STA 610L: Module 4.2b

Poststratification and weighting (Part II)

Dr. Olanrewaju Michael Akande

STA 610L

Multilevel model

As mentioned in the previous module, we will fit a multilevel 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 the questions of interest.

```
#run individual-level opinion model
ml.mod0 <- glmer(yes.of.all ~ race.female + age.edu.cat + p.relig + kerry.04 +
                   (1|state) + (1|region), data=marriage.data,
                 family=binomial(link="logit"),
                 control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e5)))
summary(ml.mod0)
```

STA 610L

Model

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: yes.of.all ~ race.female + age.edu.cat + p.relig + kerry.04 +
##     (1 | state) + (1 | region)
##    Data: marriage.data
## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
##      AIC      BIC   logLik deviance df.resid
##   7461.8   7630.7  -3705.9   7411.8     6316
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9235 -0.7099 -0.4745  0.9889  4.1065
##
## Random effects:
##  Groups Name        Variance  Std.Dev.
##  state  (Intercept) 2.891e-09 5.377e-05
##  region (Intercept) 2.559e-02 1.600e-01
## Number of obs: 6341, groups:  state, 49; region, 5
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.267524   0.447975  -2.829 0.004663
## race.femaleBlMale     0.065898   0.158329   0.416 0.677260
## race.femaleHMale      0.267396   0.161471   1.656 0.097721
## race.femaleWhFem      0.449626   0.061674   7.290 3.09e-13
## race.femaleBlFem     -0.092411   0.135863  -0.680 0.496392
## race.femaleHFem       0.537123   0.164535   3.264 0.001097
## age.edu.cat18-29,HS   0.037365   0.240087   0.156 0.876323
## age.edu.cat18-29,SC   0.405350   0.248459   1.631 0.102794
## age.edu.cat18-29,CG   0.560073   0.252118   2.221 0.026319
## age.edu.cat30-44,<HS -0.643885   0.329296  -1.955 0.050543
## age.edu.cat30-44,HS  -0.536826   0.237163  -2.264 0.023603
## age.edu.cat30-44,SC   0.026699   0.239138   0.112 0.911105
## age.edu.cat30-44,CG   0.139527   0.226531   0.616 0.537942
## age.edu.cat45-64,<HS -1.142234   0.337787  -3.382 0.000721
## age.edu.cat45-64,HS  -0.792911   0.230721  -3.437 0.000589
## age.edu.cat45-64,SC  -0.608716   0.235038  -2.590 0.009602
## age.edu.cat45-64,CG  -0.016679   0.224501  -0.074 0.940778
## age.edu.cat65+,<HS   -1.619029   0.326969  -4.952 7.36e-07
## age.edu.cat65+,HS    -1.532324   0.251011  -6.105 1.03e-09
## age.edu.cat65+,SC    -1.099368   0.268163  -4.100 4.14e-05
## age.edu.cat65+,CG    -0.568061   0.247758  -2.293 0.021859
## p.relig              -0.014821   0.004895  -3.027 0.002466
## kerry.04              0.019578   0.006778   2.889 0.003870
## optimizer (bobyqa) convergence code: 0 (OK)
```

Model

I am a bit concerned about the amount of information available to estimate the race-gender and age-education combinations.

So, let's actually treat the race-gender and age-education combinations as random effects to borrow information across their levels.

We therefore fit the following model.

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

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

STA 610L

Model

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

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

That is,

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

and

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

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

STA 610L

Model coding

```
#run individual-level opinion model
ml.mod <- glmer(yes.of.all ~ p.relig + kerry.04 +
                  (1|race.female) + (1|age.edu.cat) + (1|state) + (1|region),
                data=marriage.data,
               family=binomial(link="logit"),
               control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e5)))
# 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
```

STA 610L

Model results

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

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## yes.of.all ~ p.relig + kerry.04 + (1 | race.female) + (1 | age.edu.cat) +
##     (1 | state) + (1 | region)
##    Data: marriage.data
## Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))
##
##      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) 1.284e-08 0.0001133
##  age.edu.cat (Intercept) 3.945e-01 0.6280828
##  race.female (Intercept) 4.959e-02 0.2226868
##  region      (Intercept) 3.519e-02 0.1875976
## 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) -1.497284   0.436674  -3.429 0.000606
## p.relig     -0.014779   0.004886  -3.025 0.002487
## kerry.04     0.019112   0.006747   2.833 0.004617
##
## Correlation of Fixed Effects:
##          (Intr) p.relg
## p.relig  -0.660
## kerry.04 -0.868  0.660
```

