STA 610L: Module 4.3

Logistic mixed effects model (wrap up)

Dr. Olanrewaju Michael Akande

STA 610L

1988 ELECTIONS ANALYSIS

The dataset includes 2193 observations from one of eight surveys (the most recent CBS News survey right before the election) in the original full data.

Variable	Description
org	cbsnyt = CBS/NYT
bush	1 = preference for Bush Sr., 0 = otherwise
state	1-51: 50 states including DC (number 9)
edu	education: 1=No HS, 2=HS, 3=Some College, 4=College Grad
age	1=18-29, 2=30-44, 3=45-64, 4=65+
female	1=female, 0=male
black	1=black, 0=otherwise
region	1=NE, 2=S, 3=N, 4=W, 5=DC
v_prev	average Republican vote share in the three previous elections (adjusted for home-state and home-region effects in the previous elections)

Given that the data has a natural multilevel structure (through `state` and `region`), it makes sense to explore hierarchical models for this data.

STA 610L

1988 ELECTIONS ANALYSIS

Both voting turnout and preferences often depend on a complex combination of demographic factors.

In our example dataset, we have demographic factors such as biological sex, race, age, education, which we may all want to look at by state, resulting in $2 \times 2 \times 4 \times 4 \times 51 = 3264$ potential categories of respondents.

We may even want to control for `region`, adding to the number of categories.

Clearly, without a very large survey (most political survey poll around 1000 people), we will need to make assumptions in order to even obtain estimates in each category.

We usually cannot include all interactions; we should therefore select those to explore (through EDA and background knowledge).

The data is in the file `polls_subset.txt` on Sakai.

1988 ELECTIONS ANALYSIS

```r
###### Load the data
polls_subset <- read.table("data/polls_subset.txt",header=TRUE)
str(polls_subset)
```

```
## 'data.frame':	2193 obs. of  10 variables:
##  $ org   : chr  "cbsnyt" "cbsnyt" "cbsnyt" "cbsnyt" ...
##  $ survey: int  9158 9158 9158 9158 9158 9158 9158 9158 9158 9158 ...
##  $ bush  : int  NA 1 0 0 1 1 1 1 0 0 ...
##  $ state : int  7 39 31 7 33 33 39 20 33 40 ...
##  $ edu   : int  3 4 2 3 2 4 2 2 4 1 ...
##  $ age   : int  1 2 4 1 2 4 2 4 3 3 ...
##  $ female: int  1 1 1 1 1 1 0 1 0 0 ...
##  $ black : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ region: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ v_prev: num  0.567 0.527 0.564 0.567 0.524 ...
```

```r
head(polls_subset)
```

```
##      org survey bush state edu age female black region    v_prev
## 1 cbsnyt   9158   NA     7   3   1      1     0      1 0.5666333
## 2 cbsnyt   9158    1    39   4   2      1     0      1 0.5265667
## 3 cbsnyt   9158    0    31   2   4      1     0      1 0.5641667
## 4 cbsnyt   9158    0     7   3   1      1     0      1 0.5666333
## 5 cbsnyt   9158    1    33   2   2      1     0      1 0.5243666
## 6 cbsnyt   9158    1    33   4   4      1     0      1 0.5243666
```

STA 610L

1988 ELECTIONS ANALYSIS

```
summary(polls_subset)
```

```
##      org                survey          bush            state      
##  Length:2193        Min.   :9158   Min.   :0.0000   Min.   : 1.00  
##  Class :character   1st Qu.:9158   1st Qu.:0.0000   1st Qu.:14.00  
##  Mode  :character   Median :9158   Median :1.0000   Median :26.00  
##                     Mean   :9158   Mean   :0.5578   Mean   :26.11  
##                     3rd Qu.:9158   3rd Qu.:1.0000   3rd Qu.:39.00  
##                     Max.   :9158   Max.   :1.0000   Max.   :51.00  
##                                    NA's   :178                     
##       edu             age            female           black        
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.00000  
##  Median :2.000   Median :2.000   Median :1.0000   Median :0.00000  
##  Mean   :2.653   Mean   :2.289   Mean   :0.5887   Mean   :0.07615  
##  3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.00000  
##  Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000  
##                                                                    
##      region          v_prev      
##  Min.   :1.000   Min.   :0.1530  
##  1st Qu.:2.000   1st Qu.:0.5278  
##  Median :2.000   Median :0.5481  
##  Mean   :2.431   Mean   :0.5550  
##  3rd Qu.:3.000   3rd Qu.:0.5830  
##  Max.   :5.000   Max.   :0.6927  
## 
```

STA 610L

1988 ELECTIONS ANALYSIS

```
polls_subset$v_prev <- polls_subset$v_prev*100 #rescale
polls_subset$region_label <- factor(polls_subset$region,levels=1:5,
                                    labels=c("NE","S","N","W","DC"))
#we consider DC as a separate region due to its distinctive voting patterns
polls_subset$edu_label <- factor(polls_subset$edu,levels=1:4,
                                 labels=c("No HS","HS","Some College","College Grad"))
polls_subset$age_label <- factor(polls_subset$age,levels=1:4,
                                 labels=c("18-29","30-44","45-64","65+"))
#the data includes states but without the names, which we will need,
#so let's grab that from R datasets
data(state)
#"state" is an R data file (type ?state from the R command window for info)
state.abb #does not include DC, so we will create ours
```

```
##  [1] "AL" "AK" "AZ" "AR" "CA" "CO" "CT" "DE" "FL" "GA" "HI" "ID" "IL" "IN" "IA"
## [16] "KS" "KY" "LA" "ME" "MD" "MA" "MI" "MN" "MS" "MO" "MT" "NE" "NV" "NH" "NJ"
## [31] "NM" "NY" "NC" "ND" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VT"
## [46] "VA" "WA" "WV" "WI" "WY"
```

```
#In the polls data, DC is the 9th "state" in alphabetical order
state_abbr <- c (state.abb[1:8], "DC", state.abb[9:50])
polls_subset$state_label <- factor(polls_subset$state,levels=1:51,labels=state_abbr)
rm(list = ls(pattern = "state")) #remove unnecessary values in the environment
```

