ST518 - Mixed effects models

for experiments with more than one factor

Jason A. Osborne

N. C. State Univ.

Outline

Topic: Mixed Models for factorial experiments

- Two factors
- crossed
- nested
- random/random
- fixed/fixed
- fixed/random

Two-factor designs
with factors that are fixed/random and nested/crossed

1. Entomologist records energy expended (y) by $N = 27$ honeybees
   - at three TEMPERATURES $(20, 30, 40^o C)$
   - consuming three levels of SUCROSE $(20\%, 40\%, 60\%)$

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

2. Experiment to study effect of drug and method of administration on fasting blood sugar in a random sample of $N = 18$ diabetic patients. (dataset on website is `blsugar.dat`)
    - First factor is drug: brand I tablet, brand II tablet, insulin injection
    - Second factor is type of administration (see table)

Drug (i)	Administration type (j)	Mean $\bar{y}_{j(i)}$	Variance $s^2_{j(i)}$
Brand I tablet	$(j = 1) 30mg \times 1$	15.7	6.3
$(i = 1)$	$(j = 2) 15mg \times 2$	19.7	9.3
Brand II tablet	$(j = 1) 20mg \times 1$	20	1
$(i = 2)$	$(j = 2) 10mg \times 2$	17.3	6.3
Insulin injection	$(j = 1)$ before breakfast	28	4
$(i = 3)$	$(j = 2)$ before supper	33	9

3. An experiment is conducted to determine variability among laboratories (interlaboratory differences) in their assessment of bacterial concentration in milk after pasteurization. Milk w/ various degrees of contamination was tested by randomly drawing four samples of milk from a collection of cartons at various stages of spoilage. Y is colony-forming units/μl. Labs think they're receiving 8 independent samples

	Sample			
Lab	1	2	3	4
1	2200	3000	210	270
	2200	2900	200	260
2	2600	3600	290	360
	2500	3500	240	380
3	1900	2500	160	230
	2100	2200	200	230
4	2600	2800	330	350
	4300	1800	340	290
5	4000	4800	370	500
	3900	4800	340	480

(Data from Oehlert, 2000)

4. An expt measures *Campylobacter* counts in $N = 120$ chickens in a processing plant, at four locations, over three days. Means (std) for $n = 10$ chickens sampled at each location tabulated below:

	Location			
Day	Before Washer	After Washer	After mic. rinse	After chill tank
1	70070.00	48310.00	12020.00	11790.00
	(79034.49)	(34166.80)	(3807.24)	(7832.05)
2	75890.00	52020.00	8090.00	8690.00
	(74551.32)	(17686.27)	(4848.01)	(5526.19)
3	95260.00	33170.00	6200.00	8370.00
	(03176.00)	(22259.08)	(5028.81)	(5720.15)

Data courtesy of Michael Bashor, General Mills

Transformation?

5. An experiment to assess the variability of a particular acid among plants and among leaves of plants:

Plant i	1			2			3			4		
Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$	11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$	11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$	12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

Data from Neter, et al (1996)

6. Plantheights from 10 pots (not 2!) randomized to 5 treatment combinations. (See Table 14.2 from Rao.)

Treatment	Dark	Source	Intensity	Pot	Seedling 1	Seedling 2
DD	1	D	D	1	32.94	35.98
DD	1	D	D	2	34.76	32.40
AL	0	A	L	1	30.55	32.64
AL	0	A	L	2	32.37	32.04
AH	0	A	H	1	31.23	31.09
AH	0	A	H	2	30.62	30.42
BL	0	B	L	1	34.41	34.88
BL	0	B	L	2	34.07	33.87
BH	0	B	H	1	35.61	35.00
BH	0	B	H	2	33.65	32.91

Six types of two-factor models

Fixed and/or random effects that are either crossed or nested

$$
\begin{array}{llcc|c}
1. & Y_{ijk} & = & \mu + A_i + B_j + (AB)_{ij} + E_{ijk} & \text{crossed/random} \\
2. & Y_{ijk} & = & \mu + \alpha_i + \beta_{j(i)} + E_{ijk} & \text{nested/fixed} \\
3. & Y_{ijk} & = & \mu + A_i + B_{j(i)} + E_{ijk} & \text{nested/random} \\
4. & Y_{ijk} & = & \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk} & \text{crossed/mixed} \\
5. & Y_{ijk} & = & \mu + \alpha_i + B_{j(i)} + E_{ijk} & \text{nested/mixed} \\
6. & Y_{ijk} & = & \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk} & \text{crossed/fixed}
\end{array}
$$

In the models above, (not ordered according to six prior datasets)

- GREEK symbols parameterize FIXED, unknown treatment means
- CAPITAL letters represent RANDOM effects
- for Model 3, $A_i, B_i, (AB)_{ij}$ are all independent
- for Model 4, $B_j, (\alpha B)_{ij}$ are all independent
- for Model 5, $A_i, B_{j(i)}$ are all independent
- RANDOM effects are used when it makes sense to think of LEVELS of factor as random sample from a population.

Identifying the appropriate model for our 6 examples:

1. Energy expended by honeybees.
   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:

$$Y_{ijk} = \mu + \qquad\qquad + E_{ijk}$$

2. Change in fasting blood sugar for diabetics
   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:

$$Y_{ijk} = \mu + \qquad\qquad + E_{ijk}$$

3. Measuring bacterial concentration in milk
    - First factor:
    - Second factor:
    - Fixed or random?
    - Crossed or nested?
    - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad\qquad\qquad + E_{ijk}$$

4. Measuring bacteria counts in chickens at processing plant
    - First factor:
    - Second factor:
    - Fixed or random?
    - Crossed or nested?
    - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad\qquad\qquad + E_{ijk}$$

5. Acids in leaves of plants
    - First factor:
    - Second factor:
    - Fixed or random?
    - Crossed or nested?
    - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad + E_{ijk}$$

6. Effect of light source and intensity on plant heights (Rao Table 14.2)
    - First factor:
    - Second factor:
    - Fixed or random?
    - Crossed or nested?
    - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad + E_{ijk}$$

Tables of expected means squares (EMS)

When A, B CROSSED, EMS tabulated below

Source	df	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{AB}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{\alpha B}^2$
B	$b-1$	$\sigma^2 + na\psi_B^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{AB}^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{\alpha B}^2$
AB	$(a-1)$ $\times(b-1)$	$\sigma^2 + n\psi_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{\alpha B}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

When factor B NESTED in factor A, EMS tabulated below:

Source	df	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{B(A)}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{B(A)}^2$
$B(A)$	$a(b-1)$	$\sigma^2 + n\psi_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

where ψ^2 and σ^2 values are defined on the next page.

1. If a factor X with index i is random then $EMS(X)$ is a linear combo of σ^2 and varcomps for all random effects ________________ index i. Coefficients for varcomps are limits of indexes ________________ listed (summed over) in random effects.
2. If a factor X is fixed. Treat it like it is random and then just replace the varcomp for X with the effect size, ψ_X^2.

$$
\begin{aligned}
\psi_A^2 &= \frac{1}{a-1}\sum_1^a \alpha_i^2 \quad \text{effect size of factor } A \\
\psi_B^2 &= \frac{1}{b-1}\sum_1^b \beta_i^2 \quad \text{effect size of factor } B \\
\psi_{AB}^2 &= \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2 \quad \text{effect size of interaction} \\
\psi_{B(A)}^2 &= \frac{1}{a(b-1)}\sum_{i=1}^a\sum_{j=1}^b \beta_{j(i)}^2 \quad \text{effect size of factor } B \\
\sigma_A^2 &= \mathrm{Var}(A_i) \quad \text{variance component for factor } A \\
\sigma_B^2 &= \mathrm{Var}(B_i) \quad \text{variance component for factor } B \\
\sigma_{AB}^2 &= \mathrm{Var}((AB)_{ij}) \quad \text{variance component for interaction} \\
\sigma_{B(A)}^2 &= \mathrm{Var}(B_{j(i)}) \quad \text{variance component for factor } B \\
\sigma^2 &= \mathrm{Var}(E_{ijk)}) \quad \text{error variance}
\end{aligned}
$$

The term *effect size* is often used in power considerations and sometimes involves division by σ^2.

