520	0.83883	0.85393	0.00264
4	0.00295	36.62942	0.93792	0.05324	0.01127	0.96107

마지막 두 행의 상태지수가 10이상이므로 다중공선성 문제가 발생함을 알 수 있다. 상태지수에 대한 체크는 마지막 1-2개 행에 대해 실시하면 된다. 어떤 설명변수로 인하여 발생하였는지는 고유치가 설명한 설명변수 변동(열의 합은 1 즉 100%) 중 유난히 큰 설명변수가 존재하면 그 변수들간에 상관 관계로 인하여 다중공선성 문제가 발생한다. 마지막 행은 MRI 변동만을 가장 많이(96%) 설명하므로 문제가 아니다. 3번째 행(고유치)의 경우에는 PIQ, VIQ의 변동을 각각 83%, 85% 설명하므로 PIQ와 VIQ에 의해 다중공선성 문제가 발생함을 알 수 있다.

6.3 다중 공선선 문제 해결

6.3.1 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.

MRI_IQ 예제에서 (MRI, VIQ, PIQ)를 설명변수로 한 경우 VIQ, PIQ에 의해 다중공선성 문제가 발생하였다. VIQ, PIQ중 어느 변수를 제외할 것인가? 종속변수를 더 잘 설명하는 변수를 남겨두는 것이 일반적이다. 종속변수를 더 잘 설명한다? 종속변수와 상관계수가 높음을 의미한다. 종속변수와 설명변수간 상관계수를 구해보자.

```
proc corr data=mri nosimple;
    var VIQ PIQ Weight Height MRI;
    with FSIQ;
run;
```

VIQ와 PIQ 중 종속변수 FSIQ와 상관 관계가 높은 것은(더 잘 설명하는 변수는) VIQ이다. 그러므로 PIQ를 제외하고 회귀모형을 추정하면 된다. 그런데 PIQ도 종속변수를 설명하는 능력이 매우 유의하고 상관 계수 크기도 VIQ와 크기가 비슷하다. 만약 FSIQ에 대한 설명에서 VIQ보다 PIQ가 더 현실적으로 부합한다면 PIQ를 남겨두어도 무방하다.

	VIQ	PIQ	Weight	Height	MRI
FSIQ	0.94511 <.0001	0.93443 <.0001	-0.05148 0.7589	-0.10501 0.5304	0.33371 0.0406

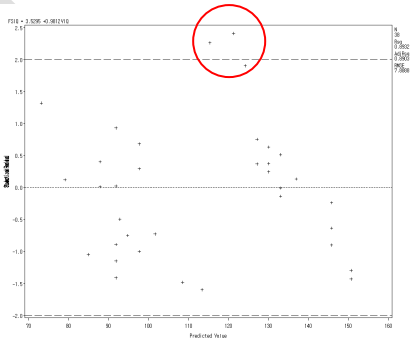
PIQ 제외하고 추정을 다시 해 보자.

```
proc reg data=mri;
    model fsiq=viq mri;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.88005	16.34802	-0.67	0.5101
VIQ	1	0.96409	0.05934	16.25	<.0001
MRI	1	0.00001801	0.00001876	0.96	0.3437

MRI 변수가 유의하지 않으므로 이를 제외하고 다시 분석하면 다음과 같다. 이상치가 2개 존재하는 문제를 제외하는 특별한 문제는 없다. 2개를 제외하니 다시 한 개가 나타나 최종적으로 3개가 제외되었다.

```
proc reg data=mri;
    model fsiq=viq/r;
    plot student.*p./vref=2 -2;
run;
```



```
proc reg data=mri;
    model fsiq=viq/r;
    reweight obs.=9;
    reweight obs.=13;
    reweight obs.=2;
    plot student.*p./vref=2 -2;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.21595	5.11142	0.82	0.4154
VIQ	1	0.96195	0.04484	21.45	<.0001

6.3.2 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.

$$Y_i = \beta_0 + \beta_1 P_{1i} + \beta_2 P_{2i} + \cdots + \beta_p P_{pi} + e_i$$

설명변수는 주성분 $P_{ki} = a_{k1} X_{1i} + a_{k2} X_{2i} + .. + a_{kp} X_{pi}, k = 1,2,.., p$ 이다.

주성분 변수는 원 변수의 변동(공분산, 상관계수)을 이용하여 변동을 가장 잘 설명하는 주성분 변수를 차례로 찾는 것이다. 물론 주성분 변수의 개수는 원 변수의 개수와 같다. 대신 첫번째 주성분의 설명력이 가장 높고 차례로 낮아지며 서로 독립이다. 주성분 변수를 이용하면 다중공선성 문제는 해결할 수 있으나 주성분 변수(z_k)에 대한 해석이 용이하지 않는 단점이 있어 자주 사용되지는 않는다.

