IDENTIFYING AN APPROPRIATE MODEL

- Given a description of a study, how do you construct an appropriate model?
- Context: more than one size of e.u.
- A made-up example, intended to be complicated
  (but far from being the most complicated I've seen)
- A study of the impact of micro-credit schemes in Ghana (West Africa). Micro-credit is the idea that by providing small loans, you can stimulate economic activity and bring people out of poverty and malnutrition. There are various ways this can be implemented, as giving loans, as giving loans and nutrition education, or by just giving nutrition education. Ghana has a mix of urban, suburban, and rural areas. Rainfall decreases from the coast (tropical rain forest) to semi-desert on the northern border with Burkina-Faso. Both are important blocking variables, so a Latin Square design was used.

- Ghana was divided into four rainfall zones; in each zone an urban, a near-suburban, a far-suburban, and a rural area were identified. The three treatments and a placebo were randomly assigned to areas subject to Latin Square constraints:

	Rainfall zone			
Area	Wet	Wet-mid	Mid	Dry
Urban	Loan	None	L+N	Nutr.
Near Suburban	None	L+N	Nutr.	Loan
Far Suburban	L+N	Nutr.	Loan	None
Rural	Nutr.	Loan	None	L+N

- These 16 areas are still quite large. Each area was divided into 3.
  - Within each 1/3'rd, one village or neighborhood was randomly chosen as the focal group. That group received the experimental treatment.
  - An adjacent village/neighborhood within the 1/3'rd received no treatment (control).
  - A somewhat distant village/neighborhood (still within the 1/3'rd) was also studied.

- The response of interest is a measure of food security
  - good: have enough food
  - bad: ate less than one meal per day.
  - continuous scale derived from 15 answers to a questionaire
- This is measured on randomly chosen families within each village/neighborhood.
- Questions concern:
  - Differences between types of assistance (none, L, N, L+N)
  - Differences between focal, control and distant control villages.
  - Is the difference between focal and control similar for all three types of assistance?
  - Differences between types of assistance in focal villages
- What's an appropriate model?

My approach: experimental study

- What are the e.u.'s? Presumably more than one.
- Are they nested? Following assumes they are.
- Start with the largest e.u. (the main plot)
  - What treatment(s) are randomly assigned to a main plot?
  - What experimental design was used?
  - Imagine **only one** observation per main plot e.u.
  - This could be the average over all split plots or just one specific split plot treatment.
  - Write out the model / ANOVA table for the main plots.
- Main plot is the area of Ghana (16 of them)
  - Loan and Nutr. assigned to area, 2 way factorial
  - Expt. design is a Latin Square
- $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl}$,
  $i \in 1, 2, 3, 4, j \in 1, 2, 3, 4, k \in 1, 2, l \in 1, 2$
  α_i: row blocks, β_j: column blocks, γ_k: None/Nutr., λ_l: None/Loan,

ANOVA table for main plots

Source	df	
Rainfall	3	
Urban	3	
Trt	3	
Loan		1
Nutr.		1
Loan*Nutr.		1
Error	6	
c.tot	15	

- Now consider the smaller e.u. = split plot
- These are randomized inside mini-blocks that are the areas fit by the main plot model.
- 9 villages per area.
- Randomly assigned to 3 treatments (focal, near, distant)
- in blocks of three villages
- Model for **one area,** i, j, k, l: $Y_{ijklmn} = \mu_{ijkl} + \nu_{ijklm} + \tau_n + \varepsilon_{ijklmn}$
  μ_{ijkl}: mean for area *ijlk*, ν_{ijlkm}: block *m* in area *ijkl*, τ_n: focal/near/distant

ANOVA for split plots in one main

Source	df
Block	2
Village	2
Error	4
c.tot	8

- There are 16 of these. Have different block effects but the same Village effect (ignoring interaction for now). Have same error variance, so will pool Error variances across the main plots

ANOVA for all split plots

Source	df
Area	15
Block(Area)	32
Village	2
Error	64+30
c.tot	143

- Error has two components:
  - Pooled error variances from the 16 “one area” models = 4*16
  - Interaction between area and village = 15*2

ANOVA for entire study

Source	df		
Rainfall	3		
Urban	3		
Trt	3		
Loan		1	
Nutr.		1	
Loan*Nutr.		1	
Trt*Rain*Urban	6		Main plot error
Block(Trt*Rain*Urban)	32		Forced to be random
Village	2		
Loan*Vill.	2		
Nutr.*Vill.	2		
L*N*V	2		
Error	64+30 - 6		Split plot error
c.tot	143		

Model for entire study

- The two pieces:
  - Main plots:
    $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl}$
  - Split plots for one area:
    $Y_{ijklmn} = \mu_{ijkl} + \nu_{ijklm} + \tau_n + \varepsilon_{ijklmn}$
- Notice that first equation is μ_{ijkl} in the second.

$$\begin{aligned} Y_{ijklmn} &= \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl} + \\ &\quad \nu_{ijklm} + \tau_n + \varepsilon_{ijklmn} \end{aligned}$$

- Treatment structure includes interactions between split treatment and main treatments. Add these to model.

$$\begin{aligned} Y_{ijklmn} &= \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl} + \\ &\quad \nu_{ijklm} + \tau_n + \gamma\tau_{kn} + \lambda\tau_{ln} + \gamma\lambda\tau_{kln} + \varepsilon_{ijklmn} \end{aligned}$$

Another way to look at the model

- The model:

$$\begin{aligned} Y_{ijklmn} &= \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl} + \\ &\quad \nu_{ijklm} + \tau_n + \gamma\tau_{kn} + \lambda\tau_{ln} + \gamma\lambda\tau_{kln} + \varepsilon_{ijklmn} \end{aligned}$$

- has two parts:
  - The treatment structure:

    $$\mu + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \tau_n + \gamma\tau_{kn} + \lambda\tau_{ln} + \gamma\lambda\tau_{kln}$$

    you should recognize this as a 3 way factorial
  - and the error structure:

    $$\alpha_i + \beta_j + \epsilon_{ijkl} + \nu_{ijklm} + \varepsilon_{ijklmn}$$

  - Blocks ν_{ijklm} are nested in main plots $\alpha_i + \beta_j + \epsilon_{ijkl}$
  - Observations (villages) ε_{ijklmn} are nested in blocks.

Choice of fixed or random

- All treatment effects are fixed
- Main plot error and split plot error are random
- blocks are forced to be random because they are nested in main plot errors
- α_i and β_j are the Latin Square blocking factors
- could be fixed or random.
- I usually treat LS factors as fixed
  (if random, what population has crossed blocks?)
- And, how do you draw “new” plots from the same population of crossed row and column effects?

More possibilities for models

- Process just described focuses on means of normally distributed random variables, while accounting for potential correlation or additional sources of variation.
- What if normal with constant variances is not appropriate?
- What if the mean varies among areas, and the variability among villages depends on the treatment applied to the area?

$$\begin{aligned}
\text{Area means: } \mu_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl} + \epsilon_{ijkl} \\
\text{Village means: } Y_{ijklmn} \mid \mu_{ijkl} &= \mu_{ijkl} + \nu_{ijklm} + \tau_n + \gamma\tau_{kn} + \\
&\qquad \lambda\tau_{ln} + \gamma\lambda\tau_{kln} + \varepsilon_{ijklmn} \\
\text{Village var: Var } \varepsilon_{ijklmn} &= \exp(\gamma_k + \lambda_l + \gamma\lambda_{kl})
\end{aligned}$$

- I.e., the variability among villages in an area depends on the treatment (loan/nutrition education) applied to that area.

What if normal is not appropriate?

- E.g., Ghana study with response being the number of days did not eat in the last 30 days.

Areas:

$$\begin{aligned} \eta_{ijkl} &\sim \beta(a, b) \\ a + b &= t \text{ (determines the variance)} \\ a &= t(\mu + \alpha_i + \beta_j + \gamma_k + \lambda_l + \gamma\lambda_{kl}) \\ &\quad \text{(determines the mean)} \end{aligned}$$

Villages:

$$\log \frac{\pi_{ijklmn}}{1 - \pi_{ijklmn}} \mid \eta_{ijkl} = \log \frac{\eta_{ijkl}}{1 - \eta_{ijkl}} + \nu_{ijklm} + \tau_n$$

Obs.:

$$Y_{ijklmn} \sim \text{Bin}(30, \pi_{ijklmn})$$

- Details absolutely not important.
- Crucial concept is that the thought process described here is not restricted to means of normal distributions.
- Even though old-fogey's like me often behave as if it is.