Using expected mean squares to analyze data in mixed models

- *EMS* tables dictate which F-ratios test which effects
- *EMS* tables yield estimating equations for variance components

$\boxed{\text{Milk example:}}$ F-tests and estimating variance components.

1. To test for interaction effect, use $F_{AB} = \frac{MS[AB]}{MS[E]}$
2. To test for main effect of A, use $F_A = \frac{MS[A]}{MS[AB]}$
3. To test for main effect of B, use $F_B = \frac{MS[B]}{MS[AB]}$

Note the departure from fixed effects analysis, where $MS[E]$ is always used in the denominator.

The SAS System 1

The GLM Procedure

Dependent Variable: ly = log(y)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	56.03510844	2.94921623	191.44	<.0001
sample	3	53.18978788	17.72992929	1150.89	<.0001
lab	4	2.30248803	0.57562201	37.37	<.0001
sample*lab	12	0.54283253	0.04523604	2.94	0.0161
Error	20	0.30810726	0.01540536		
Corrected Total	39	56.34321569			

The wrong F-ratio and p-value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[E]} = \frac{0.5756}{0.0154} = 37.37 (p < 0.0001)$$

The correct F-ratio and p-value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{\qquad\qquad\qquad} = \frac{0.5756}{\qquad\qquad\qquad} = 12.72 (p = 0.0003)$$

Estimating variance components

The estimated variance components satisfy the following system of equations:

$$
\begin{aligned}
MS[E] &= \hat{\sigma}^2 \\
MS[AB] &= \hat{\sigma}^2 + n\hat{\sigma}^2_{AB} \\
&= \hat{\sigma}^2 + 2\hat{\sigma}^2_{AB} \\
MS[A] &= \hat{\sigma}^2 + nb\hat{\sigma}^2_{A} + n\hat{\sigma}^2_{AB} \\
&= \hat{\sigma}^2 + 8\hat{\sigma}^2_{A} + 2\hat{\sigma}^2_{AB} \\
MS[B] &= \hat{\sigma}^2 + na\hat{\sigma}^2_{B} + n\hat{\sigma}^2_{AB} \\
&= \hat{\sigma}^2 + 10\hat{\sigma}^2_{B} + 2\hat{\sigma}^2_{AB}
\end{aligned}
$$

Substitution of

$$\begin{aligned} MS[E] &= 0.0154 \\ MS[AB] &= 0.0452 \\ MS[A] &= 0.5756 \\ MS[B] &= 17.7299 \end{aligned}$$

into the system of equations yields estimated variance components:

$$\begin{array}{cccccccc} \hat{\sigma}^2 & = & MS[E] & = & & = & 0.0154 \\ \hat{\sigma}^2_{AB} & = & \frac{MS[AB]-MS[E]}{n} & = & \frac{0.0452-0.0154}{2} & = & 0.01492 \\ \hat{\sigma}^2_{A} & = & \frac{MS[A]-MS[AB]}{nb} & = & \frac{0.5756-0.0452}{8} & = & 0.0663 \\ \hat{\sigma}^2_{B} & = & \frac{MS[B]-MS[AB]}{na} & = & \frac{17.7299-0.0452}{10} & = & 1.768 \end{array}$$

```
data one;
   infile "milk.dat" firstobs=4;
   input sample lab y;
   ly=log(y);
run;

proc glm;
   class lab sample;
   model ly=sample|lab;
   random sample lab sample*lab;
   test h=lab sample e=sample*lab;
   lsmeans sample*lab;
run;
```

(We have to tell the software what the appropriate error term (denominator) is for testing for lab and sample effects.)

(We have to tell the software what the appropriate error term (denominator) is for testing for lab and sample effects.)

```
                              The GLM Procedure
Dependent Variable: ly
                                        Sum of
 Source                    DF          Squares    Mean Square   F Value   Pr > F
 Model                     19      56.03510844     2.94921623    191.44   <.0001
 Error                     20       0.30810726     0.01540536
 Corrected Total           39      56.34321569

            R-Square      Coeff Var       Root MSE        ly Mean
            0.994532       1.821098       0.124118       6.815577

 Source                    DF        Type I SS    Mean Square   F Value   Pr > F
 sample                     3      53.18978788    17.72992929   1150.89   <.0001
 lab                        4       2.30248803     0.57562201     37.37   <.0001
 lab*sample                12       0.54283253     0.04523604      2.94   0.0161

Source                   Type III Expected Mean Square
sample                   Var(Error) + 2 Var(lab*sample) + 10 Var(sample)
lab                      Var(Error) + 2 Var(lab*sample) + 8 Var(lab)
lab*sample               Var(Error) + 2 Var(lab*sample)

 Tests of Hypotheses Using the Type III MS for lab*sample as an Error Term

 Source                    DF      Type III SS    Mean Square   F Value   Pr > F
 lab                        4       2.30248803     0.57562201     12.72   0.0003
 sample                     3      53.18978788    17.72992929    391.94   <.0001
```

```
proc varcomp;
   class sample lab;
   model y=sample|lab;
run;
```

```
              Variance Components Estimation Procedure

                  Variance Component               ly

                  Var(sample)                 1.76847
                  Var(lab)                    0.06630
                  Var(sample*lab)             0.01492
                  Var(Error)                  0.01541
```

1. At the end of the day, what is the conclusion from the analysis of this crossed, random effects experiment?
2. For which experimental factors can the observed variation be declared significant?
3. What are the estimated variance components associated with these factors?
4. For a randomly sampled lab and degree of contamination, what is $\widehat{\mu}$ and its associated standard error?

1. There is evidence of variabilility due to laboratory$\times$ sample interaction; interlaboratory effects vary by sample.
2. The estimated parameters (μ+ variance components) of the model

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$$

$$\begin{aligned} \hat{\sigma}^2 &= 0.0154 \\ \hat{\sigma}^2_{AB} &= 0.0149 \\ \hat{\sigma}^2_A &= 0.0663 \\ \hat{\sigma}^2_B &= 1.7685 \\ \hat{\mu} &= 6.82(\text{log scale}) \end{aligned}$$

3. The standard error of $\bar{Y}_{+++}$ can be derived by

$$\begin{aligned} \bar{Y}_{+++} &= \mu + \bar{A}_+ + \bar{B}_+ + \overline{(AB)}_{++} + \bar{E}_{+++} \\ \text{Var}(\bar{Y}_{+++}) &= \text{Var}(\bar{A}_+) + \text{Var}(\bar{B}_+) + \text{Var}(\overline{(AB)}_{++}) + \text{Var}(\bar{E}_{+++}) \\ &= \frac{\sigma^2_A}{a} + \frac{\sigma^2_B}{b} + \frac{\sigma^2_{AB}}{ab} + \frac{\sigma^2}{abn} \quad \text{(how to estimate?)} \end{aligned}$$

Estimation of standard error and approximation of *df*

$$SE(\bar{Y}_{+++}) = \sqrt{\frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn}}$$

can be estimated by substitution of estimated variance components ($\hat{\sigma}^2$), which leads to

$$\begin{aligned} \widehat{SE}(\bar{Y}_{+++}) &= \sqrt{\frac{\hat{\sigma}_A^2}{a} + \frac{\hat{\sigma}_B^2}{b} + \frac{\hat{\sigma}_{AB}^2}{ab} + \frac{\hat{\sigma}^2}{abn}} \\ &= \text{lots of algebra and cancellations} \\ &= \sqrt{\frac{1}{nab}\left(MS[A] + MS[B] - MS[AB]\right)} \end{aligned}$$

For the milk data, we have

$$\widehat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)} = 0.6757$$

For a 95% confidence interval, we have a problem: we don't know how many *df* are associated with a t statistic based on this estimated SE.