주성분 분석을 실시하여 주성분 변수를 얻어보자. OUT 옵션에 의해 주성분 결과가 저장된다.

```
proc princomp data=mri out=prin;
    var viq piq mri;
run;
```

고유치(eigen value, λ_i)는 원 변수의 상관계수 행렬(R)로부터 구해진 것이다. 만약 공분산 행렬을 이용할 경우에는 `data=mri covariance out=prin;`을 쓰면 된다. 일반적으로 공분산 행렬이 더 유용한데 설명변수의 단위가 많이 다를 때는 사용하지 않는 것이 좋다. 물론 설명변수의 측정 단위가 유사하다면 공분산 행렬을 사용하자.

Correlation Matrix

	VIQ	PIQ	MRI
VIQ	1.0000	0.7760	0.3003
PIQ	0.7760	1.0000	0.3778
MRI	0.3003	0.3778	1.0000

▶ $R\underline{e}_i = \lambda_i \underline{e}_i$

각 주성분 변수의 원 변수 변동(상관계수)에 대한 설명력(proportion)은 첫번째가 가장 높고(0.68) 갈수록 떨어진다. 마지막 열의 누적 설명력(Cumulative)은 당연히 1이다. 고유치(e_i)는 주성분 변수를 만들 때 사용되는 원 변수의 선형계수이다.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.00511727	1.22947053	0.6684	0.6684
2	0.77564674	0.55641076	0.2585	0.9269
3	0.21923598		0.0731	1.0000

Eigenvectors

	Prin1	Prin2	Prin3
VIQ	0.628579	-.367488	0.685450
PIQ	0.647337	-.241305	-.722998
MRI	0.431095	0.898179	0.086210

$$P_{ki} = a_{k1}X_{1i} + a_{k2}X_{2i} + .. + a_{kp}X_{pi}, k = 1,2,.., p$$

이제 저장해 둔 주성분 분석 결과를 출력해 보자.

```
proc print data=prin;
run;
```

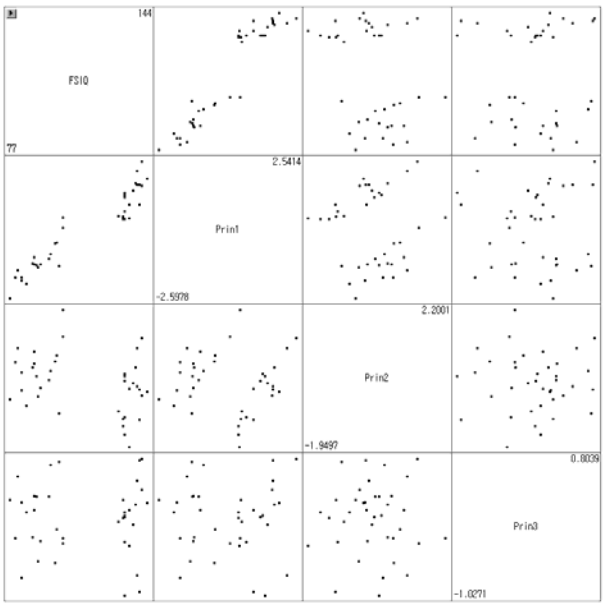
FSIQ	VIQ	PIQ	Weight	Height	MRI	Prin1	Prin2	Prin3
133	132	124	118	65	816932	0.37338	-1.56528	0.08199
139	123	150	143	73	1038437	2.18754	1.04308	-0.75562
133	129	128	172	69	965353	1.28754	0.27724	0.04070
137	132	134	147	65	951545	1.45959	-0.00581	-0.07802
99	90	110	146	69	928799	-0.51391	0.64175	-0.59217
138	136	131	138	66	991305	1.71947	0.45430	0.18422

주성분 변수의 변수 이름은 PRIN1, PRIN2로 자동 설정되어 있다. 정말 주성분 변수들은 서로 독립일까? 당근이네. 그러므로 주성분 변수를 설명변수로 사용하면 다중공선성 문제 해결된다.

```
proc corr data=prin;
    var prin1 prin2 prin3;
run;
```

	Prin1	Prin2	Prin3
Prin1	1.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin2	0.00000 1.0000	1.00000	0.00000 1.0000
Prin3	0.00000 1.0000	0.00000 1.0000	1.00000

우선 종속변수와 설명변수들의(주성분 변수 PRIN1, PRIN2, PRIN3) 산점도 행렬을 그려 보자.



