STAT 8630, Mixed-Effect Models and Longitudinal Data Analysis — Lecture Notes

Introduction to Longitudinal Data

Terminology:

Longitudinal data consist of observations (i.e., measurements) taken repeatedly through time on a sample of experimental units (i.e., individuals, subjects).

- The experimental units or subjects can be human patients, animals, agricultural plots, etc.

- Typically, the terms "longitudinal data" and "longitudinal study" refer to situations in which data are collected through time under uncontrolled circumstances. E.g., subjects with torn ACLs in their knees are assigned to one of two methods of surgical repair and then followed through time (examined at 6, 12, ,18, 24 months for knee stability, say).

- Longitudinal data are to be contrasted with **cross-sectional data**. Cross-sectional data contain measurements on a sample of subjects at only one point in time.

Repeated measures: The terms "repeated measurements" or, more simply, "repeated measures" are sometimes used as rough synonyms for "longitudinal data", however, there are sometimes slight differences in meaning for these terms.

- Repeated measures are also multiple measurements on each of several individuals, but they are not necessarily through time. E.g., measurements of chemical concentration in the leaves of a plant taken at different locations (low, medium and high on the plant, say) can be regarded as repeated measures.

- In addition, repeated measures may occur across the levels of some controlled factor. E.g., **crossover studies** involve repeated measures. In a crossover study, subjects are assigned to multiple treatments (usually 2 or 3) sequentially. E.g., a two period crossover experiment involves subjects who each get treatments A and B, some in the order AB, and others in the order BA.

Another rough synonym for longitudinal data is **panel data.**

- The term panel data is more common in econometrics, the term longitudinal data is most commonly used in biostatistics, and the term repeated measures most often arises in an agricultural context.

In all cases, however, we are referring to multiple measurements of essentially the same variable(s) on a given subject or unit of observation. We'll often use the more generic term **clustered data** to refer to this situation.

Advantages and Disadvantages of Longitudinal Data:

Advantages:

1. Although time effects can be investigated in cross-sectional studies in which different subjects are examined at different time points, only longitudinal data give information on individual patterns of change.

2. Again, in contrast to cross-sectional studies involving multiple time points, longitudinal studies economize on subjects.

3. In investigating time effects in a longitudinal design or treatment effects in a crossover design, each subject can "serve as his or her own control". That is, comparisons can be made within a subject rather than between subjects. This eliminates between-subjects sources of variability from the experimental error and makes inferences more efficient/powerful (think paired t-test versus two-sample t-test).

4. Since the same variables are measured repeatedly on the same subjects, the reliability of those measurements can be assessed, and purely from a measurement standpoint, reliability is higher.

Disadvantages:

1. For longitudinal or, more generally, clustered data it is typically reasonable to assume independence across clusters, but repeated measures within a cluster are almost always correlated, which complicates the analysis.

2. Clustered data are often unbalanced or partially incomplete (involve missing data), which also complicates the analysis. For longitudinal data, this may be due to loss to follow-up (some subjects move away, die, miss appointments, etc.). For other types of clustered data, the cluster size may vary (e.g., familial data, where family size varies).

3. As a practical matter, methods and/or software may not exist or may be complex, so obtaining results and interpretting them may be difficult.

Data Structure for Clustered Data:

The general data structure of clustered data is given in the table below. Here, y_{ij} represents the j^{th} observation from the i^{th} cluster, where $i = 1, \ldots, n$, and $j = 1, \ldots, t_i$.*

TABLE 1.1. General layout for repeated measurements

Subject	Time Point	Missing Indicator	Response	Covariates		
1	1	δ_{11}	y_{11}	x_{111}	$\cdots$	x_{11p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{1j}	y_{1j}	x_{1j1}	$\cdots$	x_{1jp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_1	δ_{1t_1}	y_{1t_1}	x_{1t_11}	$\cdots$	x_{1t_1p}
i	1	δ_{i1}	y_{i1}	x_{i11}	$\cdots$	x_{i1p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{ij}	y_{ij}	x_{ij1}	$\cdots$	x_{ijp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_i	δ_{it_i}	y_{it_i}	x_{it_i1}	$\cdots$	x_{it_ip}
n	1	δ_{n1}	y_{n1}	x_{n11}	$\cdots$	x_{n1p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	δ_{nj}	y_{nj}	x_{nj1}	$\cdots$	x_{njp}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_n	δ_{nt_n}	y_{nt_n}	x_{nt_n1}	$\cdots$	x_{nt_np}

- Associated with each observation y_{ij} we may have a $p \times 1$ vector of explanatory variables, or covariates, $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$.

- In addition, to indicate missing data, we may sometimes define a missing value indicator:

$$\delta_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ and } \mathbf{x}_{ij} \text{ are observed,} \\ 0, & \text{otherwise.} \end{cases}$$

- We will often write the set of responses from the i^{th} subject as a vector: $\mathbf{y}_i = (y_{i1}, \ldots, y_{it_i})^T$. In addition, let $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ be the combined response vector from all subjects and time points and let $N = \sum_{i=1}^{n} t_i$ be the total sample size.

* Note that our text uses slightly different notation in which n_i is the cluster size and N is the number of clusters.

Often subjects will be grouped somehow into treatment groups. In this case, we will need additional subscripts to index the groups. For example, the data layout below in Table 1.2 represents a one-way layout with s groups with repeated measures over t time points. Here, y_{hij} represents the j^{th} measurement on the i^{th} subject from the h^{th} group.

TABLE 1.2. Layout for the special case of multiple samples

		Time Point				
Group	Subject	1	...	j	...	t
1	1	y_{111}	$\cdots$	y_{11j}	$\cdots$	y_{11t}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	i	y_{1i1}	$\cdots$	y_{1ij}	$\cdots$	y_{1it}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	n_1	y_{1n_11}	$\cdots$	y_{1n_1j}	$\cdots$	y_{1n_1t}
.....						
h	1	y_{h11}	$\cdots$	y_{h1j}	$\cdots$	y_{h1t}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	i	y_{hi1}	$\cdots$	y_{hij}	$\cdots$	y_{hit}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	n_h	y_{hn_h1}	$\cdots$	y_{hn_hj}	$\cdots$	y_{hn_ht}
.....						
s	1	y_{s11}	$\cdots$	y_{s1j}	$\cdots$	y_{s1t}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	i	y_{si1}	$\cdots$	y_{sij}	$\cdots$	y_{sit}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	n_s	y_{sn_s1}	$\cdots$	y_{sn_sj}	$\cdots$	y_{sn_st}

- Note that the s groups might correspond to s levels of a single treatment factor, or s combinations of the levels of two or more factors. In the latter case, it may be convenient to introduce additional subscripts.

  - E.g., for a two-way layout with repeated measures involving factors A and B, we might index the data as y_{hkij} to represent the j^{th} observation on the i^{th} subject in the h, k^{th} treatment (at the h^{th} level of A and the k^{th} level of B).

For the single group case we can drop the index h from table 1.2 to represent the data as follows:

TABLE 1.3. Layout for the one-sample case

	Time Point				
Subject	1	...	j	...	t
1	y_{11}	...	y_{1j}	...	y_{1t}
⋮	⋮	⋱	⋮	⋱	⋮
i	y_{i1}	...	y_{ij}	...	y_{it}
⋮	⋮	⋱	⋮	⋱	⋮
n	y_{n1}	...	y_{nj}	...	y_{nt}

For longitudinal data, the response variable can be laid out in a single column as in Table 1.1, or with one column per time point as in table 1.2 amd 1.3. The two layouts suggest that such data can be conceptualized as univariate or multivariate data.

In fact, there are classical normal-theory approches to analyzing continuous repeated measures data of each type:

— univariate methods (most notably the repeated measures analysis of variance); and

— multivariate methods (profile and growth curve analysis).

Repeated Measures ANOVA

The classical repeated measures analysis of variance (RM-ANOVA) situation is a one-way layout with repeated measures over t time points. This situation is displayed in Table 1.2.

Here, there are are $n_1, n_2, \ldots, n_s$ subjects, respectively, in s treatment groups. If there measurements are taken at just a single time point, we have a (unbalanced) one-way layout. However, subjects are followed up over t time points to yield a repeated measures design.

Example – Methemoglobin in Sheep:

An experiment was designed to study trends in methemoglobin (M) in sheep following treatment with 3 equally spaced levels of NO_2 (factor A). Four sheep were assigned to each level and each animal was measured at 6 sampling times (factor B), 5 of them following treatment. The response was $\log(M + 5)$.

The data from this experiment are as follows:

		Sampling Time					
NO_2	Sheep	1	2	3	4	5	6
1	1	2.197	2.442	2.542	2.241	1.960	1.988
1	2	1.932	2.526	2.526	2.152	1.917	1.917
1	3	1.946	2.251	2.501	1.988	1.686	1.841
1	4	1.758	2.054	2.588	2.197	2.140	1.686
2	5	2.230	3.086	3.357	3.219	2.827	2.534
2	6	2.398	2.580	2.929	2.874	2.282	2.303
2	7	2.054	3.243	3.653	3.811	3.816	3.227
2	8	2.510	2.695	2.996	3.246	2.565	2.230
3	9	2.140	3.896	4.246	4.461	4.418	4.331
3	10	2.303	3.822	4.109	4.240	4.127	4.084
3	11	2.175	2.907	3.086	2.827	2.493	2.230
3	12	2.041	3.824	4.111	4.301	4.206	4.182

- Here $s = 3$, $n_1 = n_2 = n_3 = 4$, and $t = 6$.

The basic RM-ANOVA approach is based upon a similarity between a repeated measures design and a **split-plot experimental design**. The RM-ANOVA approach uses the split-plot model with modifications to the split-plot analysis, if necessary, to account for differences between the two designs.

A split-plot experimental design is one in which (at least) two sizes of experimental unit are used. The larger experimental unit is known as the **whole plot**, and is randomized to some experimental design (a one-way layout, say).

The whole plot is then subdivided into smaller units known as **split plots,** which are then assigned to a second experimental design within each whole plot.

Example – Chocolate Cake:

An experiment was conducted to determine the effect of baking temperature on quality for three recipes of chocolate cake. Recipes I and II differed in that the chocolate was added at 40° C. and 60° C., respectively, while recipe III contained extra sugar. Six different baking temperatures were considered: 175° C., 185° C., 195° C., 205° C., 215° C., and 225° C. 45 batches of cake batter were prepared using each of the 3 recipes 15 times in a completely random order. Each batch was large enough for 6 cakes, and the six baking temperatures were randomly assigned to the 6 cakes per batch in a random manner. One of several measurements of quality made on each cake was the breaking angle of the cake.

The data from this experiment are as follows:

TABLE 7.5 Breaking angles (degrees)

		Temperature							
	Rep.	175°	185°	195°	205°	215°	225°	Unit totals	Rep. totals
Recipe I	1	42	46	47	39	53	42	269	843
	2	47	29	35	47	57	45	260	820
	3	32	32	37	43	45	45	234	665
	4	26	32	35	24	39	26	182	599
	5	28	30	31	37	41	47	214	583
	6	24	22	22	29	35	26	158	517
	7	26	23	25	27	33	35	169	492
	8	24	33	23	32	31	34	177	493
	9	24	27	28	33	34	23	169	492
	10	24	33	27	31	30	33	178	519
	11	33	39	33	28	33	30	196	538
	12	28	31	27	39	35	43	203	557
	13	29	28	31	29	37	33	187	574
	14	24	40	29	40	40	31	204	502
	15	26	28	32	25	37	33	181	479
Totals		437	473	462	503	580	526	2981	8673
Recipe II	1	39	46	51	49	55	42	282	
	2	35	46	47	39	52	61	280	
	3	34	30	42	35	42	35	218	
	4	25	26	28	46	37	37	199	
	5	31	30	29	35	40	36	201	
	6	24	29	29	29	24	35	170	
	7	22	25	26	26	29	36	164	
	8	26	23	24	31	27	37	168	
	9	27	26	32	28	32	33	178	
	10	21	24	24	27	37	30	163	
	11	20	27	33	31	28	33	172	
	12	23	28	31	34	31	29	176	
	13	32	35	30	27	35	30	189	
	14	23	25	22	19	21	35	145	
	15	21	21	28	26	27	20	143	
Totals		403	441	476	482	517	529	2848	
Recipe III	1	46	44	45	46	48	63	292	
	2	43	43	43	46	47	58	280	
	3	33	24	40	37	41	38	213	
	4	38	41	38	30	36	35	218	
	5	21	25	31	35	33	23	168	
	6	24	33	30	30	37	35	189	
	7	20	21	31	24	30	33	159	
	8	24	23	21	24	21	35	148	
	9	24	18	21	26	28	28	145	
	10	26	28	27	27	35	35	178	
	11	28	25	26	25	38	28	170	
	12	24	30	28	35	33	28	178	
	13	28	29	43	28	33	37	198	
	14	19	22	27	25	25	35	153	
	15	21	28	25	25	31	25	155	
Totals		419	434	476	463	516	536	2844	
Temp. totals		1259	1348	1414	1448	1613	1591	8673	

- Here, there are 45 batches of cake, which occur in a balanced one-way layout. The batches, are the whole plots and recipe, with 3 levels, is the whole plot factor.

- Each batch is then split into 6 cakes, which are randomly assigned to one of 6 temperatures. The cakes are the split plots, and temperature is the split plot factor.

Let y_{hij} represent the response at the j^{th} level of the split-plot factor for the i^{th} whole plot in the h^{th} group. The model traditionally used for the split-plot design exemplified by the chocolate cake example is

$$y_{hij} = \mu + \alpha_h + e_{i(h)} + \beta_j + (\alpha\beta)_{hj} + \varepsilon_{hij}, \qquad (*)$$

Here, μ is a grand mean, α_h is an effect for the h^{th} level of the whole plot factor (e.g., recipe), β_j is an effect for the j^{th} level of the split-plot factor (temperature), and $(\alpha\beta)_{hj}$ is an interaction term for the whole and split plot factors. $e_{i(h)}$ is a random effect for the i^{th} whole plot nested in the h^{th} level of the whole plot factor, and ε_{hij} is the overall error term.

One way to think about the split plot model is that it is the union of the model appropriate for the whole plots and the model appropriate for the split plots.

- The whole plots occur in a one-way layout, so the one-way layout model
$$\mu + \alpha_h + e_{i(h)}$$
is appropriate for batches.

- The split plots occur in a randomized complete block design, so the RCBD model
$$\mu + \underbrace{\text{block}_{hi}}_{=\alpha_h + e_{i(h)}} + \beta_j + \varepsilon_{hij}$$
is appropriate for cakes.

- Putting these portions of the models together and adding an interaction term $(\alpha\beta)_{hj}$ to capture interactions between the whole and split plot factors, leads to model (*).

- The random whole plot effect $e_{i(h)}$ can be thought of as the whole plot error term and ε_{hij} as the split plot error term. Since there are two experimental units with two separate randomizations, there are two error terms in the model.

In fact, the split plot model described above is an example of a linear mixed-effects model (LMM).

- It includes fixed effects for the whole plot factor (the α_h's for recipes), split plot factor (the β_j's for temperatures), and their interaction (the $(\alpha\beta)_{hj}$'s). These are the regression parameters of the model.

- It also includes random effects: $e_{i(h)}$, a whole plot (or batch) effect, in addition to the overall error term ε_{hij}, which is always present in any linear model, so is typically not categorized as a random effect, even though it is.

- The term "mixed-effects model" or sometimes simply "mixed model" refers to the fact that the *linear predictor* of the model (the right side of the model equation, excluding the overall error term) includes both fixed and random effects.

Fixed vs. random effects: The effects in the model account for variability in the response across levels of treatment and design factors. The decision as to whether fixed effects or random effects should be used depends upon what the appropriate scope of generalization is.

- If it is appropriate to think of the levels of a factor as randomly drawn from, or otherwise representative of, a population to which we'd like to generalize, then random effects are suitable.

  - Design or grouping factors are usually more appropriately modeled with random effects.

  - E.g., blocks (sections of land) in an agricultural experiment, days when an experiment is conducted over several days, lab technician when measurements are taken by several technicians, subjects in a repeated measures design, locations or sites along a river when we desire to generalize to the entire river, etc.

- If, however, the specific levels of the factor are of interest in and of themselves then fixed effects are more appropriate.

  - Treatment factors are usually more appropriately modeled with fixed effects.

  - E.g., In experiments to compare drugs, amounts of fertilizer, hybrids of corn, teaching techniques, and measurement devices, all of these factors are most appropriately modeled with fixed effects.

- A good litmus test for whether the level of some factor should be treated as fixed is to ask whether it would be of broad interest to report a mean for that level. For example, if I'm conducting an experiment in which each of four different classes of third grade students are taught with each of three methods of instruction (e.g., in a crossover design) then it will be of broad interest to report the mean response (level of learning, say) for a particular method of instruction, but not for a particular classroom of third grades.

  - Here, fixed effects are appropriate for instruction method, random effects for class.

Since the whole plot error term represents random whole plot effects (batch effects), the $e_{i(h)}$'s are random effects. Therefore, we must make some assumptions about their distribution (and the distribution of the overall error term ε_{hij}) to complete the split-plot model. The following assumptions are typical:

$$y_{hij} = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj} + e_{i(h)} + \varepsilon_{hij} = \mu_{hj} + e_{i(h)} + \varepsilon_{hij},$$

where

$$\{e_{i(h)}\} \overset{iid}{\sim} N(0, \sigma_e^2)$$
$$\{\varepsilon_{hij}\} \overset{iid}{\sim} N(0, \sigma^2)$$
$$\text{cov}(\varepsilon_{hij}, e_{i'(h')}) = 0, \quad \text{for all } h, i, j, h', i'.$$

The chocolate cake experiment is an example of a balanced split plot design with whole plots arranged in a one-way layout. More complex split-plot designs are possible. E.g., whole plots are often arranged in a RCBD, split plots could be split once again to create split-split-plots, etc.

However, the classical analysis of all of these designs is relatively straightforward *provided that the design is balanced.*

- By "balanced" here, we mean that there is an equal number of replicates for each whole plot treatment (recipe), and the same set of subplot treatments (temperatures) was observed within each whole plot (batch).

The classical analysis of the balanced split plot model with whole plots in a one-way layout is based on the model at the bottom of p.12 and the following decomposition of the deviations $y_{hij} - \bar{y}_{...}$ of each observation from the grand sample mean:

$$\begin{aligned} y_{hij} - \bar{y}_{...} = (\bar{y}_{h\cdot\cdot} - \bar{y}_{...}) + (\bar{y}_{hi\cdot} - \bar{y}_{h\cdot\cdot}) + (\bar{y}_{\cdot\cdot j} - \bar{y}_{...}) \\ + (\bar{y}_{h\cdot j} - \bar{y}_{h\cdot\cdot} - \bar{y}_{\cdot\cdot j} + \bar{y}_{...}) + (y_{hij} - \bar{y}_{h\cdot j} - \bar{y}_{hi\cdot} + \bar{y}_{h\cdot\cdot}), \quad (*) \end{aligned}$$

where $n = \sum_h n_h$, we assume n_h is constant over h, and

$$\bar{y}_{...} = (nt)^{-1} \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} y_{hij} \qquad \bar{y}_{h\cdot\cdot} = (n_h t)^{-1} \sum_{i=1}^{n_h} \sum_{j=1}^{t} y_{hij}$$

$$\bar{y}_{\cdot\cdot j} = n^{-1} \sum_{h=1}^{s} \sum_{i=1}^{n_h} y_{hij} \qquad \bar{y}_{h\cdot j} = n_h^{-1} \sum_{i=1}^{n_h} y_{hij} \qquad \bar{y}_{hi\cdot} = t^{-1} \sum_{j=1}^{t} y_{hij}$$

are sample means over all observations ($\bar{y}_{...}$), over observations in the h^{th} whole plot treatment group ($\bar{y}_{h\cdot\cdot}$), etc.

This decomposition leads to the following analysis of variance:

Source of Variation	Sum of Squares	d.f.	$E(MS)$
Whole plot groups	SS_{WPG}	$s-1$	$\sigma^2 + t\sigma_e^2 + Q(\alpha, \alpha\beta)$
Whole plot error	SS_{WPE}	$n-s$	$\sigma^2 + t\sigma_e^2$
Split plot groups	SS_{SPG}	$t-1$	$\sigma^2 + Q(\beta, \alpha\beta)$
Interaction	$SS_{WPG\times SPG}$	$(s-1)(t-1)$	$\sigma^2 + Q(\alpha\beta)$
Split plot error	SS_{SPE}	$(n-s)(t-1)$	σ^2
Total	SS_T	$nt-1$	

The sums of squares in the ANOVA table are simply the sums, over all observations, of the terms in decomposition (*). That is,

$$SS_{WPG} = \sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t}(\bar{y}_{h\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SS_{WPE} = \sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t}(\bar{y}_{hi\cdot} - \bar{y}_{h\cdot\cdot})^2$$

$$SS_{SPG} = \sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t}(\bar{y}_{\cdot\cdot j} - \bar{y}_{\cdot\cdot\cdot})^2$$

$$SS_{WPG\times SPG} = \sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t}(\bar{y}_{h\cdot j} - \bar{y}_{h\cdot\cdot} - \bar{y}_{\cdot\cdot j} + \bar{y}_{\cdot\cdot\cdot})^2$$

$$SS_{SPE} = \sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t}(y_{hij} - \bar{y}_{h\cdot j} - \bar{y}_{hi\cdot} + \bar{y}_{h\cdot\cdot})^2.$$

In addition, the quantities $Q(\alpha, \alpha\beta), Q(\beta, \alpha\beta)$, and $Q(\alpha\beta)$ are quadratic forms representing differences across whole plot groups, split plot groups, and the whole plot group $\times$ split plot group interaction, respectively.

