ST518 - Mixed effects models

Mixed Effects Models

Jason A. Osborne

N. C. State Univ.

Outline

Topic: Mixed effects models

- One-way random effects model to study *variances*
- Mixed effects models
- Subsampling
- Expected mean squares for mixed models

One-way random effects model

Example:

- Genetics study w/ beef animals. Measure birthweight Y (lbs).
- $t = 5$ sires, each mated to a separate group of $n = 8$ dams.
- $N = 40$, completely randomized.

Birthweights

Sire #	Level	Sample								$\bar{y}_{i\cdot}$	s_i
177	1	61	100	56	113	99	103	75	62	83.6	22.6
200	2	75	102	95	103	98	115	98	94	97.5	11.2
201	3	58	60	60	57	57	59	54	100	63.1	15.0
202	4	57	56	67	59	58	121	101	101	77.5	25.9
203	5	59	46	120	115	115	93	105	75	91.0	28.0

Q: Statistical model for these data? $Y_{ij} = \qquad\qquad\qquad + E_{ij}$

Random effects model

The one-way random effects model:

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $T_1, T_2, \ldots, T_t \overset{iid}{\sim} N(0, \sigma_T^2)$
- $E_{11}, \ldots, E_{tn} \overset{iid}{\sim} N(0, \sigma^2)$
- $T_1, T_2, \ldots, T_t$ independent of $E_{11}, \ldots, E_{tn}$

Features

- $T_1, T_2, \ldots$ denote *random* effects, drawn from some population of interest. That is, $T_1, T_2, \ldots$ is a $\boxed{\text{random sample}}$!
- σ_T^2 and σ^2 are called $\boxed{\text{variance components}}$
- conceptually different from one-way fixed effects model

Beef animal genetic study, continued
With $t = 5$ and $n = 8$, the random effects $T_1, T_2, \ldots, T_5$ reflect sire-to-sire variability.

No particular interest in $\tau_1, \tau_2, \ldots, \tau_5$ from the (misspecified) fixed effects model:

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{\tau_i}_{\text{fixed}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

- $\tau_1, \tau_2, \ldots, \tau_t$ unknown model parameters
- $E_{11}, \ldots, E_{tn} \overset{iid}{\sim} N(0, \sigma^2)$

We're not trying to *estimate* linear combos of fixed effects such as $\mu + \tau_1$. Instead, we care about the population from which T_1 was sampled, which is $N(0, \sigma_T^2)$.

One-way random effects model continued

Exercise: Using the random effects model, specify

$$E(Y_{ij}) \text{ and } \mathrm{Var}(Y_{ij})$$

- Two *components* to variability in data: σ^2, σ_T^2
- T_1, T_2, T_3, T_4, T_5 a ____________________________ of sire effects
- Sire effects is a population in its own right.

Contrast this situation with the binding fractions. Why not model antibiotic effects as random? Why fixed?

Model parameters: $\sigma^2, \sigma_T^2, \mu$

Sums of squares, mean squares - same as in fixed effects ANOVA:

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

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

$$SS[Tot] = \sum \sum (y_{ij} - \bar{y}_{..})^2$$

The ANOVA table is almost the same, it just has a different expected mean squares column:

Source	SS	df	MS	Expected MS
Treatment	$SS[T]$	$t-1$	$MS[Trt]$	$\sigma^2 + n\sigma_T^2$
Error	$SS[E]$	$N-t$	$MS[E]$	σ^2
Total	$SS[Tot]$	$N-1$		

BTW, if $H_0 : \sigma_T^2 = 0$, what is $E(MS(Trt))/E(MS(E))$ ____________

Estimating parameters of one-way random effects model

(Solve a linear system of *estimating equations* obtained by equating statistics to their expected values and solving for unknown parameters:)

$$\begin{aligned} E(\hat{\mu}) &= \\ E(MS(T)) &= \\ E(MS(E)) &= \end{aligned}$$

leading to the solution

$$\begin{aligned} \widehat{\mu} &= \\ \widehat{\sigma}^2 &= \\ \widehat{\sigma}^2_T &= \end{aligned}$$

