STA 610L: Module 4.6

Poststratification and weighting

Dr. Olanrewaju Michael Akande

Election predictions

One interesting application of hierarchical models has been in the realm of election predictions.

We are focusing on this subject area because of interesting elements involving hierarchical logistic regression models, missing data, and biased samples.

Common data sources for election data:

- Polls,
- Historical Data (who won),
- Voter Records and Turnout (see Case Study II).

Predictions based on current polling data

One strategy for predicting the winner in an election would be to use the latest aggregate polling data from a reputable source, e.g. FiveThirtyEight or The New York Times.

- Polls may use land lines (robocalls), cell phones, or web surveys.
- Some polls are of likely voters, while others may not restrict to this group (prediction opinions of the population is great -- but voters determine an election outcome).
- Polls vary in quality (you can find some quality ratings online).

Predictions based on current polling data

- Polls may be subject to bias (e.g., nonresponse bias, assumptions involved in determining "likely voters").
- Polls are usually associated with margins of error.
- More polls focus on national sentiments than on state-specific sentiments (state-specific sentiments are important for state and local elections, e.g. the House and Senate representatives, as well as the presidential election, which is decided not by popular vote but by the electoral college).

Historical data incorporation

- Use historical data from past elections, e.g. within a House district, within a state, nationally.
- Useful when a state consistently votes for the same party but less useful for swing states, which are closer to call.
- Check out the NC Voter Record.

Voter turnout data

- States like NC make available data on who votes in each election (Case Study II).
- Voter turnout data can be used to construct a voting history for certain districts, demographic groups, etc.
- Locations with low or variable turnout often harder to predict.
- Many better predictions take voter turnout data into account in some way.

Sources of election uncertainty

- Sample sizes of polls (often $n = 1000$ in national polls).
- Individual changes in turnout.
- Systematic changes in turnout (different turnout patterns from historical records, e.g. more young people vote).
- Individual variation in support (undecided voters).
- Unmeasured bias in polls.

Working with polls

- Polls are a great source of information!
- Polls are fraught with challenges (do you hang up when a pollster calls?).
- If our polling data were perfectly representative of the population of voters, predicting election outcomes would be easier.
- Unfortunately, polls are often too small to cover as many heterogeneous population subgroups as we'd wish.
- Let's consider some important methods for using polling data: poststratification and weighting.

Why use poststratification and weighting?

- Sometimes we would like to stratify on a key variable, e.g. political party affiliation, but we cannot place the units into their correct strata until the units are sampled.

  For example, in aggregate polling data we cannot determine if an individual is a Democrat or Republican until after they've been polled (at which point we ask their affiliation).

- So, we often use poststratification, or stratification after the selection of a sample, to handle this type of data.

- We can also use poststratification and weighting if the sample is not representative of the population such as when we have nonresponse bias.

- In general, poststratification and weighting are used to obtain more accurate estimates from survey data.

How it works - population distribution

The first step in poststratification is knowing the ratio of the size of a stratum in question to that of the relevant population size.

For example, what proportion of eligible voters in Durham County, North Carolina who are Asian females aged 20-25 identify as Republicans?

We will denote this as $\frac{N_h}{N}$ where N_h is the number of individuals in stratum h, and N is the total number of individuals in the full population.

How it works - poststratifying and weighting

Suppose we have only two strata: Democrats and Republicans.

Once we know the distribution of factors of interest in our population, we can apply poststratifying and weighting.

Essentially, we calculate a weighted sum in which the weights are the relevant proportions:

$$\bar{y}_{poststrat.} = \frac{N_{democrat}}{N} \cdot \bar{y}_{democrat} + \frac{N_{republican}}{N} \cdot \bar{y}_{republican}.$$

Now we have a weighted estimate, $\bar{y}_{poststrat.}$, of our population mean.

This adjustment is important if, for example, Democrats are more likely than Republicans to participate in a survey, but we want to obtain an estimate that is generalizable to the entire population.

Example

Suppose a middle school math club wants to estimate the proportion of undergraduate students in the RTP area who agree with the statement Mike Krzyzewski (Coach K) is the all-time best college basketball coach.

They take a small sample of 100 undergraduate students at the Target on 15-501 to address this question.

Do you see any challenges that may arise due to the study design?

Strata

Suppose the RTP area contains N=60,000 undergraduates, with 24,000 at NC State, 19,000 at UNC, 6500 each at Duke and NCCU, and 4000 elsewhere (Meredith, Shaw, etc.).

Then the proportions $\frac{N_h}{N}$ of interest are roughly 0.40 for NC State, 0.32 for UNC, 0.11 for Duke, 0.11 for NCCU, and 0.06 for other universities (proportions rounded and forced to sum to 1 for convenience).

Results

Suppose the math club carries out the survey, obtaining the following *proportions* of 100 surveyed students endorsing the statement:

- 1.0 among 50 Duke students surveyed.
- 0.0 among 25 UNC students surveyed.
- 0.6 among 20 NCCU students surveyed.
- 0.5 among 4 NC State students surveyed.
- 1.0 from the single student from another university surveyed.

This means that overall among the 100 students surveyed, 65 endorsed the statement.

Is 65% a valid estimate of the percent of undergraduates in RTP who believe Coach K is the best ever?

Probably not!

Students who live close to the 15-501 Target are more likely to attend the nearby schools, and Duke is one of the closest schools to the Target.

What happens if we use poststratification to re-weight our estimate by the representation of each university among area undergraduates (rather than its representation in our biased sample)?

$$\widehat{\pi} = 1.0\left(\frac{6500}{60000}\right) + 0.0\left(\frac{19000}{60000}\right) + 0.6\left(\frac{6500}{60000}\right)$$
$$+0.5\left(\frac{24000}{60000}\right) + 1.0\left(\frac{4000}{60000}\right) = 0.44$$

Notes

Much smaller than we had before.

