Multilevel Generalized Linear Models

cont'd

04.13.22

Click here for PDF of slides

Announcements

- Quiz 04 open due **Fri, April 15 at 11:59pm**
- Final project - optional draft due
  - **Fri, Apr 15 at 11:59pm**
  - final report due **Wed, Apr 27 at 11:59pm**
- Please fill out course evaluations!
- Click here for answers to questions about multilevel models submitted on Quiz 03. Thanks to Jose for putting this document together!

Learning goals

- Exploratory data analysis for multilevel data with non-normal response variable
- Write One, Level Two and composite models for mutlilevel GLM
- Fit and interpret multilevel GLM
- Understand crossed random effects and incorporate them in the multilevel model

Data: College Basketball referees

The dataset **`basketball0910.csv`** contains data on 4972 fouls in 340 NCAA basketball games from the Big Ten, ACC, and Big East conferences during the 2009-2010 season. The goal is to determine whether the data from this season support a conclusion from Anderson and Pierce (2009) that referees tend to "even out" foul calls in a game. The variables we'll focus on are

- **`foul.home`**: foul was called on home team (1: yes, 0: no)
- **`foul.diff`**: difference in fouls before current foul was called (home - visitor)
- **`game`**: Unique game ID number
- **`visitor`**: visiting team abbreviation
- **`home`**: home team abbreviation

See BMLR: Section 11.3.1 for full codebook.

Data: College basketball referees

game	visitor	hometeam	foul.num	foul.home	foul.vis	foul.diff	foul.type	time
1	IA	MN	1	0	1	0	Personal	14.167
1	IA	MN	2	1	0	-1	Personal	11.433
1	IA	MN	3	1	0	0	Personal	10.233
1	IA	MN	4	0	1	1	Personal	9.733
1	IA	MN	5	0	1	0	Shooting	7.767
1	IA	MN	6	0	1	-1	Shooting	5.567
1	IA	MN	7	1	0	-2	Shooting	2.433
1	IA	MN	8	1	0	-1	Offensive	1.000
2	MI	MIST	1	0	1	0	Shooting	18.983
2	MI	MIST	2	1	0	-1	Personal	17.200

Composite model

Composite model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \beta_0 \, \text{foul.diff}_{ij} + [u_i + v_i \, \text{foul.diff}_{ij}]$$

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

Fit the model in R

Use the **`glmer`** function in the **lme4** package to fit multilevel GLMs.

```
model1 <- glmer(foul.home ~ foul.diff + (foul.diff|game),
                data = basketball, family = binomial)
```

```
## boundary (singular) fit: see help('isSingular')
```

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	-0.157	0.046	-3.382	0.001
fixed	NA	foul.diff	-0.285	0.038	-7.440	0.000
ran_pars	game	sd__(Intercept)	0.542	NA	NA	NA
ran_pars	game	cor__(Intercept).foul.diff	-1.000	NA	NA	NA
ran_pars	game	sd__foul.diff	0.035	NA	NA	NA

Boundary constraints

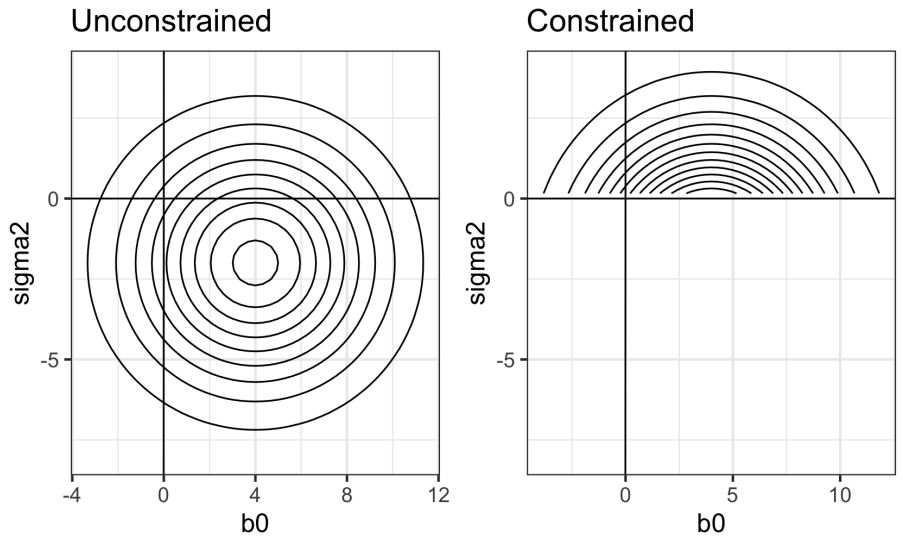
- The estimates of the parameters $\alpha_0, \beta_0, \sigma_u, \sigma_v, \rho_{uv}$ are those that maximize the likelihood of observing the data
- The fixed effects, e.g., α_0 and β_0, can take any values, but the terms associated with the error terms are constrained to a set of "allowable" values