We've derived these estimators:

$$\begin{aligned}
\widehat{\mu} &= \overline{y}_{..} \\
\widehat{\sigma}^2 &= MS[E] \\
\widehat{\sigma}_T^2 &= \frac{MS[T] - MS[E]}{n}
\end{aligned}$$

For sires data, we observed $\overline{y}_{..} = 82.6$ and

Source	SS	df	MS	Expected MS
Sire	5591	4	1398	$\sigma^2 + 8\sigma_T^2$
Error	16233	35	464	σ^2
Total	21824	39		

leading to the observed estimates

$$\begin{aligned}
\widehat{\mu} &= \underline{\hspace{4cm}} (lbs) \\
\widehat{\sigma}^2 &= \underline{\hspace{4cm}}\ (lbs^2) \\
\widehat{\sigma}_T^2 &= \\
&= \underline{\hspace{4cm}}\ (lbs^2)
\end{aligned}$$

Questions pertaining to this type of study:

Consider the birthweight of a randomly sampled calf.

1. What is the estimated variance of such a calf?
2. Estimate how much of this variation is due to the sire effect.
3. Estimate how much of this variation is not due to the sire effect.

General questions:

1. Is it possible for an estimated variance component to be negative?
2. How?
3. What do you do in that case?

$$
\begin{aligned}
\text{Var}(Y_{ij}) &= \\
\widehat{\text{Var}}(Y_{ij}) &= \\
\text{Var}(T_i)/\text{Var}(Y_{ij}) &= \\
\text{Var}(E_{ij})/\text{Var}(Y_{ij}) &=
\end{aligned}
$$

1. Yes, it is possible for $\widehat{\sigma}^2_T < 0$ even though $\sigma^2_T \geq 0$.
2. $\widehat{\sigma}^2_T < 0 \Leftrightarrow$ ________________________?
3. Inference concerning σ^2_T? ________________________

Other parameters of interest in random effects models

$\boxed{\text{Coefficient of variation}}$ (CV):

$$CV(Y_{ij}) = \frac{\sqrt{\text{Var}(Y_{ij})}}{|E(Y_{ij}|)} = ?$$

Note: this is *not* estimated by `Coeff Var` in PROC GLM output.

$\boxed{\text{Intraclass correlation coefficient}}$

$$\rho_I = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ik})}} = \underline{\hspace{6cm}}$$

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of ρ_I correspond to (bigger/smaller?) random treatment effects.
- Answers questions like: How much of this variation is due to the sire effect?

For sires,

$$\widehat{CV} = \qquad = 0.29$$
$$\hat{\rho}_I = \qquad = 0.20$$

Interpretations:

- The estimated standard deviation of a birthweight, 24.1 is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20

Using PROC GLM for random effects models

```
data one;
   input sire @;
   do i=1 to 8;
      input bw @; output;
   end;
   cards;
177 61 100 56 113 99 103 75 62
200 75 102 95 103 98 115 98 94
201 58 60 60 57 57 59 54 100
202 57 56 67 59 58 121 101 101
203 59 46 120 115 115 93 105 75
;
run;

proc glm data=one;   *PROC MIXED recommended;
   class sire;
   model bw=sire;
   random sire;
run;
```

```
The GLM Procedure

Class          Levels    Values
sire                5    177 200 201 202 203

                                 Sum of
Source               DF         Squares     Mean Square    F Value    Pr > F
Model                 4      5591.15000      1397.78750       3.01    0.0309
Error                35     16232.75000       463.79286
Corrected Total      39     21823.90000

R-Square     Coeff Var      Root MSE      bw Mean
0.256194      26.08825      21.53585     82.55000

Source               DF       Type I SS     Mean Square    F Value    Pr > F
sire                  4     5591.150000     1397.787500       3.01    0.0309

Source               DF     Type III SS     Mean Square    F Value    Pr > F
sire                  4     5591.150000     1397.787500       3.01    0.0309

Source                Type III Expected Mean Square
sire                  Var(Error) + 8 Var(sire)
```

$CV = \frac{\sqrt{\widehat{Var}(y_{ij})}}{|\hat{\mu}|} = \frac{\sqrt{\hat{\sigma}^2 + \hat{\sigma}_T^2}}{|\bar{Y}_{..}|}$

$100 \times \frac{21.5}{82.5}$

error: $Var(Error) = \sigma^2$

$= \sigma^2 + 8\sigma_\tau^2$

(σ^2 =Var(Error) and σ_T^2 =Var(sire).)

