פרק ב-5

הקשר של ניתוח שונות למודל רגרסיה ליניארית.

המודלים

מודל ניתוח שונות חד-כווני

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \tilde{\varepsilon}_{ij}$$
$$i = 1,\dots,a$$
$$j = 1,\dots,n_i$$
$$\sum_{i=1}^{a} \tau_i = 0$$
$$\tilde{\varepsilon}_{ij} \sim NID(0,\sigma^2)$$

$$N = \sum_{i=1}^{a} n_i$$

נזכר במודל רגרסיה לינארית

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \dots + \beta_{p-1} x_{(p-1)k} + \tilde{\varepsilon}_k$$
$$\tilde{\varepsilon}_k \sim NID(0,\sigma^2)$$
$$k = 1,\dots,N$$

או בכתיב מטריציוני

$$Y = X\beta + \tilde{\varepsilon}$$

$$Y_{N\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad X_{N\times p} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{(p-1)1} \\ 1 & x_{12} & x_{22} & \cdots & x_{(p-1)2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1N} & x_{2N} & \cdots & x_{(p-1)N} \end{pmatrix} \quad \beta_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \tilde{\varepsilon}_{N\times 1} = \begin{pmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_N \end{pmatrix}$$

$$\tilde{\varepsilon}_k \sim NID(0,\sigma^2)$$

שימושי המודלים

מודל ניתוח שונות חד-כווני

$$\hat{\mu}_i = \bar{y}_{i\bullet}$$

$$\begin{cases} H_0: & \tau_1 = \tau_2 = \ldots = \tau_a = 0 \\ H_1: & otherwise \end{cases}$$

- אמידה של תוחלות האוכלוסיות
- הסקה לגבי שוני בתוחלות האוכלוסיות
- מבחנים ורווחי סמך נוספים (השוואות מרובות).

מודל רגרסיה ליניארית

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$$\hat{y}(x_1,\ldots,x_{p-1}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_{p-1} x_{p-1}$$

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_1: & otherwise \end{cases}$$

- אמידה של מקדמי הרגרסיה
- חיזוי של המשתנה המוסבר.
- הסקה לגבי מובהקות מודל הרגרסיה.
- מבחנים ורווחי סמך נוספים.

מודל רגרסיה לדוגמא

נסתכל על דוגמא בעלת המימדים הללו:

$$N = 9$$
$$p = 3$$

מודל הרגרסיה

$$\begin{aligned} & y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \tilde{\varepsilon}_k \\ & \tilde{\varepsilon}_k \sim NID(0,\sigma^2) \\ & k = 1,\dots,9 \end{aligned}$$

או בכתיב מטריציוני

$$Y = X\beta + \tilde{\varepsilon}$$

$$Y_{9\times1} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{pmatrix} \quad X_{9\times3} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \\ 1 & x_{15} & x_{25} \\ 1 & x_{16} & x_{26} \\ 1 & x_{17} & x_{27} \\ 1 & x_{18} & x_{28} \\ 1 & x_{19} & x_{29} \end{pmatrix} \quad \beta_{3\times1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \tilde{\varepsilon}_{N\times1} = \begin{pmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \tilde{\varepsilon}_3 \\ \tilde{\varepsilon}_4 \\ \tilde{\varepsilon}_5 \\ \tilde{\varepsilon}_6 \\ \tilde{\varepsilon}_7 \\ \tilde{\varepsilon}_8 \\ \tilde{\varepsilon}_9 \end{pmatrix}$$

$$\tilde{\varepsilon}_k \sim NID(0,\sigma^2) \quad k = 1,\dots,9$$

נניח נתונים עבור הדוגמא

נתונים

$$Y_{9\times1}=\begin{pmatrix}17.3\\15.4\\13.1\\21.2\\23.2\\19.5\\18.8\\3.9\\5.7\end{pmatrix}\qquad X_{9\times3}=\begin{pmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\\1&0&1\\1&0&0\\1&0&0\end{pmatrix}$$

