Chapter 4 다중회귀

3장에서는 설명변수가 하나인 단순회귀모형에 대한 추론과 분석 방법에 관해 다루었다.(간편성, 회귀분석 개념에 대한 이해) 그러나 현실 세계에서는 (1)설명변수 하나만으로 설명력이 부족하고 (2)유의한 설명변수간 영향력 비교가 요구된다. 이로 인하여 설명변수가 2개 이상인 회귀모형에 대한 분석이 필요하게 된다. 이를 다중회귀분석이라 한다.

동일 설명변수의 1차항과 2차항이 동시에 있는 모형은 다중회귀모형이라기보다는 다항회귀모형(Polynomial Regression)이라 한다. 동일변수의 1차항과 2차항이 동시에 들어간 회귀모형은 설명변수간 다중공선성 문제가 발생하므로 이를 해결하기 위하여 설명변수를 표준화 하여 사용해야 한다고 언급하였다.

4.1 모형과 가정

설명변수의 개수가 $p(\geq 2)$이고 관측치 개수가 n인 경우 다중회귀모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i, \quad i = 1,2,...,n \quad \text{--- ①}$$

○ $e_i \sim iidN(0,\sigma^2)$: 오차항 e_i는 (1)독립성 (2)정규성 (3)등분산성 (단순회귀분석과 동일)

○ $\beta_0, \beta_1, ..., \beta_p$는 회귀계수이며 모수(parameter)이다. β_i는 i-번째 설명변수의 편미분계수로 다른 설명변수의 값이 고정일 때 영향력을 의미한다.

○설명변수 $X_{1i}, X_{2i}, ..., X_{pi}$는 deterministic이고(확률변수가 아니다, 그러므로 종속변수의 분포는 오차항의 분포와 동일하다) 적어도 하나 이상은 측정형 변수이어야 한다. 설명변수가 모두 분류형(범주형)이면 분산분석(ANOVA)이다. 회귀분석에서 분류형 설명변수를 지시변수(indicator variable) 혹은 가변수(dummy variable)라 한다.

식①을 행렬로 표현하면 다음과 같다.

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots x_{1p} \\ 1 & x_{21} & x_{22} \cdots x_{2p} \\ & \cdots & \\ 1 & x_{n1} & x_{n2} \cdots x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix} \Rightarrow \underline{y} = X\underline{\beta} + \underline{e}, \quad \underline{e} \sim N(\underline{0}, \sigma^2 I),$$

$$\text{데이터} \begin{pmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \cdots & & \\ y_2 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$\underline{y}$: 종속변수 벡터(차수 $n\times 1$), X: 데이터 행렬(차수 $n\times(p+1)$)

$\underline{\beta}$: 회귀계수 벡터(차수 $(p+1)\times 1$), $\underline{e}$: 오차 벡터(차수 $n\times 1$)

이로부터 $E(\underline{y}) = \underline{X\beta}$, $V(\underline{y}) = \sigma^2 I_n$ 이고 종속변수는 정규분포를 따른다.

4.2 산점도 행렬

단순회귀의 시작은 종속변수와 설명변수의 산점도를 그려 (1)변수간의 함수 관계 (2)이상치나 영향치 존재 여부를 판단할 수 있다. 다중회귀분석의 시작도 산점도를 그리는 것이다. 종속변수와 설명변수간의 산점도(이는 단순회귀분석에서 산점도를 그리는 이유와 동일하다), 설명변수간의 산점도(다중공선성 문제 미리 진단)를 그린다. 설명변수가 p 개인 경우 산점도의 개수는 $_{p+1}C_2 = \frac{p(p+1)}{2}$ 이다. 이 산점도들을 행렬처럼 그려 놓은 것을 산점도 행렬(scatter plot matrix)이라 한다.

EXAMPLE 4-1 산점도 행렬 그리기

학생의 성별, IQ(3종류, FSIQ(Full scale IQ), VIQ(Verbal, 언어), PIQ(Performance, 수행능력)와 신체조건(키, 몸무게)과 두뇌의 크기(MRI 개수)을 조사하였다.(MRI_IQ.xls) 종속변수를 FSIQ라 하고 설명변수를 VIQ, PIQ, HEIGHT, WEIGHT, MRI라 하자. ($n=38$) **MRI_IQ.xls**

	A	B	C	D	E	F	G
1	Gender	FSIQ	VIQ	PIQ	Weight	Height	MRI
2	Female	133	132	124	118	65	816932
3	Male	139	123	150	143	73	1038437
4	Male	133	129	128	172	69	965353
5	Female	137	132	134	147	65	951545

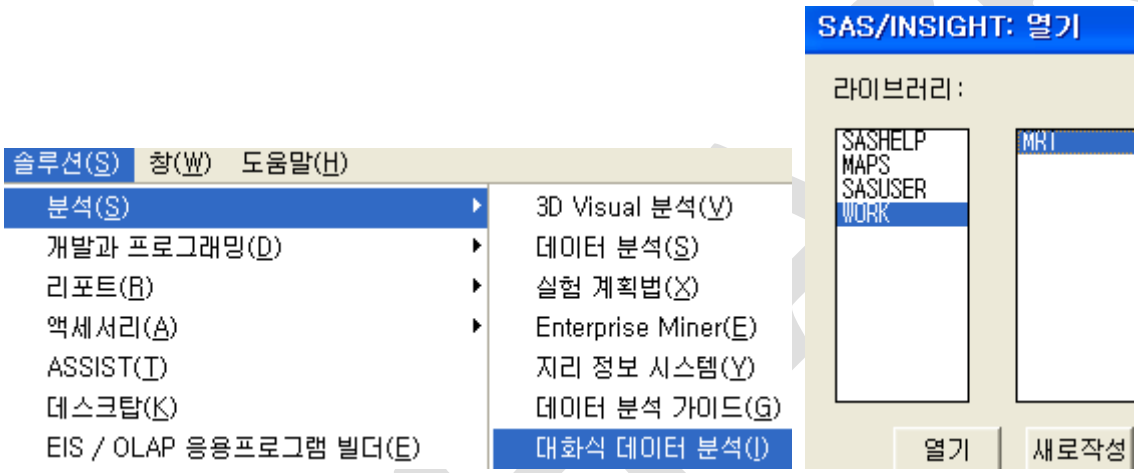
SAS에서 산점도 행렬을 그리려면 SAS/INSIGHT가 있어야 한다. 이 제품에 대한 라이센스가 없으면 아래 에러 메시지가 LOG 윈도우에 출력된다. 해결방법은? 여러 개의 산점도를 그릴 수 밖에 없다.

```
ERROR: The SAS/INSIGHT product with which Interactive Data Analysis is associated
```

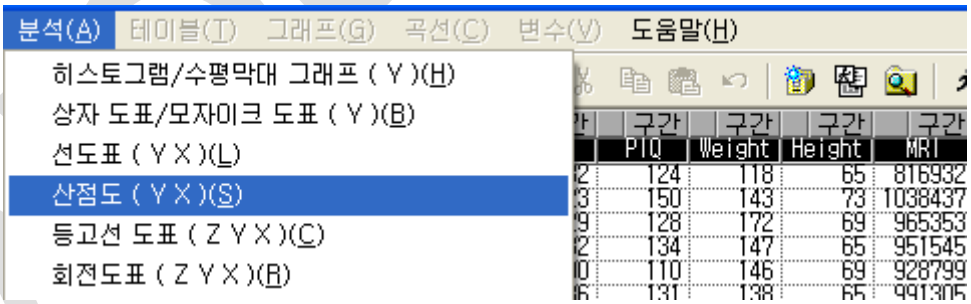
그러나... 다소 어려운 점이 있다. 아래 프로그램을 실행하면 종속변수와 설명변수들의 산점도가 출력되고 각각 하나씩 출력된다. T.T

```
proc reg data=mri;
    model fsiq=viq piq weight height mri/r;
    plot fsiq*(viq piq weight height mri);
run;
```

MRI_IQ.xls 데이터를 SAS 데이터 MRI로 만든 후 메뉴에서 다음 절차를 밟아 대화식 데이터 분석(SAS/INSIGHT)을 실시하면 된다.



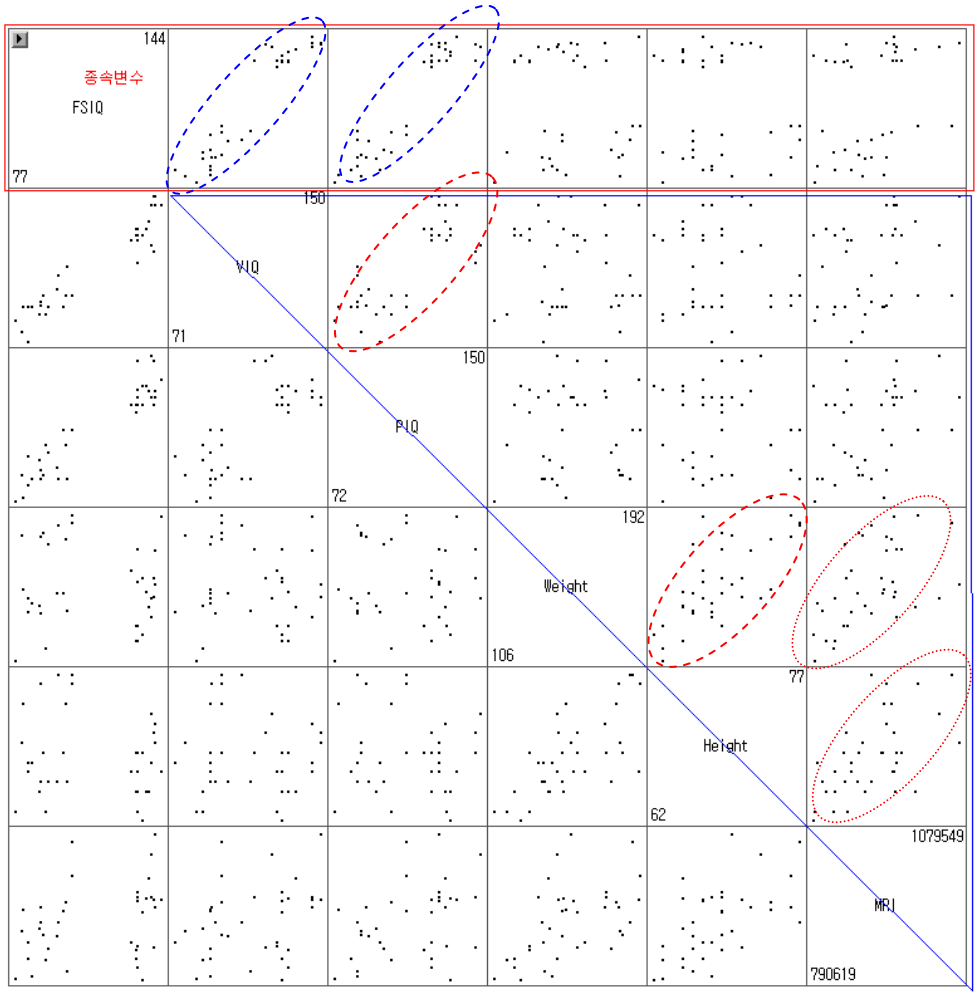
산점도 그릴 변수를 선택한 후(CRTL을 누른 상태에서 마우스로 선택한다) 분석 메뉴에서 산점도 옵션을 선택하면 된다.