- Coeff Var different from coefficient of variation defined several slides ago.

Distributional results

$Y_{ij} = \mu + T_i + E_{ij}$ $E(Y_{ij}) = \mu$

$\bar{Y}_{i.} = \mu + T_i + \bar{E}_{i.}$ $E(\bar{Y}_{i.}) = \mu$

- $(t-1)\frac{MS[T]}{\sigma^2 + n\sigma_T^2} \sim ?$ χ^2_{t-1}
- $(N-t)\frac{MS[E]}{\sigma^2} \sim ?$ $\frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i.})^2}{\sigma^2} \sim \chi^2_{N-t}$

  $N = nt$
- Ratio of independent χ^2 RVs divided by df has an F distribution
- $\frac{\frac{MS[T]}{\sigma^2 + n\sigma_T^2}}{\frac{MS[E]}{\sigma^2}} \sim ?$ $F_{t-1,\,N-t}$

  conf int for $\frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$ = $P(F$

Testing a variance component - $H_0 : \sigma_T^2 = 0$

Recall that $\sigma_T^2 = \text{Var}(T_i)$, population variance of treatment effects.

$$F = \frac{MS[T]}{MS[E]} = \frac{\frac{MS(T)}{\sigma^2}}{\frac{MS(E)}{\sigma^2}} \sim F_{t-1, N-t}$$

reject H_0 at level α if $F > F(\alpha, t-1, N-t)$

For the sires,

$$F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)$$

so H_0 is rejected at $\alpha = 0.05$. (The p-value is 0.0309)

Note that this is the same as the F-test in the Fixed effects model

Interval Estimation of some model parameters

$Y_{ij} = \mu + T_i + E_{ij}$

A 95% confidence interval for μ derived by consideration of $T = (\overline{Y}_{\cdot\cdot} - \mu)/\widehat{SE}(\overline{Y}_{\cdot\cdot})$:

$$
\begin{aligned}
\overline{Y}_{\cdot\cdot} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij} = \frac{1}{n \cdot t}\sum_{i=1}^{t}\sum_{j=1}^{n}(\mu + T_i + E_{ij}) \\
&= \mu + \frac{1}{nt}\sum_{i=1}^{t} n T_i + \frac{1}{N}\sum\sum E_{ij} \\
&= \mu + \overline{T}_{\cdot} + \overline{E}_{\cdot\cdot}
\end{aligned}
$$

where $\bar{T}_{\cdot} = (T_1 + \cdots + T_t)/t$ and $\bar{E}_{\cdot\cdot} = (\sum\sum E_{ij})/N$,

$$
\begin{aligned}
\text{Var}(\overline{Y}_{\cdot\cdot}) &= \text{Var}(\bar{T} + \bar{E}_{\cdot\cdot}) = Var(\bar{T}) + Var(\bar{E}_{\cdot\cdot}) \\
&= \frac{\sigma_T^2}{t} + \frac{\sigma^2}{nt} \\
&= \frac{1}{nt}(n\sigma_T^2 + \sigma^2) \\
&= \frac{1}{nt}(n * var(sire) + var(Error))
\end{aligned}
$$

$$
\widehat{Var}(\bar{Y}_{\cdot\cdot}) = \frac{1}{nt} MS(T) = \frac{1}{nt}(n\hat{\sigma}_T^2 + \hat{\sigma}^2)
$$

(show this for exercise)