STA 610L

1988 ELECTIONS ANALYSIS

```r
###### View properties of the data
head(polls_subset)
```

```
##      org survey bush state edu age female black region   v_prev region_label
## 1 cbsnyt   9158   NA     7   3   1      1     0      1 56.66333           NE
## 2 cbsnyt   9158    1    39   4   2      1     0      1 52.65667           NE
## 3 cbsnyt   9158    0    31   2   4      1     0      1 56.41667           NE
## 4 cbsnyt   9158    0     7   3   1      1     0      1 56.66333           NE
## 5 cbsnyt   9158    1    33   2   2      1     0      1 52.43666           NE
## 6 cbsnyt   9158    1    33   4   4      1     0      1 52.43666           NE
##      edu_label age_label state_label
## 1 Some College     18-29          CT
## 2 College Grad     30-44          PA
## 3           HS       65+          NJ
## 4 Some College     18-29          CT
## 5           HS     30-44          NY
## 6 College Grad       65+          NY
```

```r
dim(polls_subset)
```

```
## [1] 2193   14
```

STA 610L

1988 ELECTIONS ANALYSIS

```
###### View properties of the data
str(polls_subset)
```

```
## 'data.frame':    2193 obs. of  14 variables:
##  $ org         : chr  "cbsnyt" "cbsnyt" "cbsnyt" "cbsnyt" ...
##  $ survey      : int  9158 9158 9158 9158 9158 9158 9158 9158 9158 9158 ...
##  $ bush        : int  NA 1 0 0 1 1 1 1 0 0 ...
##  $ state       : int  7 39 31 7 33 33 39 20 33 40 ...
##  $ edu         : int  3 4 2 3 2 4 2 2 4 1 ...
##  $ age         : int  1 2 4 1 2 4 2 4 3 3 ...
##  $ female      : int  1 1 1 1 1 1 0 1 0 0 ...
##  $ black       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ region      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ v_prev      : num  56.7 52.7 56.4 56.7 52.4 ...
##  $ region_label: Factor w/ 5 levels "NE","S","N","W",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ edu_label   : Factor w/ 4 levels "No HS","HS","Some College",..: 3 4 2 3 2 4 2 2 4 1 ...
##  $ age_label   : Factor w/ 4 levels "18-29","30-44",..: 1 2 4 1 2 4 2 4 3 3 ...
##  $ state_label : Factor w/ 51 levels "AL","AK","AZ",..: 7 39 31 7 33 33 39 20 33 40 ...
```

STA 610L

1988 ELECTIONS ANALYSIS

I will not do any meaningful EDA here.

I expect you to be able to do this yourself.

Let's just take a look at the amount of data we have for "bush" and the age:edu interaction.

```
###### Exploratory data analysis
table(polls_subset$bush) #well split by the two values
```

```
##
##    0    1
##  891 1124
```

```
table(polls_subset$edu,polls_subset$age)
```

```
##
##       1   2   3   4
##   1  44  42  67  96
##   2 232 283 223 116
##   3 141 205  99  54
##   4 119 285 125  62
```

STA 610L

1988 ELECTIONS ANALYSIS

As a start, we will consider a simple model with fixed effects of race and sex and a random effect for state (50 states + the District of Columbia).

$$\text{bush}_{ij}|\boldsymbol{x}_{ij} \sim \text{Bernoulli}(\pi_{ij}); \quad i = 1, \ldots, n; \quad j = 1, \ldots, J = 51;$$
$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + b_{0j} + \beta_1 \text{female}_{ij} + \beta_2 \text{black}_{ij};$$
$$b_{0j} \sim N(0, \sigma^2).$$

In R, we have

```
#library(lme4)
model1 <- glmer(bush ~ black+female+(1|state_label),
                family=binomial(link="logit"),
                data=polls_subset)
summary(model1)
```

1988 ELECTIONS ANALYSIS

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: bush ~ black + female + (1 | state_label)
##    Data: polls_subset
##
##      AIC      BIC   logLik deviance df.resid
##   2666.7   2689.1  -1329.3   2658.7     2011
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7276 -1.0871  0.6673  0.8422  2.5271
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  state_label (Intercept) 0.1692   0.4113
## Number of obs: 2015, groups:  state_label, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.44523    0.10139   4.391 1.13e-05
## black       -1.74161    0.20954  -8.312  < 2e-16
## female      -0.09705    0.09511  -1.020    0.308
##
## Correlation of Fixed Effects:
##        (Intr) black
## black  -0.119
## female -0.551 -0.005
```

STA 610L

1988 ELECTIONS ANALYSIS

Looks like we dropped some NAs.

```
c(sum(complete.cases(polls_subset)),sum(!complete.cases(polls_subset)))
```

```
## [1] 2015  178
```

Not ideal; we'll learn about methods for dealing with missing data soon.

Interpretation of results:

- For a fixed state (or across all states), a non-black male respondent has odds of $e^{0.45} = 1.57$ of supporting Bush.
- For a fixed state and sex, a black respondent as $e^{-1.74} = 0.18$ times (an 82% decrease) the odds of supporting Bush as a non-black respondent; you are much less likely to support Bush if your race is black compared to being non-black.
- For a given state and race, a female respondent has $e^{-0.10} = 0.91$ (a 9% decrease) times the odds of supporting Bush as a male respondent. However, this effect is not actually statistically significant!

1988 ELECTIONS ANALYSIS

The state-level standard deviation is estimated at 0.41, so that the states do vary some, but not so much.

I expect that you will be able to interpret the corresponding confidence intervals.