However, our estimate is really sensitive to having only 1 student from the other colleges -- if that student had not liked Coach K, our estimate would have been 0.37 instead of 0.44.

For this reason, many surveys will oversample individuals from small or highly variable groups in order to obtain more stable estimates (e.g., we could have taken fewer Duke and UNC students, large groups who are more homogeneous in their thoughts about Coach K, and instead focused our efforts on less predictable groups).

This is the basic principle underlying many methods of election prediction.

Multilevel regression and poststratification

- It is often of interest to researchers to consider state-level opinion, in addition to/instead of national-level opinion.
- Finding surveys that are uniform across all or most states is extremely challenging, and states done for one state sometimes are of lower quality than national-level surveys.
- One method of estimating state-level opinion using national survey data is called **multilevel regression and poststratification ("Mr. P")**.
- We will compare this approach with a simple approach of using empirical means and poststratifying without borrowing information across groups (that's what we did on the previous slides).

Load packages

```
library(tidyverse)
library(lme4)
library(brms)
library(rstan)
library(cowplot) # plotting
library(dplyr)
library(directlabels)
library(tidybayes) #work easily with posterior samples
rstan_options(auto_write=TRUE)
options(mc.cores=parallel::detectCores())
```

Multilevel modeling with poststratification

First, we use multilevel regression to model individual survey responses as a function of demographic and geographic predictors.

Then we use poststratification, in which we weight (poststratify) the estimates for each demographic-geographic respondent type by the percentages of each type in the actual state populations.

This tutorial draws heavily on Jonathan Kastellec's MrP primer and Tim Mastny's version using Stan.

You may find the paper at their website useful in addition to the shorter version presented here.

First, download three important datasets from Sakai: gay_marriage_megapoll.dta, state_level_update.dta, and poststratification 2000.dta.

Analysis goal

The goal is to estimate support for gay marriage in each state based on survey data that are potentially non-representative.

Because not all subgroups of the population are equally likely to respond to polls, we worry that relying only on a survey could lead to biased estimates of support for gay marriage.

For example, younger people may be more likely than older people to answer questions about gay marriage.

The US Census is a good source of information about characteristics of the full US population.

Using Census data, we can scale the average support of population subgorups of interest in proportion to their representation in the state population.

Data

- A compilation of national gay marriage polls will provide information on support of gay marriage.

  Five national polls were conducted in 2004 and include information on state, gender, race/ethnicity, education, age, and party identification.

- State level data provide information including % of religious voters, voting history (Democrat or Republican), etc.

- Census data will be used to estimate the % of voters in subgroups in the state, given that poll respondents may not mirror the demographics of voting-age citizens.

  Ultimately, we need a dataset of the population counts for each subgroup, e.g. how many African Americans aged 18-25 went to college and reside in NC.

  For this tutorial, we will use the 5% Public Use Microdata Sample from the 2000 census.

Data

```
marriage.data <- foreign::read.dta('data/gay_marriage_megapoll.dta',
                                   convert.underscore=TRUE)

Statelevel <- foreign::read.dta('data/state_level_update.dta',
                                convert.underscore=TRUE)

Census <- foreign::read.dta('data/poststratification 2000.dta',
                            convert.underscore=TRUE)
```

Data munging

```
#rename to state in preparation for merging
Statelevel  <- Statelevel %>% rename(state=sstate)
marriage.data <- Statelevel %>%
  select(state,p.evang,p.mormon,kerry.04) %>%
  right_join(marriage.data)
```

```
## Joining, by = "state"
```

In this step we are combining state-level data on the percentage of evangelical Christians (often conservative), the percentage of members of the LDS church (also often conservative), and the vote share for John Kerry in 2004 (losing Democratic nominee for President) with the individual-level polling data on gay marriage.

More data munging

```
# combine demographic groups and label them
marriage.data$race.female <- (marriage.data$female *3) + marriage.data$race.wbh
marriage.data$race.female <- factor(marriage.data$race.female,levels=1:6,
                  labels=c("WhMale","BlMale","HMale","WhFem","BlFem","HFem"))
marriage.data$age.edu.cat <- 4*(marriage.data$age.cat -1) + marriage.data$edu.cat
marriage.data$age.edu.cat <- factor(marriage.data$age.edu.cat,levels=1:16,
                        labels=c("18-29,<HS","18-29,HS","18-29,SC","18-29,CG",
                                 "30-44,<HS","30-44,HS","30-44,SC","30-44,CG",
                                 "45-64,<HS","45-64,HS","45-64,SC","45-64,CG",
                                 "65+,<HS","65+,HS","65+,SC","65+,CG"))
marriage.data$p.evang <- Statelevel$p.evang[marriage.data$state.initnum]
# proportion of evangelicals in respondent's state
marriage.data$p.mormon <-Statelevel$p.mormon[marriage.data$state.initnum]
# proportion of LDS church members in respondent's state
marriage.data$p.relig <- marriage.data$p.evang + marriage.data$p.mormon
# combined evangelical + LDS proportions
marriage.data$kerry.04 <- Statelevel$kerry.04[marriage.data$state.initnum]
# John Kerry's % of 2-party vote in respondent's state in 2004
marriage.data <- marriage.data %>%
  filter(state!="")
```

