제 14 장 회귀분석

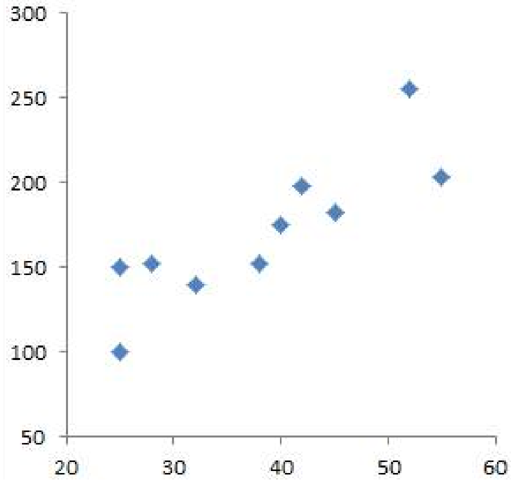
제1절 단순회귀분석 (Simple Regression Analysis)

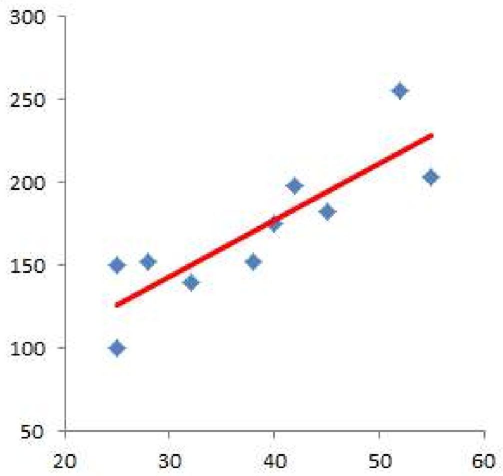
두 변수, 즉 하나의 독립변수와 하나의 종속변수 사이의 관계를 알아내는 것

[표 14-1] 평수와 전기소모량

가구	평수	전기소모량	가구	평수	전기소모량
1	25	100	6	45	183
2	52	256	7	40	175
3	38	152	8	55	203
4	32	140	9	28	152
5	25	150	10	42	198

산포도와 선형회귀선





1. 단순회귀직선모형 (Simple Linear Regression Model)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

여기서, y : 종속변수(dependent variable)

x : 독립변수(independent variable)

β_0 : 절편 모수(true y intercept for the population)

β_1 : 기울기 모수(true slope for the population)

ϵ_i: 오차항 (random error in y for observation i), 분포는 $N(0, \sigma^2)$

가정: $Cov(\epsilon_i, \epsilon_j) = 0$, 단, $i \neq j$

회귀모형의 가정

① x는 확률변수가 아니라 확정된 값이다.

② 모든 오차는 정규분포를 이루며, 평균이 0, 분산은 σ^2이며, x값에 관계없이 동일하다. → $\epsilon_i \sim N(0, \sigma^2)$

③ 서로 다른 관찰치의 오차는 독립적이다. → $Cov(\epsilon_i, \epsilon_j) = 0$, 단, $i \neq j$

④ $y \sim N(\beta_0 + \beta_1 \cdot x, \sigma^2)$

단순회귀식 (The Simple (Linear) Regression Equation)

$$\hat{y}_i = b_0 + b_1 x_i$$

$\hat{y}_i$: i번째 관찰값의 종속변수 추정치(the predicted value of y for observation i)

x_i: i번째 관찰값의 독립변수값 (the value of x for observation i)

b_0: 절편 추정치 (모수가 아니라 표본에서 구한 값)

b_1: 기울기 추정치 (모수가 아니라 표본에서 구한 값)

e_i: 잔차 (residual, $(y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$)

최소자승법 (Least Squares Method)

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$\sum e_i^2$을 최소화하는 b_0와 b_1의 값

$$b_1 = \frac{SS_{xy}}{SS_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

where SS_x: $\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

SS_y: $\sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$

SS_{xy}: $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

회귀분석식 구하기: b_0와 b_1 구하기

가구	x_i	y_i	x_i-$\overline{x}$	y_i-$\overline{y}$	$(x_i-\overline{x})^2$	$(y_i-\overline{y})^2$	$(x_i-\overline{x})(y_i-\overline{y})$
1	25	100	-13.20	-70.90	174.24	5,026.81	935.88
2	52	256	13.80	85.10	190.44	7,242.01	1,174.38
3	38	152	-0.20	-18.90	0.04	357.21	3.78
4	32	140	-6.20	-30.90	38.44	954.81	191.58
5	25	150	-13.20	-20.90	174.24	436.81	275.88
6	45	183	6.80	12.10	46.24	146.41	82.28
7	40	175	1.80	4.10	3.24	16.81	7.38
8	55	203	16.80	32.10	282.24	1,030.41	539.28
9	28	152	-10.20	-18.90	104.04	357.21	192.78
10	42	198	3.80	27.10	14.44	734.41	102.98
합계	382	1,709	-0.00	-0.00	1,027.60	16,302.90	3,506.20
평균	38.2	170.9			$= SS_x$	$= SS_y$	$= SS_{xy}$

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{3,506.20}{1,0273.60} = 3.41$$

$$b_0 = \overline{y} - b_1\overline{x} = 170.9 - (3.41)38.2 = 40.56$$

$$\hat{y_i} = b_0 + b_1 x_i = 40.56 + 3.41 x_i$$

추정치와 잔차 구하기: $\hat{y_i}$와 e_i 구하기

가구	x_i (평수)	y_i (전기료)	$\hat{y_i}$ $40.56 + 3.41x_i$	e_i $y_i - \hat{y_i}$	e_i^2 $(y_i - \hat{y_i})^2$
1	25	100	125.86	-25.86	668.80
2	52	256	217.99	38.01	1,445.07
3	38	152	170.22	-18.22	331.88
4	32	140	149.75	-9.75	94.97
5	25	150	125.86	24.14	582.68
6	45	183	194.10	-11.10	123.25
7	40	175	177.04	-2.04	4.17
8	55	203	228.22	-25.22	636.15
9	28	152	136.10	15.90	252.90
10	42	198	183.87	14.13	199.78
합계	382	1,709	1,709.00	-	4,339.65

Least Squares Method로 b_0, b_1을 구하면

① $\sum_{i=1}^{n} e_i^2$이 최소화된다.

② $\sum_{i=1}^{n} e_i = 0$이 항상 성립한다.

