Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

10 CHAPTER

相关分析与回归分析

1

第10章 相关分析与回归分析

1. 相关分析
2. 回归分析的概念
3. 简单回归分析
4. 多重回归分析*
5. 回归模型诊断*

2/52

2

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

相关分析与回归分析

- 预测因果关系 → 好的决策
- 相关分析与回归分析: 分析变量之间的相关性
- 相关分析
  - 对两个变量之间的线性关系进行计量分析
- 回归分析
  - 变量分为独立变量和从属变量，分析从属变量是否可以通过独立变量的某种函数形式来说明

3/52

3

1. 相关分析

- 相关系数(correlation coefficient)
  - 用来表示两随机变量X与Y的相关关系(线性关系)的符号和强弱的尺度

$$\rho_{XY} \equiv Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

[定理 14-1] 相关系数的特性

① ρ_{XY}取值范围为 $-1 \le \rho_{XY} \le 1$。

② 若两变量互为独立，则两变量之间不存在相关关系，$\rho_{XY}=0$。

③ 若$\rho_{XY}=0$，则两变量之间没有相关关系(线性关系)。
但是，由于有可能是非线性关系，因此无法保证两变量相互独立。

④ X与Y符合正态分布的情况下，若$\rho_{XY}=0$，则X与Y互为独立。

4/52

4

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

1. 相关分析

- 样本相关系数(sample correlation coefficient)
  - 利用样本来估计相关系数的统计量

$$r_{XY}=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\qquad S_{XX}=\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-\frac{(\sum_{i=1}^{n}X_i)^2}{n}$$

$$S_{YY}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}Y_i^2-\frac{(\sum_{i=1}^{n}Y_i)^2}{n}$$

$$S_{XY}=\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}X_iY_i-\frac{(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{n}$$

[定理 14-2] 样本相关系数的特性

① r_{XY}取值范围为$-1\le r_{XY}\le 1$。

② r_{XY} 取值越接近+1或-1，散点图上的点分布越接近直线。

③ r_{XY} 取值为+1或者-1的情况下，散点图上的所有点都在直线上。

5/52

5

1. 相关分析

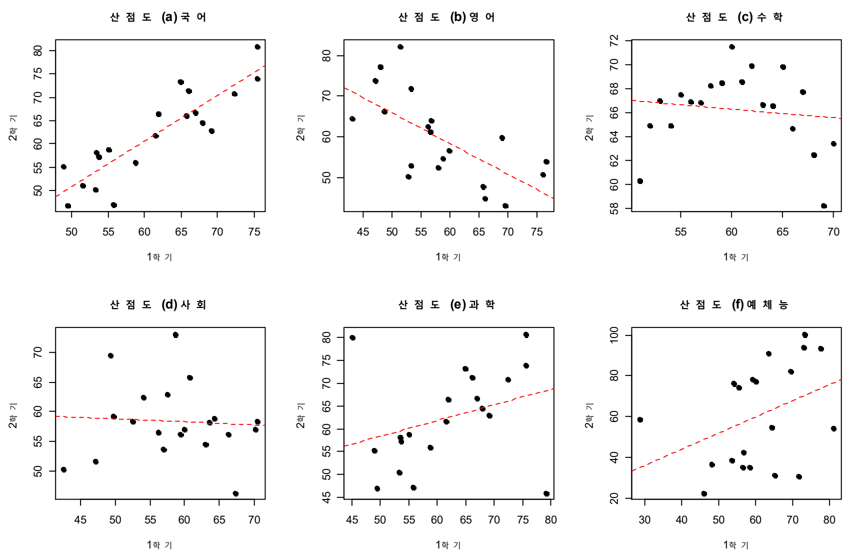
[例 14-1] 下面是20名学生第1、2学期各科目的成绩。绘制六门学科数据的散点图，求相应的相关系数并比较。

(a) 국어		(b) 영어		(c) 수학		(d) 사회		(e) 과학		(f) 예체	
1	2	1	2	1	2	1	2	1	2	1	2
53.4	58.1	69.5	43.2	51	60.3	62.9	54.5	53.4	58.1	56.7	35.3
55	58.9	59.9	56.6	52	64.9	64.3	58.9	55	58.9	81.2	54.2
53.6	57.3	58.9	54.6	53	67	60.7	65.8	53.6	57.3	46	22.6
67	66.8	76.6	53.9	54	64.9	47.1	51.6	67	66.8	48.1	36.6
69.1	62.9	53.4	71.8	55	67.5	56.2	56.5	69.1	62.9	71.8	30.8
65.7	66	69	59.7	56	66.9	60	57.1	45	80	64.5	54.9
72.3	70.8	56.2	62.4	57	66.8	49.3	69.5	72.3	70.8	53.7	38.6
49.4	47	51.4	82.2	58	68.2	59.4	56.2	49.4	47	65.2	31.2
48.9	55.3	52.9	50.3	59	68.5	56.9	53.7	48.9	55.3	56.8	42.3
51.5	51.2	43.1	64.5	60	71.5	67.3	46.3	79	46	58.4	34.9
61.5	61.7	66.1	44.9	61	68.6	54	62.4	61.5	61.7	73.4	100.4
53.3	50.4	76.2	50.7	62	69.9	66.3	56.1	53.3	50.4	28.7	58.5
64.9	73.3	56.8	63.9	63	66.7	70.5	58.3	64.9	73.3	60.2	77.4
67.9	64.5	48	77.3	64	66.6	42.6	50.2	67.9	64.5	55.6	74.2
66	71.3	58	52.3	65	69.8	58.6	73.1	66	71.3	73	93.7
75.5	80.8	56.6	61.3	66	64.7	57.5	62.9	75.5	80.8	69.5	82.1
58.7	56	65.7	47.8	67	67.7	52.5	58.4	58.7	56	59.3	78.3
75.5	73.9	47.1	73.8	68	62.5	63.6	58.2	75.5	73.9	77.6	93.2
61.9	66.4	48.7	66.1	69	58.3	49.8	59.3	61.9	66.4	54.1	76.3
55.7	47.2	53.4	52.8	70	63.4	70.1	57.1	55.7	47.2	63.6	91

6/52

6

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org



7/52

7

1. 相关分析

1.2 关于是否存在相关关系的检验

$$H_0 : \rho_{XY} = 0$$

$$T_0 \equiv r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \sim t(n-2) \Big| H_0$$

