Nov 24

2. Two categorical variables

1) Models

Model I (Regression Model)

Categorical variable I (Factor A) a levels $\Rightarrow$ $(a-1)$ dummy variables Group

Categorical variable II (Factor B) b levels $\Rightarrow$ $(b-1)$ dummy variables Class

$$y_k = \beta_0 + \sum_{i=1}^{a-1} \alpha_i * g_{i,k} + \sum_{j=1}^{b-1} \beta_j * c_{j,k} + \sum_{i=1}^{a-1}\sum_{j=1}^{b-1} \gamma_{ij} * g_{i,k} * c_{j,k} + e_k$$

for $k = 1, \ldots, N$, where $g_{i,k} = 1$ if k^{th} observation is in i^{th} level of Factor A and $g_{i,k} = 0$ otherwise; $c_{j,k} = 1$ if k^{th} observation is in j^{th} level of Factor B and $c_{j,k} = 0$ otherwise.

Model II (ANOVA)

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n_{ij}$, $\alpha_a = 0$, $\beta_b = 0$, $\gamma_{aj} = 0$ for each j, $\gamma_{ib} = 0$ for each i.

Re-parameterization of Model II

$$\Rightarrow y_{ijk} = \mu_{ij} + e_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, n_{ij}$

Inference (model II). $\begin{cases} \text{Estimation} \\ \text{HT.} \end{cases}$ unknown param: μ_{ij}, σ^2.

2. Est.

① Point est.

$$\hat{\mu}_{ij} = \bar{y}_{ij\cdot}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\cdot})^2}{\sum_{i=1}^{a}\sum_{j=1}^{b} (n_{ij} - 1)}$$

Properties:

$$E\hat{\mu}_{ij} = \mu_{ij} \qquad Var\,\hat{\mu}_{ij} = \frac{\sigma^2}{n_{ij}}, \quad Cov(\hat{\mu}_{ij}, \hat{\mu}_{kl}) = 0, \ i \neq k, j \neq l.$$

$$E\hat{\sigma}^2 = \sigma^2.$$

② $(1-\alpha)$ C.I. for μ_{ij}: $\bar{y}_{ij\cdot} \mp t_{\frac{\alpha}{2}}(N-ab)\,\hat{\sigma}\sqrt{\frac{1}{n_{ij}}}$.

3. HT

① single param

$H_0: \mu_{ij} = \mu_{ij0}$ (given)

Test stat $\quad t = \dfrac{\bar{y}_{ij\cdot} - \mu_{ij0}}{\hat{\sigma}/\sqrt{n_{ij}}}$.

Reject H_0 if $|t_{obs}| > t_{\frac{\alpha}{2}}(N-ab)$

② multi-param

- $H_0: \mu_{ij} = \mu \quad (\Leftrightarrow H_0: \alpha_i = 0, \beta_j = 0, \gamma_{ij} = 0$ in Model I).
- Sum of Squares.

$$\text{Res. S. S.} \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij.})^2 \quad \text{with d.f.} \; = \; \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1)$$

$$\text{Total S. S.} \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{...})^2 \quad \text{with d.f.} \; = \; \sum_{i=1}^{a}\sum_{j=1}^{b}n_{ij} - 1$$

$$\text{Reg.S.S.} = \text{Total S.S. - Res.S.S.} \quad \text{with d.f.} = ab - 1.$$

- Test stat $\quad F = \dfrac{\text{Reg.SS}/(ab-1)}{\text{Res.SS}/(N-ab)}$

  Reject H_0 if $F_{obs} > F_\alpha(ab-1, N-ab)$.

③ interaction effect

- $H_0: \gamma_{ij} = 0, \quad i = 1, \cdots, a-1, \quad j = 1, \cdots, b-1.$

  $\Leftrightarrow \forall i = 2, \cdots, a-1, \quad \mu_{11} - \mu_{i1} = \mu_{12} - \mu_{i2} = \cdots = \mu_{1b} - \mu_{ib}$.

- Generalized linear hypothesis

Test stat: $\quad F = \dfrac{(C\hat{\beta})^T [C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta})}{r\hat{\sigma}^2} \overset{H_0}{\sim} F(r, N-ab)$.

Reject H_0 if $F > F_\alpha(r, N-ab) \qquad r = (a-1)(b-1)$

④ Main effect test (insignificant interaction)

1. No Difference in Means Due to Factor A
   $H_0^1 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{a.}$
2. No Difference in Means Due to Factor B
   $H_0^2 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.b}$

4. Balanced design $\quad n_{ij} = n \;\; \forall i, j$

Unbalanced design $\quad$ o.w.