```
## Computing profile confidence intervals ...

##                   2.5 %      97.5 %
## .sig01       0.2608567  0.60403428
## (Intercept)  0.2452467  0.64871247
## black       -2.1666001 -1.34322366
## female      -0.2837100  0.08919986
```

1988 ELECTIONS ANALYSIS

We can definitely fit a more sophisticated model that includes other relevant survey factors, such as

- region
- prior vote history (note that this is a state-level predictor),
- age, education, and the interaction between them.

Given the structure of the data, it makes sense to include region as a second grouping variable.

We are yet to discuss that, so I will return to this later.

1988 ELECTIONS ANALYSIS

For now, let's just fit two models, one with the main effects for age and education, and the second with the interaction between them.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: bush ~ black + female + edu_label + age_label + (1 | state_label)
##    Data: polls_subset
## 
##      AIC      BIC   logLik deviance df.resid 
##   2662.2   2718.3  -1321.1   2642.2     2005 
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -1.8921 -1.0606  0.6420  0.8368  2.7906 
## 
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  state_label (Intercept) 0.1738   0.4168  
## Number of obs: 2015, groups:  state_label, 49
## 
## Fixed effects:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.31206    0.19438   1.605  0.10841
## black                  -1.74378    0.21124  -8.255  < 2e-16
## female                 -0.09681    0.09593  -1.009  0.31289
## edu_labelHS             0.23282    0.16569   1.405  0.15998
## edu_labelSome College  0.51598    0.17921   2.879  0.00399
## edu_labelCollege Grad  0.31585    0.17454   1.810  0.07036
## age_label30-44         -0.29222    0.12352  -2.366  0.01800
## age_label45-64         -0.06744    0.13738  -0.491  0.62352
## age_label65+           -0.22509    0.16142  -1.394  0.16318
```

Can you interpret the results?

STA 610L

1988 ELECTIONS ANALYSIS

```
model3 <- glmer(bush ~ black + female + edu_label*age_label + (1|state_label),
                family=binomial(link="logit"),data=polls_subset)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00802313 (tol = 0.002, component 1)
```

Why do we have a rank deficient model? Also, it looks like we have a convergence issue.

These issues can happen. We have so many parameters to estimate from the interaction terms `edu_label*age_label` (16 actually), and it looks like that's causing a problem.

We will revisit this example in a bit.

Note on estimation

ML estimation is carried out typically using adaptive Gaussian quadrature.

To improve accuracy over many package defaults (Laplace approximation), increase the number of quadrature points to be greater than one.

Note that some software packages (including the glmer function in the lme4 package) require Laplace approximation with Gaussian quadrature if the number of random effects is more than 1 for the sake of computational efficiency.

It is possible to tweak the optimizer in the glmer function in particular. Read more about the BOBYQA optimizer at your leisure.

Quick review: aggregated binary outcomes

In the context of logistic regression (and the mixed effect versions), we often observe the binary outcomes for each observation, that is, each $y_i \in \{0, 1\}$.

Of course this is not always the case. Sometimes, we get an aggregated version, with the outcome summed up by combinations of other variables. For example, suppose we had

response	0	0	1	1	1	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	0	1	1	0	1
predictor	3	3	2	1	2	3	2	2	2	2	3	1	3	1	1	2	2	2	2	1	3	3	3	1	3

where predictor is a factor variable with 3 levels: 1,2,3.

Quick review: aggregated binary outcomes

The aggregated version of the same data could look then like

predictor	n	successes
1	31	17
2	35	16
3	34	14

STA 610L

Quick review: aggregated binary outcomes

Recall that if $Y \sim \text{Bin}(n, p)$ (that is, Y is a random variable that follows a binomial distribution with parameters n and p), then Y follows a $\text{Bernoulli}(p)$ distribution when $n = 1$.

Alternatively, we also have that if $Z_1, \ldots, Z_n \sim \text{Bernoulli}(p)$, then $Y = \sum_i^n Z_i \sim \text{Bin}(n, p)$.

That is, the sum of n "iid" $\text{Bernoulli}(p)$ random variables gives a random variable with the $\text{Bin}(n, p)$ distribution.

Quick review: aggregated binary outcomes

The logistic regression model can be used either for Bernoulli data (as we have done so far) or for data summarized as binomial counts (that is, aggregated counts).

In the aggregated form, the model is a Binomial logistic regression:

$$y_i|x_i \sim \text{Bin}(n_i, \pi_i); \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

STA 610L

Quick review: Bernoulli versus binomial outcomes

Normally, for individual-level data, we would have

```
##   response predictor
## 1        0         3
## 2        0         3
## 3        1         2
## 4        1         1
## 5        1         2
## 6        0         3
```

```
M1 <- glm(response~predictor,data=Data,family=binomial)
summary(M1)
```

```
##
## Call:
## glm(formula = response ~ predictor, family = binomial, data = Data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.261  -1.105  -1.030   1.251   1.332
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1942     0.3609   0.538    0.591
## predictor2   -0.3660     0.4954  -0.739    0.460
## predictor3   -0.5508     0.5017  -1.098    0.272
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.27  on 99  degrees of freedom
## Residual deviance: 137.02  on 97  degrees of freedom
## AIC: 143.02
##
## Number of Fisher Scoring iterations: 4
```

