22 Oct

# Chapter 2 Categorical variables

Model $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$ $\qquad \underset{\sim}{e} \overset{\text{Assumption}}{\sim} N(0, \sigma^2 \underset{\sim}{I})$

$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{Y}$ $\qquad \hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2 I (\underset{\sim}{X}^T\underset{\sim}{X})^{-1})$

$\hat{\sigma}^2 = \frac{\text{Res S.S.}}{n-p'}$ $\qquad \frac{\text{Res S.S.}}{\sigma^2} \sim \chi^2(n-p')$ $\qquad$ Chapter 1

where $\text{Res S.S.} = \underset{\sim}{Y}^T(\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{X}^T)\underset{\sim}{Y}$

$\Rightarrow$ C.I. of $\underset{\sim}{\beta}$, t-test, F-test

## One categorical variable (One-way ANOVA)

e.g. p.4

Fit y on group = 1, 2, 3, 4, 5

Model $y = \beta_0 + \beta_1 X + e$ $\leftarrow$ Lack-of-fit

| group | Model | E(y) | |
|---|---|---|---|
| 1 | $y=\beta_0+\beta_1+e$ | $\beta_0+\beta_1=\mu_1$ | |
| 2 | $y=\beta_0+2\beta_1+e$ | $\beta_0+2\beta_1=\mu_2$ | $\beta_1$ |
| 3 | $y=\beta_0+3\beta_1+e$ | $\beta_0+3\beta_1=\mu_3$ | $\beta_1$ |
| 4 | $y=\beta_0+4\beta_1+e$ | $\beta_0+4\beta_1=\mu_4$ | $\beta_1$ |
| 5 | $y=\beta_0+5\beta_1+e$ | $\beta_0+5\beta_1=\mu_5$ | $\beta_1$ |

$\mu_2-\mu_1=\mu_3-\mu_2=\mu_4-\mu_3=\mu_5-\mu_4$

$\Downarrow$

It is not a general model

## General model

Dummy variable (indicator variable)

| group | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 |

group $= i \Rightarrow g_i = 1$ $\qquad i = 1, 2, 3, 4, 5$

$$
\underset{n\times \cancel{6}\ 5}{\underset{\sim}{X}} =
\begin{array}{c}
\begin{array}{cccccc} \beta_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \end{array} \\
\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & \cancel{0} \\
1 & 0 & 1 & 0 & 0 & \cancel{0} \\
1 & 0 & 0 & 1 & 0 & \cancel{0} \\
1 & 0 & 0 & 0 & 1 & \cancel{0} \\
1 & 0 & 0 & 0 & 0 & \cancel{1}
\end{pmatrix}
\end{array}
\begin{array}{l} \text{group} = 1 \\ \text{group } 2 \\ \text{group} = 3 \\ \text{group} = 4 \\ \text{group} = 5 \end{array}
$$

# of levels of group $= \cancel{5}\ m$

$\Rightarrow$ create $(\cancel{5}\ m - 1)$ dummy variables

$\det(\underset{\sim}{X}^T \underset{\sim}{X}) = 0$

$\underset{\sim}{X}$ is singular matrix

we need to add one constraint, e.g. delete the last column

## $\boxed{\text{Model I}}$ (Regression model)

$$y_i = \beta_0 + \alpha_1 g_{i1} + \alpha_2 g_{i2} + \alpha_3 g_{i3} + \alpha_4 g_{i4} + \cdots + \alpha_{m-1} g_{i,m-1} + e_i \qquad i = 1, \text{—}, n$$

$\alpha_k$ – regression coeff. of $g_k$ $\qquad k = 1, 2, \cdots, m-1$

Analysis of variance

## ANOVA ~~tat~~ model

$$y_{ij} = \beta_0 + \alpha_i + e_{ij} \qquad i = 1, \text{—}, m \quad j = 1, \text{—}, n_i$$

| Group | Model | | |
|---|---|---|---|
| 1 | $E(y_{1j}) = \beta_0 + \alpha_1 + \cancel{e_{1j}}$ | $\mu_1$ | $j = 1, \cdots, n_1$ |
| 2 | $E(y_{2j}) = \beta_0 + \alpha_2 + \cancel{e_{2j}}$ | $\mu_2$ | $j = 1, \text{—}, n_2$ |
| 3 | $E(y_{3j}) = \beta_0 + \alpha_3 + \cancel{e_{3j}}$ | $\mu_3$ | $j = 1, \text{—}, n_3$ |
| 4 | $E(y_{4j}) = \beta_0 + \alpha_4 + \cancel{e_{4j}}$ | $\mu_4$ | $j = 1, \text{—}, n_4$ |
| 5 | $E(y_{5j}) = \beta_0 + \cancel{\alpha_5} + \cancel{e_{5j}}$ | $\mu_5$ | $j = 1, \text{—}, n_5$ |