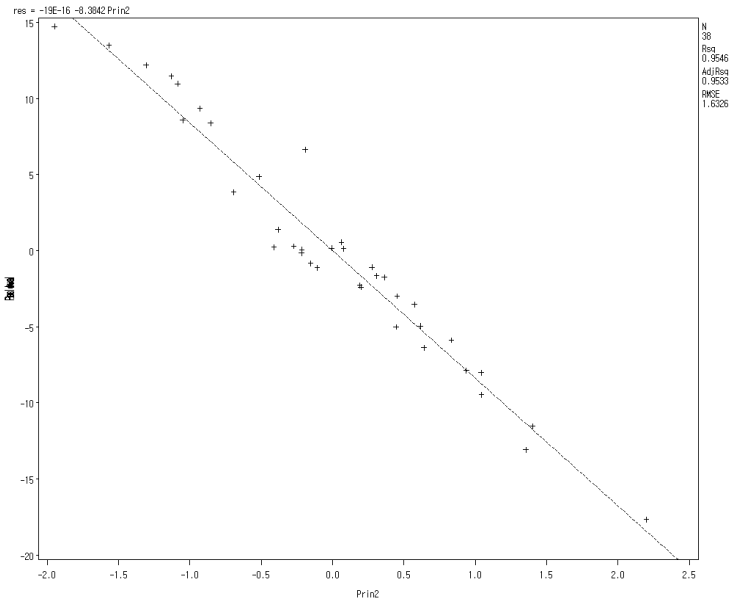
변수 순서(FSIQ, PRIN1, PRIN2, PRIN3)

설명변수 PRIN1, PRIN2, PRIN3의 상관 관계는 없어 보이고 PRIN1과 FSIQ는 직선 관계가 강해 보인다. PRIN2와 PRIN3의 경우 FSIQ와 직선 관계는 없어 보인다. 그러나 나중에 회귀모형 추정을 통한 유의한 변수를 선택해 보면 주성분 변수 PRIN2도 유의하다. 유의성은 PRIN1이 설명하고 남은 부분에 대한 설명 정도이므로 FSIQ와 관계가 직선 관계처럼 보이지 않으나 FSIQ에 영향을 미친다.

진짜일까? 이것을 확인해 보자. 어떻게 확인하지? PRIN1이 설명하고 남은 부분이란? 이것은 FSIQ를 종속변수, PRIN1을 설명변수로 하여 얻은 잔차이다. 그러므로 잔차와 PRIN2의 산점도를 그리면 직선 관계가 나타난다. 그러므로 PRIN2가 유의한 것이다. 이처럼 산점도에서는 유의하지 않아 보이더라도 포기 말고 회귀계수의 유의성 검정은 반드시 필요하다.

```
proc reg data=prin;
    model fsiq=prin1;
    output out=out1 r=res;
run;
```

```
proc reg data=out1;
    model res=prin2;
    plot res*prin2;
run;
```



주성분 변수를 설명변수로 하여 회귀분석을 실시해 보자.

$$FSIQ = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 P_3 + e$$

```
proc reg data=prin;
    model fsiq=prin1 prin2 prin3;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	113.55263	0.27238	416.89	<.0001
Prin1	1	15.94921	0.19494	81.82	<.0001
Prin2	1	-8.38416	0.31343	-26.75	<.0001
Prin3	1	0.10957	0.58954	0.19	0.8537

세 번째 주성분 PRIN3가 유의하지 않으므로 이를 제외하고 모형을 다시 추정하면 다음과 같다. 질문? PRIN1, PRIN2의 계수 추정치가 바뀌었을까요? 아니죠. 설명변수는 서로 독립인 경우에는 항상 동일한 추정치임을 잊지 말기 바랍니다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	113.55263	0.26860	422.76	<.0001	0
Prin1	1	15.94921	0.19223	82.97	<.0001	0.94831
Prin2	1	-8.38416	0.30907	-27.13	<.0001	-0.31005

$$FSIQ = 113.55 + 15.94 * \Pr in1 - 8.38 * \Pr in2$$

FSIQ에 미치는 영향 정도도 PRIN1이 크다(표준화 회귀계수). 그럼 PRIN1과 PRIN2 변수의 의미는 무엇인가? 즉 설명변수의 의미가 무엇인가 하는 것이다. 주성분 변수에 대한 해석은 원 변수 선형 결합에 사용되는 선형계수(고유벡터)를 이용한다.

	Prin1	Prin2	Prin3
VIQ	0.628579	-.367488	0.685450
PIQ	0.647337	-.241305	-.722998
MRI	0.431095	0.898179	0.086210

고유치는 PRIN1은 변수가 유사하게 영향을 미치고 PRIN2는 MRI가 PRIN3는 VIQ와 PIQ(역으로 작용)가 주된 역할을 하므로 PRIN2는 “두뇌 크기” 주성분 변수, PRIN3는 “IQ” 주성분 변수로 명명할 수 있으나 PRIN1의 이름은 무엇으로 할 것인가? 이처럼 FSIQ에 영향을 주는 설명변수의 의미가 명확하지 않아 자주 사용하는 방법은 아니다.

주성분 변수를 이용한 회귀분석의 장점은 모든 설명변수의 정보를 이용하므로(PRIN1, PRIN2 각각에 설명변수 VIQ, PIQ, MRI) 예측력이 높다. 그러므로 회귀분석의 주 목적 예측하는데 있다면 다중공선성 문제 해결로 주성분 분석을 이용해도 무방하다.

