Lecture 25

Diagnostics & Remedial Measures for ANOVA

STAT 512
Spring 2011

Background Reading
KNNL: Chapter 18
Topic Overview

- ANOVA Diagnostics
- Remedial Measures
Regression vs ANOVA

- Basic assumptions on errors are the same (independence, normality, constancy of variance for errors)
- Recall that for ANOVA we do NOT have the linearity assumption.
- Diagnostics and remedial measures often similar or the same; we will focus only on key differences
Diagnostic Procedure

- Review model diagnostics as early as possible in the analysis
  - First check residual plots
  - If any sign of problems, can use various statistical tests for some confirmation.

- If any serious problems, try appropriate remedial measures
Residuals

- Predicted values are cell means, $\hat{Y}_{ij} = \bar{Y}_i$.
- Residuals are differences between observed values and cell means: $e_{ij} = Y_{ij} - \bar{Y}_i$.
- Residual Plots
  - Plot against fitted values (cell means) or factor levels (check constant variance)
  - Sequence Plot (check independence, when sequence is available/reasonable)
  - Normal Probability Plot (check normality)
Non-constant variance

• Since there is generally no ordering to the levels of the predictor variable, it doesn’t make sense to look for a “megaphone”.

• Rather, simply look for large differences in vertical spreads.

• If sample sizes differ greatly between factor levels, use studentized residuals.
Non-constant Variance (2)

• If residual plots indicate potential problems, can use statistical tests to check.
  - \( H_0 : \sigma^2_i = \sigma^2_{i'} \) for all \( i \)
  - \( H_A : \sigma^2_i \neq \sigma^2_{i'} \) for some \( i \)

• Brown-Forsythe test. SAS: after the “/” in the MEANS statement use HOVTEST=BF

• Rejecting the null indicates there is evidence that not all of the factor level variances are equal. So, we are looking for a P-value larger than \( \alpha \) for the assumption to be met.
Non-constant Variance (3)

- Hartley test – simpler test, but requires equal sample sizes and is quite sensitive to departures from normality. Not available in SAS.
- Levene test – commonly used test for equality of variances. Similar to Brown-Forsythe, but not discussed in book. Use HOVTEST=LEVENE.
Non-constant Variance (4)

- ANOVA F-test only slightly affected by non-constant variance as long as sample sizes are equal.
- Scheffe multiple comparison procedure is also fairly robust to unequal sample variances if cell sizes are equal.
- Other pairwise comparisons CAN BE greatly affected by unequal variances – use equal sample sizes to minimize this effect.
Non-constant Variance (5)

- Easiest remedial measure is usually a transformation (can help both non-constant variance and non-normality)
  - If variance proportional to $\mu_i$ then try $\sqrt{Y}$ (sometimes occurs if $Y$ is a count)
  - If standard deviation proportional to $\mu_i$, try log transformation.
  - If standard deviation proportional to $\mu_i^2$, try $\frac{1}{Y}$
  - If response is a proportion, try arcsine transformation $Y' = 2 \arcsin \sqrt{Y}$
Non-constant Variance (6)

- To check whether one of these is applicable, calculate sample factor level variances ($s^2_i$) and means $\bar{Y}_i$.
- Create plots:
  \[
  \bar{Y}_i \text{ vs. } s^2_i, \quad \bar{Y}_i \text{ vs. } s_i, \quad \text{and } \bar{Y}^2_i \text{ vs. } s_i
  \]
- If any of the previously mentioned trends appear, use the corresponding transformation.
- Box Cox can also be used to find transformation.
Non-constant Variance (7)

- Weighted Least Squares can also be used as a remedial measure
  - Estimate sample variances
  - Use reciprocals as weights

- See section 18.4 for more information. To do such an analysis in SAS, utilize a WEIGHT statement (and carefully read the SAS help concerning this)
Non-normality

- Use a Normal Probability Plot to check this.
- If unequal variances, then often non-normality will be falsely indicated by using regular residuals; should transform first and then recheck.
- Normality is the least important assumption; almost all of ANOVA procedures robust to minor departures from normality
Non-Independence

- If data obtained in time sequence, plot residuals against time
- If pattern, then may have non-independence.
  - *Positive Serial Correlation* (adjacent residuals tend to have the same sign)
  - *Negative Serial Correlation* (adjacent residuals tend to have opposite signs)
- Non-independence usually has serious effects on inferences (making them invalid)
Outliers