STA 610L

Quick review: Bernoulli versus binomial outcomes

But we could also do the following with the aggregate level data instead

```
M2 <- glm(cbind(successes,n-successes)~predictor,data=Data_agg,family=binomial)
summary(M2)
```

```
##
## Call:
## glm(formula = cbind(successes, n - successes) ~ predictor, family = binomial,
##     data = Data_agg)
##
## Deviance Residuals:
## [1]  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1942     0.3609   0.538    0.591
## predictor2   -0.3660     0.4954  -0.739    0.460
## predictor3   -0.5508     0.5017  -1.098    0.272
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.2524e+00  on 2  degrees of freedom
## Residual deviance: 1.3323e-14  on 0  degrees of freedom
## AIC: 17.868
##
## Number of Fisher Scoring iterations: 2
```

Same results overall! Deviance and AIC are different because of the slightly different likelihood functions.

Note that some glm functions use n in the formula instead of n-successes.

STA 610L

Another example: Berkeley admissions

With that in mind, we can move forward to our next example.

We will use this next example to also start to illustrate how to fit Bayesian versions of generalized linear mixed effects models.

However, note that we can fit the frequentist versions of the same models using the lme4 package.

In fall 1973, the University of California, Berkeley's graduate division admitted 44% of male applicants and 35% of female applicants.

School administrators were concerned about the potential for bias (and lawsuits!) and asked statistics professor Peter Bickel to examine the data more carefully.

We have a subset of the admissions data for 6 departments.

Berkeley admissions

```
library(rethinking)
data(UCBadmit)
d <- UCBadmit
detach(package:rethinking,unload=T)
library(brms)
d <-
  d%>%
  mutate(male=ifelse(applicant.gender=="male",1,0),
         dept_id = rep(1:6, each = 2))
d$successrate=d$admit/d$applications
sum(d$admit[d$male==1])/sum(d$applications[d$male==1])
```

```
## [1] 0.4451877
```

```
sum(d$admit[d$male==0])/sum(d$applications[d$male==0])
```

```
## [1] 0.3035422
```

We see in this subset of departments that roughly 45% of male applicants were admitted, while only 30% of female applicants were admitted.

Berkeley admissions

Because admissions decisions for graduate school are made on a departmental level (not at the school level), it makes sense to examine results of applications by department.

```
d[,c(1,2,3,4,7)]
```

```
##    dept applicant.gender admit reject dept_id
## 1     A             male   512    313       1
## 2     A           female    89     19       1
## 3     B             male   353    207       2
## 4     B           female    17      8       2
## 5     C             male   120    205       3
## 6     C           female   202    391       3
## 7     D             male   138    279       4
## 8     D           female   131    244       4
## 9     E             male    53    138       5
## 10    E           female    94    299       5
## 11    F             male    22    351       6
## 12    F           female    24    317       6
```

Hmm, what's going on here?

Berkeley admissions

Following McElreath's analysis in *Statistical Rethinking*, we start fitting a simple logistic regression model and examine diagnostic measures.

The model for department i and gender j with $n_{admit,ij}$ of n_{ij} applicants admitted is given as:

$$n_{admit,ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$$
$$\text{logit}(\pi_{ij}) = \alpha + \beta \text{male}_{ij},$$

where $\alpha \sim N(0, 10)$ and $\beta \sim N(0, 1)$.

STA 610L

Another example:

```r
adm1 <-
  brm(data = d, family = binomial,
      admit | trials(applications) ~ 1 + male ,
      prior = c(prior(normal(0, 10), class = Intercept),
                prior(normal(0, 1), class = b)),
      iter = 2500, warmup = 500, cores = 2, chains = 2,
      seed = 10)
summary(adm1)
```

```
##  Family: binomial
##   Links: mu = logit
## Formula: admit | trials(applications) ~ 1 + male
##    Data: d (Number of observations: 12)
## Samples: 2 chains, each with iter = 2500; warmup = 500; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -0.83      0.05    -0.93    -0.73 1.00     2207     2217
## male          0.61      0.07     0.48     0.73 1.00     2837     2702
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Here it appears male applicants have $e^{0.61} = 1.8$ (95% credible interval (1.6, 2.1)) times the odds of admission as female applicants.

STA 610L

Another example:

We can also put this on the probability scale.

```
post <- posterior_samples(adm1)

post %>%
  mutate(p_admit_male   = inv_logit_scaled(b_Intercept + b_male),
         p_admit_female = inv_logit_scaled(b_Intercept),
         diff_admit     = p_admit_male - p_admit_female) %>%
  summarise(`2.5%`  = quantile(diff_admit, probs = .025),
            `50%`   = median(diff_admit),
            `97.5%` = quantile(diff_admit, probs = .975))
```

```
##        2.5%       50%     97.5%
## 1 0.1122369 0.1414303 0.1690868
```

Overall it appears the median probability of admission was 14 percentage points higher for males.

STA 610L

Model Checking

Here we take some posterior predictions and plot against the observed proportions in the data.

Here's the code to do that:

