STAT 705 Chapter 18: ANOVA diagnostics and remedial measures

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

The raw residual is

$$e_{ij} = Y_{ij} - \bar{Y}_{i\cdot} = Y_{ij} - \hat{Y}_{ij}.$$

The studentized residual is

$$r_{ij} = \frac{e_{ij}}{\hat{\sigma}\{e_{ij}\}} = \frac{e_{ij}}{\sqrt{MSE(1-h_{ij})}} = \frac{e_{ij}}{\sqrt{MSE(n_i-1)/n_i}}.$$

The studentized deleted residual is

$$t_{ij} = \frac{e_{ij}}{\sqrt{MSE_{(ij)}(1-h_{ij})}} = e_{ij}\left[\frac{n_T - r - 1}{SSE(1-\frac{1}{n_i}) - e_{ij}^2}\right]^{1/2}.$$

Recall if the model is correct, then

$$t_{ij} \sim t(n_T - r - 1).$$

Question: Is the model reasonable?

Use residuals to check (p.778):

1. Constant error variance.
2. Outliers.
3. Normality.

To check these,

1. SAS default graphics plots e_{ij} vs. $\hat{Y}_{ij}$. Plots should show roughly constant spread.
2. SAS plots t_{ij} vs. $\hat{Y}_{ij}$. We can formally determine if ijth observation is outlier by checking
$$|t_{ij}| > t(1 - \frac{\alpha}{2n_T}; n_T - r - 1).$$
3. The normal probability plot of $\{e_{ij}\}$ should be reasonably straight.

Also, if data are collected over time, we can plot e_{ij} vs time observation was recorded to check for serial correlation.

Feral hog rooting activity example

```
data rooting;
input habitat $ activity @@;
rootroot=sqrt(activity);
datalines;
BLH 139 BLH 228 BLH 275 CTS 45 CTS 127 CTS .
U 0 U 45 U 16  MS 145 MS 124 MS 240
 ;
run;

proc glm data=rooting plots=all; class habitat;
 model rootroot=habitat;
run;
```

18.2 Tests for constant variance

For the expanded model

$$Y_{ij} \overset{ind.}{\sim} N(\mu_i, \sigma_i^2),$$

we want to test $H_0 : \sigma_1 = \cdots = \sigma_r$.

In `proc glm`, use `means factorname / hovtest;` "hov" stands for homogeneity of variance. The options are `hovtest=Bartlett` (not robust to non-normality), `hovtest=Levene` (Default), `hovtest=Obrien`, and `hovtest=BF` (Brown-Forsythe, pp. 784–785).

Levene's test is a simple oneway ANOVA F-test on "dispersion variables" $z_{ij}^2 = (Y_{ij} - \bar{Y}_{i\cdot})^2$ (default) or $z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$ (`type=abs`). O'Brien and Brown-Forsythe simply use different z_{ij} – Brown-Forsythe uses $z_{ij} = |Y_{ij} - \tilde{Y}_{i\cdot}|$, also called the *modified Levene test*.

ABT electronics example pp. 783–784 & 787-788

Five different types of flux, response is amount of force (lbs) required to break the soldered joint.

```
data abt;
input force flux @@;
datalines;
  14.87  1  16.81  1  15.83  1  15.47  1  13.60  1  14.76  1  17.40  1  14.62  1
  18.43  2  18.76  2  20.12  2  19.11  2  19.81  2  18.43  2  17.16  2  16.40  2
  16.95  3  12.28  3  12.00  3  13.18  3  14.99  3  15.76  3  19.35  3  15.52  3
   8.59  4  10.90  4   8.60  4  10.13  4  10.28  4   9.98  4   9.41  4  10.04  4
  11.55  5  13.36  5  13.64  5  12.16  5  11.62  5  12.39  5  12.05  5  11.95  5
;

proc glm data=abt plots=all;
 class flux;
 model force=flux;
 means flux / hovtest=bf;
run;
```

18.3 Remedial measures

- Weighted least squares for non-constant variance, but otherwise normal data (Section 18.4). Alternatively, add `welch` in a `means` statement in `proc glm` to perform an ANOVA generalization of Satterthwaite's two-sample approach.
- Transformation to normality, e.g. Box-Cox for non-normal data with non-constant variance (Section 18.5). Carried out the same way as in STAT 704.
- A nonparametric test (Section 18.7), e.g. the Kruskall-Wallis test. Available in the `npar1way` procedure in SAS. Does not assume normality or constant variance. Essentially replaces the Y_{ij} with ranks obtained from ranking all observations without regard to group and computes regular ANOVA F-test.

