Experimental Design
Note 3-1
Model Adequacy Checking

회귀진단 ...?

Keunbaik Lee

Sungkyunkwan University

Model checking and diagnostics I

- **Checking assumptions** is important
  - Have we fit the right model?
  - Normality
  - Independence
  - Constant variance

$$
\begin{aligned}
y_{ij} &= (\overset{\hat{\mu}}{\bar{y}_{\cdot\cdot}} + \overset{\hat{\tau}_i}{(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}) + \overset{\hat{\varepsilon}_{ij}}{(y_{ij} - \bar{y}_{i\cdot})} \\
y_{ij} &= \hat{y}_{ij} + e_{ij} \\
\text{observed} &= \text{predicted} + \text{residual}
\end{aligned}
$$

- Note that the predicted response at treatment i is $\hat{y}_{ij} = \bar{y}_{i\cdot}$.
- Diagnostics use predicted responses and residuals.

Model checking and diagnoistics II

- Normality
  - Histogram of residuals
  - Normal probability plot / QQ plot
  - Shapiro-Wilk Test
- Constant Variance
  - Plot $\hat{\epsilon}_{ij}$ vs $\hat{y}_{ij}$ (residual plot)
  - Bartlett's or Levene's Test
- Independence
  - Plot $\hat{\epsilon}_{ij}$ vs time/space
  - Plot $\hat{\epsilon}_{ij}$ vs variable of interest
- Outliers

Normality Checking in the ANOVA

- Examination of residuals

$$
\begin{aligned}
e_{ij} &= y_{ij} - \hat{y}_{ij} \\
&= y_{ij} - \bar{y}_{i\cdot}
\end{aligned}
$$

- Residual plots are very useful - e.g., Q-Q plot
- Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling Tests

Outliers Checking

- Use standardized residuals to check if there is outliers

$$d_{ij} = \frac{e_{ij}}{\sqrt{MSE}} \quad = y_{ij} - \bar{y}_{i\cdot}$$

- > 3 or < -3 is a potential outlier
- Be careful for removing outliers

Constant variance checking I

- In some experiments, error variance (σ_i^2) depends on the mean response

$$E(y_{ij}) = \mu_i = \mu + \tau_i$$

  So the constant variance assumption is violated.
- Size of error (residual) depends on mean response (predicted value)
- Residual plot
  - Plot $\hat{\epsilon}_{ij}$ vs $\hat{y}_{ij}$
  - Is the range constant for different levels of $\hat{y}_{ij}$
  - More formal tests: Bartlett's Test, Modified Levene's Test.
- Modified Levene's Test
  - For each fixed i, calculate the median m_i of $y_{i1}, y_{i2}, \cdots, y_{in_i}$.

Constant variance checking II

- Compute the absolute deviation of observation from sample median:

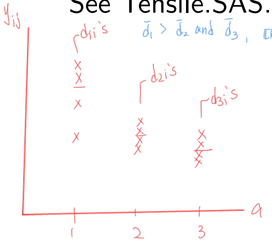
$$d_{ij} = |y_{ij} - m_i|$$

  for $i = 1, 2, \cdots, a$ and $j = 1, 2, \cdots, n_i$.
- Apply ANOVA to the deviations: d_{ij} ~ Suppose $d_{ij} = \mu + \tau_i + \varepsilon_{ij}$
- Use the usual ANOVA F-statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_a^2$

τ이 값이 크다는건 해당 i그룹의 값들이 median에서 많이 떨어져 있다는걸 말한다.

See Tensile.SAS.

$\bar{d}_1 > \bar{d}_2$ and $\bar{d}_3$, 때문에 σ_1^2이 σ_2^2와 σ_3^2보다 클 것이다 ~ reject H_0



Non-constant Variance: Impact and Remedy I

- Why concern?
  - Comparison of treatments depends on MSE
  - Incorrect intervals and comparison results
- Variance-Stabilizing Transformations
  - Common transformations

$$\sqrt{x},\ \log(x),\ 1/x,\ \arcsin(\sqrt{x}),\ \text{and } 1/\sqrt{x}$$

  - Box-Cox transformations
    - approximate the relationship $\sigma_i = \theta\mu_i^{\beta}$, then the transformation is $X^{1-\beta}$
    - use maximum likelihood principle
  - Ideas for finding proper transformations