```
proc reg data=prin;
    model fsiq=prin1 prin2/p;
run;

proc reg data=prin;
    model fsiq=viq/p;
run;
```

```
R-Square      0.9954
Adj R-Sq      0.9952

Sum of Squared Residuals            95.95260
Predicted Residual SS (PRESS)      110.71009

R-Square      0.8932
Adj R-Sq      0.8903

Sum of Squared Residuals          2240.37920
Predicted Residual SS (PRESS)     2471.05084
```

6.3.3 능형 회귀분석

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.

EXAMPLE MRI_IQ 데이터에서 설명변수로 (VIQ, PIQ, MRI) 만을 고려해 보자.

```
PROC REG DATA=MRI OUTVIF OUTEST=OUT1 RIDGE=0 TO 1 BY 0.05;
    MODEL FSIQ=VIQ PIQ MRI;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

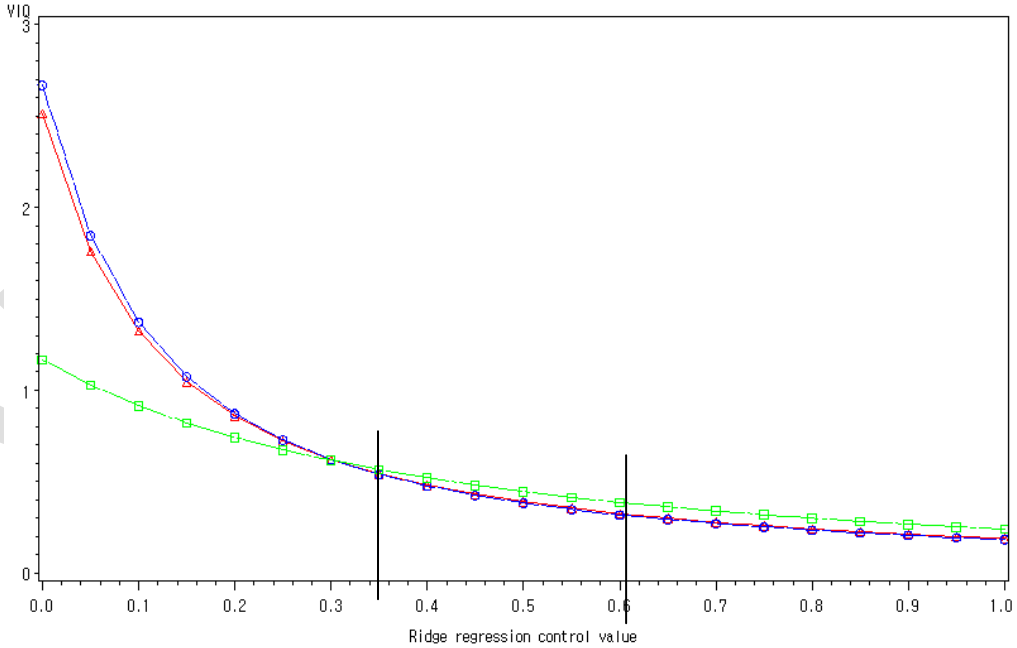
첫 행은 OLS 추정치를 보여준다.(왜냐하면 $c=0$) _TYPE_="RIDGEVIF"는 주어진 c 에 있어 각 설명변수의 VIF 값을 나타내고 _TYPE_="RIDGE"는 각 설명변수의 능형 추정치를 보여준다. 다음 산점도를 그려 VIF 와 능형 추정량이 안정화 되는 c 값 중 가장 작은 것을 택하면 된다. 적은 값을 택하는 이유는 c 가 커질수록 불편성은 증가하기 때문이다.

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	VIQ	PIQ	MRI	FSIQ
1	MODEL1	PARMS	FSIQ	.	.	1.67907	-3.2630	0.57462	0.54290	-0.00001	-1
2	MODEL1	RIDGEVIF	FSIQ	0.00	.	.	.	2.51425	2.66837	1.16665	-1
3	MODEL1	RIDGE	FSIQ	0.00	.	1.67907	-3.2630	0.57462	0.54290	-0.00001	-1
4	MODEL1	RIDGEVIF	FSIQ	0.05	.	.	.	1.76225	1.84615	1.02862	-1
5	MODEL1	RIDGE	FSIQ	0.05	.	1.83454	-2.9081	0.55524	0.52702	-0.00000	-1
6	MODEL1	RIDGEVIF	FSIQ	0.10	.	.	.	1.32613	1.37302	0.91615	-1
7	MODEL1	RIDGE	FSIQ	0.10	.	2.20134	-2.2958	0.53749	0.51207	-0.00000	-1
8	MODEL1	RIDGEVIF	FSIQ	0.15	.	.	.	1.04836	1.02420	0.82248	-1

```
DATA TEMPO;
    SET OUT1;
    IF (_TYPE_="RIDGEVIF");
RUN;
```

