36-617: Applied Linear Models

Lmer estimation and model selection

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- HW10 due Fri (updated!)*
  - I will post some guidance about calculating ICC's for part 2 of the project / technical appendix later today
- No Quiz today; no reading this week
- Project 02 Schedule:
  - **Fri Nov 19*:** Draft Technical Appendix with HW 10.
  - **Mon Nov 29 (or earlier):** Full IDMRAD paper first draft.
  - **Fri Dec 3:** Peer reviews due.
  - **Fri Dec 10 (or earlier):** Full IDMRAD paper final draft!

*These updated due dates were determined in class discussion today.

Plan for rest of semester

- M Nov 15 – estimation and model selection
- W Nov 17 – shrinkage, crash course on Bayes
- M Nov 22 – catch-up, or multilevel glm's
- **W Nov 24 – Thanksgiving break!**
- M Nov 29 – ?? Likely spline smoothing
- W Dec 1 – ?? Likely spline smoothing

Outline

- Estimation
  - ML: Full maximum likelihood
  - REML: Restricted or Residual maximum likelihood
  - Sheather’s recommendations
- AIC, BIC
  - MLE vs REML for AIC, BIC
- DIC
- Variable selection: Practical Advice
- An improved model for the London Schools data
- Automatic and Exact Methods…

Estimation: Maximum Likelihood

Consider the general Laird-Ware formulation

$$Y = X\beta + Z\eta + \varepsilon$$

Assume β is constant over subjects, ε is iid between subjects, and the variance-covariance matrix $\Psi = \mathsf{Var}(\eta)$ depends on only a few free parameters ω: $\Psi = \Psi(\omega)$. Assuming $\mathsf{Cov}(\eta, \varepsilon) = 0$,

$$\begin{aligned} Y &\sim N(X\beta, \Sigma(\omega)) \\ \text{where} \quad \Sigma(\omega) &= \mathsf{Var}(\varepsilon) + Z\Psi(\omega)Z^T \end{aligned}$$

so $-2\log(\textit{likelihood})$ is[1] (proportional to)

$$(Y - X\beta)^T \Sigma^{-1}(\omega)(Y - X\beta) + \log|\Sigma(\omega)| \qquad (*)$$

To find MLE's we can iterate[2] between minimizing in ω given β, and minimizing in β given ω; the latter is generalized least-squares (GLS)...

[1]Here we define |A| = det(A).
[2](an example of a Gauss-Seidel algorithm)

Estimation: REML

To reduce the amount of iteration for ML, we can compute a linear transformation AY whose distribution is independent of β, e.g.[1] $AY = (I - H_{OLS})Y = Y - X\hat{\beta}_{OLS}$.

Since we have changed the data (from Y to AY) we also change the likelihood from $(*)$ to

$$(Y - X\hat{\beta})^T \Sigma^{-1}(\omega)(Y - X\hat{\beta}) + \log|\Sigma(\omega)| + \log|X^T\Sigma(\omega)X| \qquad (**)$$

REML (REstricted or REsidual Maximum Likelihood) obtains $\hat{\omega}_{REML}$ by minimizing $(**)$ and then re-estimating $\hat{\beta}_{REML}$ by GLS as in $(*)$.

It can be shown that:

- $\Sigma(\hat{\omega}_{MLE})$ is biased, but $\Sigma(\hat{\omega}_{REML})$ is unbiased
- $-\frac{1}{2}(**)$ is (proportional to) a legitimate likelihood for ω
- $\hat{\beta}_{REML}$ are not maximum likelihood estimates

[1] Indeed, $AY = A(X\beta + Z\eta + \epsilon) = 0 + A(Z\eta + \epsilon) \sim N(0, A\Sigma(\omega)A^T)$ does not depend on β!

Sheather's Recommendations

- The best estimates of
  - Fixed effects β come from full maximum likelihood (MLE)
  - Variance components (τ^2's and σ^2) come from REML
- To compare models with *nested fixed effects* but *same random effects*, use LRT with MLE.
  - I agree!
- To compare nested models with *same fixed effects* but *nested random effects*, use LRT with REML.
  - I disagree!
  - Problem: H_0: $\tau^2=0$ occurs at the edge of the parameter space where LRT may not be chi-squared[1] under H_0.

[1]Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605-610.

AIC, BIC….