- Can use studentized or studentized deleted residuals as before to classify outliers.
- Leverage values: For ANOVA model it can be shown that the leverage of $Y_{ij}$ is $1/n_i$.
- If cell sizes equal, leverage values equal so no one point has more leverage than another (even though it may be outlying in the response).
**Cash Offers Example**  
*(cashoffers.sas)*

```sas
proc glm data=cash;
  class age;
  model offer=age;
  means age/hovtest=bf;
  output out=diag p=pred r=resid rstudent=SDR;
run;

symbol1 v=dot c=blue;
proc gplot data=diag;
plot resid*age;
run;

proc univariate noprint;
  var resid;
  qqplot resid/normal (L=1 mu=est sigma=est);
run;

proc print data=diag;
  var offer age SDR;
run;
```
Residuals vs. Age
### HOVTEST=BF Output

Brown and Forsythe's Test for Homogeneity of offer Variance
ANOVA of Absolute Deviations from Group Medians

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>2</td>
<td>0.3889</td>
<td>0.1944</td>
<td>0.21</td>
<td>0.8132</td>
</tr>
<tr>
<td>Error</td>
<td>33</td>
<td>30.8333</td>
<td>0.9343</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- P-value quite large so no evidence of non-constant variance
Normal Probability Plot
Cash Offers Example

• Constant variance and normality assumption appear to be satisfied.

• There do not appear to be any outliers.
Winding Speeds Example

- Problem 18.17 in the text.
- SAS code: windingspeed.sas
- Interested in determining the effect of winding speed of thread (slow, normal, fast, maximum) on the number of thread breaks during a production run – the response variable is a “count”, so we should already have some concerns.
- 64 total observations (16 each on four different speeds)
# GLM Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>1588</td>
<td>529</td>
<td>47.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>60</td>
<td>669</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>2257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grp</th>
<th>Mean</th>
<th>N</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16.563</td>
<td>16</td>
<td>4_max</td>
</tr>
<tr>
<td>B</td>
<td>10.688</td>
<td>16</td>
<td>3_fast</td>
</tr>
<tr>
<td>C</td>
<td>5.875</td>
<td>16</td>
<td>2_norm</td>
</tr>
<tr>
<td>C</td>
<td>3.563</td>
<td>16</td>
<td>1_slow</td>
</tr>
</tbody>
</table>
Residual Plot
Brown-Forsythe

Brown and Forsythe's Test

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed</td>
<td>3</td>
<td>111.5</td>
<td>37.1823</td>
<td>9.54</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>60</td>
<td>233.8</td>
<td>3.8969</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Clearly issues here – variance increases as speed increases
- Typical with count data since true distribution is probably Poisson
Normal Probability Plot
Consider Cell Means/Variances

```
proc means data=ws;
  class speed;
  var breaks;
```

<table>
<thead>
<tr>
<th>speed</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_slow</td>
<td>16</td>
<td>3.5625</td>
<td>1.0935</td>
<td>2.000</td>
<td>6.000</td>
</tr>
<tr>
<td>2_norm</td>
<td>16</td>
<td>5.8750</td>
<td>1.9958</td>
<td>2.000</td>
<td>9.000</td>
</tr>
<tr>
<td>3_fast</td>
<td>16</td>
<td>10.6875</td>
<td>3.2397</td>
<td>6.000</td>
<td>17.000</td>
</tr>
<tr>
<td>4_max</td>
<td>16</td>
<td>16.5625</td>
<td>5.3786</td>
<td>7.000</td>
<td>25.000</td>
</tr>
</tbody>
</table>
Plot Est. Means vs. VAR’s

Linear would suggest SQRT transformation
Plot Est. Means vs SD’s

Linear suggests log-transformation
Plot Squared Means vs. SD

Linear would suggest Inverse Transformation
Alternative Transformation Check

• Calculate: \[ \frac{s_i^2}{\bar{Y}_i}, \frac{s_i}{\bar{Y}_i}, \text{ and } \frac{s_i}{\bar{Y}_i}^2 \]

• If any of these are fairly constant over all factor levels, apply corresponding transformation (sqrt, log, inverse)

• In SAS:

*Alternate Way of Checking for Appropriate Transformation (if plots are difficult to interpret);

```sas
data a2; set a1;
  if _Type_ = 1;
  var_mean=(sighat*sighat)/muhat;
  sd_mean=sighat/muhat;
  sd_meansq=sighat/(muhat*muhat);
proc print data=a2;
  var speed var_mean sd_mean sd_meansq;
run;
```
Alternative Transformation Check