ST511 Flashback - Unequal variances independent samples t-test

Example: Suspended particulate matter Y (in micrograms per cubic meter) in homes with smokers (Y_1) and without smokers (Y_2):

smokers	133	128	136	135	131	131	130	131	131	132	147
no smokers	106	85	84	95	104	79	72	115	95		

$\bar{y}_1 = 133.2, s_1^2 = 26.0,\ \bar{y}_2 = 92.8, s_2^2 = 195.4,\ n_1 = 11, n_2 = 9.$

- $Y_{11}, \ldots, Y_{1n_1}$ and $Y_{21}, \ldots, Y_{2n_2}$ iid samples from $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$,

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v. } H_1 : \mu_1 - \mu_2 \neq 0$$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

For small n_1, n_2, T not $N(0,1)$, nor does the version with S_p^2.

Satterthwaite approximation: $T \sim t_{df}$ with

$$\widehat{df} = \frac{(c_1 MS_1 + c_2 MS_2)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2}$$

where $MS_i = S_i^2$ and $c_i = 1/n_i$.

ST511 Flashback continued:

For the air pollution in homes with a smoking occupant data, $c_1 MS_1 = 26/11 = 2.36, c_2 MS_2 = 195.4/9 = 21.71$ and

$$\widehat{df} = \frac{(2.36 + 21.71)^2}{\frac{2.36^2}{10} + \frac{21.71^2}{8}} = 9.74$$

97.5^{th} percentile of t distn w $df = 9.74$ is $t(0.025, 9.74) = 2.236$.

95% conf. interval for $\mu_1 - \mu_2$ given by

$$133.2 - 92.8 \pm 2.236\sqrt{26/11 + 195.4/9}$$

or

$$40.4 \pm 2.236(4.91) \quad \text{or} \quad 40.4 \pm 10.97 \quad \text{or} \quad (29.4, 51.4)$$

These data would lead to the rejection of $H_0 : \mu_1 = \mu_2 = 0$ versus the two-tailed alternative. The observed test statistic is given by

$$t_{obs} = \frac{133.2 - 92.8}{\sqrt{26/11 + 195.4/9}} = \frac{40.4}{4.91} = 8.2 \quad (p < 0.0001)$$

This problem aka the Behrens-Fisher problem.

```
proc ttest;
   class smoke;
   var y;
```

```
                       The TTEST Procedure

                                    Lower CL             Upper CL
  Variable   smoke             N      Mean      Mean       Mean

  y                  0         9    82.032    92.778     103.52
  y                  1        11    129.76    133.18      136.6
  y          Diff (1-2)             -49.91     -40.4      -30.9

                            Lower CL           Upper CL
  Variable   smoke           Std Dev  Std Dev   Std Dev   Std Err

  y                  0         9.443    13.98    26.783      4.66
  y                  1        3.5603   5.0955    8.9422    1.5363
  y          Diff (1-2)       7.6046   10.064    14.883    4.5235

                              T-Tests

Variable    Method          Variances      DF    t Value    Pr > |t|

y           Pooled          Equal          18      -8.93     <.0001
y           Satterthwaite   Unequal      9.74*     -8.23*    <.0001

                       Equality of Variances

 Variable    Method       Num DF    Den DF    F Value    Pr > F

 y           Folded F          8        10       7.53    0.0045
```

Two-way random effects, milk data, Satterthwaite's approximation (cont'd)

To approximate *df* associated with *t* statistic based on std err of the form

$$\sqrt{c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k}$$

(linear combination of MS terms), use $\boxed{\text{Satterthwaite approximation:}}$

$$\begin{aligned}\widehat{df} &= \frac{(\sum c_i MS_i)^2}{\sum (c_i MS_i)^2/df_i} \\ &= \frac{(c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2 + \cdots + (c_k MS_k)^2/df_k}\end{aligned}$$

Recall that for the milk data, we have

$$\widehat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40}(MS[A] + MS[B] - MS[AB])}$$
$$= \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)} = 0.6757$$

$$\widehat{df} = \frac{(0.6757)^4}{(\frac{1}{40}17.73)^2/3 + (\frac{1}{40}0.58)^2/4 + (\frac{1}{40}0.045)^2/12} = 3.18$$

Using $t(0.025, 3.18) = 3.08$, a 95% confidence interval for the mean μ among the population of all labs and samples is given by

$$6.82 \pm 3.08(0.6757)$$

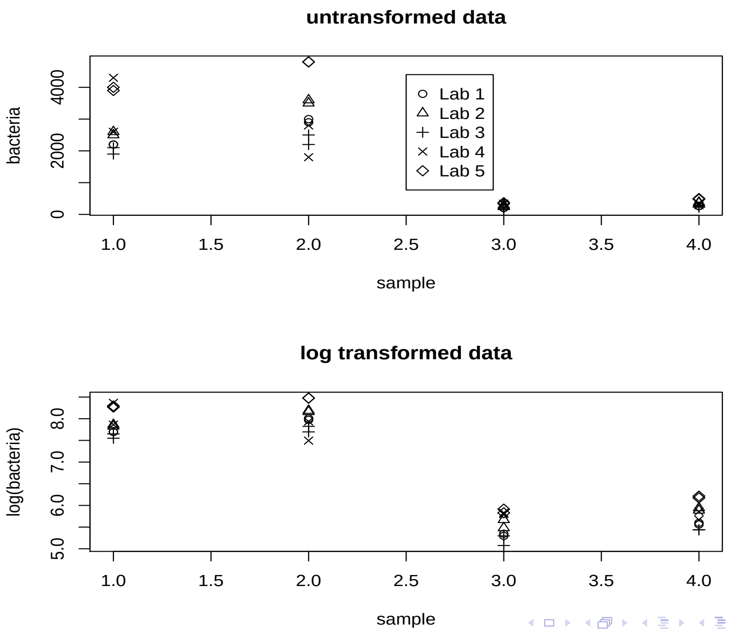
(plus or minus 3 standard errors!)

