Sections 27.3-4: Repeated measures models with two factors

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Recall a simple repeated measures model

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \epsilon_{ij}.$$

A repeated measures model with two factors, assuming the factors interact, is

$$Y_{ijk} = \mu_{..} + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}.$$

Here, there are $i = 1, \ldots, s$ experimental units, $j = 1, \ldots, a$ levels of A, and $k = 1, \ldots, b$ levels of B.

The random effects are assumed normal $\rho_1, \ldots, \rho_s \overset{iid}{\sim} N(0, \sigma_\rho^2)$ and independent of the errors ϵ_{ijk}.

In Section 27.4, the text studies the case in which both factors are repeated measures factors, but uses a model with additional terms:

$$Y_{ijk} = \mu_{...} + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + (\rho\alpha)_{ij} + (\rho\beta)_{ik} + \epsilon_{ijk}.$$

There are three sets of random effects: $\rho_i \overset{iid}{\sim} N(0, \sigma_\rho^2)$, $(\rho\alpha)_{ij} \overset{iid}{\sim} N(0, \sigma_\rho^2)$, $(\rho\beta)_{ik} \overset{iid}{\sim} N(0, \sigma_\rho^2)$, all independent of the errors ϵ_{ijk}.

Most modern approaches to repeated measures analysis group the random effects into the error term, which has s(ab-1) degrees of freedom. In the book's approach, each fixed effect is tested against its interaction with the subject effect.

Example where $n = 10$, $a = 2$, and $b = 3$

$ab = 6$ pieces of data collected from each subject.

	Factor A					
	$j = 1$			$j = 2$		
	Factor B			Factor B		
Block	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1	Y_{111}	Y_{112}	Y_{113}	Y_{121}	Y_{122}	Y_{123}
2	Y_{211}	Y_{212}	Y_{213}	Y_{221}	Y_{222}	Y_{223}
3	Y_{311}	Y_{312}	Y_{313}	Y_{321}	Y_{322}	Y_{323}
4	Y_{411}	Y_{412}	Y_{413}	Y_{421}	Y_{422}	Y_{423}
5	Y_{511}	Y_{512}	Y_{513}	Y_{521}	Y_{522}	Y_{523}
6	Y_{611}	Y_{612}	Y_{613}	Y_{621}	Y_{622}	Y_{623}
7	Y_{711}	Y_{712}	Y_{713}	Y_{721}	Y_{722}	Y_{723}
8	Y_{811}	Y_{812}	Y_{813}	Y_{821}	Y_{822}	Y_{823}
9	Y_{911}	Y_{912}	Y_{913}	Y_{921}	Y_{922}	Y_{923}
10	$Y_{10,11}$	$Y_{10,12}$	$Y_{10,13}$	$Y_{10,21}$	$Y_{10,22}$	$Y_{10,23}$

Randomized complete block design; each block gets every treatment combination.

Example

A national retail chain wants to study the effect on advertising campaign (factor A, two levels) on the volume of athletic shoe sales over time (factor B, three levels). Ten similar test markets (blocks) were chosen at random to participate in the study.

The campaigns were similar in all respects except a different sports personality was used in each, $j = 1, 2$. Sales data were collected for three two-week periods: $k = 1, 2, 3$ for two weeks prior to campaign, two weeks during campaign, and two weeks after campaign over.

To complete a full repeated measures design requires two entire six-week periods, or 12 weeks (3 months!) total. Instead, the retail chain cut expenses in half by (randomly) assigning each test market *only one* of the two possible campaigns, at half the time and cost. This results in an randomized *incomplete* block design.

Thus we have repeated measures on only one factor, here time.

	Campaign					
	$j = 1$			$j = 2$		
	Time			Time		
Block	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1	Y_{111}	Y_{112}	Y_{113}			
2	Y_{211}	Y_{212}	Y_{213}			
3	Y_{311}	Y_{312}	Y_{313}			
4	Y_{411}	Y_{412}	Y_{413}			
5	Y_{511}	Y_{512}	Y_{513}			
6				Y_{621}	Y_{622}	Y_{623}
7				Y_{721}	Y_{722}	Y_{723}
8				Y_{821}	Y_{822}	Y_{823}
9				Y_{921}	Y_{922}	Y_{923}
10				$Y_{10,21}$	$Y_{10,22}$	$Y_{10,23}$

The model is *the same* as a RCBD design, but factor A (the campaign effect) is confounded (is it serious?):

$$Y_{ijk} = \mu_{...} + \underbrace{\rho_i}_{\text{market}} + \underbrace{\alpha_j + \beta_k + (\alpha\beta)_{jk}}_{\text{campaign \& time}} + \epsilon_{ijk}.$$

Randomized incomplete block design: blocks receive only one level of A, but all levels of B.