בחרנו את כל המשתנים המסבירים
להיות בעלי ערכים 0 או 1 (דיכוטומים)

על פי הנתונים, להלן המודל:

$$\begin{aligned}
17.3 &= y_1 = \beta_0 + \beta_1 + \hat{\varepsilon}_1\\
15.4 &= y_2 = \beta_0 + \beta_1 + \hat{\varepsilon}_2\\
13.1 &= y_3 = \beta_0 + \beta_1 + \hat{\varepsilon}_3\\
21.2 &= y_4 = \beta_0 + \beta_2 + \hat{\varepsilon}_4\\
23.2 &= y_5 = \beta_0 + \beta_2 + \hat{\varepsilon}_5\\
19.5 &= y_6 = \beta_0 + \beta_2 + \hat{\varepsilon}_6\\
18.8 &= y_7 = \beta_0 + \beta_2 + \hat{\varepsilon}_7\\
3.9 &= y_8 = \beta_0 + \hat{\varepsilon}_8\\
5.7 &= y_9 = \beta_0 + \hat{\varepsilon}_9
\end{aligned}$$

מודל ניתוח שונות ונתונים לדוגמא

נסתכל על דוגמא בעלת המימדים הללו:

$$a = 3$$
$$n_1 = 3$$
$$n_2 = 4$$
$$n_3 = 2$$
$$N = \sum_{i=1}^{3} n_i = 9$$

מודל ניתוח שונות

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \tilde{\varepsilon}_{ij}$$
$$i = 1,2,3$$
$$j = 1,\ldots,n_i$$
$$\sum_{i=1}^{a} \tau_i = 0$$
$$\tilde{\varepsilon}_{ij} \sim NID(0,\sigma^2)$$

נתונים

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} & \cdot \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} 17.3 & 15.4 & 13.1 & \cdot \\ 21.2 & 23.2 & 19.5 & 18.8 \\ 3.9 & 5.7 & \cdot & \cdot \end{pmatrix}$$

על פי הנתונים, להלן המודל:

$$17.3 = y_{11} = \mu_1 + \hat{\varepsilon}_{11}$$
$$15.4 = y_{12} = \mu_1 + \hat{\varepsilon}_{12}$$
$$13.1 = y_{13} = \mu_1 + \hat{\varepsilon}_{13}$$
$$21.2 = y_{21} = \mu_2 + \hat{\varepsilon}_{21}$$
$$23.2 = y_{22} = \mu_2 + \hat{\varepsilon}_{22}$$
$$19.5 = y_{23} = \mu_2 + \hat{\varepsilon}_{23}$$
$$18.8 = y_{24} = \mu_2 + \hat{\varepsilon}_{24}$$
$$3.9 = y_{31} = \mu_3 + \hat{\varepsilon}_{31}$$
$$5.7 = y_{32} = \mu_3 + \hat{\varepsilon}_{32}$$

נשווה בין שני המודלים

תצפיות של משתנים מוסברים

תוחלת של אוכלוסיה 1

תוחלת של אוכלוסיה 2

תוחלת של אוכלוסיה 3

שגיאות

מודל הרגרסיה

$$
\begin{aligned}
17.3 &= y_1 = \beta_0 + \beta_1 + \hat{\varepsilon}_1\\
15.4 &= y_2 = \beta_0 + \beta_1 + \hat{\varepsilon}_2\\
13.1 &= y_3 = \beta_0 + \beta_1 + \hat{\varepsilon}_3\\
21.2 &= y_4 = \beta_0 + \beta_2 + \hat{\varepsilon}_4\\
23.2 &= y_5 = \beta_0 + \beta_2 + \hat{\varepsilon}_5\\
19.5 &= y_6 = \beta_0 + \beta_2 + \hat{\varepsilon}_6\\
18.8 &= y_7 = \beta_0 + \beta_2 + \hat{\varepsilon}_7\\
3.9 &= y_8 = \beta_0 + \hat{\varepsilon}_8\\
5.7 &= y_9 = \beta_0 + \hat{\varepsilon}_9
\end{aligned}
$$