위의 경우 추정식은 $\hat{y_i} = b_0 + b_1 x_i = 40.56 + 3.41 x_i$이며,

잔차는 $e_i = y_i - \hat{y_i} = y_i - (b_0 + b_1 x_i) = y_i - (40.56 + 3.41 x_i)$이다.

① 어떤 b_0, b_1값을 대입하여도 $\sum_{i=1}^{n} e_i^2$는 4,339.65보다 작아지지 않는다. (특수한 경우, b_0, b_1 값을 변경해도 $\sum_{i=1}^{n} e_i^2$이 동일할 수 있다.)

→ $b_0 = 40.56$, $b_1 = 3.41$은 $\sum_{i=1}^{n} e_i^2$를 최소화시키는 b_0와 b_1이다.

② LSM으로 b_0와 b_1을 구하면, 잔차의 합은 항상 0이 된다.

연습문제 1.

i	x_i	y_i	i	x_i	y_i
1	3	32	5	6	55
2	4	45	6	4	35
3	2	19	7	6	65
4	7	65	8	3	39

문제 1. 다음의 값들을 구하시오.

① $\overline{x}$, $\overline{y}$ ② SS_x, SS_y, SS_{xy} ③ b_1, b_0

문제 2. 위의 자료로부터 Simple Linear Regression의 공식을 도출하시오.

문제 3. 위의 자료로부터 추정치와 잔차, 잔차제곱을 구하시오.

문제 4. 잔차들의 합은?

문제 5. 잔차 제곱의 합은?

문제 6. 산포도를 작성하고 선형회귀선을 삽입하시오.

2. 회귀선의 정도 (Precision)

(1) 추정의 표준오차

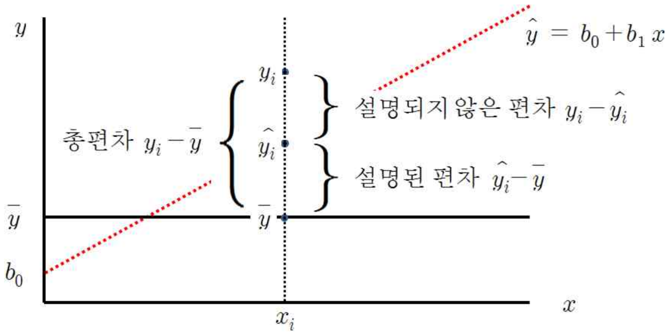
Error Sum of Squares, SSE $\quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$

MSE $\quad \dfrac{SSE}{n-2} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \dfrac{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{n-2}$

추정의 표준오차, $s_{y \cdot x}$ $\quad \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{n-2}}$

(2) 결정계수 (Coefficient of Determination)

$(y_i - \bar{y})$	=	$(y_i - \hat{y}_i)$	+	$(\hat{y}_i - \bar{y})$
(총편차)		(설명 안되는 편차)		(설명되는 편차)



$\sum_{i=1}^{n}(y_i - \bar{y})^2$	=	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	+	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
SST	=	SSE	+	SSR
(총변동)		(설명 안되는 변동)		(설명되는 변동)

Note: SST = SS_y $\leftarrow$ $SS_y = \sum (y_i - \bar{y})^2$

[표 14-1] 자료의 SST, SSR, SSE 구하기 $\bar{y}$ = 170.9, $\hat{y}_i$ = 40.56 + 3.41x_i

	x_i	y_i	$\hat{y}_i$	$(y_i-\bar{y})^2$	$(\hat{y}_i-\bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	25	100	125.86	5,026.81	2,028.49	668.80
2	52	256	217.99	7,242.01	2,217.09	1,445.07
3	38	152	170.22	357.21	0.47	331.88
4	32	140	149.75	954.81	447.52	94.97
5	25	150	125.86	436.81	2,028.49	582.68
6	45	183	194.10	146.41	538.32	123.25
7	40	175	177.04	16.81	37.72	4.17
8	55	203	228.22	1,030.41	3,285.82	636.15
9	28	152	136.10	357.21	1,211.23	252.90
10	42	198	183.87	734.41	168.11	199.78
합계	382	1,709	1,709.00	16,302.90 SST (SS_y)	11,963.25 SSR	4,339.65 SSE

표본결정계수 $r^2 = \dfrac{SSR}{SST} = \dfrac{SST-SSE}{SST} = 1-\dfrac{SSE}{SST}$

Note: $0 \le r^2 \le 1$

분산분석표 Analysis of Variance

원천 Source	제곱합 SS	자유도 df	제곱평균 MS	F비	F기각치
회귀 Regression	SSR	1	MSR $= \dfrac{SSR}{1}$	F $= \dfrac{MSR}{MSE}$	$F_{\alpha,\,1,n-2}$
잔차 Error	SSE	n-2	MSE $= \dfrac{SSE}{n-2}$		
계 Total	SST	n-1			

연습문제 2. [표 14-1]의 자료를 사용하시오.

문제 1. r^2

문제 2. 분산분석표(Analysis of Variance)를 완성하시오. $\alpha = 0.05$

3. 회귀선의 적합성 Goodness of Fit

회귀선이 유의한가에 대한 질문 (주의: H_0는 유의하지 않음이다.)

H_0: 회귀선은 유의하지 않다.

H_A: 회귀선은 유의하다.

Test Statistic: $F = \dfrac{MSR}{MSE}$

Rejection Region: $F > F_{\alpha, 1, n-2}$

[표 14-1]의 자료 α= 0.05

Source	SS	df	MS	F비	F기각치	p-value
Regression	11,963.25	1	11,963.25	22.05	5.3177	0.0015
Error	4,339.65	8	542.46			
Total	16,302.90	9				

$F = 22.05 > F_{\alpha, 1, n-2} = F_{0.05, 1, 8} = 5.32$이므로 H_0를 기각한다. (회귀선은 유의하다.)

4. 회귀분석의 추론

(1) β_1의 신뢰구간 추정과 가설검정

$$E(b_1) = \beta_1$$

$$V(b_1) = \frac{\sigma^2}{\sum(x_i - \overline{X})^2} = \frac{\sigma^2}{SS_x}, \quad \sigma_{b_1} = \frac{\sigma}{\sqrt{SS_x}}$$

회귀분석의 가정들이 성립하면, b_1은 정규분포를 따른다.