$$6.82 \pm 2.08$$

(log scale)

```
proc mixed cl;
   class sample lab;
   model ly=/s ddfm=satterth cl; *ly=log(y);
   random sample lab sample*lab;
run;
```

```
                             The Mixed Procedure

           Dependent Variable             ly
           Covariance Structure           Variance Components
           Estimation Method              REML
           Degrees of Freedom Method      Satterthwaite

                       Covariance Parameter Estimates

          Cov Parm         Estimate      Alpha       Lower       Upper

          sample             1.7685       0.05      0.5664     24.8486
          lab               0.06630       0.05     0.02233      0.7260
          sample*lab        0.01492       0.05    0.005761     0.09261
          Residual          0.01541       0.05    0.009017     0.03213

                         Solution for Fixed Effects

                           Standard
Effect       Estimate         Error       DF    t Value    Pr > |t|     Alpha
Intercept      6.8156        0.6757     3.18      10.09      0.0016      0.05

                       Effect          Lower       Upper
                       Intercept      4.7325      8.8987
```


untransformed data
bacteria
0
2000
4000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
sample
Lab 1
Lab 2
Lab 3
Lab 4
Lab 5
log transformed data
log(bacteria)
5.0
6.0
7.0
8.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
sample

```
milk.data <- read.table("milk.dat",skip=3,
  col.names=c("sample","lab","bacteria"))
attach(milk.data)
par(mfrow=c(2,1))
interaction.plot(sample,lab,(bacteria),
   main="Interaction plot, raw counts",col=1:5)
interaction.plot(sample,lab,log(bacteria),col=1:5,
   main="Interaction plot, log scale")
dev.copy2pdf(file="milkplot-color.pdf")
```

A nested design

Experiment to study effect of drug and method of administration on fasting blood sugar in diabetic patients

- First factor is drug: brand I tablet, brand II tablet, insulin injection
- Second factor is type of administration (see table)

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s^2_{j(i)}$	Mean $\bar{y}_{+(i)}$
Brand I tablet	$30mg \times 1$	15.7	6.3	17.7
	$15mg \times 2$	19.7	9.3	
Brand II tablet	$20mg \times 1$	20	1	18.7
	$10mg \times 2$	17.3	6.3	
Insulin injection	before breakfast	28	4	30.5
	before supper	33	9	

Definition: Factor B is __________ in factor A if there is a new set of levels of factor B for every different level of factor A.

Analysis of variance in nested designs

Two-factor design with factor B nested in factor A. Y_{ijk} denotes k^{th} response at level j of factor B within level i of factor A.

$$\boxed{Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk}}$$

for $i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b_i,\ k = 1, 2, \ldots, n$

$$SS[Tot] = SS[A] + SS[B(A)] + SS[E]$$

$$\sum_i \sum_j \sum_k (\qquad\qquad\qquad\qquad)^2 = SS[\qquad]$$

$$\sum_i \sum_j \sum_k (\qquad\qquad\qquad\qquad)^2 = SS[\qquad]$$

$$\sum_i \sum_j \sum_k (\qquad\qquad\qquad\qquad)^2 = SS[\qquad]$$

$$\sum_i \sum_j \sum_k (\qquad\qquad\qquad\qquad)^2 = SS[\qquad]$$

The ANOVA table looks like

Source	d.f.	Sum of squares	Mean Square	F
A	$a-1$	$SS[A]$	$MS[A]$	$\frac{MS[A]}{MS[E]}$
$B(A)$	$\sum_i(b_i-1)$	$SS[B(A)]$	$MS[B(A)]$	$\frac{MS[B(A)]}{MS[E]}$
Error	$N-\sum b_i$	$SS[E]$	$MS[E]$	
Total	$N-1$	$SS[TOT]$		

If $b_1 = b_2 = \cdots b_a = b$ then $\sum(b_i - 1) = a(b-1)$ and $df_E = ab(n-1)$.

- To test $H_0 : \alpha_i \equiv 0$, use F_A on $a-1$ and df_E degrees of freedom.
- To test $H_0 : \beta_{j(i)} \equiv 0$, for all i, j, use $F_{B(A)}$ on $\sum(b_i - 1)$ and df_E degrees of freedom.

For the diabetics blood sugar data, with $\bar{y}_{...} = 22.3$ and means

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s^2_{j(i)}$	Mean $\bar{y}_{+(i)}$
Brand I tablet	$30mg \times 1$	15.7	6.3	17.7
	$15mg \times 2$	19.7	9.3	
Brand II tablet	$20mg \times 1$	20	1	18.7
	$10mg \times 2$	17.3	6.3	
Insulin injection	before breakfast	28	4	30.5
	before supper	33	9	

$$
\begin{aligned}
SS[A] &= 2(3)[(17.7-22.3)^2 + (18.7-22.3)^2 + (30.5-22.3)^2 = 611.4 \\
SS[B(A)] &= 3[(15.7-17.7)^2 + (19.7-17.7)^2 + (20.0-18.7)^2 \\
&\quad +(17.3-18.7)^2 + (28-30.5)^2 + (33-30.5)^2] = 72.2 \\
SS[E] &= 72
\end{aligned}
$$

Q1: How many *df* associated with $SS[A]$?

Q2: How many *df* associated with $SS[B(A)]$?

Q3: How many *df* associated with $SS[E]$?

```
proc glm;
   class a b;
   model y=a b(a);
   output out=two p=p r=r;
   means a b(a)/lsd;
   estimate "effect of B within A=1" b(a) -1 1;
   estimate "effect of B within A=2" b(a) 0 0 -1 1;
   estimate "effect of B within A=3" b(a) 0 0 0 0 -1 1;
   estimate "A=1 mean - A=2 mean" a 1 -1;
   estimate "A=1 mean - A=3 mean" a 1 0 -1;
   estimate "A=2 mean - A=3 mean" a 0 1 -1;
run;
```

```
                                    Sum of
Source                    DF       Squares    Mean Square   F Value   Pr > F
Model                      5   683.6111111    136.7222222     22.79   <.0001
Error                     12    72.0000000      6.0000000
Corrected Total           17   755.6111111

Source                    DF     Type I SS    Mean Square   F Value   Pr > F
a                          2   611.4444444    305.7222222     50.95   <.0001
b(a)                       3    72.1666667     24.0555556      4.01   0.0344

                                               Standard
Parameter                       Estimate          Error    t Value    Pr > |t|
effect of B within A=1         4.0000000     2.00000000       2.00      0.0687
effect of B within A=2        -2.6666667     2.00000000      -1.33      0.2072
effect of B within A=3         5.0000000     2.00000000       2.50      0.0279
A=1 mean - A=2 mean           -1.0000000     1.41421356      -0.71      0.4930
A=1 mean - A=3 mean          -12.8333333     1.41421356      -9.07      <.0001
A=2 mean - A=3 mean          -11.8333333     1.41421356      -8.37      <.0001
```

Conclusions?

- The administration effect B (nested in the type of drug effect A) is statistically significant ($p = 0.0344$). This is due mostly to the before breakfast/supper difference, which is estimated to be

$$\bar{y}_{32+} - \bar{y}_{31+} = 5mg/dl$$

  with an (estimated) standard error of $SE = 2 =?$.
- Drug type effect (factor A) highly significant ($p < 0.0001$). Unadjusted pairwise comparisons indicate that insulin injections yield greater changes, on average, in blood sugar than either pill. Mean changes brought by the pills don't differ significantly.
- The following contrasts may be of interest:

$$\begin{aligned}\theta_1 &= \mu_{1(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)}) \\ \theta_2 &= \mu_{2(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)})\end{aligned}$$

  Exercise: Estimate them and test their significance ($H_0 : \theta_i = 0$).