Comment on these types of designs

When A is a characteristic of the blocks (if human, called subjects), such as salary range or gender then randomization is obviously not necessary, or even possible. In this case factor A is said to be "observational." We will briefly discuss this soon in the context of Section 27.3.

When B is time, as in longitudinal studies, randomization is also not possible. In fact, it is typically of interest to discern time effects, i.e. effect of drug over time.

Inferential focus

Under the model $\mu_{jk} = E(Y_{ijk}) = \mu_{...} + \alpha_j + \beta_k + (\alpha\beta)_{jk}$ because $E(\epsilon_{ijk}) = E(\rho_i) = 0$.

Look at the Type III test of $H_0 : (\alpha\beta)_{jk} = 0$. If you accept this, then refit the model without the A*B interaction and focus on the main effects for A and the main effects for B as usual.

If you reject $H_0 : (\alpha\beta)_{jk} = 0$, consider looking at slices.

The usual tools `estimate`, `lsmeans`, and `lsmestimate` are at your disposal in SAS. Do not include the blocks, they are random!

If the additive model fits you will use, e.g., `lsmeans A;` or `lsmeans B;` If the interaction model fits, use something like `slice A*B / sliceby=B adjust=tukey;`

A note on longitudinal study inference

In *longitudinal studies* where Factor B is time, it is common to focus on how factor A levels differ at each of the b time points $k = 1, \ldots, b$.

In these studies, $k = 1$ often corresponds to baseline, before differences in factor A levels have had a chance to "kick in." In this case, we *expect* to see no difference at time $k = 1$, i.e. accept $H_0 : \mu_{j_1 1} - \mu_{j_2 1} = 0$, but do hope to see differences at later times. You may want to use observations at $k = 1$ to develop a differential response measured on a relative scale.

You do not want to use "averaged effects" here such as $\bar{\mu}_{j_1 \cdot} - \bar{\mu}_{j_2 \cdot}$ because the initial absence of a baseline effect can wash out later differences in A.

Note that if factor A differences have *leveled off* for $k \geq k^*$ then you may want to look at an averaged effect $\frac{1}{b - k^* + 1} \sum_{k=k^*}^{b} [\mu_{j_1 k} - \mu_{j_2 k}]$. A graph here helps.

Conditional plot

Your book favors a particular conditional plot, namely Y_{ijk} versus k (factor B) for each block i, stratified by j (Factor A);

All of the a different panels will look like the same function shifted up and down if $H_0 : (\alpha\beta)_{jk} = 0$ is true. If the interaction is *not* zero, then each panel will have identical vertically shifted curves, but the shape of the curves will change from panel to panel (will change with factor A).

Just as in RCB designs with random blocks, there are two sets of residuals that should be checked for normality:

- $e_{ijk} = Y_{ijk} - [\hat{\mu}_{\cdots} + \hat{\rho}_i + \hat{\alpha}_j + \hat{\beta}_k + \widehat{(\alpha\beta)}_{jk}]$.
- $\hat{\rho}_i$.

Furthermore, the $\{e_{ijk}\}$ should show constant variance when plotted against the (conditional) fitted values $\hat{Y}_{ijk}$, or any of i, j, or k.

Campaign data in SAS

```
data sales;
input sales market campaign time @@;
datalines;
    958  1  1  1    1005  2  1  1     351  3  1  1     549  4  1  1     730  5  1  1
   1047  1  1  2    1122  2  1  2     436  3  1  2     632  4  1  2     784  5  1  2
    933  1  1  3     986  2  1  3     339  3  1  3     512  4  1  3     707  5  1  3
    780  6  2  1     229  7  2  1     883  8  2  1     624  9  2  1     375 10  2  1
    897  6  2  2     275  7  2  2     964  8  2  2     695  9  2  2     436 10  2  2
    718  6  2  3     202  7  2  3     817  8  2  3     599  9  2  3     351 10  2  3
;

* figure 27.6 (p. 1146);
proc sgpanel;
 panelby campaign / rows=1 columns=2;
 series x=time y=sales / group=market;

* test that H0: sigma_rho=0 is a bit different than book;
* can (sequentially) drop both campaign and campaign*time;
* can formally and elegantly test this with one F-test via contrast, but not easy;
proc mixed plots=all;              * conditional residual plots are what you want;
 class market campaign time;
 model sales=campaign|time;        * factorial structure;
 random market;                    * 10 distinct markets each with own index;
 lsmeans time / pdiff adjust=tukey cl;
```

Given that we only need the β_k's, it may make sense to look at
$\beta_1 - \beta_3$ and $\beta_2 - 0.5(\beta_1 + \beta_3)$. Why?

A note on notation

Your textbook makes use of *nesting* in Section 27.3. In the Feral Hog Rooting example, we had four habitats with three plots in each habitat. Plots 1,2, and 3 in a given habitat (e.g., Cypress-Tupelo Swamp) are not associated by level with plots 1, 2 and 3 in a different habitat (Bottomland Hardwood Forest); plot i needs to be tagged by habitat j.