Data manipulation

Here we prepare the Census data for merging.

```
# Census variables
Census <- Census %>%
  rename(state=cstate, age.cat=cage.cat, edu.cat=cedu.cat,region=cregion)
Census$race.female <- (Census$cfemale *3) + Census$crace.WBH
Census$race.female <- factor(Census$race.female,levels=1:6,
                labels=c("WhMale","BlMale","HMale","WhFem","BlFem","HFem"))
Census$age.edu.cat <- 4 * (Census$age.cat-1) + Census$edu.cat
Census$age.edu.cat <- factor(Census$age.edu.cat,levels=1:16,
                      labels=c("18-29,<HS","18-29,HS","18-29,SC","18-29,CG",
                               "30-44,<HS","30-44,HS","30-44,SC","30-44,CG",
                               "45-64,<HS","45-64,HS","45-64,SC","45-64,CG",
                               "65+,<HS","65+,HS","65+,SC","65+,CG"))
Census <- Statelevel %>%
  select(state,p.evang,p.mormon,kerry.04) %>%
  right_join(Census)
```

```
## Joining, by = "state"
```

```
Census <- Census %>% mutate(p.relig=p.evang+p.mormon)
```

Who participated in the polls?

Let's consider South Dakota as an example. Of the poll responders, 18% were white males aged 45-64 with a high school degree, and 13.6% were white females aged 65+ with a high school degree. None of the poll responders identified as black or Hispanic.

```
marriageSD <- subset(marriage.data,state=="SD")
table(marriageSD$age.edu.cat,marriageSD$race.female)/length(marriageSD$race.female)
```

```
##
##                 WhMale     BlMale      HMale      WhFem      BlFem       HFem
##   18-29,<HS 0.00000000 0.00000000 0.00000000 0.04545455 0.00000000 0.00000000
##   18-29,HS  0.04545455 0.00000000 0.00000000 0.04545455 0.00000000 0.00000000
##   18-29,SC  0.04545455 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   18-29,CG  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   30-44,<HS 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   30-44,HS  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   30-44,SC  0.04545455 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   30-44,CG  0.04545455 0.00000000 0.00000000 0.09090909 0.00000000 0.00000000
##   45-64,<HS 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   45-64,HS  0.18181818 0.00000000 0.00000000 0.09090909 0.00000000 0.00000000
##   45-64,SC  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   45-64,CG  0.04545455 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   65+,<HS   0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##   65+,HS    0.00000000 0.00000000 0.00000000 0.13636364 0.00000000 0.00000000
##   65+,SC    0.00000000 0.00000000 0.00000000 0.04545455 0.00000000 0.00000000
##   65+,CG    0.00000000 0.00000000 0.00000000 0.13636364 0.00000000 0.00000000
```

Who participated in the polls?

According to the Census, there are some black and Hispanic residents of South Dakota.

White males aged 45-64 with a high school degree make up 5.5% of the population, and white females aged 65+ with a high school degree make up 4.6% of the population.

These groups were vastly overrepresented in the survey.

```
CensusSD <- subset(Census,state=="SD")
CensusSD[,c(11,13,14)]
```

Obtain estimates based on empirical averages

First we calculate the mean responses within each state -- we will call these disaggregated estimates.

```
# Get state averages
mod.disag <- marriage.data%>%
  group_by(state) %>%
  summarise(support=mean(yes.of.all)) %>%
  mutate(model="no_ps")
```

These averages will not be representative of the actual statewide means if the sampled respondents are not in proportion to each group's percentage of the total state population and the groups differ with respect to their support level.

So we will next poststratify.

Poststratifying sample estimates

First, we find within-group averages in each state.

```
# Find average within each group
grp.means <- marriage.data%>%
  group_by(state,region,race.female,age.edu.cat,p.relig,kerry.04) %>%
  summarize(support=mean(yes.of.all,na.rm=TRUE))
```

Next we add the group's percentage in each state.

```
grp.means <- Census %>%
  select(state, region, kerry.04, race.female, age.edu.cat, p.relig,
         cpercent.state) %>%
  right_join(grp.means)
```

Sum scaled average and get total state averages.

```
mod.disag.ps <- grp.means %>%
  group_by(state) %>%
  summarize(support=sum(support * cpercent.state, na.rm=TRUE)) %>%
  mutate(model="ps")
```

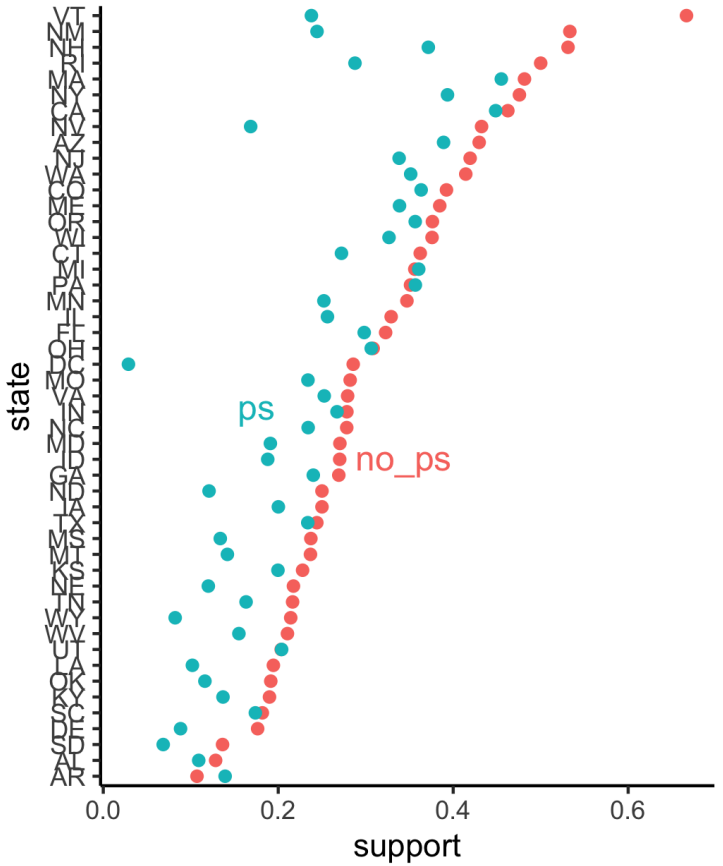
Define function for plotting (may want to use again)