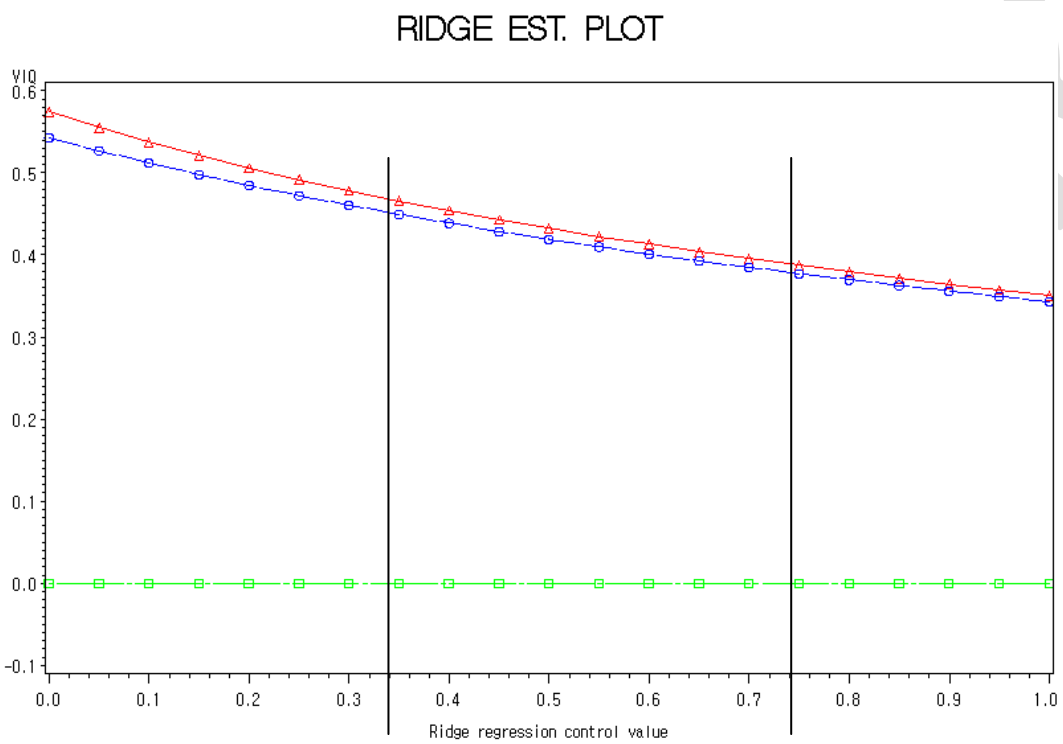
```
PROC GPLOT DATA=TEMPO;
    TITLE 'VIF PLOT';
    SYMBOL1 V=TRIANGLE I=JOIN C=RED;
    SYMBOL2 V=CIRCLE I=JOIN C=BLUE L=5;
    SYMBOL3 V=SQUARE I=JOIN C=GREEN L=10;
    PLOT (VIQ PIQ MRI)*_RIDGE_/OVERLAY;
RUN;
```

VIF PLOT



```
DATA TEMP1;
    SET OUT1;
    IF (_TYPE_="RIDGE");
RUN;
```

```
PROC GPLOT DATA=TEMP1;
    TITLE 'RIDGE EST. PLOT';
    SYMBOL1 V=TRIANGLE I=JOIN C=RED;
    SYMBOL2 V=CIRCLE I=JOIN C=BLUE L=5;
    SYMBOL3 V=SQUARE I=JOIN C=GREEN L=10;
    PLOT (VIQ PIQ MRI)*_RIDGE_/OVERLAY;
RUN;
```



MRI는 다중공선성 문제를 일으키지 않으므로 시작부터 안정화 되어 있다. 각 설명변수의 VIF와 능형 추정량이 안정화 되는 값 중 가장 적은 값은 0.3로 보이므로(주관적인 판단) $c = 0.3$ 능형 추정량을 사용한다. OUTEST 옵션을 사용하지 않으면 능형 추정치를 얻을 수 없다. OLS 추정치만 출력된다.

```
PROC REG DATA=MRI RIDGE=0.3 OUTEST=OUT1;
    MODEL FSIQ=VIQ PIQ MRI;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

RIDGE	_PCOMIT_	_RMSE_	Intercept	VIQ	PIQ	MRI
.	.	1.67907	-3.26299	0.57462	0.54290	-.000008894
0.3	.	4.10331	1.70720	0.47839	0.46050	0.000007643

$$FSIQ = 1707 + 0.47839 * VIQ + 0.4605 * PIQ + 0.000007643 * MRI$$

모든 설명변수를 고려한 모형에서도 다중공선성 문제로 인하여 MRI 부호가 음인 문제는

해결되었다. 이제 변수 제거 방법(다중공선 문제 일으키는 변수와 유의하지 변수 제외), 주성분 분석 방법, 그리고 능형 추정에 의한 예측치를 비교해 보자.