β_1의 100(1-α)% 신뢰구간 $\quad b_1 \pm t_{n-2, \frac{\alpha}{2}} \cdot \dfrac{s_e}{\sqrt{SS_x}}$

가설검정 (Drawing Inferences About b_1)

β_1 검정

$H_A : \beta_1 \neq 0$	x와 y 간에 선형관계가 존재하는 가에 대한 검증
$H_A : \beta_1 > 0$	x와 y 간에 양의 선형관계가 존재하는 가에 대한 검증
$H_A : \beta_1 < 0$	x와 y 간에 음의 선형관계가 존재하는 가에 대한 검증

모든 경우에 H_0: $\beta_1 = 0$ (x와 y 간에 선형관계가 존재하지 않는다.)

The Test Statistic: $t_{n-2} = \dfrac{b_1 - \beta_1}{s_{b_1}}$, where $s_{b_1} = \dfrac{s_\epsilon}{\sqrt{SS_x}}$

For α, Rejection Region: $t < -t_{n-2,\frac{\alpha}{2}}$, $t_{n-2,\frac{\alpha}{2}} < t$ if $H_A : \beta_1 \neq 0$

$t_{n-2,\alpha} < t$ if $H_A : \beta_1 > 0$

$t < -t_{n-2,\alpha}$ if $H_A : \beta_1 < 0$

예제.

1. $\hat{y}_i = 40.56 + 3.41x_i$에서 β_1의 95% 신뢰구간을 구하시오.

   MSE = 542.46, SS_x = 1,027.60, n = 10

   Ans. $b_1 = 3.41$, $t_{8,\,0.025} = 2.3060$, $s_{b_1} = \dfrac{s_e}{\sqrt{SS_x}} = \dfrac{\sqrt{MSE}}{\sqrt{SS_x}} = \dfrac{\sqrt{542.46}}{\sqrt{1,027.60}} = 0.7266$

   신뢰구간 = 3.41 ± 2.3060 × 0.7266 = 3.41 ± 1.68 → (1.735, 5.086)

2. β_1에 대한 가설검정을 수행하시오. $\alpha = 0.05$

   H_0: $\beta_1 = 0$, H_A: $\beta_1 > 0$

   Ans. Test Statistic $t_8 = \dfrac{b_1 - \beta_1}{s_{b_1}}$, where $s_{b_1} = \dfrac{s_\epsilon}{\sqrt{SS_x}}$

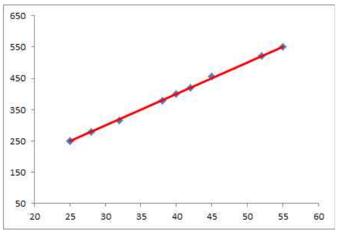
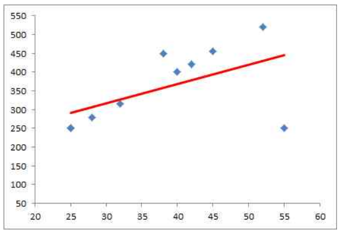
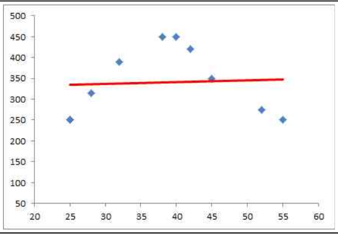
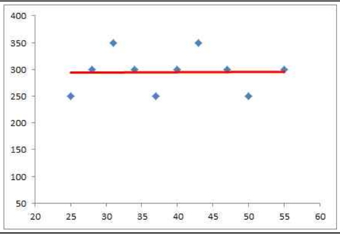
   For 0.05, Rejection Region $t > t_{8,\,0.05} = 1.8595$

   Value of the Test Statistic $t = \dfrac{3.41 - 0}{0.7266} = 4.6931$

   Conclusion Reject H_0

   Note: p-value = $P(t_8 > 4.6931) = 0.00078$

모의실험 $n = 10$, H_0: $\beta_1 = 0$, H_A: $\beta_1 \neq 0$, $\alpha = 0.05$, $t_{8,0.025} = 2.3060$

	$\hat{y} = -4.94+10.13x$ $t = 140.41$ + 선형관계 존재		$\hat{y} = 161.09+5.18x$ $t = 1.8313$ 선형관계가 매우 약함
	$\hat{y} = 323+0.45x$ $t = 0.1628$ 규칙적이지만, 선형이 아님		$\hat{y} = -293+0.06x$ $t = 0.0434$ 규칙적이지만, 선형이 아님

5. 상관계수

모집단 상관계수 $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

$$r = \frac{\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

$$r^2 = \frac{SS_{xy}^2}{SS_x \cdot SS_y} = \frac{SSR}{SST}$$ [1]

---

1) Note: $SSR = \sum (\hat{y}_i - \bar{y})^2 = \sum (b_0 + b_1 x_i - b_0 - b_1 \bar{x})^2$ since $\hat{y}_i = b_0 + b_1 x_i$, $\bar{y} = b_0 + b_1 \bar{x}$

$= b_1^2 \cdot \sum (x_i - \bar{x})^2 = b_1^2 \cdot SS_x$ since $SS_x = \sum (x_i - \bar{x})^2$

$SST = SS_y$

$\dfrac{SSR}{SST} = \dfrac{b_1^2 \cdot SS_x}{SS_y}$ since $SSR = b_1^2 \cdot SS_x$ and $SST = SS_y$

$= b_1^2 \cdot \dfrac{SS_x}{SS_y} = \dfrac{SS_{xy}^2}{SS_x^2} \cdot \dfrac{SS_x}{SS_y}$ since $b_1 = \dfrac{SS_{xy}}{SS_x}$

$= \dfrac{SS_{xy}^2}{SS_x \cdot SS_y} = r^2$

연습문제 3. $\hat{y}_i = 135.63 + 1.05x_i$

Analysis of Variance ($\alpha = 0.05$)

Source	SS	df	MS	F비	F기각치	p-value
Regression	1,017.01	1			5.3177	0.4865
Error						
Total	16,302.90	9				

문제 1. 위의 ANOVA 테이블을 완성하시오.