$$\sigma_u \geq 0 \quad \sigma_v \geq 0 \quad -1 \leq \rho_{uv} \leq 1$$

- Because of these boundaries, a "constrained" search is used to find the MLEs.
- The warning message **`"## boundary (singular) fit"`**, means the estimate of one or more terms was set at the maximum (or minimum) allowable value, not the value it would've been if an unconstrained search were allowable

Illustrating boundary constraints

Contour plots from a hypothetical likelihood $L(\beta_0, \sigma^2)$



- In the unconstrained search, the likelihood $L(\beta_0, \sigma^2)$ is maximized at $\hat{\beta}_0 = 4, \hat{\sigma}^2 = -2$
- In reality σ^2 must be non-negative, so the search for the MLE is restricted to the region such that $\sigma^2 \geq 0$.
- The constrained likelihood is maximized at $\hat{\beta}_0 = 4, \hat{\sigma}^2 = 0$

From BMLR Figure 10.14

Original model

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	-0.157	0.046	-3.382	0.001
fixed	NA	foul.diff	-0.285	0.038	-7.440	0.000
ran_pars	game	sd__(Intercept)	0.542	NA	NA	NA
ran_pars	game	cor__(Intercept).foul.diff	-1.000	NA	NA	NA
ran_pars	game	sd__foul.diff	0.035	NA	NA	NA

1. Which term(s) should we remove to try to address boundary constraint?
2. Refit the model with these terms removed.
3. Which model do you choose? Use AIC to help make your choice.

04:00

Interpret model coefficients

Interpret the following coefficients in the selected model (if applicable):

- $\hat{\alpha}_0$
- $\hat{\beta}_0$
- $\hat{\sigma}_u$
- $\hat{\sigma}_v$
- $\hat{\rho}_{uv}$

Crossed random effects

- The Level Two covariates are the home team and visiting team
- There is some evidence in the EDA that there may be differences in the probability of a foul depending on the home team
- We will account for this difference by treating home team and visiting team as random effects in the model
    - *Issue*: Home and visiting team are not nested within game, since a single home and visiting team can be in multiple games
- The random effects for game, home team, and visiting team are **crossed random effects**

Notation

$Y_{i[gh]j}$: Indicator of whether the j^{th} foul in Game i was called on home team h instead of visiting team g

$$Y_{i[gh]j} \sim Bernoulli(p_{i[gh]j})$$

where $p_{i[gh]j}$ is the true probability a foul in Game i was called on home team h instead of visiting team g

Models

Level One

$$\log\left(\frac{p_{i[gh]j}}{1-p_{i[gh]j}}\right) = a_i + b_i \,\text{foul.diff}_{ij}$$

Level Two

$$a_i = \alpha_0 + u_i + v_h + w_g$$
$$\beta_i = \beta_0$$

$$u_i \sim N(0, \sigma_u^2) \quad v_h \sim N(0, \sigma_v^2) \quad w_g \sim N(0, \sigma_w^2)$$

Composite model

$$\log\left(\frac{p_{i[gh]j}}{1-p_{i[gh]j}}\right) = \alpha_0 + \beta_0 \,\text{foul.diff}_{ij} + [u_i + v_h + w_g]$$

$$u_i \sim N(0, \sigma_u^2) \quad v_h \sim N(0, \sigma_v^2) \quad w_g \sim N(0, \sigma_w^2)$$

Why add additional random effects?