```
proc reg data=mri;
    model fsiq=viq;
    output out=out1 p=yhat_o;
run;

proc reg data=prin;
    model fsiq=prin1 prin2;
    output out=out2 p=yhat_p;
run;
```

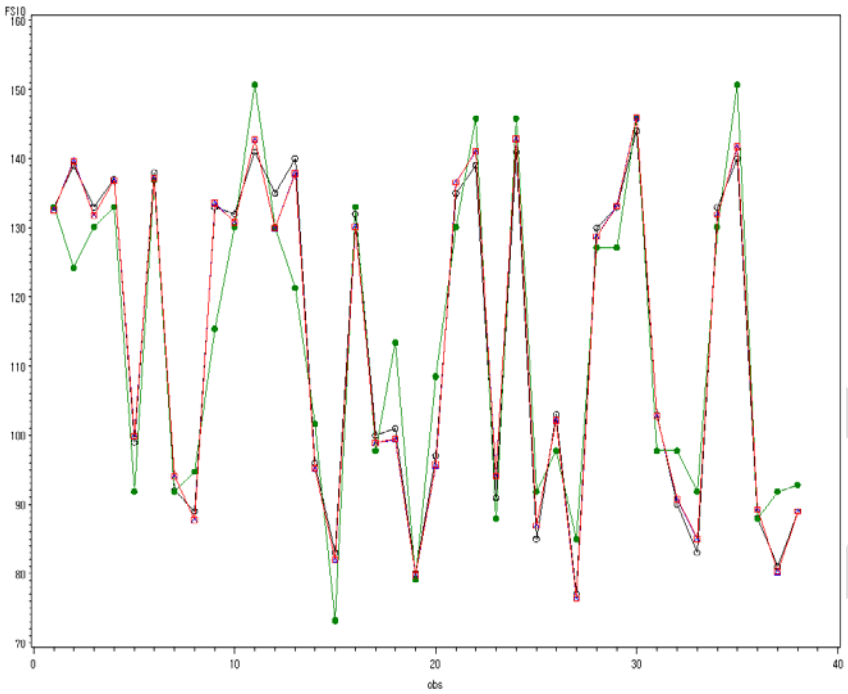
```
PROC REG DATA=MRI RIDGE=0.3;
    MODEL FSIQ=VIQ piq mri;
    output out=out3 p=yhat_r;
RUN;

data all;
    merge out1 out2 out3;obs+1;
    res_o=fsiq-yhat_o;
    res_p=fsiq-yhat_p;
    res_r=fsiq-yhat_r;
run;
```

OBS+1;은 OBS는 관측치 순서의 1, 2, 3,... 번호 부여하기 위함이다. RES_*는 잔차이다.