Taylor's Theorem :

If the $(n-1)^{st}$ derivative of $f(x)$, $f^{(n-1)}(x)$ is continuous on $[a,b]$ and the n^{th} derivative $f^{(n)}(x)$ exists on (a,b), then for each $x \in [a,b]$, we have

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + \frac{f^{(n)}(\xi)}{n!}(x-a)^n, \text{ where } a < \xi < x$$

위 이론을 통해 $f(x) = e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ 가 유도된다

$$\left.\begin{array}{l} f(x) = e^x \\ f'(x) = e^x \\ \vdots \\ f^{(n)}(x) = e^x \end{array}\right\} f^{(i)}(0) = 1$$

$\Rightarrow$ given $a=0$, $f(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^{n-1}}{(n-1)!} + \frac{x^n}{n!} + \cdots$

~ Delta-method :

In Taylor's Theorem, we only consider 1^{st} order of $f(x)$,

$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a)$, now let Y be a random variable with $E(Y) = \mu$ and $Var(Y) = \sigma^2$. Then we can find the mean and variance of $f(Y)$.

Suppose $a = \mu$, then $f(Y) = f(\mu) + \frac{f'(\mu)}{1!}(Y-\mu)$.

$$\Rightarrow E[f(Y)] = f(\mu) \quad \& \quad Var(f(Y)) = (f'(\mu))^2 \sigma^2$$

Variance - Stabilizing Transformation

1. If $\sigma_i = \theta \mu_i^{\beta}$, then $Y_i^{1-\beta}$ is the VST

pf) Let $f(Y_i) = Y_i^{1-\beta} \Rightarrow f'(Y_i) = (1-\beta) Y_i^{-\beta}$, $var(Y_i^{1-\beta}) \approx (f'(\mu_i))^2 \sigma_i^2 = (1-\beta)^2 \mu_i^{-2\beta} \theta^2 \mu_i^{2\beta} = (1-\beta)^2 \theta^2$

2. If $Y_i \sim Poisson(\mu_i)$, then $f(Y_i) = \sqrt{Y_i}$ is the VST

pf) We have $E(Y_i) = Var(Y_i) = \mu_i$

$$\Rightarrow Var\{f(Y_i)\} = Var(\sqrt{Y_i}) = \left(\frac{1}{2}\mu_i^{-\frac{1}{2}}\right)^2 \cdot \mu_i = \frac{1}{4}\mu_i^{-1} \cdot \mu_i = \frac{1}{4}$$

Non-constant Variance: Impact and Remedy II

- Consider response Y with mean $E(Y) = \mu$ and variance $var(Y) = \sigma^2$.
- That σ^2 depends on μ leads to nonconsistent variances for different μ.
- Let f be a transformation and $\tilde{Y} = f(Y)$. What is the mean and variance of $\tilde{Y}$?
- Approximate $f(Y)$ by a linear function (Delta Method):

$$f(Y) \approx f(\mu) + (Y - \mu)f^{'}(\mu)$$

Then

$$\begin{aligned} \text{Mean: } \tilde{\mu} = E(\tilde{Y}) &= E(f(Y)) \approx E(f(\mu)) + E((Y-\mu)f^{'}(\mu)) \\ &= f(\mu) \end{aligned}$$

$$\text{Variance: } \tilde{\sigma}^2 = var(\tilde{Y}) \approx \left[f^{'}(\mu)\right]^2 var(Y) = \left[f^{'}(\mu)\right]^2 \sigma^2$$

Non-constant Variance: Impact and Remedy III

- f is a good transformation if $\tilde{\sigma}^2$ does not depend on $\tilde{\mu}$ anymore. So, $\tilde{Y}$ has constant variance for different $f(\mu)$.