```
library(wesanderson)
library(dutchmasters)
library(ggplot2)
d <-
  d %>%
  mutate(case = factor(1:12))

p <-
  predict(adm1) %>%
  as_tibble() %>%
  bind_cols(d)

d_text <-
  d %>%
  group_by(dept) %>%
  summarise(case  = mean(as.numeric(case)),
            admit = mean(admit / applications) + .05)
```

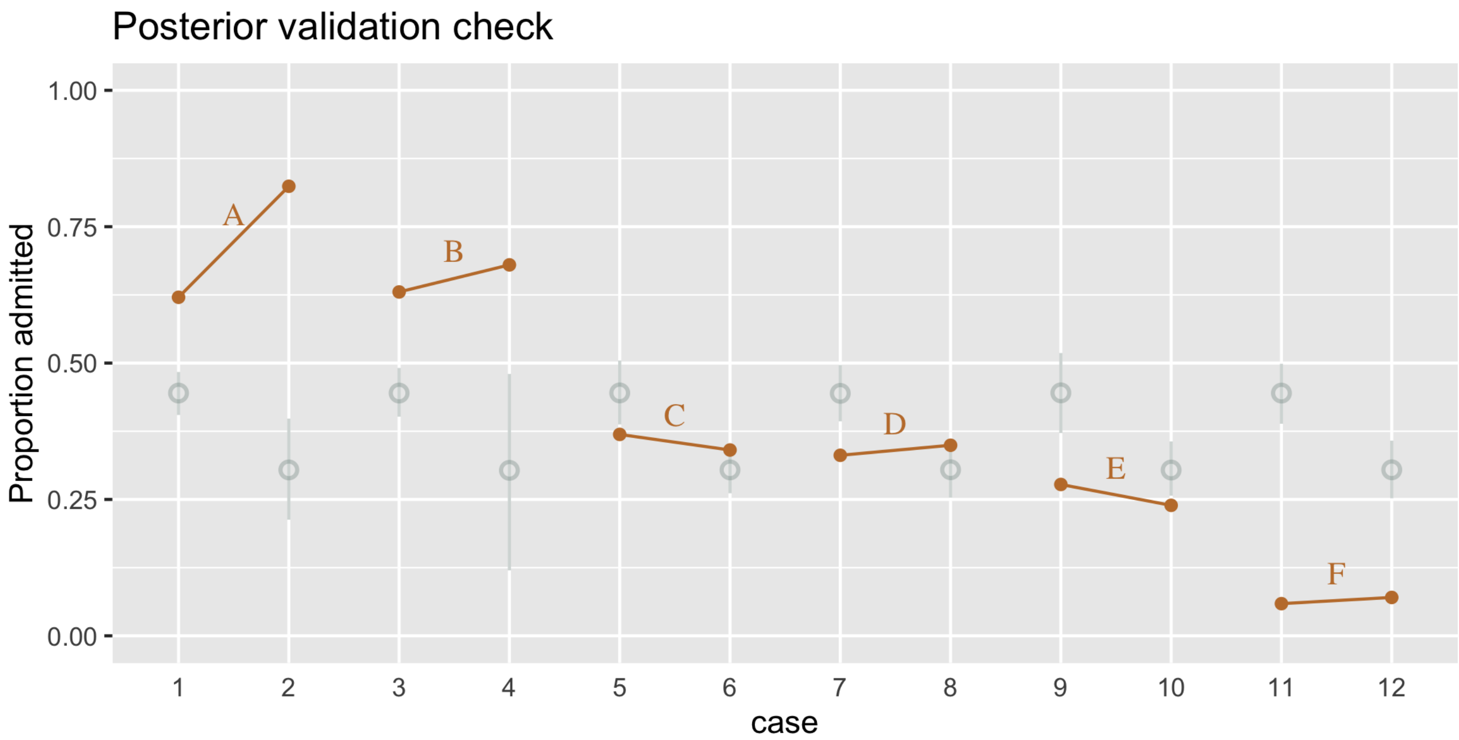
STA 610L

Model Checking

..and the rest of the code:

```
ggplot(data = d, aes(x = case, y = admit / applications)) +
  geom_pointrange(data = p,
                  aes(y    = Estimate / applications,
                      ymin = Q2.5     / applications ,
                      ymax = Q97.5    / applications),
                  color = wes_palette("Moonrise2")[1],
                  shape = 1, alpha = 1/3) +
  geom_point(color = wes_palette("Moonrise2")[2]) +
  geom_line(aes(group = dept),
            color = wes_palette("Moonrise2")[2]) +
  geom_text(data = d_text,
            aes(y = admit, label = dept),
            color = wes_palette("Moonrise2")[2],
            family = "serif") +
  coord_cartesian(ylim = 0:1) +
  labs(y     = "Proportion admitted",
       title = "Posterior validation check") +
  theme(axis.ticks.x = element_blank())
```

Model Checking



The orange lines connect observed proportions admitted in each department (odd numbers indicate males; even females).

The grey circles indicate point and interval estimates of the model-predicted proportion admitted. Clearly the model fits the data poorly.

Varying/random intercepts

Based on the plot, we have some big departmental differences.

Let's specify department as a random effect in the model.

$$n_{admit,ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \alpha_{0i} + \beta \text{male}_{ij}$$

$$\alpha_{0i} \sim N(\alpha, \sigma^2); \quad \sigma^2 \sim \text{HalfCauchy}(0, 1)$$

$$\alpha \sim N(0, 10) \quad \text{and} \quad \beta \sim N(0, 1).$$

Varying/random intercepts

```
adm2 <-
  brm(data = d, family = binomial,
      admit | trials(applications) ~ 1 + male + (1 | dept_id),
      prior = c(prior(normal(0, 10), class = Intercept),
                prior(normal(0, 1), class = b),
                prior(cauchy(0, 1), class = sd)),
      iter = 4500, warmup = 500, chains = 3, cores = 3,
      seed = 13,
      control = list(adapt_delta = 0.99))
```

STA 610L

Varying/random intercepts

```
## Compiling Stan program...