- In order to properly use AIC or BIC in R, ***must*** calculate the true maximum log-likelihood.
  - Does not depend on chi-squared distribution
  - Works for nested or non-nested models
- Still need to be sure you are working with the same data and model family!
  - For this reason, we tend to work on fixed effects and random effects separately…
- By default, lmer() calculates REML estimates.
  - For AIC(), BIC(), logLik() functions in R, need full MLEs!

REML vs MLE

- Can use lmer() or update() function to get MLE fit
  - lmer.1 <- lmer(Y ~ 1 + LRT + (1 + LRT|school), data=school.frame, REML=F)
  - lmer.1 <- update(lmer.1, . ~ ., REML=F)
- Can produce substantial differences in likelihood
  - Use AIC(), BIC(), logLik() to extract these values directly from fitted model

lmer.1	logLik	AIC	BIC
REML	−2376.19	4764.38	4794.92
MLE	−2368.68	4749.36	4782.90

Not valid to compare models with these numbers

OK to compare models with these numbers

  - anova() always refits using MLEs so that comparisons are valid

DIC (Deviance Information Criterion)

- We know
  - AIC = -2logLik(M) + 2 k
  - BIC = -2logLik(M) + k log(n)
- DIC = -2logLik(M) + 2 k_{eff}
- In *multilevel models* k is not always obvious. For example:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2) \\ \alpha_j &= \beta_0 + \eta_j, \ \eta_j \sim N(0, \tau_0^2) \end{aligned}$$

  - τ_0^2 large $\Rightarrow$ one-way ANOVA with J cells (df=J)
  - τ_0^2 small $\Rightarrow$ fitting grand mean only (df=1)
  - $1 \leq k_{eff} \leq J$, depending on size of $\tau_0{}^2$

Spiegelhalter et al. (2002). Bayesian measures of model complexity and fit. *JRSSB, 64,* 583-639.

Variable Selection: Practical Advice

- Start with multilevel model that represents your initial guesses about group structure in the data
- Do variable selection on all the fixed effects first, using AIC, BIC or DIC
  - AIC will result in *bigger models* that predict better
  - BIC will result in *smaller models* that interpret better
  - DIC usually results in models between AIC and BIC sizes…
  - LRT only valid if models have nested fixed effects and same random effects
- Then go back and use AIC, BIC or DIC (or parametric boostrap[1]) to do selection on random effects

[1]We discuss library(RLRsim) later in the lecture.

Back to the London Schools Data

- Student (1..1978)
  - Gender (0=Female, 1=Male), per student
  - VR = verbal reasoning level (High/Med/Low)
  - LRT = London Reading test (at beginning of year)
  - Y = end-of-year test
- School (1..38)
  - School.gender (All.Boy, All.Girl, Mixed)
  - School.denom (Other,CofE,RomCath,State)
- So far, we have fitted the model