우 상단 삼각형과 죄 하단 삼각형은 동일한 산점도이므로 상단 산점도만 해석하면 된다. 빨간 박스 안은(가능하면 종속변수가 제일 위에 올라 오게 데이터를 만드는 것이 유리하다) 종속변수와 설명변수들간에 산점도이므로 종속변수에 유의한 영향(직선 관계가 존재)을 미치는 설명변수를 예상할 수 있다. VIQ와 PIQ는 FSIQ에 양의 영향을 미침을(상관계수 양, 회귀계수 부호 양, 파란 타원) 알 수 있다. 영향치나 이상치는 존재하는 것 같지 않다.

파란 역 삼각형 부분은 설명변수들간의 (함수) 관계를 나타내는 산점도이므로 변수들간의

직선(상관) 관계가 존재하면 다중공선성 문제(상관 관계가 높은 설명변수들에 의해 종속변수를 설명하는 부분이 겹친다. 이로 인하여 회귀계수 추정치의 분산이 커진다)가 발생하므로 미리 주의해야 한다. 이에 대해서는 나중에 자세히 다루기로 한다. (PIQ, VIQ), (키, 몸무게, 뇌의 크기) 사이에는 양의 상관 관계가 존재함을 알 수 있다.

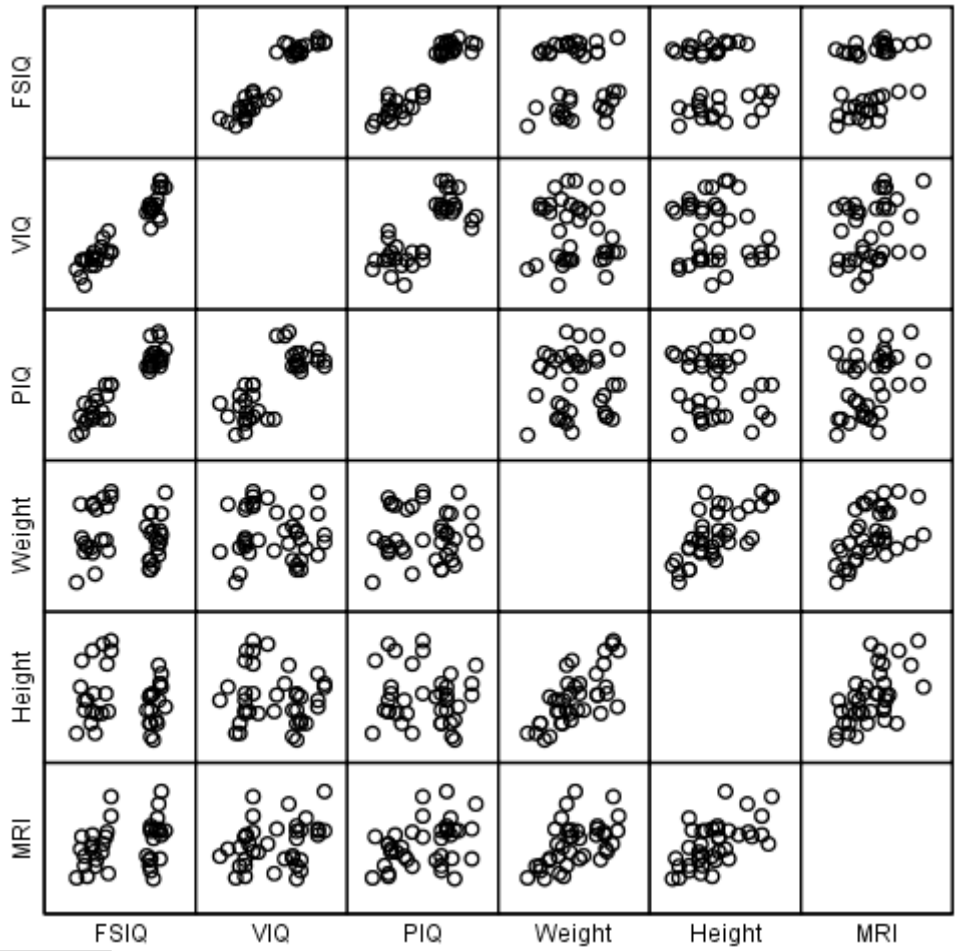


SPSS에서 산점도 행렬을 그리려면 다음 방법을 사용하면 된다.

Gender	FSIQ	VIQ	PIQ	Weight	Height	MRI
Female	133	132	124	118	65	816932
Male	139	123	150	143	73	1038437
Male	133	129	128	172	69	965353

SPSS 그래프(G) ▶ 산점도(S)... 메뉴를 선택한 후 아래와 같이 설정하면 된다.

종속변수와 설명변수를 행렬변수에 넣는다. 종속변수를 제일 위에 오게 하는 정도의 "센스"는 갖자. 그래야 종속변수와 설명변수들의 산점도가 가자 위에 나오게 되고



HOMEWORK #6-1 DUE 4월 13일(수)

CARS.txt (텍스트 데이터)

연비에 영향을 주는 변수로 자동차 무게, 운전 비율, 마력, 배기량, 실린더 수를 고려하였다. 우선 산점도 행렬을 먼저 그리고 해석하시오. 데이터 CARS.txt

①Country: 제조 국가(U.S., Japan) ②Car: 자동차 이름

③MPG: Miles per gallon(연비) ④Weight: 자동차 무게

⑤Drive_Ratio: Lead-screw(회전 운동을 직선운동으로 바꿈) 회전당 모터 회전 비율

⑥Horsepower: 마력 ⑦Displacement: 배기량 ⑧Cylinder: 실린더 수, (지시 변수)

4.3 추론 및 분산 분석

4.3.1 회귀계수에 대한 OLS 추정

OLS 추정치는 오차항의 제곱합($\min_{\beta_0,\beta_1,..\beta_p} \sum_{i=1}^{n} e_i^2$)을 최소화 하는 추정치이다. 그러므로 다중회귀모형으로부터 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i$, 오차의 제곱합을 최소화하는 추정치가 OLS 추정치이다. 즉 $\min_{\beta_0.\beta_1,...,\beta_p} \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - ... - \beta_p x_{pi})^2$ 으로부터 회귀계수 $(\beta_0, \beta_1, \beta_2, ..., \beta_p)$ 의 OLS 추정치를 구하려면 각 회귀계수에 의해 편미분을 하여 0 이라 놓은 후 $(p+1)$ 개의 방정식을 풀어야 한다.

행렬에 의해 OLS 추정치를 구해보자. 회귀모형 $\underline{y} = X\underline{\beta} + \underline{e}$ 로부터 다음 식에 의해 회귀계수 벡터를 추정하면 된다.

$$\min_{\alpha,\beta_1,\beta_2,...\beta_p} \sum_{i=1}^{n} e_i^2 = \min_{\underline{\beta}} \underline{e}' \underline{e} = \min_{\underline{\beta}} (\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}) \rightarrow \text{OLS 추정치 } \hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$$

Gauss Markov Theorem에 의해 $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$ 는 BLUE(Best Linear Unbiased Estimator) 이고 $E(\hat{\underline{\beta}}) = \underline{\beta}$ 이고 $V(\hat{\underline{\beta}}) = \sigma^2(X'X)^{-1}$ 이다. $V(\hat{\underline{\beta}})$ 의 추정치는 $s^2(\hat{\underline{\beta}}) = MSE(X'X)^{-1}$ 이다.

About Matrix

차수가 $n \times p$ 인 행렬 $X_{n\times p}$ (matrix X of order $n \times p$)라 부른다. i 는 행을, j 는 열을 나타내며 행렬의 간편 기호는 $X_{n\times p} = \{x_{ij}\}$ 이다. x_{ij} 를 원소(element)라 한다.

$$X_{n\times p} = \begin{bmatrix} x_{11}\, x_{12} \cdots x_{1p} \\ x_{21}\, x_{22} \cdots x_{2p} \\ \cdots\cdots x_{ij} \cdots\cdots \\ x_{n1}\, x_{n2} \cdots x_{np} \end{bmatrix}$$

정방행렬: 행과 열의 차수가 같은 행렬(즉, $n = p$) $A_{3\times3} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 2 & 3 \\ 2 & 1 & 1 \end{bmatrix}$

대각행렬: 정방 행렬에서 대각선에 위치한 원소를 대각 원소(diagonal element)라 하며 대각 원소를 제외한 모든 원소가 0인 행렬

대각합: 정방 행렬의 대각 원소의 합을 대각합(trace)이라 하고 $tr(A) = \sum_{i=1}^{n} A_{ii}$ 로 정의한다.