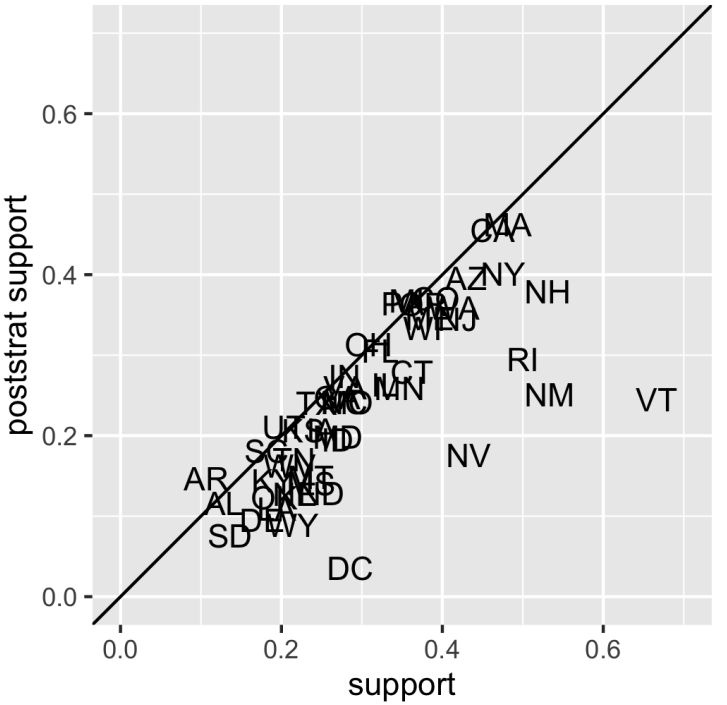
```
#make a function so we don't have to type over and over
compare_scat <- function(d){
  return(
    ggplot(data=d, aes(x=support...2 ,y=support...3))+
      geom_text(aes(label=state),hjust=0.5,vjust=0.25) +
      geom_abline(slope=1,intercept=0) +
      xlim(c(0,0.7)) + ylim(c(0,0.7)) +
      xlab("support") + ylab("poststrat support") +
      coord_fixed()
  )
}
```

Plotting empirical and poststratified means

```
#compare poststratified and empirical means -- nice plot!
disag.point <- bind_rows(mod.disag,mod.disag.ps) %>%
  group_by(model) %>%
  arrange(support, .by_group=TRUE) %>%
  ggplot(aes(x=support,y=forcats::fct_inorder(state),color=model)) +
  geom_point() + theme_classic() +
  theme(legend.position='none') +
  directlabels::geom_dl(aes(label=model),method='smart.grid') +
  ylab('state')
disag.scat <- bind_cols(mod.disag[,1:2],mod.disag.ps[,2]) %>% compare_scat()
plot_grid(disag.point,disag.scat)
```

Plots





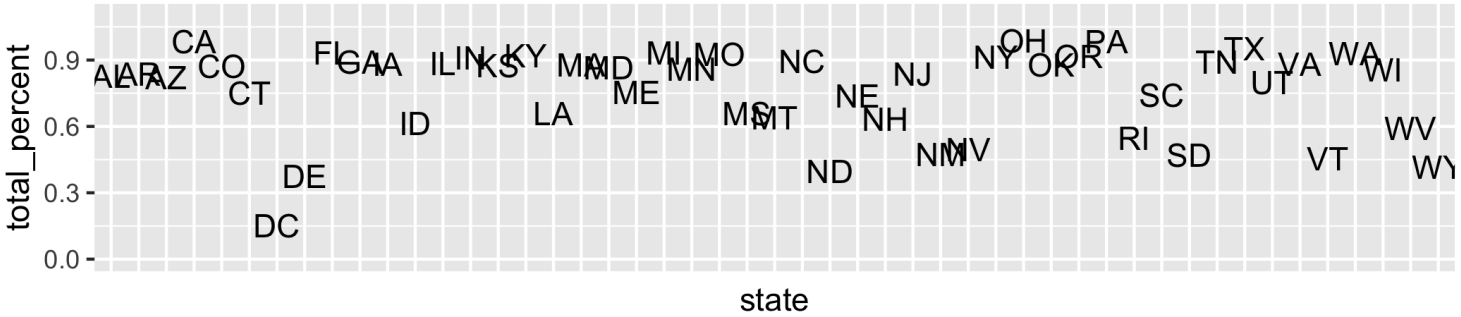
Variation in poststratified estimates is pretty large. Also, the poststratified estimates appear closer to zero -- what is going on?

Demographic representation by state

Let's sum the percentages in the poll data by state to make sure they all sum to 1.

```
grp.means %>%
  group_by(state) %>%
  summarize(total_percent=sum(cpercent.state, na.rm=TRUE)) %>%
  filter(state != "") %>%
  ggplot(aes(x=state,y=total_percent)) +
  geom_text(aes(label=state),hjust=0.5,vjust=0.25) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  coord_fixed(ratio=8) + ylim(c(0,1.1))
```

Demographic representation by state



Ahh, the surveys do not have responses from each demographic group in each state.