More Two-factor mixed models

- *campylobacter* counts in $N = 120$ chickens in processing plant
  - Crossed design with two factors (Michael Bashor, General Mills)
    - Location (4 levels)
    - Day (3 levels)
  - 4×3 layout, $n = 10$ chickens per combo

Day	Location: Before Washer	After Washer	After mic. rinse	After chill tank
1	70070.00	48310.00	12020.00	11790.00
	(79034.49)	(34166.80)	(3807.24)	(7832.05)
2	75890.00	52020.00	8090.00	8690.00
	(74551.32)	(17686.27)	(4848.01)	(5526.19)
3	95260.00	33170.00	6200.00	8370.00
	(03176.00)	(22259.08)	(5028.81)	(5720.15)

- An experiment to assess the variability of a particular acid among plants and among leaves of plants:

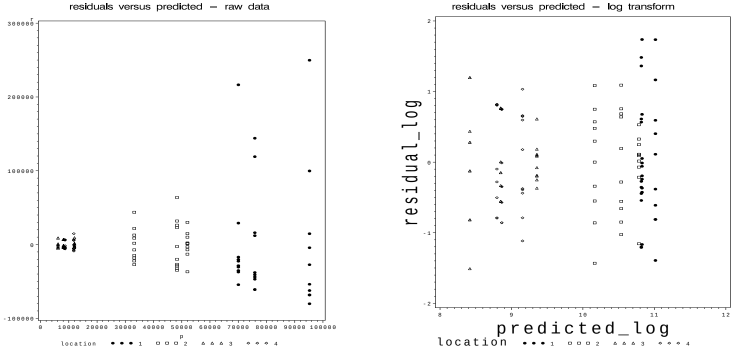
Plant i	1			2			3			4		
Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$	11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$	11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$	12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

- Study of light source and intensity on plant height.

Treatment	Dark	Source	Intensity	Pot	Seedling 1	Seedling 2
DD	1	D	D	1	32.94	35.98
DD	1	D	D	2	34.76	32.40
AL	0	A	L	1	30.55	32.64
AL	0	A	L	2	32.37	32.04
AH	0	A	H	1	31.23	31.09
AH	0	A	H	2	30.62	30.42
BL	0	B	L	1	34.41	34.88
BL	0	B	L	2	34.07	33.87
BH	0	B	H	1	35.61	35.00
BH	0	B	H	2	33.65	32.91

Analysis of *Campylobacter* counts on chickens data

Residual plots (resid .vs $\hat{y}$) for bacteria counts, after fitting two factor fixed effects models (similar plots for mixed models):



```
data one;   infile "bashor.dat" firstobs=3; input day location y;ly=log(y);

proc glm;
   class day location;
   model y ly=location|day;
   output out=two r=residual residual_log p=predicted predicted_log;
run;
/*
symbol1 value=dot color=black;      symbol2 value=square color=black;
symbol3 value=triangle color=black; symbol4 value=diamond color=black;

axis1 offset=(1,1) label=(height=3);
axis2 offset=(1,1) label=(height=3 angle=90);
legend1 label=(height=2);

proc gplot data=two;
   title "residuals versus predicted";
   plot residual*predicted=location/haxis=axis1 vaxis=axis2 legend=legend1;
   plot residual_log*predicted_log=location/haxis=axis1 vaxis=axis2 legend=legend1;
run; */
proc mixed method=type3 cl;
   class day location;
   model ly=location/ddfm=satterth outp=predz;
   random day day*location;
   lsmeans location/adj=tukey;
run;

proc mixed method=type3;  * to get ANOVA table with EMS terms;
*proc mixed cl;  * to get asymmetric confidence intervals ;
   class day location;
   model ly=location/ddfm=satterth;
   random day day*location;
   lsmeans location/adj=tukey;
run;
```

```
------------------------------------------------------------------------------------
                                   The SAS System                                  1
                                 The Mixed Procedure
                                 Model Information

               Data Set                        WORK.ONE
               Dependent Variable              ly
               Covariance Structure            Variance Components
               Estimation Method               Type 3
               Fixed Effects SE Method         Model-Based
               Degrees of Freedom Method       Satterthwaite

                             Type 3 Analysis of Variance

                               Sum of
Source              DF        Squares    Mean Square   Expected Mean Square
location             3      97.865388      32.621796   Var(Residual) + 10
                                                       Var(day*location) + Q(location)
day                  2       2.787355       1.393677   Var(Residual) + 10
                                                       Var(day*location) + 40 Var(day)
day*location         6       4.533565       0.755594   Var(Residual) + 10
                                                       Var(day*location)
Residual           108      59.254946       0.548657   Var(Residual)

                              Type 3 Analysis of Variance

                                                           Error
  Source          Error Term                                  DF    F Value     Pr > F
  location        MS(day*location)                             6      43.17     0.0002
  day             MS(day*location)                             6       1.84     0.2375
  day*location    MS(Residual)                               108       1.38     0.2303
------------------------------------------------------------------------------------
```

(generated by 2nd run of PROC MIXED)

*	Cov Parm	Estimate	Alpha	Lower	Upper
*					
*	day	0.01595	0.05	0.002071	1156981
*	day*location	0.02069	0.05	0.002844	145734
*	Residual	0.5487	0.05	0.4274	0.7303

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
location	3	6	43.17	0.0002

Least Squares Means

Effect	location	Estimate	Standard Error	DF	t Value	Pr > |t|
location	1	10.8870	0.1747	7.33	62.33	<.0001
location	2	10.4953	0.1747	7.33	60.09	<.0001
location	3	8.8745	0.1747	7.33	50.81	<.0001
location	4	8.9394	0.1747	7.33	51.18	<.0001

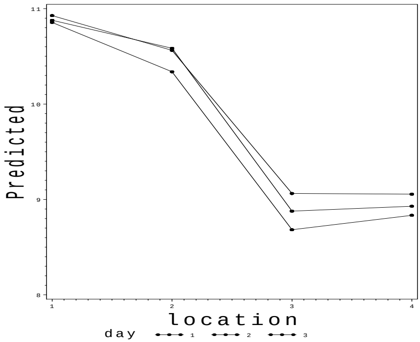
Differences of Least Squares Means

Effect	location	_location	Estimate	Standard Error	DF	t Value	Pr > |t|
location	1	2	0.3917	0.2244	6	1.75	0.1316
location	1	3	2.0125	0.2244	6	8.97	0.0001
location	1	4	1.9476	0.2244	6	8.68	0.0001
location	2	3	1.6208	0.2244	6	7.22	0.0004
location	2	4	1.5559	0.2244	6	6.93	0.0004
location	3	4	-0.06488	0.2244	6	-0.29	0.7823

```
------------------------------------------------------------------------------------
                         Differences of Least Squares Means

              Effect      location    _location    Adjustment       Adj P
              location    1           2            Tukey-Kramer    0.3801
              location    1           3            Tukey-Kramer    0.0004
              location    1           4            Tukey-Kramer    0.0005
              location    2           3            Tukey-Kramer    0.0015
              location    2           4            Tukey-Kramer    0.0018
              location    3           4            Tukey-Kramer    0.9907
```

Theory for mixed/crossed model used to analyze *Campylobacter* data

Discussion of MIXED output



Model

$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$$

w/ variance components $\sigma_B^2, \sigma_{\alpha B}^2, \sigma^2$.