That is, $Q(\alpha, \alpha\beta)$ is a sum of squares in the α_h's and $(\alpha\beta)_{hj}$'s that equals zero under the null hypothesis of no differences across whole plot groups (hypothesis (1) below). Similarly, $Q(\beta, \alpha\beta)$ and $Q(\alpha\beta)$ are sums of squares that are zero under no differences across split plot groups (hypothesis (2)), and under no interaction (hypothesis (3)), respectively.

- The $Q(\cdot)$ notation is often used because it is more convenient than writing the term out exactly. The exact form of these terms are not important, it only matters that these terms are zero under the null hypothesis of no effect, and positive under the alternative.

F tests appropriate for the hypotheses of interest can be determined by examination of the expected mean squares.

Let $\mu_{hj} = \mathrm{E}(y_{hij}) = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj}$ and let

$$\bar{\mu}_{h\cdot} = t^{-1} \sum_{j=1}^{t} \mu_{hj} = \mu + \alpha_h + \bar{\beta}_{\cdot} + (\bar{\alpha\beta})_{h\cdot}$$

be the marginal mean for whole plot group h and let

$$\bar{\mu}_{\cdot j} = s^{-1} \sum_{h=1}^{s} \mu_{hj} = \mu + \bar{\alpha}_{\cdot} + \beta_j + (\bar{\alpha\beta})_{\cdot j}$$

be the marginal mean for the j^{th} split plot group. Then the hypotheses of interest and their corresponding test statistics are as follows:

1. $H_1 : \bar{\mu}_{1\cdot} = \cdots = \bar{\mu}_{s\cdot}$ (no main effect of the whole plot factor), which is tested with

$$F = \frac{MS_{WPG}}{MS_{WPE}} \sim F(s-1, n-s);$$

2. $H_2 : \bar{\mu}_{\cdot 1} = \cdots = \bar{\mu}_{\cdot t}$ (no main effect of the split plot factor), which is tested with

$$F = \frac{MS_{SPG}}{MS_{SPE}} \sim F(t-1, (n-s)(t-1));$$

3. and $H_3 : (\alpha\beta)_{hj} = 0$ for all h, j (no interaction between whole plot and split plot factors), which is tested with

$$F = \frac{MS_{WPG \times SPG}}{MS_{SPE}} \sim F((s-1)(t-1), (n-s)(t-1)).$$

Note that side conditions are often placed on the split plot model to avoid the complications introduced by having an overparameterized model and non-full-rank design matrix. The usual sum-to-zero side conditions are

$$\sum_h \alpha_h = \sum_j \beta_j = \sum_h (\alpha\beta)_{hj} = \sum_j (\alpha\beta)_{hj} = 0.$$

Such conditions are not strictly necessary to derive the F tests given above, but they do simplify things somewhat.

Without the side conditions, H_1 can be expressed in terms of the fixed effects in the model as

$$H_1 : \alpha_1 + (\bar{\alpha\beta})_{1\cdot} = \cdots = \alpha_s + (\bar{\alpha\beta})_{s\cdot}$$

which reduces to

$$H_1 : \alpha_1 = \cdots = \alpha_s = 0$$

under the sum-to-zero constraints. Note that under these constraints the $Q(\alpha, \alpha\beta)$ term in $\mathrm{E}(MS_{WPG})$ reduces to $Q(\alpha) = \sum_{h=1}^{s} \alpha_h^2/(s-1)$, which, of course, is 0 under H_1 and > 0 otherwise.

- Similar comments apply to H_2, which under the sum-to-zero constraints is equivalent to $H_2 : \beta_1 = \cdots = \beta_t = 0$ and $Q(\boldsymbol{\beta}, \alpha\beta) = Q(\beta) = \sum_{j=1}^{t} \beta_j^2/(t-1)$.

Example — Chocolate Cake:

- See the handout labelled choccake.sas.

- In choccake.sas we use PROC MIXED to perform the analysis just described. A call to PROC GLM is also included which reproduces the basic PROC MIXED results (e.g., the ANOVA table and expected mean squares). However, PROC GLM is not designed for mixed models, and cannot, in general, be trusted to produce correct results for split plot models and other LMMs.

- Note that method=type3 requests the classical ANOVA-type analysis. This is not the default, which is a REML analysis.

- The basic results are that there are not significant interactions between recipe and baking temperature, there are not significant main effects of recipe, but there are significant main effects of temperature.

  - From the contrasts and profile plot, we see that mean breaking angle increases linearly with baking temperature.

- Note that the expected mean squares are printed on the bottom of p.1 and the top of p.2. These results agree with the ANOVA table given previously in these notes.

- Method of moment estimators of the variance components σ^2 and σ_e^2 are easily derived from the expressions for $\mathrm{E}(MS_{SPE})$ and $\mathrm{E}(MS_{WPE})$. Equating MS_{SPE} with its expectation σ^2 yields the estimator

$$\hat{\sigma}^2 = MS_{SPE} = 20.4709.$$

  Similarly, equating MS_{WPE} with its expectation $\sigma^2 + t\sigma_e^2$ yields

$$\hat{\sigma}_e^2 = \frac{MS_{WPE} - MS_{SPE}}{t} = 41.8370.$$

Estimation and Inference on Means in the Split-plot Model:

According to the model on p.12,

$$\begin{aligned}\text{var}(y_{hij}) &= \text{var}(\mu_{hj} + e_{i(h)} + \varepsilon_{hij}) = \text{var}(e_{i(h)} + \varepsilon_{hij}) \\ &= \text{var}(e_{i(h)}) + \text{var}(\varepsilon_{hij}) + 2\underbrace{\text{cov}(e_{i(h)}, \varepsilon_{hij})}_{=0} = \sigma_e^2 + \sigma^2.\end{aligned}$$

- Because the variance of y_{hij} is the sum of two components, σ_e^2 and σ^2, these quantities are often called **variance components.**

In addition,

$$\begin{aligned}&\text{cov}(y_{hij}, y_{hik}) \\ &\quad= \text{cov}(\mu_{hj} + e_{i(h)} + \varepsilon_{hij}, \mu_{hk} + e_{i(h)} + \varepsilon_{hik}) = \text{cov}(e_{i(h)} + \varepsilon_{hij}, e_{i(h)} + \varepsilon_{hik}) \\ &\quad= \text{cov}(e_{i(h)}, e_{i(h)}) + \underbrace{\text{cov}(e_{i(h)}, \varepsilon_{hik})}_{=0} + \underbrace{\text{cov}(\varepsilon_{hij}, e_{i(h)})}_{=0} + \underbrace{\text{cov}(\varepsilon_{hij}, \varepsilon_{hik})}_{=0} \\ &\quad= \text{var}(e_{i(h)}) = \sigma_e^2\end{aligned}$$

for $j \neq k$, and

$$\text{cov}(y_{hij}, y_{h'i'j'}) = 0,$$

for $h \neq h'$ or $i \neq i'$ (i.e., the covariance is 0 between subjects).

From these results we see that the correlation is zero between observations on different subjects, but

$$\text{corr}(y_{hij}, y_{hik}) = \frac{\sigma_e^2}{\sigma^2 + \sigma_e^2} \equiv \rho, \quad j \neq k.$$

- The within-subject correlation ρ here is called the **intra-class correlation**.

The within subject covariance structure we have just described can be represented succinctly as

$$\begin{aligned}\text{var}(\mathbf{y}_{hi}) &= \begin{pmatrix} \sigma^2+\sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \cdots & \sigma_e^2 \\ \sigma_e^2 & \sigma^2+\sigma_e^2 & \sigma_e^2 & \cdots & \sigma_e^2 \\ \sigma_e^2 & \sigma_e^2 & \sigma^2+\sigma_e^2 & \cdots & \sigma_e^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \cdots & \sigma^2+\sigma_e^2 \end{pmatrix} \\ &= (\sigma^2+\sigma_e^2)[(1-\rho)\mathbf{I}_t + \rho\mathbf{J}_{tt}],\end{aligned}$$

where $\mathbf{I}_t$ is the $t \times t$ identity matrix and $\mathbf{J}_{tt}$ is a $t \times t$ matrix of ones.

- This variance covariance structure is called **compound symmetry**.

Recall that for a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

where $\mathbf{V}$ is a known positive definite matrix, the best linear unbiased estimator (BLUE) of a vector of estimable functions $\mathbf{C}\boldsymbol{\beta}$ is given by $\mathbf{C}\hat{\boldsymbol{\beta}}_{GLS}$ where $\hat{\boldsymbol{\beta}}_{GLS}$ is a generalized least squares (GLS) estimator of the form

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

- Under compound symmetry it is not hard to show that the GLS estimator is equivalent to the ordinary least squares (OLS) estimator $\mathbf{C}\hat{\boldsymbol{\beta}}_{OLS}$ where

  $$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$$

  (see Rencher, Example 7.8.1; or Graybill, Corollary 6.8.1.2).

For balanced split-plot experiments, it is easy to show that the marginal mean $\bar{\mu}_{h\cdot}$, $\bar{\mu}_{\cdot j}$ and the joint means μ_{hj} are all estimable, and have BLUEs

$$\hat{\bar{\mu}}_{h\cdot} = \bar{y}_{h\cdot\cdot}, \quad \hat{\bar{\mu}}_{\cdot j} = \bar{y}_{\cdot\cdot j}, \quad \hat{\mu}_{hj} = \bar{y}_{h\cdot j}.$$

Standard errors for these quantities are defined as the estimated standard deviations. Therefore, we need the variances of these estimators.

$$\begin{aligned}
\text{var}(\bar{y}_{h\cdot\cdot}) &= \text{var}\left(\frac{1}{n_h t}\sum_i \mathbf{j}_t^T \mathbf{y}_{hi}\right) = \frac{1}{(n_h t)^2}\sum_i \mathbf{j}_t^T \text{var}(\mathbf{y}_{hi})\mathbf{j}_t \\
&= \frac{1}{(n_h t)^2}\sum_i [t(\sigma^2 + \sigma_e^2) + (t^2 - t)\sigma_e^2] = \frac{1}{(n_h t)^2}\sum_i [t\sigma^2 + t^2\sigma_e^2] \\
&= \frac{n_h t}{(n_h t)^2}(\sigma^2 + t\sigma_e^2) = \frac{1}{n_h t}(\sigma^2 + t\sigma_e^2)
\end{aligned}$$

In addition,

$$\begin{aligned}
\text{var}(\bar{y}_{\cdot\cdot j}) &= \text{var}\left(\frac{1}{s n_h}\sum_h\sum_i y_{hij}\right) = \frac{1}{(s n_h)^2}\sum_h\sum_i \text{var}(y_{hij}) \\
&= \frac{1}{(s n_h)^2}\sum_h\sum_i (\sigma^2 + \sigma_e^2) = \frac{s n_h}{(s n_h)^2}(\sigma^2 + \sigma_e^2) = \frac{\sigma^2 + \sigma_e^2}{s n_h}
\end{aligned}$$

and, similarly,

$$\text{var}(\bar{y}_{h\cdot j}) = \text{var}\left(\frac{1}{n_h}\sum_i y_{hij}\right) = \frac{n_h}{n_h^2}\text{var}(y_{hij}) = \frac{1}{n_h}(\sigma^2 + \sigma_e^2).$$

In the case of $\bar{y}_{h\cdot\cdot}$, its variance is easy to estimate because $\mathrm{E}(MS_{WPE}) = \sigma^2 + t\sigma_e^2$. So,

$$\text{s.e.}(\bar{y}_{h\cdot\cdot}) = \sqrt{\hat{\text{var}}(\bar{y}_{h\cdot\cdot})} = \sqrt{\frac{MS_{WPE}}{n_h t}}.$$

However, $\text{var}(\bar{y}_{\cdot\cdot j})$ and $\text{var}(\bar{y}_{h\cdot j})$ both involve $\sigma^2 + \sigma_e^2$, which is not the expected value of any mean square in the ANOVA.

We do have estimates of σ^2 and σ_e^2 individually, though; namely, MS_{SPE}, and $(MS_{WPE} - MS_{SPE})/t$ (bottom of p.17). So,

$$\text{s.e.}(\bar{y}_{\cdot\cdot j}) = \sqrt{\frac{\widehat{\sigma^2 + \sigma_e^2}}{sn_h}}, \quad \text{and} \quad \text{s.e.}(\bar{y}_{h\cdot j}) = \sqrt{\frac{\widehat{\sigma^2 + \sigma_e^2}}{n_h}}$$

where

$$\begin{aligned}\widehat{\sigma^2 + \sigma_e^2} &= MS_{SPE} + (MS_{WPE} - MS_{SPE})/t \\ &= \frac{t-1}{t} MS_{SPE} + \frac{1}{t} MS_{WPE}.\end{aligned}$$

CIs and contrasts for marginal means for whole plot factor:

Confidence intervals and hypothesis tests on $\bar{\mu}_{h\cdot\cdot}$ are based on the pivotal quantity

$$t = \frac{\bar{y}_{h\cdot\cdot} - \bar{\mu}_{h\cdot\cdot}}{\text{s.e.}(\bar{y}_{h\cdot\cdot})} = \frac{\bar{y}_{h\cdot\cdot} - \bar{\mu}_{h\cdot\cdot}}{\sqrt{MS_{WPE}/(n_h t)}} \sim t(\text{d.f.}_{WPE}) = t(n-s).$$

This leads to a $100(1-\alpha)\%$ CI for $\bar{\mu}_{h\cdot\cdot}$ given by

$$\bar{y}_{h\cdot\cdot} \pm t_{1-\alpha/2}(\text{d.f.}_{WPE})\text{s.e.}(\bar{y}_{h\cdot\cdot}) \quad = \quad \bar{y}_{h\cdot\cdot} \pm t_{1-\alpha/2}(n-s)\sqrt{\frac{MS_{WPE}}{n_h t}}$$

For a contrast $\psi = \sum_h c_h \bar{\mu}_{h\cdot\cdot}$ with sample estimator $C = \sum_h c_h \bar{y}_{h\cdot\cdot}$, we test $H_0 : \psi = 0$ via t or F tests.

The appropriate t test statistic is

$$t = \frac{|C|}{\text{s.e.}(C)}, \quad \text{which we compare to} \quad t_{1-\alpha/2}(\text{d.f.}_{WPE})$$

for an α-level test. Equivalently, we can use an F test:

$$F = t^2, \quad \text{which we compare to} \quad F_{1-\alpha}(1, \text{d.f.}_{WPE}).$$

The standard error for a contrast in the whole plot groups is given by

$$\text{s.e.}(C) = \sqrt{\hat{\text{var}}\left(\sum_h c_h \bar{y}_{h\cdot\cdot}\right)} = \sqrt{\sum_h c_h^2 \hat{\text{var}}(\bar{y}_{h\cdot\cdot})} = \sqrt{\frac{MS_{WPE}}{n_h t}\sum_h c_h^2}.$$

A $100(1-\alpha)\%$ CI for ψ is given by

$$C \pm t_{1-\alpha/2}(\text{d.f.}_{WPE})\text{s.e.}(C).$$

CIs and contrasts for marginal means for split plot factor:

Confidence intervals and hypothesis tests on $\bar{\mu}_{\cdot\cdot j}$ are based on the pivotal quantity

$$t = \frac{\bar{y}_{\cdot\cdot j} - \bar{\mu}_{\cdot\cdot j}}{\text{s.e.}(\bar{y}_{\cdot\cdot j})} = \frac{\bar{y}_{\cdot\cdot j} - \bar{\mu}_{\cdot\cdot j}}{\sqrt{[(t-1)MS_{SPE} + MS_{WPE}]/(sn_h t)}}.$$

- However, now this quantity is not distributed exactly as a student's t!

  Why?

Because the denominator doesn't involve a single mean square (χ^2 divided by its d.f.), but instead involves a linear combination of mean squares.

What's the distribution of a linear combination of independent mean squares in normally distributed random variables?

An approximate answer is given by **Satterthwaite's formula.** Satterthwaite showed that a linear combination of independent mean squares of the form $MS = a_1 MS_1 + \cdots + a_k MS_k$ is *approximately* χ^2 with approximate degrees of freedom given by

$$\text{d.f.} = \frac{(MS)^2}{\frac{(a_1 MS_1)^2}{\text{d.f.}_1} + \cdots + \frac{(a_k MS_k)^2}{\text{d.f.}_k}},$$

where here d.f._i is the d.f. associated with MS_i and the a_i's are constants.

In our case, MS is $(t-1)MS_{SPE} + MS_{WPE}$ so we have

$$\nu = \frac{[(t-1)MS_{SPE} + MS_{WPE}]^2}{\frac{[(t-1)MS_{SPE}]^2}{\text{d.f.}_{SPE}} + \frac{(MS_{WPE})^2}{\text{d.f.}_{WPE}}}.$$

Thus the pivotal quantity has an approximate t distribution:

$$t = \frac{\bar{y}_{\cdot\cdot j} - \bar{\mu}_{\cdot\cdot j}}{\text{s.e.}(\bar{y}_{\cdot\cdot j})} = \frac{\bar{y}_{\cdot\cdot j} - \bar{\mu}_{\cdot\cdot j}}{\sqrt{[(t-1)MS_{SPE} + MS_{WPE}]/(sn_h t)}} \dot{\sim} t(\nu).$$

This leads to an approximate $100(1-\alpha)\%$ CI for $\bar{\mu}_{\cdot\cdot j}$ given by

$$\bar{y}_{\cdot\cdot j} \pm t_{1-\alpha/2}(\nu)\text{s.e.}(\bar{y}_{\cdot\cdot j}).$$

For a contrast $\psi = \sum_j c_j \bar{\mu}_{\cdot\cdot j}$, we have the sample estimator $C = \sum_j c_j \bar{y}_{\cdot\cdot j}$. Despite the fact that s.e.$(\bar{y}_{\cdot\cdot j})$ involves two mean squares, it turns out that s.e.(C) involves only one, so no Satterthwaite approximation is necessary for a contrast in the $\bar{\mu}_{\cdot\cdot j}$'s. To see this, note

$$\begin{aligned} C &= \sum_j c_j \bar{y}_{\cdot\cdot j} = \sum_j c_j \frac{1}{sn_h} \sum_h \sum_i y_{hij} = \sum_j c_j \frac{1}{sn_h} \sum_h \sum_i (\mu_{hj} + e_{i(h)} + \varepsilon_{hij}) \\ &= \sum_j c_j (\bar{\mu}_{\cdot j} + \bar{e}_{\cdot(\cdot)} + \bar{\varepsilon}_{\cdot\cdot j}) = \sum_j c_j \bar{\mu}_{\cdot j} + \bar{e}_{\cdot(\cdot)} \underbrace{\sum_j c_j}_{=0} + \sum_j c_j \bar{\varepsilon}_{\cdot\cdot j} \end{aligned}$$

So,

$$\text{var}(C) = \text{var}(\sum_j c_j \bar{\varepsilon}_{\cdot\cdot j}) = \sum_j c_j^2 \text{var}(\bar{\varepsilon}_{\cdot\cdot j}) = \sum_j c_j^2 \frac{\sigma^2}{sn_h} = \frac{\sigma^2}{sn_h} \sum_j c_j^2.$$

Therefore, for this kind of contrast,

$$\text{s.e.}(C) = \sqrt{\frac{MS_{SPE}}{sn_h} \sum_j c_j^2}.$$

Based on this result, we can test $H_0 : \psi = 0$ with an exact t or F test. Our test statistic is

$$t = \frac{|C|}{\text{s.e.}(C)}, \quad \text{which we compare to} \quad t_{1-\alpha/2}(\text{d.f.}_{SPE})$$

for an α-level test. Equivalently, we can compute:

$$F = t^2, \quad \text{which we compare to} \quad F_{1-\alpha}(1, \text{d.f.}_{SPE}).$$

A $100(1-\alpha)\%$ CI for ψ is given by

$$C \pm t_{1-\alpha/2}(\text{d.f.}_{SPE})\text{s.e.}(C).$$

Joint Means:

As for marginal means of the split plot groups, the variance of the joint mean estimator $\bar{y}_{h\cdot j}$ involves $\sigma^2 + \sigma_e^2$, which we must estimate with $[(t-1)MS_{SPE} + MS_{WPE}]/t$, rather than just a single mean square.