문제 2. F 기각치를 해석하시오. ($\alpha = 0.05$)

문제 3. $r^2 =$

문제 4. $s_e^2 =$

문제 5. $\hat{y}_i = 135.63 + 1.05x_i$에서 β_1의 95% 신뢰구간을 구하시오. ($SS_x = 918.50$)

문제 6. β_1에 대한 가설검정을 수행하시오. $\alpha = 0.05$

H_0: $\beta_1 = 0$, $H_{A:}$ $\beta_1 > 0$

6. Using the Regression Equation

6.1 Predicting the Particular Value of y for a Given x_g

Prediction Interval: $\hat{y} \pm t_{\frac{\alpha}{2},n-2}\, s_\epsilon \sqrt{1+\frac{1}{n}+\frac{(x_g-\bar{x})}{SS_x}}$

6.2 Estimating the Expected Value of y for a Given x_g

Confidence Interval: $\hat{y} \pm t_{\frac{\alpha}{2},n-2}\, s_\epsilon \sqrt{\frac{1}{n}+\frac{(x_g-\bar{x})}{SS_x}}$

모의실험 β_1 **= 0,** $\epsilon_i \sim N(0,10^2)$**,** n **= 20**

실험 **1.** $\alpha = 0.05$

$\hat{y}_i = 0.33 - 0.07x_i$, $r^2 = 0.0095$, $s_{b_1} = 0.1799$, $t = -0.4159$

Source	SS	df	MS	F비	F기각치	p-value
Regression	14.88	1	14.883	0.1729	4.4139	0.6824
Error	1,548.96	18	86.053			
Total	1,563.84	19				

실험 **2.** $\alpha = 0.05$

$\hat{y}_i = -1.41 + 0.07x_i$, $r^2 = 0.0053$, $s_{b_1} = 0.2286$ $t = 0.3084$

Source	SS	df	MS	F비	F기각치	p-value
Regression	13.22	1	13.215	0.0951	4.4139	0.7613
Error	2,501.08	18	138.949			
Total	2,514.29	19				

실험 **3.** $\alpha = 0.05$

$\hat{y}_i = 10.26 - 0.29x_i$, $r^2 = 0.1187$, $s_{b_1} = 0.1860$ $t = -1.5573$

Source	SS	df	MS	F비	F기각치	p-value
Regression	223.21	1	223.212	2.4250	4.4139	0.1368
Error	1,656.80	18	92.045			
Total	1,880.02	19				

모의실험 $\beta_1 = 10$, $\epsilon_i \sim N(0, 10^2)$, $n = 20$

실험 **1.** $\alpha = 0.05$

$\hat{y}_i = 10.67 + 9.69x_i$, $r^2 = 0.9937$, $s_{b_1} = 0.1814$, $t = 53.4091$

Source	SS	df	MS	F비	F기각치	p-value
Regression	249,647.32	1	249,647.320	2,852.5331	4.4139	0.00E+00
Error	1,575.32	18	87.518			
Total	251,222.64	19				

실험 2. $\alpha = 0.05$

$\hat{y}_i = 5.98 + 9.86 x_i$, $r^2 = 0.9937$, $s_{b_1} = 0.1854$, $t = 53.1556$

Source	SS	df	MS	F비	F기각치	p-value
Regression	258,404.73	1	258,404.726	2,825.5152	4.4139	0.00E+00
Error	1,646.17	18	91.454			
Total	260,050.90	19				

실험 3. $\alpha = 0.05$

$\hat{y}_i = -1.83 + 10.04 x_i$, $r^2 = 0.9921$, $s_{b_1} = 0.2113$, $t = 47.5270$

Source	SS	df	MS	F비	F기각치	p-value
Regression	268,314.45	1	268,314.453	2,258.8118	4.4139	0.00E+00
Error	2,138.14	18	118.786			
Total	270,452.59	19				

제 2 절 중회귀분석

1. 중회귀모형

중회귀모형 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$

여기서 $\epsilon \sim N(0, \sigma^2)$이며 독립적이다.

회귀방정식 $\hat{y_i} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$

Error $e_i = y_i - \hat{y_i} = y_i - (b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)$

Least Squares Method (1)

(1) Q = 잔차제곱의 합 (편의상 잔차제곱의 합을 Q로 표현)

$= \sum e^2 = \sum (y - \hat{y})^2 = \sum (y - b_o - b_1 x_1 - b_2 x_2 - \cdots - b_k x_k)^2$

(2) Q를 각각의 b_i들로 편미분한 값들이 0이 될 때, Q는 최소화된다.

$$\frac{\partial Q}{\partial b_0} = 2 \sum (y - b_o - b_1 x_1 - b_2 x_2 - \cdots - b_k x_k)(-1) = 0$$

$$\rightarrow \sum (y - b_o - b_1 x_1 - b_2 x_2 - \cdots - b_k x_k) = 0$$

$$\frac{\partial Q}{\partial b_i} = 2 \sum (y - b_o - b_1 x_1 - b_2 x_2 - \cdots - b_k x_k)(-x_i) = 0, \ j = 1, 2, ..., k$$

$$\rightarrow \sum (y - b_o - b_1 x_1 - b_2 x_2 - \cdots - b_k x_k) \cdot x_j = 0, \ j = 1, 2, ..., k$$

(3) 정규방정식(normal equation)을 구하면 다음과 같다.

$$\sum(y-b_o-b_1x_1-b_2x_2-\ \cdots\ -b_kx_k)\ =\ 0$$
$$\rightarrow\ \sum y\ =\ \sum b_o\ +\ \sum b_1x_1\ +\ \sum b_2x_2\ +\ \cdots\ +\ \sum b_kx_k$$
$$\rightarrow\ \sum y\ =\ n\cdot b_0\ +\ b_1\sum x_1\ +\ b_2\sum x_2\ +\ \cdots\ +\ b_k\sum x_k$$

$$\sum(y-b_o-b_1x_1-b_2x_2-\ \cdots\ -b_kx_k)\cdot x_j\ =\ 0,\ j\ =\ 1,\ 2,\ ...,\ k$$
$j\ =\ 1$인 경우 $\rightarrow\ \sum x_1\,y\ =\ b_0\sum x_1\ +\ b_1\sum x_1^2\ +\ b_2\sum x_1\,x_2\ +\ \cdots\ +\ b_k\sum x_1\,x_k$

$j\ =\ 2$인 경우 $\rightarrow\ \sum x_2\,y\ =\ b_0\sum x_2\ +\ b_1\sum x_1\,x_2\ +\ b_2\sum x_2^2\ +\ \cdots\ +\ b_k\sum x_2\,x_k$