Observational studies

- All of the above was possible because of the way treatments were randomized to eu's.
- What if no randomization?
- Much harder to justify a specific model
- May be multiple reasonable models, each corresponding to a particular view of the study
- My approach:
    - If a factor were randomly assigned, what would it have been assigned to?
    - i.e. reconstruct a reasonable assignment scheme even though observational groups not assigned to treatments by the investigator.

- Example: A study of the impact of “welfare reform”. States may or may not adopt a package of policies called “welfare reform”. Data are collected on individuals categorized by race/ethnic group and where they live (rural, suburban, urban).
- Decide to use a 3 way treatment structure adopt$\times$race$\times$location
- What are the random effects: 3 of the many possibilities
  1. Individuals “assigned” to adopt, race, urban
     All individuals independent: one error term
  2. States “assigned” to adopt
     individuals to race, urban
     Two random effects: state(adopt), individual(state, adopt)
  3. States “assigned” to adopt
     counties “assigned” to urban, suburban, rural
     individuals “assigned” to race
     Three random effects: state(adopt), county(state, adopt), individual(county, state, adopt)

- Which is the more appropriate?

- I generally consider 1) wrong, because states likely to differ
  That is a subject-matter reason, not a statistical reason
- Deciding between 2) and 3): Not clear (to me).
- Are there subject-matter reasons that:
  - Counties receiving same treatment will have different means
  - Or equivalently, individuals correlated within a county
- If subject-matter not clear, can use the data:
  - if there is a positive correlation among individuals within a county, need 3)
  - if individuals independent within states, 2) adequate
  - if data suggests no correlation of indidividuals in same state, 1) may be appropriate
- Can use data to evaluate “fit” of various correlation models using REML-AIC/BIC

REML ESTIMATION IN THE GENERAL LINEAR MODEL

Reminder of some comments made earlier:

- The MLE of the variance component vector γ is often biased.
- For example, for the case of $\epsilon = \sigma^2$ I, where $\gamma = \sigma^2$, the MLE of σ^2 is $\frac{(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})}{n}$ with expectation $\frac{n-p}{n}\sigma^2$.
- The MLE of σ^2 is often criticized for "failing to account for the loss of degrees of freedom needed to estimate $\boldsymbol{\beta}$. "

$$\begin{aligned} E[\frac{(\boldsymbol{y}-\boldsymbol{X}\beta)'(\boldsymbol{y}-\boldsymbol{X}\beta)}{n}] &= \sigma^2 \\ \text{but, } E[\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})'(\boldsymbol{y}-\boldsymbol{X}\hat{\beta})}{n}] &= \frac{n-p}{n}\sigma^2 < \sigma^2 \end{aligned}$$

- REML is an approach that produces unbiased estimators for many special cases and produces less biased estimates than ML in general.
- Depending on who you ask, REML stands for REsidual Maximun Likelihood or REstricted Maximum Likelihood.
- The REML method:
  1. Find $m \equiv$ n - rank(x) linearly independent vectors $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$ such that $\boldsymbol{a}_i'X = \mathbf{0}'$ for all $i = 1, ..., m$.
  2. Find the maximum likelihood estimate of γ using $w_1 = \boldsymbol{a}_1'\boldsymbol{y}, ..., w_m \equiv \boldsymbol{a}_m'\boldsymbol{y}$ as data.

$$A = [\boldsymbol{a}_1, ..., \boldsymbol{a}_m] \qquad \boldsymbol{w} = A'\boldsymbol{y} = \begin{bmatrix} \boldsymbol{a}_1\boldsymbol{y} \\ \vdots \\ \boldsymbol{a}_m\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

- If $\boldsymbol{a}'\boldsymbol{X} = \mathbf{0}'$, $\boldsymbol{a}'\boldsymbol{y}$ is called an error contrast.
- Thus, $w_1, ..., w_m$ are a set of m error contrasts.
- Because $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{0}$, the elements of $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ are each error contrasts.
- Because rank $(\boldsymbol{I} - \boldsymbol{P_X}) = n-$ rank $(x) = m$, there exists a set of m linearly independent rows of $\boldsymbol{I} - \boldsymbol{P_X}$ that can be used in step 1 of the REML method to get $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$
- If we do use a subset of rows of $\boldsymbol{I} - \boldsymbol{P_X}$ to get $\boldsymbol{a}_1, ..., \boldsymbol{a}_m$; the error contrasts $w_1 = \boldsymbol{a}_1'\boldsymbol{y}, ..., w_m = \boldsymbol{a}_m'\boldsymbol{y}$ will be a subset of the elements of $(I - P_x)\boldsymbol{y} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, the residual vector.
- This is why it makes sense to call the procedure Residual Maximum Likelihood.

- Note that

$$\begin{aligned}\boldsymbol{w} &= A'\boldsymbol{y} \\ &= A'(X\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= A'X\boldsymbol{\beta}+A'\boldsymbol{\epsilon} \\ &= \mathbf{0}+A'\boldsymbol{\epsilon} \\ &= A'\boldsymbol{\epsilon}\end{aligned}$$

- Thus $\boldsymbol{w} = A'\boldsymbol{\epsilon} \sim N(A'\mathbf{0},\ A'\boldsymbol{\Sigma}A) \overset{d}{=} N(\mathbf{0},\ A'\boldsymbol{\Sigma}A)$ and the distribution of $\boldsymbol{w}$ depends on γ but not $\boldsymbol{\beta}$.
- The log likelihood function in this case is

$$l(\gamma|\boldsymbol{w}) = -\frac{1}{2}\log|A'\boldsymbol{\Sigma}A| - \frac{1}{2}w'(A'\boldsymbol{\Sigma}A)^{-1}\boldsymbol{w} - \frac{m}{2}\log(2\pi)$$

- Find $\hat{\gamma}$ using by numerically maximizing the REML lnL(γ).
- Choice of m error contrasts is arbitrary.
- Can prove that every set of m linearly independent error contrasts yields the same REML estimator of γ. (611)

- For example, suppose $y_1, ..., y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$.
Then $X = \mathbf{1}$ and $\mathbf{\Sigma} = \sigma^2 I$. It follows that

$$\begin{array}{lcl} \boldsymbol{a}_1' = (1, -1, 0, 0, ..., 0) & & \boldsymbol{b}_1' = (1, 0, ..., 0, -1) \\ \boldsymbol{a}_2' = (0, 1, -1, 0, ..., 0) & \text{and} & \boldsymbol{b}_2' = (0, 1, 0, ..., 0, -1) \\ \vdots & & \vdots \\ \boldsymbol{a}_{n-1} = (0, ..., 0, -1, 1) & & \boldsymbol{b}_{n-1} = (0, ..., 0, 1, -1) \end{array}$$

are each a set of $m = n - 1 = n - \text{rank}(\boldsymbol{X})$ linear independent vectors that can be used to form error contrasts. Either

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{y} \\ \boldsymbol{a}_2'\boldsymbol{y} \\ \vdots \\ \boldsymbol{a}_{n-1}'\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} y_1 - y_2 \\ y_2 - y_3 \\ \vdots \\ y_{n-1} - y_n \end{bmatrix} \text{ or } \boldsymbol{v} = \begin{bmatrix} \boldsymbol{b}_1'\boldsymbol{y} \\ \boldsymbol{b}_2'\boldsymbol{y} \\ \vdots \\ \boldsymbol{b}_{n-1}'\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} y_1 - y_n \\ y_2 - y_n \\ \vdots \\ y_{n-1} - y_n \end{bmatrix}$$

could be used to obtain the same REML estimator of σ^2

- For the normal theory Gauss-Markov linear model, $\boldsymbol{y} = X\beta + \epsilon, \epsilon \sim N(\mathbf{0}, \sigma^2 I)$ the REML estimator of σ^2 is $\hat{\sigma}^2 = \frac{\boldsymbol{y}'(I - P_x)\boldsymbol{y}}{n - \text{rank}(x)}$, the commonly used unbiased estimator, s^2.
- Outline of REML-based inference:
  1. Use REML to estimate γ (and thus $\boldsymbol{\Sigma}$).
  2. Use estimated GLS, to estimate an estimable $\boldsymbol{C}\beta$ by $\boldsymbol{C}\hat{\beta}_g = \boldsymbol{C}(X'\hat{\boldsymbol{\Sigma}}^{-1}X)^{-}X'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}$
  3. Conditional on $\hat{\gamma}$, this is the BLUE of an estimable $\boldsymbol{C}\beta$.
  4. Estimate Var $\boldsymbol{C}\hat{\beta}$ by $\boldsymbol{C}(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-}\boldsymbol{C}'$
  5. Use
     $$T = \frac{\boldsymbol{C}\hat{\beta} - \boldsymbol{C}\beta}{\sqrt{\text{Var } \boldsymbol{C}\hat{\beta}}}$$
     for tests and confidence intervals
- But, what distribution? Not clear. Don't know d.f.