STA 610L

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)
```

STA 610L

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 default priors to help some more 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,
                   control = list(adapt_delta = 0.99,max_treedepth = 12),
                 family=bernoulli())
```

STA 610L

Bayesian model results

```r
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)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 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)     0.70      0.15     0.47     1.05 1.00      936     1884
##
## ~race.female (Number of levels: 6)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.31      0.16     0.13     0.71 1.00     1662     2588
##
## ~region (Number of levels: 5)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.37      0.29     0.11     1.20 1.00     1519     2344
##
## ~state (Number of levels: 49)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.07      0.04     0.00     0.16 1.00     1549     2023
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -1.52      0.53    -2.60    -0.51 1.00     2199     2764
## p.relig      -0.01      0.01    -0.02    -0.00 1.00     4960     3140
## kerry.04      0.02      0.01     0.01     0.03 1.00     4802     3061
##
## Draws were sampled 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

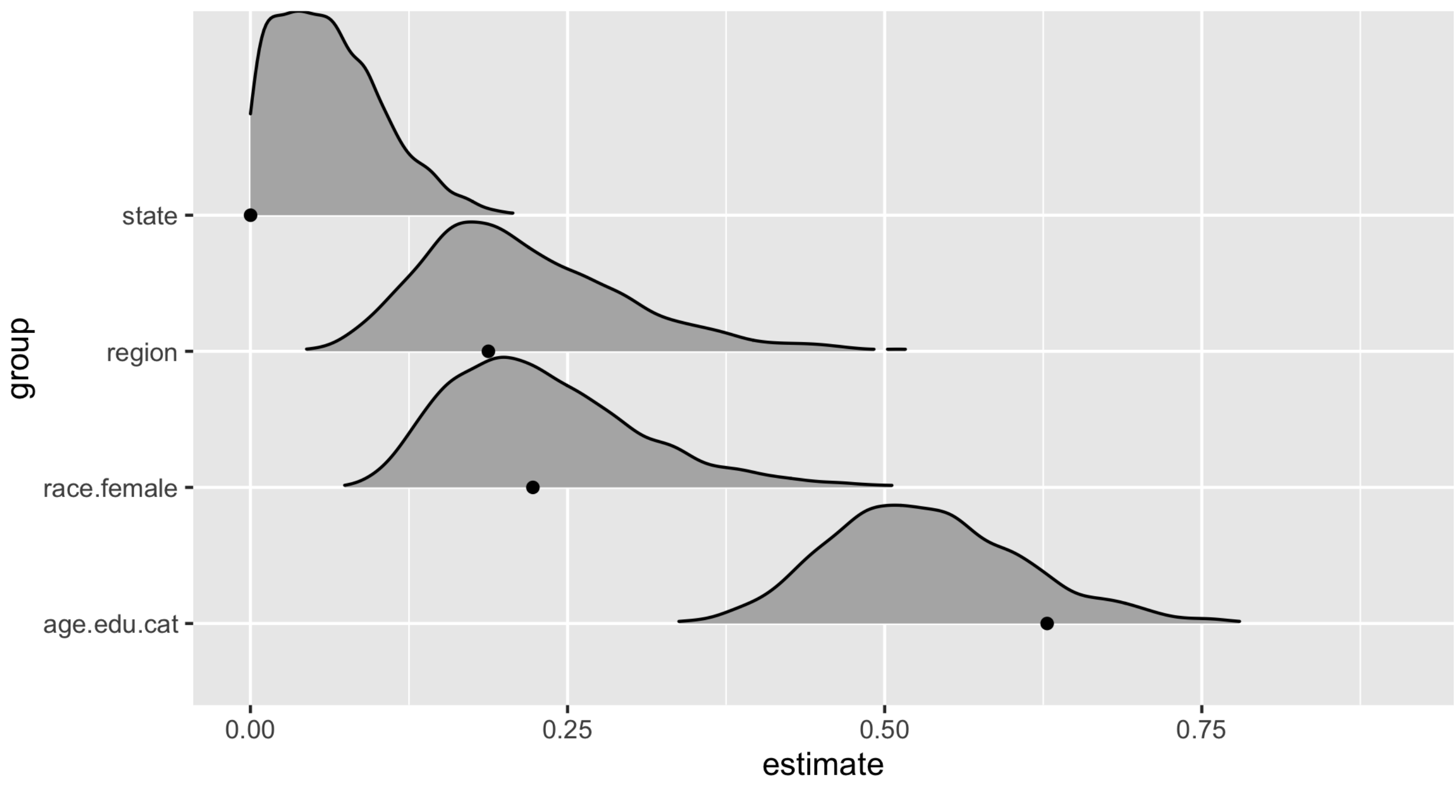
The most obvious benefit of a Bayesian approach is the total accounting of uncertainty, as we can easily see by plotting the estimated SD's of the group-level intercepts in the frequentist model against the posteriors from the Bayesian model.