Our poststratification is assuming the missing demographic groups have 0% support, which is not good -- even though we have no black men from South Dakota in the polls, there are some in the state (1.7% of the SD population identifies as black or African-American). We underestimate the level of support by assuming no black men in SD support gay marriage.

Multilevel model

One advantage of fitting a multilevel model is that we can borrow information to get better estimates.

In the case of African-American men from South Dakota, we do have responses from black men in nearby states (North Dakota has roughly 3x the African-American population of South Dakota) and other states in the region, which we can use to make a better guess (than 0%!) about the level of support for gay marriage among black men in South Dakota.

Fitting Individual-level Regression Model

Now we fit a regression model for individual survey responses on gay marriage rights given demographics and geography, i.e. each individual's response will be a function of their demographics and state.

Let i index each individual, j index the race-gender combination, k index the age-education combination, s index each state, and r index region.

We denote $y_{ijksr} = 1$ for supporters of same-sex marriage and $y_{ijksr} = 0$ for opponents and those with no opinion.

We model the mean for the state effect as a function of 3 state level variables: the region into which the state falls, the state's conservative (defined as evangelical+LDS) religious percentage, and its Democratic 2004 presidential vote share.

Model

We will not do any model selection here; this model is based on domain/scientific knowledge. We also treat the race-gender and age-education combinations to borrow information across the levels.

We fit the following model.

$$\begin{aligned}\text{logit}\left[\Pr(y_{ijksr}=1)\right] &= \beta_0 + \beta^{relig}\cdot relig_s + \beta^{vote}\cdot vote_s \\ &\quad + \alpha_r^{region} + \alpha_s^{state} + \alpha_j^{race,gender} + \alpha_k^{age,edu};\end{aligned}$$

$$\begin{aligned}\alpha_r^{region} &\sim N(0,\sigma^2_{region}), \quad \text{r} = 1,\ldots,5; \\ \alpha_s^{state} &\sim N(0,\sigma^2_{state}), \quad \text{s} = 1,\ldots,51; \\ \alpha_j^{race,gender} &\sim N(0,\sigma^2_{race,gender}), \quad \text{j} = 1,\ldots,6; \\ \alpha_k^{age,edu} &\sim N(0,\sigma^2_{age,edu}), \quad \text{k} = 1,\ldots,16.\end{aligned}$$

Model

Using a slightly different notation, we can also write the model as

$$\text{logit}\left(\Pr(y_i = 1)\right) = \beta_0 + \alpha^{race,gender}_{j[i]} + \alpha^{age,edu}_{k[i]} + \alpha^{state}_{s[i]}.$$

That is,

$$\begin{aligned}
\alpha^{race,gender}_{j} &\sim N(0, \sigma^2_{race,gender}), \quad \text{j} = 1, \ldots, 6, \\
\alpha^{age,edu}_{k} &\sim N(0, \sigma^2_{age,edu}), \quad \text{k} = 1, \ldots, 16,
\end{aligned}$$

and

$$\begin{aligned}
\alpha^{state}_{s} &\sim N(\alpha^{region}_{r[s]} + \beta^{relig} \cdot relig_s + \beta^{vote} \cdot vote_s, \sigma^2_{state}), \\
\alpha^{region}_{r} &\sim N(0, \sigma^2_{region}),
\end{aligned}$$

where $\text{r} = 1, \ldots, 5$ and $\text{s} = 1, \ldots, 51$.

Model coding

```
#run individual-level opinion model
ml.mod <- glmer(yes.of.all ~ (1|race.female)+(1|age.edu.cat)+
               (1|state)+(1|region)+
               p.relig+kerry.04, data=marriage.data,
             family=binomial(link="logit"))
# just checking scale of these proportions
summary(marriage.data$p.relig)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.839   8.718  12.823  16.287  25.012  68.090
```

```
summary(marriage.data$kerry.04)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0    42.2    48.7    47.7    54.3    89.2
```

Model results

```
summary(ml.mod)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: yes.of.all ~ (1 | race.female) + (1 | age.edu.cat) + (1 | state) +
##     (1 | region) + p.relig + kerry.04
##    Data: marriage.data
##
##      AIC      BIC   logLik deviance df.resid
##   7504.8   7552.1  -3745.4   7490.8     6334
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8404 -0.7100 -0.4845  0.9989  3.8023
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  state       (Intercept) 0.00000  0.0000
##  age.edu.cat (Intercept) 0.39449  0.6281
##  race.female (Intercept) 0.04959  0.2227
##  region      (Intercept) 0.03519  0.1876
## Number of obs: 6341, groups:
## state, 49; age.edu.cat, 16; race.female, 6; region, 5
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.497277   0.437339  -3.424 0.000618
## p.relig     -0.014779   0.004889  -3.023 0.002503
## kerry.04     0.019112   0.006755   2.829 0.004664
##
## Correlation of Fixed Effects:
##          (Intr) p.relg
## p.relig  -0.660
## kerry.04 -0.868  0.661
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

Model results

Note we have no responses from AK or HI.

```
# note nobody from AK or HI in survey
marriage.data %>%
  filter(state=="AK",state=="HI")
```

```
##  [1] state                 p.evang               p.mormon
##  [4] kerry.04              poll                  poll.firm
##  [7] poll.year             id                    statenum
## [10] statename             region.cat            female
## [13] race.wbh              edu.cat               age.cat
## [16] age.cat6              age.edu.cat6          educ
## [19] age                   democrat              republican
## [22] black                 hispanic              weight
## [25] yes.of.opinion.holders yes.of.all           state.initnum
## [28] region                no.of.all             no.of.opinion.holders
## [31] race.female           age.edu.cat           p.relig
## <0 rows> (or 0-length row.names)
```

Predictions

We make predictions in states, broken out by the demographic groups of interest, which will allow us to poststratify down the road.

For now we calculate the predictions, and we'll examine them closely later.

