1 Statistics 249: Categorical Data/GLM Lecture 6

1.0.1 Final Projects

Meet with Rima about projects by March 30. Presentation for 25 minutes. Can be a paper or a chapter out of the book that we have not covered yet. Will be judged on strength of presentation, use of equipment, choice of material.

Rather than the last class being a class, we may choose to have a discussion of articles. Read a paper, provide a summary. This discussion is in the air.

1.1 One Way Anova

\[ Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \]
\[ i = 1, \ldots, I \quad \sum n_i = N \]
\[ j = 1, \ldots, n_i \]

Balanced design or sampling case is defined as \( n_i = n \) for all \( i \).

Estimates for \( \mu_i = \bar{y}_i \) and \( \hat{\tau}_i = \bar{y}_i - \bar{y}_.. \).

1.1.1 The Squared Errors

- \( SS_T = \sum_{ij} (Y_{ij} - \bar{Y}_..)^2 \)
- \( SST_{\text{reat}} = \sum_i n_i (\bar{Y}_i. - \bar{Y}_..)^2 \)
- \( SSE_{\text{error}} = \sum_{ij} (Y_{ij} - \bar{Y}_i.)^2 \)

\[ SS_T = SST_{\text{reat}} + SSE_{\text{error}} \]

1.1.2 The Hypothesis

\( H_0 : \tau_1 = \tau_2 = \tau_3 = \ldots \tau_I = 0 \)
\( H_A : \) at least one \( \tau_i \neq 0 \)

If \( \frac{msT_{\text{reat}}}{msE_{\text{error}}} > F_{I-1,N-I,\alpha} \) Then reject \( H_0 \) at level \( \alpha \) at this time.

1.1.3 Constrast Measurements

Constraints are a linear combination of our treatments \( \sum c_i \tau_i \) such that \( \sum c_i = 0 \).

ex.: \( \tau_1 - \tau_2 \) is a constrast with estimate \( \bar{Y}_1. - \bar{Y}_2. \)

One’s estimate of a contrast is not affected by the contrast that one chooses to make at the beginning.
1.2 Estimating \( \sigma^2 \)

Estimate is always \( \sum c_i \bar{Y}_i \). Method of Moments, minimizing SSE, all minimized. The variance is:

\[
\text{Var} \left( \sum_i c_i \bar{Y}_i \right) = \sum_i \frac{c_i^2}{n_i} \sigma^2
\]

1.2 Estimating \( \sigma^2 \)

Most often we wish to use an unbiased estimate for \( \sigma^2 \)

\[
\mathbb{E} \left[ \frac{\text{SSE}_{\text{error}}}{N-I} \right] = \sigma^2
\]

Thus we use \( \hat{\sigma}^2 = \frac{\text{SSE}_{\text{error}}}{N-I} \). \( \hat{\sigma}^2 \sim \frac{\sigma^2}{N-I} \chi^2_{N-I} \). Thus for our confidence interval for \( \sum c_i \tau_i \) one should use a \( t \) statistic, not a \( Z \) statistic:

1.2.1 C.I. for \( \sum c_i \tau_i \)

\[
\sum c_i \bar{Y}_i \pm t_{N-I, \alpha/2} \cdot \sqrt{\text{MSE}_{\text{error}} \sum \frac{c_i^2}{n}}
\]

Use this confidence interval ONLY if one is estimating one confidence interval. Multiple pairwise tests will cause mistakes, since running two 95% confidence interval tests (assuming independent) have a 8% chance of failing, even if the null is true.

1.3 The Three Methods of Multiple Hypothesis Tests

Tests for \( m \) contrasts. \( \sum c_{1i} \tau_i, \sum c_{2i} \tau_i, \ldots, \sum c_{mi} \tau_i \)

The form of the confidence interval becomes:

\[
\sum c_i \bar{Y}_i \pm \underbrace{w}_{\text{Correction}} \sqrt{\text{MSE}_{\text{error}} \sum \frac{c_i^2}{n}}
\]

1. Bonferroni \( w = t_{N-I, \frac{\alpha}{m}} \)
   Best when \( m \) is small

2. Scheffe \( w = \sqrt{(I-1)F_{I-1, N-I, \alpha}} \)
   Best when \( m \) is large.

3. Tukey Only effective for \( m \) pairwise differences.

\[
\hat{w}_T = \frac{Q_{I,n-I, \alpha}}{\sqrt{2}}
\]

Where \( Q \) stands for the “Studentized” range distribution.