Campylobacter analysis, continued

Fixed Factor A: location Random Factor B: day

To test $H_0 : \sigma^2_{\alpha B} = 0$, use

$$F_{AB} = \frac{MS[AB]}{MS[E]} = \frac{0.76}{0.55} = 1.38$$

on $(a-1)(b-1) = 6$ and $ab(n-1) = 108$ df. The p-value is 0.2303. providing no evidence of a random day $\times$ location interaction effect. The variance component for this random effect is estimated by

$$\hat{\sigma}^2_{\alpha B} = \frac{MS[AB] - MS[E]}{n} = \frac{0.76 - 0.55}{10} = 0.021$$

Intepretation: there is no evidence that day-to-day variability varies by location. The estimated variance component is itself very small.

$$\begin{aligned} \hat{\sigma}^2 &= MS[E] = \boxed{0.55} \\ \hat{\sigma}^2_B &= \frac{MS[B] - MS[AB]}{na} \\ &= \frac{1.39 - 0.76}{40} = \boxed{0.016} \end{aligned}$$

Implied correlation structure

What is the correlation of two observations taken on the same day

- at the same location?
- at different locations?

Recall that $\boxed{Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}.}$

$$\begin{aligned} Corr(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\ &= \frac{\text{Cov}(B_i, B_i) + \text{Cov}((\alpha B)_{ij}, (\alpha B)_{ij})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\ &= \frac{\sigma^2_B + \sigma^2_{\alpha B}}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \end{aligned}$$

$$
\begin{aligned}
Corr(Y_{1jk_1}, Y_{2jk_2}) &= \frac{\text{Cov}(Y_{1jk_1}, Y_{2jk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\text{Cov}(B_i, B_i)}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\sigma^2_B}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}}
\end{aligned}
$$

Estimates of these correlations are

- $\frac{0.016+0.021}{0.016+0.021+0.55} = \frac{0.037}{.587} = 0.063$
- $\frac{0.016}{0.016+0.021+0.55} = \frac{0.016}{.587} = 0.027$

Which is which?

What about the correlation of two observations on different days?

Some analysis of fixed effects

Consider testing for a fixed effect of location. That is, test the hypothesis that average bacteria counts are constant across the locations,

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$F_A = \frac{MS[A]}{MS[AB]} = \frac{32.6}{0.76} = 43.2$$

on $a - 1 = 3$ and $(a-1)(b-1) = 6$ df, which is significant ($p = 0.0002$).

To estimate the a pairwise comparison among location means, such as, $\alpha_4 - \alpha_3$, consider

$$\hat{\theta} = \bar{y}_{4++} - \bar{y}_{3++} = 8.940 - 8.875 = -0.065 \quad (SE = ?)$$

Note that $\text{Var}(\bar{Y}_{4++} - \bar{Y}_{3++}) \neq \sigma^2(\frac{1}{nb} + \frac{1}{nb})$ (Why not?)

What is $SE(\hat{\theta})$ and how can it be estimated?

$$\begin{aligned}
\hat{\theta} &= \bar{Y}_{2++} - \bar{Y}_{1++} \\
&= \\
&= \\
&= \\
\text{Var}(\widehat{\theta}) &= \text{Var}\left(\overline{\alpha B}_{2+}\right) + \text{Var}(\overline{\alpha B}_{1+})\big) + \text{Var}(\overline{E}_{2++}) + \text{Var}(\overline{E}_{1++}) \\
&= \\
&=
\end{aligned}$$

So far we have

$$\text{Var}(\overline{Y}_{4++} - \overline{Y}_{3++}) = \text{Var}(\frac{2}{nb}(\sigma^2 + n\sigma^2_{\alpha B}))$$

which can be estimated nicely on $(a-1)(b-1) = 6df$ by

$$\widehat{\text{Var}}(\widehat{\theta}) = \frac{2}{nb} MS[\qquad\qquad]$$

for the chickens, where $\overline{y}_{4++} - \overline{y}_{3++} = -0.06$ the SE is

$$\sqrt{\widehat{\text{Var}}(\widehat{\theta})} = \sqrt{\frac{2}{3*10}0.76} = 0.22$$

Since $t(0.025, 6) = 2.45$, a 95% c.i. for θ given by $-0.06 \pm 2.45(0.22)$.

Campylobacter analysis, continued

Reporting standard errors for sample means of levels of fixed factor, like LOCATION means, is a little messier:

$$
\begin{aligned}
\overline{Y}_{i++} &= \mu + \alpha_i + \overline{B} + \overline{\alpha B}_{i+} + \overline{E}_{i++} \\
\text{Var}(\overline{Y}_{i++}) &= \text{Var}(\overline{B}) + \text{Var}(\overline{\alpha B}_{i+}) + \text{Var}(\overline{E}_{i++}) \\
&= \\
&= \frac{1}{nb}\left(\qquad\qquad\qquad\qquad \right) \\
&\quad \text{estimated by} \\
\widehat{\text{Var}}(\overline{Y}_{i++}) &= \frac{1}{nb}(n\hat{\sigma}^2_B + n\hat{\sigma}^2_{\alpha B} + \hat{\sigma}^2) \\
&= \text{algebra yields a linear combo of multiple EMS terms} \\
&= \frac{1}{nab}\{(a-1)EMS[AB] + EMS[B]\}
\end{aligned}
$$

The standard error is estimated easily enough:

$$\begin{aligned}\widehat{SE}(\overline{Y}_{i++}) &= \sqrt{\frac{1}{nab}\{(a-1)MS[AB]+MS[B]\}}\\ &= \sqrt{\frac{1}{120}\{(4-1)0.76+1.39\}}\\ &= \sqrt{0.03} = 0.175\end{aligned}$$

but the *df* must be approximated using the Satterthwaite approach

$$\hat{df} = \frac{0.175^4}{\frac{1}{120^2}\left(\frac{((4-1)0.76)^2}{6}+\frac{1.39^2}{2}\right)} = 7.33$$

with $df_{AB} = 6, df_B = 2$. Since $t(0.025, 7.33) = 2.34$, a 95% c.i. for the population mean of location 1, for example, is $\boxed{10.9 \pm 2.34(0.175)}$.

SAS code to fit two-factor random effects model for plant acid data

Nested or crossed?

```
proc mixed cl method=type3;
*proc mixed cl;
   class plant leaf;
   model y=/s cl;
   random plant leaf(plant);
run;

goptions colors=(black) dev=pslepsf;
*goptions colors=(black);

axis1 value=(h=2) offset=(10);

symbol1 value=dot h=1.5;
symbol2 value=diamond h=1.5;
symbol3 value=plus h=1.5;

proc gplot; title "plant acids";
   plot y*plant=leaf/haxis=axis1;
run;
```

```
                         The Mixed Procedure
                       Class Level Information

             Class     Levels     Values

             plant          4     1 2 3 4
             leaf           3     1 2 3

                     Type 3 Analysis of Variance

                                          Sum of
            Source              DF       Squares      Mean Square

            plant                3    343.178889       114.392963
            leaf(plant)          8    187.453333        23.431667
            Residual            24      3.033333         0.126389
                                                                         Error
Source         Expected Mean Square                  Error Term             DF

plant          Var(Residual) + 3 Var(leaf(plant))    MS(leaf(plant))         8
               + 9 Var(plant)
leaf(plant)    Var(Residual) + 3 Var(leaf(plant))    MS(Residual)           24
Residual       Var(Residual)                         .                       .

                         Source         F Value     Pr > F

                         plant             4.88     0.0324
                         leaf(plant)     185.39     <.0001
```