[定理 14-3] 关于是否存在相关关系的检验

原假设: “两变量之间不存在相关关系”

귀무가설 H_0: $\rho_{XY}=0$, 검정통계량: $T_0 = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}}$

대립가설 H_1 : $\rho_{XY} > 0 \Rightarrow$ 기각역: $T_0 > t_{1-\alpha;\,n-2}$

대립가설 H_1 : $\rho_{XY} < 0 \Rightarrow$ 기각역: $T_0 < t_{\alpha;\,n-2} = -t_{1-\alpha;\,n-2}$

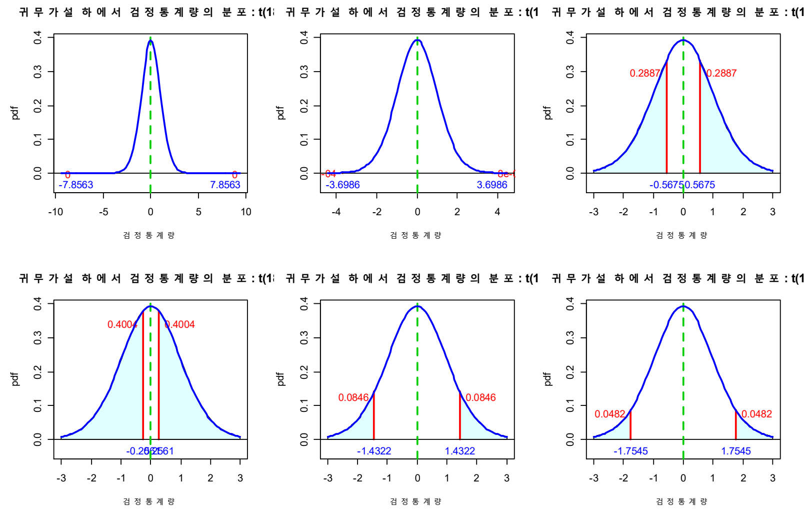
대립가설 H_1 : $\rho_{XY} \neq 0 \Rightarrow$ 기각역: $|T_0| > t_{1-\alpha/2;\,n-2}$

8/52

8

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

[例 14-2] 20名学生第1、2学期的各科成绩数据与[例 14-1]一样时，在各科目成绩的显著性水平为5%的情况下，求检验第1学期成绩与第2学期成绩之间是否存在相关关系。



[例 14-3] 某一出口公司为了检验韩币对美元汇率与出口额(亿韩元)之间的关系而收集了某一分店近10个月的相关数据。请在显著性水平为5%的情况下，检验韩币对美元汇率与出口额之间是否存在相关关系。

월	1	2	3	4	5	6	7	8	9	10
환율	1095	1110	1086	1074	1098	1105	1163	1124	1088	1064
수출액	49	52	48	49	50	51	50	51	49	48

귀무가설 $H_0: \rho_{XY}=0$, 대립가설 $H_1: \rho_{XY}\neq 0$

$$\sum_{i=1}^{10} x_i = 11{,}007,\ \sum_{i=1}^{10} y_i = 497,\ \sum_{i=1}^{10} x_i^2 = 12{,}122{,}411,\ \sum_{i=1}^{10} y_i^2 = 24{,}711$$

$$S_{XX} = 12{,}122{,}411 - 11{,}007^2/10 = 7{,}006.1 \qquad \sum_{i=1}^{10} x_i y_i = 547{,}242$$

$$S_{YY} = 24{,}711 - 497^2/10 = 16.1 \qquad S_{XY} = 547{,}242 - 11{,}007\times 497/10 = 194.1$$

$$\Rightarrow r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{194.1}{\sqrt{7006.1\times 16.1}} \doteq 0.578$$

$$T_0 = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \doteq 0.578\times\sqrt{\frac{10-2}{1-0.578^2}} \doteq 2.003 < t_{0.975;8} \doteq 2.306$$

→ 原假设被接受

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

1. 相关分析

1.3 关于相关系数的检验

$$H_0 : \rho_{XY} = \rho_0$$

$$Z_0 \equiv \sqrt{n-3}\left[\frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right)-\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right]\overset{a}{\sim}N(0,1)\ \Big|\ H_0$$

[定理 14-4] 关于相关系数的检验

귀무가설: $H_0 : \rho_{XY} = \rho_0$

검정통계량: $Z_0 \equiv \sqrt{n-3}\left[\frac{1}{2}\ln\left(\frac{1+r_{XY}}{1-r_{XY}}\right)-\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right]$

대립가설 $H_1 : \rho_{XY} > \rho_0 \Rightarrow$ 기각역: $Z_0 > z_{1-\alpha}$

대립가설 $H_1 : \rho_{XY} < \rho_0 \Rightarrow$ 기각역: $Z_0 < z_\alpha = -z_{1-\alpha}$

대립가설 $H_1 : \rho_{XY} \neq \rho_0 \Rightarrow$ 기각역: $|Z_0| > z_{1-\alpha/2}$

11/52

11

1. 相关分析

[例 14-4] 在[例 14-3]中，为了排除汇兑损益，决定将出口额单位‘亿韩元’改为’10万USD’，对数据进行分析。

(1)请检验在显著性水平为5%的情况下，韩币对美元汇率与出口额之间是否存在相关关系。 (2)请检验在显著性水平为5%的情况下，韩币对美元汇率与出口额之间的相关系数是否为0.9。

월	1	2	3	4	5	6	7	8	9	10
환율	1095	1110	1086	1074	1098	1105	1163	1124	1088	1064
수출액	53.655	57.72	52.128	52.626	54.9	56.355	58.15	57.324	53.312	51.072

(1) 귀무가설 H_0 : $\rho_{XY}=0$, 대립가설 H_1 : $\rho_{XY}\neq 0$

$$\sum_{i=1}^{10} x_i = 11{,}007,\ \sum_{i=1}^{10} x_i^2 = 12{,}122{,}411$$

$$\sum_{i=1}^{10} y_i = 547.242,\ \sum_{i=1}^{10} y_i^2 \doteq 30{,}055.160$$

$$\sum_{i=1}^{10} x_i y_i = 602{,}909.922$$

12/52

12

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

(1) 귀무가설 $H_0: \rho_{XY}=0$, 대립가설 $H_1: \rho_{XY}\neq 0$ 기각치 $t_{0.975;8} \doteq 2.306$

$$S_{XX} = 12{,}122{,}411 - 11{,}007^2 / 10 = 7{,}006.1$$

$$S_{YY} \doteq 30{,}055.160 - 547.242^2 / 10 \doteq 57.779$$

$$S_{XY} \doteq 602{,}909.922 - 11{,}007 \times 547.242 / 10 \doteq 560.653$$

$$\Rightarrow r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \doteq \frac{560.653}{\sqrt{7006.1 \times 57.779}} \doteq 0.881$$

$$T_0 = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \doteq 0.881 \times \sqrt{\frac{10-2}{1-0.881^2}} \doteq 5.267 > t_{0.975;8} \doteq 2.306$$