Best for Pairwise differences.

All of these are very conservative intervals.
1.4 Transforming data, scale.

An issue for debate: On what range should you analyze a dataset, transform it, to maintain the assumptions of the model?

1.5 Two Way ANOVA

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \varepsilon_{ijk} \]

For rat study \( Y_{ijk} \) the weight of rats, from different litters and different mothers. Consider \( \alpha_i \) as the effect of litter (genetic effect) and \( \beta_j \) as the effect of the environment (mother \( j \) feeding rat)

1.5.1 Interaction Term

1.5.2 Main Effect Model (No interaction)

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \]

1.5.3 Constraints for this model

\[ \sum_{ij} d_{ij} \tau_{ij} : \sum_{ij} d_{ij} = 0 \]

Or:

\[ \sum_i c_i \alpha_i^* \text{ for } \sum_i c_i = 0 \text{ and } \alpha_i^* = \bar{\tau}_i. \]

\[ \sum_j d_j \beta_j^* \text{ for } \sum_j d_j = 0 \text{ and } \alpha_j^* = \bar{\tau}_j. \]

Note that \( \bar{\tau}_i = \frac{1}{J} \left[ \sum_j \mu + \alpha_i + \beta_j \right] \)
1.6 Estimate of of $\alpha_i^*$ and $\beta_j^*$

For $\sum c_i \alpha_i^*$:

$$\sum c_i \left( \frac{1}{J} \sum_j \bar{Y}_{ij} \right)$$

For $\sum d_j \beta_j^*$:

$$\sum d_j \left( \frac{1}{I} \sum_i \bar{Y}_{ij} \right)$$

1.6.1 An unbiased estimate for the error estimate $\hat{\sigma}^2$

$\hat{\sigma}^2$ is the MSE

1.7 Anova: (balanced design)

$$SS_T = SS_1 + SS_2 + SS_{12} + SSE_{\text{error}} \quad \text{(full model)}$$

These Sum of squares are then defined as:

- $SS_T = \sum_{ijk} (Y_{ijk} - \bar{Y}_{..})^2$
- $SS_1 = \sum_i J \cdot n (\bar{Y}_{i..} - \bar{Y}_{..})^2$
- $SS_2 = \sum_j I \cdot n (\bar{Y}_{.j.} - \bar{Y}_{..})^2$
- $SS_{12} = n \sum_{ij} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{..})^2$
- $SSE_{\text{error}} = \sum_{ijk} (Y_{ijk} - \bar{Y}_{ij.})^2$

1.8 For the Main Effect Model

$$SS_T = \frac{SS_1 + SS_2}{\sum_{ijk} (Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{..})^2} + SSE_{\text{error}} \quad \text{(main effect model)}$$

Same as before

If interaction term is not significant, then the main effect model might be valid.

1.9 The Rat Data

<table>
<thead>
<tr>
<th>Litter</th>
<th>Mother</th>
<th>Wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>17</td>
<td>Min. :36.30</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>1st Qu.:48.20</td>
</tr>
<tr>
<td>I</td>
<td>14</td>
<td>Median :54.00</td>
</tr>
<tr>
<td>J</td>
<td>15</td>
<td>Mean :53.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3rd Qu.:60.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Max. :69.80</td>
</tr>
</tbody>
</table>
The data is definitely not parallel, we should probably look at an interaction term in the model.

It appears that the residuals in the data are lowering in variance with higher fitted values. Cook’s Distance is a “leave one out” diagnostic, that checks what the result is when one does the analysis “missing” one of the data points. *Design of Analysis and Experiments* is the book being used for this.