- For many mixed effect models (perhaps all), when the data are balanced and the ANOVA estimates of $\hat{\gamma}$ are positive, the REML estimates and the ANOVA estimates are exactly the same.
- REML estimates are forced to respect the parameter space, i.e. forced to be non-negative
- Hence, ANOVA and REML estimates are not same when one or more ANOVA estimates is negative.
- e.g. subsampling, σ_u^2 is the variance component for e.u.'s σ_e^2 is variance among measurements of the same e.u.
- In the ANOVA analysis, $\text{MS}_{eu} < \text{MS}_{ou}$, so $\hat{\sigma}_u^2 = \frac{MS_{eu} - MS_{ou}}{m} < 0$
- for which:
  ANOVA: $\hat{\sigma}_u^2 = -0.0321$, $\hat{\sigma}_e^2 = 0.856$,
  but REML: $\hat{\sigma}_u^2 = 0$, $\hat{\sigma}_e^2 = 0.827$
- Note: REML "fixes" the negative $\hat{\sigma}_u^2$, but because the estimates are not independent, that forces an decrease in $\hat{\sigma}_e^2$
- I believe $\hat{\sigma}_e^2$ is well-defined value: the variability among ou's in one eu. That is the ANOVA estimate. It is not the REML estimate.

- Should I use REML or ANOVA?
- Lots of different opinions.
- In favor of REML:
  - ANOVA is old fashioned
  - estimates are non-negative
  - easily extended to more complicated models, e.g. autoregressive errors, where no ANOVA estimates
  - and to non-normal data
  - so provides consistent framework for estimation and infererence
- In favor of ANOVA:
  - Estimates are unbiased
  - ANOVA estimate of σ_e^2 clearly connected to variability in data
  - REML estimate of σ_e^2 may be shifted to compensate for non-negative constraint on other parameters (e.g. 0.856 to 0.827 on previous page)
  - Matches inferences from eu averages, when averaging appropriate.

- What decides the issue for me is the consequences for the important inferences.
- For me, that's usually inference on estimable $\boldsymbol{C\beta}$.
- For designed experiments, i.e. known structure for $\boldsymbol{\Sigma}$, ANOVA estimation of γ leads to tests that are closer to nominal type-I error rate and intervals that are closer to nominal coverage
- Details, e.g. liberal or conservative tests, depend on design and test.
- σ_e^2 tends to be underestimated by REML because of the adjustment induced by the non-negative constraint

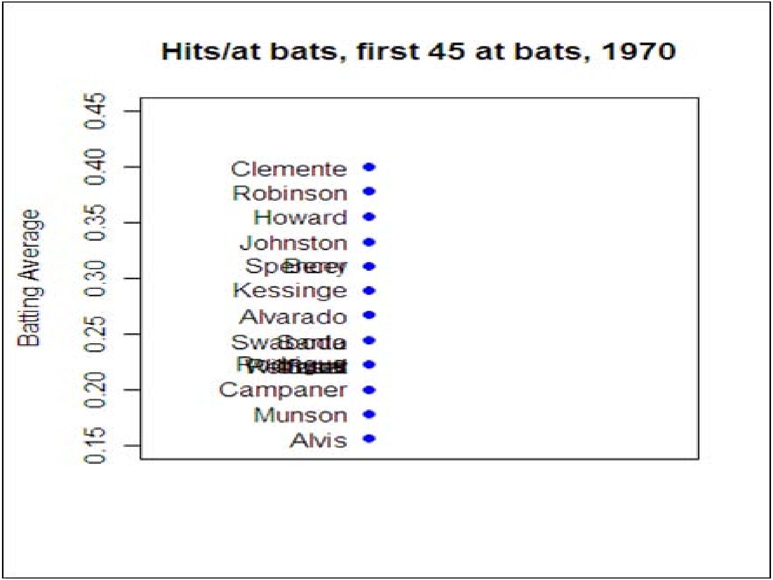
- Example: Two treatments, 5 eu's per trt, 5 subsamples per eu 1000 data sets

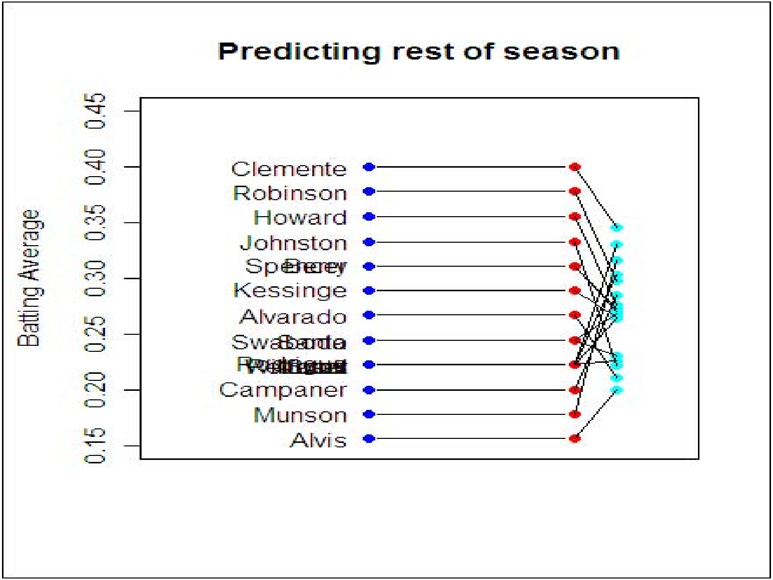
	empirical P[rej] for	
Estimator	$\alpha = 0.05$	$\alpha = 0.01$
REML	2.2%	0.5%
ML	2.6%	0.5%
ANOVA	5.7%	1.3%

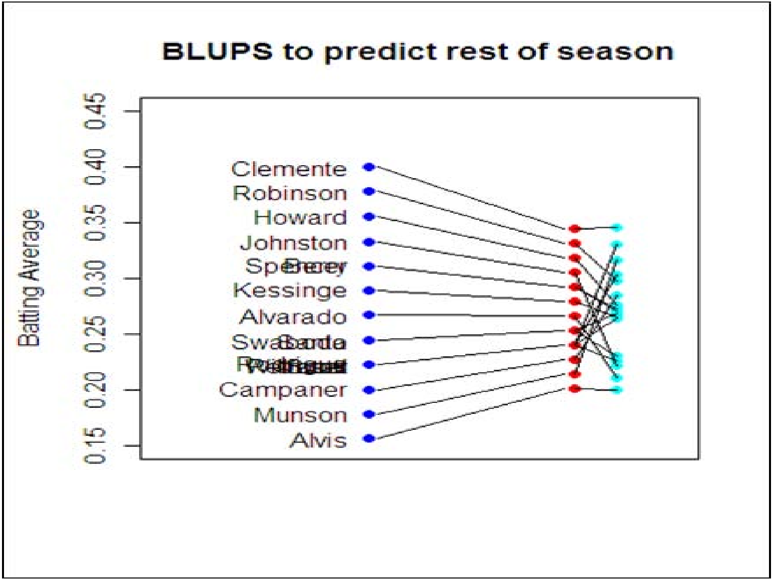
- More extensive evaluation in Stroup and Little, 2002, Impact of Variance Component Estimates on Fixed Effect Inference in Unbalanced Mixed Models, *Proceedings 14'th Annual Kansas State Conference in Applied Statistics in Agriculture*, pp 32-48.
- Summarized in Littell et al., SAS for Mixed Models, p 150.
- Many others simply prefer REML.
- I use REML when no alternative (e.g. autocorrelated errors)

Prediction of random variables

- Key distinction between fixed and random effects:
  - Estimate means of fixed effects
  - Estimate variance of random effects
- But in some instances, want to predict FUTURE values of a random effect
- Example (from Efron and Morris, 1975, JASA 70:311-319):
- Baseball players. Given a player's performance in the beginning of the season, predict performance in rest of season.