Satterthwaite's formula yields

$$\frac{\bar{y}_{h\cdot j} - \bar{\mu}_{h\cdot j}}{\text{s.e.}(\bar{y}_{h\cdot j})} = \frac{\bar{y}_{h\cdot j} - \bar{\mu}_{h\cdot j}}{[(t-1)MS_{SPE} + MS_{WPE}]/(sn_h t)} \dot{\sim} t(\nu).$$

This leads to an approximate $100(1-\alpha)\%$ CI for the joint mean $\bar{\mu}_{h\cdot j}$ given by

$$\bar{y}_{h\cdot j} \pm t_{1-\alpha/2}(\nu)\text{s.e.}(\bar{y}_{h\cdot j}).$$

Contrasts:

For contrast in the joint means of the form $\psi = \sum_h \sum_j c_{hj}\bar{\mu}_{h\cdot j}$, we estimate the contrast with $C = \sum_h \sum_j c_{hj}\bar{y}_{h\cdot j}$. and we form test statistics in the usual way. That is, t and F statistics for $H_0 : \psi = 0$ are given by

$$t = \frac{|C|}{\text{s.e.}(C)}, \quad \text{and} \quad F = t^2, \quad \text{respectively.}$$

However, the formula for s.e.(C) and the distribution of the test statistics depend on the nature of the contrast in the joint means. For certain contrasts in the joint means, s.e.(C) involves only one mean squares; for others, s.e.(C) involves two mean squares.

The former case is easy and yields an exact distribution, the latter case is harder because we need Satterthwaite's formula again.

Case 1 (the easy case): Contrasts across split plot groups but within a single whole plot group:

In this case, $C = \sum_h \sum_j c_{hj} \bar{y}_{h\cdot j}$ has standard error

$$\text{s.e.}(C) = \sqrt{\frac{MS_{SPE}}{n_h} \sum_h \sum_j c_{hj}^2}$$

Exact tests and confidence intervals for these contrasts are then based on the fact that the t and F test statistics are distributed as

$$t \sim t(\text{d.f.}_{SPE}), \qquad F = t^2 \sim F(1, \text{d.f.}_{SPE}).$$

Case 2 (the harder case): Contrasts involving more than one level of the whole plot factor:

In this case, $C = \sum_h \sum_j c_{hj} \bar{y}_{h\cdot j}$ has standard error

$$\text{s.e.}(C) = \sqrt{\frac{(t-1)MS_{SPE} + MS_{WPE}}{n_h t} \sum_h \sum_j c_{hj}^2}$$

The t and F statistics for contrasts of this type do not have exact t and F distributions. However, approximate tests and confidence intervals for these contrasts can be based on the approximations

$$t \dot{\sim} t(\nu), \qquad F \dot{\sim} (1, \nu),$$

where, again, ν is given by Satterthwaite's formula (bottom of p. 23).

Example — Chocolate Cake (Continued)

- The DDFM=SATTERTH option on the MODEL statement tells SAS to use the Satterthwaite formula to obtain the correct degrees of freedom for F and t tests. These are denominator d.f. (hence DDFM) for F tests and d.f. for t tests.

- By default, PROC MIXED will use the "containment" method for computing these d.f. if a RANDOM statement is included in the call to PROC MIXED. This is a method which attempts to guess the right d.f. from the syntax of the MODEL statement (see the SAS documentation for details). However, the containment method can be wrong. Its advantage is that it takes fewer computational resources than the Satterthwaite method.

- As an example of the Satterthwaite method, consider estimation of $\mu_{11} - \mu_{21}$, the difference between the mean response for cakes made with recipe 1, temperature 1 versus cakes made with recipe 2, temperature 1.

- This is a "case 2" scenario, so we must use Satterthwaite's formula. From the formula on the bottom of p.23,

$$\nu = \frac{[(t-1)MS_{SPE} + MS_{WPE}]^2}{\frac{[(t-1)MS_{SPE}]^2}{\text{d.f.}_{SPE}} + \frac{(MS_{WPE})^2}{\text{d.f.}_{WPE}}} = \frac{[(6-1)20.471 + 271.493]^2}{\frac{[(6-1)20.471]^2}{210} + \frac{(271.493)^2}{42}} = 77.4.$$

The estimate is $\bar{y}_{1\cdot 1} - \bar{y}_{2\cdot 1} = 29.1333 - 26.8667 = 2.2667$ with a standard error of

$$\sqrt{\frac{(t-1)MS_{SPE} + MS_{WPE}}{n_h t} \sum_h \sum_j c_{hj}^2}$$

$$= \sqrt{\frac{(6-1)20.471 + 271.493}{15(6)}[1^2 + (-1)^2]} = 2.8823$$

- So, an approximate 95% CI for $\mu_{11} - \mu_{21}$ is

$$2.2667 \pm t_{.975}(77.4)(2.8823) = (-3.47, 8.01)$$

and an approximate .05-level t test of $H_0 : \mu_{11} = \mu_{21}$ compares

$$t = 2.2667/(2.8823) = 0.79$$

to a $t(77.4)$ distribution. Equivalently, an F test can be used where we compare $F = t^2 = 0.79^2 = 0.62$ to an $F(1, 77.4)$ distribution. Both of these tests give a p-value of .4340, so we fail to reject H_0 at level $\alpha = .05$.

We have presented the simplest split plot design in which whole plots occur in a one-way layout. Other more complicated split plot designs are possible. In practice, the most common split plot design is one in which the whole plots occur in a randomized complete block design.

- For example, suppose the chocolate cake experiment was conducted over 15 days, and the batches were blocked by day. That is, on day one 1 batch from each recipe was prepared and the resulting cakes baked. On day two 1 more batch from each recipe was prepared, etc.

In such a situation, the appropriate model becomes

$$y_{hij} = \mu + \alpha_h + b_i + e_{hi} + \beta_j + (\alpha\beta)_{hj} + \varepsilon_{hij},$$

where y_{hij} is the response on j^{th} split-plot in the h^{th} whole plot group in the i^{th} whole plot block.

Here, e_{hi} and ε_{hij} are the whole plot and split plot error terms, respectively, with similar assumptions as in the previous model. In addition, b_i represents a whole plot block effect. If considered random, which is typically most appropriate, the b_i's are assumed i.i.d. $N(0, \sigma_b^2)$ and independent of the e_{hi}'s and the ε_{hij}'s.

- Rather than treat this model in any detail, we move on to the repeated measures ANOVA (RM-ANOVA). This example will be covered by the general theory of LMMs, which we are working toward.

RM-ANOVA:

The repeated measures ANOVA (RM-ANOVA) is based upon the similarity between repeated measures designs and split-plot designs.

Consider again the Methomoglobin in Sheep example. This is a one-way layout with repeated measures over 6 time points. It has much the same structure as the chocolate cake example. This suggests using the same analysis.

- However, there is one important difference: In the RM design, the "split-plot" factor, time, is not randomized!

Instead, each experimental unit is measured under "time 1" first, "time 2" second, etc. This means that observations are subject to serial correlation as well as shared-characteristics, or clustering-type, correlation.

Recall that in the split-plot model, $\mathbf{y}_{hi}$, the vector of observation on the h, i^{th} whole plot, had the compound symmetry variance-covariance structure:

$$\text{var}(\mathbf{y}_{hi}) = (\sigma_e^2 + \sigma^2) \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

where $\rho = \sigma_e^2/(\sigma^2 + \sigma_e^2)$.

- Often, this seems an inappropriate variance-covariance structure for repeated measures. Typically, we would expect observations taken close together in time to be more highly correlated than observations taken far apart.

- That is, we often expect a decaying correlation structure through time, rather than constant correlation through time.

Sphericity:

It turns out that compound symmetry is a sufficient but not necessary condition for the F tests from the split-plot analysis to be valid for the RM-ANOVA design.

A more general condition, known as **sphericity,** is necessary and sufficient. Sphericity can be expressed in several different, but equivalent ways. In particular, sphericity is equivalent to

1. the variances of all pairwise differences between repeated measures are equal; that is,

$$\text{var}(y_{hij} - y_{hik}) \quad \text{is constant for all } j, k.$$

2. $\epsilon = 1$, where

$$\epsilon = \frac{t^2(\bar{\sigma}_{jj} - \bar{\sigma}_{..})^2}{(t-1)(\sum_{j=1}^{t}\sum_{j'=1}^{t}\sigma_{jj'}^2 - 2t\sum_{j=1}^{t}\bar{\sigma}_{j.}^2 + t^2\bar{\sigma}_{..}^2)},$$

where

$$\begin{aligned}\bar{\sigma}_{..} &= \text{mean of all elements of } \Sigma \equiv \text{var}(\mathbf{y}_{hi}) \\ \bar{\sigma}_{jj} &= \text{mean of elements on main diagonal of } \Sigma \\ \bar{\sigma}_{j.} &= \text{mean of elements in row } j \text{ of } \Sigma\end{aligned}$$

- Since compound symmetry means that $\text{var}(y_{hij})$ and $\text{cov}(y_{hij}, y_{hij'})$ are constant for all i, j, j', it is clear that

$$\text{var}(y_{hij} - y_{hij'}) = \text{var}(y_{hij}) + \text{var}(y_{hij'}) - 2\text{cov}(y_{hij}, y_{hij'})$$

is constant for all j, j'. Therefore, compound symmetry is a special case of sphericity.

- Sphericity is a more general (weaker) condition than compound symmetry, and mathematically, it is all that is needed. However, it is difficult to envision a realistic form for Σ in which sphericity would hold and compound symmetry would not.

Mauchly has proposed a test for sphericity. This test is of limited practical use for several reasons:

- Low power in small samples.
- In large samples test is likely to reject sphericity when non-sphericity has little effect on validity of the split-plot F tests.
- Sensitive to departures from normality.
- Very sensitive to outliers.

It can be shown that sphericity holds when $\epsilon = 1$ and maximum non-sphericity holds when $\epsilon = 1/(t-1)$.

Under non-sphericity, it can be shown that the F test statistics for the repeated measures factor (usually time) and interactions involving the repeated measures factor have approximate F distributions where the degrees of freedom are the usual degrees of freedom multiplied by ϵ.

- Therefore, the "fix" of the split-plot analysis is to multiply the numerator and denominator degrees of freedom by $\hat{\epsilon}$ for all F tests on time and on and interactions involving time.

Two estimators of ϵ are commonly used: Greenhouse-Geisser and Hunyh-Feldt. $\hat{\epsilon}_{GG}$ is simply ϵ computed on the sample variance-covariance matrix S rather than Σ, the population (true) variance-covariance matrix. $\hat{\epsilon}_{HF}$ is defined as

$$\hat{\epsilon}_{HF} = \min \left(1, \underbrace{\frac{n.(t-1)\hat{\epsilon}_{GG} - 2}{(t-1)(\text{d.f.}_{WPE} - (t-1)\hat{\epsilon}_{GG})}}_{\text{value given by SAS}} \right),$$

where $n.$ =the total number of "whole plots".

- Note that $\epsilon \leq 1$, so the value given by SAS should always be rounded down to 1, if it is > 1.

Which $\hat{\epsilon}$? For true $\epsilon \leq 0.5$ (greater non-sphericity) $\hat{\epsilon}_{GG}$ is better. For true $\epsilon \geq 0.75$ (less non-sphericity) $\hat{\epsilon}_{HF}$ is better. In practice we don't know the true value of ϵ so often it is hard to say which is better.

- $\hat{\epsilon}_{GG}$ tends to give the larger adjustment (makes it harder to reject H_0) than $\hat{\epsilon}_{HF}$, so if we desire to be conservative (slow to reject) then we should use $\hat{\epsilon}_{GG}$.

If we have a program like SAS that can compute $\hat{\epsilon}_{GG}$ and $\hat{\epsilon}_{HF}$ and corresponding adjusted p-values easily, then we should always go ahead and do the adjustment for non-sphericity.

- There is no down-side here, because if the data are spherical, then we should get $\hat{\epsilon}_{GG} = \hat{\epsilon}_{HF} = 1$, and the adjustment will end up not altering the split-plot analysis at all (which is what we would want). If the data are non-spherical, then an appropriate adjustment will be done.

- If we want to avoid computing $\hat{\epsilon}$, then a very conservative approach is to use the adjustment for maximum non-sphericity. That is, multiply numerator and denominator d.f. of the F tests by $(t-1)^{-1}$.

- Alternatively, use the Greenhouse-Geisser algorithm (see Davis, §5.3.2).

Example — Methemoglobin in Sheep:

- See sheep1.sas. In this SAS program PROC MIXED is used first to perform the split-plot analysis exactly as in the chocolate cake example, and then the REPEATED statement in PROC GLM is used to reproduce this split-plot analysis as a RM-ANOVA analysis. That is, both PROCs fit the model

$$y_{hij} = \mu_{hj} + e_{i(h)} + \varepsilon_{hij},$$

where $\mu_{hj} = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj}$, $\{e_{i(h)}\} \overset{iid}{\sim} N(0, \sigma_e^2)$, $\{\varepsilon_{hij}\} \overset{iid}{\sim} N(0, \sigma^2)$, and the $e_{i(h)}$'s and ε_{hij}'s are independent of each other.

- The two sets of results are basically the same, but PROC GLM gives Mauchly's test and Greenhouse-Geisser and Hunyh-Feldt adjusted p-values for tests involving time.

- For the basic ANOVA table and F tests, both procedures give the correct RM-ANOVA results. However, PROC GLM will still not give the correct inferences on means and contrasts.

- PROC MIXED would give the correct inferences under the assumption of sphericity if we were to add ESTIMATE and CONTRAST statements to the program in sheep1.sas. In addition, PROC MIXED can be made to "fix" the split-plot analysis for non-sphericity, but with a more sophisticated fix than that we've discussed and that which is implemented in PROC GLM with the REPEATED statement. We'll get to this later.

- The basic results of the analysis are as follows:

  - According to Mauchly's test (the one labelled, "Orthogonal Components" on p.8), there is significant evidence of non-sphericity, so we should use adjusted p-values for tests on time and no2×time. The estimates of ϵ are $\hat{\epsilon}_{GG} = .2610$ and $\hat{\epsilon}_{HF} = .3551$, which are both pretty far from 1, indicating non-sphericity.

– At level $\alpha = .05$, we reject the hypothesis of no interaction with either adjustment (HF or GG, see p.10). The nature of the interaction can be seen in the profile plot on p.5. From the profile plot it appears that we should make inferences on time separately within each of the no2 groups, but it does seem meaningful to test for a main effect of no2.

– The main effect test of no2 is significant ($p = .0026$, see p.9) and needs no adjustment. It is clear that the mean response is higher (at least after time 1) for increasing levels of no2.

– The main effect of time is significant, but, as noted above, the pattern over time seems to be different enough from one no2 group to the next that it really would be more appropriate to compare times separately within each no2 group.

– A natural set of contrasts to examine here would be linear and nonlinear contrasts over time separately in each no2 group. E.g., for no2 group 1, assuming that the times were equally spaced, these contrasts would look like this in SAS:

```
contrast 'linear time, no2=1'          time -5 -3 -1 1 3 5
                                     no2*time -5 -3 -1 1 3 5
                                     0 0 0 0 0 0 0 0 0 0 0 0;
contrast 'nonlinear time, no2=1'       time 5 -1 -4 -4 -1 5
                                     no2*time 5 -1 -4 -4 -1 5
                                     0 0 0 0 0 0 0 0 0 0 0 0,
                                       time -5 7 4 -4 -7 5
                                     no2*time -5 7 4 -4 -7 5
                                     0 0 0 0 0 0 0 0 0 0 0 0,
                                       time 1 -3 2 2 -3 1
                                     no2*time 1 -3 2 2 -3 1
                                     0 0 0 0 0 0 0 0 0 0 0 0,
                                       time -1 5 -10 10 -5 1
                                     no2*time -1 5 -10 10 -5 1
                                     0 0 0 0 0 0 0 0 0 0 0 0;
```

Multivariate Methods for Repeated Measures

The Multivariate Linear Model:

Suppose we have a t component response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})^T$ on the i^{th} of n subjects, and suppose that y_{ij} is generated from the linear model

$$y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, t,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of p explanatory variables specific to the i^{th} subject (but constant over the t components of the response), and $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})^T$ is a vector of unknown parameters specific to the j^{th} component of the response.

Let $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{it})^T$ denote the vector of error terms for the i^{th} subject.

We assume that the t components of the response are correlated within a given subject, so we assume

$$\boldsymbol{\varepsilon}_i \sim N_t(\mathbf{0}, \Sigma).$$

- In applications to repeated measures data, the t components of the response vector correspond to the response measured at t distinct time points. So, Σ describes the variance-covariance structure through time.

- To ensure that Σ is positive definite, we assume $p \leq n - t$.

We also assume independence between subjects so

$$\boldsymbol{\varepsilon} \equiv \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} \sim N_{nt}(\mathbf{0}, \mathbf{I}_n \otimes \Sigma).$$

- Here, $\otimes$ denotes the Kronecker (aka direct) product. $\mathbf{W} \otimes \mathbf{Z}$, the Kronecker product of matrices $\mathbf{W}_{a\times b}, \mathbf{Z}_{c\times d}$, results in the $ac \times bd$ matrix

$$\begin{pmatrix} w_{11}\mathbf{Z} & w_{12}\mathbf{Z} & \cdots & w_{1b}\mathbf{Z} \\ w_{21}\mathbf{Z} & w_{22}\mathbf{Z} & \cdots & w_{2b}\mathbf{Z} \\ \vdots & \vdots & \ddots & \vdots \\ w_{a1}\mathbf{Z} & w_{a2}\mathbf{Z} & \cdots & w_{ab}\mathbf{Z} \end{pmatrix}$$

  Thus, $\mathbf{I}_n \otimes \Sigma$ is the $nt \times nt$ block-diagonal matrix

$$\begin{pmatrix} \Sigma & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma \end{pmatrix}$$

- Thus, according to this model, $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independent random vectors with $\mathbf{y}_i \sim N_t(\boldsymbol{\mu}_i, \Sigma)$ where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{it})^T$, $\mu_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j$.

To express this model in matrix terms, let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1t} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nt} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix},$$

where $\mathbf{X}$ is of rank $p \leq (n - t)$. Also, let

$$\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_t) = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1t} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pt} \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} \boldsymbol{\varepsilon}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_n^T \end{pmatrix} = \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1t} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nt} \end{pmatrix}$$

Then the multivariate linear model takes the form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \quad \text{vec}(\mathbf{E}^T) \sim N_{nt}(\mathbf{0}, \mathbf{I}_n \otimes \Sigma). \qquad (*)$$

Estimation:

The maximum likelihood and ordinary least squares estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Note that $\hat{\mathbf{B}}$ is equal to $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_t)$, where $\hat{\boldsymbol{\beta}}_j = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{u}_j$ is the usual least squares estimator based just on $\mathbf{u}_j$, the j^{th} column of $\mathbf{Y}$.

- $\hat{\mathbf{B}}$ is the BLUE.

That the OLS estimator is the BLUE can be seen by writing the multivariate model as a linear model. Let vec($\mathbf{M}$) denote the vector formed by stacking the columns of its matrix argument $\mathbf{M}$. The the multivariate linear model (*) can be written in a univariate form as

$$\text{vec}(\mathbf{Y}) = (\mathbf{I}_p \otimes \mathbf{X})\text{vec}(\mathbf{B}) + \text{vec}(\mathbf{E}).$$

It is easily seen that the error term has moments $\text{E}\{\text{vec}(\mathbf{E})\} = \mathbf{0}$ and $\text{var}\{\text{vec}(\mathbf{E})\} = \Sigma \otimes \mathbf{I}_n$. Therefore, this model has the form of a GLS model, which implies that the GLS estimator would be BLUE.

However, the same theorem we alluded to in claiming OLS to yield BLUEs in the split-plot model (see p.19) applies here. That theorem says that in the univariate linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V},$$

the OLS estimator of $\boldsymbol{\beta}$ is BLUE iff $C(\mathbf{VX}) \subset C(\mathbf{X})$ (see Graybill, Thm 6.8.1, or Christensen, Thm 10.4.5).

This condition is easy to show here because it is a property of Kronecker products that $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ for suitably conformable matrices. Therefore,

$$(\Sigma \otimes \mathbf{I}_n)(\mathbf{I}_p \otimes \mathbf{X}) = \Sigma \otimes \mathbf{X} = (\mathbf{I}_p \otimes \mathbf{X})(\Sigma \otimes \mathbf{I}).$$

So,

$$C([\Sigma \otimes \mathbf{I}_n][\mathbf{I}_p \otimes \mathbf{X}]) = C([\mathbf{I}_p \otimes \mathbf{X}][\Sigma \otimes \mathbf{I}]) \subset C([\mathbf{I}_p \otimes \mathbf{X}]).$$

The MLE of Σ is

$$\frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}).$$

However, this estimator is biased. An unbiased estimator of Σ is

$$\mathbf{S} = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}).$$

Estimation of linear functions of $\mathbf{B}$ is often of interest. Let $\psi = \mathbf{a}^T\mathbf{B}\mathbf{c}$ where $\mathbf{a}$ and $\mathbf{c}$ are $p \times 1$ and $t \times 1$ vectors of constants, respectively.