$$\Rightarrow \quad y_{ij} = \mu_i + e_{ij} \qquad i = 1, \text{—}, m \quad j = 1, \text{—}, n_i$$

$$\begin{aligned}
\beta_0 + \alpha_1 &= \mu_1 \\
\beta_0 + \alpha_2 &= \mu_2 \\
\beta_0 + \alpha_3 &= \mu_3 \\
\beta_0 + \alpha_4 &= \mu_4 \\
\beta_0 &= \mu_5
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
\alpha_1 &= \mu_1 - \mu_5 \\
\alpha_2 &= \mu_2 - \mu_5 \\
\alpha_3 &= \mu_3 - \mu_5 \\
\alpha_4 &= \mu_4 - \mu_5
\end{aligned}$$

group = 5 is a reference group

$\boxed{\text{model without intercept}}$

Model:

$$y_{ij} = \mu_i + e_{ij}$$

$i = 1, \ldots, \cancel{m}\ 5$

$j = 1, \ldots, \cancel{n_i}\ 6$

$$\underset{n \times 5}{\underset{\sim}{X}} = \begin{array}{c}
\begin{array}{ccccc} \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{array} \\
\left(\begin{array}{ccccc}
1 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 0 & 0 \\ \hline
0 & 1 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & 0 & 0 & 0 \\ \hline
0 & 0 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 1 & 0 & 0 \\ \hline
0 & 0 & 0 & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 1 & 0 \\ \hline
0 & 0 & 0 & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 1
\end{array}\right)
\end{array}
\begin{array}{l}
\text{group}=1 \\ \\ \text{group}=2 \\ \\ \text{group}=3 \\ \\ \text{group}=4 \\ \\ \text{group}=5
\end{array}$$

$$\underset{\sim}{Y} = \left(\begin{array}{c}
y_{11} \\ \vdots \\ y_{1n_1} \\ \hline y_{21} \\ \vdots \\ y_{2n_2} \\ \hline \vdots \\ \hline \vdots \\ \hline y_{51} \\ \vdots \\ y_{5n_5}
\end{array}\right)
\begin{array}{l} \text{group}=1 \\ \\ \text{group}=2 \\ \\ \text{group}=3 \\ \\ \text{group}=4 \\ \\ \text{group}=5 \end{array}
= \left(\begin{array}{c}
551 \\ \vdots \\ 632 \\ \hline 595 \\ \vdots \\ 517 \\ \hline \vdots \\ \hline 563 \\ \vdots \\ 679
\end{array}\right)
\begin{array}{l} \text{group}=1 \\ \\ \text{group}=2 \\ \\ \\ \text{group}=5 \end{array}$$

$$\underset{5 \times 5}{\underset{\sim}{X}^T \underset{\sim}{X}} = \begin{pmatrix}
n_1 & 0 & 0 & 0 & 0 \\
0 & n_2 & 0 & 0 & 0 \\
0 & 0 & n_3 & 0 & 0 \\
0 & 0 & 0 & n_4 & 0 \\
0 & 0 & 0 & 0 & n_5
\end{pmatrix}$$

$$(\underset{\sim}{X}^T \underset{\sim}{X})^{-1} = \begin{pmatrix}
\frac{1}{n_1} & 0 & 0 & 0 & 0 \\
0 & \frac{1}{n_2} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{n_3} & 0 & 0 \\
0 & 0 & 0 & \frac{1}{n_4} & 0 \\
0 & 0 & 0 & 0 & \frac{1}{n_5}
\end{pmatrix}$$

$$\underset{\sim}{X}^T \underset{\sim}{Y} = \begin{pmatrix}
\sum_{j=1}^{n_1} y_{1j} \\
\sum_{j=1}^{n_2} y_{2j} \\
\vdots \\
\sum_{j=1}^{n_5} y_{5j}
\end{pmatrix}$$

$$\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{Y} = \begin{pmatrix}
\bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \bar{y}_{3\cdot} \\ \bar{y}_{4\cdot} \\ \bar{y}_{5\cdot}
\end{pmatrix}$$

$$\hat{\underset{\sim}{\beta}} = \begin{pmatrix}
\hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \\ \hat{\mu}_5
\end{pmatrix}$$

$$\hat{\mu}_i = \bar{y}_{i\cdot} \quad i = 1, \ldots, m$$

Regression model

$$y_i = \beta_0 + \alpha_1 g_{i1} + \alpha_2 g_{i2} + \alpha_3 g_{i3} + \alpha_4 g_{i4} + e_i$$

