Repeated measures on one factor

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 506: Introduction to Experimental Design

Repeated measures model with two factors

Recall a simple repeated measures model from last time

$$y_{ij} = \mu + \rho_i + \tau_j + \epsilon_{ij}.$$

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 $i = 1, \ldots, n$ experimental units, $j = 1, \ldots, a$ levels of A, and $k = 1, \ldots, b$ levels of B.

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

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 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.

Repeated measures on time, not campaign

	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 *the same* as a RCBD; there's just less information for factor A (the campaign effect):

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

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."

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 – more coming up.

Note that this is a *nested* design; markets are nested within campaign (more coming up). It makes sense to consider the market effects as random (but not strictly necessary).

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 tool `lsmeans`, `pairs`, `contrast`, `pairwise` (if additive model with no random effects is being fit), etc. are at your disposal in R. Do not include the blocks, they are random!

If the additive model fits you can, e.g., consider pairwise comparisons of A and B. If the interaction model fits, use slices, i.e. differences in A for each level of B.

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 do not want to use "averaged effects" here such as $\bar{\mu}_{j_1 \bullet} - \bar{\mu}_{j_2 \bullet}$ 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 picture here helps.

Diagnostics

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

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

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 R

```
library(cfcdae); library(lsmeans); library(lme4); library(car)
sales=c( 958,1005, 351, 549, 730,1047,1122, 436, 632, 784, 933, 986, 339, 512, 707,
         780, 229, 883, 624, 375, 897, 275, 964, 695, 436, 718, 202, 817, 599, 351)
market=factor(c(rep(1:5,3),rep(6:10,3)))
campaign=factor(rep(1:2,each=15))
time=factor(c(rep(1:3,each=5),rep(1:3,each=5)))
d=data.frame(sales,market,campaign,time)
with(d,interactplot(time,campaign,sales,confidence=0.95))

f1=lmer(sales~campaign*time+(1|market),REML=F)
Anova(f1,type=3)
f2=lmer(sales~time+(1|market),REML=F)
Anova(f2,type=3)
anova(f2,f1) # note little 'a'

pairs(lsmeans(f2,"time"))
```

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

The shoe data/design are *nested*: there are 10 unique markets. Standard notation takes their random effects to simply be $\rho_1, \rho_2, \rho_3, \rho_4, \rho_5, \rho_6, \rho_7, \rho_8, \rho_9, \rho_{10}$.

We can also associate five markets with campaign $i = 1$, and the other five with campaign $i = 2$. There are still 10 unique markets, but we've now renumbered them 1,2,3,4,5 within each campaign. The same random effects listed above now become $\rho_{1(1)}, \rho_{2(1)}, \rho_{3(1)}, \rho_{4(1)}, \rho_{5(1)}, \rho_{1(2)}, \rho_{2(2)}, \rho_{3(2)}, \rho_{4(2)}, \rho_{5(2)}$. The model and analysis stays the same; what mainly changes is how the data are coded and the addition of "nested notation" in SAS.

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

Campaign data, nested

	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_{121}	y_{122}	y_{123}
7				y_{221}	y_{222}	y_{223}
8				y_{321}	y_{322}	y_{323}
9				y_{421}	y_{422}	y_{423}
10				y_{521}	y_{522}	y_{523}

Making use of nested notation. Here

$$\rho_{1(1)} = \rho_1, \ \rho_{2(1)} = \rho_2, \ \rho_{3(1)} = \rho_3, \ \rho_{4(1)} = \rho_4, \ \rho_{5(1)} = \rho_5,$$

$$\rho_{1(2)} = \rho_6, \ \rho_{2(2)} = \rho_7, \ \rho_{3(2)} = \rho_8, \ \rho_{4(2)} = \rho_9, \ \rho_{5(2)} = \rho_{10},$$

from previous notation.

Campaign data in R, nested

Market now goes from 1 to 5 instead of 1 to 10. Still 10 unique markets though.

```
market=factor(rep(1:5,6))
f=aov(sales~campaign*time+campaign/market)
anova(f)
```

This treats the market effects as fixed, not random:

$$y_{ijk} = \mu + \rho_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}.$$

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

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

$$\mathbf{y}_{ij} = \begin{bmatrix} y_{ij1} \\ y_{ij2} \\ y_{ij3} \end{bmatrix} \overset{ind.}{\sim} N_3 \left(\begin{bmatrix} \mu + \alpha_j + \beta_1 + (\alpha\beta)_{j1} \\ \mu + \alpha_j + \beta_2 + (\alpha\beta)_{j2} \\ \mu + \alpha_j + \beta_3 + (\alpha\beta)_{j3} \end{bmatrix}, \begin{bmatrix} \sigma^2_\rho + \sigma^2 & \sigma^2_\rho & \sigma^2_\rho \\ \sigma^2_\rho & \sigma^2_\rho + \sigma^2 & \sigma^2_\rho \\ \sigma^2_\rho & \sigma^2_\rho & \sigma^2_\rho + \sigma^2 \end{bmatrix} \right).$$

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^2_ρ 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. MANOVA is covered in more detail in STAT 730 and STAT 771.

Split plot designs

Much of statistics came from agricultural experiments. The term "split plot" refers to one type of experiment that leads to repeated measures within one factor.

Feasible to treat entire field with one level of factor A (irrigation method, insecticide sprayed by plane, et cetera). Sometimes fields separated by many miles or many hundreds of miles.

Can further split up fields (split the plots of land) into pieces for administration of factor B (seed type, hybrid/non-hybrid, weeded or not, et cetera).

y_{ijk} is then yield of whatever grown. The $\rho_1, \ldots, \rho_n$ are field-specific proxies for combined overall, unmeasured aspects affecting growth: soil quality, moisture levels, history of other crops grown, phosphorus, sunlight, elevation, etc.

A bit on fitting the random effects version

The mean parameters are the μ, 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, these estimated mean parameters to estimate mean-zero residuals, then 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.

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.