$j\ =\ 3$인 경우 $\rightarrow\ \sum x_3\,y\ =\ b_0\sum x_3\ +\ b_1\sum x_1\,x_3\ +\ b_2\sum x_2^2\ +\ \cdots\ +\ b_k\sum x_3\,x_k$

$\vdots$

$j\ =\ k$인 경우 $\rightarrow\ \sum x_k\,y\ =\ b_0\sum x_k\ +\ b_1\sum x_1\,x_k\ +\ b_2\sum x_2\,x_k\ +\ \cdots\ +\ b_k\sum x_k^2$

행렬로 표현하면

$$\begin{pmatrix} n & \sum x_1 & \sum x_2 & \sum x_3 & \cdots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_1x_2 & \sum x_1x_3 & \cdots & \sum x_1x_k \\ \sum x_2 & \sum x_2x_1 & \sum x_2^2 & \sum x_2x_3 & \cdots & \sum x_2x_k \\ \sum x_3 & \sum x_3x_1 & \sum x_3x_2 & \sum x_3^2 & \cdots & \sum x_3x_k \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_k & \sum x_kx_1 & \sum x_kx_2 & \sum x_kx_3 & \cdots & \sum x_k^2 \end{pmatrix}\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum x_1y \\ \sum x_2y \\ \sum x_3y \\ \vdots \\ \sum x_ky \end{pmatrix}$$

$$\rightarrow\ \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} n & \sum x_1 & \sum x_2 & \sum x_3 & \cdots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_1x_2 & \sum x_1x_3 & \cdots & \sum x_1x_k \\ \sum x_2 & \sum x_2x_1 & \sum x_2^2 & \sum x_2x_3 & \cdots & \sum x_2x_k \\ \sum x_3 & \sum x_3x_1 & \sum x_3x_2 & \sum x_3^2 & \cdots & \sum x_3x_k \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum x_k & \sum x_kx_1 & \sum x_kx_2 & \sum x_kx_3 & \cdots & \sum x_k^2 \end{pmatrix}^{-1}\begin{pmatrix} \sum y \\ \sum x_1y \\ \sum x_2y \\ \sum x_3y \\ \vdots \\ \sum x_ky \end{pmatrix}$$

Least Squares Method (2) - 다른 표현법

(1) $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$, $\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$, $\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

$$\begin{aligned}
y_1 &= 1\times b_o + x_{11}\times b_1 + x_{12}\times b_2 + \cdots + x_{1j}\times b_j + \cdots + x_{1k}\times b_k + e_1 \\
y_2 &= 1\times b_o + x_{21}\times b_1 + x_{22}\times b_2 + \cdots + x_{2j}\times b_j + \cdots + x_{2k}\times b_k + e_2 \\
\vdots & \qquad\qquad\qquad \vdots \\
y_i &= 1\times b_o + x_{i1}\times b_1 + x_{i2}\times b_2 + \cdots + x_{ij}\times b_j + \cdots + x_{ik}\times b_k + e_i \\
\vdots & \qquad\qquad\qquad \vdots \\
y_n &= 1\times b_o + x_{n1}\times b_1 + x_{n2}\times b_2 + \cdots + x_{nj}\times b_j + \cdots + x_{nk}\times b_k + e_n
\end{aligned}$$

$$e_i = y_i - 1\times b_o - x_{i1}\times b_1 - x_{i2}\times b_2 - \cdots - x_{ij}\times b_j - \cdots - x_{ik}\times b_k$$

(2) $\displaystyle\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_o - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_j x_{ij} - \cdots - b_k x_{ik})^2$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_0} = 2\sum_{i=1}^{n} (y_i - b_o - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_j x_{ij} - \cdots - b_k x_{ik})(-1) = 0$$

$$\begin{aligned}
\rightarrow \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} b_0 + \sum_{i=1}^{n} b_1 x_{i1} + \sum_{i=1}^{n} b_2 x_{i2} + \cdots + \sum_{i=1}^{n} b_j x_{ij} + \cdots + \sum_{i=1}^{n} b_k x_{ik} \\
&= n b_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \cdots + b_j \sum_{i=1}^{n} x_{ij} + \cdots + b_k \sum_{i=1}^{n} x_{ik}
\end{aligned}$$

$j = 1, 2, ..., k$에 대하여

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_j} = 2\sum_{i=1}^{n} (y_i - b_o - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_j x_{ij} - \cdots - b_k x_{ik})(-x_{ij}) = 0$$

$$\begin{aligned}
\rightarrow \sum_{i=1}^{n} x_{ij} y_i &= \sum_{i=1}^{n} b_0 x_{ij} + \sum_{i=1}^{n} b_1 x_{i1} x_{ij} + \cdots + \sum_{i=1}^{n} b_j x_{ij}^2 + \cdots + \sum_{i=1}^{n} b_k x_{ik} x_{ij} \\
&= b_0 \sum_{i=1}^{n} x_{ij} + b_1 \sum_{i=1}^{n} x_{i1} x_{ij} + \cdots + b_j \sum_{i=1}^{n} x_{ij}^2 + \cdots + b_k \sum_{i=1}^{n} x_{ik} x_{ij}
\end{aligned}$$

행렬로 표현하면

$$
\begin{vmatrix}
n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ij} & \cdots & \sum_{i=1}^{n} x_{ik} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ij} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ik} \\
\sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2} x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2} x_{ij} & \cdots & \sum_{i=1}^{n} x_{i2} x_{ik} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
\sum_{i=1}^{n} x_{ij} & \sum_{i=1}^{n} x_{ij} x_{i1} & \sum_{i=1}^{n} x_{ij} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ij}^2 & \cdots & \sum_{i=1}^{n} x_{ij} x_{ik} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} x_{ij} & \cdots & \sum_{i=1}^{n} x_{ik}^2
\end{vmatrix}
\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_k \end{bmatrix}
=
\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}\, y_i \\ \sum_{i=1}^{n} x_{i2}\, y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ij}\, y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\, y_i \end{bmatrix}
$$

Note.