## Start sampling

## Inference for Stan model: f9cec24254cb76a5ed974b425b0c8035.
## 3 chains, each with iter=4500; warmup=500; thin=1;
## post-warmup draws per chain=4000, total post-warmup draws=12000.
##
##                          mean se_mean   sd   2.5%    25%    50%    75%  97.5%
## b_Intercept             -0.60    0.01 0.61  -1.81  -0.95  -0.59  -0.24   0.61
## b_male                  -0.10    0.00 0.08  -0.26  -0.15  -0.10  -0.04   0.07
## sd_dept_id__Intercept    1.39    0.01 0.54   0.76   1.04   1.26   1.59   2.79
## r_dept_id[1,Intercept]   1.27    0.01 0.61   0.04   0.92   1.27   1.63   2.50
## r_dept_id[2,Intercept]   1.23    0.01 0.61   0.00   0.87   1.22   1.58   2.46
## r_dept_id[3,Intercept]   0.01    0.01 0.61  -1.21  -0.34   0.02   0.37   1.25
## r_dept_id[4,Intercept]  -0.02    0.01 0.61  -1.24  -0.37  -0.02   0.34   1.22
## r_dept_id[5,Intercept]  -0.46    0.01 0.61  -1.70  -0.82  -0.46  -0.10   0.77
## r_dept_id[6,Intercept]  -2.01    0.01 0.62  -3.26  -2.36  -2.00  -1.64  -0.77
## lp__                   -62.06    0.05 2.48 -67.82 -63.47 -61.69 -60.27 -58.22
##                        n_eff Rhat
## b_Intercept             2125    1
## b_male                  4830    1
## sd_dept_id__Intercept   1813    1
## r_dept_id[1,Intercept]  2124    1
## r_dept_id[2,Intercept]  2133    1
## r_dept_id[3,Intercept]  2125    1
## r_dept_id[4,Intercept]  2124    1
## r_dept_id[5,Intercept]  2148    1
## r_dept_id[6,Intercept]  2224    1
## lp__                    2701    1
##
## Samples were drawn using NUTS(diag_e) at Wed Mar 24 08:50:18 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

In this model we see no evidence of a difference in admissions probabilities by gender though we do see big departmental variability.

Random slopes?

How about random slopes for gender by department?

```
adm3 <-
  brm(data = d, family = binomial,
      admit | trials(applications) ~ 1 + male + (1 + male | dept_id),
      prior = c(prior(normal(0, 10), class = Intercept),
                prior(normal(0, 1), class = b),
                prior(cauchy(0, 1), class = sd),
                prior(lkj(2), class = cor)),
      iter = 5000, warmup = 1000, chains = 4, cores = 4,
      seed = 13,
      control = list(adapt_delta = .99,
                     max_treedepth = 12))
adm3$fit
```

STA 610L

Random slopes?

```
## Compiling Stan program...

## Start sampling

## Inference for Stan model: a035d956cf1fd75687fe3dffeff8956b.
## 4 chains, each with iter=5000; warmup=1000; thin=1;
## post-warmup draws per chain=4000, total post-warmup draws=16000.
##
##                              mean se_mean   sd   2.5%    25%    50%    75%
## b_Intercept                 -0.51    0.01 0.68  -1.84  -0.91  -0.50  -0.11
## b_male                      -0.16    0.00 0.22  -0.61  -0.29  -0.15  -0.03
## sd_dept_id__Intercept        1.56    0.01 0.57   0.86   1.17   1.43   1.78
## sd_dept_id__male             0.46    0.00 0.23   0.15   0.31   0.42   0.56
## cor_dept_id__Intercept__male -0.33   0.00 0.34  -0.86  -0.59  -0.36  -0.10
## r_dept_id[1,Intercept]       1.79    0.01 0.71   0.43   1.36   1.78   2.22
## r_dept_id[2,Intercept]       1.25    0.01 0.72  -0.16   0.80   1.23   1.68
## r_dept_id[3,Intercept]      -0.13    0.01 0.68  -1.47  -0.53  -0.15   0.27
## r_dept_id[4,Intercept]      -0.11    0.01 0.68  -1.44  -0.51  -0.11   0.29
## r_dept_id[5,Intercept]      -0.62    0.01 0.68  -1.96  -1.02  -0.63  -0.21
## r_dept_id[6,Intercept]      -2.09    0.01 0.69  -3.47  -2.50  -2.08  -1.67
## r_dept_id[1,male]           -0.61    0.00 0.31  -1.28  -0.80  -0.59  -0.39
## r_dept_id[2,male]           -0.05    0.00 0.33  -0.71  -0.25  -0.05   0.15
## r_dept_id[3,male]            0.24    0.00 0.24  -0.22   0.08   0.22   0.38
## r_dept_id[4,male]            0.07    0.00 0.24  -0.41  -0.08   0.06   0.21
## r_dept_id[5,male]            0.27    0.00 0.26  -0.21   0.10   0.26   0.43
## r_dept_id[6,male]            0.04    0.00 0.31  -0.58  -0.15   0.04   0.23
## lp__                       -65.53    0.07 3.72 -73.90 -67.78 -65.14 -62.84
##                             97.5% n_eff Rhat
## b_Intercept                  0.83  3751    1
## b_male                       0.27  6301    1
## sd_dept_id__Intercept        3.03  4867    1
## sd_dept_id__male             1.01  5224    1
## cor_dept_id__Intercept__male 0.41  9857    1
## r_dept_id[1,Intercept]       3.20  3771    1
## r_dept_id[2,Intercept]       2.68  4215    1
## r_dept_id[3,Intercept]       1.20  3737    1
## r_dept_id[4,Intercept]       1.23  3747    1
## r_dept_id[5,Intercept]       0.72  3820    1
## r_dept_id[6,Intercept]      -0.72  3962    1
## r_dept_id[1,male]           -0.06  7500    1
## r_dept_id[2,male]            0.63 11973    1
## r_dept_id[3,male]            0.75  7256    1
## r_dept_id[4,male]            0.56  6909    1
## r_dept_id[5,male]            0.83  7388    1
## r_dept_id[6,male]            0.65 10417    1
## lp__                       -59.40  3279    1
##
```