```
ps.ml.mod <- Census %>%
  mutate(support=predict(ml.mod,newdata=.,allow.new.levels=TRUE,type='response')) %>%
  mutate(support=support*cpercent.state) %>%
  group_by(state) %>%
  summarize(support=sum(support))
```

Bayesian model

Now we fit a fully Bayesian model, with same data model as the ML model but with some weakly informative priors on the SD's of varying intercepts that will help with borrowing of information and convergence.

```
bayes.mod <- brm(yes.of.all~(1|race.female)+(1|age.edu.cat)+(1|state)+(1|region)+
                   p.relig+kerry.04, data=marriage.data,
                 family=bernoulli(),
                 prior=c(set_prior("normal(0,0.2)",class='b'), #0.2 is SD (not variance)
                         set_prior("normal(0,0.2)",class='sd',group="race.female"),
                         set_prior("normal(0,0.2)",class='sd',group="age.edu.cat"),
                         set_prior("normal(0,0.2)",class='sd',group="state"),
                         set_prior("normal(0,0.2)",class='sd',group="region")))
```

Bayesian model results

```
summary(bayes.mod)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: yes.of.all ~ (1 | race.female) + (1 | age.edu.cat) + (1 | state) + (1 | region) + p.relig + kerry.04
##    Data: marriage.data (Number of observations: 6341)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~age.edu.cat (Number of levels: 16)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.54      0.08     0.40     0.69 1.00     1288     2022
##
## ~race.female (Number of levels: 6)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.23      0.08     0.12     0.41 1.00     1986     2154
##
## ~region (Number of levels: 5)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.22      0.08     0.10     0.42 1.00     1870     2466
##
## ~state (Number of levels: 49)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.06      0.04     0.00     0.15 1.00     1321     1849
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -1.49      0.47    -2.44    -0.59 1.00     1983     2146
## p.relig      -0.01      0.01    -0.03    -0.00 1.00     2860     3223
## kerry.04      0.02      0.01     0.00     0.03 1.00     3154     2453
##
## 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).
```

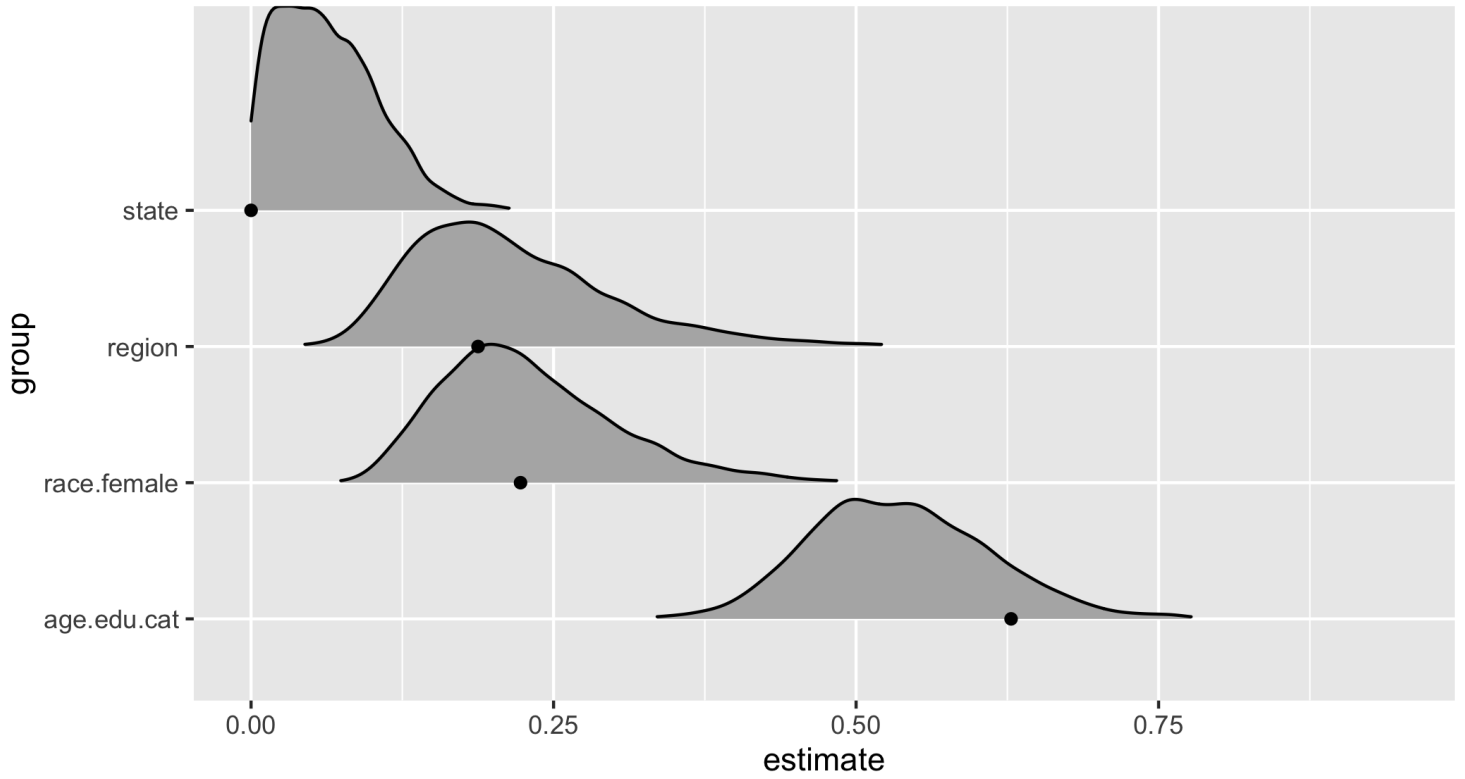
Benefits of Bayesian approach

The most obvious benefit of a Bayesian approach is the total accounting of uncertainty, as we can easily see by plotting the estimated SD's of the group-level intercepts in the frequentist model against the posteriors from the Bayesian model.