מודל ניתוח שונות

$$
\begin{aligned}
17.3 &= y_{11} = \mu_1 + \hat{\varepsilon}_{11}\\
15.4 &= y_{12} = \mu_1 + \hat{\varepsilon}_{12}\\
13.1 &= y_{13} = \mu_1 + \hat{\varepsilon}_{13}\\
21.2 &= y_{21} = \mu_2 + \hat{\varepsilon}_{21}\\
23.2 &= y_{22} = \mu_2 + \hat{\varepsilon}_{22}\\
19.5 &= y_{23} = \mu_2 + \hat{\varepsilon}_{23}\\
18.8 &= y_{24} = \mu_2 + \hat{\varepsilon}_{24}\\
3.9 &= y_{31} = \mu_3 + \hat{\varepsilon}_{31}\\
5.7 &= y_{32} = \mu_3 + \hat{\varepsilon}_{32}
\end{aligned}
$$

קבלנו:

$$
\boxed{\begin{aligned}
\mu_1 &= \beta_0 + \beta_1\\
\mu_2 &= \beta_0 + \beta_2\\
\mu_3 &= \beta_0
\end{aligned}}
$$

בעצם נעזרנו במשתנים דיכוטומים במודל הרגרסיה....

- בעצם ניסחנו את בעיית ניתוח השונות כבעיית רגרסיה ע"י שימוש במשתנים דיכוטומים במודל הרגרסיה.
- עבור הטיפולים: 1,2,...,a-1 ,
- הגדרנו את המשתנים:

$$D_{1,k}=\begin{cases}1 & Treatment\,1\,in\,use\\0 & otherwise\end{cases}$$

$$D_{2,k}=\begin{cases}1 & Treatment\,2\,in\,use\\0 & otherwise\end{cases}$$

$$\vdots$$

$$D_{a-1,k}=\begin{cases}1 & Treatment\,a-1\,in\,use\\0 & otherwise\end{cases}$$

$$\boxed{p=a}$$

- והשתמשנו במודל הרגרסיה:

$$y_k=\beta_0+\beta_1 D_{1k}+\beta_2 D_{2k}+\ldots+\beta_{a-1}D_{a-1,k}+\tilde{\varepsilon}_k$$
$$\tilde{\varepsilon}_k\sim NID(0,\sigma^2)$$
$$k=1,\ldots,N$$

ובאופן מטריציוני...

$$
Y_{N\times 1}=\begin{pmatrix} y_{11}\\ \vdots \\ y_{1n_1} \\ -- \\ y_{21} \\ \vdots \\ y_{2n_2} \\ -- \\ \vdots \\ \vdots \\ \vdots \\ -- \\ y_{a1} \\ \vdots \\ y_{an_a} \end{pmatrix}
\qquad
X_{N\times a}=\begin{pmatrix}
1 & 1 & 0 & \cdots & \cdots & 0 \\
\vdots & & & \vdots & & \\
\vdots & 1 & 0 & \cdots & \cdots & 0 \\
\vdots & -- & -- & -- & -- & -- \\
\vdots & 0 & 1 & 0 & \cdots & 0 \\
\vdots & & & \vdots & & \\
\vdots & 0 & 1 & 0 & \cdots & 0 \\
\vdots & -- & -- & -- & -- & -- \\
\vdots & & & & & \\
\vdots & & & \vdots & & \\
\vdots & & & & & \\
\vdots & -- & -- & -- & -- & -- \\
\vdots & 0 & 0 & \cdots & 0 & 1 \\
\vdots & & & \vdots & & \\
\vdots & 0 & 0 & \cdots & 0 & 1 \\
\vdots & -- & -- & -- & -- & -- \\
\vdots & 0 & & \cdots & & 0 \\
\vdots & & & \vdots & & \\
1 & 0 & & \cdots & & 0
\end{pmatrix}
$$