→ 귀무가설 기각

(2) 귀무가설 $H_0: \rho_{XY}=0.9$, 대립가설 $H_1: \rho_{XY}\neq 0.9$

$$Z_0 \doteq \sqrt{10-3}\left[\frac{1}{2}\ln\left(\frac{1+0.881}{1-0.881}\right) - \frac{1}{2}\ln\left(\frac{1+0.9}{1-0.9}\right)\right] \doteq \sqrt{7}(1.380-1.472) \doteq -0.243$$

$|Z_0| \doteq 0.243 < z_{0.975} \doteq 1.96$ → 귀무가설 채택

→ **在显著性水平为5%的情况下，无足够的证据表明韩币对美元汇率与出口额之间的相关系数不是0.9。**

13/52

13

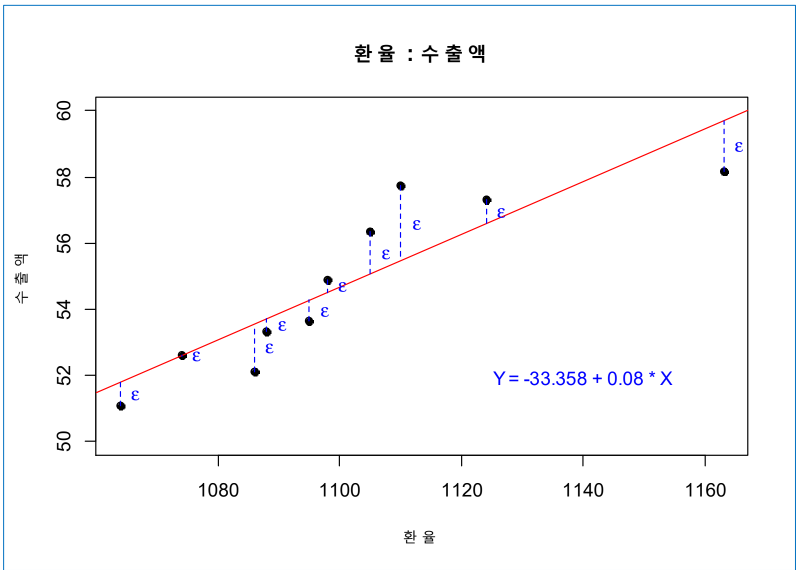
2. 回归分析的概念

- 简单回归分析(simple regression analysis)
  - 用一个自变量来说明一个因变量的模型
  - (例如) 用父亲的身高来说明某一子女的身高的情况
- 多重回归分析(multiple regression analysis)
  - 用两个以上的自变量来说明一个因变量的模型
  - (例如) 用父亲与母亲的身高来说明某一子女的身高
- 曲线回归分析(cuvilinear regression analysis)
  - 用二元以上的函数来说明自变量和因变量的关系
  - (例如) 二元函数关系 → 自变量=(x, x^2) → 使用多重回归分析的方法 → 注意自变量间的从属性
- 多元回归分析(multivariate regression analysis)
  - 用两个以上的因变量的模型
  - (例如) 用父亲与母亲的身高来说明两个子女的身高的情况

14/52

14

[例 14-5] 利用[例 14-4]的数据，绘制散点图，并标示直线回归公式。



15/52

15

3. 简单回归分析

- 简单线性回归模型 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n$
- 回归系数 $\beta_0 =$ 절편, $\beta_1 =$ 기울기
- 误差项 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- 简单线性回归模型的特性

  (1) $E(\varepsilon_i) = 0 \ \Rightarrow E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i$

  (2) $Var(y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i) \ = Var(\varepsilon_i) = \sigma^2$

  (3) ε_i's are independent $\Rightarrow y_i$'s are independent

  (4) $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$
- 简单线性回归估计模型 $y_i = \hat{y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,\ i = 1, 2, \ldots, n$

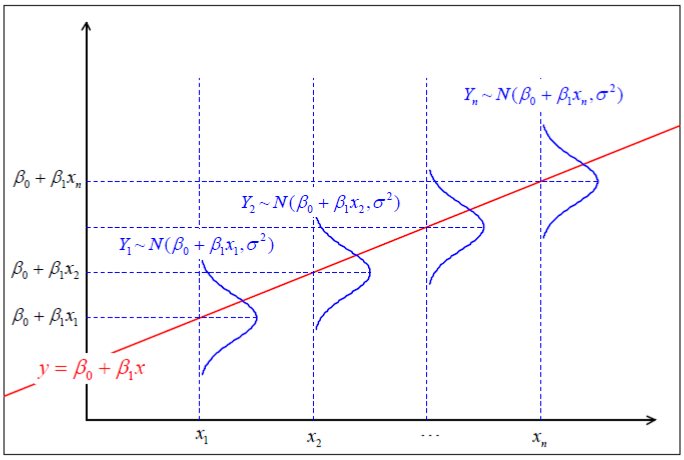
  e_i =残差(residual) → 误差的观测值

16/52

16

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

단순선형회귀 모형의 개념



17/52

17

3. 简单回归分析

3.1 回归系数的估计

- 最小二乘估计(least square estimation; LSE)
  - 通过最小化误差平方和寻求回归系数值得方法

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1}=-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

$$\left.\frac{\partial Q}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1}=-2\sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

- 正规方程(normal equation)

$$\sum_{i=1}^{n}y_i=n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i$$

$$\sum_{i=1}^{n}x_i y_i=\hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2$$

18/52

18

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

- 正规方程(normal equation)

$$\left[\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\right] \times n$$

$$-\left[\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i\right] \times \sum_{i=1}^{n} x_i$$

$$\Rightarrow n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = \hat{\beta}_1 \left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]$$

$$\boxed{\Rightarrow \hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{S_{XY}}{S_{XX}}}$$

