STAT 705 Chapters 23 and 24: Two factors, unequal sample sizes; multi-factor ANOVA

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Balanced vs. unbalanced data

Balance is nice if calculating by hand! Typically, data are not balanced. Why?

- Observational studies – don't get to impose treatments on groups of same size.
- Subjects may "drop out" of a planned experiment.
- Cost considerations – some treatments more expensive.

Notation and model is exactly the same for balanced ($n_{ij} = n$) and unbalanced:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}.$$

Here, $i = 1, \dots, a$, $j = 1, \dots, b$, and $k = 1, \dots, n_{ij}$.

I have already covered two-way ANOVA assuming unbalanced data.

Fitting

The model is fit as a regression model. There are $(a-1)$ binary predictors for factor A, $(b-1)$ binary predictors for factor B, and $(a-1)(b-1)$ interaction predictors obtained by multiplying factor A predictors by factor B predictors. See example, pp. 954–957.

In general, SSTR $\neq$ SSA $+$ SSB $+$SSAB as defined in Chapter 19; orthogonality is lost in unbalanced designs.

Type III tests

We treat model as regression model with $(a-1)+(b-1)+(a-1)(b-1)=ab-1$ predictors, but we only test dropping blocks of predictors from this full model V corresponding to A, B, or AB, using general nested linear hypotheses ("big model / little model"), as in regression. Recall $n_T = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$. SAS gives Type III tests for

$$F_A = \frac{MSE(A|B, AB)}{MSE} \sim F(a-1, n_T - dfE) \text{ if } H_0 : \alpha_i = 0.$$

$$F_B = \frac{MSE(B|A, AB)}{MSE} \sim F(b-1, n_T - dfE) \text{ if } H_0 : \beta_i = 0.$$

$$F_{AB} = \frac{MSE(AB|A, B)}{MSE} \sim F((a-1)(b-1), n_T - dfE) \text{ if } H_0 : (\alpha\beta)_{ij} = 0.$$

Only the last test leaves a hierarchical model (additive model IV).

Modeling strategy

Say $a = 2$ and $b = 3$. If accept $H_0 : (\alpha\beta)_{ij} = 0$ then can look at, e.g., $L_1 = \beta_1 - \frac{1}{2}(\beta_2 + \beta_3)$ and $L_2 = \beta_2 - \beta_3$ via

```
lsmestimate B "L1" 1.0 -0.5 -0.5,
              "L2" 0.0  1.0 -1.0 / adjust=bonf;
```

If reject $H_0 : (\alpha\beta)_{ij} = 0$ then look at linear combinations of $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. For example, maybe $\mu_{21} - \mu_{11}$, $\mu_{22} - \mu_{12}$, and $\mu_{23} - \mu_{13}$ (differences in A over levels of B).

```
lsmestimate A*B "mu21-mu11" -1  0  0  1  0  0,
                "mu22-mu12"  0 -1  0  0  1  0,
                "mu23-mu13"  0  0 -1  0  0  1 / adjust=bonf;
```

Note that Tukey still works for pairwise comparisons, but $FER < \alpha$ rather than $FER = \alpha$.

Note: can also work directly with model parameters using `estimate`.

Bone growth, pp. 954–959

- Synthetic growth hormone given to $n_T = 14$ children, all hormone deficient and short.
- Y_{ijk} is difference in growth rate during month of treatment vs. previous non-treatment in cm/month.
- $i = 1, 2$ is gender (male/female) and $j = 1, 2, 3$ is bone development (severely depressed, moderately depressed, mildly depressed).
- No randomization of treatments employed; treatments (gender and level of depression) are observational here. Every child gets growth hormone.

Analysis in SAS

```
data growth;
input ratediff gender bonedev @@;
datalines;
  1.4  1  1  2.4  1  1  2.2  1  1  2.1  1  2  1.7  1  2  0.7  1  3  1.1  1  3
  2.4  2  1  2.5  2  2  1.8  2  2  2.0  2  2  0.5  2  3  0.9  2  3  1.3  2  3
;

* get interaction plot;
* table 23.4 (p. 959) is Type III SS Table;
proc glm plots=all;
 class gender bonedev;
 model ratediff=gender|bonedev;

* with p=0.80 we can drop the interaction and look at main effects;
proc glm plots=all;
 class gender bonedev;
 model ratediff=gender bonedev;
 lsmeans bonedev / pdiff adjust=tukey alpha=0.05 cl;

* removing gender as well;
proc glm plots=all;
 class gender bonedev;
 model ratediff=bonedev;
 lsmeans bonedev / pdiff adjust=tukey alpha=0.05 cl;
```

Chapter 24: Multi-factor studies

Say we have three factors: A, B, and C. A full, three-way interaction model is

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}.$$

Here $i = 1, \dots, a$, $j = 1, \dots, b$, and $k = 1, \dots, c$; have replicates $l = 1, \dots, n_{ijk}$. Balanced if $n_{ijk} = n$ for all i, j, k.