$$
\begin{aligned}
& X^T X \\
&= \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}^T
\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix}
\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \\
&= \begin{vmatrix}
n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ij} & \cdots & \sum_{i=1}^{n} x_{ik} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ij} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ik} \\
\sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2} x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2} x_{ij} & \cdots & \sum_{i=1}^{n} x_{i2} x_{ik} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
\sum_{i=1}^{n} x_{ij} & \sum_{i=1}^{n} x_{ij} x_{i1} & \sum_{i=1}^{n} x_{ij} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ij}^2 & \cdots & \sum_{i=1}^{n} x_{ij} x_{ik} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} x_{ij} & \cdots & \sum_{i=1}^{n} x_{ik}^2
\end{vmatrix}
\end{aligned}
$$

$$X^T y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}\, y_i \\ \sum_{i=1}^{n} x_{i2}\, y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\, y_i \end{bmatrix}$$

정리하면 $(X^T X)\, b = X^T y$

(3) 정규방정식으로 b를 구하면,

$(X^T X)\, b = X^T y$

$\rightarrow (X^T X)^{-1} \times (X^T X) \times b = (X^T X)^{-1} \times X^T y$

$\rightarrow b = (X^T X)^{-1} \times X^T y$

Example 평수(x_1), 가족 수(x_2), 전기소모량(y) 자료

	y	x_1	x_2	$x_1\, y$	$x_2\, y$	$x_1\, x_2$	x_1^2	x_2^2
1	100	25	3	2,500	300	75	625	9
2	256	52	6	13,312	1,536	312	2,704	36
3	152	38	5	5,776	760	190	1,444	25
4	140	32	5	4,480	700	160	1,024	25
5	150	25	4	3,750	600	100	625	16
6	183	45	7	8,235	1,281	315	2,025	49
7	175	40	5	7,000	875	200	1,600	25
8	203	55	4	11,165	812	220	3,025	16
9	152	28	2	4,256	304	56	784	4
10	198	42	4	8,316	792	168	1,764	16
Sum	1,709	382	45	68,790	7,960	1,796	15,620	221

$$\sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \cdots + b_k \sum_{i=1}^{n} x_{ik}$$

$$\sum_{i=1}^{n} x_{ij} y_i = b_0 \sum_{i=1}^{n} x_{ij} + b_1 \sum_{i=1}^{n} x_{i1}\, x_{ij} + \cdots + b_k \sum_{i=1}^{n} x_{ik}\, x_{ij}, \quad j = 1,\ 2,\ \ldots,\ k$$ 이므로

$1{,}709 = 10 b_0 + 382 b_1 + 45 b_2$

$68{,}790 = 328 b_0 + 15{,}620 b_1 + 1{,}796 b_2$

$7{,}960 = 45 b_0 + 1{,}796 b_1 + 221 b_2$

$$\rightarrow \begin{bmatrix} 10 & 328 & 45 \\ 328 & 15{,}620 & 328 \\ 45 & 1{,}796 & 221 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1{,}709 \\ 68{,}790 \\ 7{,}960 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 10 & 328 & 10 \\ 328 & 15{,}620 & 328 \\ 45 & 1{,}796 & 45 \end{bmatrix}^{-1} \begin{bmatrix} 1{,}709 \\ 68{,}790 \\ 7{,}960 \end{bmatrix} = \begin{bmatrix} 1.7307 & -0.0275 & -0.1286 \\ -0.0275 & 0.0014 & -0.0059 \\ -0.1286 & -0.0059 & 0.0786 \end{bmatrix} \begin{bmatrix} 1{,}709 \\ 68{,}790 \\ 7{,}960 \end{bmatrix} = \begin{bmatrix} 39.69 \\ 3.37 \\ 0.53 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 25 & 3 \\ 1 & 52 & 6 \\ 1 & 38 & 5 \\ \vdots & \vdots & \vdots \\ 1 & 42 & 4 \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 25 & 52 & 38 & \cdots & 42 \\ 3 & 6 & 5 & \cdots & 4 \end{bmatrix}, \quad y = \begin{bmatrix} 100 \\ 256 \\ 152 \\ \vdots \\ 198 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 25 & 52 & 38 & \cdots & 42 \\ 3 & 6 & 5 & \cdots & 4 \end{bmatrix} \begin{bmatrix} 1 & 25 & 3 \\ 1 & 52 & 6 \\ 1 & 38 & 5 \\ \vdots & \vdots & \vdots \\ 1 & 42 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 328 & 45 \\ 328 & 15{,}620 & 328 \\ 45 & 1{,}796 & 221 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 10 & 328 & 45 \\ 328 & 15{,}620 & 328 \\ 45 & 1{,}796 & 221 \end{bmatrix}^{-1} = \begin{bmatrix} 1.7307 & -0.0275 & -0.1286 \\ -0.0275 & 0.0014 & -0.0059 \\ -0.1286 & -0.0059 & 0.0786 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 25 & 52 & 38 & \cdots & 42 \\ 3 & 6 & 5 & \cdots & 4 \end{bmatrix} \begin{bmatrix} 100 \\ 256 \\ 152 \\ \vdots \\ 198 \end{bmatrix} = \begin{bmatrix} 1{,}709 \\ 68{,}790 \\ 7{,}960 \end{bmatrix}$$

$$b = (X^T X)^{-1} \times X^T y = \begin{bmatrix} 1.7307 & -0.0275 & -0.1286 \\ -0.0275 & 0.0014 & -0.0059 \\ -0.1286 & -0.0059 & 0.0786 \end{bmatrix} \begin{bmatrix} 1{,}709 \\ 68{,}790 \\ 7{,}960 \end{bmatrix} = \begin{bmatrix} 39.6892 \\ 3.3722 \\ 0.5321 \end{bmatrix}$$

중회귀식: $\hat{y} = 39.69 + 3.37x_1 + 0.53x_2$

평수(x_1) = 30, 가족 수(x_2) = 4일 때의 전기소모량 추정치
= 39.69 + 3.37(30) + 0.53(4) = 142.91

2. 중회귀식의 정도

(1) 추정의 표준오차

추정의 분산 (variance of the errors): $S_e^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n-k-1}$

추정의 표준오차(the regression standard error/the residual standard error):

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}}$$

(2) 결정계수 (Coefficient of Determination)

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

(3) 수정표본결정계수 (Adjusted Sample Coefficient of Determination)

$$r_a^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{SSE/n-k-1}{SST/n-1}$$

(4) 상관계수

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}}$$

독립변수들 사이에 상관관계가 없으면 ($r_{12}^2 = 0$), $R_{y\cdot 12}^2 = r_{y\cdot 1}^2 + r_{y\cdot 2}^2$