- Note that $\mathbf{a}$ operates as a contrast within time points. $\mathbf{c}$ operates as a contrast across time points.

The BLUE of ψ is

$$\hat{\psi} = \mathbf{a}^T\hat{\mathbf{B}}\mathbf{c}$$

and has variance

$$\text{var}(\hat{\psi}) = (\mathbf{c}^T\Sigma\mathbf{c})[\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}].$$

Hypothesis Testing:

Most hypotheses of interest in the multivariate linear model can be expressed as

$$H_0 : \mathbf{ABC} = \mathbf{D}, \qquad (\dagger)$$

where

$\mathbf{A}_{a\times p}$	has rank $a \leq p$, and operates across subjects (w/in time)
$\mathbf{C}_{t\times c}$	has rank $c - p$, and operates across time (w/in subjects)
$\mathbf{D}_{a\times c}$	is a matrix of constants; often, $\mathbf{D} = \mathbf{0}_{a\times c}$.

- This framework is very general. E.g., setting $\mathbf{A} = \mathbf{I}$, $\mathbf{D} = \mathbf{0}$ yields contrasts in time, $\mathbf{C} = \mathbf{I}$, $\mathbf{D} = \mathbf{0}$ yields contrasts across subjects, $\mathbf{A} = \mathbf{I}$, $\mathbf{C} = \mathbf{I}$, $\mathbf{D} = \mathbf{0}$ yields the hypothesis $\mathbf{B} = \mathbf{0}$, etc.

There are four tests commonly used to test hypotheses of this form, and all of the test statistics are defined in terms of the **hypothesis SSCP** matrix and the **error, or residual, SSCP** matrix

- Here, SSCP stands for sum of squares and cross-products. A SSCP is the multivariate analog of a sum of squares in the univariate linear model. Note that it is a matrix, not a scalar.

Recall that in the univariate linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

an F test statistic for the hypothesis $H_0 : \mathbf{A}_{a\times p}\boldsymbol{\beta}_{p\times 1} = \mathbf{d}_{\mathbf{a}\times 1}$ was given by

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{d})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{d})}{[\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}]}\frac{n-p}{a} = \frac{SSH}{SSE}\frac{n-p}{a}.$$

In the multivariate context, SSH becomes a matrix, the hypothesis SSCP, given by

$$Q_h = (\mathbf{A}\hat{\mathbf{B}}\mathbf{C} - \mathbf{D})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{C} - \mathbf{D}),$$

and SSE becomes a matrix, the error SSCP, given by

$$Q_e = \mathbf{C}^T[\mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{B}}^T(\mathbf{X}^T\mathbf{X})\hat{\mathbf{B}}]\mathbf{C}.$$

- Since these quantities are matrices in the multivariate context, we can no longer compare them as simply as in the univariate F test. That is, we cannot simply take their ratio. Even computing $\mathbf{Q}_h\mathbf{Q}_e^{-1}$ is not an option, because the result is not a scalar test statistic.

How to do we compare the "sizes" of $\mathbf{Q}_h$ and $\mathbf{Q}_e$ with a scalar quantity?

Several answers have been put forward, leading to several different test statistics:

- Roy's test: based on the largest eigenvalue of $\mathbf{Q}_h\mathbf{Q}_e^{-1}$.
- Lawley and Hotelling's test: test statistic is $\text{tr}(\mathbf{Q}_h\mathbf{Q}_e^{-1})$.
- Pillai's test: test statistic is a function of $\text{tr}[\mathbf{Q}_h(\mathbf{Q}_h + \mathbf{Q}_e)^{-1}]$.
- Wilks' likelihood ratio test: Wilks' test statistic depends on $\Lambda = |\mathbf{Q}_e|/|\mathbf{Q}_h + \mathbf{Q}_e|$ and is equivalent to the LRT.

Unfortunately, none of these tests is "best" in all situations. However, all of these tests are approximately equivalent in large samples, and differ very little in power for small samples.

- Because LRTs have good properties in general, **we will confine attention to Wilks' test.**

There is no general result that gives the exact distribution of Wilks' test statistic. That is, the exact reference distribution for Wilks' test is not known, in general.

However, for some multivariate ANOVA (MANOVA) models that arise in profile analysis, there is an equivalence between Wilks' test and an F test, where the exact reference distribution is an F distribution.

In particular, in a one-way MANOVA model for comparing (s) groups (treatments) based upon a t-variate response, exact distributions are available for the following cases:

Case	Test Statistic	Distribution
t=1 $s\geq 2$	$\left(\frac{n-s}{s-1}\right)\left(\frac{1-\Lambda}{\Lambda}\right)$	$F(s-1, n-s)$
t=2 $s\geq 2$	$\left(\frac{n-s-1}{s-1}\right)\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)$	$F(2(s-1), 2(n-s-1))$
$t\geq 1$ s=2	$\left(\frac{n-t-1}{t}\right)\left(\frac{1-\Lambda}{\Lambda}\right)$	$F(t, n-t-1)$
$t\geq 1$ s=3	$\left(\frac{n-t-2}{t}\right)\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)$	$F(2t, 2(n-t-2))$

where $n = \sum_{h=1}^{s} n_h$ is the total number of subjects.

- In other cases we rely on approximations to obtain $p-$values for Wilks' Lambda.

- A large sample approximation due to Bartlett gives the following rejection rule to obtain an approximate $\alpha-$level test: Reject H_0 if

$$-\left(n-1-\frac{t+s}{2}\right)\log(\Lambda) > \chi^2_\alpha(t(s-1)).$$

- Other approximations are available to obtain approximate $p-$values when the total sample size n is small. These approximation are implemented in SAS and other computer programs and are quite good even for small sample sizes.

Profile Analysis:

We maintain the same notation and set-up: suppose repeated measures at t time points from s groups of subjects. Let n_h = number of subjects in group h, and let $n = \sum_{h=1}^{s} n_h$. Let y_{hij} denote the observation at time j for subject i in group h.

We assume the response vectors $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hit})^T$ are independent, with

$$\mathbf{y}_{hi} \sim N(\boldsymbol{\mu}_h, \Sigma), \quad \text{where} \quad \boldsymbol{\mu}_h = \begin{pmatrix} \mu_{h1} \\ \vdots \\ \mu_{ht} \end{pmatrix},$$

and $\mu_{hj} = \mathrm{E}(y_{hij})$.

The profile analysis model is

$$y_{hij} = \mu_{hj} + \varepsilon_{hij}, \quad \text{where} \quad \boldsymbol{\varepsilon}_{hi} = \begin{pmatrix} \varepsilon_{hi1} \\ \vdots \\ \varepsilon_{hit} \end{pmatrix} \sim N(\mathbf{0}, \Sigma).$$

In terms of the multivariate general linear model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \qquad \text{or}$$

$$\begin{pmatrix} \mathbf{y}_{11}^T \\ \vdots \\ \mathbf{y}_{1n_1}^T \\ \hline \mathbf{y}_{21}^T \\ \vdots \\ \mathbf{y}_{2n_2}^T \\ \hline \vdots \\ \hline \mathbf{y}_{s1}^T \\ \vdots \\ \mathbf{y}_{sn_s}^T \end{pmatrix} = \left(\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{array}\right) \begin{pmatrix} \mu_{11} & \cdots & \mu_{1t} \\ \mu_{21} & \cdots & \mu_{2t} \\ \vdots & \ddots & \vdots \\ \mu_{s1} & \cdots & \mu_{st} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{11}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{1n_1}^T \\ \hline \boldsymbol{\varepsilon}_{21}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{2n_2}^T \\ \hline \vdots \\ \hline \boldsymbol{\varepsilon}_{s1}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{sn_s}^T \end{pmatrix}$$

Three general hypotheses are of interest in profile analysis:

H_{01} : the mean profiles (over time) for the s groups are parallel (i.e., no group$\times$time interaction);
H_{02} : no differences among groups;
H_{03} : no differences among time points.

- Note that H_{01} should be tested first, because the result of this test affects what form the other two hypotheses should take (H_{02} and H_{03} have been expressed in a purposely vague way here).

If H_{01} is accepted, then, under the assumption that no interaction is present, it is appropriate to test for no difference between groups by comparing the mean response in each group averaged over all time points, and it is appropriate to test no difference across time points by comparing the mean response at each time, averaged over groups.

If, however, we reject H_{01} then it may be more appropriate to test hypotheses of the form

H_{04} : no difference among groups within some subset of the measurement occasions;
H_{05} : no difference among time points in a particular group, or subset of groups;
H_{06} : no difference within some subset of measurement occasions in a particular group or subset of groups.

Test of parallelism:

The hypothesis of parallelism is

$$H_{01}: \begin{pmatrix} \mu_{11} - \mu_{12} \\ \mu_{12} - \mu_{13} \\ \vdots \\ \mu_{1,t-1} - \mu_{1t} \end{pmatrix} = \begin{pmatrix} \mu_{21} - \mu_{22} \\ \mu_{22} - \mu_{23} \\ \vdots \\ \mu_{2,t-1} - \mu_{2t} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{s1} - \mu_{s2} \\ \mu_{s2} - \mu_{s3} \\ \vdots \\ \mu_{s,t-1} - \mu_{st} \end{pmatrix}.$$

- Testing this hypothesis is equivalent to conducting a one-way multivariate analysis of variance (MANOVA) on the $t-1$ differences between adjacent time points from each subject.

In terms of the general form of the hypothesis, H_{01} can be expressed as $\mathbf{ABC} = \mathbf{D}$ where

$$\mathbf{A}_{(s-1)\times s} = (\mathbf{I}_{s-1}, -\mathbf{j}_{s-1}),$$

$$\mathbf{C}_{t\times(t-1)} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix}$$

$$\mathbf{D}_{(s-1)\times(t-1)} = \mathbf{0}_{(s-1)\times(t-1)}$$

Test of no difference among groups:

Depending on the result of the test of H_{01}, two tests of no difference among groups are possible.

First, if H_{01} is accepted, then we would test for differences across groups averaging over (or equivalently, summing over) time points. In this case H_{02} takes the form $H_{02a} : \mathbf{ABC} = \mathbf{D}$ where

$$\begin{aligned} \mathbf{A}_{(s-1)\times s} &= (\mathbf{I}_{s-1}, -\mathbf{j}_{s-1}), \\ \mathbf{C}_{t\times 1} &= \mathbf{j}_t \\ \mathbf{D}_{(s-1)\times 1} &= \mathbf{0}_{(s-1)\times 1}. \end{aligned}$$

- This test is equivalent to doing a one-way ANOVA on the totals (or means) across time, for each subject.

Second, if H_{01} is rejected, we would not want to assume parallelism in testing across groups. In this case the null hypothesis is

$$H_{02b} : \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1t} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2t} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2t} \end{pmatrix},$$

or $H_{02b} : \mathbf{ABC} = \mathbf{D}$, where

$$\begin{aligned} \mathbf{A}_{(s-1)\times s} &= (\mathbf{I}_{s-1}, -\mathbf{j}_{s-1}), \\ \mathbf{C}_{t\times 1} &= \mathbf{I}_t \\ \mathbf{D}_{(s-1)\times t} &= \mathbf{0}_{(s-1)\times t}. \end{aligned}$$

- This is the one-way MANOVA test on the vector of means at each time point.

Test of no difference among time points:

Similar to testing no difference among groups, the appropriate test here depends upon the result of testing H_{01}. If H_{01} is accepted, we will typically want to test no difference across time points, averaging (or equivalently, summing) across groups. This hypothesis is $H_{03a} : \mathbf{ABC} = \mathbf{D}$ where

$$\begin{aligned}\mathbf{A}_{1\times s} &= \mathbf{j}_s^T \quad \text{or} \quad \frac{1}{s}\mathbf{j}_s^T \\ \mathbf{C}_{t\times(t-1)} &= \begin{pmatrix} \mathbf{I}_{t-1} \\ -\mathbf{j}_{t-1}^T \end{pmatrix} \\ \mathbf{D}_{1\times(t-1)} &= \mathbf{0}_{1\times(t-1)}.\end{aligned}$$

If H_{01} is rejected, then we can compare time points without assuming parallelism. That is, we can test the hypothesis

$$H_{03b} : \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{s1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{s2} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \\ \vdots \\ \mu_{st} \end{pmatrix}.$$

This hypothesis can also be written in the form $\mathbf{ABC} = \mathbf{D}$ where

$$\begin{aligned}\mathbf{A}_{s\times s} &= \mathbf{I}_s, \\ \mathbf{C}_{t\times(t-1)} &= \begin{pmatrix} \mathbf{I}_{t-1} \\ -\mathbf{j}_{t-1}^T \end{pmatrix}, \\ \mathbf{D}_{s\times(t-1)} &= \mathbf{0}_{s\times(t-1)}.\end{aligned}$$

Example — Methemoglobin in Sheep (again):

Recall that there are $t = 6$ measurements through time on each of $n_1 = n_2 = n_3 = 4$ sheep in NO_2 groups 1, 2 and 3 ($s = 3$). The profile analysis model for these data is

$$\begin{pmatrix} \mathbf{y}_{11}^T \\ \vdots \\ \mathbf{y}_{14}^T \\ \mathbf{y}_{21}^T \\ \vdots \\ \mathbf{y}_{24}^T \\ \mathbf{y}_{31}^T \\ \vdots \\ \mathbf{y}_{34}^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{16} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{26} \\ \mu_{31} & \mu_{32} & \cdots & \mu_{36} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{11}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{14}^T \\ \boldsymbol{\varepsilon}_{21}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{24}^T \\ \boldsymbol{\varepsilon}_{31}^T \\ \vdots \\ \boldsymbol{\varepsilon}_{34}^T \end{pmatrix},$$

or $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ where $\mathbf{Y}$ and $\mathbf{E}$ are 12×6 matrices, $\mathbf{X}$ is 12×3, and $\mathbf{B}$ is 3×6.

The hypothesis of parallelism is $H_{01} : \mathbf{ABC} = \mathbf{0}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

- See handout sheep2.sas. In this handout, PROC GLM is used to fit the MANOVA model and to illustrate how to test the profile analysis hypotheses.

- In the first call to PROC GLM, we test for parallelism. Note that **A** is specified with the CONTRAST statement, the transpose of **C** is specified with the MANOVA statement, and **D** is assumed to be equal to **0**.

- According to Wilks' test, the hypothesis of parallelism is rejected at $\alpha = .05$ ($p = .0164$).

- Given that H_{01} is rejected, we should compare groups and times without assuming parallelism, and/or compare times within each group separately and compare groups with each time separately. However, for illustration purposes, I've given the tests of H_{02a}, H_{02b}, H_{03a}, and H_{03b} in sheep2.sas.

  - The hypothesis of no group effect assuming parallelism (H_{02a}) is rejected ($p = .0026$),
  - the hypothesis of no group effect without assuming parallelism (H_{02b}) is rejected ($p = .0350$),
  - the hypothesis of no time effect assuming parallelism (H_{03a}) is rejected ($p < .0001$), and
  - the hypothesis of no time effect without assuming parallelism (H_{03b}) is rejected ($p < .0001$).

- Finally, I tested for no time effect in group 1 only. This hypothesis was also rejected ($p = .0006$).

- All of these results are consistent with the profile plot obtained in sheep1.sas.

Growth Curve Analysis:

We have seen that repeated measures of a single variable (methemoglobin, say) over time can be analyzed with multivariate methods (e.g., MANOVA) by regarding each time-specific measurement of the variable as a distinct variable.

- E.g., if we measure methemoglobin at 10, 20, 30, 40, 50, and 60 minutes after treatment for each subject in each of three treatment groups, we can compare the groups with a MANOVA based on $t = 6$ variables: methemoglobin at 10 min., methemoglobin at 20 min., ..., methemoglobin at 60 min.

- Such an approach does not recognize any ordering of the repeated measurements and fits no model to describe time trends or **growth curves.**

- In fact, repeated measurements through time are naturally ordered.

In this case, it may be of interest to characterize trends over time using low-order polynomials (e.g., linear or quadratic curves in time).

By *modelling* the time trend, we hope to summarize the mean response at the t time points with $q < t$ parameters, rather than allowing for t separate time-specific means.

- The use of polynomials to describe time trend *within the context of a multivariate linear model* is known as **growth curve analysis,** and is usually attributed to Potthoff and Roy.

- Not to be confused with the use of nonlinear models of growth (e.g., Richards' model, von Bertalanffy's model, etc.).

- This approach is seldom used these days, so we will not discuss it further. The use of polynomials in time to describe patterns of change is still common, but this is more commonly done in the framework of linear mixed models these days.

Linear Mixed Effects Models (LMMs)

There are several disadvantages/limitations to multivariate methods (profile analysis, growth curves) for longitudinal data analysis.

- Methods assume same set of measurement times for each subject, so cannot handle missing data, varying measurement times, varying cluster size (number of repeated measures) easily.

- Cannot handle time-varying covariates easily.

- Models make no assumptions on within-subject var-cov matrix. This makes these methods broadly valid, but not powerful.

- M'variate methods don't model sources of heterogeneity/correlation in the design generating the data. No quantification of heterogeneity, little flexibility to model multiple sources of heterogeneity and correlation.

A much more flexible class of models is the class of linear mixed effects models (LMMs).

We have already seen examples of LMMs: the split-plot model (chocolate cake example), RM-ANOVA model (methemoglobin in sheep example).

- In these cases, a cluster-specific random effect (whole plot error term) was included to model whole plot to whole plot or subject to subject variability and to imply correlation within a whole-plot/subject.

In general, the inclusion of random effects into the linear model allows for modeling (and quantification) of multiple sources of heterogeneity and complex patterns of correlation.

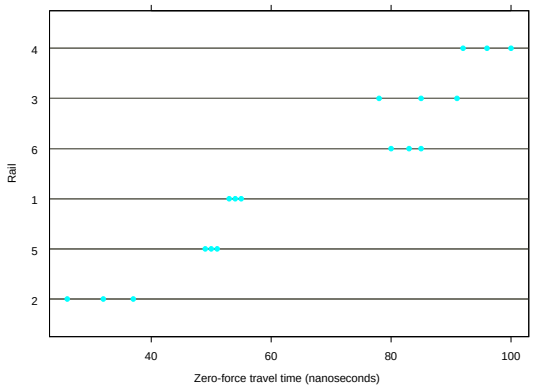
Further flexibility is achieved in this class of models by also letting the error term have a general, non-spherical variance-covariance matrix.

The result is a very rich, flexible and useful class of models.

Some Simple LMMs:

The one-way random effects model — Railway Rails:

(See Pinheiro and Bates, §1.1) The data displayed below are from an experiment conducted to measure longitudinal (lengthwise) stress in railway rails. Six rails were chosen at random and tested three times each by measuring the time it took for a certain type of ultrasonic wave to travel the length of the rail.



Clearly, these data are grouped, or clustered, by rail. This clustering has two closely related implications:

1. (within-cluster correlation) we should expect that observations from the same rail will be more similar to one another than observations from different rails; and

2. (between cluster heterogeneity) we should expect that the mean response will vary from rail to rail in addition to varying from one measurement to the next.

These ideas are really flip-sides of the same coin.

Although it is fairly obvious that clustering by rail must be incorporated in the modeling of these data somehow, we first consider a naive approach.

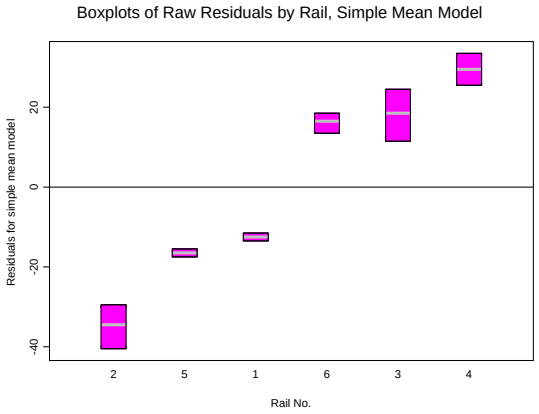
The primary interest here is in measuring the mean travel time. Therefore, we might naively consider the model

$$y_{ij} = \mu + \varepsilon_{ij}, \quad i = 1, \ldots, 6, j = 1, \ldots, 3,$$

where y_{ij} is the travel time for the j^{th} trial on the i^{th} rail, and we assume $\varepsilon_{11}, \ldots, \varepsilon_{63} \overset{iid}{\sim} N(0, \sigma^2)$.

Here, μ is the mean travel time which we wish to estimate. Its ML/OLS estimate is $\bar{y}_{\cdot\cdot} = 66.5$ and the MSE is $s^2 = 23.645^2$.