That's a lot of parameters!

SAS sets parameters equal to zero that have indices $i = a$, $j = b$, or $k = c$.

Section 24.2 has some text on interpreting the model; pp. 998–1002.

A model-based approach to multi-way ANOVA

Use interaction plots and Type III tests for find a simpler *hierarchical* model to explain the data. Check residuals vs. fitted values (heteroscedascity?), histogram of residuals (skew? bimodality?), and normal probability plot to assess model adequacy. Decide what sorts of paired differences or linear combinations you want to look at.

For example, if end up with A, B, C, and B*C, you can look at main effects in A, and B*C interaction effects. These might include looking at all differences in main effects of A $\bar{\mu}_{i_1 \bullet\bullet} - \bar{\mu}_{i_2 \bullet\bullet} = \alpha_{i_1} - \alpha_{i_2}$ (use Tukey), and looking at slices $\bar{\mu}_{\bullet j_1 k} - \bar{\mu}_{\bullet j_2 k}$ for pairs $1 \leq j_1 < j_2 \leq b$.

The lowest order interactions in the effect left in the model determine which pairwise differences make sense to look at!

Hierarchical model building

Recall with hierarchical model building, if we have an interaction, we must include all lower order effects that comprise the interaction. So if we have a three way interaction $A * C * D$, we must also include the effects A, C, D, $A * C$, $A * D$, and $C * D$. In SAS this is accomplished including $A|C|D$ in the `model` statement.

A reasonable approach to model building is pare down higher order interactions until you have a model with largely significant effects in it, i.e. "backwards elimination." This approach incurs the problem of multiple hypothesis testing, but can be somewhat eleviated using Kimball's inequality, or else by considering one overall test for dropping several effects at once; I suggest the latter.

Averaged effects

Regardless of the final model chosen, one can always resort to the examination of so-called "averaged effects." Let's consider three factors A, B, and C for simplicity. The averaged effect for $A = i$ is given by

$$\bar{\mu}_{i\bullet\bullet} = \frac{1}{bc} \sum_{j=1}^{b} \sum_{k=1}^{c} \mu_{ijk},$$

where $\mu_{ijk} = E(Y_{ijkm})$ under your final model. This is the mean response at $A = i$ averaged over the levels of B and C, and are provided by SAS `lsmeans A`. This averaging assumes that all factors levels are "weighted equally."

Differences in averaged effects

We may furthermore look at differences in averaged effects, e.g. $\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet}$. These are also interpreted as *treatment differences averaged over the other effects*, e.g.

$$\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c}[\mu_{2jk} - \mu_{1jk}].$$

You can obtain these from adding `pdiff` to your `lsmeans` statement. Both `lsmeans` and `lsmestimate` deal with *averaged effects*. The rest of the averaged effects for the three-factor model are

$$\bar{\mu}_{\bullet j\bullet} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c}\mu_{ijk}, \quad \bar{\mu}_{\bullet\bullet k} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ijk},$$

$$\bar{\mu}_{ij\bullet} = \frac{1}{c}\sum_{k=1}^{c}\mu_{ijk}, \quad \bar{\mu}_{i\bullet k} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ijk}, \quad \bar{\mu}_{\bullet jk} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ijk}.$$

You can look at these effects and obtain pairwise differences by including, e.g. `lsmeans A*B / pdiff adjust=tukey cl;`

Simplification when higher order interactions are dropped

Note, if A does not share any interactions with other factors, e.g. the model $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk}$ fits, then $\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet} = \alpha_2 - \alpha_1$. This idea generalizes to the other factors as well.

However, if the model $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk}$ fits, then $\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet} \neq \alpha_2 - \alpha_1$. In fact, $\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet} = \alpha_2 - \alpha_1 + \frac{1}{b}\sum_{j=1}^{b}(\alpha\beta)_{2j} - \frac{1}{b}\sum_{j=1}^{b}(\alpha\beta)_{1j}$.

In general, differences in the A treatments can vary across the other two factors in a complex way. Ideally, one would then look at, e.g. $\bar{\mu}_{2jk} - \bar{\mu}_{1jk}$ for different values of j and k to see where treatment A differences occur. It can happen that the averaged difference $\bar{\mu}_{2\bullet\bullet} - \bar{\mu}_{1\bullet\bullet}$ is *not significantly non-zero*, yet one or more of the $\bar{\mu}_{2jk} - \bar{\mu}_{1jk}$ *are* significantly non-zero. You can examine the individual differences using `estimate`.