```
proc gplot data=all;
    goptions reset=all;
    symbol1 i=join v=circle c=black;
    symbol2 i=join v=dot c=green;
    symbol3 i=join v=triangle c=blue;
    symbol4 i=join v=square c=red;
    plot (fsiq yhat_o yhat_p yhat_r)*obs/overlay;
run;
```

주성분 변수를 이용한 회귀분석 방법과 능형 추정 방법이 예측을 잘한다. 물론 설명변수 세 개 모두를 사용하기 때문에 예측력을 높일 수 있는 것일 수도 있다. 그러므로 다중공선성 문제가 발생하는 경우 예측이 주 목적이라면 주성분 변수를 이용한 회귀분석이나 능형 추정 방법이 적당하다. 그리고 설명변수 해석까지 생각하면 능형 추정방법이 적절하다. 그러나 설명변수 개수가 충분하다면 다중공선성 문제를 발생시키는 설명변수를 제외하고 OLS 회귀 추정하는 것을 권한다.

res_r	res_p	res_o	FSIQ
0.35969	0.36867	-0.0475	133
-0.61403	-0.69682	14.7833	139
-1.23206	-1.23652	2.8961	133
0.12798	0.11943	3.9525	137
-0.91071	-0.97560	-2.1628	99
0.81186	0.83210	-1.0278	138
-2.05889	-2.09141	0.1628	92
-1.26795	-1.30091	-5.7808	89
-0.55174	-0.66429	12.6141	133
1.23418	1.23555	1.8961	132
-1.81920	-1.73113	-9.7090	141
-5.08638	-5.04949	4.8961	135
2.12005	2.01426	18.7269	140
0.75730	0.78877	-5.6492	96
1.04343	0.95716	9.8055	83
1.84538	1.87298	-1.0475	132
1.11968	1.11460	-2.2756	100
1.48888	1.57144	-12.4235	101
0.23572	0.20724	-0.9184	80
1.23292	1.31186	-11.5175	97
-1.57952	-1.61884	4.8961	135
-2.05389	-1.99839	-6.8030	139
-3.13203	-3.19459	3.0876	91
-1.85611	-1.81367	-4.8030	141
-1.95320	-1.94389	-6.8372	85
-0.83044	-0.80260	5.2756	103
0.53892	0.56672	-7.9688	77
-1.24972	-1.24554	-2.8397	130
-0.17248	-0.20587	5.8397	133
-1.98816	-1.96492	-1.8030	144
0.25642	0.22018	5.2756	103
-0.76293	-0.73039	-7.7244	90
-2.00668	-1.98220	-8.8372	83
-1.07831	-1.08051	2.8961	133
-1.80525	-1.72009	-10.7090	140
-1.23548	-1.26190	-0.0876	88
0.64459	-0.70607	-10.8372	81
-0.02155	-0.00863	-3.8184	89



HOMEWORK #9-1 DUE 5월 11일(수)

FITNESS 데이터에서 종속변수는 Oxygen(산소량)이다. 나머지 변수(6개) 설명변수로 하여 다중회귀모형을 실시한다고 하자. 🖫FITNESS_IQ.xls SPSS

①상관계수, 선점도 행렬에 의해 다중공선성 문제를 일으킬 변수가 있는지 판단하시오.

②수작업에 의해 다중공선성 문제를 일으키는 변수를 제외하고 모형을 추정하시오.

③VIF와 상태지수를 이용하여 다중공선성 문제를 발견하고 ①의 결과와 비교하시오.

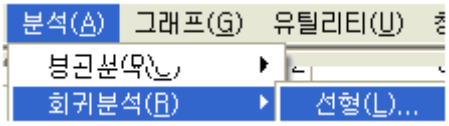
HOMEWORK #9-2 DUE 5월 11일(수)

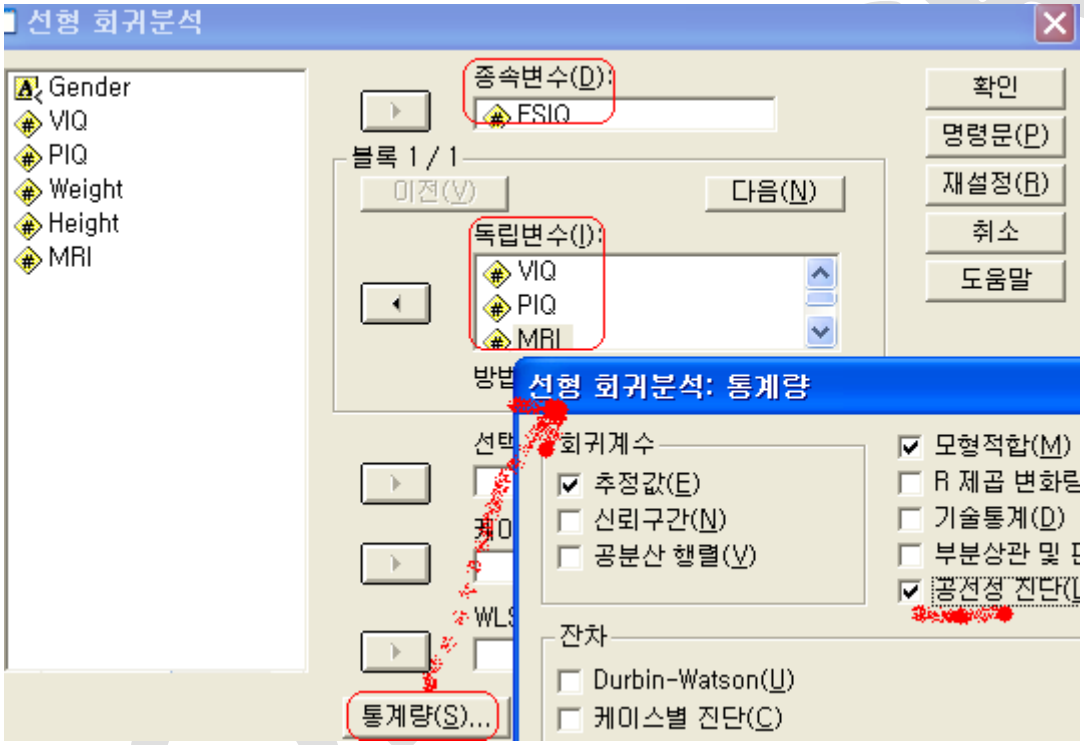
🖫SALES.txt 데이터에서 SALE(매출)에 영향을 미치는 요인으로 광고비(A), 기업 홍보비(P), 매출 관련 비용(E)을 생각하여 22개년 자료를 수집하였다. 회귀분석 하시오.

모형: $Y_t = \alpha + \beta_1 * A_t + \beta_2 * A_{t-1} + \beta_3 * P_t + \beta_4 * P_{t-1} + +\beta_5 * E_t$ SAS

SPSS에서 다중공선성 진단하기

회귀분석 메뉴를 선택하고 종속변수와 독립변수(설명변수)를 선택하고 "통계량" 설정 창을 선택하여 "공성선 진단"을 선택한다. 결과는 페이지 147의 SAS 결과와 동일하다.





계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
		B	표준오차	베타			공차한계	VIF
1	(상수)	-3,263	3,487		-,936	,356		
	VIQ	,575	,019	,553	30,116	,000	,398	2,514
	PIQ	,543	,020	,515	27,208	,000	,375	2,668
	MRI	-8,89E-06	,000	-,027	-2,165	,038	,857	1,167

공선성 진단[a]

모형	차원	고유값	상태지수	분산비율			
				(상수)	VIQ	PIQ	MRI
1	1	3,957	1,000	,00	,00	,00	,00
	2	,031	11,257	,06	,13	,11	,04
	3	,009	21,260	,01	,85	,84	,00
	4	,003	36,629	,94	,01	,05	,96

설명변수가 서로 다른 모형의 비교

설명변수의 군이 다르거나 개수가 다른 경우 어떤 회귀모형을 선택할지 무엇을 기준으로 선택할 것인가?

①수정 결정계수(adjusted determin): 설명변수가 추가적으로 삽입될 때 설명변수의 종속변수 설명력(결정계수)은 항상 증가하는 문제가 있다. 이에 대한 보완책으로 수정결정계수를 이용한다. $R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$, 이 방법은 사용되는 설명 변수가 같을 경우 사용된다. 그럼 변수 군이 다른 경우에는? 즉, 주성분을 이용한 회귀분석, 능형 회귀모형, 변수 제거한 추정치 중 어느 것이 더 예측 정도가 높은가? 아래 3가지 통계량을 이용한다.

②AIC(Akaike Information Criteria)= $n\ln(SSE/n) + 2(p-1)$ 작을수록 적합도가 높다.

③SBC(Schwarz's Bayesian criterion)= $n\ln(SSE/n) + (p-1)\ln(n)$ 작을수록 적합도가 높다.

④PRESS: $\frac{r_i}{1-h_i}$ (h_i는 Hat 행렬의 i-번째 대각원소)의 제곱합이다. 이 값이 크다는 것은 회귀모형의 예측이 부정확하다는 것이다.

SAS에서만 가능하다. AIC, SBC는 PLOT 옵션에서, PRESS는 MODEL의 R 옵션을 이용하면 된다. 수정 결정계수는 MODEL 문에 의해 자동 출력된다.

