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

Adapted from 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 – we don't get to impose treatments on groups of same size.
- Subjects may "drop out" of a planned experiment.
- Cost considerations – some treatments are more expensive.
- Sample sizes are chosen to be representative of a factor level's importance or presence in a population

The notation and model are exactly the same for balanced ($n_{ij} = n$) and unbalanced data:

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

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

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 the model as a 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 ("full model / reduced model"), as in regression. Recall $n_T = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$. SAS gives Type III tests for

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

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

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

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

Modeling strategy

Say $a = 2$ and $b = 3$. If we accept $H_0 : (\alpha\beta)_{ij} = 0$, then we 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 we 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: we can also work directly with model parameters using `estimate`, but we need to be careful.

Teaching fractions

- Five teachers in a Pee Dee school district were assigned three different methods to teach fractions; the experimental unit was a class of students.
- Y_{ijk} is the class average improvement on a pre- and post-test.
- $i = A, B, C$ is method and $j = 1, 2, 3, 4, 5$ is teacher
- The study is based on an experiment that was poorly randomized, and badly balanced with missing cells. Teacher is actually more of a blocking variable here.

Analysis in SAS

```
data fraction;
input teacher method $ diff @@;
datalines;
 1 a 10 1 a  7 1 b  4 1 c  9 2 a  6 2 b  7 2 c 13
 3 a 11 3 b  5 3 c 16 4 a  6 4 b  7 4 b  8 4 c 14
 5 a  6 5 b  3 5 c  2
;
run;

* Only two df available for error;
proc glm data=fraction plots=all;
 class teacher method;
 model diff=teacher|method;
 run;

* Drop interaction and focus on teaching method;
proc glm plots=all;
class teacher method;
model diff=teacher method;
 lsmeans method / pdiff adjust=tukey alpha=0.05 cl;
 run;
```

Missing/empty cells: $n_{ij} = 0$

Missing or empty cells are a special form of unbalanced factorial studies that require a different approach. E.g., for a two-way factorial study, some marginal means can no longer be estimated, and Type III analyses can be difficult to interpret (and even irrelevant). Simple remedies include:

- Construct "sliceby" contrasts that study one factor's effects within levels of the other factor
- Study Type IV hypotheses (which are based on the types of contrasts mentioned above)
- For additive models, the usual analysis is possible provided designs are connected

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 $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, c$; we have replicates $l = 1, \ldots, n_{ijk}$. The design is 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 to 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 you 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, $\mu_{i_1 \cdot \cdot} - \mu_{i_2 \cdot \cdot} = \alpha_{i_1} - \alpha_{i_2}$ (use Tukey), and looking at slices $\mu_{\cdot j_1 k} - \mu_{\cdot j_2 k}$ 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 alleviated 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

$$\mu_{i\cdot\cdot} = \frac{1}{bc} \sum_{j=1}^{b} \sum_{k=1}^{c} \mu_{ijk},$$

where $\mu_{ijk} = E(Y_{ijkl})$ under your final model. This is the mean response at $A = i$ averaged over the levels of B and C, and is 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. $\mu_{2..} - \mu_{1..}$. These are also interpreted as *treatment differences averaged over the other effects*, e.g.

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

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

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

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

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 $\mu_{2..} - \mu_{1..} = \alpha_2 - \alpha_1$. 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 $\mu_{2..} - \mu_{1..} \neq \alpha_2 - \alpha_1$. In fact, $\mu_{2..} - \mu_{1..} = \alpha_2 - \alpha_1 + \frac{1}{b}\sum_{j=1}^{b}(\alpha\beta)_{2j} - \frac{1}{b}\sum_{j=1}^{b}(\alpha\beta)_{1j}$.

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. $\mu_{2jk} - \mu_{1jk}$ for different values of j and k to see where treatment A differences occur. It can happen that the averaged difference $\mu_{2..} - \mu_{1..}$ is *not significantly non-zero*, yet one or more of the $\mu_{2jk} - \mu_{1jk}$ *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 $(2, 2)$, as are 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 obtain these either from `estimate` directly, or `lsmestimate` by noting that for this model
$\mu_{2j\cdot} - \mu_{1j\cdot} = \mu_{2jk} - \mu_{1jk} = \alpha_2 - \alpha_1 + (\alpha\beta)_{2j} - (\alpha\beta)_{1j}$.

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 $\mu_{i\cdot\cdot}$, $\mu_{\cdot j\cdot}$, $\mu_{\cdot\cdot k}$, $\mu_{ij\cdot}$, 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. look at each of `model=A|B;`, `model=A|C;`, and `model=B|C;`

It is also possible to obtain 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 $\bar{Y}_{ij\cdot\cdot}$; there are two conditional plots for A and B, one at $k = 1$ uses $\bar{Y}_{ij1\cdot}$ and the other at $k = 2$ uses $\bar{Y}_{ij2\cdot}$. 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

$$\begin{aligned}
\mu_{ij\cdot} &= \mu_{\cdots} + \alpha_i + \beta_j + \bar{\gamma} \text{ parallel} \\
\mu_{i\cdot k} &= \mu_{\cdots} + \alpha_i + \bar{\beta} + \gamma_k \text{ parallel} \\
\mu_{\cdot jk} &= \mu_{\cdots} + \bar{\alpha} + \beta_j + \gamma_k \text{ parallel}
\end{aligned}$$

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

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

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

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

Body fat

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

The 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;
```