- Transformations
  - Suppose σ^2 is a function of μ, that is $\sigma^2 = g(\mu)$
  - Want to find transformation f such that $\tilde{Y} = f(Y)$ has constant variance: $var(\tilde{Y})$ does not depend on μ.
  - Have shown $var(\tilde{Y}) \approx \left[f'(\mu)\right]^2 \sigma^2 \approx \left[f'(\mu)\right]^2 g(\mu)$
  - Want to choose f such that $[f'(\mu)]^2 g(\mu) \approx c$

Distribution	Variance	Transformation
Poisson	$g(\mu) = \mu$	$f(\mu) = \int \frac{1}{\sqrt{\mu}} d\mu \longrightarrow f(X) = \sqrt{X}$
Binomial	$g(\mu) = \mu(1-\mu)$	$f(\mu) = \int \frac{1}{\sqrt{\mu(1-\mu)}} d\mu \longrightarrow f(X) = \arcsin(\sqrt{X})$
Box-Cox	$g(\mu) = \mu^{2\beta}$	$f(\mu) = \int \mu^{-\beta} d\mu \longrightarrow f(X) = X^{1-\beta}$
Box-Cox	$g(\mu) = \mu^2$	$f(\mu) = \int \frac{1}{\mu} d\mu \longrightarrow f(X) = \log X$

Non-constant Variance: Impact and Remedy IV

- Identify Box-Cox Transformation using Data: Approximate Method
  - From the previous slide, if $\sigma_i = \theta\mu_i^\beta$, the transformation is

    $$f(Y) = \begin{cases} Y^{1-\beta}, & \beta \neq 1; \\ \log Y, & \beta = 1. \end{cases}$$

    So it is crucial to estimate β based on data y_{ij}, $i = 1, \cdots, a$.
  - We have $\log \sigma_i = \log \theta + \beta \log \mu_i$.
  - Let s_i and $\bar{y}_{i\cdot}$ be the sample standard deviations and means. Because $\hat{\sigma}_i = s_i$ and $\hat{\mu}_i = \bar{y}_{i\cdot}$, approximately,

    $$\log s_i = \text{constant} + \beta \log \bar{y}_{i\cdot},$$

    where $i = 1, \cdots, a$.
  - We can plot $\log s_i$ against $\log \bar{y}_{i\cdot}$, fit a straight line and use the slope to estimate β.

Non-constant Variance: Impact and Remedy V

- Identify Box-Cox Transformation: Formal Method
  - For a fixed λ, perform analysis of variance on

$$y_{ij}(\lambda) = \begin{cases} \frac{y_{ij}^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0; \\ \dot{y}\log y_{ij}, & \lambda = 0, \end{cases}$$

    where $\dot{y} = \prod_{i=1}^{a}\prod_{j=1}^{n_i} y_{ij}^{1/N}$. geometric mean of obs
    Notice that we cannot select the value of λ by directly comparing the SSEs from the ANOVA on y^{λ} because for each value of λ, the SSEs are measured on different scales.
  - Step 1 generates a transformed data $y_{ij}(\lambda)$. Apply ANOVA to the new data and obtain SS_E. Because SS_E depends on λ, it is denoted by $SS_E(\lambda)$.
  - Repeat 1 and 2 for various λ in an interval, e.g., $[-2, 2]$, and record $SS_E(\lambda)$

Non-constant Variance: Impact and Remedy VI

- Find λ_0 which minimizes $SS_E(\lambda)$ and pick up a meaningful λ in the neighborhood of λ_0. Denote it again by λ.
- Now the selected transformation is:

$$f(y_{ij}) = \begin{cases} y_{ij}^{\lambda_0}, & \text{if } \lambda_0 \neq 0; \\ \log y_{ij}, & \text{if } \lambda_0 = 0. \end{cases}$$

See Transformation.SAS.

Nonparametric methods for ANOVA I

H_0 : a treatments are equal vs H_a: at least one not equal.
(But normality assumption is unsatisfied)

- Kruskal-Wallis Test
  - Rank the observations y_{ij} in ascending order
  - Replace each observation by its rank R_{ij} (assign average for tied observations)
  - Test statistic

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i\cdot}^2}{n_i} - \frac{N(N+1)^2}{4}\right] \approx \chi^2_{a-1}$$

where $S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right]$

Nonparametric methods for ANOVA II

- Decision Rule: reject H_0 if $H > \chi^2_{\alpha,a-1}$.

See Nonparametric.SAS.