אמידה וחיזוי

$$\boxed{p = a}$$

בדוגמא:

$$\mu_1 = \beta_0 + \beta_1$$
$$\mu_2 = \beta_0 + \beta_2$$
$$\mu_3 = \beta_0$$

$$\hat{\mu}_1 = \widehat{\beta_0 + \beta_1} = \bar{y}_{1\bullet}$$
$$\hat{\mu}_2 = \widehat{\beta_0 + \beta_2} = \bar{y}_{2\bullet}$$
$$\hat{\mu}_3 = \hat{\beta}_0 = \bar{y}_{3\bullet}$$
$$\hat{y}(x_1, x_2) = \begin{cases} \bar{y}_{1\bullet} & (x_1, x_2) = (1,0) \\ \bar{y}_{2\bullet} & (x_1, x_2) = (0,1) \\ \bar{y}_{3\bullet} & (x_1, x_2) = (0,0) \end{cases}$$

באופן כללי:

$$\mu_i = \beta_0 + \beta_i \quad i = 1, \dots, p-1$$
$$\mu_a = \beta_0$$

$$\hat{\mu}_i = \widehat{\beta_0 + \beta_i} = \bar{y}_{i\bullet} \quad i = 1, \dots, p-1$$
$$\hat{\mu}_a = \hat{\beta}_0 = \bar{y}_{a\bullet}$$
$$\hat{y}(x_1, \dots, x_{p-1}) = \begin{cases} \bar{y}_{1\bullet} & (x_1, \dots, x_{p-1}) = (1, 0, \dots, 0) \\ \bar{y}_{2\bullet} & (x_1, \dots, x_{p-1}) = (0, 1, \dots, 0) \\ \vdots & \\ \bar{y}_{(a-1)\bullet} & (x_1, \dots, x_{p-1}) = (0, \dots, 1) \\ \bar{y}_{a\bullet} & (x_1, \dots, x_{p-1}) = (0, \dots, 0) \end{cases}$$

פרוק סכום ריבועים

ניתוח שונות:

$$\overset{N-1}{\overbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2}^{SS_{Total}}} = \overset{a-1}{\overbrace{\sum_{i=1}^{a}n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2}^{SS_{Treatments}}} + \overset{N-a}{\overbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2}^{SS_E}}$$

רגרסיה:

$$\underset{N-1}{\underbrace{\sum_{i=1}^{N}(y_i-\bar{y})^2}_{SS_T}} = \underset{p-1}{\underbrace{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}_{SS_{reg}}} + \underset{N-p}{\underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}_{SS_E}}$$

ולכן טבלאות ניתוח השונות של מודל הרגרסיה המתאים
למודל ניתוח שונות זהות לטבלאות של מודל ניתוח השונות.

הדוגמא ב SAS

- דגשים:
  - טבלאות ניתוח השונות הן זהות
  - RootMSE, R-Square, Mean, CV הם זהים.
  - אמדי קו הרגרסיה נותנים את הממוצעים:

$$\boxed{\begin{aligned}\mu_1 &= \beta_0 + \beta_1\\ \mu_2 &= \beta_0 + \beta_2\\ \mu_3 &= \beta_0\end{aligned}}$$

מסקנות

- מודל ניתוח שונות הוא מקרה פרטי של מודל הרגרסיה.
  - עבור a אוכלוסיות בניתוח שונות, ניתן להגדיר מודל רגרסיה עם a-1 משתנים מסבירים (זהו מודל רגרסיה ובו $p=a$).
  - כל התצפיות מכל קבוצת אוכלוסיה במודל ניתוח השונות יתוארו ע"י התצפיות של המשתנים המוסברים במודל הרגרסיה.
  - למשתנים המסבירים יותאמו משתנים קטגורים.
- אמדי הריבועים הפחותים של מודל הרגרסיה תואם את האמדים של מודל ניתוח שונות.
- טבלת ניתוח השונות היא בעצם אותה טבלה.