

## Ch 18 – ANOVA Diagnostics and Remedial Measures

We use various diagnostic tools to check model assumptions. If the diagnostics indicate that some model assumptions fail to hold, and if the methods of analysis are not sufficiently robust to these departures from the assumptions, then we consider remedial measures.

“LINE” is a good acronym for the model assumptions, (though better for simple linear regression):

- L: **L**inear model correctly specified mean response  $E[Y_{ij}]$ . If not, we have **L**ack of fit of our model.
- I: **I**ndependent errors  $\epsilon_{ij}$
- N: **N**ormally distributed errors  $\epsilon_{ij}$
- E: **E**qual variances for errors  $\epsilon_{ij}$ —i.e. homoscedasticity

### Residual Analysis

Residual analysis is used to evaluate the model assumptions. Recall, the residuals are the differences between the observed and fitted values—namely,  $e_{ij} = y_{ij} - \hat{y}_{ij}$ .

The permutation “LIEN” indicates a better order to check the model assumptions. Generally speaking, one should check first for lack of fit because, given lack of fit, the residuals are poor surrogates for the errors so are not useful for evaluating the other assumptions. That said, one cannot have lack of fit of the one-way ANOVA model, since the residuals necessarily sum to zero for each treatment—namely, for each  $i$ ,  $\sum_j e_{ij} = \sum_j (y_{ij} - \bar{y}_{i.}) = y_{i.} - n_i \bar{y}_{i.} = 0$ . Nonetheless, if one has omitted other important explanatory variables from the model, this may be evident by plotting the residuals against said variables.

It is reasonable to check independence second, since inference are not robust to dependence, and since tests for normality and homoscedasticity require independence. However, it’s difficult to detect dependence, unless it’s due to a clear time trend or time or spatial pattern. Fortunately, independence will be a reasonable assumption for a well-run experiment unless the experimental units induce correlations and, even then, randomization tends to mitigate the effects of dependence. Dependence may be evident if the observations are collected sequentially over time or if experimental units are spatially related.

The equal variances assumption is more critical than the normality assumption because statistical methods for analysis of fixed effects are generally rather robust to the normality assumption, largely because of the central limit theorem, but less robust to the homoscedasticity assumption. The  $F$ -test, and correspondingly Scheffé’s method of multiple comparisons, are rather robust to heteroscedasticity if samples sizes  $n_i$  are equal (or nearly equal) and if the largest ratio of treatment variances is at most three—namely,  $\max_i(\sigma_i^2)/\min_i(\sigma_i^2) \leq 3$ . One can expect similar robustness of Tukey’s method under similar circumstances. However, individual confidence intervals (and tests) can be quite affected by unequal variances. For example, if two treatments have relatively large variances, MSE will tend to be too small, so confidence intervals will tend to be too narrow, or equivalently, to have a lower level of confidence than claimed. Correspondingly, the Bonferroni method can be quite affected by unequal variances depending on the pre-specified treatment contrasts.

## Residual Plots

Residual plots (with which you should already be familiar) are the primary tool for evaluation model assumptions, though certain statistical tests may also be useful. A plot of the residuals versus time (or spatial order) can be used to look for dependence, though these plots if available tend not to be very insightful unless a clear pattern is evident. Homogeneity of variances can be evaluated by plotting the residuals versus either the factor levels or the fitted values. Plotting the residuals versus the fitted values (i.e.  $e_{ij}$  vs  $\hat{y}_{ij}$ ) is the more useful of the two plots, because it can clearly reveal if variability increases (or decreases) with mean response—a case of heteroscedasticity that we can often repair, using a Box-Cox transformation for example. If homoscedasticity seems reasonable, then a normal probability plot of the collective residuals can be used to evaluate normality, (plotting residuals versus Blom's normal scores,  $z((k - 3/8)/(n_T + 1/4))$ , for example, from (3.6), p. 111). Inferences on fixed effects are relatively robust to asymmetry but more affected by non-normal kurtosis—i.e. heavier or lighter tails. Light tails are suggestive of conservative inferences and heavy tails suggestive of liberal inferences when applying normality-based procedures. Outliers correspond to heavy tails.