However, an examination of the residuals form this model plotted separately by rail reveals the inadequacy of the model:



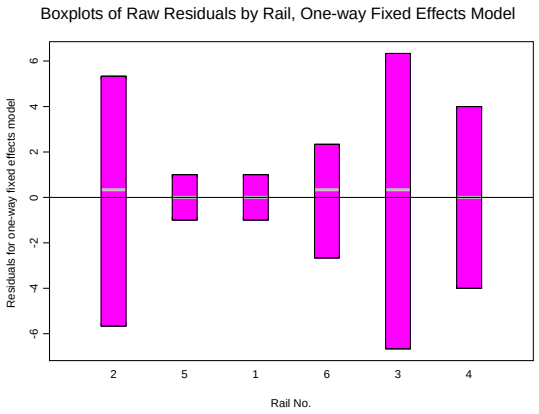
Clearly, the mean response is changing from rail to rail. Therefore, we consider a one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}. \qquad (*)$$

Here, μ is a grand mean across the rails included in the experiment, and α_i is an effect up or down from the grand mean specific to the i^{th} rail. Alternatively, we could define $\mu_i = \mu + \alpha_i$ as the mean response for the i^{th} rail and reparameterize this model as

$$y_{ij} = \mu_i + \varepsilon_{ij}.$$

The OLS estimates of the parameters of this model are $\hat{\mu}_i = \bar{y}_{i\cdot}$, of $(\hat{\mu}_1, \ldots, \hat{\mu}_6) = (54.00, 31.67, 84.67, 96.00, 50.00, 82.67)$ and $s^2 = 4.02^2$. The residual plot looks much better:



However, there are still drawbacks to this one-way fixed effects model:

- It only models the specific sample of rails used in the experiment, while the main interest is in the population of rails from which these rails were drawn.

- It does not produce an estimate of the rail-to-rail variability in travel time, which is a quantity of significant interest in the study.

- The number of parameters increases linearly with the number of rails used in the experiment.

These deficiencies are overcome by the one-way random effects model.

To motivate this model, consider again the one-way fixed effects model. Model (*) can be written as

$$y_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$$

where, under the usual constraint $\sum_i \alpha_i = 0$, $(\mu_i - \mu)$ has mean 0 when averaged over the groups (rails).

The one-way random effects model, replaces the fixed parameter $(\mu_i - \mu)$ with a random effect b_i, a random variable specific to the i^{th} rail, which is assumed to have mean 0 and an unknown variance σ_b^2. This yields the model

$$y_{ij} = \mu + b_i + \varepsilon_{ij},$$

where $b_1, \ldots, b_6$ are independent random variables, each with mean 0 and variance σ_b^2. Often, the b_i's are assumed normal, and they are usually assumed independent of the ε_{ij}'s. Thus we have

$$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_b^2), \quad \text{independent of} \quad \varepsilon_{11} \ldots, \varepsilon_{nt_n} \overset{iid}{\sim} N(0, \sigma^2),$$

where n is the number of rails, t_i the number of observations on the i^{th} rail.

– Note that now the interpretation of μ changes from the mean over the 6 rails included in the experiment (fixed effects model) to the mean over the population of all rails from which the six rails were sampled.

– In addition, we are not estimating μ_i the mean response for a single rail, which is not of interest. Instead we are estimating the population mean μ and the variance from rail to rail in the population, σ_b^2.

– That is, now our scope of inference is the population of rails, rather than the six rails included in the study.

– In addition, we can estimate rail to rail variability σ_b^2; and

– The number of parameters no longer increases with the number of rails tested in the experiment.

The one-way random effects model is really a simplified version of the split-plot model and it implies a similar variance-covariance structure. It is easy to show that for the one-way random effects model

$$\begin{aligned}
\text{var}(y_{ij}) &= \sigma_b^2 + \sigma^2 \\
\text{cov}(y_{ij}, y_{ij'}) &= \sigma_b^2, \quad j \neq j' \\
\text{corr}(y_{ij}, y_{ij'}) &= \rho \equiv \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}, \quad j \neq j', \quad \text{and} \\
\text{cov}(y_{ij}, y_{i'j'}) &= 0, \quad i \neq i'.
\end{aligned}$$

That is, if $\mathbf{y}_i = (y_{i1}, \ldots, y_{it_i})^T$, then $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independent, with

$$\text{var}(\mathbf{y}_i) = (\sigma_b^2 + \sigma^2) \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

(cf. the split-plot var-cov structure on p.29).

– In the rails example, the one-way random effects model again leads to a BLUE of $\bar{y}_{\cdot\cdot} = 65.5$ for μ.

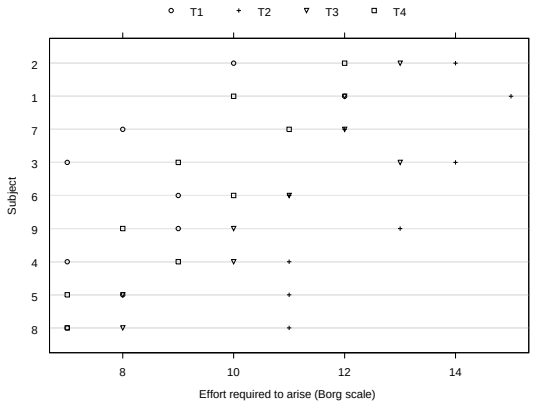
– The **restricted maximum likelihood** (REML) estimators of σ^2 and σ_b^2 coincide with the method of moment type estimators we derived in the split-plot model. These estimates are $\hat{\sigma}_b^2 = 24.805^2$ and $\hat{\sigma}^2 = 4.021^2$.

The randomized complete block model — Stool Example:

In the last example, the data were grouped by rail and we were interested in only one treatment (there was only one experimental condition under which the travel time along the rail was measured).

Often, several treatments are of interest and the data are grouped. In a randomized complete block design (RCBD), each of s treatments are observed in each of n blocks.

As an example, consider the data displayed below. These data come from an experiment to compare the ergonomics of four different stool designs. $b = 9$ subjects were asked to sit in each of $s = 4$ stools. The response measured was the amount of effort required to stand up.



Let y_{ij} be the response for the j^{th} stool type tested by the i^{th} subject. The classical fixed effects model for the RCBD assumes

$$\begin{aligned} y_{ij} &= \mu + \alpha_j + \beta_i + \varepsilon_{ij}, \\ &= \mu_j + \beta_i + \varepsilon_{ij}, \end{aligned} \qquad i = 1, \ldots, n, j = 1, \ldots, s,$$

where $\varepsilon_{11}, \ldots, \varepsilon_{ns} \overset{iid}{\sim} N(0, \sigma^2)$.

Here, μ_j is the mean response for the j^{th} stool type, which can be broken apart into a grand mean μ and a stool type effect α_j. β_i is a fixed subject effect.

Again, the scope of inference for this model is the set of 9 subjects used in this experiment. If we wish to generalize to the population from which the 9 subjects in this experiment were drawn, a more appropriate model would consider the subject effects to be random.

The RCBD model with random subject effects is

$$y_{ij} = \mu_j + b_i + \varepsilon_{ij},$$

where

$$b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_b^2) \quad \text{independent of} \quad \varepsilon_{11}, \ldots, \varepsilon_{ns} \overset{iid}{\sim} N(0, \sigma^2).$$

An equivalent representation is

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{is} \end{pmatrix}, \mathbf{X}_i = \mathbf{I}_s, \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_s \end{pmatrix}, \mathbf{Z}_i = \mathbf{j}_s = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{is} \end{pmatrix}.$$

From this model representation it is clear that the variance-covariance structure here is quite similar to that in the one-way random effects and split plot models. In particular,

$$\begin{aligned}
\text{cov}(\mathbf{y}_i, \mathbf{y}_{i'}) &= \text{cov}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i, \mathbf{X}_{i'}\boldsymbol{\beta} + \mathbf{Z}_{i'} b_{i'} + \boldsymbol{\varepsilon}_{i'}) = 0, \quad i \neq i', \\
\text{var}(\mathbf{y}_i) &= \text{var}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i) = \text{var}(\mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i) \\
&= \mathbf{Z}_i \underbrace{\text{var}(b_i)}_{=\sigma_b^2} \mathbf{Z}_i + \underbrace{\text{var}(\boldsymbol{\varepsilon}_i)}_{=\sigma^2\mathbf{I}_s} = \sigma_b^2 \mathbf{J}_{s,s} + \sigma^2 \mathbf{I}_s \\
&= \begin{pmatrix} \sigma^2 + \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma^2 + \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma^2 + \sigma_b^2 \end{pmatrix}
\end{aligned}$$

It is often stated that whether block effects are assumed random or fixed does not affect the analysis of the RCBD. This is not completely true. It is true that whether or not blocks are treated as random does not affect the ANOVA F test for treatments. The ANOVA table for the RCBD with random block effects is

Source of Variation	Sum of Squares	d.f.	Mean Squares	E(MS)	F
Treat's	$n\sum_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	$s-1$	$\frac{SS_{Trt}}{s-1}$	$\sigma^2 + \frac{n\sum_j(\mu_j - \bar{\mu}_\cdot)^2}{s-1}$	$\frac{MS_{Trt}}{MS_E}$
Blocks	$s\sum_j(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$n-1$	$\frac{SS_{Blocks}}{n-1}$	$\sigma^2 + s\sigma_b^2$	
Error	SS_E (by subtr.)	$(s-1)(n-1)$	$\frac{SS_E}{(s-1)(n-1)}$	σ^2	
Total	$\sum_i\sum_j(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$sn-1$			

– This table is identical to that with blocks fixed except for the expected MS for blocks. The F tests for the two situations are identical.

However, there are important differences in the analysis of the two designs. These differences affect inferences on treatment means.

For instance, in the fixed block effects model, the variance of a treatment mean is

$$\text{var}(\bar{y}_{\cdot j}) = \text{var}\{n^{-1}\sum_i(\mu_j + \beta_i + \varepsilon_{ij})\} = \text{var}(\bar{\varepsilon}_{\cdot j}) = \frac{\sigma^2}{n},$$

whereas in the random block effects model

$$\begin{aligned}\text{var}(\bar{y}_{\cdot j}) &= \text{var}\{n^{-1}\sum_i(\mu_j + b_i + \varepsilon_{ij})\} = \text{var}(\bar{b}_{\cdot} + \bar{\varepsilon}_{\cdot j}) \\ &= \text{var}(\bar{b}_{\cdot}) + \text{var}(\bar{\varepsilon}_{\cdot j}) = \frac{\sigma_b^2}{n} + \frac{\sigma^2}{n} = \frac{\sigma^2 + \sigma_b^2}{n}.\end{aligned}$$

From the expressions for expected MS, method of moment (aka anova) estimators for σ^2 and σ_b^2 are easily derived (cf. p.17, the analogous results for the split-plot model):

$$\begin{aligned}\hat{\sigma}^2 &= MS_E \\ \hat{\sigma}_b^2 &= \frac{MS_{Blocks} - MS_E}{s}\end{aligned}$$

This leads to a standard error of

$$\text{s.e.}(\bar{y}_{\cdot j}) = \sqrt{\hat{\text{var}}(\bar{y}_{\cdot j})} = \sqrt{\frac{MS_{Blocks} + (s-1)MS_E}{ns}}$$

in the random block effects model and a standard error of

$$\text{s.e.}(\bar{y}_{\cdot j}) = \sqrt{\frac{MS_E}{n}}$$

in the fixed block effects model.

– See stool1.sas. Note that the s.e.'s on LSMEANS are larger for the random blocks model. This makes sense, since the scope of inference for this model is broader.

The General LMM — Theory:

In general, we can write the linear mixed model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known matrices (the **model** or **design matrices** for the fixed and random effects, respectively), $\boldsymbol{\beta}$ is a vector of unknown fixed effects (parameters), $\mathbf{b}$ is a vector of random effects, and $\boldsymbol{\varepsilon}$ is a vector of random error terms.

We assume for the random vectors $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ that

$$\begin{aligned} \mathrm{E}(\mathbf{b}) &= \mathbf{0}, \qquad \mathrm{var}(\mathbf{b}) = \mathbf{D}, \\ \mathrm{E}(\boldsymbol{\varepsilon}) &= \mathbf{0}, \qquad \mathrm{var}(\boldsymbol{\varepsilon}) = \mathbf{R}, \\ \text{and} \quad & \mathrm{cov}(\mathbf{b}, \boldsymbol{\varepsilon}) = \mathbf{0}. \end{aligned}$$

- For statistical inference and for likelihood-based estimation we must add distributional assumptions on $\mathbf{b}$ and $\boldsymbol{\varepsilon}$. We make the usual assumptions that

$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}).$$

- Notice that the variance-covariance matrices $\mathbf{D}$ and $\mathbf{R}$ are not assumed to be known, and are of general form. In special cases we will assume spherical errors ($\mathbf{R} = \sigma^2 \mathbf{I}_n$) and/or special forms for $\mathbf{D}$.

For example, in the RCBD model with random block effects, suppose there are $n = 3$ random blocks and $s = 2$ treatments. Then

$$y_{ij} = \mu_j + b_i + \varepsilon_{ij}$$

can be written in the general form (1) as follows:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}}_{=\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{=\mathbf{Z}} \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}}_{=\mathbf{b}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

For estimation of the fixed effects $\boldsymbol{\beta}$, the mixed model can be written as a generalized least-squares model. Define

$$\mathbf{V} \equiv \text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}) = \text{var}(\mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}.$$

- We assume that $\mathbf{V}$ is nonsingular.

Then model (1) is equivalent to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\zeta}, \quad \text{E}(\boldsymbol{\zeta}) = 0, \quad \text{var}(\boldsymbol{\zeta}) = \mathbf{V}.,$$

or

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

If $\mathbf{V}$ were known (at least up to a multiplicative constant), then our results on GLS estimation would apply here, and we would obtain $\hat{\boldsymbol{\beta}}$ as a solution to the equation

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

and then the BLUE of any estimable function $\mathbf{c}^T\boldsymbol{\beta}$ would be given by $\mathbf{c}^T\hat{\boldsymbol{\beta}}$.

However, $\mathbf{V}$ is typically unknown. In that case, suppose we have an estimator $\hat{\mathbf{V}}$ of $\mathbf{V}$. Then a natural approach for estimating $\mathbf{c}^T\boldsymbol{\beta}$ is to treat $\hat{\mathbf{V}}$ as the true value $\mathbf{V}$ and then use the (estimated) GLS estimator

$$\mathbf{c}^T\hat{\boldsymbol{\beta}} = \mathbf{c}^T(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}. \qquad (*)$$

- If $\hat{\mathbf{V}}$ is "close to" $\mathbf{V}$, then $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ should be close to the BLUE of $\mathbf{c}^T\boldsymbol{\beta}$. However, corresponding standard errors based on $\hat{\text{var}}(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \mathbf{c}^T(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{c}$ are known to be underestimated somewhat because they don't account for the error in estimating $\mathbf{V}$ by $\hat{\mathbf{V}}$.

- We will see that the estimator defined by (*) is the ML (REML) estimator of $\boldsymbol{\beta}$ when $\hat{\mathbf{V}}$ is the ML (REML) estimator of $\mathbf{V}$.

Prediction of **b**:

Before considering our specific problem of predicting **b** in the LMM, we need to know a little bit about prediction of random variables, in general.

Suppose we have random variables $y_1, \ldots, y_n$ from which we'd like to predict the random variable y_0. What is the best predictor of y_0?

If we use the mean squared error of prediction as our criterion of optimality, then we can show that the best (minimum mse) predictor is

$$m_{\text{bp}}(\mathbf{y}) \equiv \text{E}(y_0|\mathbf{y}), \quad \text{where} \quad \mathbf{y} = (y_1, \ldots, y_n)^T.$$

- Here the mean squared error of prediction for a predictor $t(\mathbf{y})$ is defined to be $\text{E}[\{y_0 - t(\mathbf{y})\}^2]$. Note this is a criterion of optimality for a predictor $t(\mathbf{y})$. Don't confuse this with the MS_E of a fitted regression model.

This result is stated in the following theorem:

Theorem: Let $m_{\text{bp}}(\mathbf{y}) = \text{E}(y_0|\mathbf{y})$. Then, for any predictor $t(\mathbf{y})$,

$$\text{E}[\{y_0 - t(\mathbf{y})\}^2] \geq \text{E}[\{y_0 - m_{\text{bp}}(\mathbf{y})\}^2].$$

Thus $m_{\text{bp}}(\mathbf{y}) = \text{E}(y_0|\mathbf{y})$ is the **best predictor** of y_0 in the sense of minimizing the mean squared error of prediction.

Proof:

$$\begin{aligned}
\text{E}\left[\{y_0 - t(\mathbf{y})\}^2\right] &= \text{E}\left[\{y_0 - m_{\text{bp}}(\mathbf{y}) + m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}^2\right] \\
&= \text{E}\left[\{y_0 - m_{\text{bp}}(\mathbf{y})\}^2\right] + \text{E}\left[\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}^2\right] \\
&\quad + 2\text{E}\left[\{y_0 - m_{\text{bp}}(\mathbf{y})\}\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\right].
\end{aligned}$$

Since both $\text{E}[\{y_0 - m_{\text{bp}}(\mathbf{y})\}^2]$ and $\text{E}[\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}^2]$ are nonnegative, it suffices to show that $\text{E}\left[\{y_0 - m_{\text{bp}}(\mathbf{y})\}\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\right] = 0$. This is indeed the case because

$$\begin{aligned}
\text{E}\big[\,\{y_0 - m_{\text{bp}}(\mathbf{y})\}\,\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\,\big] &= \text{E}\Big(\text{E}\big[\{y_0 - m_{\text{bp}}(\mathbf{y})\}\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}|\mathbf{y}\big]\Big) \\
&= \text{E}\Big(\text{E}\big[\{y_0 - m_{\text{bp}}(\mathbf{y})\}|\mathbf{y}\big]\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\Big) \\
&= \text{E}\big(\{\underbrace{\text{E}(y_0|\mathbf{y})}_{=m_{\text{bp}}(\mathbf{y})} - m_{\text{bp}}(\mathbf{y})\}\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\big) \\
&= \text{E}\big(0\,\{m_{\text{bp}}(\mathbf{y}) - t(\mathbf{y})\}\,\big) = 0. \qquad \blacksquare
\end{aligned}$$

- To form the best predictor $E(y_0|\mathbf{y})$ we, in general, require knowledge of the joint distribution of $(y_0, y_1, \ldots, y_n)^T$ which may not be available.

- It requires substantially less information to form the **best linear predictor** of y_0 based on $\mathbf{y}$. For the best predictor in the class of linear predictors, we need only the means, variances and covariances of y_0 and $\mathbf{y}$.

Limiting ourselves to the class of linear predictors, we seek a predictor of the form $\gamma_0 + \mathbf{y}^T\boldsymbol{\gamma}$ for some vector $\boldsymbol{\gamma}$ that minimizes $\mathrm{E}\{(y_0 - \gamma_0 - \mathbf{y}^T\boldsymbol{\gamma})^2\}$.

Let $\mu_{y_0} = \mathrm{E}(y_0)$, $\sigma^2_{y_0} = \mathrm{var}(y_0)$, $\boldsymbol{\mu}_\mathbf{y} = \mathrm{E}(\mathbf{y})$, $\mathbf{V}_{\mathbf{yy}} = \mathrm{var}(\mathbf{y})$ and $\mathbf{v}_{\mathbf{y}y_0} = \mathrm{cov}(\mathbf{y}, y_0)$.

Let $\boldsymbol{\gamma}^*$ denote a solution to $\mathbf{V}_{\mathbf{yy}}\boldsymbol{\gamma} = \mathbf{v}_{\mathbf{y}y_0}$. I.e. $\boldsymbol{\gamma}^* = \mathbf{V}^-_{\mathbf{yy}}\mathbf{v}_{\mathbf{y}y_0}$ ($= \mathbf{V}^{-1}_{\mathbf{yy}}\mathbf{v}_{\mathbf{y}y_0}$ in the case that $\mathbf{V}_{\mathbf{yy}}$ is nonsingular). Then the following theorem holds:

Theorem: The function

$$m_{\mathrm{blp}}(\mathbf{y}) \equiv \mu_{y_0} + (\boldsymbol{\gamma}^*)^T(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})$$

is a best linear predictor of y_0 based on $\mathbf{y}$.

Proof: Denote an arbitrary linear predictor as $t(\mathbf{y}) = \gamma_0 + \mathbf{y}^T\boldsymbol{\gamma}$. Then

$$\begin{aligned}\mathrm{E}[\{y_0 - t(\mathbf{y})\}^2] &= \mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y}) + m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}^2] \\ &= \mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y})\}^2] + \mathrm{E}[m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}^2] \\ &\quad + 2\mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y})\}\{m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}].\end{aligned}$$

Again, it suffices to show that $\mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y})\}\{m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}] = 0$, because if this cross-product term is 0 then

$$\mathrm{E}[\{y_0 - t(\mathbf{y})\}^2] = \mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y})\}^2] + \mathrm{E}[m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}^2].$$

To find $t(\mathbf{y})$ that minimizes the left hand side (the mse criterion), observe that both terms on the right hand side are nonnegative, the first term does not depend on $t(\mathbf{y})$, and the second term is minimized when it is zero, which happens when $t(\mathbf{y}) = m_{\mathrm{blp}}(\mathbf{y})$.