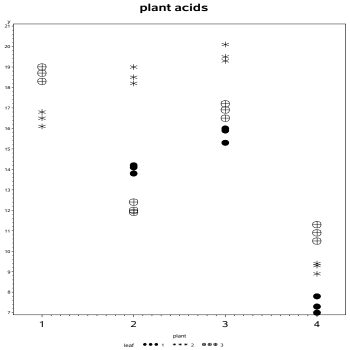
```
                 Covariance Parameter Estimates

      Cov Parm          Estimate      Alpha        Lower        Upper

      plant              10.1068       0.05     -10.3930      30.6066
      leaf(plant)         7.7684       0.05       0.1142      15.4227
      Residual            0.1264       0.05      0.07706       0.2446

               /*Covariance Parameter Estimates*/

      Cov Parm          Estimate      Alpha        Lower        Upper

      plant              10.1068       0.05       2.6599       499.70
      leaf(plant)         7.7684       0.05       3.5322      28.7787
      Residual            0.1264       0.05      0.07706       0.2446
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Intercept	14.2611	1.7826	3	8.00	0.0041	0.05

Solution for Fixed Effects

Effect	Lower	Upper
Intercept	8.5882	19.9341



Discussion of MIXED output and analysis of plant acid data

Random, nested model

$$Y_{ijk} =$$

w/ variance components

To test for random effect of nested factor B (leaf), $H_0 : \sigma^2_{B(A)} = 0$,

$$F = \frac{MS[B(A)]}{MS[E]} = \frac{23.4}{0.13} = 185.4$$

on $(b-1)a = 8$ and $(n-1)ab = 24$ df (p-value < 0.0001).

To test for random effect of factor A (plant), $H_0 : \sigma^2_A = 0$,

$$F = \frac{MS[A]}{MS[B(A)]} = \frac{114.4}{23.4} = 4.88$$

on $a - 1 = 3$ and $(b-1)a = 8df$ with $p = 0.0324$.
Reminder: Watch that denominator MS!

How big are the variance components?

$$
\begin{aligned}
\hat{\sigma}^2 &= MS[E] = \boxed{0.13} \\
\hat{\sigma}^2_{B(A)} &= \\
&= \frac{23.4 - 0.13}{3} = \boxed{7.8} \\
\hat{\sigma}^2_A &= \\
&= \frac{114.4 - 23.4}{9} = \boxed{10.1}
\end{aligned}
$$

So there is some evidence of both a random plant effect and a random leaf effect, nested in plant. The magnitudes of these effects are quantified by the estimated variance components. The statistical significance addressed by the *p*-values.

Implied correlation structure for plant acids

Correlation of two observations taken from same plant?

- also same leaf?
- different leaves?

Recall that $\boxed{Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}.}$

$$
\begin{aligned}
Corr(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} \\
&=
\end{aligned}
$$

Estimated correlation:

$$
\frac{10.1 + 7.8}{10.1 + 7.8 + 0.13} = \frac{17.9}{18.0} = 0.99
$$

$$
\begin{aligned}
Corr(Y_{ij_1k_1}, Y_{ij_2k_2}) &= \frac{\mathrm{Cov}(Y_{ij_1k_1}, Y_{2j_2k_2})}{\sigma^2+\sigma_A^2+\sigma_{B(A)}^2} \\
&=
\end{aligned}
$$

Estimated correlation :

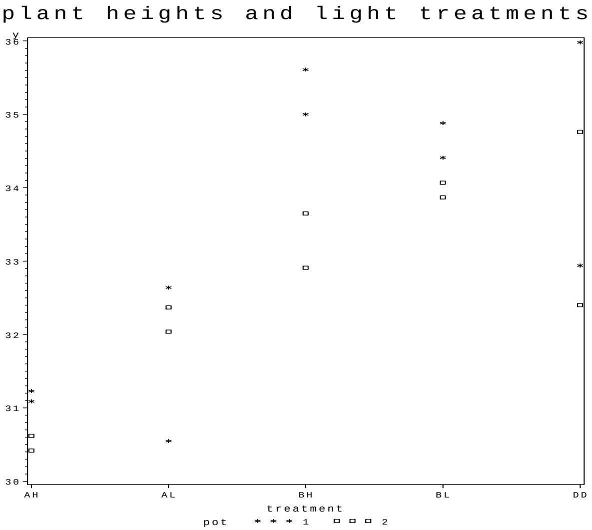
$$
\frac{10.1}{10.1+7.8+0.13} = \frac{10.1}{18.0} = 0.56
$$

- $\frac{10.1+7.8}{10.1+7.8+0.13} = \frac{17.9}{18.0} = 0.99$
- $\frac{10.1}{10.1+7.8+0.13} = \frac{10.1}{18.0} = 0.56$

This means that two measurements taken on the same leaf are almost perfectly correlated. Almost all the variation in any measurement can be explained by the leaf and plant effects.

Treatment	Dark	Source	Intensity	Pot	Seedling 1	Seedling 2
DD	1	D	D	1	32.94	35.98
DD	1	D	D	2	34.76	32.40
AL	0	A	L	1	30.55	32.64
AL	0	A	L	2	32.37	32.04
AH	0	A	H	1	31.23	31.09
AH	0	A	H	2	30.62	30.42
BL	0	B	L	1	34.41	34.88
BL	0	B	L	2	34.07	33.87
BH	0	B	H	1	35.61	35.00
BH	0	B	H	2	33.65	32.91

- Response (y) is seedling height,
- treatments are light sources, intensities,
- experimental units are 10 pots (points on graph).


plant heights and light treatments
y
36
35
34
33
32
31
30
AH
AL
BH
BL
DD
treatment
pot * * * 1 □ □ □ 2

Experiment with light treatments on seedlings

$$Y_{ijk} = \mu + \alpha_i + P_{j(i)} + E_{ijk}$$

α_i - treatment effects for $i = 1, 2, 3, 4, 5$
$P_{j(i)}$ - pot effects, nested in treatments, $j = 1, 2$ for each i.
E_{ijk} - seedling/experimental errors, $k = 1, 2$

$$P_{j(i)} \overset{iid}{\sim} N(0, \sigma_P^2), \quad E_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \;\; (P_{j(i)} \perp E_{ijk})$$

For treatment effects, use $MS(Pot(treatments))$ as error term.
For example, for $H_0 : \alpha_1 = \alpha_2 = \cdots = 0$, use

$$F = \frac{MS(\text{treatment})}{MS(\text{Pot}(\text{treatment}))} \sim F_{5-1,5(2-1)} \text{ or } F_{4,5}$$

Be careful not to use

$$F = \frac{MS(\text{treatment})}{MS(E)}$$

For these data, we get $F = \frac{10.27}{1.22} = 8.4(df = 4, 5, p = .0192)$, providing evidence of a treatment effect on plant heights.

SAS code to fixed the mixed effects model (output follows):

```
proc mixed method=type3 cl;
*proc mixed data=planthts cl;
   class pot treatment;
   model y=treatment;
   random pot(treatment);
   *lsmeans treatment/diffs adj=tukey;
   lsmeans treatment/diffs;
   estimate "main effect of source" treatment 1 1 -1 -1/divisor=2;
   estimate "main effect of intensity" treatment 1 -1 1 -1/divisor=2;
   estimate "interaction " treatment 1 -1 -1 1;
   contrast "main effect of source" treatment 1 1 -1 -1;
   contrast "main effect of intensity" treatment 1 -1 1 -1;
   contrast "interaction " treatment 1 -1 -1 1;
run;
```

```
The SAS System                                                              1
The Mixed Procedure
               Class Level Information

Class         Levels    Values
pot                2    1 2
treatment          5    AH AL BH BL DD
             Type 3 Analysis of Variance

                                 Sum of
Source                DF        Squares      Mean Square
treatment              4      41.080770        10.270192
pot(treatment)         5       6.112350         1.222470
Residual              10      10.264200         1.026420
```

```
                          Type 3 Analysis of Variance
                                                                               Error
Source          Expected Mean Square                   Error Term                 DF
treatment       Var(Residual) + 2 Var(pot(treatment))  MS(pot(treatment))          5
                + Q(treatment)
pot(treatment)  Var(Residual) + 2 Var(pot(treatment))  MS(Residual)               10
Residual        Var(Residual)                          .                           .

