STA 610L: Module 2.7

Random effects ANOVA (estimation)

Dr. Olanrewaju Michael Akande

Estimation methods

We briefly consider the following estimation methods for random intercept models.

- Maximum likelihood (ML)
- Restricted maximum likelihood (REML)
- Empirical Bayes estimation

Maximum likelihood estimation

We can also think of this formulation in the framework of the general linear mixed effects model, where

$$y = X\beta + Zb + \varepsilon.$$

In the random effects ANOVA case,

- X is just a column of 1's specifying the intercept $\beta = \mu$
- Z is a matrix of indicator variables indicating group membership
- Assume the random effects $b \sim N(0, G)$ where $G = \tau^2 I$ and the errors $\varepsilon \sim N(0, R)$ where $R = \sigma^2 I$

  The covariance matrix Σ is then given by

$$\begin{aligned}\Sigma &= \mathrm{Var}(y) = \mathrm{Var}(X\beta + Zb + \varepsilon) \\ &= \mathrm{Var}(X\beta) + \mathrm{Var}(Zb) + \mathrm{Var}(\varepsilon) \\ &= Z\mathrm{Var}(b)Z' + \mathrm{Var}(\varepsilon) = ZGZ' + R = \tau^2 ZZ' + \sigma^2 I\end{aligned}$$

Maximum likelihood estimation

Assuming our N outcomes follow the multivariate Gaussian distribution, our likelihood is given by

$$(2\pi)^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y-X\beta)'\Sigma^{-1}(y-X\beta)\right),$$

and we often work with the log-likelihood, given by

$$\begin{aligned}\ell(y,\beta,\Sigma) &= -\frac{1}{2}\left\{N\log(2\pi)+\log|\Sigma|+(y-X\beta)'\Sigma^{-1}(y-X\beta)\right\}\\ &\propto \log|\Sigma|+(y-X\beta)'\Sigma^{-1}(y-X\beta),\end{aligned}$$

which we then minimize (as I took the negative) in order to find the MLE.

MLE FOR SIMPLEST CASE

Recall a one-sample setting in which we wish to estimate the sample mean μ and variance σ^2 using the model

$$y_i = \mu + \varepsilon_i, \;\; i = 1, \ldots, n$$

with $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

Then our log-likelihood is proportional to $n \log \sigma^2 + \frac{\sum (y_i - \mu)^2}{\sigma^2}$, and to find the MLE's of μ and σ^2, when you take derivatives and solve for zero, you obtain $\widehat{\mu} = \overline{y}$ and $\widehat{\sigma}^2 = \frac{\sum (y_i - \overline{y})^2}{n}$.

Typically we don't use the MLE to estimate σ^2 because of its well-known small-sample bias, instead using the unbiased estimate $s^2 = \frac{\sum (y_i - \overline{y})^2}{n-1} = \frac{n}{n-1} \widehat{\sigma}^2$.

MLE FOR BIKE DATA

Recall our ANOVA model for the bike data.

Let's add one more level to make it a hierarchical model.

That is,

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

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \perp \alpha_j \overset{iid}{\sim} N(0, \tau^2)$.

y_{ij} indicates the passing distance between the car and the bike, and α_j represent effects of different distances between the bike and the curb.

MLE FOR BIKE DATA

Instead of directly maximizing the log-likelihood in R, we can use the lme4 library to do the work for us.

```
load("data/PsychBikeData.RData")
library(lme4)
fit.ml=lmer(`passing distance` ~ (1 | kerb), REML=FALSE, data = PsychBikeData)
summary(fit.ml)
```

MLE FOR BIKE DATA

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: `passing distance` ~ (1 | kerb)
##    Data: PsychBikeData
##
##      AIC      BIC   logLik deviance df.resid
##   2028.7   2046.0  -1011.4   2022.7     2352
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5113 -0.6674 -0.0948  0.5511  6.3949
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  kerb     (Intercept) 0.009206 0.09595
##  Residual             0.137203 0.37041
## Number of obs: 2355, groups:  kerb, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.54023    0.04363    35.3
```

Our ML estimates of (μ, τ^2, σ^2) for the bike data are $(\widehat{\mu}, \hat{\tau}^2, \widehat{\sigma}^2) = (1.540, 0.009, 0.137)$.

REML

REML (restricted or residual maximum likelihood) estimation is quite popular for variance component estimation.

Features of REML estimation include the following

- it is based on a likelihood function that only uses information that does not depend on fixed effects (we define new outcomes orthogonal to the mean)
- it is generally less biased than ML estimates (and unbiased in certain special cases)

REML FOR SIMPLEST CASE

Recall our one-sample setting in which we wish to estimate the sample mean μ and variance σ^2 using the model

$$y_i = \mu + \varepsilon_i, \;\; i = 1, \ldots, n, \;\; \varepsilon_i \sim N\left(0, \sigma^2\right).$$

REML estimates are often based on a full-rank set of error contrasts -- the basic idea is to retain the information in the data about the variance while eliminating the fixed effects.

Because the residuals sum to zero in our simple model by definition, a full-rank set of residuals contains $n - 1$ residuals for our mean-only model.

If you've taken advanced linear models, you can show that the REML estimator maximizes the likelihood of a full rank set of error contrasts.

REML FOR SIMPLEST CASE

Because not everyone took STA 721, we'll consider a heuristic derivation.

The full residuals $\varepsilon_i = y_i - \mu$ contain all the information in the likelihood about the variance parameter σ^2.

Because the residuals are independent of the fixed effect μ and the sample mean $\overline{y}$ is independent of the sample variance, we can re-express our log likelihood to isolate the residual likelihood:

$$\ell(y, \mu, \sigma^2) = \ell(\varepsilon, \mu, \sigma^2) + \ell(\overline{y}, \mu, \sigma^2)$$

We know $\widehat{\mu} = \overline{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and so $\ell(\overline{y}, \mu, \sigma^2) \propto \log \frac{\sigma^2}{n} + \frac{(\overline{y}-\mu)^2}{\frac{\sigma^2}{n}}$ which reduces to $\log \sigma^2 - \log n$ once we plug in the MLE $\overline{y}$ for μ.

REML FOR SIMPLEST CASE

Then

$$\ell(\varepsilon, \mu, \sigma^2) \propto n \log \sigma^2 + \frac{\sum (y_i - \mu)^2}{\sigma^2} - \log \sigma^2 + \log n$$

which is proportional to

$$(n-1) \log \sigma^2 + \frac{\sum (y_i - \mu)^2}{\sigma^2},$$

which looks just like our ML likelihood with the exception of the multiplier $n-1$ instead of n, and it's straightforward to show the maximum is $\widehat{\sigma}^2_{REML} = \frac{\sum (y_i - \mu)^2}{n-1}$.

Because they are generally less biased than ML estimates, REML estimates are typically the default frequentist estimates provided by many software packages.

REML ESTIMATES FOR THE BIKE DATA