```
library(broom.mixed)
ml_sd <- broom::tidy(ml.mod) %>%
  filter(stringr::str_detect(term,"sd_"))

bayes.mod %>%
  gather_draws(`sd.*`,regex=TRUE) %>%
  ungroup() %>%
  mutate(group=stringr::str_replace_all(.variable,c("sd_"="","__Intercept"="")),
         estimate=.value) %>%
  ggplot(aes(y=group,x=estimate)) +
  ggridges::geom_density_ridges(aes(height=..density..),
                                rel_min_height=0.01,stat="density",
                                scale=1.5) +
  geom_point(data=ml_sd)
```

Benefits of Bayesian approach



The dots are the point estimates from the frequentist model, but the Bayesian model gives you an idea of the full posterior distribution of values, from which we can sample.

Poststratifying Bayes

```
#next let's get the point estimate and poststratify from the Bayesian model
ps.bayes.mod <- bayes.mod %>%
  add_predicted_samples(newdata=Census, allow_new_levels=TRUE) %>%
  rename(support = pred) %>%
  mean_qi() %>%
  mutate(support = support * cpercent.state) %>%
  group_by(state) %>%
  summarize(support = sum(support))
```

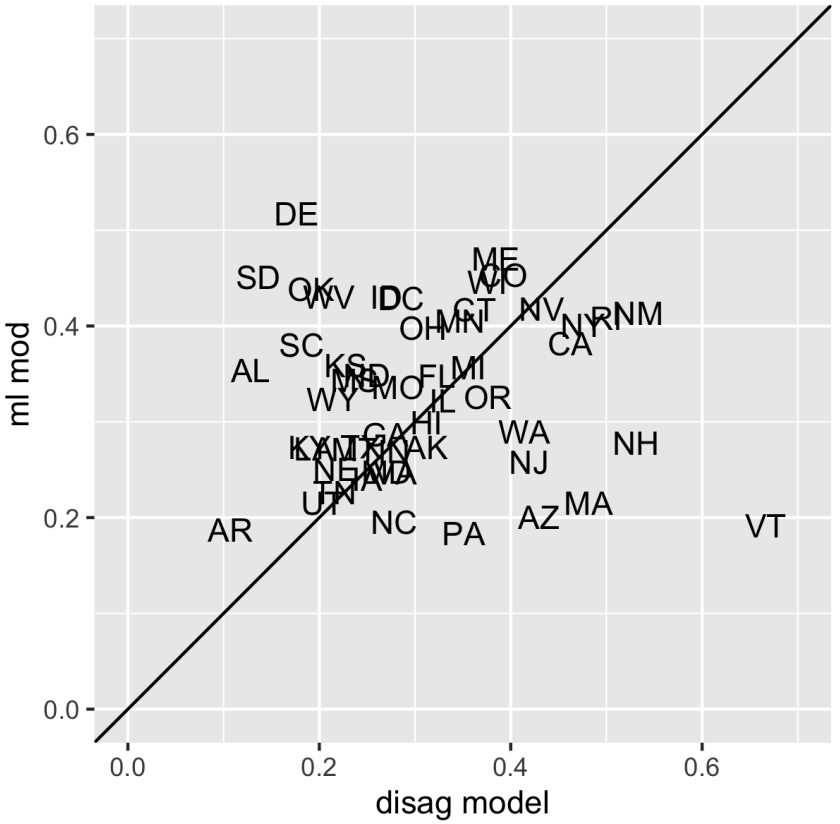
Comparing results

Now we consider comparisons across the 3 approaches.

```
mod.disag[nrow(mod.disag) + 1,] = list("AK", mean(mod.disag$support), "no_ps")
mod.disag[nrow(mod.disag) + 1,] = list("HI", mean(mod.disag$support), "no_ps")
disag.ml <- bind_cols(mod.disag[,1:2], ps.ml.mod[,2]) %>% compare_scat() +
  xlab("disag model") + ylab("ml mod")
disag.bayes <- bind_cols(mod.disag[,1:2], ps.bayes.mod[,2]) %>% compare_scat() +
  xlab("disag model") + ylab("bayes model")
ml.bayes <- bind_cols(ps.ml.mod[,1:2], ps.bayes.mod[,2]) %>% compare_scat() +
  xlab("ml model") + ylab("bayes model")
```

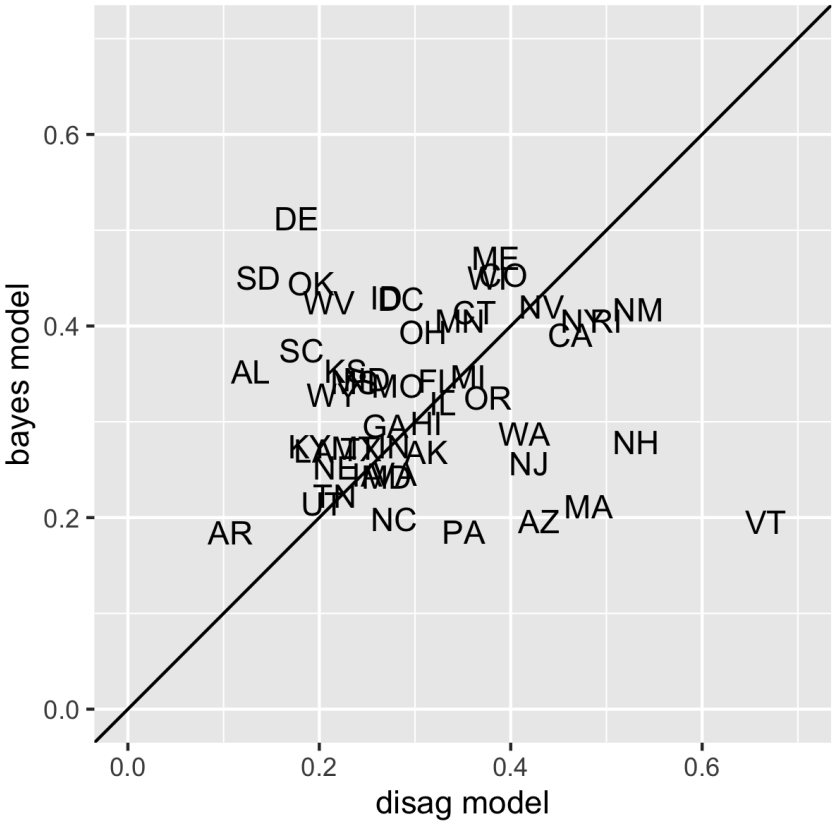
Plots