항등행렬: 정방 행렬 중 대각 원소가 모두1이고 다른 원소는 모두 0인 행렬을 항등 행렬(Identity Matrix)라 하고 I_n라 표시한다. 항등 행렬은 선형대수(Linear Algebra)의 곱에서 1의 역할과 동일하다. 행렬대수(matrix algebra)의 역수의 개념은 역행렬(inverse matrix)이며 정방 행렬 A에 대해 $AA^{-1} = A^{-1}A = I$ 가 성립하는 A^{-1}을 역행렬이라 한다.

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

동일: (1)차수가 동일하고 (2)대응 원소가 같으면 두 행렬은 동일(equal)하다고 한다. 즉 $A = B$ 이면 $a_{ij} = b_{ij}$, $for\ \ all\ \ i, j$ 이다.

II EXAMPLE II $A = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 2 & 2 \\ 3 & -1 & 1 \end{bmatrix}$ 인 경우 $A = B$ 이나 $A \neq C$ 이다.

전치: 행의 원소를 열로 보내고 열의 원소를 행으로 보내어 만들어진 행렬을 전치 행렬이라 하고 이 과정을 전치(transpose)라 하다. 행렬 $X_{n\times p}$ 의 전치 행렬은 $X'_{p\times n}$ 이고 차수는 ($p\times n$)이다. 이를 간편 기호로 나타내면 다음과 같다. $X' = \{x_{ji}\}$

II EXAMPLE II $X_{4\times3} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 2 & 3 \\ 2 & 1 & 1 \\ 5 & 2 & 2 \end{bmatrix}$ 의 전치 행렬 $X'_{3\times4}$을 구하면 $X'_{3\times4} = \begin{bmatrix} 2 & 4 & 2 & 5 \\ 3 & 2 & 1 & 2 \\ 2 & 3 & 1 & 2 \end{bmatrix}$

전치 성질: ① $(A')' = A$ ② $A + B) = A' + B'$ ③ $AB)' = B'A'$

대칭행렬: 행렬과 전치행렬이 동일한 행렬, 즉 $A = A'$ ($\Leftrightarrow \{a_{ij}\} = \{a_{ji}\}$)인 경우 행렬 A 을 대칭행렬(Systematic Matrix)이라 한다. 대칭행렬이 되려면 반드시 정방행렬이어야 한다.

합 연산: 행렬의 합을 구하는 경우 두 행렬의 차수는 동일해야 하며(conformable for addition: 합 연산 적합) 각 행렬에서 대응하는 원소들의 합을 그 위치에 적으면 된다.

합 성질: ① $tr(A + B) = tr(A) + tr(B)$ ②결합법칙(associate law): $(A + B) + C = A + (B + C)$

(행렬)x(행렬): 앞 행렬의 열의 차수와 뒤 행렬의 행의 차수가 동일해야 행렬의 곱이 성립하며, 결과는 앞 행렬의 행의 차수와 뒤 행렬의 열의 차수가 된다.

$$\text{행렬 } A_{n\times p}=\begin{bmatrix} a_{11}\ a_{12}\ \dots\ a_{1p} \\ a_{21}\ a_{22}\ \dots\ a_{2p} \\ \dots\dots\dots a_{ij}\dots\dots\dots \\ a_{n1}\ a_{n2}\ \dots\ a_{np} \end{bmatrix},\quad B_{p\times q}=\begin{bmatrix} b_{11}\ b_{12}\ \dots\ b_{1q} \\ b_{21}\ b_{22}\ \dots\ b_{2q} \\ \dots\dots\dots b_{ij}\dots\dots\dots \\ b_{p1}\ b_{p2}\ \dots\ b_{pq} \end{bmatrix} \text{ 이라면}$$

$$AB_{n\times q}=\begin{bmatrix} a_{11}b_{11}+a_{12}b_{21}+\cdots+a_{1p}b_{p1} & \cdots & a_{11}b_{1q}+a_{12}b_{2q}+\cdots+a_{1p}b_{pq} \\ \vdots & \vdots & \vdots \\ a_{n1}b_{11}+a_{n2}b_{21}+\cdots+a_{np}b_{p1} & \cdots & a_{n1}b_{1q}+a_{n2}b_{2q}+\cdots+a_{np}b_{pq} \end{bmatrix}$$

곱 성질: $A_{n\times p}$, $B_{p\times q}$

(1) BA의 연산이 가능하더라도 일반적으로 $AB\neq BA$이다.

(2) $(AB)'=B'A'$ 이 성립한다. (단 곱의 연산이 적합한 경우 가능하다)

(3)A, B가 대칭 행렬이면 $(AB)'=B'A'=BA$

(4) $tr(AB)=tr(BA)$ 단, AB가 정방 행렬일 때만 성립한다.

(5)결합 법칙(Associate law): $(A+B)+C=A+(B+C)$, $(AB)C=A(BC)=ABC$

(6)배분 법칙(Distribution law): $A(B+C)=AB+AC$

(7)교환 법칙(Communication law): $(A+B)=(B+A)$

멱등행렬: $M^2=MM=M$ 이면 행렬 M은 멱등행렬(Idempotent matrix)이다. M 이 멱등행렬이면 $M^k=M$ (k는 양의 정수)이 성립한다.

직교행렬: $AA'=A'A=I$ 이면 행렬 A는 직교 행렬(Orthogonal matrix)이라 한다.

행렬식

차수가 2일 경우: 행렬 $A=\begin{bmatrix}7 & 3\\ 4 & 6\end{bmatrix}$의 행렬식(Determinant)은 $|A|=7\times6-3\times4=30$ (scalar이다.) 차수가 3일 경우: $A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 7\\ 8 & 9 & 10\end{bmatrix}$의 행렬식은

$$|A|=1(-1)^{1+1}\begin{vmatrix}5 & 6\\ 7 & 10\end{vmatrix}+2(-1)^{1+2}\begin{vmatrix}4 & 7\\ 8 & 10\end{vmatrix}+3(-1)^{1+3}\begin{vmatrix}4 & 5\\ 8 & 9\end{vmatrix}=-3 \quad \text{혹은}$$

$$|A|=4(-1)^{2+1}\begin{vmatrix}2 & 3\\ 9 & 10\end{vmatrix}+5(-1)^{2+2}\begin{vmatrix}1 & 3\\ 8 & 10\end{vmatrix}+7(-1)^{2+3}\begin{vmatrix}1 & 2\\ 8 & 9\end{vmatrix}=-3$$ (2번째 행을 이용) 혹은

$|A| = 1(-1)^{1+1}\begin{vmatrix}5 & 7\\ 9 & 10\end{vmatrix} + 4(-1)^{2+1}\begin{vmatrix}2 & 3\\ 9 & 10\end{vmatrix} + 8(-1)^{3+1}\begin{vmatrix}2 & 3\\ 5 & 7\end{vmatrix} = -3$ (1번째 열 이용) 모두 같은 값이다.

이를 확장하면 차수 n의 행렬의 행렬식은 $|A| = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}|$ 이다.

$|M_{ij}|$ 를 minor라 하고 $(-1)^{i+j}|M_{ij}|$ 를 cofactor라 한다.

행렬식 성질

(1) $|A'| = |A|$, $|AB| = |A||B|$, $|AB| = |BA|$

(2)행렬 A의 두 행이 같으면 행렬식은 0이다.

(3)한 행(열)의 상수를 곱하여 다른 행에 더해도 행렬식 값은 변하지 않는다.

(4)한 행(열)을 다른 행들의 선형결합으로 표현할 수 있으면 행렬식의 값은 0이다. (예: 다중공선성)

역행렬: 정방행렬 A에서 $AB = BA = I$ 를 만족하는 행렬 B를 A의 역행렬이라 하고 A^{-1} 로 나타낸다. $A^{-1} = \frac{1}{|A|}adjA = \frac{1}{|A|}$ [A 원소를 cofactor로 대치]'

역행렬 성질

(1)역행렬은 unique하다.

(2) $|A^{-1}| = 1/|A|$, $(A^{-1})^{-1} = A$, $(A')^{-1} = (A^{-1})'$, $(AB)^{-1} = B^{-1}A^{-1}$

정의(LIN: linearly independent vector): $a_1\underline{x}_1 + a_2\underline{x}_2 + \ldots + a_p\underline{x}_p = \underline{0}$ 가 모든 $a_i = 0$ 일 때만 만족한다면 벡터 $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ 는 선형 독립(linearly independent) 벡터라 하고, 0이 아닌 a_i 에 대해서 만족한다면 선형 종속(linearly dependent)인 벡터라 한다. 상호 종속인 벡터는 하나의 벡터를 다른 벡터들의 선형 결합으로 표시할 수 있다는 것을 의미한다.

정의(full rank): (nxn)정방 행렬에서 선형 독립인 행(열)의 개수($rank(A)$)가 행렬의 차수 n와 같다면 이 행렬은 full-rank 행렬이라 한다. 즉 $rank(A_{n\times n}) = n$ 이면 full-rank이다.

▪ 역행렬이 존재한다.	▪ 역행렬이 존재하지 않는다
▪ full-rank 이다. rank(A)=n	▪ full-rank 아니다. rank(A)<n
▪ A 는 non-singular 이다.	▪ A 는 singular 이다.
▪ $\|A\|\neq 0$	▪ $\|A\|=0$
▪ Ax=b 의 해가 존재한다.	▪ Ax=b 의 해가 존재하지 않는다.