Example where $a = b = c = 2$

Say through backward elimination, the model A, B, C, $A * B$, $B * C$ is shown to adequately describe the data; i.e. $\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk}$. Then in SAS the $A * B$ effects are listed in the order $(A, B) = (1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$, same with the $B * C$ effects. Since A does not interact with C, we *only need to examine A differences over B*. To visualize this, note that *for any k*

$$\begin{aligned} \mu_{2jk} - \mu_{1jk} &= \mu + \alpha_2 + \beta_j + \gamma_k + (\alpha\beta)_{2j} + (\beta\gamma)_{jk} - [\mu + \alpha_1 + \beta_j + \gamma_k + (\alpha\beta)_{1j} + (\beta\gamma)_{jk}] \\ &= \alpha_2 - \alpha_1 + (\alpha\beta)_{2j} - (\alpha\beta)_{1j}, \end{aligned}$$

which is independent of k, i.e. independence of factor C.

There are only two of these to look at, namely $\alpha_2 - \alpha_1 + (\alpha\beta)_{21} - (\alpha\beta)_{11}$, how treatment A differs when $B = 1$, and $\alpha_2 - \alpha_1 + (\alpha\beta)_{22} - (\alpha\beta)_{12}$, how treatment A differs when $B = 2$.

Example, continued

You can get these either out of `estimate` directly, or `lsmestimate` by noting that for this model $\bar{\mu}_{2j\bullet} - \bar{\mu}_{1j\bullet} = \mu_{2jk} - \mu_{1jk} = \alpha_2 - \alpha_1 + (\alpha\beta)_{2j} - (\alpha\beta)_{1j}$.

The command `estimate` works with all of the effects in the model that you list, i.e. the α_i's, β_j's, γ_k's, $(\alpha\beta)_{ij}$'s, etc., whereas `lsmestimate` works with the averaged effects $\bar{\mu}_{i\bullet\bullet}$, $\bar{\mu}_{\bullet j\bullet}$, $\bar{\mu}_{\bullet\bullet k}$, $\bar{\mu}_{ij\bullet}$, etc.

Final comments

Output from `lsmestimate` is essentially *always interpratable*, either as a conditional or averaged linear combination, depending on the interaction structure. Similarly, output from `lsmeans` is also interpretable in terms of averaged effects. Output from `estimate` will be interpretable if you are careful.

Don't be afraid to *write down the model* and play around with the math. This is how you can find out when `estimate` and `lsmestimate` give you the same results!

Interaction plots for multi-way models

For multi-factor models, we can look at averaged (or marginal) interaction plots obtained in `proc glm` by simply fitting a model with only two of the factors, e.g. get each of `model=A|B;`, `model=A|C;`, and `model=B|C;`

It is also possible to get conditional interaction plots directly out of SAS.

Say you have three factors, A, B, and C, each with two levels. The averaged plot for A and B uses $\bar{Y}_{ij\bullet\bullet}$; there are two conditional plots for A and B, one at $k = 1$ uses $\bar{Y}_{ij1\bullet}$ and the other at $k = 2$ uses $\bar{Y}_{ij2\bullet}$. Averaged plots can tell you whether two-way interactions are necessary; conditional plots can tell you whether two-way and higher interactions are necessary, but are a pain to interpret without some practice. See Section 24.2 (pp. 998–1000).

Averaged interaction plots for some models

Model A B C has A/B, A/C, and B/C averaged plots

$$\begin{aligned}
\bar{\mu}_{ij\bullet} &= \mu + \alpha_i + \beta_j + \bar{\gamma} \text{ parallel} \\
\bar{\mu}_{i\bullet k} &= \mu + \alpha_i + \bar{\beta} + \gamma_k \text{ parallel} \\
\bar{\mu}_{\bullet jk} &= \mu + \bar{\alpha} + \beta_j + \gamma_k \text{ parallel}
\end{aligned}$$

Model A B C A*B has A/B, A/C, and B/C averaged plots

$$\begin{aligned}
\bar{\mu}_{ij\bullet} &= \mu + \alpha_i + \beta_j + \bar{\gamma} + (\alpha\beta)_{ij} \text{ not parallel} \\
\bar{\mu}_{i\bullet k} &= \mu + \alpha_i + \bar{\beta} + \gamma_k + (\overline{\alpha\beta})_{i\bullet} \text{ parallel} \\
\bar{\mu}_{\bullet jk} &= \mu + \bar{\alpha} + \beta_j + \gamma_k + (\overline{\alpha\beta})_{\bullet j} \text{ parallel}
\end{aligned}$$