```
plot_grid(disag.ml)
```



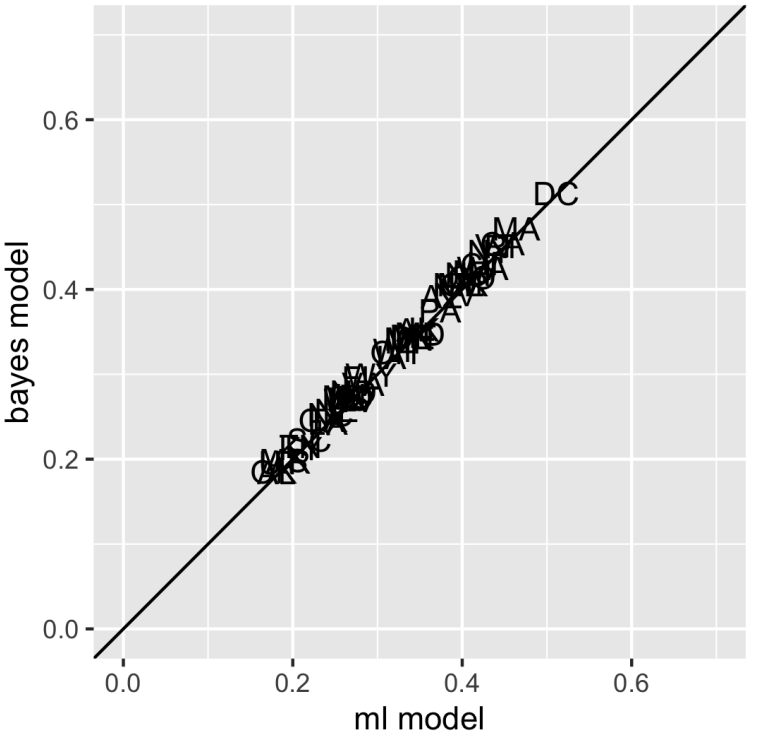
Plots

```
plot_grid(disag.bayes)
```



Plots

```
plot_grid(ml.bayes)
```



Note our predictions from the ML and Bayesian approaches are similar, and the models disagree with the disaggregated model, which does not borrow information.

Prediction

Now we can evaluate predictions, taking advantage of the uncertainty quantification advantages of the Bayesian approach.

We will sample from the posterior to get predicted probabilities for each group of interest based on proportions obtained from the Census data.

```
predict_val <- predict(bayes.mod, newdata=Census, allow_new_levels=TRUE,
            nsamples=500, summary=FALSE)
```

Prediction

```
dim(Census)
```

```
## [1] 4896   15
```

```
head(Census)
```

```
##   state p.evang p.mormon kerry.04 crace.WBH age.cat edu.cat cfemale .freq
## 1    AK   12.44 3.003126     35.5         1       1       1       0   467
## 2    AK   12.44 3.003126     35.5         1       2       1       0   377
## 3    AK   12.44 3.003126     35.5         1       3       1       0   419
## 4    AK   12.44 3.003126     35.5         1       4       1       0   343
## 5    AK   12.44 3.003126     35.5         1       1       2       0   958
## 6    AK   12.44 3.003126     35.5         1       2       2       0  1359
##   cfreq.state cpercent.state region race.female age.edu.cat  p.relig
## 1       21222     0.02200547   west      WhMale   18-29,<HS 15.44313
## 2       21222     0.01776458   west      WhMale   30-44,<HS 15.44313
## 3       21222     0.01974366   west      WhMale   45-64,<HS 15.44313
## 4       21222     0.01616247   west      WhMale     65+,<HS 15.44313
## 5       21222     0.04514183   west      WhMale    18-29,HS 15.44313
## 6       21222     0.06403732   west      WhMale    30-44,HS 15.44313
```

We'll focus on the first four subgroups: white Alaskan men with <HS education in the 4 age groups (18-29, 30-44, 45-64, 65+).

The first 6 sampled support values for those men are in columns 1-4 here....

```
dim(predict_val)
head(predict_val)
```

POSTSTRATIFICATION AGAIN

We could then use these predicted probabilities to estimate public opinion under a variety of assumptions (opinion of all residents, or applying other data on how frequently people in each demographic group vote, to get opinions of likely voters).

These predictions based on data from the Census can be combined with information on how often people in each group vote to predict election outcomes.

What's next?

Move on to the readings for the next module!