Hits/at bats, first 45 at bats, 1970
Batting Average
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Clemente
Robinson
Howard
Johnston
Spencer
Berry
Kessinge
Alvarado
Swaboda
Santo
Rodriguez
Petrocelli
Scott
Unser
Williams
Campaner
Munson
Alvis


Predicting rest of season
Batting Average
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Clemente
Robinson
Howard
Johnston
Berry
Kessinge
Alvarado
Santo
Campaner
Munson
Alvis


BLUPS to predict rest of season
Batting Average
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Clemente
Robinson
Howard
Johnston
Kessinge
Alvarado
Campaner
Munson
Alvis

- Best predictor is found by “Shrinking” obs. performance towards overall mean.
- So, how much shrinkage is needed? How do we compute optimal predictor?
- General answer using a linear mixed effects model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon$, where

$$\begin{bmatrix} \boldsymbol{u} \\ \epsilon \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{bmatrix} \right)$$

- Given data $\boldsymbol{y}$, what is our best guess for values in the unobserved vector $\boldsymbol{u}$?

- Because $\boldsymbol{u}$ is a random vector rather than a fixed parameter, we talk about predicting $\boldsymbol{u}$ rather than estimating $\boldsymbol{u}$.
- We seek a Best Linear Unbiased Predictor (BLUP) for $\boldsymbol{u}$, which we will denote by $\hat{\boldsymbol{u}}$
- To be a BLUP, we require
  1. $\hat{\boldsymbol{u}}$ is a linear function of $\boldsymbol{y}$.
  2. $\hat{\boldsymbol{u}}$ is unbiased for $\boldsymbol{u}$ so that $E(\hat{\boldsymbol{u}} - \boldsymbol{u}) = \mathbf{0}$.
  3. $Var(\hat{\boldsymbol{u}} - \boldsymbol{u})$ is no "larger" than $Var(\boldsymbol{v} - \boldsymbol{u})$, where $\boldsymbol{v}$ is any other linear and unbiased predictor.
- It turns out that the BLUP for $\boldsymbol{u}$ is the BLUE of $E(\boldsymbol{u}|\boldsymbol{y})$.

- What does $E(\boldsymbol{u}|\boldsymbol{y})$ look like?
- We will use the following result about conditional distributions for multivariate normal vectors.
  Suppose
  $$\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$$
  where $\boldsymbol{\Sigma} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ is a positive definite variance matrix.

  Then the conditional distribution of $\boldsymbol{w}_2$ given $\boldsymbol{w}_1$ is as follows
  $(\boldsymbol{w}_2|\boldsymbol{w}_1) \sim N(\mu_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{w}_1 - \mu_1), \ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$

Now note that

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}\beta \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \epsilon \end{bmatrix}$$

Thus,

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{X}\beta \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}' & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \right)$$

$$\stackrel{d}{=} N\left(\begin{bmatrix} \boldsymbol{X}\beta \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{ZGZ}' + \boldsymbol{R} & \boldsymbol{ZG} \\ \boldsymbol{GZ}' & \boldsymbol{G} \end{bmatrix} \right)$$

- Thus,

$$\begin{aligned} E(\boldsymbol{u}|\boldsymbol{y}) &= \mathbf{0} + \boldsymbol{GZ}'(\boldsymbol{ZGZ}' + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta) \end{aligned}$$

- Thus, the BLUP of $\boldsymbol{u}$ is

$$\begin{aligned} \boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_g) &= \boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}) \\ &= \boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})\boldsymbol{y} \end{aligned}$$

- For the usual case in which $\boldsymbol{G}$ and $\boldsymbol{\Sigma} = \boldsymbol{ZGZ}' + \boldsymbol{R}$ are unknown, we replace the matrices by estimates and approximate the BLUP of $\boldsymbol{u}$ by $\hat{G}\boldsymbol{Z}'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\hat{\boldsymbol{\Sigma}}})$

- Often we wish to make predictions of quantities like $\boldsymbol{C}\beta + \boldsymbol{d}\boldsymbol{u}$ for some estimable $\boldsymbol{C}\beta$.
- The BLUP of such a quantity is $\boldsymbol{C}\hat{\beta}_g + \boldsymbol{d}\hat{\boldsymbol{u}}$, the BLUE of $\boldsymbol{C}\beta$ plus $\boldsymbol{d}$ times the BLUP of $\boldsymbol{u}$

- Baseball players example is slightly complicated because quantities of interest are proportional to Binomial random variables.
- Simpler example, using Normal random variables:
  An old problem, a variation of a homework problem (# 23 on p. 164) of Mood, A.M. (1950) *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

- Suppose intelligence quotients (IQs) for a population of students are normally distributed with a mean μ and variance σ_u^2
- An IQ test was given to an i.i.d. sample of such students.
- Given the IQ of a student, the test score for that student is normally distributed with a mean equal to the student's IQ and a variance σ_e^2 and is independent of the test score of any other students.
- Suppose it is known that $\sigma_u^2/\sigma_e^2 = 9$
- If the sample mean of the students' test scores was 100, what is the best prediction of the IQ of a student who scored 130 on the test?

- Suppose $u_1, \ldots, u_n, \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ independent of $e_1, \ldots, e_n, \overset{i.i.d.}{\sim} N(0, \sigma_e^2)$.
- If we let $\mu + u_i$ denote the IQ of student $i (i = 1, ..., n)$, then the IQs of the students are $N(\mu, \sigma_u^2)$ as in the statement of the problem.
- If we let $y_i = \mu + u_i + e_i$ denote the test score of student $i (i = 1, \ldots, n)$, then $(y_i | \mu + u_i) \ \sim \ N(\mu + u_i, \sigma_e^2)$ as in the problem statement.
- We have $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $X = \mathbf{1}, \boldsymbol{\beta} = \mu, \boldsymbol{Z} = \boldsymbol{I}, \boldsymbol{G} = \sigma_u^2 \boldsymbol{I}, \boldsymbol{R} = \sigma_e^2 \boldsymbol{I}, \boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} = (\sigma_u^2 + \sigma_e^2)\boldsymbol{I}$.
- Thus $\hat{\beta}_g = (X'\boldsymbol{\Sigma}^{-1}X)^{-}X'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = (\mathbf{1}'\mathbf{1})^{-}\mathbf{1}'\boldsymbol{y} = \bar{y}_{.}$ and $\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \boldsymbol{I}$
- Thus, the BLUP for $\boldsymbol{u}$ is $\hat{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_g) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}(\boldsymbol{y} - \mathbf{1}\bar{y}_{.})$
- The i^{th} element of this vector is $\hat{u}_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}(y_i - \bar{y}_{.})$.

- Thus, the BLUP for the IQ of student $i, \mu + u_i$, is $\hat{\mu} + \hat{u}_i = \bar{y}_. + \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}(y_i - \bar{y}_.) = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} y_i + \frac{\sigma_e^2}{\sigma_u^2+\sigma_e^2} \bar{y}_.$
- Note that the BLUP is a convex combination of the individual score and the overall mean score $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} y_i + \frac{\sigma_e^2}{\sigma_u^2+\sigma_e^2} \bar{y}_.$
- Because $\frac{\sigma_u^2}{\sigma_e^2}$ is assumed to be 9, the weights are $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} = \frac{\frac{\sigma_u^2}{\sigma_e^2}}{\frac{\sigma_u^2}{\sigma_e^2}+1} = \frac{9}{9+1} = 0.9$ and $\frac{\sigma_e^2}{\sigma_u^2+\sigma_e^2} = 0.1$. We predict $0.9(130) + 0.1(100) = 127$ to be the IQ of a student who scored 130 on the test.
- US College test results (e.g. SAT) now include information about "If you take this test again, your score is predicted to be ..."
- If above average, predicted to drop from current score. I suspect these predictions are BLUPs.