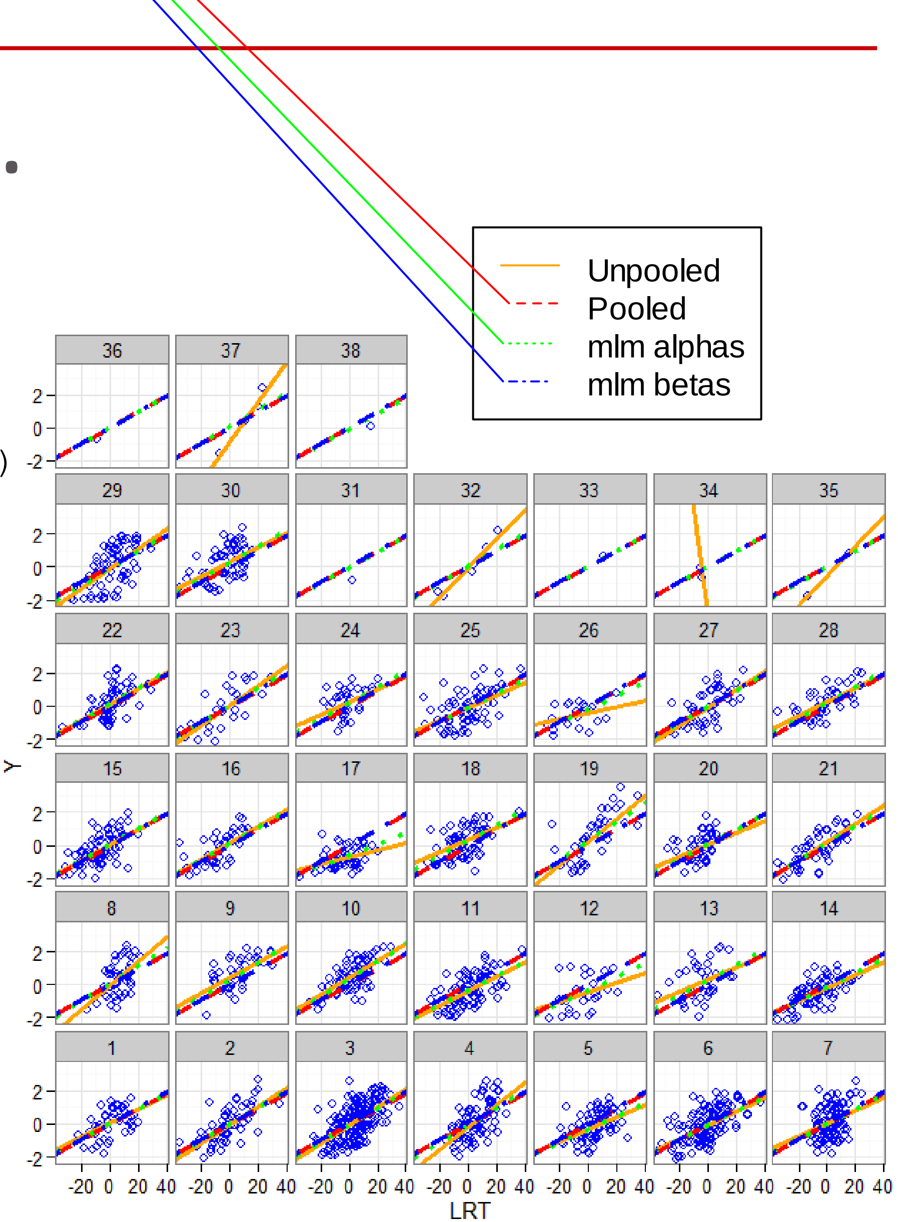
$$Y \sim 1 + LRT + (1 + LRT | school)$$

Our initial model…

```
> display(lmer.1
+ <- lmer(Y ~ 1 + LRT +
+ (1 + LRT|school),data=school.frame))

              coef.est  coef.se
(Intercept)   0.01      0.05
LRT           0.05      0.00

Error terms:
 Groups   Name   Std.Dev. Corr
 school   (Int)  0.23
          LRT    0.01     0.56
 Residual        0.79
---
number of obs: 1978,
groups: school, 38
AIC = 4764.4, DIC = 4722.4
deviance = 4737.4
```



The London Schools Data – Variable Selection

- How can we improve the model?

- We have a bunch of other variables lying around:
  - Unit-level (student): Gender, VR
  - Group-level (school): School.denom, School.gender

- Which ones to include? Fixed effects or random effects? Interactions? Etc.

Back to London Schools Data

```
> names(tmp) # main variabes in school.frame…
# [1] "Y"     "LRT"   "Gender"   "School.gender"
# [5] "School.denom"   "VR"

> lmer.2 <- update(lmer.1, . ~ . + Gender)
> anova(lmer.1,lmer.2)
# refitting model(s) with ML (instead of REML)
#        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
# lmer.1    6 4749.4 4782.9 -2368.7   4737.4
# lmer.2    7 4738.2 4777.3 -2362.1   4724.2 13.202  1  0.0002797 ***
# --> AIC, BIC prefer lmer.2

> lmer.3 <- update(lmer.2, . ~ . + School.gender)
> anova(lmer.2,lmer.3)
# refitting model(s) with ML (instead of REML)
#        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
# lmer.2    7 4738.2 4777.3 -2362.1   4724.2
# lmer.3    9 4736.4 4786.7 -2359.2   4718.4 5.7284  2    0.05703 .
# --> AIC, BIC disagree; LR test weakly in favor of lmer.3
```

Etc!

London Schools Data

- Tried Gender, School.gender, School.denom, and VR as *fixed effects*, found that Gender, School.gender and VR seem to improve the model.