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) / n$$

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \;\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \;\boxed{\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

19/52

19

3. 简单回归分析

[定理 14-5] 简单线性回归模型的最小二乘估计

단순선형회귀 모형: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n$

최소제곱추정치(LSE): $\hat{\beta}_1 = S_{XY} / S_{XX},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

추정 회귀식: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

잔차의 특성: $\sum_{i=1}^{n} e_i = 0,\ \sum_{i=1}^{n} x_i e_i = 0$

- 残差的特性 $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i e_i = 0$$

추정된 회귀식은 다음과 같이 평균점 $(\bar{x}, \bar{y})$을 항상 지나게 된다.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 (x_i - \bar{x})$$

$$= (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 (x_i - \bar{x}) = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

20/52

20

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

[例 14-6] 将[例 14-4]中的汇率作为自变量，出口额作为因变量，在简单线性回归模型中求回归系数的最小二乘法估计值。

$$\sum_{i=1}^{10} x_i = 11{,}007,\ \sum_{i=1}^{10} x_i^2 = 12{,}122{,}411 \quad \Rightarrow \bar{x} = 1{,}100.7,\ S_{XX} = 7{,}006.1$$

$$\sum_{i=1}^{10} y_i = 547.242,\ \sum_{i=1}^{10} x_i y_i = 602{,}909.922 \quad \Rightarrow \bar{y} \doteq 54.724,\ S_{XY} \doteq 560.653$$

$$\hat{\beta}_1 = S_{XY} / S_{XX} \doteq 560.653 / 7006.1 \doteq 0.080$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \doteq 54.724 - 0.080 \times 1100.7 = -33.332$$

$$\Rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \doteq -33.332 + 0.080x$$

```
x <- c(1095, 1110, 1086, 1074, 1098, 1105, 1163, 1124, 1088, 1064)
y2 <- c(53.655,57.72,52.128,52.626,54.9,56.355,58.15,57.324,53.312,51.072)
rg1 <- lm(y2 ~ x)
rg1$coef
(Intercept) x
-33.35765964 0.08002349
```

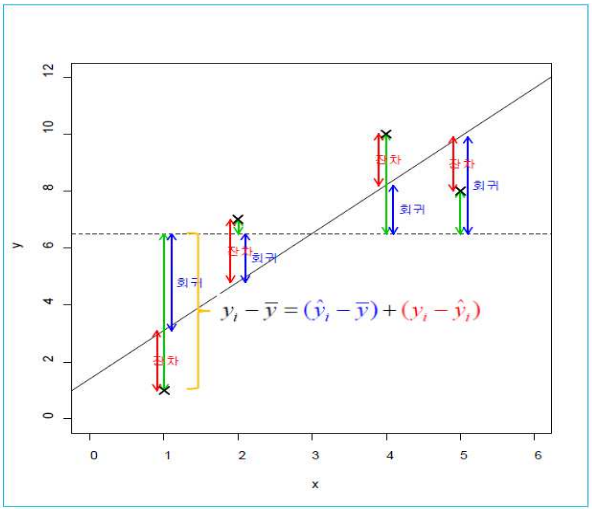
☞ 与前面的计算式相比，由于回归系数估计值的四舍五入误差，导致截距估计值多少发生了差异。

21/52

21

3.2 模型的拟合优度检验(方差分析)

- 总偏差(total deviation) $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$



22/52

22

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

- 总变异量(total variation), 总平方和(total sum of squares)

$$\boxed{SS_T \equiv \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$$

$$\hat{y}_i-\bar{y}=\hat{\beta}_1(x_i-\bar{x}),\ e_i=y_i-\hat{y}_i$$

$$\Rightarrow \sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=\sum_{i=1}^{n}\hat{\beta}_1(x_i-\bar{x})e_i=\hat{\beta}_1\sum_{i=1}^{n}x_ie_i+\hat{\beta}_1\bar{x}\sum_{i=1}^{n}e_i=0$$

$$\boxed{\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

- 回归平方和(regression sum of squares)

$$\boxed{SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\hat{\beta}_1^2S_{XX}=S_{XY}^2/S_{XX}}$$

- 误差平方和(error sum of squares)

$$\boxed{SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=SS_T-SS_R}$$

23/52

23

3. 简单回归分析

[定理 14-6] 简单线性回归模型的平方和分解

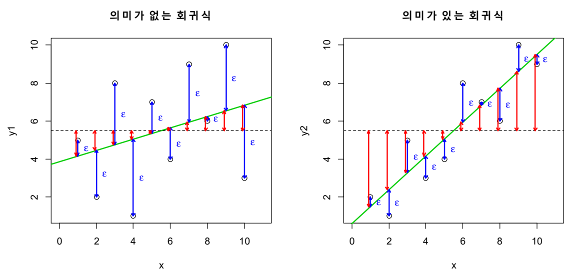
총제곱합: $SS_T \equiv \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n=SS_R+SS_E$

회귀제곱합: $SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\hat{\beta}_1^2S_{XX}=S_{XY}^2/S_{XX}$

오차제곱합: $SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=SS_T-SS_R$

SS_T의 자유도: $\phi_T=n-1$ SS_R의 자유도: $\phi_R=1$

SS_E의 자유도: $\phi_E=n-2$ $\phi_T=n-1=1+(n-2)=\phi_R+\phi_E$



24/52

24

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

[表 14-1] 简单线性回归分析的方差分析表

요인	제곱합(SS)	자유도	평균제곱(MS)	검정통계량	기각역
회귀	SS_R	1	$MS_R = SS_R$	$\frac{MS_R}{MS_E}$	$F_{1-\alpha;(1,n-2)}$
잔차	SS_E	$n-2$	$MS_E = \frac{SS_E}{n-2}$		
계	SS_T	$n-1$			

- 均值平方(mean square)：用自由度除以平方和后所得的值
- 回归公式的显著性检验(F-检验)

$$F_0 = \frac{MS_R}{MS_E} > F_{1-\alpha;(1,n-2)} \Rightarrow \text{Reject } H_0 : \beta_1 = 0$$

- 决定系数(coefficient of determination)：用来体现所估计得到的回归直线能够说明因变量的变异的程度的尺度。

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

25/52

25

3. 简单回归分析

[定理 14-7] 样本相关系数与决定系数的关系

$$r_{XY}^2 = R^2$$

[证明] $r_{XY}^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}}$

$$SS_R = \hat{\beta}_1^2 S_{XX} = \left(\frac{S_{XY}}{S_{XX}}\right)^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}}$$

$$SS_T = S_{YY}$$

$$\Rightarrow R^2 = \frac{SS_R}{SS_T} = \frac{S_{XY}^2}{S_{XX}S_{YY}} = r_{XY}^2$$