- An extension that illustrates an important property of BLUP's
- IQ problem, except now, y_i is the average of n_i test scores for a student
- Some student's scores based on $n_i = 1$ test, others on $n_i = 5$ tests
- Now, $(y_i|\mu + u_i) \sim N(\mu + u_i, \sigma_e^2/n_i)$
- Now, $\hat{\beta}_g$ is a weighted average of student scores
- The BLUP for u_i is

$$\hat{u}_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_i}(y_i - \hat{\beta}_g)$$

- Again a convex combination of y_i and overall average, $\hat{\beta}_g$.
- But now weights are $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/n_i}$ and $\frac{\sigma_e^2/n_i}{\sigma_u^2+\sigma_e^2/n_i}$

- Numerical illustration of weights

	$\sigma^2_u/\sigma^2_e = 9$		$\sigma^2_u/\sigma^2_e = 1$	
Number of tests	y_i	$\hat{\beta}_g$	y_i	$\hat{\beta}_g$
1	0.9	0.1	0.5	0.5
2	0.947	0.053	0.667	0.333
3	0.964	0.036	0.75	0.25
4	0.973	0.027	0.80	0.2
10	0.989	0.011	0.909	0.0909

- More tests = more precise information about an individual: BLUP is closer to the data value, y_i
- Fewer tests = less precise information about an individual: BLUP is closer to the estimated population mean, $\hat{\beta}_g$
- More variability between individuals (larger σ^2_u relative to σ^2_e): BLUP is closer to the data value, y_i

A collection of potentially useful models

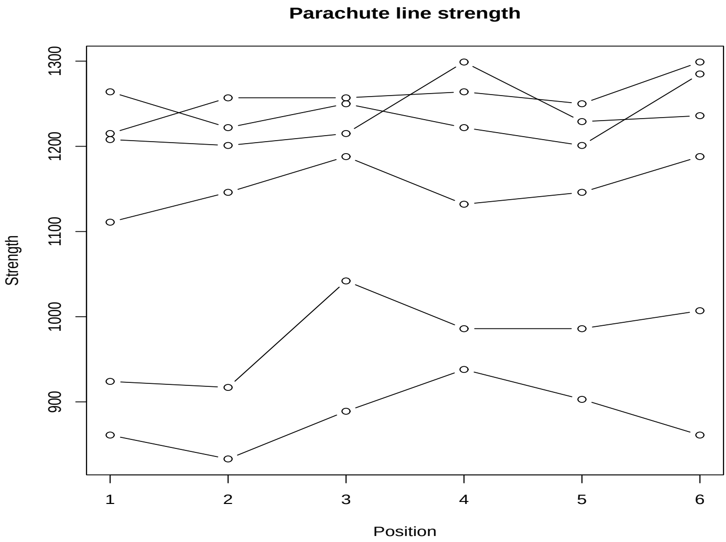
- We’ve already seen two very common mixed models:
  - for subsampling
  - for designed experiments with multiple experimental units
- Here are three more general classes of models
  - Random coefficient models, aka multi-level models
  - Models for repeated experiments
  - Models for repeated measures data
- For complicated problems, may need to combine ideas

Random coefficient models

- A regression where all coefficients vary between groups
- Example: Strength of parachute lines.
  - Measure strength of a parachute line at 6 positions
  - Physical reasons to believe that strength varies linearly with position
  - Model by $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where X is the position, y is the strength, and i indexes the measurement
- What if have 6 lines, each with 6 observations?
- Measurements nested in line
- suggests:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij} + \nu_j + \epsilon_{ij} \\ &= (\beta_0 + \nu_j) + \beta_1 X_{ij} + \epsilon_{ij}, \end{aligned}$$

  where j indexes the line.
- Intercept varies between lines, but slope does not



Parachute line strength
Strength
900
1000
1100
1200
1300
Position
1
2
3
4
5
6

- Random coefficient regression models allow slope to also vary

$$Y_{ij} = (\beta_0 + \alpha_{j0}) + (\beta_1 + \alpha_{j1})X_{ij} + \epsilon_{ij}$$

$$\boldsymbol{u}' = [\alpha_{10}, \alpha_{11}, \alpha_{20}, \alpha_{21}, \ldots, \alpha_{60}, \alpha_{61}]$$

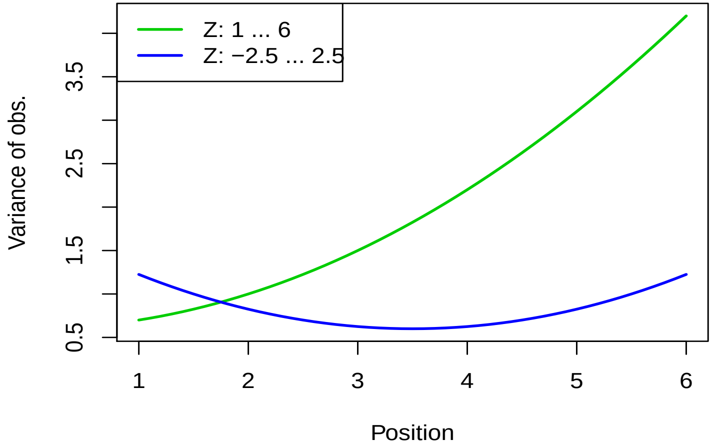
$$\boldsymbol{u} = \begin{bmatrix} 1 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 2 & 0 & 0 & \ldots & 0 \\ 1 & 3 & 0 & 0 & \ldots & 0 \\ 1 & 4 & 0 & 0 & \ldots & 0 \\ 1 & 5 & 0 & 0 & \ldots & 0 \\ 1 & 6 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 1 & 0\ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 1 & 6 & \end{bmatrix}_{36\times 12}$$

- $\begin{bmatrix} \alpha_{j0} \\ \alpha_{j1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{G} \right)$
- $\boldsymbol{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}$
- sometimes see model written as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij},$$

$$\begin{bmatrix} \beta_{j0} \\ \beta_{j1} \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{G} \right)$$

- $\boldsymbol{R}$ usually assumed $\sigma_e^2 \boldsymbol{I}$.
- $\boldsymbol{\Sigma} = \boldsymbol{ZGZ}' + \boldsymbol{R}$ is quite complicated, can write out but not enlightening. Features:
  - Var Y_{ij} not constant, depends on X_{ij}, even if $\boldsymbol{R}$ is $\sigma_e^2 \boldsymbol{I}$
  - Cov $Y_{ij}, Y_{i'j}$ not constant, depends on X_{ij} and $X_{i'j}$
  - Cov $Y_{ij}, Y_{i'j'} = 0$, since obs. on different lines assumed independent
  - so $\boldsymbol{\Sigma}$ is block diagonal, with non-zero blocks for obs. on same line


Z: 1 ... 6
Z: −2.5 ... 2.5
Variance of obs.
0.5
1.5
2.5
3.5
1
2
3
4
5
6
Position

- Customary to include a parameter for the covariance between intercept and slope random effects.
  - if omit, then model is not invariant to translation of X
  - i.e., fixed effect part of the regression is the same model even if shift X, e.g. $X - 3$.
  - random effects part is the same only if include the covariance
  - some parameter values will change if X shifted, but structure stays the same.
- Can extend model in at least two ways:
  1. More parameters in regression model
     e.g. quadratic polynomial: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \beta_{2j} X_{ij}^2 + \epsilon_{ij}$,
     - Example: Allan Trapp's MS. Longevity of stored seed, quadratic, 2833 seed lots of maize, each with 3 to 7 observations.
  2. More levels of random effects
     - 6 Measurements per line, 6 lines per parachute, 4 chutes
     - Measurements nested within Lines, Lines nested within Chutes

Repeated Experiments

- In some scientific fields, it is expected that you will repeat the entire experiment
- Plant Pathology, Agronomy. Often journals will not publish unless repeated.
- Concern is with erroneous effects that occur once but not repeatable.

- Repetitions like blocks, but have replication within each repetition.
- In the medical literature, often called multi-center clinical trials
- Example: Systolic Blood Pressure Intervention Trial
  - Two treatments: control: medication to maintain systolic BP < 140, and treatment: medication to reduce systolic BP < 120.
  - Randomly assigned to individuals.
  - Approx. 7500 individuals spread across 80 clinical centers.

Analysis of multicenter trial

- Possible ways to analyze these data
  1. Average subjects within treatment and center. Analyze 160 averages, RCBD
Source	df
Center	79
Treatment	1
Error	79
     - Note important point: Error variance is the center*treatment interaction
     - Measures the consistency of treatment effects across centers

Analysis of multicenter trial

2. Individual subjects (total = 7520) with centers as blocks

Source	df
Center	79
Treatment	1
Error	7439

    - Note: Error variance is mostly the variability among subjects within a center
    - Very different from previous analysis
    - We'll see why "mostly" shortly

Analysis of multicenter trial

3. A more careful version of the previous analysis
Source	d.f.
Center	79
Treatment	1
C*T	79
Error	7360
   - Now Error really is pooled variability among individuals in a center
   - Analysis 2 pooled C*T and error
   - Can estimate both because there is replication within a center.
   - Treatment is clearly fixed.
   - Center may be fixed or random - no effect if balanced.
   - Is C*T random or fixed? This really matters.