```
fit.reml=lmer(`passing distance` ~ (1 | kerb), REML=TRUE, data = PsychBikeData)
summary(fit.reml)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: `passing distance` ~ (1 | kerb)
##    Data: PsychBikeData
##
## REML criterion at convergence: 2027
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5132 -0.6647 -0.0940  0.5498  6.3978
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  kerb     (Intercept) 0.01157  0.1076
##  Residual             0.13720  0.3704
## Number of obs: 2355, groups:  kerb, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.54008    0.04876   31.59
```

Our REML estimates for the bike data are
$(\widehat{\mu}, \hat{\tau}^2, \widehat{\sigma}^2) = (1.540, 0.012, 0.137).$

REML ESTIMATES FOR THE BIKE DATA

Note that the "REML" estimate of the mean μ is not obtained using the REML likelihood, which does not involve parameters in the linear predictor.

We will return to this point later in the course when we deal with more complex models than this simple case, in which the mean estimate is $\overline{y}$ regardless of whether we use ML or REML for σ^2.

In more complicated models, the estimates of mean parameters will be a function of variance parameters, and iterative methods are required.

Empirical Bayes

When we have random effects in a model, the standard frequentist effects of these random quantities are called *empirical Bayes* estimates, regardless of whether we obtain other estimates using ML or REML.

Empirical Bayes

Recall our group means formulation:

$$\begin{aligned} y_{ij} &= \mu_j + \varepsilon_{ij} \\ \mu_1, \cdots, \mu_J &\overset{iid}{\sim} N(\mu, \tau^2) \\ \varepsilon_{ij} &\overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

Suppose (μ, τ^2, σ^2) are known exactly and consider estimating μ_j with an estimator that is a linear function of the group sample mean $\widehat{\mu}_j = a\bar{y}_j + b$.

Then one can show that the MSE $E[(\mu_j - \widehat{\mu}_j)^2]$ is minimized if $a = \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$

and $b = (1-a)\mu$, so that $\widehat{\mu}_j = w_j \bar{y}_j + (1 - w_j)\mu$, where $w_j = \frac{\frac{n_j}{\sigma^2}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$

Empirical Bayes

If we knew (μ, τ^2, σ^2) this estimate would be the *Bayes estimate*; however, we do not know these parameters and are instead estimating them from the data, so that

$$\widehat{\mu}_j = \widehat{w}_j \overline{y}_j + (1 - \widehat{w}_j)\widehat{\mu}, \text{ where } \widehat{w}_j = \frac{\frac{n_j}{\widehat{\sigma}^2}}{\frac{n_j}{\widehat{\sigma}^2} + \frac{1}{\hat{\tau}^2}}$$

is called an *empirical Bayes estimate* because our unknown parameters have been replaced by "empirical" estimates from the data.

While this estimate is widely-used, it has several unsatisfactory qualities, including a standard variance estimate known to be an underestimate.

This is great motivation for consideration of Bayesian approaches when formal comparisons among groups modeled with random effects are desired.

EB Estimates of group means for bike data

```
table(PsychBikeData$kerb); mean(PsychBikeData$`passing distance`)
```

```
##
## 0.25  0.5 0.75    1 1.25
##  670  545  339  469  332
```

```
## [1] 1.563912
```

```
tapply(PsychBikeData$`passing distance`,PsychBikeData$kerb,mean)
```

```
##     0.25      0.5     0.75        1     1.25
## 1.698054 1.590473 1.505519 1.490584 1.412813
```

```
coef(fit.ml)
```

```
## $kerb
##      (Intercept)
## 0.25    1.694619
## 0.5     1.589136
## 0.75    1.506981
## 1       1.492113
## 1.25    1.418287
##
## attr(,"class")
## [1] "coef.mer"
```

EB Estimates of group means for bike data

```
tapply(PsychBikeData$`passing distance`,PsychBikeData$kerb,mean)
```

```
##     0.25      0.5     0.75        1     1.25
## 1.698054 1.590473 1.505519 1.490584 1.412813
```

```
coef(fit.reml)
```

```
## $kerb
##      (Intercept)
## 0.25    1.695307
## 0.5     1.589401
## 0.75    1.506687
## 1       1.491805
## 1.25    1.417201
##
## attr(,"class")
## [1] "coef.mer"
```

Here we see only a slight shrinkage back towards the overall mean, due in large part to the large sample sizes within curb distances.

What's next?

Move on to the readings for the next module!