26/52

26

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

[例 14-7] 将[例 14-6]中的汇率作为自变量，出口额作为因变量，进行用于检验简单线性回归模型的拟合优度的方差分析，求出决定系数。

$$SS_T = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n \doteq 30,055.160 - (547.242)^2/10 \doteq 57.779$$

$$SS_R = \hat{\beta}_1^2 S_{XX} = S_{XY}^2 / S_{XX} = (560.653)^2/(7,006.1) \doteq 44.865$$

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_T - SS_R \doteq 12.914$$

要因	平方和( )	自由度	均值平方	检验统计量	拒绝域
回归	44.865	1	44.865	27.797	5.318
残差	12.914	8	1.614		
合计	57.779	9			

$$R^2 = \frac{SS_R}{SS_T} \doteq \frac{44.865}{57.779} \doteq 0.776$$

$$F_0 \doteq 27.797 > F_{0.95;(1,8)} \doteq 5.318$$

→ 原假设被拒绝

27/52

27

3. 简单回归分析

14.3.3 对于回归系数的估计

(1) 关于斜率 β1的检验和置信区间

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \equiv \sum_{i=1}^{n} c_i y_i \qquad c_i \equiv \frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$E(y_i) = \beta_0 + \beta_1 x_i,\ Var(y_i) = \sigma^2$$

$$\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i} = \beta_1$$

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 Var(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} = \frac{\sigma^2}{S_{XX}}$$

28/52

28

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

(1)关于斜率 β_1的检验和置信区间 (承上)

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{MS_E/S_{XX}}}\sim t(n-2)$$

β_1에 대한 $100(1-\alpha)\%$의 신뢰구간

$$\left[\hat{\beta}_1-t_{1-\alpha/2;n-2}\sqrt{\frac{MS_E}{S_{XX}}},\ \hat{\beta}_1+t_{1-\alpha/2;n-2}\sqrt{\frac{MS_E}{S_{XX}}}\right]$$

귀무가설 $H_0:\beta_1=b_1$에 대한 검정 통계량

$$T_0\equiv\frac{\hat{\beta}_1-b_1}{\sqrt{MS_E/S_{XX}}}\sim t(n-2)\ \Big|\ H_0$$

29/52

29

3. 简单回归分析

(2) 关于截距β_0的检验和置信区间

$$E(\hat{\beta}_0)=E(\bar{y}-\hat{\beta}_1\bar{x})=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0$$

$$Var(\hat{\beta}_0)=Var(\bar{y}-\hat{\beta}_1\bar{x})=\frac{\sigma^2}{n}+\bar{x}^2Var(\hat{\beta}_1)=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)$$

β_0에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left[\hat{\beta}_0\pm t_{1-\alpha/2;n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}\right]$$

가설 $H_0:\beta_0=b_0$

$$T_0\equiv\frac{\hat{\beta}_0-b_0}{\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}}\sim t(n-2)\ \Big|\ H_0$$

30/52

30

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

[定理 14-8] 对于回归系数的检验

$$H_0:\beta_1=b_1 \Rightarrow \quad T_0 \equiv \frac{\hat{\beta}_1-b_1}{\sqrt{MS_E/S_{XX}}} \sim t(n-2) \mid H_0$$

$$H_0:\beta_0=b_0 \Rightarrow \quad T_0 \equiv \frac{\hat{\beta}_0-b_0}{\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}} \sim t(n-2) \mid H_0$$

① $H_1:\beta_1>b_1\ (\beta_0>b_0) \Rightarrow$ 기각역: $T_0>t_{1-\alpha;n-2}$

② $H_1:\beta_1<b_1\ (\beta_0<b_0) \Rightarrow$ 기각역: $T_0<t_{\alpha;n-2}=-t_{1-\alpha;n-2}$

③ $H_1:\beta_1\neq b_1\ (\beta_0\neq b_0) \Rightarrow$ 기각역: $|T_0|>t_{1-\alpha/2;n-1}$

31/52

31

3. 简单回归分析

[例 14-8] 前面[例 14-7]中汇率为自变量，出口额为因变量，对于简单线性回归模型的斜率与截距的95%的置信区间，检验在显著性水平为5%时斜率和截距是否不为0。

$$\hat{\beta}_1 \doteq 0.080,\ S_{XX}\doteq 7006.1,\ MS_E\doteq 1.614,\ t_{0.975;8}\doteq 2.306$$

$$\Rightarrow \beta_1:\left[0.080\pm 2.306\times\sqrt{1.614/7006.1}\right]\doteq[0.080\pm 0.035]=[0.045,\ 0.115]$$

$$\bar{x}=1100.7,\ \hat{\beta}_0\doteq 33.332$$

$$\Rightarrow \beta_0:\left[-33.332\pm 2.306\times\sqrt{1.614\left(\frac{1}{10}+\frac{1100.7^2}{7006.1}\right)}\right]$$

$$\doteq[-33.332\pm 38.533]=[-71.865,\ 5.201]$$

$$H_0:\beta_1=0 \quad T_0\doteq\frac{0.080}{\sqrt{1.614/7006.1}}\doteq 5.271>t_{0.975;8}\doteq 2.306$$ → 귀무가설 기각

$$H_0:\beta_0=0 \quad |T_0|\doteq\frac{33.332}{\sqrt{1.614\left(\frac{1}{10}+\frac{1100.7^2}{7006.1}\right)}}\doteq 1.995<t_{0.975;8}\doteq 2.306$$ → 귀무가설 채택

32/52

32

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

3.4 回归公式的应用

(1) 用于求取自变量在某一特定取值时，相应因变量期望的置信区间。

$$E(y|x_0)\text{에 대한 추정량} \Rightarrow \hat{y}|x_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$E(\hat{y}|x_0) = E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

$$Var(\hat{y}|x_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0)$$

$$= Var(\bar{y}) + (\bar{x} - x_0)^2 Var(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

$E(y|x_0)$의 $100(1-\alpha)\%$ 신뢰구간

$$\left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{1-\alpha/2;n-2} \sqrt{MS_E \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)} \right]$$