So, it remains to show that $\mathrm{E}[\{y_0 - m_{\mathrm{blp}}(\mathbf{y})\}\{m_{\mathrm{blp}}(\mathbf{y}) - t(\mathbf{y})\}] = 0$, which we leave as an exercise. ∎

- In general, the best linear predictor and best predictor differ. However, in the special case in which $(y_0, y_1, \ldots, y_n)^T$ is multivariate normal, the best linear predictor and best predictor coincide.

- It can also be shown that the BLP is essentially unique, so that it makes sense to speak of *the* BLP.

Now let's return to the LMM context. Suppose $\mathbf{y}$ satisfies the LMM (1) from p.69. Then

$$\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{V}_{\mathbf{yy}} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R} \quad \text{(assumed nonsingular)}$$

so that the BLP of y_0 is

$$\mu_{y_0} + \mathbf{v}_{y_0\mathbf{y}}(\mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- However, this predictor is typically not of much use, because $\boldsymbol{\beta}$ is unknown. In addition, μ_{y_0}, $\mathbf{D}$ and $\mathbf{R}$ may be unknown as well.

For now, suppose that $\mathbf{D}$ and $\mathbf{R}$ and $\mathbf{v}_{y_0\mathbf{y}}$ (which is often a function of $\mathbf{D}$ and $\mathbf{R}$) are known, but μ_{y_0} and $\boldsymbol{\beta}$ are not. Then, since the BLP is not available, a natural predictor to consider is

$$m_{\text{blup}}(\mathbf{y}) \equiv \hat{\mu}_{y_0} + \mathbf{v}_{y_0\mathbf{y}}(\mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\mu}_{y_0}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$ are BLUEs of μ_{y_o} and $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, respectively.

- It can be shown that $m_{\text{blup}}(\mathbf{y})$ is the **best linear unbiased predictor** of y_0 (see Christensen, Ch.12). That is, in the class of unbiased predictors that are linear in $\mathbf{y}$, $m_{\text{blup}}(\mathbf{y})$ has the minimum mse of prediction.

Unbiasedness of a Predictor: A predictor $t(\mathbf{y})$ of y_0 is said to be unbiased if

$$\mathrm{E}\{t(\mathbf{y})\} = \mathrm{E}(y_0).$$

In a LMM context, it is typically of interest to predict $\mathbf{c}^T\mathbf{b}$, a linear combinations of the vector of random effects, based upon $\mathbf{y}$, the observed data vector.

- That is, we now let $\mathbf{c}^T\mathbf{b}$ play the role of y_0 in our description of BLUP above.

Since $\mathrm{E}(\mathbf{c}^T\mathbf{b}) = \mathbf{c}^T\mathrm{E}(\mathbf{b}) = 0$, $\hat{\mu}_{y_0}$ in $m_{\mathrm{blup}}(\mathbf{y})$ becomes 0. In addition,

$$\begin{aligned}\mathbf{v}_{y_0\mathbf{y}} &= \mathrm{cov}(\mathbf{c}^T\mathbf{b}, \mathbf{y}) = \mathbf{c}^T\mathrm{cov}(\mathbf{b}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}) \\ &= \mathbf{c}^T\mathrm{cov}(\mathbf{b}, \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}) = \mathbf{c}^T\{\mathrm{cov}(\mathbf{b}, \mathbf{b})\mathbf{Z}^T + \underbrace{\mathrm{cov}(\mathbf{b}, \boldsymbol{\varepsilon})}_{=\mathbf{0}}\} \\ &= \mathbf{c}^T\mathrm{var}(\mathbf{b})\mathbf{Z}^T = \mathbf{c}^T\mathbf{D}\mathbf{Z}^T\end{aligned}$$

Therefore, the BLUP of $\mathbf{c}^T\mathbf{b}$ is given by

$$\mathbf{c}^T\mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

where $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V} = \mathrm{var}(\mathbf{y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}$.

If we are interested in the BLUP of a vector of such functions, $(\mathbf{c}_1^T\mathbf{b}, \ldots, \mathbf{c}_r^T\mathbf{b})^T = \mathbf{C}\mathbf{b}$ this result extends in the obvious way: The BLUP of $\mathbf{C}\mathbf{b}$ is given by

$$\mathbf{C}\mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

It is sometimes convenient to write this BLUP in the equivalent form

$$\mathrm{BLUP}(\mathbf{C}\mathbf{b}) = \mathbf{C}\mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{C}\mathbf{D}\mathbf{Z}^T\mathbf{P}\mathbf{y}, \qquad (\dagger)$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^T\mathbf{V}^{-1}.$$

The "Mixed Model Equations":

At this point we have seen that for $\mathbf{D}$ and $\mathbf{R}$ known (i.e., for $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}$ known), a BLUE of $\boldsymbol{\beta}$ and the BLUP of $\mathbf{b}$ are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}^T)^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \quad \hat{\mathbf{b}} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

respectively.

In the classical linear model, the BLUE of $\boldsymbol{\beta}$ is obtained as the solution of the normal equations. In the LMM there is an analogous set of equations that yield the BLUE and BLUP of $\boldsymbol{\beta}$ and $\mathbf{b}$. These equations are called the **mixed model equations** or, sometimes, Henderson's equations.

We now present the mixed model equations. We assume $\mathbf{R}$ and $\mathbf{D}$ are nonsingular, known matrices.

Recall the LMM:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b}+\boldsymbol{\varepsilon}, \quad \text{E}(\mathbf{b}) = \text{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{var}(\mathbf{b}) = \mathbf{D}, \text{var}(\boldsymbol{\varepsilon}) = \mathbf{R}, \text{cov}(\mathbf{b}, \boldsymbol{\varepsilon}) = \mathbf{0}.$$

If $\mathbf{b}$ was fixed instead of random, the normal equations (based on GLS) for the model would be

$$\begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{R}^{-1}(\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} \mathbf{R}^{-1}\mathbf{y}$$

which may be written equivalently as

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

Of course, in the mixed model, $\mathbf{b}$ is random, which leads to a slightly different set of equations, known as the mixed model equations:

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \qquad (*)$$

Theorem: If $(\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{b}}^T)^T$ is a solution to the mixed model equations, then $\mathbf{X}\hat{\boldsymbol{\beta}}$ is a BLUE of $\mathbf{X}\boldsymbol{\beta}$ and $\hat{\mathbf{b}}$ is a BLUP of $\mathbf{b}$.

Proof: Recall that the LMM is equivalent to the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\zeta}, \quad \mathrm{E}(\boldsymbol{\zeta}) = \mathbf{0}, \quad \mathrm{var}(\boldsymbol{\zeta}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R} \equiv \mathbf{V}.$$

Therefore, $\mathbf{X}\hat{\boldsymbol{\beta}}$ will be a BLUE of $\mathbf{X}\boldsymbol{\beta}$ if $\hat{\boldsymbol{\beta}}$ is a solution to $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$. It can be shown (see Theorem B.56 in Christensen, for example) that

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}.$$

If $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are solutions, then the second row of (*) gives

$$\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}\hat{\mathbf{b}} = \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y}$$

$$\Rightarrow \qquad \hat{\mathbf{b}} = \{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \qquad (**)$$

The first row of (*) is

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}.$$

Substituting the expression for $\hat{\mathbf{b}}$ gives

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y},$$

or

$$\begin{aligned}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \\ = \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y},\end{aligned}$$

or

$$\begin{aligned}\mathbf{X}^T \underbrace{(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1})}_{=\mathbf{V}^{-1}}\mathbf{X}\hat{\boldsymbol{\beta}} \\ = \mathbf{X}^T \underbrace{(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1})}_{=\mathbf{V}^{-1}}\mathbf{y}.\end{aligned}$$

Thus, $\hat{\boldsymbol{\beta}}$ is a GLS solutions so that $\mathbf{X}\hat{\boldsymbol{\beta}}$ is BLUE.

Now to show $\hat{\mathbf{b}}$ is a BLUP: $\hat{\mathbf{b}}$ in (**) can be rewritten as

$$\begin{aligned}
\hat{\mathbf{b}} &= (\mathbf{D}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\} - \mathbf{D}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}) \\
&\quad \times \{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{D}\mathbf{Z}^T\mathbf{R}^{-1} - \mathbf{D}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{D}\mathbf{Z}^T \underbrace{(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\{\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\}^{-1}\mathbf{Z}^T\mathbf{R}^{-1})}_{=\mathbf{V}^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),
\end{aligned}$$

which is the BLUP of $\mathbf{b}$ by result (†) on p.75 (here $\mathbf{I}$ is playing the role of $\mathbf{C}$ since we're interested in the BLUP of $\mathbf{I}\mathbf{b} = \mathbf{b}$). ∎

Sampling Variance of BLUE and BLUP for $\mathbf{V}$ known:

Just as it is useful for inference on $\boldsymbol{\beta}$ to known the variance of our estimator $\hat{\boldsymbol{\beta}}$, it is useful to known the prediction variance of the BLUP.

For $\mathbf{V}$ known and $\mathbf{C}\boldsymbol{\beta}$ a vector of estimable functions, the estimator $\mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ has variance-covariance matrix

$$\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{C}^T.$$

The analogous result for $\mathbf{C}\hat{\mathbf{b}}$ is as follows:

$$\begin{aligned}\text{var}(\mathbf{C}\hat{\mathbf{b}}) &= \text{var}(\mathbf{CDZ}^T\mathbf{Py}) = \mathbf{CDZ}^T\mathbf{PVP}^T\mathbf{ZD}^T\mathbf{C}^T \\ &= \mathbf{CDZ}^T\mathbf{PZDC}^T\end{aligned}$$

Here, we have used the fact that $\mathbf{PVP} = \mathbf{P}$ and that $\mathbf{P}$ and $\mathbf{D}$ are symmetric.

If we are interested in the variance in the prediction error $\mathbf{Cb} - \mathbf{C}\hat{\mathbf{b}}$, then we have

$$\begin{aligned}\text{var}(\mathbf{Cb} - \mathbf{C}\hat{\mathbf{b}}) &= \mathbf{C}\text{var}(\mathbf{b} - \hat{\mathbf{b}})\mathbf{C}^T \\ &= \mathbf{C}\{\text{var}(\mathbf{b}) + \text{var}(\hat{\mathbf{b}}) - \text{cov}(\mathbf{b}, \hat{\mathbf{b}}) - \text{cov}(\mathbf{b}, \hat{\mathbf{b}})^T\}\mathbf{C}^T\end{aligned}$$

where

$$\begin{aligned}\text{cov}(\mathbf{b}, \hat{\mathbf{b}}) &= \text{cov}(\mathbf{b}, \mathbf{DZ}^T\mathbf{Py}) = \text{cov}(\mathbf{b}, \mathbf{y})\mathbf{PZD} \\ &= \text{cov}(\mathbf{b}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon})\mathbf{PZD} = \text{cov}(\mathbf{b}, \mathbf{Zb})\mathbf{PZD} \\ &= \text{cov}(\mathbf{b}, \mathbf{b})\mathbf{Z}^T\mathbf{PZD} = \mathbf{DZ}^T\mathbf{PZD} = \text{var}(\hat{\mathbf{b}}).\end{aligned}$$

Therefore,

$$\begin{aligned}\text{var}(\mathbf{Cb} - \mathbf{C}\hat{\mathbf{b}}) &= \mathbf{C}\{\text{var}(\mathbf{b}) + \text{var}(\hat{\mathbf{b}}) - \text{var}(\hat{\mathbf{b}}) - \text{var}(\hat{\mathbf{b}})^T\}\mathbf{C}^T \\ &= \mathbf{C}\{\mathbf{D} - \mathbf{DZ}^T\mathbf{PZD}\}\mathbf{C}^T.\end{aligned}$$

In addition, note that for $\mathbf{C}$ a matrix of constants such that $\mathbf{C}\boldsymbol{\beta}$ is a vector of estimable functions, then

$$\text{cov}(\mathbf{C}\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) = \mathbf{0}.$$

Maximum Likelihood Estimation:

- We have already seen that for known $\mathbf{V}$, $\mathbf{C}\boldsymbol{\beta}$ has BLUE $\mathbf{C}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and $\mathbf{b}$ has BLUP $\mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

- These results do not depend upon any distributional assumption on $\mathbf{b}$ and $\boldsymbol{\varepsilon}$.

- In addition, to this point we have concentrated on the case when $\mathbf{V}$ is known. We now relax that assumption to consider the $\mathbf{V}$ unknown case.

- Note that $\mathbf{b}$ is not a parameter of the model, so while we may be interested in $\hat{\mathbf{b}}$, predicting $\mathbf{b}$ is not part of fitting the model.

- The unknown parameters of the LMM are $\boldsymbol{\beta}$, $\mathbf{D}$, and $\mathbf{R}$.

- Typically, some structure is placed on $\mathbf{D}$ and $\mathbf{R}$ so that their forms are known, and they are assumed to be matrix functions of a relatively small number of parameters.

Let $\boldsymbol{\theta}$ be the $q \times 1$ vector of unknown parameters describing $\mathbf{D}$ and $\mathbf{R}$, and hence $\mathbf{V}$.

- We will often write these matrices as $\mathbf{D}(\boldsymbol{\theta})$, $\mathbf{R}(\boldsymbol{\theta})$, $\mathbf{V}(\boldsymbol{\theta})$ to emphasize this dependence.

So, fitting the LMM involves estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. After the model has been fit it may also be of interest to predict $\mathbf{b}$.

A unified framework for estimation of these parameters is provided by maximum likelihood, which requires that make distributional assumptions on $\mathbf{b}$ and $\boldsymbol{\varepsilon}$.

- Such assumptions will be necessary anyway for inference, so there is not much cost in making them at the estimation phase.

Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$, where $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))$ and $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are independent. Then

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})), \quad \text{where} \quad \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R},$$

so the loglikelihood for $\boldsymbol{\beta}, \boldsymbol{\theta}$ is just the log of a multivariate normal density:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\{|\mathbf{V}(\boldsymbol{\theta})|\} - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The ML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be found by taking partial derivatives of $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y})$ with respect to $\boldsymbol{\beta}$ and the components of $\boldsymbol{\theta}$ and setting the resulting functions equal to zero, and solving.

To take these partial derivatives, we need some results on matrix and vector differentiation. The following four results appear in Christensen (*Plane Answers to Complex Questions*) as Proposition 12.4.1, but can also be found in McCulloch et al., 2008, Appendix M, and other standard references.

1. $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}$.

2. $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T\mathbf{A}$.

3. If $\mathbf{A}$ is a function of a scalar s,

$$\frac{\partial \mathbf{A}^{-1}}{\partial s} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial s}\mathbf{A}^{-1}.$$

4. If $\mathbf{A}$ is a function of a scalar s,

$$\frac{\partial \log|\mathbf{A}|}{\partial s} = \mathrm{tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial s}\right).$$

Back to our problem. Recall the loglikelihood:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\{|\mathbf{V}(\boldsymbol{\theta})|\} - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Using the matrix and vector differentiation results above, we obtain the following partial derivatives:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\boldsymbol{\beta}^T\mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{X} + \mathbf{y}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{X}, \quad \text{and}$$
$$\frac{\partial \ell}{\partial \theta_j} = -\frac{1}{2}\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

$j = 1, \ldots, q$.

Setting these partials equal to zero, we get the following set of estimating equations which can be solved to obtain the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} \mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{y} \\ \text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\right) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned} \qquad (\heartsuit)$$

$j = 1, \ldots, q$.

- Although these equations do not, in general, have simple closed-form solutions, they can be solved simultaneously by any one of several numerical techniques (e.g., Newton-Raphson, EM algorithm)

Instead of solving the equations (♡) simultaneously, an alternative method of maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y})$, which is often more convenient, is the method of profile (log)likelihood.

i. First, treat $\boldsymbol{\theta}$ as fixed and maximize the loglikelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y})$ with respect to $\boldsymbol{\beta}$.

ii. Second, plug the estimator of $\boldsymbol{\beta}$, call it $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ (a function of $\boldsymbol{\theta}$), back into the loglikelihood. This yields $p\ell(\boldsymbol{\theta}; \mathbf{y}) \equiv \ell(\boldsymbol{\beta}_{\boldsymbol{\theta}}, \boldsymbol{\theta}; \mathbf{y})$, which is a function of $\boldsymbol{\theta}$ only (called the profile loglikelihood for $\boldsymbol{\theta}$). Maximize $p\ell(\boldsymbol{\theta}; \mathbf{y})$ with respect to $\boldsymbol{\theta}$ to obtain the MLE $\hat{\boldsymbol{\theta}}$.

iii. Finally, the MLE of $\boldsymbol{\beta}$ is obtained by plugging $\hat{\boldsymbol{\theta}}$ into our estimator for $\boldsymbol{\beta}$ obtained in step 1. That is, the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}}$.

- Notice that for fixed $\boldsymbol{\theta}$, maximizing $\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{y})$ with respect to $\boldsymbol{\beta}$ is equivalent to minimizing

  $$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \{\mathbf{V}(\boldsymbol{\theta})\}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

  which is the GLS criterion. Therefore, step 1 gives

  $$\boldsymbol{\beta}_{\boldsymbol{\theta}} = [\mathbf{X}^T \{\mathbf{V}(\boldsymbol{\theta})\}^{-1} \mathbf{X}]^- \mathbf{X}^T \{\mathbf{V}(\boldsymbol{\theta})\}^{-1} \mathbf{y}.$$

The real work is done in step 2, where we obtain $\hat{\boldsymbol{\theta}}$ by maximizing

$$p\ell(\boldsymbol{\theta}; \mathbf{y}) = -\frac{1}{2} \left[\log\{|\mathbf{V}(\boldsymbol{\theta})|\} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})^T \{\mathbf{V}(\boldsymbol{\theta})\}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}}) \right].$$

Once this step is accomplished, it is clear that the MLE of $\boldsymbol{\beta}$ will then be

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^- \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y},$$

where $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$.

- For a presentation of efficient computational methods for maximizing $p\ell(\boldsymbol{\theta}; \mathbf{y})$, see Pinheiro and Bates (2000, §2.2) and McCulloch, Searle, and Neuhaus (2008, Ch. 14).

Variance Component Models:

While the parameterization of $\mathbf{V}$ through $\boldsymbol{\theta}$ can, in the general LMM, take on a wide variety of forms, in the important subclass of LMMs known as **variance component models**, $\mathbf{V}(\boldsymbol{\theta})$ has a specific simple form.

In variance component models, the levels of any particular random effect are assumed to be independent with the same variance. Different random effects are allowed different variances and are assumed independent. In addition, the errors are assumed to be homoscedastic so that $\mathbf{R} = \sigma^2\mathbf{I}_n$.

- The one-way random effects model, the RCB model, and the split-plot model are all examples of variance component models.

Another example is the model for an $s \times s$ Latin Square design in which both blocking factors (rows and columns in the Latin square) are thought of as random. The appropriate model for y_{ijk}, the response in the i^{th} treatment, j^{th} row, k^{th} column, would be

$$y_{ijk} = \mu + \alpha_i + r_j + c_k + \varepsilon_{ijk}, \quad \begin{array}{l} r_1, \ldots, r_s \overset{iid}{\sim} N(0, \sigma_r^2) \\ c_1, \ldots, c_s \overset{iid}{\sim} N(0, \sigma_c^2) \\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \end{array}$$

In variance component models, $\mathbf{Z}$ can be partitioned into $q-1$ submatrices ($q = 2$ in the one-way, RCB, and split-plot models, $q = 3$ in the LSD) as $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_{q-1})$ and $\mathbf{b}$ can be partitioned accordingly as $\mathbf{b} = (\mathbf{b}_1^T, \mathbf{b}_2^T, \ldots, \mathbf{b}_{q-1}^T)^T$.

Let $m(i)$ denote the number of columns in $\mathbf{Z}_i$ (= number of element in $\mathbf{b}_i$). We assume $\text{var}(\mathbf{b}_i) = \sigma_i^2 \mathbf{I}_{m(i)}$, $i = 1, \ldots, q-1$, and $\text{cov}(\mathbf{b}_i, \mathbf{b}_j) = \mathbf{0}$, for $i \neq j$.

Then $\mathbf{D}$ takes on a block-diagonal structure as follows:

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{m(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I}_{m(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_{q-1}^2 \mathbf{I}_{m(q-1)} \end{pmatrix}$$

We also assume $\mathbf{R} = \sigma_q^2 \mathbf{I}_n$. Putting these assumptions together, the matrix $\mathbf{V}$ is assumed to be of the form

$$\mathbf{V} = \sum_{i=1}^{q-1} \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T + \sigma_q^2 \mathbf{I}_n = \sum_{i=1}^{q} \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T, \qquad (\diamondsuit)$$

where $\mathbf{Z}_q \equiv \mathbf{I}_n$.