## Testing Normality

A simple test of normality, assuming independence and homoscedasticity, rejects normality if the sample correlation between the residuals and their normal scores is too small. See Section 3.5 for details and Table B.6 for critical values.

**Example:** ch18eg1.txt

**Homework:** 18.5abc, 18.6abc. (For part (a) of each problem, plot residuals versus fitted values and factor levels rather than what's requested.)

## 18.2 Tests for Constancy of Error Variance

We consider two tests of  $H_0 : \sigma_1^2 = \dots = \sigma_r^2$ , for  $\sigma_i^2$  the variance for the  $i$ th treatment, versus  $H_a : \text{not all } \sigma_i^2 \text{ are equal}$ .

### Hartley's Test

Hartley's test assumes independent samples from Normal distributions, requires equal samples sizes (more or less), and is *not* very robust to non-normality. A level- $\alpha$  test rejects the null hypothesis if  $H^* = \max_i(s_i^2)/\min_i(s_i^2) > H(1 - \alpha; r, \nu)$ . Critical values  $H(1 - \alpha; r, \nu)$  are in Table B.10, and  $\nu$  is the common number of degrees of freedom for each sample variance  $s_i^2$ .

**Example:** ch18eg2.txt, using exercise 18.17 data (p. 805)

— end of class #5, 1/20/09 —

### Brown-Forsythe Test

This test assumes independent samples from Normal distributions, but it is rather robust to non-normality, and it does not require equal sample sizes.

Compute each treatment (sample) median,  $\tilde{y}_i$  ( $i = 1, \dots, r$ ), then compute the absolute deviation,  $d_{ij} = |y_{ij} - \tilde{y}_i|$ , of each observation from the treatment median. Compute the usual ANOVA  $F$ -statistic but using these absolute deviations  $d_{ij}$  as the data, yielding  $F_{BF}^*$ . If samples sizes are not extremely small, then under the null hypothesis  $F_{BF}^*$  has approximately the  $F(r - 1, n_T - r)$  distribution, so an approximate level- $\alpha$  test rejects  $H_0$  if  $F_{BF}^* > F(1 - \alpha; r - 1, n_T - r)$ .

**Example:** ch18eg2.txt, using exercise 18.17 data (p. 805)

## Remedial Measures for Lack-of-Fit

If there is a clear time (or spatial) pattern in the residuals, then a more complicated model may be used. For example, if there is a clear linear time trend, a linear time covariate may be added to the model. Changing the model based on the data makes the data analysis exploratory, however.

## Remedial Measures for Heteroscedasticity

### Box-Cox Transformations

One option for clearly unequal variances is to seek a variance stabilizing transformation. A suitable Box-Cox transformation is generally available if variability increases (or decreases) with mean response. The down side is that the analysis compares effects for transformed data rather than on the original scale.

### Box-Cox Transformations handout

**Example:** ch18eg2.txt, using exercise 18.17 data (p. 805)

### Satterthwaite's Approximation

Let  $\sigma_i^2 = \sigma^2(\epsilon_{ij})$ , and let  $S_i^2$  denote the sample variance of the observations on the  $i$ th treatment. For a treatment contrast  $L = \sum_i c_i \tau_i$ ,  $\hat{L} = \sum_i c_i \bar{Y}_i$ ,  $\sigma^2(\hat{L}) = \sum_i c_i^2 \sigma_i^2 / n_i$ , and  $S^2(\hat{L}) = \sum_i c_i^2 S_i^2 / n_i$ . By Satterthwaite's approximation,  $\nu S^2(\hat{L}) / \sigma^2(\hat{L})$  is approximately  $\chi^2(\nu)$ , for

$$\nu = \frac{(\sum_i c_i^2 S_i^2 / n_i)^2}{\sum_i \frac{(c_i^2 S_i^2 / n_i)^2}{n_i - 1}}.$$