18.3 Kruskal-Wallis Test

The Kruskal-Wallis test replaces $\{Y_{ij}\}$ with $\{R_{ij}\}$ where ranking occurs across all r factor levels. You may see results in exact test, F-test and χ^2-test form (the latter two are shown below):

$$X^2_{KW} = \frac{SSTR}{SSTO/(n_T - 1)} \sim \chi^2_{r-1} \text{ under } \mathrm{H}_o$$

$$F = \frac{MSTR}{MSE} \sim F_{r-1, n_T - r} \text{ under } \mathrm{H}_o$$

Alternative nonparametric tests replace the *Wilcoxon scores* for the ranks–$a(R_{ij}) \equiv R_{ij}$–with more general forms for $a(\cdot)$. E.g., the Van der Waerden scores are

$$a\left(R_{ij}\right) = \Phi^{-1}\left(\frac{R_{ij}}{n_{T+1}}\right)$$

Servo data

Servo-Data, Inc., operates mainframes at three locations, all the identical make/model. Recorded are lengths of time (hours) between failures at the three locations.

```
data servo;
input time loc @@;
datalines;
    4.41  1  100.65  1   14.45  1   47.13  1   85.21  1
    8.24  2   81.16  2    7.35  2   12.29  2    1.61  2
  106.19  3   33.83  3   78.88  3  342.81  3   44.33  3
;

* look at diagnostic plots;
proc glm data=servo plots=all; class loc;
 model time=loc;
 means loc;
run;

* Box-Cox transformation;
proc transreg data=servo;
 model boxcox(time) = class(loc);
run;

* Kruskall-Wallis test;
proc npar1way data=servo;
 class loc; var time;
run;
```

Special case of Section 11.1 (pp. 421–427). Define weights $w_{ij} = 1/s_i^2$ where s_i is sample standard deviation from ith factor level.

Tim prefers to fit the model $Y_{ij} = \mu_i + \epsilon_{ij}$, $\epsilon_{ij} \stackrel{ind.}{\sim} N(0, \sigma_i^2)$ directly in `proc mixed`. This approach uses maximum likelihood and normal approximations for testing. However, WLS also uses an approximation. `proc mixed` will at least take variability of estimating σ_i by s_i into account!

Carried out in `proc mixed` by adding

```
repeated / group=factorname;
```

Carried out in `proc glimmix` by adding

```
random _residual_ / group=factorname;
```

ABT electronics example

Fitting

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{ind.}{\sim} N(0, \sigma_i^2).$$

```
data abt;
input force flux @@;
datalines;
  14.87  1  16.81  1  15.83  1  15.47  1  13.60  1  14.76  1  17.40  1  14.62  1
  18.43  2  18.76  2  20.12  2  19.11  2  19.81  2  18.43  2  17.16  2  16.40  2
  16.95  3  12.28  3  12.00  3  13.18  3  14.99  3  15.76  3  19.35  3  15.52  3
   8.59  4  10.90  4   8.60  4  10.13  4  10.28  4   9.98  4   9.41  4  10.04  4
  11.55  5  13.36  5  13.64  5  12.16  5  11.62  5  12.39  5  12.05  5  11.95  5
;

proc sort data=abt; by flux; run;

proc mixed data=abt;
 class flux;
 model force=flux / solution;
 repeated / group=flux; * different variance for each flux;
run;

* another way using Welch-Satterthwaite approximation;
proc glm data=abt; class flux;
 model force=flux;
 means flux / welch;
run;
```

The "Null Model Likelihood Ratio Test" tests $H_0 : \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5$ by comparing two models, the model with different factor level variances, and the model with one overall variance.

The Type III test looks at $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, whether there are significant differences in group means. This is of course the same as testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ in the cell-means model.

What happens if you run the usual one-way ANOVA?

According to a Minitab White Paper by Rob Kelly:

- **Normality**: Simulations show accuracy with non-normal data with reasonably large sample sizes. If r is 2–9 levels, need $n_i \geq 15$. If r is 10-12 levels, need $n_i \geq 20$. The book notes the F test is generally robust to departures from normality
- **Non-constant variance**. The Type I error is best for equal group sizes: true α ranges from $0.02 - 0.08$ when $\alpha = 0.05$. Otherwise α goes as high as 0.22 in some cases. Again, the F test is generally robust, though sensitive in the case of unequal sample sizes.

If you have *many* comparisons to make, check out `proc multtest`. Among other techniques, `multtest` can implement the false discovery rate control approach in Benjamini and Hochberg (1995).

What about multiple comparisons?

The Games-Howell (Games and Howell, 1996) method has been shown to work very well for normal data with non-constant variance and different groups sizes as long as $n_i \geq 5$. Simulation can also work well.

```
proc mixed data=abt; class flux;
 model force=flux / ddfm=satterth; * denom. degrees freedom for MODEL and LSMEANS;
 repeated / group=flux;            *
 *ROW adjusts for varying df in pairwise differences;
 *SMM is more conservative than Tukey/Games-Howell;
 lsmeans flux / adjust=smm adjdfe=row;
run;

proc glimmix data=abt; class flux;
 model force=flux;
 random _residual_ / group=flux;
 covtest homogeneity;
 lsmestimate flux "1 vs 2" -1  1  0  0  0,
                  "3 vs 4"  0  0 -1  1  0 / adjust=simulate;
run;
```

What about multiple comparisons?

We can also follow up Kruskall-Wallis with two-sample Mann-Whitney-Wilcoxon tests, using Bonferroni to bound the FER. We need to have moderate to small r to achieve any power. A better approach is to use the `dscf` option in `proc npar1way` to compute the Dwass, Steel, Critchlow-Fligner multiple comparison analysis, based on pairwise two-sample standardized MWW comparisons.

```
proc npar1way data=abt dscf; class flux;
var force;
run;
```