Analysis of multicenter trial

- Skeleton ANOVA corresponds to the effects model:

$$Y_{ijk} = \mu + \alpha_i + \tau_j + \alpha\tau_{ij} + \epsilon_{ijk},$$

  where *i* indexes centers, *j* indexes treatments, and *k* indexes subjects.
- All three analyses are reasonable.
- Note: Analyses 1 and 2 answer subtly different questions.
- Differ in what the treatment difference is compared to:
  1. Is the treatment effect large relative to its repeatability (consistency) across centers?
  2. Is the treatment effect large relative to the variability among individuals in these specific centers.
- Correspond to choices of fixed or random C*T in model 3.

Choice of fixed or random C*T

- C*T is fixed:

Source	df	E MS
Center	79	
Treatment	1	$Q(t) + \sigma_e^2$
C*T	79	$Q(ct) + \sigma_e^2$
Error	7360	σ_e^2

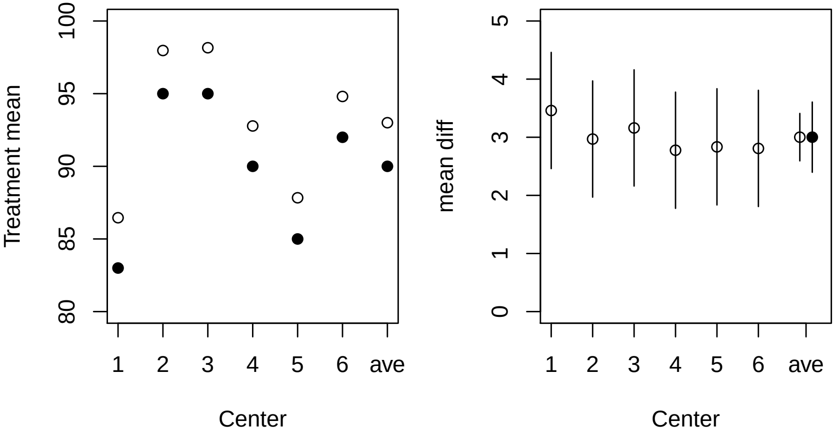
  - Denominator for F test of treatments is the residual variance.

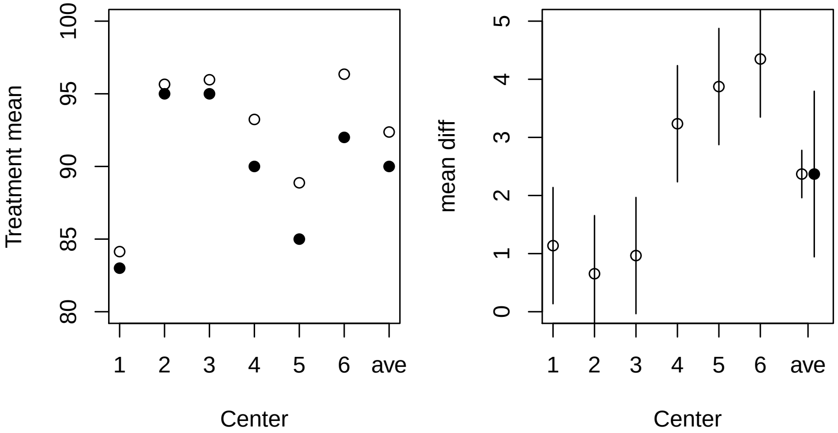
- C*T is random:

Source	df	E MS
Center	79	
Treatment	1	$Q(t) + 47\sigma_{ct}^2 + \sigma_e^2$
C*T	79	$47\sigma_{ct}^2 + \sigma_e^2$
Error	7360	σ_e^2

  - Denominator is the C*T interaction.

- Comparison of fixed C*T analysis and random C*T analysis
  - Very similar conclusions if C*T interaction is small, but fewer df for C*T
  - Quite different if C*T interaction is large
  - Plot below and on next page illustrate the difference




Treatment mean
80
85
90
95
100
1
2
3
4
5
6
ave
Center
mean diff
0
1
2
3
4
5
1
2
3
4
5
6
ave
Center

Inference space

- Inference space (McLean, Sanders, and Stroup 1991, Am. Stat. 45:54-64)
  - What set of centers are we making conclusions about?
  - Narrow sense inference: for these specific centers
    Fixed C*T
    or Random C*T, but conditional on observed values of C*T effects
  - Intermediate/Broad sense inference:
    to a larger population of centers
    Random C*T
  - (MSS make a distinction between intermediate- and broad-sense inference, not relevant here)

- A few additional thoughts:
  1. Think about whether error variance same at all centers
     - expect larger variance if more heterogeneous patient population
     - use Aitken model (known variance ratios)
     - or estimate separate variances for each center
  2. If more than two treatments, e.g. 2 way factorial., more than one way to proceed when using random C*T
     - Big issue is pooling Center*something interactions
     - Interaction quantifies consistency of treatment components across centers
     - What treatments have similar consistency?
       Pool those interactions.

- Example: 10 centers, 6 treatments with a 2 x 3 factorial structure

Source	df	Source	df	Source	df
Center	9	Center	9	Center	9
Trt	5				
A	1	A	1	A	1
B	2	B	2	B1	1
				B2	1
A*B	2	A*B	2	AB1	1
				AB2	1
C*Trt	45	C*A	9	C*A	9
		C*B	18	C*B1	9
				C*B2	9
		C*A*B	18	C*AB1	9
				C*AB2	9

- Standard advice is to use the middle column without thinking.
- My advice: think!
- Illustrated by three examples

- Example 1:
  - 2 levels of P fertilizer, 3 levels of N fertilizer,
    repeated at 10 locations.
  - Effects of N and effects of P likely to be similar across locations
  - suggests left analysis,
    using a single consistency term (C*Trt) is reasonable

- Example 2:
  - repeated at 10 locations, at each:
    2 levels of fertilizer,
    3 levels of insecticide (none, product A, product B)
  - Fertilizer has similar effects at all locations
    small Center*Fertilizer interaction
  - Insecticide effect depends on # nasty insects in area
    - very few: no effect of insecticide (either A or B)
    - lots of insects: big diff. between none and either product
    - Large Center*Insecticide interaction
  - Middle analysis: Center*Fertilizer different from Center*Insecticide
    Don't pool over all 6 treatments

- Example 3: Same study as example 2
    - repeated at 10 locations, at each:
      2 levels of fertilizer,
      3 levels of insecticide (none, product A, product B)
    - But, is middle analysis appropriate?
    - Two insecticides (A, B) may be approx. equally effective
      Product A-Product B consistent across locations
    - But None - (A+B)/2 may vary widely across locations
      Large effect when many insects, little effect when few
    - Perhaps right-most analysis (C*individual components) is most appropriate

- Sometimes have multiple locations and multiple years
- Example: 3 locations, each with 3 years.
  New experiments each year, 2 treatments
- Do you consider locations and years differently or combine into 9 “environments”
-

Source	d.f.	Source	d.f.
Environment	8		
		Year	2
		Location	2
		Y*L	4
Treatment	1	Treatment	1
Env * Trt	8		
		Year * trt	2
		Loc * trt	2
		Y*L*T	4

- Same sort of thinking needed: what is appropriate to pool?
- May need to use data to help decide (will see shortly)

Repeated Measures data

- Repeated measures: multiple observations (usually over time) on same e.u.
- Example: ISU Blossom Project. Excess weight gain during pregnancy is a serious health issues.
- Walking more often may alleviate the problem.
- Study: Pregnant moms randomly assigned to one of three treatments: usual care, weak encouragement to walk, strong encouragement to walk.
- Weight measured at 16 weeks (start of study), 20 weeks, 24 weeks, 28 weeks, 32 weeks, and 36 weeks.
- Response is weight gain at time *i* = weight at time *i* - weight at 16 weeks.
    - 5 responses for each e.u.
    - N.B. Treatment randomly assigned; Time not!
- What is the correlation among the 5 obs. made on the same subject?