Confidence interval for μ, continued

If the data are normally distributed, then

$$T = \frac{\overline{Y}_{..} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim ?$$

$$\frac{\dfrac{\overline{Y}_{..} - \mu}{\sqrt{\frac{1}{nt}(n\sigma_T^2 + \sigma^2)}}}{\sqrt{\dfrac{MS(T)}{(n\sigma_T^2+\sigma^2)}}} \stackrel{d}{=} \frac{Z}{\sqrt{\frac{\chi^2}{t-1}}}$$

$$\sim t_{t-1}$$

$$1-\alpha = P\left(-t_{\alpha/2,t-1} < \frac{\overline{Y}_{..} - \mu}{\sqrt{\frac{MS(T)}{nt}}} < t_{\alpha/2,t-1}\right)$$

and a 95% confidence interval for μ given by . .

$$\overline{Y}_{..} \pm t_{\alpha/2} \cdot \sqrt{\frac{MS(T)}{nt}}$$

$$\boxed{\overline{Y}_{..} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}}$$

Sires data: $\overline{y}_{..} = 82.6$, $MS[T] = 1398$, $nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the interval

$$82.6 \pm 2.78(5.91) \quad \text{or} \quad (66.1, 99.0).$$

captures μ w/ high conf.

Confidence interval for ρ_I

A 95% confidence interval for ρ_I can be obtained from the expression

$$\frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} < \rho_I < \frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}$$

where $F_{\alpha/2} = F(\frac{\alpha}{2}, t-1, N-t)$ and F_{obs} is the observed F-ratio for treatment effect from the ANOVA table. right tail

left tail (2.5th percentile)

For the sires, $F_{obs} = 3.01$ and $F_{0.025} = 3.179, F_{0.975} = 0.119$. The formula gives $(-0.01, 0.75)$.

97.5th percentile

Note the asymmetry and disagreement with test of $H_0 : \sigma^2_T = 0$
Derivation: Rearranging the probability statement below

$$1 - \alpha = \Pr\left(F(1-\frac{\alpha}{2}, t-1, N-t) < \frac{\frac{MS[T]}{\sigma^2 + n\sigma^2_T}}{\frac{MS[E]}{\sigma^2}} < F(\frac{\alpha}{2}, t-1, N-t) \right)$$

so that ρ_I gets left in the middle yields the confidence interval yields the c.i. at the top o' the page.

Using PROC MIXED for random effects models

```
proc mixed cl;
   class sire;
   model bw=;
   random sire;
   estimate "mean" intercept 1/cl;
run;
```

fixed effects

random effects

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

     Dependent Variable              bw
     Covariance Structure            Variance Components
     Estimation Method               REML
     Residual Variance Method        Profile
     Fixed Effects SE Method         Model-Based
     Degrees of Freedom Method       Containment