행렬 미분

상수 벡터 $\underline{a}=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$, 확률 변수 벡터 $\underline{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ 라 하면

(1) $\frac{\partial}{\partial \underline{x}}(\underline{a}'\underline{x})=\underline{a}$ (2) $\frac{\partial}{\partial \underline{x}}(\underline{x}'\underline{a})=\underline{a}$

(3) $\frac{\partial}{\partial \underline{x}}(\underline{x}'A\underline{x})=A\underline{x}+A'\underline{x}$ (A는 정방 행렬) ➔ 만약 A가 대칭 행렬이면 $\frac{\partial}{\partial \underline{x}}(\underline{x}'A\underline{x})=2A\underline{x}$

END of Matrix

HOMEWORK #6-2 DUE 4월 13일(수)

① $\sum_{i=1}^{n} e_i^2=\underline{e}'\underline{e}=(\underline{Y}-X\underline{\beta})'(\underline{Y}-X\underline{\beta})$ 임을 보이시오.

② $X'X$ 가 대칭행렬임을 보이시오.

③ $(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})=\underline{y}'\underline{y}-\underline{y}'X\underline{\beta}-\underline{\beta}'X'\underline{y}+\underline{\beta}'X'X\underline{\beta}$ 임을 보이시오.

④OLS 추정치 $\hat{\underline{\beta}}=(X'X)^{-1}X'\underline{y}$ 을 구하시오.

⑤ $V(\underline{a}'X)=\underline{a}V(X)\underline{a}'$ 을 이용하여 $E(\hat{\underline{\beta}})=\underline{\beta}$, $V(\hat{\underline{\beta}})=\sigma^2(X'X)^{-1}$ 을 증명하시오.

⑥ $H=X(X'X)^{-1}X'$ 을 HAT 행렬이라 정의한다. H, $(I-H)$ 가 멱등행렬임을 보이시오.

⑦예측치 $\hat{\underline{y}}=H\underline{y}$ 이고 잔차 $\underline{r}=\hat{\underline{e}}=(I-H)y$ 임을 보이시오.

4.3.2 예측치와 잔차

추정치 $\hat{\underline{\beta}}$ 가 구해지면 종속변수에 대한 예측치는 $\hat{\underline{y}} = X\hat{\underline{\beta}} = X(X'X)^{-1}X'\underline{y} = H\underline{y}$ (H 는 Hat 행렬이라 부른다) 이고 잔차 벡터는 $\hat{\underline{e}} = \underline{r} = \underline{y} - \hat{\underline{y}} = (I-H)\underline{y}$ 이다. 잔차 벡터에 대해 $E(\underline{r}) = \underline{0}$ 이고 $V(\underline{r}) = \sigma^2(I-H)$ 이다. $V(\underline{r})$ 의 추정치는 $MSE(I-H)$ 이다.

HOMEWORK #6-3 DUE 4월 13일(수)

개인 프로젝트(term paper)에 대한 1 페이지 개요 제출하기

프로젝트 분석 목적, 내용, 데이터(변수) 측정(수집) 방법

4.3.3 모형에 대한 추론(분산분석)

총변동($SST = \sum(y_i - \bar{y})^2$), 오차변동($SSE = \sum(y_i - \hat{y}_i)^2$), 회귀변동($SSR = \sum(\hat{y}_i - \bar{y})^2$)를 행렬로 표시하고 분산분석표(ANOVA)를 작성하면 다음과 같다.

변동 (source)	SS(자승합)	자유도	MS (평균자승합)	F-검정
Regression (모형)	$SSR = \underline{y}'[H - (\frac{1}{n})J]$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE} \sim F(p, n-1)$
Error (오차)	$SSE = \underline{y}'[I-H]\underline{y}$	$n-p-1$	$MSE = \frac{SSE}{n-2}$	
Total (총 변동)	$SST = \underline{y}'[I - (\frac{1}{n})J]$	$n-1$	결정계수: $R^2 = \frac{SSR}{SST}$	

$SST = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \underline{y}'\underline{y} - \frac{1}{n}\underline{y}'J\underline{y}$, J 는 모든 원소가 1인 정방행렬

$$SSE = \underline{e}'\underline{e} = (\underline{y} - \hat{\underline{\beta}}X)'(\underline{y} - \hat{\underline{\beta}}X) = \underline{y}'\underline{y} - \hat{\underline{\beta}}'X\underline{y}$$

$$SSR = SST - SSE = \hat{\underline{\beta}}'X\underline{y} - \frac{1}{n}X'JX$$

F-검정

다중회귀모형에서 F-검정은 다음 귀무가설을 검정한다.

○귀무가설 $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ (모든 설명 변수는 유의하지 않다)

○대립가설 $H_a : all\ \beta_i \neq 0$ (유의한 설명 변수가 적어도 하나는 있다.)

그러므로 F-검정 결과 귀무가설이 기각되면 유의한 설명 변수가 하나 이상 있다는 것이므로 각 설명 변수에 대한 유의성을 t-검정을 이용하여 알아 보면 된다.

결정계수(coefficient of multiple determination)

결정 계수는 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 에 의해 정의되므로 어떠한 설명 변수를(비록 유의하지 않더라도) 추가하더라도(SST는 일정하므로) 항상 증가한다. 이를 보완하기 위하여 수정된 (adjusted) 결정 계수를 구하게 된다. 수정된 결정계수, $R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$ 는 결정계수가 항상 증가하는 문제는 해결할 수 있으나 결정계수 분포를 알 수 없으므로(단순회귀의 경우 결정계수의 제곱근은 상관계수와 동일하므로 상관계수 유의성 검정에 의해 유의성 판단이 가능하다.) 설명 변수의 유의성을 판단하는 검정 통계량으로 사용되지 않는다.

4.3.4 회귀모형에 대한 t-검정

앞 절에서 OLS $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$ (정규분포)이고 $E(\underline{\hat{\beta}}) = \underline{\beta}$, $\hat{V}(\underline{\hat{\beta}}) = s^2(\underline{\hat{\beta}}) = MSE(X'X)^{-1}$ 임을 알았다. 이 사실을 이용하여 각 회귀계수의 유의성(설명변수의 유의성)을 다음에 의해 검정할 수 있다.

○귀무가설 $H_0 : \beta_k = 0$ (설명 변수 X_k 는 유의하지 않다)

○대립가설 $H_a : \beta_k \neq 0$ (설명 변수 X_k 는 유의하다)

○검정통계량 $T = \frac{\hat{\beta}_k - \beta_k}{s(\hat{\beta}_k)} \sim t(n-p-1)$

4.3.5 종속변수 평균 및 종속변수 예측치 추론

설명변수들의 관측치 $\underline{x}_h' = \begin{pmatrix} 1 & x_{1h} & x_{2h} \dots & x_{ph} \end{pmatrix}$ 가 주어지면 종속변수 평균은 $E(y_h) = x_h'\underline{\beta}$

이므로 추정치는 $\hat{E(y_h)} = x_h^{'}\hat{\underline{\beta}}$ 이고 추정 분산은 $s^2(\hat{E(y_h)}) = MSE\underline{x}_h^{'}(X'X)^{-1}\underline{x}_h$ 이다. 그러므로 평균에 대한 100(1-α)%신뢰구간은

$$x_h^{'}\hat{\underline{\beta}} \pm t(1-\alpha/2; n-p-1)MSE\underline{x}_h^{'}(X'X)^{-1}\underline{x}_h$$

설명변수들의 관측치 $\underline{x}_h^{'} = \begin{pmatrix}1 & x_{h1} & x_{h2} \cdots & x_{hp}\end{pmatrix}$ 가 주어지면 새로운 관측치에 대한 추정치는 $\hat{y_{new}} = x_h^{'}\hat{\underline{\beta}}$ 이고 추정 분산은 $s^2(\hat{y_{new}}) = MSE(1+\underline{x}_h^{'}(X'X)^{-1}\underline{x}_h)$ 이다. 그러므로 새로운 관측치에 대한 100(1-α)%신뢰구간은

$$x_h^{'}\hat{\underline{\beta}} \pm t(1-\alpha/2; n-p-1)MSE(1+\underline{x}_h^{'}(X'X)^{-1}\underline{x}_h).$$

4.3.6 분석 예제

EXAMPLE 4-2 산점도 행렬 그리기

종속변수를 FSIQ(Full scale IQ), 설명변수를 VIQ(Verbal, 언어) 와 두뇌의 크기(MRI 개수)로 하여 다중회귀분석을 실시해 보자. $FSIQ_i = \beta_0 + \beta_1 VIQ_i + \beta_2 MRI_i + e_i$

다중회귀분석은 단순과 달리 우선 각 설명변수의 유의성을 점검한 후 유의한 설명변수만으로 잔차분석과 이상치 진단을 실시하면 된다. 물론 이상치가 많이 존재하거나 산점도에서 특이한 사항이 존재하면 유의성 검정 전에 해결하는 것이 적절하다.

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

F-검정 결과 유의하므로 귀무가설 $H_0 : \beta_1 = \beta_2 = 0$ 은 기각되므로 (**VIQ, MRI**) 중 적어도 하나는 FSIQ에 영향을 준다(유의하다). 결정계수는 0.896이다.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	18802	9401.24198	150.74	<.0001
Error	35	2182.91078	62.36888		
Corrected Total	37	20985			

Root MSE	7.89740	R-Square	0.8960

t-검정 결과 유의수준 0.05에서 유의한 설명 변수는 VIQ이다. 만약 두 설명변수가 모두 유의하였다면 (임시적) 추정회귀모형은 $FSIQ_i = -10.88 + 0.96VIQ_i + 0.000018MRI_i$ 이다.

Parameter Estimates

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

유의할 것 같은 PIQ를 사용하지 않은 것은 설명변수 VIQ와 상관관계가 높아 다중공선성 문제가 발생할 것 같아 제외하였다. 물론 산점도만 보고 제외하면 안되고 다중공선성을 판단하는 VIF나 상태지수를 이용해야 하지만 여기서는 예제이므로 MRI를 사용하였다.