<table>
<thead>
<tr>
<th>Obs</th>
<th>speed</th>
<th>var_mean</th>
<th>sd_mean</th>
<th>sd_meansq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1_slow</td>
<td>0.33567</td>
<td>0.30696</td>
<td>0.086164</td>
</tr>
<tr>
<td>2</td>
<td>2_norm</td>
<td>0.67801</td>
<td>0.33972</td>
<td>0.057824</td>
</tr>
<tr>
<td>3</td>
<td>3_fast</td>
<td>0.98207</td>
<td>0.30313</td>
<td>0.028363</td>
</tr>
<tr>
<td>4</td>
<td>4_max</td>
<td>1.74667</td>
<td>0.32474</td>
<td>0.019607</td>
</tr>
</tbody>
</table>

Here, $\frac{s_i}{\bar{Y}_i}$ appears fairly constant, confirming that a log transformation may be useful.
Box Cox

• Use PROC TRANSREG as before – difference is any categorical predictor needs to be prefaced by class(*).

```plaintext
proc transreg data=ws;
  model boxcox(breaks)=class(speed);
  run;
```

• Expecting (from what we just saw) to use a log transformation (lambda = 0)
## Box Cox (2)

<table>
<thead>
<tr>
<th>Lambda</th>
<th>R-Square</th>
<th>Log Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.00</td>
<td>0.62</td>
<td>-91.231</td>
</tr>
<tr>
<td>-0.75</td>
<td>0.66</td>
<td>-80.496</td>
</tr>
<tr>
<td>-0.50</td>
<td>0.70</td>
<td>-71.714</td>
</tr>
<tr>
<td>-0.25</td>
<td>0.72</td>
<td>-65.409</td>
</tr>
<tr>
<td>0.00 +</td>
<td>0.74</td>
<td>-62.028 *</td>
</tr>
<tr>
<td>0.25</td>
<td>0.74</td>
<td>-61.764 &lt;</td>
</tr>
<tr>
<td>0.50</td>
<td>0.74</td>
<td>-64.482</td>
</tr>
<tr>
<td>0.75</td>
<td>0.72</td>
<td>-69.783</td>
</tr>
<tr>
<td>1.00</td>
<td>0.70</td>
<td>-77.169</td>
</tr>
</tbody>
</table>

< - Best Lambda
* - 95% Confidence Interval
+ - Convenient Lambda
Model for $Y' = \log(Y)$

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>21.69</td>
<td>7.23</td>
<td>56.78</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>60</td>
<td>7.64</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>29.33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P-value for BF Test is now 0.96 (so constancy of variance satisfied)
Model for $Y' = \log(Y)$
Model for $Y' = \log(Y)$
Model for $Y' = \log(Y)$

- Previously thought all groups of means were significantly different except 1 and 2, but now see that all of them are different.

<table>
<thead>
<tr>
<th>GRP</th>
<th>Mean</th>
<th>N</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.7499</td>
<td>16</td>
<td>4_max</td>
</tr>
<tr>
<td>B</td>
<td>2.3211</td>
<td>16</td>
<td>3_fast</td>
</tr>
<tr>
<td>C</td>
<td>1.7039</td>
<td>16</td>
<td>2_norm</td>
</tr>
<tr>
<td>D</td>
<td>1.2237</td>
<td>16</td>
<td>1_slow</td>
</tr>
</tbody>
</table>
Alternatively: WLS

• Instead of transforming, try Weighted Least Squares
• Need inverse cell variances for weights (get these from PROC MEANS)

```
proc means data=ws;
  class speed;
  var breaks;
  output out=a1 var=variance;
```

```
data a1; set a1; if _type_ = 1;
data ws; merge ws a1; by speed;
  weight=1/variance;
```
WLS (2)

```plaintext
proc glm data=ws;
  class speed;
  model breaks=speed;
  weight weight;
  means speed /tukey lines;
```

**Weight:** weight

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>153.95</td>
<td>51.32</td>
<td>51.32</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>60</td>
<td>60.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>213.95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Notice MSE = 1. This will always be the case when the cell variances are used as the weights.
Weighted Analysis

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>N</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16.563</td>
<td>16</td>
<td>4_max</td>
</tr>
<tr>
<td>B</td>
<td>10.687</td>
<td>16</td>
<td>3_fast</td>
</tr>
<tr>
<td>C</td>
<td>5.875</td>
<td>16</td>
<td>2_norm</td>
</tr>
<tr>
<td>C</td>
<td>3.563</td>
<td>16</td>
<td>1_slow</td>
</tr>
</tbody>
</table>

- Same results as no transformation.
Residual Plots (1)
Residual Plots (2)
Weighted Analysis

- Did not really help the non-constant variance/normality issue and pairwise comparison results were the same as the original data.
- If normality is an issue, a transformation is generally better than WLS.
- If normality is not an issue, WLS is appropriate.
Upcoming in Lecture 26...

- Two-way ANOVA (Chapter 19)