- Get more precise estimates of fixed effects
- Can make comparisons of game-to-game and team-to-team variability
- Can get estimated random effects for each team and use them to compare odds of a foul on the home team for different teams

Model

```
model2 <- glmer(foul.home ~ foul.diff + (1|game) + (1|hometeam) + (1 | visitor),
                data = basketball, family = binomial)
```

```
tidy(model2) %>% kable(digits = 3)
```

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	-0.188	0.063	-2.967	0.003
fixed	NA	foul.diff	-0.264	0.039	-6.795	0.000
ran_pars	game	sd__(Intercept)	0.414	NA	NA	NA
ran_pars	hometeam	sd__(Intercept)	0.261	NA	NA	NA
ran_pars	visitor	sd__(Intercept)	0.152	NA	NA	NA

About what percent of the variability in the intercepts is due to...

- game-to-game differences?
- differences among home teams?
- differences among visiting teams?

Keep the crossed random effects?

- Given a large proportion of the variability in the intercepts is explained by game-to-game differences, we can assess if the random effects for home team and visiting team are providing useful information.
- To do so, we will compare the following models

```
modela <-  glmer(foul.home ~ foul.diff + (1|game),
                 data = basketball, family = binomial)

modelb <- glmer(foul.home ~ foul.diff + (1|game) + (1 | hometeam) + (1|visitor),
                data = basketball, family = binomial)
```

Write the null and alternative hypotheses to test these models.

Keep the crossed random effects?

Likelihood ratio test using χ^2 distribution (potentially unreliable when testing variance components)

```
anova(modela, modelb, test = "Chisq") %>% kable(digits = 3)
```

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
modela	3	6792.540	6812.075	-3393.270	6786.540	NA	NA	NA
modelb	5	6780.466	6813.024	-3385.233	6770.466	16.074	2	0

Parametric bootstrap (long computational time!)

Keep the crossed random effects?

AIC or BIC

```
glance(modela) %>% kable(digits = 3)
```

sigma	logLik	AIC	BIC	deviance	df.residual
1	-3393.27	6792.54	6812.075	6397.136	4969

```
glance(modelb) %>% kable(digits = 3)
```

sigma	logLik	AIC	BIC	deviance	df.residual
1	-3385.233	6780.466	6813.024	6420.537	4967

Estimated random effects for each team

- We will use **`model B`** with (crossed random effects) to get the estimated random effect for each team.
- These are **empirical Bayes estimates** ("shrinkage estimates").
  - Combine individual-specific information with information from all teams
  - "Shrinks" the individual estimates toward the group averages

See On the Use of Empirical Bayes Estimates as Measures of Individual Traits

for more detail on empirical Bayes estimates.

Estimated random effects for each team

We can get these effects using the **`ranef`** function in the lmer R package.