유의하지 않은 설명변수는 유의하지 않은 순서대로(유의확률이 큰 순서대로) 하나씩 제외하면 된다. MRI를 제외하자.

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

설명변수가 모두 유의하면 잔차분석과 이상치 진단을 한다.

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

P, R, CLM 옵션에 의해 예측치, 잔차, 평균에 대한 신뢰구간이 출력된다.(단순회귀와 동일) 유의한 설명변수를 선택하면 다중공성선 문제 진단, 이상치 혹은 영향치 진단, 그 후에 잔차분석을 실시한다. 잔차분석에서는 단순회귀분석과 유사하나 다음과 같은 차이가 있다.

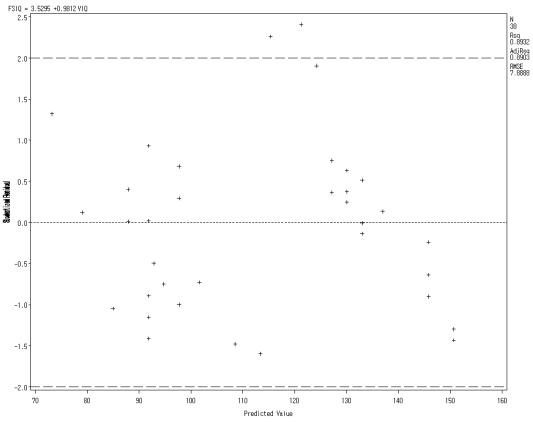
○스튜던트 잔차와 예측치의 산점도는 등분산성 체크에만 사용된다. 이에 대한 해결책으로 WLS 방법(가중치로 $1/\hat{y}_i^2$)이나 종속변수를 변환한다.

○잔차는 정규성 검정에 사용된다. 정규성 가정이 무너지면 그냥 뭉개거나(표본의 크기가 크면 정규성 가정이 무너져도 큰 문제가 되지 않는다. 중심극한정리와 유사) LOG 변환이나 제곱 변환을 실시한다.

○스튜던트 잔차와 설명변수의 산점도를 그려 설명변수로 인한 이분산성을 문제를 진단한다. 이에 대한 해결책으로는 설명변수를 변환시킨다.

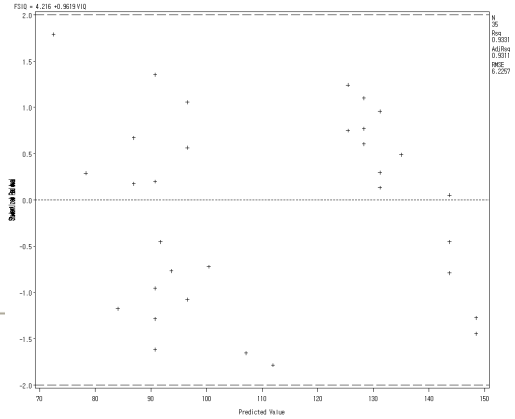
○이상치 진단은 스튜던트 잔차만으로 하는 것이 아니라 다양한 통계량을 이용한다.

이상치가 2개 존재한다. 출력 창에서 어느 관측치가 이상치인지 알아보았더니 9, 13번째 관측치이다.



```
proc reg data=mri;
     model fsiq=viq/r p clm;
     reweight obs.=9;
     reweight obs.=13;
     reweight obs.=2;
     plot student.*p./vref=-2 2;
     output out=out1 student=sres;
run;

proc univariate data=out1 normal plot;
run;
```



최종적으로 3개 관측치(9, 13, 2)가 이상치로 제외되었다. 산점도는 가운데가 비여 이상해 보이지만 랜덤으로 보기에 문제는 없고 정규성 가정(잔차가 아니라 스튜던트 잔차를 사용한 것은 REWEIGHT 문을 사용해고 이상치에 대한 잔차는 계산되어 있기 때문이다)도 만족한다. 가운데가 비어 있는 것은 종속변수 FSIQ 특성이다.

정규성 검정

검정		----통계량----		-------p-값-------
Shapiro-Wilk	W	0.939844	Pr < W	0.0553
Kolmogorov-Smirnov	D	0.123896	Pr > D	>0.1500

```
줄기 잎                #       상자그림
   1 8                 1          |
   1 0111124           7          |
   0 5667888           7       +-----+
   0 112233            6       +-----+
  -0                           |  +  |
  -0 88755             5       |     |
  -1 433210            6       +-----+
  -1 876               3          |
     ----+----+----+----+
```

4.4 추가자승합

4.3절에서 다중 회귀 모형의 유의성 검정은 F -검정($H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ 모든 설명 변수는 유의하지 않다.)을 이용하였고 각 회귀계수(설명변수)에 대한 유의성 검정은 t-검정($H_0 : \beta_k = 0$)을 이용하였다. 귀무가설 $H_0 : \beta_i = \beta_j = 0, i \neq j$ (두 설명변수는 모두 유의하지 않다) 혹은 두 설명변수의 회귀계수는 동일하다는 $H_0 : \beta_i = \beta_j, i \neq j$ 가설을 검정하려면 추가 자승합(ESS, Extra sum of Square) 개념이 필요하다. 고려한 모든 변수가 있는 모형(이를 Full 모형)의 SSE(혹은 SSR)와 귀무가설 하에서 모형(이를 Reduced 모형)의 SSE(SSR)을 비교하여 귀무가설의 유의성을 검정한다.

4.4.1 기본 개념

EXAMPLE 4-3 추가 자승합 개념

MRI_IQ.xls 예제 데이터를 이용하여 추가 자승합 개념을 살펴보자.

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2337.03506	2337.03506	4.51	0.0406
Error	36	18648	518.00999		
Corrected Total	37	20985			

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	18802	9401.24198	150.74	<.0001
Error	35	2182.91078	62.36888		
Corrected Total	37	20985			

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20890	6963.17984	2469.85	<.0001
Error	34	95.85522	2.81927		
Corrected Total	37	20985			

설명 변수 VIQ가 들어감으로써 SSR이 2337.03에서 18802로 증가하였다. 그러므로 SSE는 18648에서 2182.91로 동일한 값만큼 감소하였다. 즉, 총변동 $SST = \sum(y_i - \bar{y}_i)^2$)는 동일하다. 설명 변수가 추가됨으로써 증가된 SSR의 한계 증가량 기호를 $SSR(\text{기존}|\text{추가})$ 이라 하자.

$SSR(VIQ|MRI) = SSR(VIQ, MRI) - SSR(MRI) = 18802 - 2337.04 = 16465$ ➔"FSIQ=VIQ"에 설명변수 MRI을 삽입하였을 때 증가되는 한계(marginal) SSR(설명력)이다.

$SSE(VIQ|MRI) = SSE(MRI) - SSE(VIQ, MRI)$ ➔ "FSIQ=VIQ"에 설명변수 MRI을 삽입하였을 때 감소되는 한계(marginal) SSE(설명되지 않는)이다.

같은 이유로 $SSR(PIQ \mid VIQ, MRI) = SSR(PIQ, VIQ, MRI) - SSR(VIQ, MRI) = 20890 - 18802 = 2088$ 이다. 이것 역시 $SSE(PIQ \mid VIQ, MRI) = SSE(VIQ, MRI) - SSR(PIQ, VIQ, MRI)$ 표시할 수 있다.

정의

$$SSR(X_1 \mid X_2) = SSR(X_1, X_2) - SSR(X_2) = SSE(X_2) - SSE(X_1, X_2)$$
$$SSR(X_1 \mid X_2, X_3) = SSR(X_1, X_2, X_3) - SSR(X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3)$$
$$SSR(X_1, X_2 \mid X_3) = SSR(X_1, X_2, X_3) - SSR(X_3) = SSE(X_3) - SSE(X_1, X_2, X_3)$$

4.4.2 SSR을 ESS 분해

$$\begin{aligned} SST &= SSE(X_1, X_2, ..., X_p) + SSR(X_1, X_2, ..., X_p) \\ &= SSE(X_1, X_2, ..., X_p) + SSR(X_1) + SSR(X_2 \mid X_1) \\ &+ SSR(X_3 \mid X_1, X_2) + ... + SSR(X_p \mid X_1, X_2, ..., X_{p-1}) \end{aligned}$$

EXAMPLE 4-4 추가 자승합 개념(2)

MRI_IQ.xls 데이터에서 VIQ, PIQ만 설명변수로 사용했을 경우 SSR를 분해해 보자.

<table>
<tr><td>$SSR(VIQ)$
18745</td><td rowspan="2">$SSR(VIQ, MRI)$
18802</td><td>$SSR(MRI)$
2337</td><td rowspan="2">SST
20985</td></tr>
<tr><td rowspan="2">$SSR(VIQ \mid MRI)$
16465</td></tr>
<tr><td>$SSR(MRI \mid VIQ)$
57</td><td></td></tr>
<tr><td>SSE
2183</td><td>SSE
2183</td><td>SSE
2183</td><td></td></tr>
</table>

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18745	18745	301.21	<.0001
Error	36	2240.37920	62.23276		
Corrected Total	37	20985			

MRI 변수가 혼자 단독으로 설명하는 설명력은 $SSR(MRI) = 2337$ 이나 되지만 VIQ 변수가 설명하고 남은 부분에 대한 설명력은 $SSR(MRI \mid VIQ) = 57$ 밖에는 되지 않는다. 그러므로 MRI 하나만 설명변수인 경우에는 지적 수준($FSIQ$)을 설명하는 것이 유의하지만 VIQ 변수가 추가된 상태에서는 MRI의 설명력은 유의하지 않다(아래 결과, p-값은 0.3437).