33/52

33

3. 简单回归分析

(2) 用于求取对于未来响应值的预测区间

$$y_0 \equiv \beta_0 + \beta_1 x_0 + \varepsilon\text{에 대한 추정량} \Rightarrow \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\varepsilon}$$

$$E(\hat{y}_0) = \beta_0 + \beta_1 x_0$$

$$Var(\hat{y}_0) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

$(y|x_0)$의 $100(1-\alpha)\%$ 예측구간

$$\left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{1-\alpha/2;n-2} \sqrt{MS_E \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)} \right]$$

34/52

34

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 简单回归分析

[例 14-9] 在前面[例 14-8]中，当汇率取值为1200时，求关于所预想的出口额的95%的置信区间和95%的预测区间，并绘图展示关于简单线性回归模型的95%的置信区间和预测区间。

$$\hat{\beta}_1 \doteq -33.332,\ \hat{\beta}_1 \doteq 0.080,\ S_{XX} \doteq 7006.1,\ MS_E \doteq 1.614,\ t_{0.975;8} \doteq 2.306$$

$E(y \mid x_0 = 1200)$의 $100(1-\alpha)\%$ 신뢰구간

$$\left[(-33.332 + 0.080 \times 1200) \pm 2.306 \times \sqrt{1.614\left(\frac{1}{10} + \frac{(1200-1100.7)^2}{7006.1}\right)}\right]$$

$$\doteq [62.668 \pm 3.600] = [59.068,\ 66.268]$$

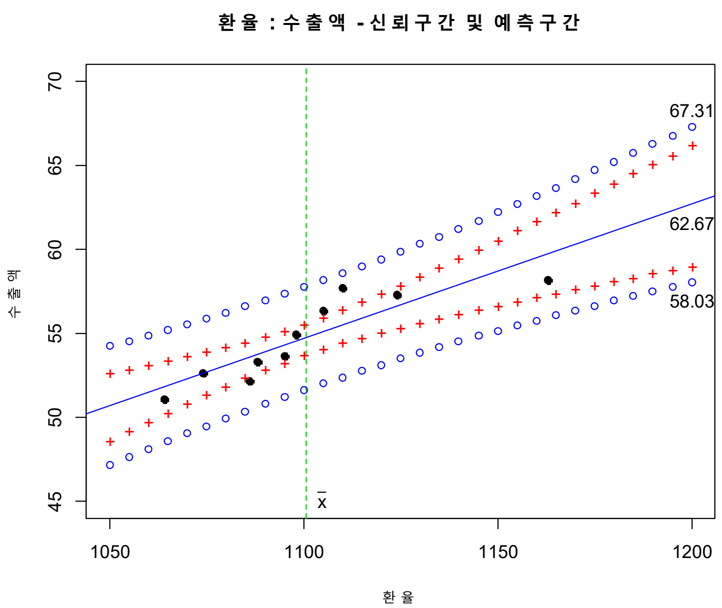
$(y \mid x_0 = 1200)$의 $100(1-\alpha)\%$ 예측구간

$$\left[62.668 \pm 2.306 \times \sqrt{1.614\left(1 + \frac{1}{10} + \frac{(1200-1100.7)^2}{7006.1}\right)}\right]$$

$$\doteq [62.668 \pm 4.642] = [58.026,\ 67.310]$$

35/52

35



36/52

36

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

- 多重回归模型 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,\ i = 1, 2, \ldots, n$
- 回归系数 $\beta_0 =$ 절편, $\beta_1, \beta_2, \ldots, \beta_k =$ 각 독립변수의 기울기
- 误差项 $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- 多重回归模型的特性

  (1) $E(\varepsilon_i) = 0 \ \Rightarrow E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$

  (2) $Var(y_i) = Var(\varepsilon_i) = \sigma^2$

  (3) ε_i's are independent $\Rightarrow y_i$'s are independent

  (4) $y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}, \sigma^2)$
- 多重回归估计模型

$$y_i = \hat{y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} + e_i,\ i = 1, 2, \ldots, n$$

e_i =残差(residual) → 误差的观测值

37/52

37

4. 多重回归分析

4.1 回归系数的估计

- 具有k个自变量的多重回归模型

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,\ i = 1, 2, \ldots, n$$

- 用矩阵、向量表示 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{n \times 1}{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}},\ \underset{n \times (k+1)}{\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}},\ \underset{(k+1) \times 1}{\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}},\ \underset{n \times 1}{\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

38/52

38

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

4.1 回归系数的估计

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- 最小二乘法(method of least squares; MSE)

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}^T - \boldsymbol{\beta}^T \mathbf{X}^T)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

$$= \mathbf{y}^T \mathbf{y} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}$$

$$= \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} \qquad \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta}$$

- 正规方程式(normal equation)

$$\left.\frac{dQ}{d\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \quad \Rightarrow \mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

39/52

39

4. 多重回归分析

- 向量的微分 $\boldsymbol{x} \equiv (x_1, x_2)^T$

일차함수 $f(x_1, x_2) = a_1 x_1 + a_2 x_2 = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{a}^T \mathbf{x}$

$$\frac{df(x_1, x_2)}{dx} \equiv \begin{vmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{vmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \qquad \boxed{\Rightarrow \frac{d(\mathbf{a}^T \mathbf{x})}{d\mathbf{x}} = \mathbf{a}}$$

이차함수 $f(x_1, x_2) = c_1 x_1^2 + c_2 x_2^2 + 2c_3 x_1 x_2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} c_1 & c_3 \\ c_3 & c_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}^T \mathbf{C} \mathbf{x}$

$$\frac{df(x_1, x_2)}{dx} = \begin{bmatrix} 2c_1 x_1 + 2c_3 x_2 \\ 2c_2 x_2 + 2c_3 x_1 \end{bmatrix} = 2 \begin{bmatrix} c_1 & c_3 \\ c_3 & c_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{C}\mathbf{x}$$

$$\boxed{\Rightarrow \frac{d(\mathbf{x}^T \mathbf{C} \mathbf{x})}{d\mathbf{x}} = 2\mathbf{C}\mathbf{x}}$$

40/52

40

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

- 正规方程式 $Minimize\ Q = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$

$$\frac{d(y^Ty)}{d\beta}=0,\quad \frac{d(2y^TX\beta)}{d\beta}=2(y^TX)^T=2X^Ty,\quad \frac{d(\beta^TX^TX\beta)}{d\beta}=2X^TX\beta$$

$$\left.\frac{dQ}{d\beta}\right|_{\beta=\hat\beta} = -2X^Ty+2X^TX\hat\beta=0 \qquad \boxed{\Rightarrow \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}^T\mathbf{y}}$$