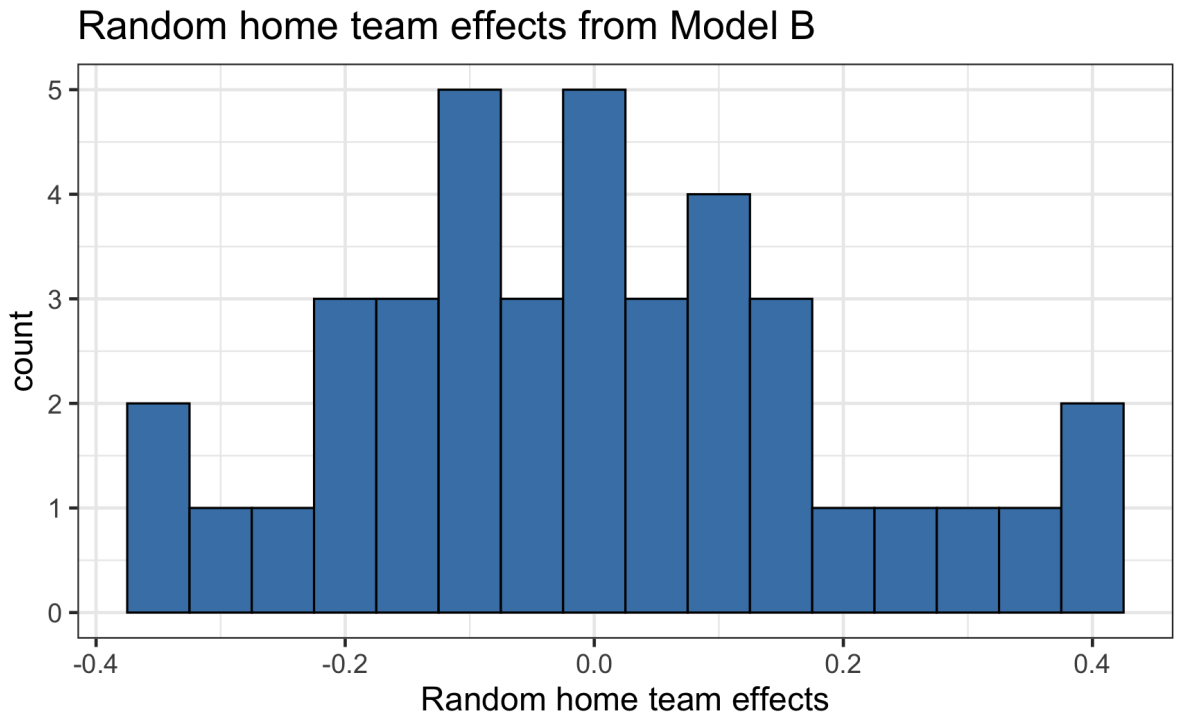
```
reffect_game <- ranef(modelb)$game %>% select(`(Intercept)`) %>% pull()
reffect_home <- ranef(modelb)$hometeam %>% select(`(Intercept)`) %>% pull()
reffect_visitor <- ranef(modelb)$visitor %>% select(`(Intercept)`) %>% pull()
team_names <- rownames(ranef(modelb)$visitor)
reffect_team <- tibble(team = team_names,
                       reffect_home = reffect_home,
                      reffect_visitor = reffect_visitor)
```

```
reffect_team %>%
  slice(1:5)
```

```
## # A tibble: 5 × 3
##   team   reffect_home reffect_visitor
##   <chr>         <dbl>           <dbl>
## 1 BC          -0.0900          0.0355
## 2 CIN          0.261          -0.0944
## 3 CLEM        -0.105          -0.110
## 4 CT          -0.290           0.0360
## 5 DEPAUL       0.416          -0.154
```