Diagnostics

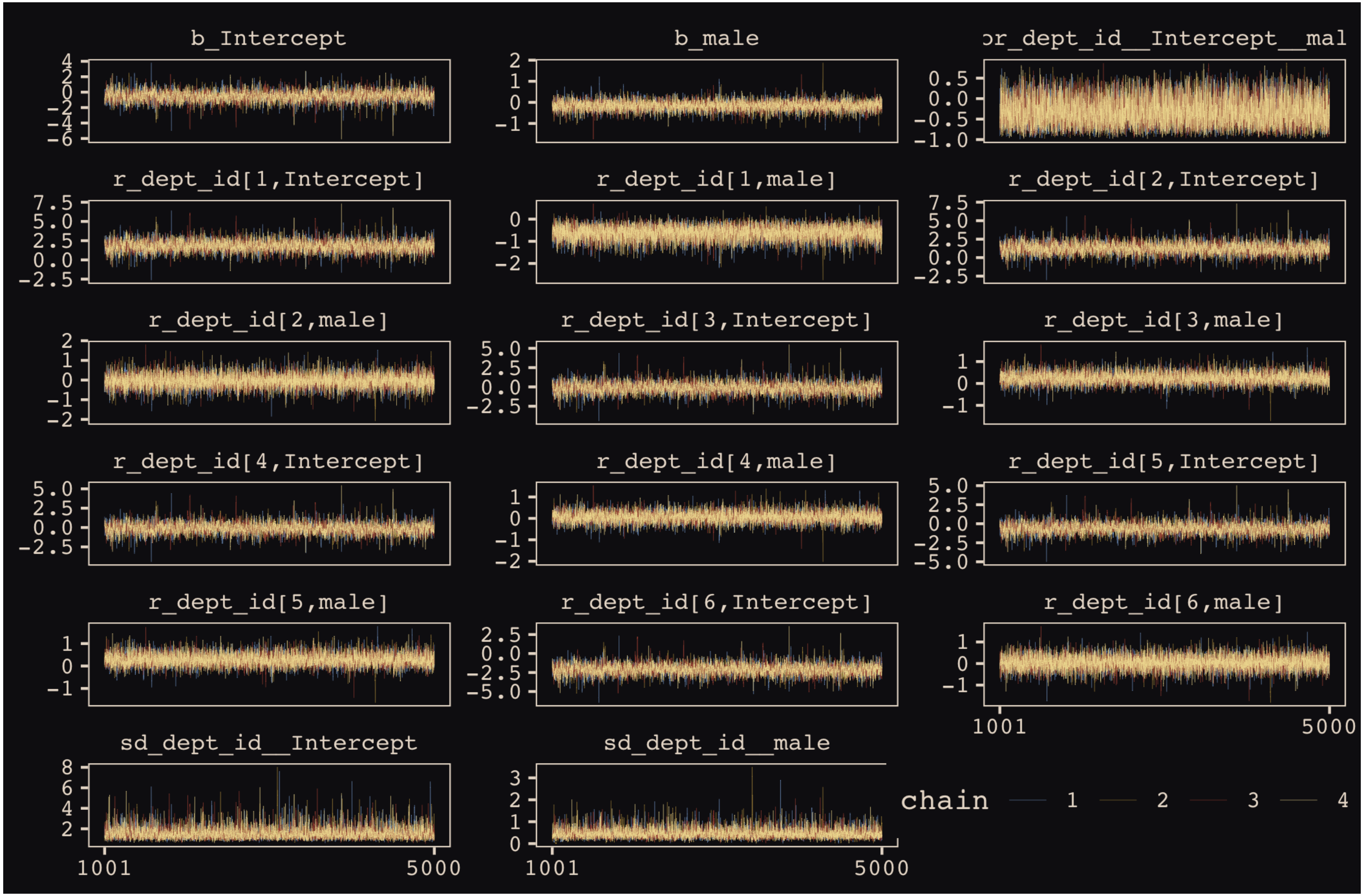
Before we get too excited let's take a quick look at the trace plots.

```
post <- posterior_samples(adm3, add_chain = T)
post <- post[,!is.element(colnames(post),c("lp__"))]

post %>%
  gather(key, value, -chain, -iter) %>%
  mutate(chain = as.character(chain)) %>%
  ggplot(aes(x = iter, y = value, group = chain, color = chain)) +
  geom_line(size = 1/15) +
  scale_color_manual(values = c("#80A0C7", "#B1934A", "#A65141", "#EEDA9D")) +
  scale_x_continuous(NULL, breaks = c(1001, 5000)) +
  ylab(NULL) +
  theme_pearl_earring +
  theme(legend.position  = c(.825, .06),
        legend.direction = "horizontal") +
  facet_wrap(~key, ncol = 3, scales = "free_y")
```