   Type 3 Analysis of Variance

Source          F Value    Pr > F
treatment          8.40    0.0192
pot(treatment)     1.19    0.3793
Residual            .        .

               Covariance Parameter Estimates

Cov Parm           Estimate     Alpha       Lower       Upper
pot(treatment)      0.09802      0.05     -0.7831      0.9792
Residual             1.0264      0.05      0.5011      3.1612

/*
pot(treatment)      0.09802      0.05    0.008606    3.993E31
Residual             1.0264      0.05      0.5011      3.1612
*/
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
main effect of source	-2.9300	0.5528	5	-5.30	0.0032
main effect of intensity	-0.5375	0.5528	5	-0.97	0.3756
interaction	-1.0450	1.1057	5	-0.95	0.3880

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
main effect of source	1	5	28.09	0.0032
main effect of intensity	1	5	0.95	0.3756
interaction	1	5	0.89	0.3880

Output for plant heights and light sources, cont'd

Least Squares Means

Effect	treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
treatment	AH	30.8400	0.5528	5	55.79	<.0001
treatment	AL	31.9000	0.5528	5	57.70	<.0001
treatment	BH	34.2925	0.5528	5	62.03	<.0001
treatment	BL	34.3075	0.5528	5	62.06	<.0001
treatment	DD	34.0200	0.5528	5	61.54	<.0001

Differences of Least Squares Means

Effect	treatment	_treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
treatment	AH	AL	-1.0600	0.7818	5	-1.36	0.2332
treatment	AH	BH	-3.4525	0.7818	5	-4.42	0.0069
treatment	AH	BL	-3.4675	0.7818	5	-4.44	0.0068
treatment	AH	DD	-3.1800	0.7818	5	-4.07	0.0097
treatment	AL	BH	-2.3925	0.7818	5	-3.06	0.0281
treatment	AL	BL	-2.4075	0.7818	5	-3.08	0.0275
treatment	AL	DD	-2.1200	0.7818	5	-2.71	0.0422
treatment	BH	BL	-0.01500	0.7818	5	-0.02	0.9854
treatment	BH	DD	0.2725	0.7818	5	0.35	0.7416
treatment	BL	DD	0.2875	0.7818	5	0.37	0.7281

Using nested factorial effects to get SAS to produce appropriate contrast sums of squares for factorial effects analysis of plant height and light source data

```
proc mixed method=type3;
   class pot treatment source intensity dark;
   model y=dark source(dark) intensity(dark) source*intensity(dark) dark ;
   random pot(source*intensity*dark);
   lsmeans dark source(dark) intensity(dark) source*intensity(dark);
run;
```

```
The SAS System        The Mixed Procedure

Class          Levels    Values
pot                 2    1 2
treatment           5    AH AL BH BL DD
source              3    A B D
intensity           3    D H L
dark                2    0 1
```

```
                         Type 3 Analysis of Variance

                                   Sum of
Source                  DF       Squares     Mean Square   Expected Mean Square
dark                     1      4.493520        4.493520   Var(Residual) + 2
                                                           Var(pot(sour*inten*dark)) +
                                                           Q(dark,source(dark),intensity
                                                           (dark),source*intensi(dark))
source(dark)             1     34.339600       34.339600   Var(Residual) + 2
                                                           Var(pot(sour*inten*dark)) +
                                                           Q(source(dark),source*intensi(dark))
intensity(dark)          1      1.155625        1.155625   Var(Residual) + 2
                                                           Var(pot(sour*inten*dark)) +
                                                           Q(intensity(dark),source*
                                                           intensi(dark))
source*intensi(dark)     1      1.092025        1.092025   Var(Residual) + 2
                                                           Var(pot(sour*inten*dark))
                                                           + Q(source*intensi(dark))
pot(sour*inten*dark)     5      6.112350        1.222470   Var(Residual) + 2
                                                           Var(pot(sour*inten*dark))
Residual                10     10.264200        1.026420   Var(Residual)

                                                               Error
Source                  Error Term                                DF      F Value    Pr > F
dark                    MS(pot(sour*inten*dark))                   5         3.68    0.1133
source(dark)            MS(pot(sour*inten*dark))                   5        28.09    0.0032
intensity(dark)         MS(pot(sour*inten*dark))                   5         0.95    0.3756
source*intensi(dark)    MS(pot(sour*inten*dark))                   5         0.89    0.3880
pot(sour*inten*dark)    MS(Residual)                              10         1.19    0.3793
Residual                .                                          .            .         .
```

Covariance Parameter Estimates

Cov Parm	Estimate
pot(sour*inten*dark)	0.09802
Residual	1.0264

Least Squares Means

Effect	source	intensity	dark	Estimate	Standard Error	DF	t Value	Pr > |t|
dark			0	32.8350	0.2764	5	118.79	<.0001
dark			1	34.0200	0.5528	5	61.54	<.0001
source(dark)	A		0	31.3700	0.3909	5	80.25	<.0001
source(dark)	B		0	34.3000	0.3909	5	87.74	<.0001
source(dark)	D		1	34.0200	0.5528	5	61.54	<.0001
intensity(dark)		H	0	32.5662	0.3909	5	83.31	<.0001
intensity(dark)		L	0	33.1038	0.3909	5	84.68	<.0001
intensity(dark)		D	1	34.0200	0.5528	5	61.54	<.0001
source*intensi(dark)	A	H	0	30.8400	0.5528	5	55.79	<.0001
source*intensi(dark)	A	L	0	31.9000	0.5528	5	57.70	<.0001
source*intensi(dark)	B	H	0	34.2925	0.5528	5	62.03	<.0001
source*intensi(dark)	B	L	0	34.3075	0.5528	5	62.06	<.0001
source*intensi(dark)	D	D	1	34.0200	0.5528	5	61.54	<.0001

Inference for light effects

Model for treatment combination "*ijk*" and pot l, seedling m:

$$Y_{ijklm} = \mu + \delta_i + \alpha_{j(i)} + \beta_{k(i)} + (\alpha\beta)_{jk(i)} + P_{l(ijk)} + E_{ijklm}$$

For treatment effects, use $MS(Pot(treatments))$ as the error term.
e.g.: Is intensity effect is constant across light types? $(H_0 : \gamma_{1jk} \equiv 0)$

$$F = \frac{MS(\text{interaction(dark)})}{MS(\text{Pot(dark*source*intensity)})} = \frac{1.09}{1.22} = .89(p = .3880)$$

Degrees of freedom: $(df = ?, ?)$
Estimation of variance components:

$$\hat{\sigma}^2 = MS(E) = 1.02(df = 10)$$

$$\hat{\sigma}^2_{P(T)} = \frac{MS(\text{pot(treatment)}) - MS(E)}{2} = \frac{1.22 - 1.02}{2} = 0.098(df = \widehat{df})$$

Correlation structure? Intrapot correlation?

$$\widehat{\text{Corr}}(Y_{ijklm_1}, Y_{ijklm_2}) = \frac{\hat{\sigma}^2_{P(T)}}{\hat{\sigma}^2 + \hat{\sigma}^2_{P(T)}} = \frac{.098}{.098 + 1.02} = .088$$