- 最小二乘估计值(LSE) $\Rightarrow \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

[例] 简单线性回归

$$X^TX=\begin{bmatrix}1 & 1 & \cdots & 1\\ x_1 & x_2 & \cdots & x_n\end{bmatrix}\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}=\begin{bmatrix}n & \Sigma_{i=1}^n x_i\\ \Sigma_{i=1}^n x_i & \Sigma_{i=1}^n x_i^2\end{bmatrix},\quad X^Ty=\begin{bmatrix}\Sigma_{i=1}^n y_i\\ \Sigma_{i=1}^n x_iy_i\end{bmatrix}$$

$$\begin{bmatrix}n & \Sigma_{i=1}^n x_i\\ \Sigma_{i=1}^n x_i & \Sigma_{i=1}^n x_i^2\end{bmatrix}\begin{bmatrix}\hat\beta_0\\ \hat\beta_1\end{bmatrix}=\begin{bmatrix}\Sigma_{i=1}^n y_i\\ \Sigma_{i=1}^n x_iy_i\end{bmatrix}$$

$$\sum_{i=1}^n y_i = n\hat\beta_0+\hat\beta_1\sum_{i=1}^n x_i$$

$$\sum_{i=1}^n x_iy_i=\hat\beta_0\sum_{i=1}^n x_i+\hat\beta_1\sum_{i=1}^n x_i^2$$

41/52

41

4. 多重回归分析

[定理 14-9] 多重回归模型的最小二乘估计

다중회귀 모형: $y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki}+\varepsilon_i,\ i=1,2,\ldots,n$

$\Leftrightarrow \mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

최소제곱추정량(LSE): $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

추정 회귀식: $\hat{\mathbf{y}}=\mathbf{X}\hat{\boldsymbol{\beta}}$

잔차의 특성: $\mathbf{X}^T\mathbf{e}=\mathbf{0}$

$$\mathbf{X}^T\mathbf{e}=\mathbf{X}^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})=\mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$
$$=\mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{X}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right]=\mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{y}=\mathbf{0}$$

- 多重共线性(multicollinearity)
  - ✓ 自变量间存在高度相关关系的情况
  - ✓ 由于估计得到的回归系数的方差太大而导致精确的参数估计和检验存在困难，估计得到的回归模型的可靠性下降
  - ✓ 需要事先调查自变量之间的相关系数

42/52

42

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

[例 14-10] 某一高中欲进行影响语文成绩的要因分析
选取16名学生，收集其一周内的平均语文学习时间和阅读时间
期末语文考试成绩的计算结果 → 估计多重回归模型的回归系数

样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
学习时间	9	7	5	6	11	10	8	4	8	6	5	3	12	7	6	5
阅读时间	8	5	3	4	9	6	5	12	7	4	2	10	8	9	4	8
期末成绩	91	72	65	69	89	85	73	93	88	80	62	86	89	81	72	78

期末成绩=y，周平均学习时间=x_1，周平均阅读时间=x_2

- 多重回归模型 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,\ i = 1, 2, \ldots, n$

$$\Leftrightarrow \mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

$$y = \begin{bmatrix} 91 \\ 72 \\ \vdots \\ 78 \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},\ X = \begin{bmatrix} 1 & 9 & 8 \\ 1 & 7 & 5 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 8 \end{bmatrix}$$

43/52

43

4. 多重回归模型

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 9 & 7 & \cdots & 5 \\ 8 & 5 & \cdots & 8 \end{bmatrix} \begin{bmatrix} 1 & 9 & 8 \\ 1 & 7 & 5 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 8 \end{bmatrix} = \begin{bmatrix} 16 & 112 & 104 \\ 112 & 880 & 736 \\ 104 & 736 & 794 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 9 & 7 & \cdots & 5 \\ 8 & 5 & \cdots & 8 \end{bmatrix} \begin{bmatrix} 91 \\ 72 \\ \vdots \\ 78 \end{bmatrix} = \begin{bmatrix} 1273 \\ 9056 \\ 8624 \end{bmatrix}$$

최소제곱추정치 $\hat{\beta} = (X^T X)^{-1} X^T y$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 16 & 112 & 104 \\ 112 & 880 & 736 \\ 104 & 736 & 794 \end{bmatrix}^{-1} \begin{bmatrix} 1273 \\ 9056 \\ 8624 \end{bmatrix} \doteq \begin{bmatrix} 51.975 \\ 1.271 \\ 2.876 \end{bmatrix}$$

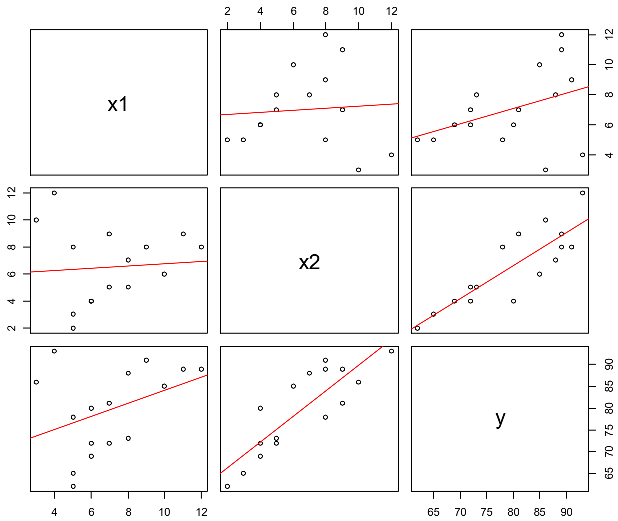
比较 $$\begin{bmatrix} 0.871271 & -0.068714 & -0.050426 \\ -0.068714 & 0.010476 & -0.000710 \\ -0.050426 & -0.000710 & 0.008523 \end{bmatrix} \begin{bmatrix} 1273 \\ 9056 \\ 8624 \end{bmatrix} \doteq \begin{bmatrix} 51.980 \\ 1.275 \\ 2.880 \end{bmatrix}$$

44/52

44

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

☞ 由于x1与x2之间的相关关系较弱，可以看出不存在多重共线性问题
☞ 可预测对于y的解释程度x2比x1更高



45/52

45

4. 多重回归分析

4.2 模型的拟合优度检验 (方差分析)

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
H_1: 모든 β_i가 0은 아니다. 즉, 적어도 한 개는 0이 아니다.