Using this value of  $\nu$ , an approximate  $100(1 - \alpha)\%$  confidence interval for  $L$  is

$$\hat{L} \pm t(1 - \alpha/2; \nu) s(\hat{L}).$$

The Bonferroni method for  $g$  pre-specified contrasts provides confidence intervals of the form

$$\hat{L} \pm t(1 - \alpha/2g; \nu) s(\hat{L}).$$

(See page 1043 for Satterthwaite's Procedure in another context.)

**Example:** ch18eg2.txt, using exercise 18.17 data (p. 805)

**Homework:** 18.13ab, 18.15 (in place of parts (d) and (e), determine an appropriate Box-Cox transformation using the method presented in class. Does  $q = 1$  appear to be reasonable?), 16ab

— end of class #6, 1/22/09; — then 1/27/09 a snow day! —

## 18.7 Nonparametric Rank F Test and Multiple Comparison Tests

Nonparametric statistical methods are methods that do not depend on parametric model assumptions such as normality. As such, they are often suggested inappropriately as a remedial measure when variances are unequal. The Nonparametric Rank  $F$  Test considered here does not assume or require normality, but it does assume that the  $r$  treatment distributions are continuous and differ only with respect to location. Note: this implies equal variances!

**Test Procedure:** Rank the observations in ascending order from 1 to  $n_T$ , and let  $R_{ij}$  denote the rank of  $Y_{ij}$ . Conduct the usual ANOVA  $F$ -test, but using the statistic  $F_R^*$ , say, computed from the ranks rather than the original data. When the treatment distributions are the same, the statistic  $F_R^*$  follows approximately the  $F(r - 1, n_T - r)$  distribution if the sample sizes  $n_i$  are not very small. One can view this as a test for equality of treatment means or effects, or equivalently as a test of equality of treatment medians.

### Multiple Comparison Tests:

One can test all pairwise comparisons using an approximate Bonferroni method if sample sizes are sufficiently large for rank sample means to be approximately normal. Treatments  $i$  and  $i'$  have significantly different means if

$$|\bar{R}_i - \bar{R}_{i'}| > B \left[ \frac{n_T(n_T + 1)}{12} \left( \frac{1}{n_i} + \frac{1}{n_{i'}} \right) \right]^{1/2},$$

where  $B = z(1 - \alpha/2g)$  and  $g = \binom{r}{2} = r(r - 1)/2$ .

Notes: One cannot get a confidence interval for treatment means based on ranks. Also, one would probably get better pairwise comparisons by ranking the treatments for each pair of treatments and adjusting the standard error accordingly.

Interesting question: can the pairwise comparisons provide directional inference?

**Example:** ch18eg2.txt, using exercise 18.17 data (p. 805)

**Homework:** 18.24abce (revised 1/29/09 to exclude part d)

### Confidence Intervals:

Another way to formulate the problem is as follows, (see Hollander, M. and Wolfe, D. A. (1973), *Nonparametric Statistical Methods*, Wiley: New York).

Assume  $r = 2$ . Consider independent, continuously distributed observations:  $Y_{1j} = \Delta + \epsilon_{1j}$  ( $j = 1, \dots, n_1$ ),  $Y_{2j} = +\epsilon_{2j}$  ( $j = 1, \dots, n_2$ ). One can then either test  $H_0 : \Delta = 0$  (for example) or construct a confidence interval for  $\Delta$ . For the  $n_1 n_2$  differences  $Y_{1j} - Y_{2j'}$ ,  $\hat{\Delta} = \text{median}\{(Y_{1j} - Y_{2j'}), j = 1, \dots, n_1, j' = 1, \dots, n_2\}$  is called the *Hodges-Lehmann estimator* of  $\Delta$ . See Hollander and Wolfe (1973) or the documentation for SAS proc NPar1Way for further details.