```
proc reg data=mri;
      model fsiq=viq mri/r;
      reweight obs.=9;
      reweight obs.=13;
      reweight obs.=2;
      plot student.*p./vref=2 -2 aic sbc;
run;
```

```
 38     1.0000    89.0000    92.2347       1.6306    -3.2347      6.073    -0.533  |     *|
```

```
Sum of Residuals                           0
Sum of Squared Residuals             1265.23685
Predicted Residual SS (PRESS)        1546.63007
```

회귀모형에서 유의한 설명변수의 종속변수 영향력 비교?

설명변수의 측정 단위가 다른 경우 종속변수에 대한 그들의 영향력을 비교할 때 사용되는 개념이 표준화 회귀계수다.

$Y_i^* = \frac{Y_i - \bar{Y}}{s_Y}$, $X_{ki}^* = \frac{X_{ki} - \bar{X}_k}{s_{X_k}}$, $(i = 1,2,...,n,\ k = 1,2,...,p)$ 표준화 하고 다음 회귀 모형의 OLS 추정치를 구하면 표준화 회귀계수라 한다. $y_i^* = \beta_0 + \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + .. + \beta_p X_{pi}^* + e_i$ 회귀계수 β_k는 편미분 계수($\frac{dy^*}{dx_k^*}$)에 해당하므로 다른 설명변수들의 값이 주어졌을 때 종속 변수 Y에 대한 설명 변수 X_k의 한계 영향력(한계 설명력)으로 해석된다.

다음은 SAS에서 표준화 회귀계수 추정치를 얻기 위한 옵션 사용방법이다. SPSS에서는 자동 출력된다.

```
PROC REG DATA=MRI;
     MODEL FSIQ=MRI VIQ PIQ/STB;
 RUN;
```

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	-3.26299	3.48702	-0.94	0.3560	0
MRI	MRI	1	-0.00000889	0.00000411	-2.16	0.0375	-0.02710
VIQ	VIQ	1	0.57462	0.01908	30.12	<.0001	0.55349
PIQ	PIQ	1	0.54290	0.01995	27.21	<.0001	0.51515

지적 능력(FSIQ)에 VIQ가 영향을 가장 많이 영향을 미치며, PIQ 설명변수도 아이 지적 능력(FSIQ) 미치는 영향력은 VIQ와 비슷하다. MRI의 영향력이 가장 낮다. 아이 지적 능력에 가장 영향을 많이 미치는 요인은 VIQ(언어 능력)이다.

물론 이런 해석은 설명변수가 유의하고, 다중공선성 문제 없고, 잔차분석 후 최종 모형을 가지고 해석해야 한다.