- Trying to convert School.gender and VR to *random effects* does not improve AIC enough to keep them, so the final model we obtain is

```
> formula(lmer.5)
Y ~ LRT + Gender + School.gender + VR + (1 + LRT |
school)
```

London Schools Data – "final" model

```
> display(lmer.5)
lmer(formula = Y ~ LRT + Gender +
School.gender + VR + (1 + LRT |
    school), data = school.frame)
                      coef.est coef.se
(Intercept)            0.47     0.09
LRT                    0.03     0.00
Gender1                0.16     0.05
School.genderAll.Girl  0.04     0.13
School.genderMixed    -0.17     0.09
VRLow                 -0.92     0.07
VRMed                 -0.57     0.05

Error terms:
 Groups   Name        Std.Dev. Corr
 school   (Intercept) 0.23
          LRT         0.01     0.75
 Residual             0.75
```

```
---
number of obs: 1978, groups: school,
38
AIC = 4599, DIC = 4509.8
deviance = 4543.4

> anova(lmer.1,lmer.5)
refitting model(s) with ML (instead
of REML)
       npar    AIC    BIC  logLik
lmer.1    6 4749.4 4782.9 -2368.7
lmer.5   11 4565.4 4626.9 -2271.7
> 2271.7-2368.7
[1] -97
> pchisq(-2*(-97),5,lower=F)
[1] 5.453246e-40
```

Some Automatic & Exact Methods

- There are a number of R packages that will do variable selection for lmer models, including:
  - `LMERConvenienceFunctions` automates backwards selection of fixed effects and forward selection of random effects, using AIC, BIC, etc.
    - `fitLMER.fnc()` is general-purpose function for this
  - `RLRsim` provides simulation-based exact likelihood ratio tests for random effects
    - `exactLRT()` performs exact LRT test for true ML fits
    - `exactRLRT()` performs exact LRT test for REML fits

Automated Variable Selection…

```
> library(LMERConvenienceFunctions) # for fitLMER.fnc() function...
# start with a "big fixed effects" model
> lmer.10 <- lmer(Y ~ LRT + VR + Gender + School.gender +School.denom +
+ (1+LRT|school), data=school.frame)
> lmer.11 <- fitLMER.fnc(lmer.10,
+ ran.effects=c("(School.gender|school)",
+ "(School.denom|school)"),method="BIC")
> anova(lmer.5,lmer.10,lmer.11)
refitting model(s) with ML (instead of REML)
Data: school.frame
Models:
lmer.11: Y ~ LRT + VR + Gender + (1 + LRT | school)
lmer.5: Y ~ LRT + School.denom + VR + (1 + LRT | school)
lmer.10: Y ~ LRT + VR + Gender + School.gender + School.denom + (1 + LRT |
lmer.10:     school)
        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
lmer.11  9 4566.9 4617.2 -2274.4   4548.9
lmer.5  11 4577.2 4638.7 -2277.6   4555.2     0      2          1
lmer.10 14 4618.9 4697.2 -2295.5   4590.9     0      3          1
```

fitLMER.fnc:

1. Backwards elimination of F.E's
2. Forward selection of R.E.'s
3. Backwards elimination of F.E.'s

Exact Test of Random Effect..

```
library(RLRsim)

m0 <- lmer(Y ~ LRT + VR + Gender + (1  | school), data=school.frame)
lmer.11a <- lmer(Y ~ LRT + VR + Gender + (1|school) + (0 + LRT | school),
                 data=school.frame) # need indep rand effects for RLRsim...
lmer.LRT.only <- lmer(Y ~ LRT + VR + Gender + (0 + LRT | school),
                      data=school.frame)

formula(m0) # formula under H0: no random slopes for LRT
formula(lmer.11a) # model under HA: yes random slopes for LRT
formula(lmer.LRT.only) # model with *only* random slopes for LRT

exactRLRT(lmer.LRT.only,lmer.11a,m0)

#          simulated finite sample distribution of RLRT.
#
#          (p-value based on 10000 simulated values)
#
# data:
# RLRT = 6.2561, p-value = 0.0055
```

Summary

- Estimation
  - ML: Full maximum likelihood
  - REML: Restricted or Residual maximum likelihood
  - Sheather's recommendations
- AIC, BIC
  - MLE vs REML for AIC, BIC
- DIC
- Variable selection: Practical Advice
- An improved model for the London Schools data
- Automatic and Exact Methods…