STA 610L

Diagnostics

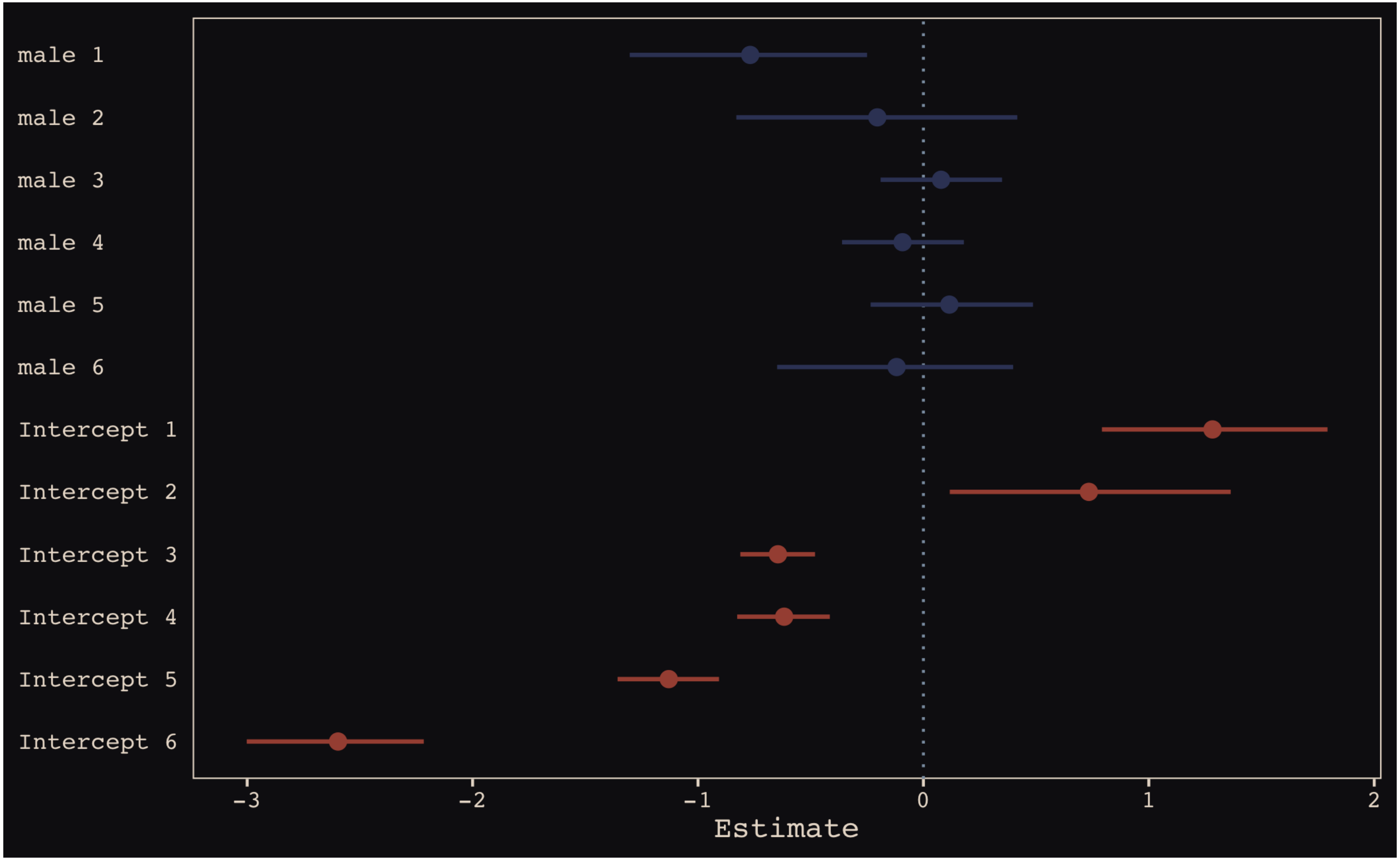


Random effects

```r
rbind(coef(adm3)$dept_id[, , 1],
      coef(adm3)$dept_id[, , 2]) %>%
  as_tibble() %>%
  mutate(param   = c(paste("Intercept", 1:6), paste("male", 1:6)),
         reorder = c(6:1, 12:7)) %>%

  # plot
  ggplot(aes(x = reorder(param, reorder))) +
  geom_hline(yintercept = 0, linetype = 3, color = "#8B9DAF") +
  geom_pointrange(aes(ymin = Q2.5, ymax = Q97.5, y = Estimate, color = reorder < 7),
                  shape = 20, size = 3/4) +
  scale_color_manual(values = c("#394165", "#A65141")) +
  xlab(NULL) +
  coord_flip() +
  theme_pearl_earring +
  theme(legend.position = "none",
        axis.ticks.y    = element_blank(),
        axis.text.y     = element_text(hjust = 0))
```

Random effects



We see much more variability in the random intercepts than in the random slopes.

What happened at Berkeley?

What happened at Berkeley? It actually doesn't require too much sophisticated modeling.

What we are seeing is just Simpson's paradox.

```
d[,c(1,2,3,4,8)]
```

```
##    dept applicant.gender admit reject successrate
## 1     A             male   512    313  0.62060606
## 2     A           female    89     19  0.82407407
## 3     B             male   353    207  0.63035714
## 4     B           female    17      8  0.68000000
## 5     C             male   120    205  0.36923077
## 6     C           female   202    391  0.34064081
## 7     D             male   138    279  0.33093525
## 8     D           female   131    244  0.34933333
## 9     E             male    53    138  0.27748691
## 10    E           female    94    299  0.23918575
## 11    F             male    22    351  0.05898123
## 12    F           female    24    317  0.07038123
```

What happened at Berkeley?

In the raw data, women had higher acceptance probabilities in 4 of the 6 departments.

However, the departments to which they applied in higher numbers were the departments that had lower overall acceptance rates.

What happened is that women were more likely to apply do departments like English, which have trouble supporting grad students, and they were less likely to apply to STEM departments, which had more plentiful funding for graduate students.

The men, on the other hand, were much more likely to apply to the STEM departments that had higher acceptance rates.

What's next?

Move on to the readings for the next module!

STA 610L