Possible models for the correlations (or Var-Cov matrix) among the 5 obs. per subject:

1. Independence: no correlation between obs. made on same subject

$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

2. Compound symmetry: two sources of variation: subjects and measurements
   - data just like a split plot study with time randomly assigned to measurement
   - σ_s^2: variation among subjects (main plot)
   - σ_m^2: variation among measurements (split plot)

$$\begin{bmatrix} \sigma_s^2+\sigma_m^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2+\sigma_m^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2+\sigma_m^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma^2-s+\sigma_m^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2+\sigma_m^2 \end{bmatrix}$$

   - Correlation matrix shows that obs. 1 time apart have same correlation as those 4 times apart

$$\begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

   - Makes sense for split-plot data, but not for temporal data

3. Autocorrelated, order 1
   - correlation among observations t units apart: ρ^t
   - The covariance matrix for five obs. on one subject:

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

3. AR(1) with compound symmetry
   - two pieces to the covariance: something to do with the subject (CS) and something that decays with time lag (AR1)
   - correlations that decay with time lag, but to a non-zero asymptote
   - σ_s^2 is the variability among subjects; σ_e^2 is the variability that decays with time.
   - The covariance matrix for five obs. on one subject:

$$\begin{bmatrix} \sigma_s^2 + \sigma_e^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2\rho^2 & \sigma_s^2 + \sigma_e^2\rho^3 & \sigma_s^2 + \sigma_e^2\rho^4 \\ \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2\rho^2 & \sigma_s^2 + \sigma_e^2\rho^3 \\ \sigma_s^2 + \sigma_e^2\rho^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2\rho^2 \\ \sigma_s^2 + \sigma_e^2\rho^3 & \sigma_s^2 + \sigma_e^2\rho^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2 & \sigma_s^2 + \sigma_e^2\rho \\ \sigma_s^2 + \sigma_e^2\rho^4 & \sigma_s^2 + \sigma_e^2\rho^3 & \sigma_s^2 + \sigma_e^2\rho^2 & \sigma_s^2 + \sigma_e^2\rho & \sigma_s^2 + \sigma_e^2 \end{bmatrix}$$

4. Toeplitz
    - correlations same among adjacent obs., but not ρ^2 for two apart
    - $$\sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

6. Autocorrelated, order 1, with heterogeneous variances
   - Variances for each time not same
   - correlation among observations t units apart: ρ^t
   - The covariance matrix for five obs. on one subject:

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 & \sigma_1\sigma_5\rho^4 \\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 & \sigma_2\sigma_5\rho^3 \\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho^2 \\ \sigma_1\sigma_4\rho^3 & \sigma_2\sigma_4\rho^2 & \sigma_3\sigma_4\rho & \sigma_4^2 & \sigma_4\sigma_5\rho \\ \sigma_1\sigma_5\rho^4 & \sigma_2\sigma_5\rho^3 & \sigma_3\sigma_5\rho^2 & \sigma_4\sigma_5\rho & \sigma_5^2 \end{bmatrix}$$

5. Antedependence, order 1
   - correlations between periods *i* and *j* are product of pairwise correletions for all adjacent intervening periods
   - $$\sigma^2 \begin{bmatrix} 1 & \rho_{12} & \rho_{12}\rho_{23} & \rho_{12}\rho_{23}\rho_{34} & \rho_{12}\rho_{23}\rho_{34}\rho_{45} \\ \rho_{12} & 1 & \rho_{23} & \rho_{23}\rho_{34} & \rho_{23}\rho_{34}\rho_{45} \\ \rho_{12}\rho_{23} & \rho_{23} & 1 & \rho_{34} & \rho_{34}\rho_{45} \\ \rho_{12}\rho_{23}\rho_{34} & \rho_{23}\rho_{34} & \rho_{34} & 1 & \rho_{45} \\ \rho_{12}\rho_{23}\rho_{34}\rho_{45} & \rho_{23}\rho_{34}\rho_{45} & \rho_{34}\rho_{45} & \rho_{45} & 1 \end{bmatrix}$$
   - Commonly has heterogeneous variances as well

7. Unstructured
    - Variances for each time not same
    - different covariances for each pair of times
    - The covariance matrix for five obs. on one subject:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 \end{bmatrix}$$

    - If you know about the MANOVA approach to repeated measures, that is equivalent to fitting an unstructured model, with slightly different hypothesis tests.

- Choice of model affects estimates of $\boldsymbol{C}\beta$; some choices really affect variance of $\boldsymbol{C}\hat{\beta}$.
- Time not randomly assigned, so really don't know which covariance model is most appropriate.
- Want a simple model that is consistent with the data
    - Estimators less efficient if use too complicated a covariance model
    - Estimators biased if use too simple a covariance model
    - Often use AIC or BIC calculated from the REML lnL to select covariance model
- One standard approach (Diggle, Liang, Zeger, and Heagerty):
    - Use data to help choose a covariance model
    - Estimate $\boldsymbol{C}\beta$ and Var $\boldsymbol{C}\hat{\beta}$ conditional on that covariance model
- Var $\boldsymbol{C}\hat{\beta}$ often biased in models like AR(1) and ANTE(1)
- Kenward-Roger have bias correction.

Choosing among possible random effects structures

- Sometimes random effects structure specified by the experimental design
  - e.g. for experimental study, need a random effect for each e.u.
- Sometimes subject matter information informs the choice
  - e.g. expect a correlation among people in the same family
- Sometimes you need to use the data to help choose an appropriate structure
- two commonly used approaches and one less commonly used.
  - AIC or BIC
  - Likelihood ratio test

INFORMATION CRITERIA: AIC and BIC

- Goal is a model that:
  - Fits the data reasonably well
  - Is not too complicated
- Deviance: $-2l(\hat{\theta})$, where $l(\hat{\theta})$ is the log likelihood function evaluated at the mle's.
- Smaller values (or more negative values) = better fit of model to data.
- Adding parameters (i.e. a more complicated model) always reduces deviance.
- Akaike's Information criterion AIC $= -2l(\hat{\theta}) + 2k$, where k is the total number of model parameters.
- The +2k portion of AIC can be viewed as a penalty for model complexity.
- Small values of AIC are preferred.
- Beware: sometimes calculated as $2l(\hat{\theta}) - 2k$, for which large is better. (Now rare)

- Schwarz's Bayesian Information Criterion BIC $= -2l(\hat{\theta}) + k\log(n)$
- Similar to AIC except the penalty for model complexity is greater (for $n \geq 8$) and grows with n.
- AIC and BIC can be used to compare any set of models.
- Pairs do not need to be nested (i.e., one is not a special case of the other)
    - reduced vs. full model comparison only works for nested models
- Require that models are fit to **same** data
- If based on REML lnlL, models **MUST have the same fixed effects structure**.
    - Different fixed effect imply different error constrasts, so different data
- Actual value of AIC or BIC is uninformative (depend $\mid \Sigma \mid$)

Interpretation of differences between two AIC/BIC values:

1. Choose the model with the smallest AIC/BIC. period.
2. Look at the difference between a model and the best model (smallest AIC/BIC).
   - Interpretations suggested by Burnham and Anderson, Model Selection and Inference, book.
   - difference < 2: both models are reasonable
   - difference > 10: strongly prefer model with smaller AIC
   - But, revisions to these cutpoints have recently been proposed (by Burnham), e.g. < 4: both models are reasonable. I'm confused.
3. So, I suggest you choose the model with the smallest AIC/BIC.
   - But, Choose a "nearly" best model if preferred by subject-matter science or previous work in the area.
   - I.e. Think!

- Example: Subjects randomly assigned to treatments, 3 repeated measurements

Model	# param	AIC	Δ AIC	BIC	Δ BIC
Independence	1	342.0	8.9	343.5	8.9
Comp. Sym.	2	337.0	3.9	338.5	3.9
AR(1)	2	333.1		334.6	
ARH(1)	4	334.8	1.7	336.2	1.6
ANTE(1)	5	335.7	2.6	340.3	5.7
UN	6	335.7	2.6	340.3	5.7

- In this case, both AIC and BIC suggest the same model: AR(1).
- If not the case, which is more appropriate?
  - Depends on reason for selecting a model
  - I've gone back and forth. How important is a simple model?
- If past work suggests that you should expect unequal variances, I would use ARH(1).
- Is it necessary to get the correct correlation model?

Consequences of choice of correlation structure

- Consider two specific comparisons of means
- between 2 treatments (btwn subjects) and between two times (w/i subjects)

	Trt A - Trt B		Time 2 - Time 3	
Model	s.e.	d.f.	s.e.	d.f.
Independence	12.09	33	10.46	33
Comp. Sym.	17.84	9	7.82	24
AR(1)	18.34	9.3	6.98	23.1
ARH(1)	18.01	8.53	6.21	20.6
ANTE(1)	18.48	8.74	7.64	12
UN	18.62	8.29	7.49	12

- Assuming independence clearly different from models w/correlation
- Choice among correlated data models not that important
- Models with more param. tend to have larger s.e. and fewer d.f. i.e. lower power.