```r
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)
```

STA 610L

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))
```

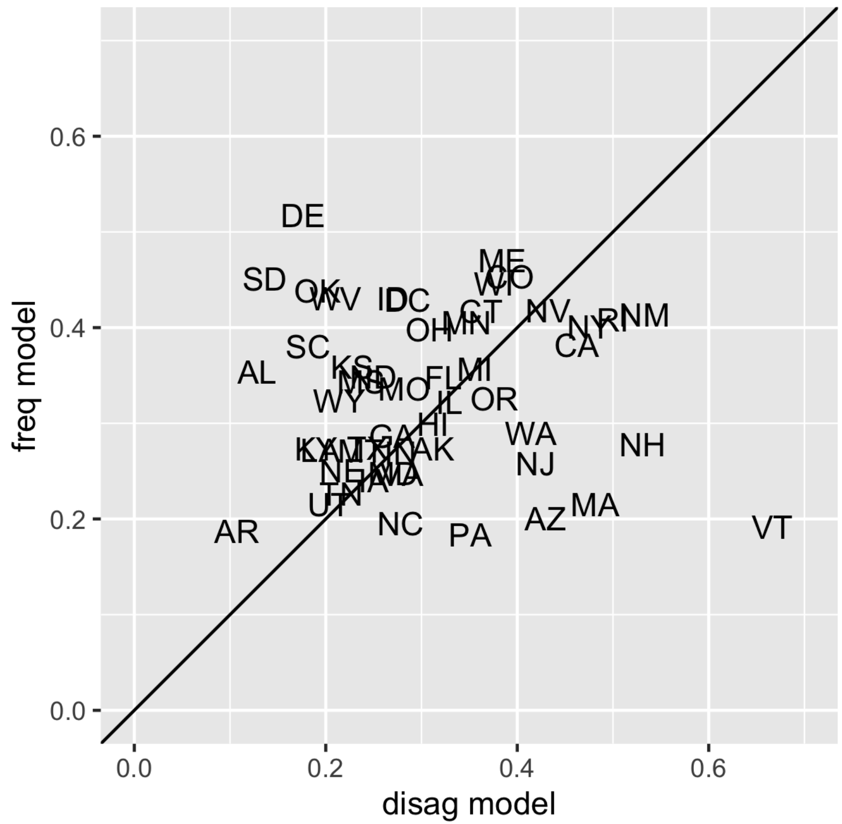
Comparing results

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("freq model")
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("freq model") + ylab("bayes model")
```

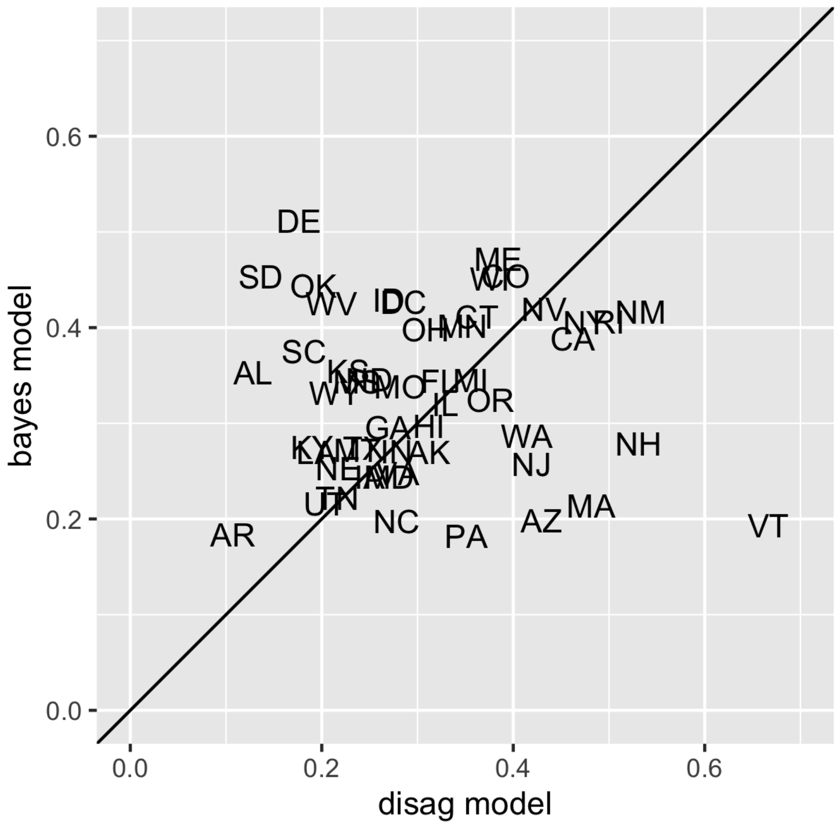
PLOTS