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

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.23792	46.90291	0.30	0.7632
MRI	1	0.00010953	0.00005157	2.12	0.0406

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

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

4.4.3 분산분석에서 SSR 분해

설명 변수가 3개인 경우 예를 들어 설명해보기로 하자.

변동	SS	df	MS
회귀	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3) = SSR(X_1, X_2, X_3)/3$
	$SSR(X_1 \mid \mu)$	1	$MSR(X_1 \mid \mu) = SSR(X_1 \mid \mu)$
	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1) = SSR(X_2 \mid X_1)$
	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2) = SSR(X_3 \mid X_1, X_2)$
오차	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3) = SSE(X_1, X_2, X_3)/(n-4)$
수정 총변동	SST	$n-1$	

ESS에 기호에 대한 몇 가지 예를 들어 보자.

$$SSR(X_3 \mid X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$
$$SSR(X_2, X_3 \mid X_1) = SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1)$$

SSR을 분해하는 방법은 두 가지 방법이 있다. 위와 같은 방법의 분해는 Sequential SS 이라 한다. 설명 변수를 하나씩 추가하면서 마지막 변수의 한계 설명력을 나타낸다. Partial SS는 다른 설명 변수들에 의해 설명되고 남은 부분에 대해 그 설명 변수가 설명하는 부분을 나타내는 것이다. 설명 변수가 $X_1, X_2, \ldots, X_n$ 인 다중 회귀에서 Type I SS, Type II SS를 표로 나타내면 다음과 같다.

모수	Sequential SS (Type I SS)	Partial SS (Type II SS)
X_1	$SSR(X_1 \mid \mu)$	$SSR(X_1 \mid \mu, X_1, X_2, \ldots, X_p)$
X_2	$SSR(X_2 \mid \mu, X_1)$	$SSR(X_2 \mid \mu, X_2, X_3, \ldots, X_p)$
X_3	$SSR(X_3 \mid \mu, X_1, X_2)$	$SSR(X_3 \mid \mu, X_1, X_2, \ldots, X_p)$
⋮	⋮	⋮
X_p	$SSR(X_p \mid \mu, X_1, X_2, \ldots, X_{p-1})$	$SSR(X_p \mid \mu, X_1, X_2, \ldots, X_{p-1})$

마지막으로 고려된 설명 변수에 대해서는 Sequential SS, Partial SS는 서로 동일하다.

EXAMPLE 4-5 Type I, III 자승합

MRI_IQ.xls 데이터에서 VIQ, PIQ, VIQ 세 개 사용했을 때 Sequential SS (Type I SS), Partial SS (Type II SS)를 구해보자.

```
PROC REG DATA=MRI;
     MODEL FSIQ=MRI VIQ PIQ/SS1 SS2;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20890	6963.17984	2469.85	<.0001
Error	34	95.85522	2.81927		
Corrected Total	37	20985			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-3.26299	3.48702	-0.94	0.3560	489980	2.46865
MRI	1	-0.00000889	0.00000411	-2.16	0.0375	2337.03506	13.21044
VIQ	1	0.57462	0.01908	30.12	<.0001	16465	2556.96484
PIQ	1	0.54290	0.01995	27.21	<.0001	2087.05556	2087.05556

각 설명 변수에 대한 유의성 검정은 각 회귀 계수의 t-검정이나 Partial SS(Type II)를 이용하면 된다. F-값(Type II SS)/MSE=(t-값)^{2}이므로 유의확률은 동일하다.

4.4.4 ESS를 이용하여 설명 변수 유의성 검정

Full model에서 Reduced Model로 줄어든 회귀계수에 대한 모수 개수를 m이라 할 때 다음 검정통계량에 의해 회귀계수에 대한 가설의 유의성을 검정한다.

SSR 이용: $T = \dfrac{(SSR_F - SSR_R)/m}{SSE_F/(n-p-1)} = \dfrac{(SSR_F - SSR_R)/m}{MSE_F} \sim F(m, n-p-1)$

SSE 이용: $T = \dfrac{(SSE_R - SSE_F)/m}{SSE_F/(n-p-1)} = \dfrac{(SSE_R - SSE_F)/m}{MSE_F} \sim F(m, n-p-1)$

▶귀무가설 $H_0 : \beta_k = 0$(설명 변수 X_k는 유의하지 않다)에 대한 검정

각 설명 변수 하나에 대한 유의성 검정은 ESS를 이용하기 보다는 t-검정을 이용하면 된다. 그러나 ESS를 이용한 유의성 검정 개념을 이해하기 위하여 살펴보기로 하자.

Full model: $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_p X_{pi} + e_i$

Reduced model: $y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_{(k-1)} X_{(k-1)i} + \beta_{(k+1)} X_{(k+1)i} .. + \beta_p X_{pi} + e_i$

줄어든 모수의 개수 $m = 1$

$$SSR(X_k \mid X_1, X_2, .., X_{k-1}, , X_{k+1}, .., X_p) = SSR(X_1, X_2, .., X_p) - SSR(X_1, X_2, .., X_{k-1}, , X_{k+1}, .., X_p)$$
$$= SSR_F - SSR_R$$
$$SSE(X_k \mid X_1, X_2, .., X_{k-1}, , X_{k+1}, .., X_p) = SSE(X_1, X_2, .., X_{k-1}, , X_{k+1}, .., X_p) - SSE(X_1, X_2, .., X_p)$$
$$= SSE_R - SSE_F$$

EXAMPLE 4-6 추가 자승합 개념

MRI_IQ.xls 데이터에서 VIQ, PIQ, VIQ 세 개 사용했을 때 MRI 설명변수 유의성 검정을 ESS 방법을 사용하여 실시해 보자.

회귀 모형 $FSIQ = \beta_0 + \beta_1 * MRI + \beta_2 * VIQ + \beta_3 * PIQ + e_i,\ for\ i = 1,2,..,38$, 설명변수 개수 $p = 3$

▣귀무가설: $H_0 : \beta_1 = 0$(MRI 설명 변수는 유의하지 않다)

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20890	6963.17984	2469.85	<.0001
Error	34	95.85522	2.81927		
Corrected Total	37	20985			

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	20876	10438	3349.69	<.0001
Error	35	109.06566	3.11616		
Corrected Total	37	20985			

검정통계량을 계산해 보자.

SSR 이용: $T = \dfrac{(SSR_F - SSR_R)/m}{SSE_F/(n-p-1)} = \dfrac{(20890-20876)/1}{2.82} = 4.96 \sim F(1,35)$ (반올림오차)

SSE 이용: $T = \dfrac{(SSE_R - SSE_F)/m}{SSE_F/(n-p-1)} = \dfrac{(109.07-95.86)/1}{2.82} = 4.68 \sim F(1,35)$

이 검정통계량은 $H_0 : \beta_1 = 0$ (설명변수 하나의 유의성 검정)에 대한 F-검정통계량이므로 페이지 107의 설명변수 MRI의 t-검정통계량 값과 동일하다 $\sqrt{4.68} = 2.16$

아이고 복잡하다. SAS를 이용하면 간단히 해결될 수 있는가?

```
PROC REG DATA=MRI;
     MODEL FSIQ=MRI VIQ PIQ;
     TEST MRI=0;
RUN;
```

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	13.21044	4.69	0.0375
Denominator	34	2.81927		

▶귀무가설 $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (설명변수 $X_1, X_2, ..., X_k$ 가 모두 유의하지 않다) 검정

EXAMPLE 4-7 Full model vs. Reduced Model

MRI_IQ.xls 데이터에서 VIQ, PIQ, VIQ 세 개 사용했을 때 (MRI, VIQ) 설명변수 군의 유의성 검정을 실시하시오.

○Full 회귀 모형 $FSIQ = \beta_0 + \beta_1 * MRI + \beta_2 * VIQ + \beta_3 * PIQ + e_i$

○귀무가설: $H_0 : \beta_1 = \beta_2 = 0$

○Reduce 회귀 모형 $FSIQ = \beta_0 + \beta_3 * PIQ + e_i$

줄어든 모수의 개수 $m = 2$ 이다.

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

$FullModel\_SSR$, $FullModel\_MSE$ 을 계산

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	20890	6963.17984	2469.85	<.0001
Error	34	95.85522	2.81927		
Corrected Total	37	20985			

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

$\mathrm{Re}\,ducedModel\_SSR$, $\mathrm{Re}\,ducedModel\_MSE$ 을 계산

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	18323	18323	247.81	<.0001
Error	36	2661.93397	73.94261		
Corrected Total	37	20985			

검정통계량을 계산해 보자.

SSR 이용: $T = \dfrac{(SSR_F - SSR_R)/m}{SSE_F/(n-p-1)} = \dfrac{(20890 - 18323)/2}{2.82} = 455.14 \sim F(2,34)$

SSE 이용: $T = \dfrac{(SSE_R - SSE_F)/m}{SSE_F/(n-p-1)} = \dfrac{(2661.9 - 95.86)/2}{2.82} = 455.14 \sim F(2,34)$

Test 문을 이용하여 실시 해보자.