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n = \mathbf{y}^T\mathbf{y} - CT$$

$$CT \equiv (\sum_{i=1}^{n} y_i)^2 / n \Rightarrow \text{보정항}$$

$$SS_E = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - 2\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$$

$$\Rightarrow SS_E = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$$

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SS_T - SS_E = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - CT$$

46/52

46

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

[定理 14-10] 多重回归模型的平方和分解

총제곱합: $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n = \mathbf{y}^T\mathbf{y} - CT$

회귀제곱합: $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y} - CT$

잔차제곱합: $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$

구분	제곱합	자유도	제곱평균	검정통계량	기각역
회귀	SS_R	k	$MS_R = \frac{SS_R}{k}$	$\frac{MS_R}{MS_E}$	$F_{1-\alpha;(k,n-k-1)}$
잔차	SS_E	$n-k-1$	$MS_E = \frac{SS_E}{n-k-1}$		
합계	SS_T	$n-1$			

- 决定系数

$$R^2 = \frac{SS_R}{SS_T}$$

- 校正决定系数

$$R_{Adj}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

47/52

47

4. 多重回归分析

[例 14-11] 对[例 14-10]的多重回归模型进行方差分析，检验其拟合优度，并求其决定系数和校正决定系数。

$$SS_T = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 / n = 102,709 - 1273^2/16 \fallingdotseq 1425.938$$

$$SS_R = \hat{\beta}^T X^T y - CT \fallingdotseq [51.974964 \quad 1.270774 \quad 2.875710] \begin{bmatrix} 1273 \\ 9056 \\ 8624 \end{bmatrix} - \frac{1273^2}{16}$$

$$\fallingdotseq 102,472.382 - 101,283.063 = 1,189.319$$

$$SS_E = y^T y - \hat{\beta}^T X^T y \fallingdotseq 102,709 - 102,472.382 = 236.618$$

구분	제곱합	자유도	제곱평균	검정통계량	기각역
회귀	1189.319	2	594.6595	32.671	3.886
잔차	236.618	13	18.2014	→ 原假设被拒绝	
합계	1425.938	15			

$$R^2 = \frac{SS_R}{SS_T} \fallingdotseq \frac{1189.319}{1425.938} \fallingdotseq 0.834 \qquad R_{Adj}^2 \fallingdotseq 1 - \frac{236.618/13}{1425.938/15} \fallingdotseq 0.809$$

48/52

48

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

4. 多重回归分析

4.3 对回归系数的估计

$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$

检验统计量： $t_0 = \dfrac{\hat{\beta}_j}{S.E.(\hat{\beta}_j)} \qquad S.E.(\hat{\beta}_j) = \sqrt{MS_E \times c_{jj}}$

c_{jj}는 행렬 $(X'X)^{-1}$의 $j+1$번째 대각 원소

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

$$\boxed{\Rightarrow E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}}$$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\right] = E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \boxed{= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}} \end{aligned}$$

49/52

49

4. 多重回归分析

[例 14-12] 检验[例 14-10]的多重回归模型中各回归系数的显著性。

$MS_E \doteq 18.201,\ \hat{\beta}_0 \doteq 51.975,\ \hat{\beta}_1 \doteq 1.271,\ \hat{\beta}_2 \doteq 2.876$

$$(X^TX)^{-1} \doteq \begin{bmatrix} 0.871271 & -0.068714 & -0.050426 \\ -0.068714 & 0.010476 & -0.000710 \\ -0.050426 & -0.000710 & 0.008523 \end{bmatrix}$$

귀무가설 $H_0 : \beta_j = 0$, 대립가설 $H_1 : \beta_j \neq 0$에 대한 검정통계량

절편 $(j=0)$: $T_0 = \dfrac{\hat{\beta}_0}{S.E.(\hat{\beta}_0)} \doteq \dfrac{51.975}{\sqrt{18.201 \times 0.871271}} \doteq 13.052$

학습시간 $(j=1)$: $T_0 = \dfrac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} \doteq \dfrac{1.271}{\sqrt{18.201 \times 0.010476}} \doteq 2.911$

독서시간 $(j=2)$: $T_0 = \dfrac{\hat{\beta}_2}{S.E.(\hat{\beta}_2)} \doteq \dfrac{2.876}{\sqrt{18.201 \times 0.008523}} \doteq 7.302$

기각치는 $t_{0.975\,;\,13} \doteq 2.160$ 모든 회귀계수에 대해 귀무가설을 기각

50/52

50

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

5. 回归模型诊断

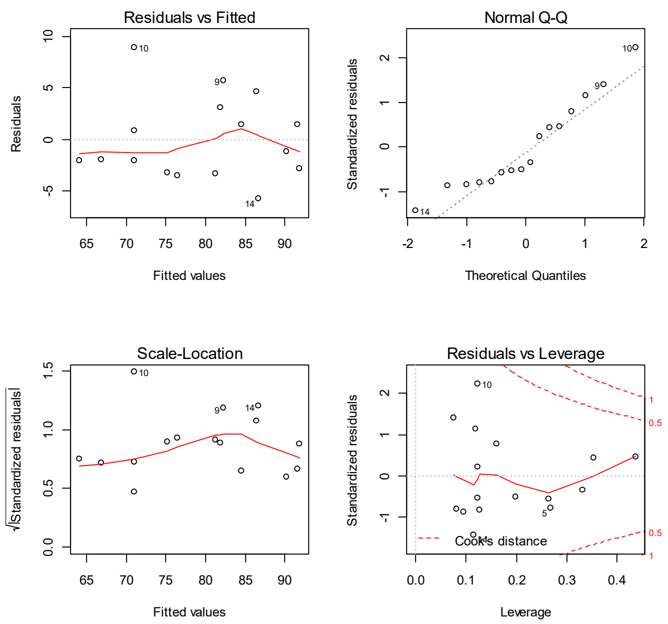
[例 14-13] 采用[例 14-10]的数据，比较下面的多重回归模型，并对所选取的模型进行诊断。

$$[모형1]\ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,\ i = 1, 2, \ldots, n$$

$$[모형2]\ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i,\ i = 1, 2, \ldots, n$$

51/52

51



52/52

52