- E.g., in the LSD, suppose that the number of treatments = number of rows = number of columns = 3. Suppose the design and data are as follows:

A	B	C
C	A	B
B	C	A

,

y_{111}	y_{212}	y_{313}
y_{321}	y_{122}	y_{223}
y_{231}	y_{332}	y_{133}

Then the model can be written as

$$\begin{pmatrix} y_{111} \\ y_{212} \\ y_{313} \\ y_{321} \\ y_{122} \\ y_{223} \\ y_{231} \\ y_{332} \\ y_{133} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

$$+ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{212} \\ \varepsilon_{313} \\ \varepsilon_{321} \\ \varepsilon_{122} \\ \varepsilon_{223} \\ \varepsilon_{231} \\ \varepsilon_{332} \\ \varepsilon_{133} \end{pmatrix}$$

or $\quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \boldsymbol{\varepsilon}$

Here, $\mathbf{R} = \sigma_q^2\mathbf{Z}_q$, where $\mathbf{Z}_q = \mathbf{I}_9$, and $\mathbf{D}$ has the form

$$\mathbf{D} = \begin{pmatrix} \sigma_r^2\mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \sigma_c^2\mathbf{I}_3 \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{V} &= \operatorname{var}(\mathbf{y}) = \operatorname{var}(\mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \boldsymbol{\varepsilon}) \\ &= \mathbf{Z}_1\sigma_r^2\mathbf{I}_3\mathbf{Z}_1^T + \mathbf{Z}_2\sigma_c^2\mathbf{I}_3\mathbf{Z}_2^T + \mathbf{Z}_q\sigma_q^2\mathbf{I}_9\mathbf{Z}_q^T \\ &= \sigma_r^2\mathbf{Z}_1\mathbf{Z}_1^T + \sigma_c^2\mathbf{Z}_2\mathbf{Z}_2^T + \sigma_q^2\mathbf{Z}_q\mathbf{Z}_q^T. \end{aligned}$$

In variance component models, $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2)^T$ and $\mathbf{V}(\boldsymbol{\theta}) = \sum_{j=1}^q \theta_j\mathbf{Z}_j\mathbf{Z}_j^T$ (cf. ($\diamondsuit$)).

This leads to some simplification in the likelihood equations ($\heartsuit$). In particular,

$$\frac{\partial\mathbf{V}}{\partial\theta_j} = \mathbf{Z}_j\mathbf{Z}_j^T.$$

Information Matrix:

Under the general theory of maximum likelihood estimation, ML estimators are consistent and asymptotically normal, under suitable regularity conditions.

- For a discussion of the regularity conditions, see, for example, Cox and Hinkley (1974, *Theoretical Statistics*, Ch.9), or Seber and Wild (1989, *Nonlinear Regression*, Ch.12).

In particular, the asymptotic variance-covariance matrix of a MLE $\hat{\boldsymbol{\phi}}$, defined as the maximizer of a loglikelihood function $\ell(\boldsymbol{\phi})$, is $\mathbf{I}(\boldsymbol{\phi})^{-1}$, where

$$\mathbf{I}(\boldsymbol{\phi}) = -\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right),$$

is known as the **Fisher information matrix.**

- In practice, we replace $\boldsymbol{\phi}$ by $\hat{\boldsymbol{\phi}}$ and use $\mathbf{I}(\hat{\boldsymbol{\phi}})$.

- Alternatively, the **observed information matrix** or negative Hessian matrix

$$-\left(\frac{\partial^2 \ell(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}\right)\Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$$

  can be used in place of $\mathbf{I}(\hat{\boldsymbol{\phi}})$ without changing the asymptotics.

In the LMM context, $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ and the loglikelihood is given on the top of p.73. The information matrix is given by

$$-\mathrm{E}\begin{pmatrix} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}^T} \\ \left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}^T}\right)^T & \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\left\{\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_k}\right)\right\} \end{pmatrix},$$

where $\left\{\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_k}\right)\right\}$ denotes the $q \times q$ matrix with the quantity inside the curly braces as its j, k^{th} element (see §6.11.a.iii of McCulloch, Searle, and Neuhaus, 2008, for the details).

Inverting this matrix leads to the following asymptotic variance-covariance matrices for the MLE $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T$:

$$\begin{aligned} \mathrm{avar}(\mathbf{X}\hat{\boldsymbol{\beta}}) &= \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^T \\ \mathrm{avar}(\hat{\boldsymbol{\theta}}) &= 2\left[\left\{\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_k}\right)\right\}\right]^{-1} \\ \mathrm{acov}(\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) &= \mathbf{0}. \end{aligned}$$

- Inference for ML in the LMM can now be based on standard asymptotic likelihood-based methods — Wald, score and LR tests and CIs; and model selection criteria such as AIC, BIC — all can be formed in the usual way.

- E.g., the Wald test statistic for the hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{C}\boldsymbol{\beta}$ is a vector of estimable functions (i.e., equals $\mathbf{AX}$ for some $\mathbf{A}$), and where $\mathrm{rank}(\mathbf{C}) = \mathrm{nrows}(\mathbf{C}) \leq \mathrm{rank}(\mathbf{X})$, is given by

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T \left[\mathbf{C}\left\{\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right\}^{-}\mathbf{C}^T\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \overset{a}{\sim} \chi^2(\mathrm{nrows}(\mathbf{C})).$$

- However, we will see that we can improve upon these asymptotic results with approximate F and t-based inference that work in small and large samples.

Restricted Maximum Likelihood Estimation:

Recall that in simple problems, ML estimation of variances produces biased estimators.

- E.g., in a one-sample problem from a $N(\mu, \sigma^2)$, the MLE of σ^2 is $\frac{1}{n}\sum_i^n (x_i - \bar{\mathbf{x}})^2$ rather than the unbiased estimator $s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{\mathbf{x}})^2$.

- Another example: in the CLM the MLE of the error variance is $\frac{1}{n}\mathrm{SS_E}$ rather than the unbiased estimator $S^2 = \frac{1}{n-\mathrm{rank}(\mathbf{X})}\mathrm{SS_E}$.

In estimating the variance, these ML estimators ignore the fact that parameters in the mean have been estimated.

- In the one-sample problem, one degree of freedom is used up in estimating μ with $\bar{\mathbf{x}}$, so the appropriate divisor is $n-1$ (number of observations minus number of "non-redundant" parameters estimated) rather than n.

- In the CLM, we use up $\mathrm{rank}(\mathbf{X})$ degrees of freedom in estimating $\boldsymbol{\beta}$. Therefore, the appropriate divisor is $n - \mathrm{rank}(\mathbf{X})$ rather than n.

We'd prefer to have a general "likelihood-based" method of estimation that produces estimators of variances that account for the degrees of freedom lost (information used up) in estimating parameters of the mean. Such a method is **restricted maximum likelihood (REML)** estimation (sometimes called residual maximum likelihood or marginal maximum likelihood).

- Note that the goal here is to improve upon ML estimators of $\boldsymbol{\theta}$, not $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. That is, REML is a method of estimating the variance-covariance parameters, not a method of estimating all of the parameters of the model.

  - However, given a REML estimator of $\boldsymbol{\theta}$ it is obvious how the ML estimator of $\boldsymbol{\beta}$ should be formed.

In REML, parameter estimates of $\boldsymbol{\theta}$ are obtained by maximizing that part of the likelihood which is invariant to $\mathbf{X}\boldsymbol{\beta}$.

- That is, we eliminate $\boldsymbol{\beta}$ from the log-likelihood by considering the loglikelihood of a set of linear combinations of $\mathbf{y}$, known as **error contrasts**, whose distribution does not depend upon $\boldsymbol{\beta}$, rather than the density of $\mathbf{y}$ itself.

Error Contrasts: A linear combination $\mathbf{k}^T\mathbf{y}$ is said to be an error contrast if $\mathrm{E}(\mathbf{k}^T\mathbf{y}) = 0$ for all $\boldsymbol{\beta}$.

- It follows that $\mathbf{k}^T\mathbf{y}$ is an error contrast if and only if $\mathbf{X}^T\mathbf{k} = \mathbf{0}$.

Let $C(\mathbf{X})$ denote the column space of $\mathbf{X}$ and $C(\mathbf{X})^\perp$ its orthogonal complement. Let $\mathbf{P}_{C(\mathbf{X})^\perp} = \mathbf{I} - \mathbf{P}_{C(\mathbf{X})} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ and suppose $\mathrm{rank}(\mathbf{X}) = s$. Then each element of the vector $\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}$ is an error contrast because

$$\mathrm{E}(\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}) = (\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathrm{E}(\mathbf{y}) = (\mathbf{I} - \mathbf{P}_{C(\mathbf{X})})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

Note, however, that $\mathrm{rank}(\mathbf{P}_{C(\mathbf{X})^\perp}) = n - s$ while the dimension of $\mathbf{P}_{C(\mathbf{X})^\perp}$ is $n \times n$.

- Therefore, there are some redundancies among the elements of $\mathbf{P}_{C(\mathbf{X})^\perp}\mathbf{y}$.

A natural question arises:

How many essentially different (non-redundant) error contrasts can be included in a single set?

Linearly Independent Error Contrasts: Error contrasts $\mathbf{k}_1^T\mathbf{y}, \mathbf{k}_2^T\mathbf{y}, \ldots, \mathbf{k}_m^T\mathbf{y}$ are said to be linearly independent if $\mathbf{k}_1, \ldots, \mathbf{k}_m$ are linearly independent vectors.

Theorem: Any set of error contrasts contains at most $n - \text{rank}(\mathbf{X}) = n - s$ linearly independent error contrasts.

Theorem: Let $\mathbf{K}$ be a $n \times (n-s)$ matrix such that $\mathbf{K}^T\mathbf{K} = \mathbf{I}$ and $\mathbf{K}\mathbf{K}^T = \mathbf{P}_{C(\mathbf{X})^\perp}$. The $(n-s) \times 1$ vector $\mathbf{w}$ defined by

$$\mathbf{w} = \mathbf{K}^T\mathbf{y}$$

is a vector of $n-s$ linearly independent error contrasts. (It is not the only vector, however.)

The REML approach consists of applying ML estimation to $\mathbf{w}$ rather than $\mathbf{y}$. If $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$ where $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{D}(\boldsymbol{\theta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\theta})$, then

$$\mathbf{w} \sim N_{n-s}(\mathbf{0}, \mathbf{K}^T\mathbf{V}(\boldsymbol{\theta})\mathbf{K}).$$

Therefore, the restricted loglikelihood for $\boldsymbol{\theta}$ is just the log density of $\mathbf{w}$:

$$\ell_R(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n-s}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}^T\mathbf{V}(\boldsymbol{\theta})\mathbf{K}| - \frac{1}{2}\mathbf{w}^T(\mathbf{K}^T\mathbf{V}(\boldsymbol{\theta})\mathbf{K})^{-1}\mathbf{w}.$$

$\hat{\boldsymbol{\theta}}$ is a REML estimate of $\boldsymbol{\theta}$ if $\ell_R(\boldsymbol{\theta}; \mathbf{y})$ attains its maximum value at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

- It can be shown that the maximizer of the restricted loglikelihood does not depend upon which vector of $n-s$ linearly independent error contrasts is used to form $\mathbf{w}$.
  - That is, the REML estimator is well-defined.

It is preferrable to express $\ell_R(\boldsymbol{\theta}; \mathbf{y})$ in terms of $\mathbf{X}$ and $\mathbf{V}$ (quantities that define our model) rather than in terms of $\mathbf{K}$ and $\mathbf{V}$. Hence, the following result:

Theorem: The log-likelihood function associated with any vector of $n-s$ linearly independent error contrasts is, apart from an additive constant that doesn't depend on $\boldsymbol{\theta}$,

$$\begin{aligned}\ell_R(\boldsymbol{\theta}; \mathbf{y}) = &-\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}}) \\ &- \frac{1}{2}\log|\tilde{\mathbf{X}}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\tilde{\mathbf{X}}|,\end{aligned} \qquad (\spadesuit)$$

where $\tilde{\mathbf{X}}$ represents any $n \times s$ matrix such that $C(\tilde{\mathbf{X}}) = C(\mathbf{X})$ and $\boldsymbol{\beta}_{\boldsymbol{\theta}}$ is any solution to $\mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{y}$.

- Note that in ordinary ML, we obtain the profile likelihood for $\boldsymbol{\theta}$ as

$$p\ell(\boldsymbol{\theta}; \mathbf{y}) = -\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}}).$$

  Notice that $\ell_R(\theta; \mathbf{y})$ differs only from $p\ell(\boldsymbol{\theta}; \mathbf{y})$ by the additional term $-\frac{1}{2}\log|\tilde{\mathbf{X}}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\tilde{\mathbf{X}}|$. This term serves as an adjustment, or penalty, for the estimation of $\boldsymbol{\beta}$. Hence REML estimation is sometimes called a *penalized likelihood* method.

To obtain $\hat{\boldsymbol{\theta}}$, the REML estimate of $\boldsymbol{\theta}$, we solve the estimating equations

$$\frac{\partial \ell_R(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}_i} = 0, \quad i = 1, \ldots, q$$

In general LMMs, these estimating equations can be written

$$\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\left(\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right)\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})-\frac{1}{2}\text{tr}\left\{\mathbf{Q}\left(\frac{\partial\mathbf{V}(\boldsymbol{\theta})}{\partial\theta_i}\right)\right\}=0,$$

$i=1,\ldots,q$, where

$$\mathbf{Q}=\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\left[\mathbf{I}-\mathbf{X}(\mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\mathbf{X})^{-}\mathbf{X}^T\{\mathbf{V}(\boldsymbol{\theta})\}^{-1}\right].$$

In variance component models where $\boldsymbol{\theta}=(\sigma_1^2,\sigma_2^2,\ldots,\sigma_q^2)^T$, these estimating equations simplify based on $\partial\mathbf{V}(\boldsymbol{\theta})/(\partial\sigma_i^2)=\mathbf{Z}_i\mathbf{Z}_i^T$, $i=1,\ldots,q$.

- REML estimators are not, in general, unbiased. However, they typically have less bias than ML estimators of variance components.

- While it is not possible to make completely general recommendations concerning REML vs. ML estimation, it does appear that REML estimators perform better than ML estimators for s large relative to n. I would recommend REML over ML for $s>4$ or so.

- As previously mentioned, REML provides an estimator of $\boldsymbol{\theta}$, it says nothing about the estimation of $\boldsymbol{\beta}$. However, ML says to estimate $\boldsymbol{\theta}$ as

  $$\hat{\boldsymbol{\theta}}_{\text{ML}}=\arg\max_{\boldsymbol{\theta}}p\ell(\boldsymbol{\theta};\mathbf{y})$$

  and then estimate $\boldsymbol{\beta}$ as

  $$\hat{\boldsymbol{\beta}}_{\text{ML}}=\boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_{\text{ML}}}=\left[\mathbf{X}^T\left\{\mathbf{V}(\hat{\boldsymbol{\theta}}_{\text{ML}})\right\}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^T\left\{\mathbf{V}(\hat{\boldsymbol{\theta}}_{\text{ML}})\right\}^{-1}\mathbf{y}.$$

  In REML, we estimate $\boldsymbol{\theta}$ by maximizing $\ell_R(\boldsymbol{\theta};\mathbf{y})$ instead of $p\ell(\boldsymbol{\theta};\mathbf{y})$. Therefore, the obvious "REML estimator of $\boldsymbol{\beta}$" is

  $$\hat{\boldsymbol{\beta}}_{\text{REML}}=\boldsymbol{\beta}_{\hat{\boldsymbol{\theta}}_{\text{REML}}}=\left[\mathbf{X}^T\left\{\mathbf{V}(\hat{\boldsymbol{\theta}}_{\text{REML}})\right\}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^T\left\{\mathbf{V}(\hat{\boldsymbol{\theta}}_{\text{REML}})\right\}^{-1}\mathbf{y},$$

  where

  $$\hat{\boldsymbol{\theta}}_{\text{REML}}=\arg\max_{\boldsymbol{\theta}}\ell_R(\boldsymbol{\theta};\mathbf{y}).$$

The asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\beta}}^T_{\text{REML}}, \hat{\boldsymbol{\theta}}^T_{\text{REML}})^T$ is given by

$$\begin{pmatrix} (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^- & \mathbf{0} \\ \mathbf{0} & 2\left[\left\{\text{tr}\left(\mathbf{P}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_k}\right)\right\}\right]^{-1} \end{pmatrix},$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}^T\mathbf{V}^{-1} = \mathbf{K}(\mathbf{K}^T\mathbf{V}\mathbf{K})^{-1}\mathbf{K}^T.$$

- As in ML estimation, this asymptotic var-cov matrix can be estimated by evaluating it at the REML estimates.

- Wald-based inference can then be done based on the estimated asymptotic var-cov matrix.

- Note however, that the restricted loglikelihood cannot be treated as an ordinary loglikelihood. In particular, LRTs, AICs, and BICs based on the restricted loglikelihood objective function should not be used to select between models with different fixed-effects specifications.

- The restricted loglikelihood given by (♠) can be derived in a number of different ways. Harville (1974) and Laird and Ware (1982) use a Bayesian approach, while Patterson and Thompson (1971) use a more traditional fequentist approach.

- It can also be derived as a modified profile likelihood function (see Pawitan, §10.6 and Ch.17). See also McCullagh and Nelder, §7.2, for connections to marginal and conditional likelihood.

Small-Sample Inference on the Fixed Effects:

As mentioned previously, ML/REML estimation provides a unified framework for estimation and inference in the LMM, and standard likelihood-based inference techniques for fixed effects are available (Wald tests, LR tests, etc.).

In addition, for many special cases of the LMM, such as anova models, exact (small and large sample) inference techniques are available **for balanced data.**

- E.g., in the one-way random effects model, or the RCB model, or the split-plot model, it is possible to obtain exact F tests to test treatment effects, do inference on treatment means, etc.

However, for unbalanced data, exact distributional results are not available, and in small samples, asymptotic variances can seriously underestimate the true variances of parameter estimators, compromising the validity of the asymptotic inferences.

- Therefore, large-sample techniques (e.g., conventional Wald and LR tests) are **not** recommended for inference on the fixed effects in LMMs, except in very large samples.

We now consider small sample inference methods which attempt to compensate for the underestimation of the sampling variance of the REML estimator of $\boldsymbol{\beta}$ in the LMM. The following presentation is based on Kenward and Roger (1997).

Recall that the REML estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \Phi(\hat{\boldsymbol{\theta}})\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{y},$$

where

$$\Phi(\boldsymbol{\theta}) = \{\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1},$$

and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\text{REML}}$ is the REML estimator of $\boldsymbol{\theta}$.

- For simplicity, we will assume that $\mathbf{X}$ is of full rank, the results presented here extend to the non-full rank case as well.

Recall that the matrix $\Phi(\boldsymbol{\theta})$ is the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, and conventionally, its estimator

$$\hat{\Phi} \equiv \Phi(\hat{\boldsymbol{\theta}}) = \{\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X}\}^{-1}$$

is used to quantify the precision of $\hat{\boldsymbol{\beta}}$.

There are two sources of bias in $\hat{\boldsymbol{\phi}}$ as a measure of the precision of $\hat{\boldsymbol{\beta}}$ in small samples:

1. $\Phi(\boldsymbol{\theta})$ takes no account of the error introduced into $\hat{\boldsymbol{\beta}}$ by having to estimate $\boldsymbol{\theta}$, so it is an underestimate of $\text{var}(\hat{\boldsymbol{\beta}})$.

    – Another way to say this is as follows: When $\boldsymbol{\theta}$ is known, we know that $\text{var}\left[\{\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{y}\right] = \Phi(\boldsymbol{\theta})$. Undoubtedly, plugging in $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$ introduces some extra variability (error), so $\text{var}(\hat{\boldsymbol{\beta}})$ must be greater than $\Phi(\boldsymbol{\theta})$.

2. $\hat{\Phi} = \Phi(\hat{\boldsymbol{\theta}})$ is a biased estimator of $\Phi(\boldsymbol{\theta})$.

To correct these deficiencies, we write the variability in $\hat{\boldsymbol{\beta}}$ as the sum of two components:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \Phi + \Lambda,$$

where Λ represents the amount by which $\Phi = \text{avar}(\hat{\boldsymbol{\beta}})$ underestimates $\text{var}(\hat{\boldsymbol{\beta}})$.

Using Taylor series expansions, it can be shown that Λ can be approximated by

$$\Lambda \approx \Phi \left\{ \sum_{i=1}^{q} \sum_{j=1}^{q} W_{ij}(\mathbf{S}_{ij} - \mathbf{T}_i \Phi \mathbf{T}_j) \right\} \Phi,$$

where

$$\mathbf{T}_i = \mathbf{X}^T \frac{\partial \mathbf{V}^{-1}}{\partial \theta_i} \mathbf{X}, \quad \mathbf{S}_{ij} = \mathbf{X}^T \frac{\partial \mathbf{V}^{-1}}{\partial \theta_i} \mathbf{V} \frac{\partial \mathbf{V}^{-1}}{\partial \theta_i} \mathbf{X},$$

and W_{ij} is the $(i,j)^{\text{th}}$ element of $\mathbf{W} = \text{var}(\hat{\boldsymbol{\theta}})$.