ANOVA model

$$y_{ij} = \mu_i + e_{ij} \qquad i = 1, \text{—}, 5$$
$$j = 1, \text{—}, n_i$$

$$\boxed{\begin{aligned} \hat\beta_0 + \hat\alpha_1 &= \hat\mu_1 \\ \hat\beta_0 + \hat\alpha_2 &= \hat\mu_2 \\ \hat\beta_0 + \hat\alpha_3 &= \hat\mu_3 \\ \hat\beta_0 + \hat\alpha_4 &= \hat\mu_4 \\ \hat\beta_0 &= \hat\mu_5 \end{aligned}} \begin{aligned} &= \bar y_{1\cdot} \\ &= \bar y_{2\cdot} \\ &= \bar y_{3\cdot} \\ &= \bar y_{4\cdot} \\ &= \bar y_{5\cdot} \end{aligned}$$

$$\Rightarrow \hat\beta_0 = \bar y_{5\cdot}, \quad \hat\alpha_1 = \bar y_{5\cdot} - \bar y_{1\cdot}, \quad \cdots$$

---

$$\text{Res S.S.} = \underset{\sim}{Y}^T \underset{\sim}{Y} - \hat{\underset{\sim}{\beta}}^T \underset{\sim}{X}^T \underset{\sim}{Y}$$

$$= \sum_{i=1}^{m} \boxed{\sum_{j=1}^{n_i} y_{ij}^2} - (\boxed{\bar y_1}, \bar y_2, \cdots, \bar y_m) \begin{pmatrix} \boxed{\sum_{j=1}^{n_1} y_{ij}} \\ \vdots \\ \sum_{j=1}^{n_m} y_{mj} \end{pmatrix}$$

For $i = 1$

$$\sum_{j=1}^{n_1} y_{1j}^2 - \bar y_{1\cdot} \sum_{j=1}^{n_1} y_{1j} \longleftarrow n_1 \bar y_1$$

$$= \sum_{j=1}^{n_1} y_{1j}^2 - n \bar y_{1\cdot}^2$$

$$= \sum_{j=1}^{n_1} (y_{1j} - \bar y_{1\cdot})^2$$

Pure Error S.S.

$$\Rightarrow \text{Res S.S.} = \boxed{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar y_i)^2}$$

MATH 3423

$$\frac{\sum_{j=1}^{n_i} (y_{ij} - \bar y_i)^2}{\sigma^2} \sim \chi^2_{(n_i - 1)}$$

$$\frac{\text{Res S.S.}}{\sigma^2} \sim \chi^2_{\underbrace{\sum_{i=1}^{m}(n_i - 1)}_{n-m}} \Rightarrow \hat\sigma^2 = \frac{\text{Res S.S}}{n-m} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar y_i)^2}{n-m}$$

$$Var(\hat{\beta}) = (X^T X)^{-1} \sigma^2$$

$$\beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{n_3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{n_4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{n_5} \end{pmatrix} \sigma^2$$

$$Var(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$$

$$\Rightarrow Var(\bar{y}_i) = \frac{\sigma^2}{n_i}$$

C.I. for $\mu_i$

$$\bar{y}_{i\cdot} \pm t_{\alpha/2,\, n-m}\ \hat{\sigma} \sqrt{\frac{1}{n_i}}$$

$H_0: \mu_i = \mu_{i0}$

$$t = \frac{\bar{y}_{i\cdot} - \mu_{i0}}{\hat{\sigma}/\sqrt{n_i}}$$

---

Fitted model: $y$ on group $= 1, 2, 3, 4, 5$

$$y_i = \beta_0 + \beta_1 * \underset{\substack{\uparrow \\ 1,2,3,4,5}}{group_i} + e \qquad i = 1, -, n$$

Section 6 of Chapter 1

$H_0$: No lack of fit

$H_1$:

Res S.S. | fitted model − Pure Error S.S.

d.f. of Res S.S. | fitted model $\underset{n-2}{}$ − d.f. of Pure Error S.S. $\underset{n-m}{}$

$$F = \frac{\boxed{\text{lack of fit S.S.}} / \boxed{\text{d.f} \mid \text{lack of fit S.S.}}}{\boxed{\text{Pure Error S.S}} / \boxed{\text{d.f} \mid \text{pure error S.S}}}$$

Pure Error S.S $= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 =$ Res S.S. | ANOVA

d.f | pure error S.S $= n - m$

$H_0 = \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ ⟵ ANOVA model

vs $H_1$ : at least ~~one~~ two of $\mu$'s are not equal

$y_{ij} = \mu_i + e_{ij}$

Under $H_0$, $y_{ij} = \mu + e_{ij}$

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{m-1} = 0$ ⟵ Regression model

$y_i = \beta_0 + \alpha_1 g_{i1} + \cdots + \alpha_{m-1} g_{i,m-1} + e_i$

Under $H_0$, $y_i = \beta_0 + e_i$

Section 4 of Chapter 1

Total S.S. = Reg S.S. + Res S.S.