The random effects are now denoted by $\rho_{i(j)}$. The analysis substantively changes since $\rho_{i(j)}$ measures between-plot error, and ϵ_{ijk} measure within-plot error.

$$\underbrace{Y_{ijk}}_{\text{disturbance}} = \mu_{...} + \underbrace{\alpha_j}_{\text{habitat}} + \underbrace{\rho_{i(j)}}_{\text{plot}} + \underbrace{\beta_k + (\alpha\beta)_{jk}}_{\text{time/time*habitat}} + \epsilon_{ijk}.$$

Longitudinal data analysis

Analysis via mixed effects models on repeated measures taken over time on individuals is termed *longitudinal data analysis*. This course is offered every Spring as STAT 771.

The campaign data is an example of longitudinal data analysis. We can group the three repeated measurements from the ith market receiving campaign j as a vector

$$\mathbf{Y}_{ij} = \left[\begin{array}{c} Y_{ij1} \\ Y_{ij2} \\ Y_{ij3} \end{array} \right] \stackrel{ind.}{\sim} N_3 \left(\left[\begin{array}{c} \mu_{...} + \alpha_j + \beta_1 + (\alpha\beta)_{j1} \\ \mu_{...} + \alpha_j + \beta_2 + (\alpha\beta)_{j2} \\ \mu_{...} + \alpha_j + \beta_3 + (\alpha\beta)_{j3} \end{array} \right], \left[\begin{array}{ccc} \sigma_\rho^2 + \sigma^2 & \sigma_\rho^2 & \sigma_\rho^2 \\ \sigma_\rho^2 & \sigma_\rho^2 + \sigma^2 & \sigma_\rho^2 \\ \sigma_\rho^2 & \sigma_\rho^2 & \sigma_\rho^2 + \sigma^2 \end{array} \right] \right).$$

The model is fit using MLE or REML in `proc mixed`; the latter disallows easy comparison among models via AIC.

An older approach to estimation and testing for this model is available in `proc glm`.

Multivariate analysis of variance (MANOVA)

The model on the previous slide can be recast as

$$\mathbf{Y}_{1j}, \ldots, \mathbf{Y}_{n_j,j} \overset{iid}{\sim} N_b(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ has *exchangeable* structure parameterized by the variance components σ_ρ^2 and σ^2. A test of no factor A effect is simply $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_a$.

If we allow $\boldsymbol{\Sigma}$ to be completely arbitrary (except that it is positive definite and symmetric), this model is a MANOVA model. This can be fit in `proc glm` by adding a `manova` subcommand, e.g. `manova h=A;`. Need to have the data as one row per experimental unit and have something like
`model sales1 sales2 sales3=campaign / nouni;`

MANOVA is covered in more detail in STAT 730 and STAT 771.

A bit on fitting

The mean parameters are the $\mu...$, α_j, β_k, and $(\alpha\beta)_{jk}$. The variance components are σ^2 and σ^2_ρ.

REML (restricted maximum likelihood) essentially uses OLS to estimate the mean parameters, then uses these estimated mean parameters to estimate mean-zero residuals, then uses maximum likelihood to estimate variance components from the residuals. The variance component estimates are then used in GLS (more general than WLS) to re-estimate the mean parameters. This results in unbiased estimates of variance components. More on this in STAT 714.

I've seen versions that plug in WLS estimates of the mean parameters assuming the variance components are known to create a likelihood for the variance components from the residuals. The mean parameters are then estimated using GLS, as above.

Maximum likelihood simply estimates both the mean parameters and variance components at the same time using maximum likelihood.

In either case the $\rho_1, \ldots, \rho_n$ *are not part of the likelihood*. These can be estimated after the population parameters are estimated from either method using Bayes rule; they are called "BLUPs" for best linear unbiased predictor.

A bit on fitting

- PROC MIXED models repeated measures effects with a _REPEATED_ statement, while GLIMMIX uses RANDOM _RESIDUAL_.
- The default fitting method for PROC MIXED for the normal linear mixed model is METHOD=REML. PROC GLIMMIX uses METHOD=RSPL; these are equivalent.
- METHOD=MSPL in PROC GLIMMIX is equivalent to METHOD=ML in PROC MIXED.
- PROC GLIMMIX has more optimization methods available, and uses a different default optimization method from PROC MIXED.

A bit on fitting

- PROC GLM uses Method-of-Moments to estimate variance components. It constructs appropriate F-tests, but doesn't build randomness of effects into tests or estimation of main effects, lsmeans, contrasts, etc.
- PROC MIXED has some more complex covariance structures (Kronecker-type for spatiotemporal models) that GLIMMIX lacks.
- Some of PROC GLIMMIX's most useful features are not available in PROC MIXED: COVTEST, LSMESTIMATE, OUTPUT.