In addition, a Taylor series expansion about $\boldsymbol{\theta}$ can be used to show that $\hat{\Phi}$ is biased as follows:

$$\text{E}(\hat{\Phi}) \approx \Phi - \Lambda + \underbrace{\frac{1}{2} \sum_{i}^{q} \sum_{j}^{q} W_{ij} \Phi \mathbf{R}_{ij} \Phi}_{=(*)},$$

where

$$R_{ij} = \mathbf{X}^T \mathbf{V}^{-1} \frac{\partial^2 \mathbf{V}}{\partial \theta_i \partial \theta_j} \mathbf{V}^{-1} \mathbf{X}.$$

Since we want to estimate $\Phi + \Lambda$, Kenward and Roger suggest an adjusted small sample var-cov matrix of $\hat{\boldsymbol{\beta}}$ given by

$$\hat{\Phi}_{\text{adj}} = \hat{\Phi} + 2\hat{\Lambda} - \frac{1}{2} \sum_{i}^{q} \sum_{j}^{q} \hat{W}_{ij} \hat{\Phi} \hat{\mathbf{R}}_{ij} \hat{\Phi},$$

where $\hat{W}_{ij}$ is the $(i,j)^{\text{th}}$ element of $\hat{\text{avar}}(\hat{\boldsymbol{\theta}})$ (top of p.85), and $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is substituted for $\mathbf{V}(\boldsymbol{\theta})$ to form $\hat{\mathbf{R}}$ and $\hat{\Lambda}$.

- Note that in variance component models, the term (*) equals 0, so the adjusted estimator of $\text{var}(\hat{\boldsymbol{\beta}})$ simplifies to $\hat{\Phi}_{\text{adj}} = \hat{\Phi} + 2\hat{\Lambda}$.

Inference and Degrees of Freedom:

For a testable hypothesis of the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{C}$ is $c \times p$ of full row rank, a reasonable test statistic for H_0 is given by

$$\frac{1}{c}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T[\hat{\text{var}}(\mathbf{C}\hat{\boldsymbol{\beta}})]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}). \qquad (**)$$

When $\mathbf{V}$ is known, we have

$$\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{C}^T,$$

or if $\mathbf{V}$ is known up to a multiplicate constant, i.e., if $\mathbf{V} = \sigma^2\mathbf{W}$ for $\mathbf{W}$ known, then

$$\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2}\mathbf{C}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^-\mathbf{C}^T,$$

which is estimated by

$$\hat{\text{var}}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \frac{1}{S^2}\mathbf{C}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^-\mathbf{C}^T,$$

where

$$S^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - \text{rank}(\mathbf{X})}$$

is the MS_E from the fitted model. In that case, (**) becomes

$$F \equiv \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T[\mathbf{C}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^-\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})}{cS^2}$$

and, under H_0, we have

$$F \sim F(c, n - \text{rank}(\mathbf{X})).$$

However, in the LMM when $\mathbf{V}$ is unknown, the estimation of $\text{var}(\mathbf{C}\hat{\boldsymbol{\beta}})$ becomes more challenging, and no longer leads necessarily to an exact F statistic.

That is, when $\mathbf{V}$ is unkown, it still makes sense to use (**) as a test statistic, but now we use the Kenward-Roger estimate $\hat{\Phi}_{\text{adj}}$ to obtain $\hat{\text{var}}(\mathbf{C}\hat{\boldsymbol{\beta}})$. This leads to the test statistic

$$\hat{F} \equiv \frac{1}{c}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^T[\mathbf{C}\hat{\Phi}_{\text{adj}}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

This quantity no longer has an exact F distribution, but similar to the Satterthwaite procedure, we can approximate the distribution of $\hat{F}$ by assuming that

$$\lambda\hat{F} \dot{\sim} F(c, d), \qquad (\dagger)$$

for some λ and d.

A λ and d that make this a good approximation can be found by equating the first and second moments of both sides of (†). This approach leads to

$$\lambda = \frac{d}{(d-2)\text{E}(\hat{F})}, \quad \text{and}$$
$$d = \frac{1 + 2/c}{\frac{\text{var}(\hat{F})}{2\{\text{E}(\hat{F})\}^2} - \frac{1}{c}}.$$

Formulas for $\text{E}(\hat{F})$ and $\text{var}(\hat{F})$ are given in Kenward and Roger(1997).

- This procedure for estimating λ and d gives approximate F statistics that will typically perform much better than asymptotic inference techniques.

- These F statistics are approximate, in general, but do reduce to the exact F statistics in those cases in which exact results are available (e.g., balanced anova models).

- The Kenward-Roger approach to inference is implemented in PROC MIXED with the DDFM=KENWARDROGER option on the MODEL statment.

- Alternatively, the DDFM=SATTERTH option implements a closely related Satterthwaite approximation to obtain denominator degrees of freedom for approximate F tests. In the Satterthwaite procedure, the unadjusted estimator $\hat{\Phi}$ is used instead of Kenward and Roger's $\hat{\Phi}_{\text{adj}}$ to form the F statistic. That is, $\hat{\text{var}}(\mathbf{C}\hat{\boldsymbol{\beta}}) = \mathbf{C}\{\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X}\}^{-}\mathbf{C}^T$ is used in the Satterthwaite procedure.

- There is not yet consensus on which approach to small sample inference is, in general, preferrable in LMMs. However, a recent simulation study (Schaalje *et al.,* 2002) suggests that the K-R procedure may be superior, and it is what I recommend.

Inference on $\boldsymbol{\theta}$:

Often, the mean structure is of more interest when fitting a LMM than the variance-covariance structure. However, adequate modeling of the var-cov structure is important for model-based inference on the mean (on $\boldsymbol{\beta}$).

- Overparameterization of the var-cov structure leads to inefficient estimation of $\boldsymbol{\beta}$, low power, and potentially poor estimation of standard errors for $\hat{\boldsymbol{\beta}}$.

- On the other hand, an over-simplified var-cov structure can invalidate inferences on $\boldsymbol{\beta}$.

In addition, the var-cov structure is sometimes of interest in and of itself, and correct modeling of the var-cov structure can be important when handling certain types of missing data.

The major obstacle to inference on $\boldsymbol{\theta}$ is that $\boldsymbol{\theta}$ parameterizes the variance-covariance matrices $\mathbf{D}$ and $\mathbf{R}$, which must be positive definite.

Therefore, the parameter space for $\boldsymbol{\theta}$ is constrained, which can strongly affect the distributional results typically used for inference.

- E.g., in variance component models, the elements of $\boldsymbol{\theta}$ have the interpretation as variances, so they are necessarily non-negative. This strongly affects the adequacy of distributional approximations for $\hat{\boldsymbol{\theta}}$.

Wald Tests:

Based on classical likelihood theory,

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \overset{a}{\sim} N\left(\boldsymbol{\theta}, \text{avar}(\hat{\boldsymbol{\theta}}_{\text{ML}})\right), \quad \text{and} \quad \hat{\boldsymbol{\theta}}_{\text{REML}} \overset{a}{\sim} N\left(\boldsymbol{\theta}, \text{avar}(\hat{\boldsymbol{\theta}}_{\text{REML}})\right),$$

where the asymptotic var-cov matrices are given at the top of pp. 79 and 85, for ML and REML, respectively.

In principle therefore, Wald based inference for a linear combination $\eta = \mathbf{c}^T\boldsymbol{\theta}$, could be based on the distributional result:

$$\frac{\mathbf{c}^T\hat{\boldsymbol{\theta}} - \eta}{\sqrt{\mathbf{c}^T \text{avar}(\hat{\boldsymbol{\theta}})\mathbf{c}}} \dot{\sim} N(0,1). \qquad (\dagger)$$

However, the adequacy of this normal approximation depends strongly on how close η is to the boundary of its parameter space.

E.g., suppose $\eta = \theta_j$ and θ_j is a variance component (or a diagonal element of $\mathbf{D}$). Then if θ_j is close to zero, (†) becomes a poor approximation. In fact, (†) breaks down altogether if $\theta_j = 0$, or, more generally, if η is on the boundary of its parameter space.

- So, if η is far from the boundary of its parameter space, (†) can be used to form Wald confidence intervals and hypothesis tests.
  - Here, how far η must be from its boundary depends upon the sample size.
- However, (†) is useless for testing whether variance components are equal to zero (i.e., for testing whether a certain random effect is necessary in the model).

Likelihood Ratio Tests:

Similar comments apply to LR tests.

Let Θ denote the parameter space for $\boldsymbol{\theta}$. Then, according to classical likelihood theory, a hypothesis of the form

$$H_0 : \boldsymbol{\theta} \in \Theta_0,$$

where Θ_0 is a subspace of Θ, can be tested with the LR statistic

$$2\{\ell(\hat{\boldsymbol{\theta}}_{\text{ML}}) - \ell(\hat{\boldsymbol{\theta}}^0_{\text{ML}}\} \dot{\sim} \chi^2(\dim(\Theta) - \dim(\Theta_0)),$$

where $\hat{\boldsymbol{\theta}}_{\text{ML}}$ and $\hat{\boldsymbol{\theta}}^0_{\text{ML}}$ are the MLEs in Θ and Θ_0, respectively.

- Similarly, if the null hypothesis does not involve $\boldsymbol{\beta}$, then a restricted LR test can be done based on

$$2\{\ell_R(\hat{\boldsymbol{\theta}}_{\text{REML}}) - \ell_R(\hat{\boldsymbol{\theta}}^0_{\text{REML}}\} \dot{\sim} \chi^2(\dim(\Theta) - \dim(\Theta_0)).$$

However, the regularity conditions that establish these results (Wilks' Theorem) assume that $\boldsymbol{\theta}$ is not on the boundary of its parameter space. So, once again, standard asymptotic theory does not apply to LR testing for the significance of variance components.

In selecting an appropriate variance-covariance specification for our model, such hypotheses often arise.

- E.g., testing whether a certain random effect is necessary in the model is equivalent to testing whether its associate variance component is zero.

Therefore, how can we go about building the var-cov structure of our model?

One possible answer is to use the same Wald and LR test statistics, but use their proper reference distributions under the null hypothesis.

- That is, when the hypothesis places the parameter on the boundary of its parameter space, that means the usual normal and chi-square limiting distributions are no longer correct, not that the test statistics are no loinger appropriate. So, use the same statistics but use the right reference distribution.

Unfortunately, figuring out what the right reference distribution is can be hard, and sometime even if the distribution is known, obtaining critical values or $p-$values from that distribution can be hard.

- The only truly simple case is when the error variance covariance matrix is of the form $\mathbf{R} = \sigma^2\mathbf{I}$ and we are testing a null model that includes i.i.d. random effects that are each q-variate versus an alternative model with i.i.d. random effects that are each $q+1$ variate, where under both the null and alternative each random effect vector has a general, unstructured variance-covariance matrix.
  - In this case, the null distribution of the LRT and restricted LRT is a 50:50 mixture of a $\chi^2(q)$ and a $\chi^2(q+1)$ distribution.

- E.g., suppose we are choosing between a model with no random effects, versus a model with a cluster specific random effect, b_i, where

$$b_1, ..., b_n \overset{iid}{\sim} N(0, \sigma_b^2).$$

Then to test $H_0 : \sigma_b^2 = 0$, the LR and restricted LR tests both have null distribution that is a 50:50 mixture of a $\chi^2(0)$ and a $\chi^2(1)$ distribution, where $\chi^2(0)$ denotes the chi-square distribution with 0 d.f., which is 0 with probability 1.

  - In this case, the correct p-value for H_0 is exactly one-half the value based on a $\chi^2(1)$ distribution.

- As a second example, suppose we are choosing between a model with a cluster specific random effect, b_i, where

$$b_1, ..., b_n \overset{iid}{\sim} N(0, \sigma_b^2)$$

versus a model with a bivariate cluster-specific random effect, $\mathbf{b}_i$, where

$$\mathbf{b}_1, ..., \mathbf{b}_n \overset{iid}{\sim} N(\mathbf{0}, \mathbf{D}), \quad \text{where} \quad \mathbf{D} = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{12} & \theta_{22} \end{pmatrix}.$$

Then the LR and restricted LR tests both have null distribution that is a 50:50 mixture of a $\chi^2(1)$ and a $\chi^2(2)$.

- A table of critical values of 50:50 mixtures of $\chi^2(q)$ and $\chi^2(q+1)$ is given in the back of our text, and can be used for testing hypotheses on variance components in this special case.

- However, more generally (e.g., when $\mathbf{R} \neq \sigma^2\mathbf{I}$, or when comparing more complex random effects structures) the null distribution of the LRT can be difficult to find and use.

  - In such cases, our textbook authors suggest testing the hypotheses using the standard asymptotics of the LRT, but using $\alpha = 0.1$ instead of $\alpha = 0.05$.

- Alternatively, if we are not interested in inference on the variance-covariance structure per se, but simply want to choose an adequate var-cov structure so that we can get valid conclusions in our inferences on the mean, then we may choose not to do formal hypothesis tests on the variance-covariance structure of the model, but simply select an adequate var-cov structure via a model selection criterion, such as AIC.

Model Selection Criteria:

Consider testing some hypothesis H_0 versus an alternative H_A where these hypotheses correspond to nested models.

Let ℓ_0 and ℓ_A denote the loglikelihood function evaluated at the MLE under the null and alternative models, respectively. Further let $\#\boldsymbol{\phi}_0$ and $\#\boldsymbol{\phi}_A$ denote the number of free parameters under H_0 and H_A.

Then the LR test rejects H_0 is $\ell_A - \ell_0$ is large in comparison to the difference in d.f. of the two models to be compared. That is, it rejects if

$$\ell_A - \ell_0 > \mathcal{F}(\#\boldsymbol{\phi}_A) - \mathcal{F}(\#\boldsymbol{\phi}_0),$$

or equivalently, if

$$\ell_A - \mathcal{F}(\#\boldsymbol{\phi}_A) > \ell_0 - \mathcal{F}(\#\boldsymbol{\phi}_0),$$

for an appropriate function $\mathcal{F}$.

- For example, for an α-level LR test under standard regularity conditions (i.e., when the null hypothesis doesn't place the parameter on the boundary of the parameter space), $\mathcal{F}$ is a function such that

$$\mathcal{F}(\#\boldsymbol{\phi}_A) - \mathcal{F}(\#\boldsymbol{\phi}_0) = \frac{1}{2}\chi^2_{1-\alpha}(\#\boldsymbol{\phi}_A - \#\boldsymbol{\phi}_0).$$

- This procedure can only be considered a formal hypothesis test if the null and alternative correspond to nested models and if $\mathcal{F}(\#\boldsymbol{\phi}_A) - \mathcal{F}(\#\boldsymbol{\phi}_0)$ gives the appropriate critical value from the reference distribution of $\ell_A - \ell_0$.

However, there is no reason why the above procedure could not be used as a rule of thumb for comparing any two (not necessarily nested) models, or why we couldn't consider other functions $\mathcal{F}(\cdot)$ for choosing between models.

- Some commonly used functions for $\mathcal{F}$ are given in Table 6.7 of Verbeke & Molenberghs (2000, p.74) reproduced below.

TABLE 6.7. *Overview of frequently used information criteria for comparing linear mixed models. We hereby define n^* equal to the total number $n = \sum_{i=1}^{N} n_i$ of observations or equal to $n - p$, depending on whether ML or REML estimation was used in the calculations.*

Criterion	Definition of $\mathcal{F}(\cdot)$
Akaike (AIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta}$
Schwarz (SBC)	$\mathcal{F}(\#\boldsymbol{\theta}) = (\#\boldsymbol{\theta}\ \ln n^*)/2$
Hannan and Quinn (HQIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta}\ \ln(\ln n^*)$
Bozdogan (CAIC)	$\mathcal{F}(\#\boldsymbol{\theta}) = \#\boldsymbol{\theta}\ (\ln n^* + 1)/2$

- These functions yield different choices of information criteria, of which AIC and BIC are by far the most common.

- The basic idea in all of these criteria is to compare models based on their maximized loglikelihood values, but to penalize for the use of too many parameters (penalize model complexity).

- The model with the largest value of whichever criterion is chosen is the winner, but note that sometimes AIC and BIC are defined as -2 times the definition given here so that smallest is best.

- Note that for discriminating between variance-covariance structures, ℓ_R may be used in place of ℓ as long as we replace n by $n^* = n - p$, the number of linearly independent error contrasts used to form ℓ_R.

- Which criterion performs best is not an easy question to answer and depends upon the nature of the data, models, and the purpose to which the models are to be put.

- AIC tends to penalize less for model complexity than BIC, so it tends to err on the side of overspecified models, whereas BIC errs on the side of underspecification.

    - Underspecification tends to lead to bias, overspecification to inefficiency (increased variance).

- For choosing an adequate variance-covariance structure when interest centers on the mean, the consequences of underspecification are more dire. Therefore I do not recommend BIC.

- For selecting the variance-covariance structure in mixed models, AIC is commonly used. However, AIC is an estimator of the expected Kullback discrepancy between the true model and a fitted model, and as such it is known to be downwardly biased by an amount that disappears asymptotically, but can be substantial in small samples or whenever p is large relative to the total sample size n.

- Therefore, a variety of alternatives to AIC have been proposed that attempt to correct this bias with the goal of better performance in small samples. One of the most popular and effective alternatives is the corrected AIC (AICc) of Sugiura (1978) and Hurvich and Tsai (1989, 1993, 1995):

$$AICc = -2\ell(\hat{\boldsymbol{\phi}}) + 2k\left(\frac{n^*}{n^* - k - 1}\right) \quad \text{vs.} \quad AIC = -2\ell(\hat{\boldsymbol{\phi}}) + 2k,$$

  where $\hat{\boldsymbol{\phi}}$ is the MLE of the model parameter $\boldsymbol{\phi}$ and $k = \#\boldsymbol{\phi}$.

- AICc is recommended in many regression contexts (CLM, GLMs, time series analysis, nonlinear regression, etc.), but has not been formally justified in a mixed model context, especially for the selection of the variance-covariance/random effects structure. Therefore, it is not recommended for this purpose. Development of bias-corrected versions of the AIC for mixed models is an active area of research. At present, most of the methods that have been proposed and validated are computationally intensive and not implemented in standard software.

The Linear Mixed Model for Clustered Data:

So far, we have presented the LMM in its general form. However, application to longitudinal or other clustered data is among the most common uses of the LMM, so we now consider the LMM in that context.

Suppose we have data on n subjects, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{it_i})^T$ is the $t_i \times 1$ vector of observations available on the i^{th} subject, $i = 1, \ldots, n$. Then the LMM for longitudinal data is given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $t_i \times p$ and $t_i \times g$ design matrices for fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{b}_i$, respectively. $\boldsymbol{\varepsilon}_i$ is a vector of error terms.

As before, it is assumed that

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i(\boldsymbol{\theta}))$$
$$\text{and} \quad \mathbf{b}_1, \ldots, \mathbf{b}_n, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \quad \text{are independent.}$$

- Note that $\text{var}(\boldsymbol{\varepsilon}_i) = \mathbf{R}_i$ depends upon i through the dimension t_i of $\mathbf{y}_i$ (the cluster size), and may also depend on i by having a different form, or at least different parameter values for different subsets of subjects.

- $\mathbf{D}$ and the dimension of $\mathbf{b}_i$, however, are assumed to the same for all i.

The specification above can be equivalently expressed as

$$\mathbf{y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i(\boldsymbol{\theta})), \quad \text{where} \quad \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})). \qquad (\clubsuit)$$

- ($\clubsuit$) is known as the hierarchical formulation of the model.

Note that the hierarchical model implies a marginal model. That is, the marginal density of $\mathbf{y}_i$ can be obtained from the conditional density of $\mathbf{y}_i|\mathbf{b}_i$ and the marginal density of $\mathbf{b}_i$ throught the relationship

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i.$$

Because both densities in the integrand are normal, the integral yields a normal so that

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\theta})), \quad \mathbf{V}_i(\boldsymbol{\theta}) = \mathbf{Z}_i\mathbf{D}(\boldsymbol{\theta})\mathbf{Z}_i^T + \mathbf{R}_i(\boldsymbol{\theta}).$$

- That is, the hierarchical (conditionally specified) model implies a corresponding marginal model.

- Note however, that for arbitrary $\mathbf{V}_i(\boldsymbol{\theta})$ ($\mathbf{V}_i(\boldsymbol{\theta})$ an arbitrary p.d. matrix), or even $\mathbf{V}_i(\boldsymbol{\theta})$ p.d. and of the form $\mathbf{Z}_i\mathbf{D}(\boldsymbol{\theta})\mathbf{Z}_i^T + \mathbf{R}_i(\boldsymbol{\theta})$ where $\mathbf{D}$ and $\mathbf{R}_i$ are not assumed p.d., the marginal model does not necessarily imply the hierarchical one.

    - That is, there is a subtle distinction between the hierarchical and marginal formulations of the model. This distinction will become more important when we study generalized linear mixed models.