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$$

overall mean of $y_{ij}$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2}_{= \sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}$$

$\frac{\text{Total S.S}}{\sigma^2} \sim \chi^2(n-1, \lambda)$

$\frac{\text{Res S.S.}}{\sigma^2} \sim \chi^2(n-m)$

$\frac{\text{Reg S.S.}}{\sigma^2} \sim \chi^2(m-1, \lambda)$

$$\Rightarrow F = \frac{\text{Reg S.S.}/(m-1)}{\text{Res S.S.}/(n-m)} \overset{\text{under } H_0}{\sim} F(m-1, n-m)$$

---

$$\text{Reg S.S.} = \sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$N = \sum_{i=1}^{m} n_i$ = total # of observations

$$= \sum_{i=1}^{m} n_i\left(\frac{T_{i.}}{n_i} - \frac{T_{..}}{N}\right)^2$$

$$= \sum_{i=1}^{m} n_i\left(\frac{T_{i.}^2}{n_i^2} + \frac{T_{..}^2}{N^2} - 2\frac{T_{i.}}{n_i}\frac{T_{..}}{N}\right)$$

$$= \sum_{i=1}^{m}\frac{T_{i.}^2}{n_i} + \frac{T_{..}^2}{n^2}\underbrace{\sum_{i=1}^{m} n_i}_{n} - 2\frac{T_{..}}{N}\underbrace{\boxed{\sum_{i=1}^{m} \cancel{n_i} \times \frac{T_{i.}}{\cancel{n_i}}}}_{= T_{..}}$$

$$= \sum_{i=1}^{m}\frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$$

$H_0: \mu_1 = \mu_2 = \cdots = \mu_m = \mu \Leftrightarrow H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_{m-1} = 0$

| Source of Variation of | Sum of Squares | Degrees of freedom | Mean Square | Computed $f$ |
|---|---|---|---|---|
| Model | $\sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ | $m-1$ | $\frac{\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{m-1}$ | $\frac{(\sum_{i=1}^{m} n_i - m)\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{(m-1)\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}$ |
| Error | $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$ | $\underbrace{\sum_{i=1}^{m} n_i}_{n} - m$ | $\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}{\sum_{i=1}^{m} n_i - m}$ | |
| Total | $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$ | $\sum_{i=1}^{m} n_i - 1$ | | |

The advantages of choosing equal sample sizes over the choice of unequal sample sizes are: 1) the $f$ ratio is insensitive to slight departures from the assumption of equal variances for the $m$ populations when the sample are of equal sizes; and 2) the choice of equal sample size minimizes the probability of committing a type II error.

**Example**

$X_1$ categorical variable

$S_i^2$ – sample variance of $y_{ij}$ in the ith group

$S_T^2$ sample variances of all $y_{ij}$

repeated measurements

| | Group | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| | 551 ($y_{11}$) | 595 | 639 | 417 | 563 ($y_{51}$) | |
| | 457 ($y_{12}$) | 580 | 615 | 449 | 631 | |
| | 450 | 508 | 511 | 517 | 522 | |
| | 731 | 583 | 573 | 438 | 613 | |
| | 499 | 633 | 648 | 415 | 656 | |
| | 632 ($y_{16}$) | 517 | 677 | 555 | 679 ($y_{56}$) | |
| Total | 3320 | 3416 | 3663 | 2791 | 3664 | 16854 |
| Mean | 553.33 | 569.33 | 610.50 | 465.19 | 610.67 | 561.80 |

$S_{yy}^{(i)}$ = Sum of squares of $y_{ij}$ in the ith group

**Solution**

$S_1^2 = 12,133.8667$, $S_2^2 = 2,302.6667$, $S_3^2 = 3593.5$, $S_4^2 = 3,318.5667$, $S_5^2 = 3,455.4667$,
$S_T^2 = 7,219.8897$

$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{m}(n_i - 1)S_i^2$

| Source of Variation of | Sum of Squares | Degrees of freedom | Mean Square | Computed $f$ |
|---|---|---|---|---|
| Group | 85,356 | 4 | 21,339 | $\boxed{4.30}$ |
| Error | 124,021 | 25 $= \sum_{i=1}^{m}(n_i - 1) = n - m$ | 4,961 | |
| Total | 209,377 | 29 | | |

① Reg S.S. = Total S.S. − Res S.S.

② $\sum_{i=1}^{m} \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{n}$

30 − 1

$\frac{21339}{4961}$

C.V. $= F_{0.05,4,25} = 2.76$

The critical value $f_{0.05}(4, 25) = 2.76$. Thus, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ is rejected.

$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = (n-1)S_T^2 = S_{yy}$

overall sample size

⇒ Reject $H_0$

What is the reason of rejecting the null hypothesis?