LIKELIHOOD RATIO BASED INFERENCE

- Suppose we need a formal hypothesis test for a covariance/correlation model.
- E.g.: repeated measures, 3 times. Willing to assume ar(1) correlation. Can we assume the same variance for all 3 times?
- test Ho: AR(1) vs Ha: ARH(1)
- restricted to nested models, i.e. situations when Ho is a special case of Ha
- If you can assume some standard regularity conditions (discussed in Shao(2003) and probably in Casella and Berger), (REDUCED MODEL DEVIANCE) - (FULL MODEL DEVIANCE) $\overset{H_0}{\sim} X^2_{k_f - k_r}$, where k_f and k_r are the number of free parameters under the full and reduced models, respectively.
- This is an asymptotic result, but usually works surprisingly well for small-ish samples **when regularity conditions hold.**

- Applies to REML lnL just as well as the usual lnL
- for our situation (3 sampling times), $k_f = 4$, $k_r = 2$
- So reject Ho at level α if Δ deviance $> \chi^2_2(1 - \alpha)$
- For data used above, Δ deviance = 329.1 - 327.8 = 1.3, p $\sim$ 0.52.
- Note that the test statistic is equal to $-2 \log \Lambda$, where

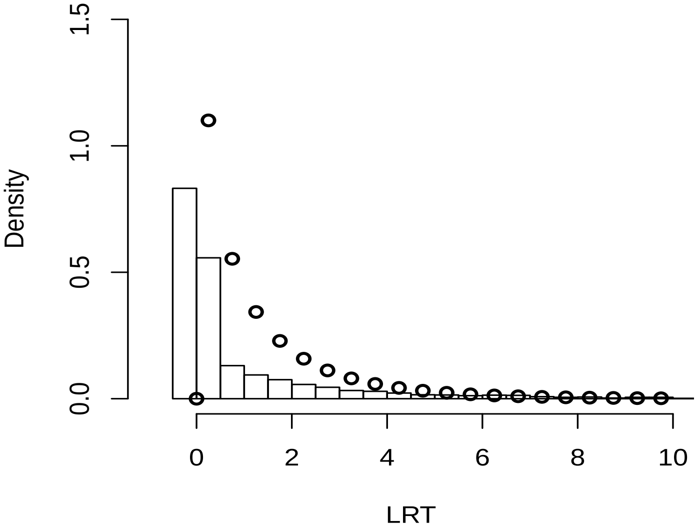
$$\Lambda = \frac{\text{LIKELIHOOD MAXIMIZED UNDER REDUCED MODEL}}{\text{LIKELIHOOD MAXIMIZED UNDER THE FULL MODEL}}$$

- Λ is known as the likelihood ratio, and tests based on $-2 log\Lambda$ are called likelihood ratio tests.

Two notes for caution

1. Regularity conditions and Distribution of Δ deviance
2. Robustness of the LRT

- Example: Soil porosity sampling study used in a HW problem
- $Y_{ijk} = \mu + \alpha_1 + \beta_{ij} + \epsilon_{ijk}$, where $\alpha_i \sim N(0, \sigma^2_{field})$, $\beta_{ij} \sim N(0, \sigma^2_{section})$, $\epsilon_{ijk} \sim N(0, \sigma^2_{location})$
- Want a formal hypothesis test of $\sigma^2_{field} = 0$
- Could fit two models and calculate two deviances
  - Ho: $\sigma^2_{field} = 0$, $Y_{ijk} = \mu + \beta_{ij} + \epsilon_{ijk}$, $D_{red} = -2 \log L(\hat{\sigma}^2_{section}, \hat{\sigma}^2_{location} | \sigma^2_{field} = 0)$
  - Ha: $\sigma^2_{field} > 0$, $Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}$, $D_{full} = -2 \log L(\hat{\sigma}^2_{field}, \hat{\sigma}^2_{section}, \hat{\sigma}^2_{location})$
- Usual theory says $D_{red} - D_{full} \overset{H_0}{\sim} \chi^2_1$
- Not correct for this application!


Density
0.0
0.5
1.0
1.5
0
2
4
6
8
10
LRT

- The regularity conditions do not hold if the true parameter falls on the boundary of the parameter space, when $\sigma_u^2 = 0$.
- $D_{red} - D_{full}$ for testing $H_0 : \sigma_u^2 = 0$ does not have a χ_1^2 distribution!
- Chernoff (1954, On the distribution of the likelihood ratio, Ann. Math. Stat. 25:573-578) develops the appropriate asymptotic distribution.
- In sufficiently large samples, T = $D_{red} - D_{full}$ for testing parameters on the boundary is a mixture of χ^2 distributions.
- For Ho: one parameter = 0, $f(T) = 0.5f(\chi_0^2) + 0.5f(\chi_1^2)$, where $f(\chi_k^2)$ is the pdf of a χ^2 distribution with k d.f. A χ_0^2 distribution is a point mass at 0.
- So $P[T > x] = P[\chi_1^2 > x]/2$
- Or, ignore the LRT and use an F test
  (Mean Squares not influenced by the parameter boundary)
- To my knowledge, no one has ever asked whether the interpretation of AIC values needs to change when some models include parameters on their boundary.

- So what might you do if you can't use an F test and don't know the appropriate theory?
- Randomization tests and non-parametric bootstraps are difficult because observations are correlated.
- not impossible, just difficult
- Use simulation. Like a parametric bootstrap but under Ho.
  1. Fit model under Ho to get estimates of remaining elements of γ.
  2. Simulate a data set assuming H0 and $\gamma = \hat{\gamma}$.
  3. Calculate T = desired test statistic from simulated data
  4. Repeat many times, e.g. N = 999
  5. Calculate empirical probability, $P[T \geq T_{obs}]$ from 1 obs. and N simulated data sets.

Robustness of LRT's to non-normality

- All the LR theory based on models that specify distributions
- Some LR tests involving variances are very sensitive to outliers
  Classic example is Bartlett's test of equal variances in two samples
  This is the LRT if errors are normally distributed.
- Little known about performance in general mixed models
- Verbyla, 1993, JRSS B 55:493-508, suggests that, in at least one case, REML estimators are more robust to outliers than are ML estimators.
- Performance of tests based on those estimators. Don't know.
- ANOVA estimators are MoM: no assumed distribution
- F tests generally robust to non-normality

Diagnostics

- What can you do to assess mixed model assumptions about distributions?
- Canonical mixed model: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \epsilon$
- Two types of residuals: $\boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ and $\boldsymbol{y} - (\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{\boldsymbol{u}})$
- ϵ is easy: look at $\hat{\epsilon} = \boldsymbol{y} - (\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{\boldsymbol{u}})$
- plot residuals against predicted means, $\boldsymbol{X}\hat{\beta}$, or predicted values, $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{\boldsymbol{u}}$.
- Both identify outliers; $\boldsymbol{X}\hat{\beta} + \boldsymbol{Z}\hat{\boldsymbol{u}}$ probably better.
- If variances are unequal, need to look at standardized residuals

$$r_i = \frac{\boldsymbol{y}_i - (\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{u}})}{\sqrt{\text{Var } y_i}}$$

or

$$r_i = \frac{\boldsymbol{y}_i - (\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{u}})}{\sqrt{\text{Var } (y_i - (\boldsymbol{X}_i\hat{\beta} + \boldsymbol{Z}_i\hat{\boldsymbol{u}}))}}$$

- What about diagnostics for $\boldsymbol{u}$?
- Sometimes called “case” or “subject” outliers.
  Occur when all obs. on a subject are unusual.
- Tempting to use BLUP's: $\hat{\boldsymbol{u}}$ to assess normality, outliers, and equal variance
- Sometimes works, sometimes doesn't.
- When lots of shrinkage (i.e. large Var ϵ and small Var u_i, empirical distribution of $\hat{u}_i$ looks normal even if true distribution of u_i far from normal (Verbeke and Lesaffre, 1997, Comp. Stat. and Data An., 23:541-556)
- Some ideas available, e.g. Verbeke and Molenberghs (2000, *Linear Mixed Models for Longitudinal Data*, section 7.8), but not yet widely available or commonly used
- Still an open issue.