        Class     Levels     Values
        sire           5     177 200 201 202 203
```

Covariance Parameter Estimates

	Cov Parm		Estimate	Alpha	Lower	Upper
σ_T^2	sire	$\hat{\sigma}_T^2$	116.75	0.05	29.9707	7051.37
σ^2	Residual	$\hat{\sigma}^2 =$	463.79	0.05	305.11	789.17

$\left(\frac{SS(E)}{\chi^2_{Hi}}, \frac{SS(E)}{\chi^2_{lo}} \right)$

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
mean	82.5500	5.9114	4	13.96	0.0002	0.05
	$\bar{Y}_{..}$	$\sqrt{\frac{1}{40} MS(sire)}$				

Estimates

Label	Lower	Upper
mean	66.1373	98.9627

$82.6 \pm 2.78 \sqrt{\frac{1}{40} MS(sire)}$

More interval estimation for variance components

The estimated residual variance component for the sire data was $\hat{\sigma}^2 = MS[E] = 464\ lbs^2$.

A 95% confidence interval for this variance component is given by

$$\left(\frac{(40-5)464}{53.2} < \sigma^2 < \frac{(40-5)464}{20.6} \right)$$

or

$$\left(\frac{35}{53.2} 464 < \sigma^2 < \frac{35}{20.6} 464 \right)$$

or $(305.2, 789.5) lbs^2$

(Derivation outlined next slide)

This can be derived using the distributional result — Pivotal Qty

$$\frac{\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2}{\sigma^2} = \quad (N-t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t} \qquad P\left(\frac{SS(E)}{\chi^2_{hi}} < \sigma^2 < \frac{SS(E)}{\chi^2_{lo}}\right)$$

setting up the probability statement $= P\left(\chi^2_{lo} < \frac{SS(E)}{\sigma^2} < \chi^2_{hi}\right)$

$$1-\alpha = \Pr\left(\chi^2(1-\frac{\alpha}{2}, N-t) < (N-t)\frac{MS[E]}{\sigma^2} < \chi^2(\frac{\alpha}{2}, N-t)\right)$$

Rearranging to get σ^2 in the middle yields the $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}}\right).$$

Q: What are the mean and variance of the χ^2_{35} distribution?

Interval estimation for σ^2_T

The estimated variance component for the random sire effect was $\hat{\sigma}^2_T = 117.$ $= \frac{MS(T)-MS(E)}{n} = c_1 MS(T) + c_2 MS(E) = c_1 MS_1 + c_2 MS_2$

Q: How can we get a 95% confidence interval for σ^2_T?

A: In a similar fashion, but the confidence level based on Satterthwaite's approximation to the degrees of freedom of the linear combination of *MS* terms:

$$\left(\frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{1-\alpha/2,\widehat{df}}} \right)$$

where

$$\widehat{df} = \frac{(n\hat{\sigma}^2_T)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}} = \frac{(\hat{\sigma}^2_T)^2}{\sum \frac{c_i^2 MS_i^2}{df_i}}$$

For the sire data,

$$\widehat{df} = \frac{(8 \times 117)^2}{\frac{1398^2}{4} + \frac{464^2}{35}} = 1.76$$

Using the `CL` option in the `MIXED` statement will request this confidence interval and will use this approximation to *df* and will not round to the nearest integer *df*:

$$\chi^2_{0.975,1.76} = 0.029, \quad \chi^2_{0.025,1.76} = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87} \; \frac{1.76(117)}{0.029}\right) = \left(\frac{\hat{df}\,\hat{\sigma}_\tau^2}{\chi^2_{hi}}, \frac{\hat{df}\,\hat{\sigma}_\tau^2}{\chi^2_{lo}}\right)$$

or

$$(30, 7051)$$

Review of one-way random effects ANOVA
The model

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}} \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

$$T_1, T_2, \ldots, T_t \overset{iid}{\sim} N(0, \sigma_T^2) \;\; \text{independent of} \;\; E_{11}, \ldots, E_{tn} \overset{iid}{\sim} N(0, \sigma^2)$$

a random sample from a pop'n of levels

Remarks:

- (T_1, $T_2, \ldots$ randomly drawn from pop'n of treatment effects.)
- Only three parameters: μ, σ, σ_T^2
- Several functions of these parameters of interest
  - $CV(Y) = \frac{\sqrt{\sigma^2+\sigma_T^2}}{\mu}$ $\quad \frac{\sigma}{|\mu|}$
  - $\rho_I = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2+\sigma_T^2}$
- Two observations from same treatment group not independent

Exercise: match up the formulas for confidence intervals below with their targets, $\rho_I, \sigma^2, \sigma_T^2, \mu$: param

$$\overline{Y}_{\cdot\cdot} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

μ

$$\left(\frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}} \quad , \quad \frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}}\right)$$

ρ_I

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}} \quad , \quad \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}}\right)$$

σ^2

$$\left(\frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{\alpha/2,\widehat{df}}} \quad , \quad \left(\frac{\widehat{df}\,\hat{\sigma}^2_T}{\chi^2_{1-\alpha/2,\widehat{df}}}\right)\right)$$

σ^2_T

A guide to modelling factorial effects: fixed, or random?

	Random	Fixed
Levels		
- selected from conceptually ∞ popn of collection of levels	X	
- finite number of possible levels		X
Another expt		
- would use same levels		X
- would involve new levels sampled from same popn	X	
Goal		
- estimate varcomps	X	
- estimate longrun means		X
Inference		
- for these levels used in this expt		X
- for the popn of levels	X	

Artificial sires