```
PROC REG DATA=MRI;
     MODEL FSIQ=MRI VIQ PIQ;
     TEST MRI=0,VIQ=0;
RUN;
```

Test 1 Results for Dependent Variable FSIQ

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1283.03938	455.10	<.0001
Denominator	34	2.81927		

PIQ 설명 변수가 있는 경우 (MRI, VIQ) 설명 변수 군을 추가하여도 그들의 설명력은 유의하다.

SPSS TEST 문이 없으므로 Full model과 Reduced model의 SSR, SSE을 구하여 수작업 계산해야 한다.

HOMEWORK #7-1 DUE 4월 27일(수)

MRI_IQ.xls 데이터 (SPSS) 활용

Reduced 모형을 쓰고 Full 모형, Reduced 모형에 의해 귀무가설의 유의성을 검정하시오. 그리고 SAS의 경우 TEST 문을 이용하여 확인하시오.

① $FSIQ = \beta_0 + \beta_1 * MRI + \beta_2 * VIQ + \beta_3 * PIQ + e_i$ 에서 $H_0 : \beta_2 = \beta_3$ 유의성 검정을 실시하시오.

② $FSIQ = \beta_0 + \beta_1 * MRI + \beta_2 * VIQ + \beta_3 * PIQ + e_i$ 에서 $H_0 : \beta_2 = 0.5$ 유의성 검정을 실시하시오. (SSE 을 이용해야 한다)

추가자승합으로 회귀계수에 대한 가설 검정의 경우 SST 가 변동이 없다면 Full 모형과 Reduced 모형의 SSR(모형 변동) 차이에 의해 가설을 검정하면 된다. 그러나 SST(총변동)가 변하는 경우에는(예: $H_0 : \beta_2 = 0.5$) SSE 의 변동에 의해 가설을 검정해야 한다. 즉 SSE(F)-SSE(R)가 분자의 변동이 된다.

4.5 Coefficient of Partial Determination (부분 결정계수)

ESS는 다중 회귀모형에서 회귀계수들의 유의성(즉 설명 변수의 유의성) 검정에 사용할 수 있을 뿐 아니라 부분 결정계수라 불리는 변수들의 선형 관계 척도를 구하는데도 유용하다. 부분 결정계수는 단순 회귀 모형의 결정계수(R^2)와 동일하게 해석되며 0과 1 사이의 값을 갖는다. 부분 결정계수의 제곱근은 부분 상관 계수이다.

▶변수가 2개일 경우 $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

$SSE(X_2)$ 는 $y_i = \beta_0 + \beta_1 X_{1i} + e_i$ 의 오차 변동이고 $SSE(X_1, X_2)$ 는 $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$ 의 오차 변동이다. 설명 변수 X_2 가 이미 들어간 상태에서 X_1 을 추가했을 때 Y 의 변동의 한계 감소(marginal reduction, 이는 Y 에 대한 설명 변수 X_1 설명력의 한계 증가)는 다음과 같다.

$$\frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1 \mid X_2)}{SSE(X_2)} \quad \text{--- ①}$$

①은 X_2 가 주어졌을 때 Y 와 X_1 의 부분 상관 계수라 하고 $r^2{}_{Y1\bullet 2}$ 라 표시한다. 그러므로 X_1 가 주어졌을 때 Y 와 X_2 의 부분 상관 계수는

$$r^2{}_{Y2\bullet 1} = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} \quad \text{--- ②}$$

설명변수 X_1 과 종속변수 Y 의 단순 회귀분석의 잔차 $(Y_i - \hat{Y}_i(X_1))$와 설명변수 X_1 와 종속변수 X_2 의 단순 회귀분석의 잔차 $(X_{2i} - \hat{X}_{2i}(X_1)$ 의 상관계수 r^2 은 $r^2{}_{Y2\bullet 1}$. 그러므로 종속변수와 설명변수의 부분 결정계수는 다른 설명 변수에 의해 수정된(adjusted) 결정계수이다.

▶일반적인 경우 $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_p X_{pi} + e_i$

$$r^2{}_{Y1\circ 23} = \frac{SSR(X_1 \mid X_2, X_3)}{SSE(X_2, X_3)}, \quad r^2{}_{Y2\circ 13} = \frac{SSR(X_2 \mid X_1, X_3)}{SSE(X_1, X_3)}, \quad r^2{}_{Y4\circ 123} = \frac{SSR(X_4 \mid X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

부분 상관 계수는 부분 결정계수의 제곱근이며 부호는 회귀계수 추정치의 부호에 의해 결정된다. 위에서 언급한 것 같이 $r^2{}_{Y4\circ 123}$ 는 설명변수 (X_1, X_2, X_3) 와 종속변수 Y 의 단순 회귀분석의 잔차 $(Y_i - \hat{Y}_i(X_1, X_2, X_3))$와 설명변수 X_4 와 종속변수 (X_1, X_2, X_3) 의 회귀분석의 잔차 $(X_{4i} - \hat{X}_{4i}(X_1, X_2, X_3)$ 사이의 단순 상관 계수의 제곱(r^2)과 같으므로 부분 상관 계수는 나머지 설명 변수들에 의해(given) 수정된 종속 변수 Y 와 설명 변수 X_4 의 상관 관계 정도를 나타낸다. 부분 결정계수는 주로 변수 선택에 사용될 수 있지만 다른 방법들이(F-검정을 이용한 stepwise) 선호되고 있으므로 자주 사용되는 개념은 아니다.

4.6 표준화 회귀계수

설명변수의 측정 단위가 다른 경우 종속변수에 대한 그들의 영향력을 비교할 때 사용되는 개념이 표준화 회귀계수다. OLS 추정치 $\hat{\beta} = (X'X)^{-1}X'Y$ 를 계산할 때 반올림(rounding-off) 문제가 발생하게 되는데 이는 $(X'X)^{-1}$ 계산하는 과정에서 발생한다. 이런 문제는 (1) $X'X$ 의 행렬식의 값이 0에 가깝거나 $|X'X| \approx 0$ (2) $X'X$ 의 측정 단위의 차이가 많은 경우 발생하게 된다. (1)은 다중공선성 문제라 하는데 이는 설명 변수들간의 높은 상관 관계로 인하여 발생한다. 다음에 다루기로 한다. (2)의 문제는 단위를 표준화함으로써 해결할 수 있다. 변수들을 표준화 한 후 구한 회귀계수를 표준화 회귀계수(standardized regression coefficient)라 하며 이는 (1)측정 단위가 달라 반올림으로 인해 발생하는 문제를 해결할 수 있을 뿐 아니라 (2)설명 변수의 종속 변수에 대한 설명력 비교(추정된 회귀계수)를 하고자 할 때 사용된다. 그러나 컴퓨터 발달로 인하여 반올림 문제는 거의 해결하였으므로 표준화 회귀계수는 설명 변수 간의 설명력 비교에 주로 사용된다.

$Y_i^* = \frac{Y_i - \bar{Y}}{s_Y}$, $X_{ki}^* = \frac{X_{ki} - \bar{X}_k}{s_{X_k}}$, $(i = 1,2,...,n,\ k = 1,2,...,p)$ 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_k 는 편미분 계수($\frac{dy^*}{dx_k^*}$)에 해당하므로 다른 설명변수들의 값이 주어졌을 때 종속 변수 Y 에 대한 설명 변수 X_k 의 한계 영향력(한계 설명력)으로 해석된다.

EXAMPLE 4-8 표준화 회귀계수

MRI_IQ.xls

지적 능력 예제 계속: 설명 변수 MRI, VIQ, PIQ 세 개를 사용했을 때 지적 능력에 가장 영향을 많이 미치는 설명 변수는 무엇인가?

설명 변수들의 측정 단위가 다르므로 표준화 회귀계수를 구하여 비교하여야 한다.

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

STB 옵션이 표준화 회귀계수를 출력한다.

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 MRI의 영향력은 가장 적을 뿐 아니라 음의 영향을 미치고 있음을 알 수 있다. 근데 이상하지 않나요? 머리가 크면 지적 능력이 떨어진다? 이는 다중공선성 문제이다. 자세한 내용은 나중에 다루기로 한다.

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	14.23792	46.90291	0.30	0.7632	0
MRI	MRI	1	0.00010953	0.00005157	2.12	0.0406	0.33371

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	-10.88005	16.34802	-0.67	0.5101	0
MRI	MRI	1	0.00001801	0.00001876	0.96	0.3437	0.05486
VIQ	VIQ	1	0.96409	0.05934	16.25	<.0001	0.92864

(MRI), (MRI, VIQ)만 있는 모형에서는 양의 영향 미친다. 어떻게 이런 일이... 짐작 하겠지만 다중공선성 문제이다. 즉 VIQ, PIQ의 상관 관계가 매우 높아 회귀계수 추정에 문제가 발생한다. PIQ가 들어 오면서 VIQ, MRI 추정치가 전혀 다른 값을 보이고 있다.

```
PROC CORR DATA=MRI;
    VAR MRI VIQ PIQ;
RUN;
```

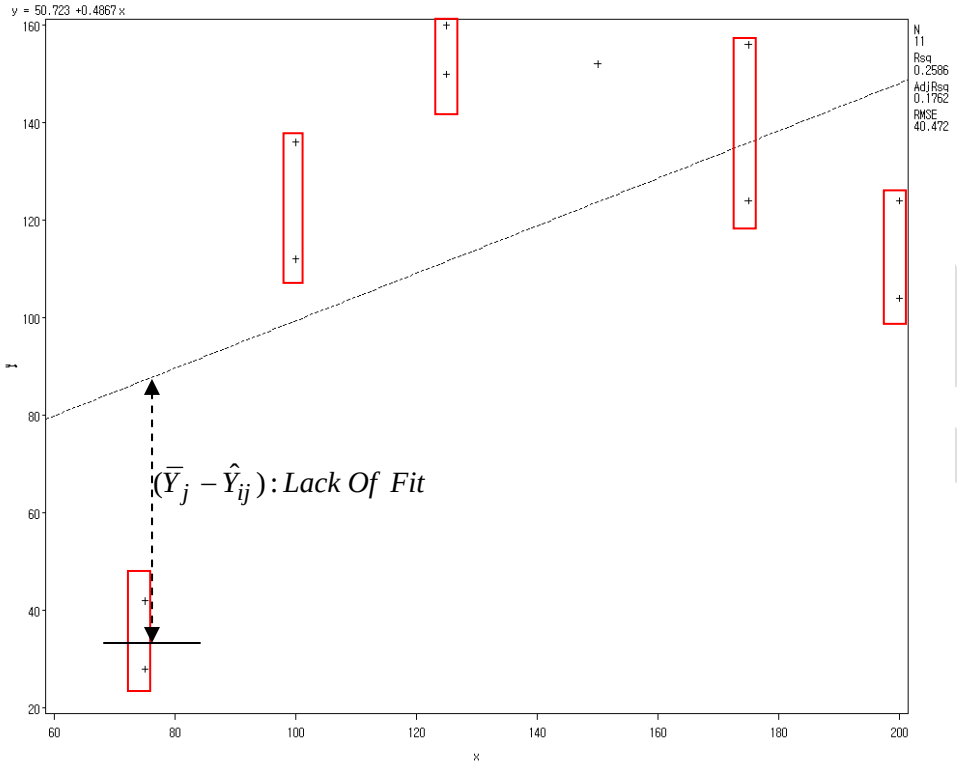
	MRI	VIQ	PIQ
MRI	1.00000	0.30028 0.0670	0.37778 0.0194
VIQ	0.30028 0.0670	1.00000	0.77602 <.0001

4.7 적합성 결여 검정

동일한 설명변수의 값에 종속변수가 2개 이상 측정이 있는 경우 회귀모형의 유의성 검정을 하는 경우 적합성 결여는 실제 직선 관계가 아니기 때문에 발생한 것이 아니라 종속변수의 분산에 의해 오차의 분산이 커지기 때문일 수 있다. 적합 결여성 분석은 설명변수가 하나인 경우 주로 실시된다. 다음 예제 데이터를 살펴보자.