3. 중회귀식의 적합성

총편차 = 설명되는 편차 + 설명 안되는 편차

$(y_i - \overline{y})$	=	$(\hat{y}_i - \overline{y})$	+	$(y_i - \hat{y}_i)$
(총편차)		(설명되는 편차)		(설명 안되는 편차)

SST = SSR + SSE

$\sum(y_i - \overline{y})^2$	=	$\sum(\hat{y}_i - \overline{y})^2$	+	$\sum(y_i - \hat{y}_i)^2 \; \left(= \sum e_i^2\right)$
SST		SSR		SSE

분산분석표 (Analysis of Variance)

원천 Source	제곱합 SS	자유도 df	제곱평균 MS	F비	F기각치
회귀 Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	$F_{\alpha,\, k,\, n-k-1}$
잔차 Error	SSE	n-k-1	$MSE = \frac{SSE}{n-k-1}$		
계 Total	SST	n-1			

회귀선의 가설검정

H_0 : $\beta_1 = \beta_2 = ... = \beta_k = 0$ (해석: 회귀선은 유의하지 않다.)

H_A: β_i 중 최소한 하나는 0이 아니다. (해석: 회귀선은 유의하다.)

Test Statistic: $F = \frac{MSR}{MSE}$, 자유도는 k와 $n-k-1$

For α, Rejection Region: $F > F_{\alpha,\, k,\, n-k-1}$

β_i의 가설검정

H_0 : $\beta_i = 0$ (해석: i번째 독립변수 x_i와 y 간에 선형관계가 존재하지 않는다.)

H_A: $\beta_i \neq 0$ (해석: i번째 독립변수 x_i와 y 간에 선형관계가 존재한다.)

Test Statistic: $t = \frac{b_i - \beta_i}{s_{b_i}}$, 자유도 n-k-1

Rejection Region: $t < -t_{\frac{\alpha}{2}, n-k-1}$ 또는 $t > t_{\frac{\alpha}{2}, n-k-1}$

F-Test, r^2 and s_e

SSE	s_e	r^2	F	Assessment of Model
0	0	1	∞	perfect
small	small	close to 1	large	good
large	large	close to 0	small	poor
SST	$\sqrt{\frac{SST}{n-k-1}}$	0	0	no linear relationship

모의실험 1. 회귀식이 전혀 유의하지 않은 경우 $\beta_1 = \beta_2 = \beta_3 = 0$, $\epsilon_i \sim N(0, 3^2)$

실험 1.

회귀분석 통계량	
다중 상관계수	0.3026
결정계수	0.0915
조정된 결정계수	-0.0133
표준오차	3.5713
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱 평균	F 비	P-값
회귀	3	33.41	11.14	0.87	0.4676
잔차	26	331.60	12.75		
계	29	365.02			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%
Y 절편	4.26	7.21	0.59	0.5595	-10.56	19.09
X1	-3.16	2.82	-1.12	0.2724	-8.95	2.63
X2	-0.00	1.40	-0.00	0.9980	-2.87	2.87
X3	1.28	1.74	0.74	0.4681	-2.30	4.87

Note: H_0: $\beta_i = 0$, H_A: $\beta_i \neq 0$, $t = \dfrac{b_i - \beta_i}{s_{b_i}}$이므로

(1) $b_i < 0$이면, $t < 0$, $b_i > 0$이면, $t > 0$

(2) $|t|$이 크면 p-value는 0에 가깝고, $|t|$이 작으면 p-value는 1에 가깝다.

(3) 양측검정이므로(H_A: $\beta_i \neq 0$) p-value $< \alpha$이면 β_i의 신뢰구간은 0을 포함하지 않고, p-value $> \alpha$이면 β_i의 신뢰구간은 0을 포함한다.

실험 2.

회귀분석 통계량	
다중 상관계수	0.2712
결정계수	0.0735
조정된 결정계수	-0.0334
표준오차	3.2207
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱평균	F비	P-값
회귀	3	21.41	7.14	0.69	0.5676
잔차	26	269.69	10.37		
계	29	291.10			

	계수	표준오차	t통계량	P-값	하위95%	상위95%
Y절편	8.40	6.50	1.29	0.2077	-4.96	21.77
X1	-2.04	2.54	-0.80	0.4290	-7.26	3.18
X2	-1.77	1.26	-1.40	0.1719	-4.36	0.82
X3	1.09	1.57	0.69	0.4940	-2.14	4.32

모의실험 2. 회귀식은 유의하고, β_2가 유의하지 않은 경우

$$\beta_1 = 10, \quad \beta_2 = 0, \quad \beta_3 = -5, \; \epsilon_i \sim N(0, 3^2)$$

실험 1.

회귀분석 통계량	
다중 상관계수	0.7917
결정계수	0.6267
조정된 결정계수	0.5836
표준오차	3.1482
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱 평균	F 비	P-값
회귀	3	432.65	144.22	14.55	0.0000
잔차	26	257.70	9.91		
계	29	690.34			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%
Y 절편	-2.05	6.36	-0.32	0.7498	-15.11	11.02
X1	12.29	2.48	4.95	0.0000	7.18	17.39
X2	**-0.06**	**1.23**	**-0.05**	**0.9621**	**-2.59**	**2.47**
X3	-6.20	1.54	-4.03	0.0004	-9.36	-3.04

실험 2.

회귀분석 통계량	
다중 상관계수	0.7251
결정계수	0.5257
조정된 결정계수	0.4710
표준오차	3.1262
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱평균	F 비	P-값
회귀	3	281.65	93.88	9.61	0.0002
잔차	26	254.10	9.77		
계	29	535.75			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%
Y절편	-10.93	6.31	-1.73	0.0952	-23.91	2.04
X1	12.36	2.47	5.01	0.0000	7.29	17.43
X2	**2.35**	**1.22**	**1.92**	**0.0659**	**-0.17**	**4.86**
X3	-6.48	1.53	-4.25	0.0002	-9.62	-3.34

모의실험 3. 회귀식과 모든 β_i가 유의한 경우;

β_1 **= 10,** β_2 **= -10,** β_3 **= -5,** $\epsilon_i \sim N(0, 3^2)$

실험 1.