Distribution of random home team effects

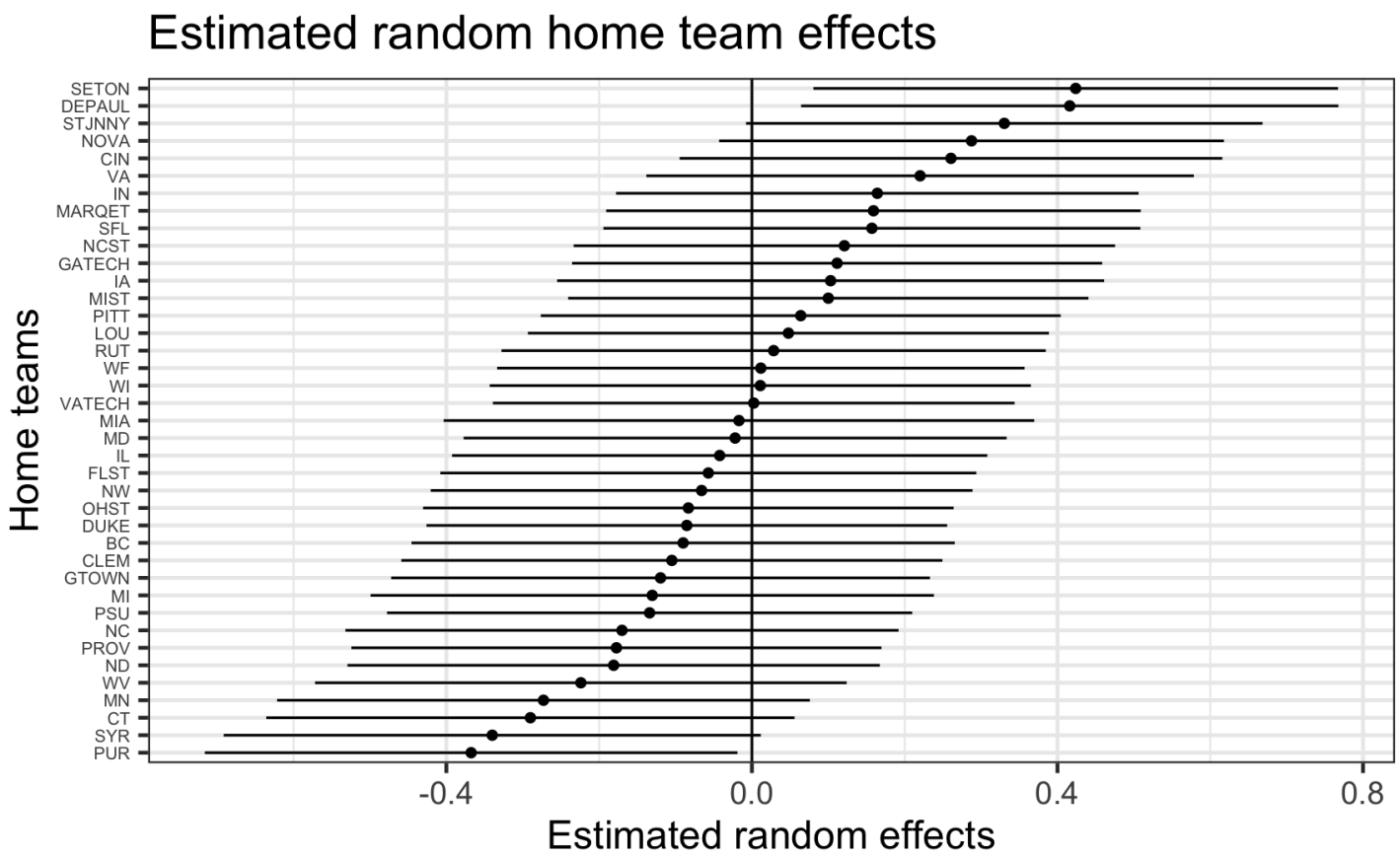
```
ggplot(data = reffect_team, aes(x = reffect_home)) +
  geom_histogram(binwidth = .05, color = "black", fill = "steelblue") +
  labs(x = "Random home team effects",
       title = "Random home team effects from Model B")
```



Estimated home random effects by team

Plot

Code



Estimated home random effects by team

Plot

Code

```
var <- attr(ranef(modelb)$hometeam, "postVar")
reffect_predict <- tibble(Intercepts = reffect_home,
              SD = 2*sqrt(var[,,1:length(var)]),
              team_names = reffect_team$team)

ggplot(data = reffect_predict, aes(fct_reorder(team_names, Intercepts),
                                   Intercepts)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_errorbar(aes(ymin = Intercepts - SD,
                    ymax = Intercepts + SD),
                width=0,color="black") +
  labs(title = "Estimated random home team effects",
       x = "Home teams",
       y = "Estimated random effects") +
 theme(axis.text.y = element_text(size = 7)) +
  coord_flip()
```

Modeling next steps

Do the data provide evidence that referees tend to "even out" foul calls?

- Adjust for additional covariates (score differential, type of foul, time left in first half)
- Hypothesize that the effect of foul differential may depend on other covariates, so consider potential interaction terms with foul differential
- Consider other random effects associated with game, home team, and visiting team

See Section 11.7: A final model for examining referee bias for final model chosen by the authors.

Acknowledgements

- BMLR: Section 10.5 - Encountering boundary constraints
- Chapter 11: Multilevel generalized linear models structure
    - Sections 11.1 - 11.4
- Anderson, Kyle J., and David A. Pierce. 2009. “Officiating Bias: The Effect of Foul Differential on Foul Calls in NCAA Basketball.” Journal of Sports Sciences 27 (7): 687–94. https://doi.org/10.1080/02640410902729733.