Model A B C A*B B*C has A/B, A/C, and B/C averaged plots

$$\begin{aligned}
\bar{\mu}_{ij\bullet} &= \mu + \alpha_i + \beta_j + \bar{\gamma} + (\alpha\beta)_{ij} + (\overline{\beta\gamma})_{j\bullet} \text{ not parallel} \\
\bar{\mu}_{i\bullet k} &= \mu + \alpha_i + \bar{\beta} + \gamma_k + (\overline{\alpha\beta})_{i\bullet} + (\overline{\beta\gamma})_{\bullet k} \text{ parallel} \\
\bar{\mu}_{\bullet jk} &= \mu + \bar{\alpha} + \beta_j + \gamma_k + (\overline{\alpha\beta})_{\bullet j} + (\beta\gamma)_{jk} \text{ not parallel}
\end{aligned}$$

Body fat

Effects of gender (A), body fat (%, B), and smoking history (C) of subjects on exercise tolerance Y_{ijkl} is minutes of bicycling until fatigue, were measured in small study of $n_T = 24$ subjects 25–35 years old.

Study happens to be balanced. Partial analysis on pp. 1005–1012.

Time until fall off bike, SAS analysis

```
data tol; * 1=male vs. female, 1=low fat vs. high, 1=light smoking vs. heavy;
 input tol gender fat smoking @@;
datalines;
  24.1  1  1  1  29.2  1  1  1  24.6  1  1  1
  20.0  2  1  1  21.9  2  1  1  17.6  2  1  1
  14.6  1  2  1  15.3  1  2  1  12.3  1  2  1
  16.1  2  2  1   9.3  2  2  1  10.8  2  2  1
  17.6  1  1  2  18.8  1  1  2  23.2  1  1  2
  14.8  2  1  2  10.3  2  1  2  11.3  2  1  2
  14.9  1  2  2  20.4  1  2  2  12.8  1  2  2
  10.1  2  2  2  14.4  2  2  2   6.1  2  2  2
;

* "conditional" interaction plot;
proc sgpanel;
 panelby gender / rows=1 columns=2;
 scatter x=fat y=tol / group=smoking;
 reg x=fat y=tol / group=smoking;

* fat*smoking averaged over gender;
proc sgplot;
 title "averaged over gender";
 scatter x=fat y=tol / group=smoking;
 reg x=fat y=tol / group=smoking;

* can also get them the usual way through proc glm;
* fat by gender interaction probably not needed;
proc glm plots=all; class gender fat smoking; model tol=gender|fat;

* gender by smoking interaction probably not needed;
proc glm plots=all; class gender fat smoking; model tol=gender|smoking;

* fat by smoking interaction probably needed;
proc glm plots=all; class gender fat smoking; model tol=fat|smoking;
```

One overall F-test for dropping effects

```
* saturated model with all possible interactions;
* we can drop any one of gender*fat, gender*smoking, or gender*fat*smoking;
proc glm data=tol outstat=full;
class gender fat smoking;
model tol=fat|smoking|gender / solution;
run;

* let's see if we can drop all three of these effects at the same time;
proc glm data=tol outstat=reduced;
class gender fat smoking;
model tol=gender fat|smoking;
run;

* p-value for nested hypothesis in SAS;
data test1; set full reduced; if _SOURCE_="ERROR";  * get error SS and DF;
proc means data=test1 noprint; var ss df; output out=test2 min=minss mindf max=maxss maxdf;
data nested;  set test2; fstar=((maxss-minss)/(maxdf-mindf))/(minss/mindf);
 pvalue=1-cdf('f',fstar,maxdf-mindf,mindf);
proc print; run;
```

Analysis of reduced model

```
proc glm plots=all;
 class gender fat smoking;
 model tol=gender fat|smoking;
 lsmeans gender / pdiff adjust=tukey alpha=0.05 cl;
 lsmeans fat*smoking / pdiff adjust=tukey alpha=0.05 cl;

proc glimmix;
 class gender fat smoking;
 model tol=gender fat|smoking;
 lsmestimate gender "gender" 1 -1 / adjust=t alpha=0.05 cl;
 lsmestimate fat*smoking "diff1" -1  1  0  0,
                         "diff2"  0  0 -1  1 / adjust=bon alpha=0.05 cl;

* for illustration consider a model with G F S G*F and F*S;
* since gender DOES NOT interact with smoking, gender *differences* only change with fat type;
proc glimmix;
 class gender fat smoking;
 model tol=gender fat smoking gender*fat fat*smoking;
 estimate "gender diff @ fat=1" gender -1  1  gender*fat -1  0  1  0;
 estimate "gender diff @ fat=2" gender -1  1  gender*fat  0 -1  0  1;
 lsmestimate gender*fat "gender diff @ fat=1" -1  0  1  0; * same as above!;
 lsmestimate gender*fat "gender diff @ fat=1"  0 -1  0 -1;
```