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.

$Y_i^*=\dfrac{Y_i-\bar{Y}}{s_Y}$, $X_{ki}^*=\dfrac{X_{ki}-\bar{X}_k}{s_{X_k}}$, $(i=1,2,\ldots,n,\ k=1,2,\ldots,p)$ 표준화 하고 다음 회귀 모형의 OLS 추정치를 구하면 표준화 회귀계수라 한다. $y_i^*=\beta_0+\beta_1X_{1i}^*+\beta_2X_{2i}^*+..+\beta_pX_{pi}^*+e_i$

회귀계수 β_k 는 편미분 계수($\dfrac{dy^*}{dx_k^*}$)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;
```

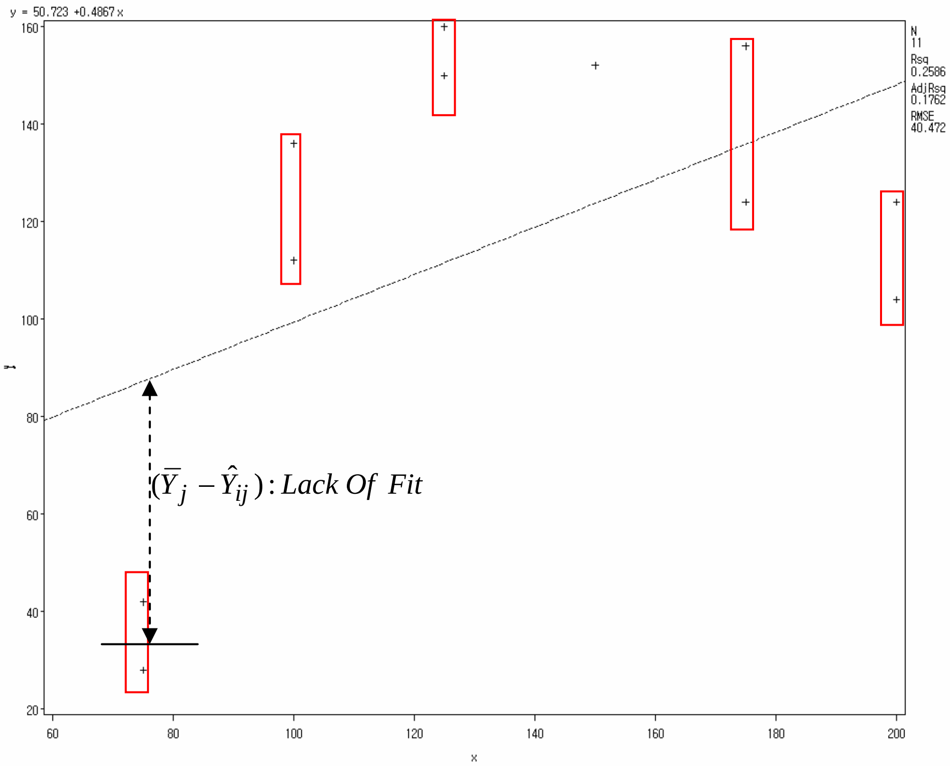
		MRI	VIQ	PIQ
MRI	MRI	1.00000	0.30028	0.37778
	MRI		0.0670	0.0194
VIQ	VIQ	0.30028	1.00000	0.77602
	VIQ	0.0670		<.0001

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.

<table>
<tr><th colspan="2">변동(source)</th><th>SS(자승합)</th><th>자유도</th><th>MS(평균자승합)</th><th>F-검정</th></tr>
<tr><td colspan="2">Regression (모형)</td><td>5141</td><td>1</td><td>5141.3</td><td rowspan="3">3398.4/229.6
=14.8</td></tr>
<tr><td rowspan="2">Error(오차)</td><td>결여적합성</td><td>13593</td><td>4</td><td>3398.4</td></tr>
<tr><td>순수 변동</td><td>1148</td><td>5</td><td>229.6</td></tr>
<tr><td colspan="2">Total (총 변동)</td><td>19883</td><td>10</td><td></td><td></td></tr>
</table>

$E(MSPE) = \sigma^2$, $E(MSLF) = \sigma^2 + \sum n_j [\mu_j - (\alpha + \beta_j x_j)]^2 / (k-2)$ (k 는 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: $H_0 : E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$

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

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.

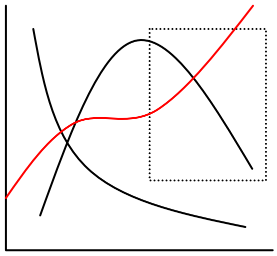
4.8 다항 회귀 모형

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),

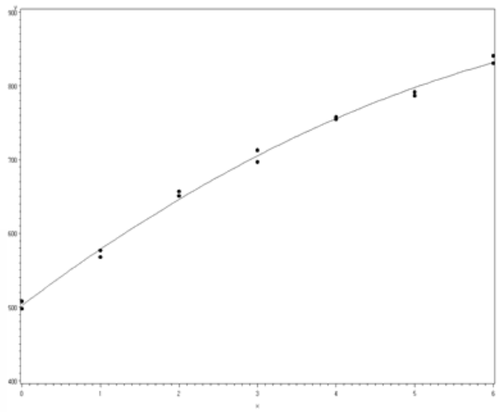
$Y_i = \beta_0 + \beta_{11}x_i + \beta_{12}x_i^2 + \beta_{13}x_i^3 + e_i$ (third order)



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