회귀분석 통계량	
다중 상관계수	0.9498
결정계수	0.9021
조정된 결정계수	0.8908
표준오차	3.1385
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱 평균	F 비	P-값
회귀	3	2,359.77	786.59	79.86	0.0000
잔차	26	256.10	9.85		
계	29	2,615.87			

	계수	표준오차	t통계량	P-값	하위95%	상위95%
Y절편	-0.87	6.34	-0.14	0.8922	-13.89	12.16
X1	12.10	2.48	4.89	0.0000	7.01	17.19
X2	-9.67	1.23	-7.88	0.0000	-12.19	-7.15
X3	-7.17	1.53	-4.68	0.0001	-10.32	-4.03

실험 2.

회귀분석 통계량	
다중 상관계수	0.9600
결정계수	0.9216
조정된 결정계수	0.9126
표준오차	2.6520
관측수	30

분산 분석 (α=0.05)

	자유도	제곱합	제곱 평균	F 비	P-값
회귀	3	2,149.88	716.63	101.89	0.0000
잔차	26	182.87	7.03		
계	29	2,332.74			

	계수	표준 오차	t 통계량	P-값	하위95%	상위95%
Y절편	-11.43	5.35	-2.13	0.0425	-22.43	-0.42
X1	14.68	2.09	7.02	0.0000	10.38	18.98
X2	-7.90	1.04	-7.63	0.0000	-10.03	-5.77
X3	-7.69	1.29	-5.94	0.0000	-10.35	-5.03

Summary of Formulas

$$SS_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_x}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

$$SSE = SS_y - \frac{SS_{xy}^2}{SS_x}$$

$$s_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{SS_x}}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

$$r^2 = \frac{SS_y - SSE}{SS_y} = \frac{SSR}{SS_y}$$

$$\hat{y} \pm t_{\frac{\alpha}{2}, n-2}\, s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \overline{x})}{SS_x}} \quad \text{(prediction interval)}$$

$$\hat{y} \pm t_{\frac{\alpha}{2}, n-2}\, s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_g - \overline{x})}{SS_x}} \quad \text{(confidence interval)}$$

연습문제

1. (1)

i	x_i	y_i	$x_i-\overline{x}$	$y_i-\overline{y}$	$(x_i-\overline{x})^2$	$(y_i-\overline{y})^2$	$(x_i-\overline{x})(y_i-\overline{y})$
1	3	32	-1.38	-12.38	1.89	153.14	17.02
2	4	45	-0.38	0.63	0.14	0.39	-0.23
3	2	19	-2.38	-25.38	5.64	643.89	60.27
4	7	65	2.63	20.63	6.89	425.39	54.14
5	6	55	1.63	10.63	2.64	112.89	17.27
6	4	35	-0.38	-9.38	0.14	87.89	3.52
7	6	65	1.63	20.63	2.64	425.39	33.52
8	3	39	-1.38	-5.38	1.89	28.89	7.39
합계	35	355	-	-	21.88	1,877.88	192.88
평균	4.375($\overline{x}$)	44.375($\overline{y}$)			$= SS_x$	$= SS_y$	$= SS_{xy}$

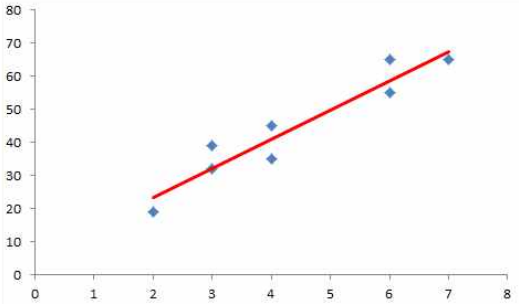
$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{192.88}{21.88} = 8.82, \quad b_0 = \overline{y} - b_1\overline{x} = 44.375 - (8.82)4.375 = 5.80$$

(2) $\hat{y_i} = b_0 + b_1 x_i = 44.375 + 5.80x_i$

(3)

i	x_i	y_i	$\hat{y_i}$ $44.375 + 5.80x_i$	e_i $y_i - \hat{y_i}$	e_i^2 $(y_i - \hat{y_i})^2$
1	3	32	32.25	-0.25	0.06
2	4	45	41.07	3.93	15.46
3	2	19	23.43	-4.43	19.66
4	7	65	67.52	-2.52	6.35
5	6	55	58.70	-3.70	13.71
6	4	35	41.07	-6.07	36.83
7	6	65	58.70	6.30	39.65
8	3	39	32.25	6.75	45.54
합계	35	355	355.00	-0.00	177.27
평균	4.375($\overline{x}$)	44.375($\overline{y}$)			

(4) $\sum_{i=1}^{8} e_i = 0$ (5) $\sum_{i=1}^{8} e_i^2 = 177.27$

(6) 

2. (1) $r^2 = \dfrac{SSR}{SST} = \dfrac{11{,}963.253}{16{,}302.900} = 0.734$

(2)

Source	SS	df	MS	F비	F기각치	p-value
Regression	11,963.25	1	11,963.25	22.05	5.3177	0.0015
Error	4,339.65	8	542.46			
Total	16,302.90	9				

3. (1)

Source	SS	df	MS	F비	F기각치	p-value
Regression	1,017.01	1	1,017.008	0.5323	5.3177	0.4865
Error	15,285.89	8	1,910.736			
Total	16,302.90	9				

(2) $\underline{\beta_1 = 0}$이고 독립변수가 1개이며, 자료가 총 10개일 때 상위 5%에 해당하는 F값

(3) $r^2 = \dfrac{SSR}{SSE} = 0.0665$ (4) $s_e^2 = \dfrac{SSE}{n-1} = MSE = 1{,}910.736$

(5) $b_1 = 1.05$, $t_{8,\,0.025} = 2.3060$, $s_{b_1} = \dfrac{s_e}{\sqrt{SS_x}} = \dfrac{\sqrt{MSE}}{\sqrt{SS_x}} = \dfrac{\sqrt{1{,}910.736}}{\sqrt{918.50}} = 1.4423$

신뢰구간 = 1.05 ± 2.3060 × 1.4423 = 1.05 ± 3.326 → (-2.274, 4.378)

(6) Test Statistic $\quad t_8 = \dfrac{b_1 - \beta_1}{s_{b_1}}$, where $s_{b_1} = \dfrac{s_\epsilon}{\sqrt{SS_x}}$

For 0.05, Rejection Region $\quad t > t_{8,\,0.05} = 1.8595$

Value of the Test Statistic $\quad t = \dfrac{1.05 - 0}{1.4423} = 0.7296$

Conclusion $\quad$ Reject H_0