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



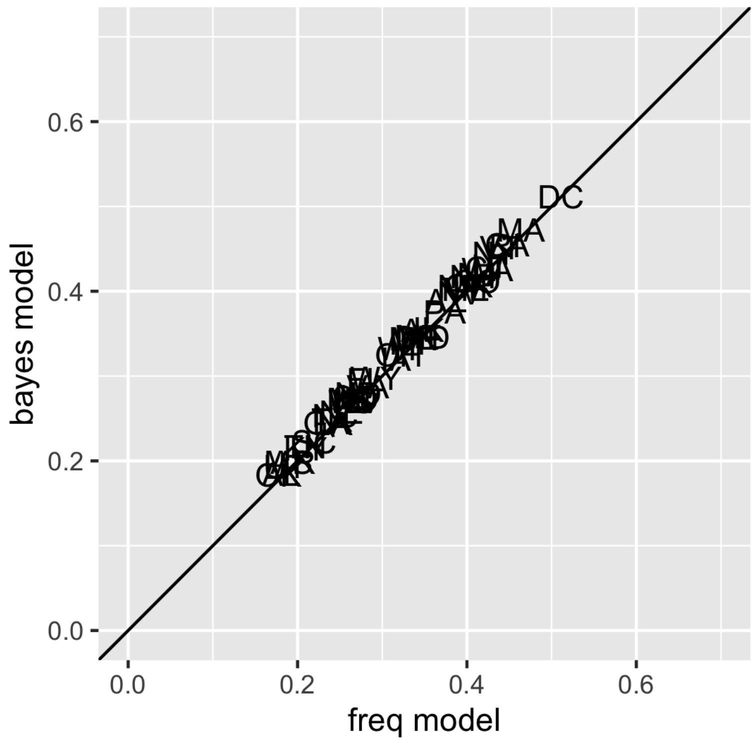
Plots

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



Plots

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



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

STA 610L

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)
```

```
## [1]  500 4896
```

```
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!

STA 610L