```
data lack;
    input x y @@;
    cards;
125 160 100 112 200 124 75 28 150 152 175 156
75 42 175 124 125 150 200 104 100 136
run;

goptions reset=all;
proc reg data=lack;
    model y=x;
    plot y*x;
run;
```



동일한 설명변수에 종속변수 값이 여러 개 측정되었으므로 설명변수간의 직선 관계 결여 때문에 오차 변동이 생길 뿐 아니라 측정 오차에 의한 변동이 생긴다. 그러므로 오차 변동은 다음과 같이 분해할 수 있다. 설명변수의 동일 값이 종속변수가 2개 이상이므로 i, j 두 개 첨자가 필요하다. 오차 변동은 다음과 같이 분해할 수 있다.

$$(Y_{ij} - \hat{Y}_{ij}) = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_{ij}) \Rightarrow \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum (Y_{ij} - \bar{Y}_j) + \sum\sum (\bar{Y}_j - \hat{Y}_{ij})$$
$$SSE = SSPE + SSLF$$

설명변수 동일 값에 대해 종속변수의 관측치가 반복되는 오차변동(SSE)은 관측치의 반복으로 생긴 순수오차변동(SSPE, SS of Pure Error)과 두 변수간의 직선 관계의 결여로 인하여 생긴 적합결여변동(SSLF, SS of Lack of Fit)으로 나눌 수 있다.

적합결여변동 $\sum\sum (\bar{Y}_j - \hat{Y}_{ij})$ 이므로 이는 일원분산분석($Y_{ij} = \mu_i + e_{ij}$)의 오차변동과 동일하다. 이런 데이터의 경우 두 변수의 직선 관계 유의성 검정은 다음과 같이 해야 한다.

Full model: $Y_{ij} = \mu_i + e_{ij}$

```
proc glm data=lack;
    class x;
    model y=x;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	18734.90909	3746.98182	16.32	0.0041
Error	5	1148.00000	229.60000		
Corrected Total	10	19882.90909			

Reduced model: $Y_{ij} = \alpha + \beta x_{ij} + e_{ij}$

```
proc reg data=lack;
    model y=x;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5141.33841	5141.33841	3.14	0.1102
Error	9	14742	1637.95230		
Corrected Total	10	19883			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50.72251	39.39791	1.29	0.2301
x	1	0.48670	0.27471	1.77	0.1102

데이터의 직선관계 유의성 검정을 위한 분산분석 표를 만들어 보자.

변동(source)		SS(자승합)	자유도	MS(평균자승합)	F-검정
Regression (모형)		5141	1	5141.3	3398.4/229.6 =14.8
Error(오차)	결여적합성	13593	4	3398.4	
	순수 변동	1148	5	229.6	
Total (총 변동)		19883	10		

$E(MSPE) = \sigma^2$, $E(MSLF) = \sigma^2 + \sum n_j [\mu_j - (\alpha + \beta_j x_j)]^2 / (k-2)$ (k는 설명변수 수준 수)이므로

$F^* = \dfrac{SSLF/(c-p+1)}{SSE/(n-c)} = \dfrac{MSLF}{MSPE}$ (c=수준 수)가 크면 두 변수 직선 관계는 유의하지 않다.

귀무가설: $H_0 : E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

대립가설: $H_a : E(Y) \neq \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

설명변수 X와 종속변수 Y의 직선 관계에 대한 유의성 검정은 MSE(=SSE/9)를 이용하는 것이 아니라 순수변동의 MSPE(=SSPE/자유도)를 이용해야 한다. 유의수준 0.05 하에서 귀무가설(설명변수와 종속변수 간에는 직선 관계가 성립한다, 기각치 $F(0.95;4,5)=5.19$)가 기각된다. 그러므로 X, Y의 직선 관계는 유의하다. OLS 추정회귀직선 $\hat{Y}_{ij}=50.7+0.48x_{ij}$ 이 위 데이터에 적합하다.

직선 적합성결여 검정과는 달리 단순회귀 분석을 하면 회귀계수(회귀모형)도 유의하지 않다고 결론이 나온다.($F=5141.3/1637.95=3.14$, 이전 페이지 분산분석 결과를 참고) 어느 것이 옳은가? 동일한 설명변수 값에 대해 관측치가 반복되므로 당연히 적합성결여 유의성 검정을 실시하는 것이 바람직하다.

HOMEWORK #7-2 DUE 4월 27일(수)

MRI_IQ.xls 데이터 (SPSS) 활용

i:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_{i1}:	4	4	4	4	6	6	6	6	8	8	8	8	10	10	10	10
X_{i2}:	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4
Y_i:	64	73	61	76	72	80	71	83	83	89	86	93	88	95	94	100

①다중회귀분석($Y_{ij}=\alpha+\beta_1 x_{1ij}+\beta_2 x_{2ij}+e_{ij}$)을 실시하고 회귀계수를 추정하시오.

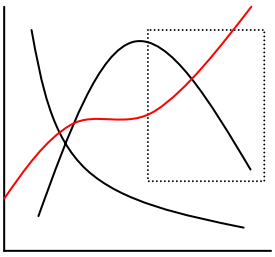
②직선 적합결여성 분석을 실시하시오.

③결과 ①, ② 이용하여 최종 회귀모형을 적고 해석하시오.

4.8 다항 회귀 모형

설명 변수가 2개 이상이고 (1)각 설명 변수의 차수가 2차 이상이거나(이고) (2) 설명 변수의 교차항이 존재하는 모형을 다항 회귀 모형(Polynomial Regression)이라 하며 이 모형은 곡선 반응(curvilinear response) 모형이다.

변수가 하나인 경우 다항 모형은 산점도(사실 쉽게 구별할 수 있는 것은 아니다, 아래 예제 참고)나 잔차 분석 결과(잔차와 예측치 산점도가 이차 형태를 갖는다) 이용하면 가능하나 변수가 2개 이상인 경우 다항 모형을 생각하는 것은 쉬운 일이 아니다. 다소 TRIAL-ERROR 방법이나 이론적 모형에 의존할 수 밖에 없다. 설명변수가 하나인 경우 다항 모형은 다음과 같다.



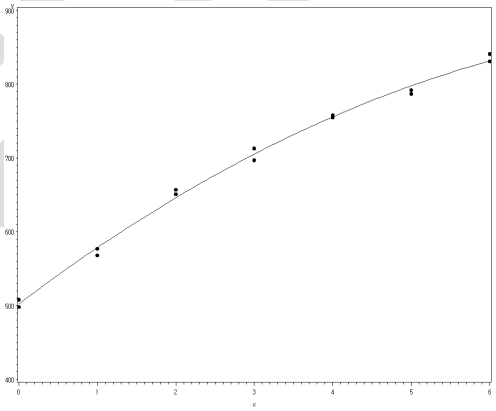
$Y_i = \beta_0 + \beta_{11}x_i + \beta_{12}x_i^2 + e_i$ (second order),

$Y_i = \beta_0 + \beta_{11}x_i + \beta_{12}x_i^2 + \beta_{13}x_i^3 + e_i$ (third order)

설명변수 x_i 와 x_i^2 는 상관 관계가 높으므로 두 설명변수를 모두 넣은 다항식 모형은 다중공선성 문제가 발생한다. 이를 완화 시키는 방법으로 $x_i^* = (x_i - \bar{x})$ (혹은 표준화, 앞에서 언급)을 사용한다. 데이터가 이차형식을 따르므로 다음과 같이 GPLOT을 이용하여 산점도를 그리면 된다. RQ는 regression quadratic의 약어이다. 이차 형식 함수를 그려준다. 직선은 RL을 사용한다.

```
data poly;
    input x y @@;
    cards;
0 508 0 498 1 568 1 577 2 651 2 657 3 713 3 697
4 755 4 758 5 787 5 792 6 841 6 831
run;
```

```
proc gplot data=poly;
    goptions reset=all;
    symbol i=rq v=dot c=black;
    plot y*x;
run;
```



```
data poly1;
    set poly;
    x1=x-3;
    x2=x1*x1;
run;
```

```
proc reg data=poly1;
    model y=x1 x2;
run;
```

설명변수 X의 평균은 3이다.

일차항(X1), 이차항(X2) 모두 유의하므로 최종 추정 회귀모형은 다음과 같다.

$\hat{y} = 705.05 + 54.8x - 4.24x^2$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	705.04762	3.21803	219.09	<.0001
x1	1	54.83929	1.05335	52.06	<.0001
x2	1	-4.24405	0.60815	-6.98	<.0001