STA 610L: Module 4.2a

Poststratification and weighting (Part I)

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.
- 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").

- They 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 (again, 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 are often harder to predict.
- 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!
- However, they 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 would like, which often means doing some correction after the fact to account for this.
- 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, relative 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)?

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

STA 610L

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

```r
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.

These slides 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 over-represented in the survey.

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

Who participated in the polls?

```
##      cpercent.state race.female age.edu.cat
## 3937   2.085007e-02      WhMale   18-29,<HS
## 3938   1.369313e-02      WhMale   30-44,<HS
## 3939   2.088659e-02      WhMale   45-64,<HS
## 3940   3.874242e-02      WhMale     65+,<HS
## 3941   3.132988e-02      WhMale    18-29,HS
## 3942   5.553932e-02      WhMale    30-44,HS
## 3943   5.477251e-02      WhMale    45-64,HS
## 3944   3.249836e-02      WhMale      65+,HS
## 3945   3.520047e-02      WhMale    18-29,SC
## 3946   4.732345e-02      WhMale    30-44,SC
## 3947   3.903454e-02      WhMale    45-64,SC
## 3948   1.299934e-02      WhMale      65+,SC
## 3949   1.044329e-02      WhMale    18-29,CG
## 3950   2.899292e-02      WhMale    30-44,CG
## 3951   3.505441e-02      WhMale    45-64,CG
## 3952   1.055284e-02      WhMale      65+,CG
## 3953   2.556051e-04      BlMale   18-29,<HS
## 3954   1.095450e-04      BlMale   30-44,<HS
## 3955   1.095450e-04      BlMale   45-64,<HS
## 3956   0.000000e+00      BlMale     65+,<HS
## 3957   2.921201e-04      BlMale    18-29,HS
## 3958   1.460600e-04      BlMale    30-44,HS
## 3959   7.303002e-05      BlMale    45-64,HS
## 3960   0.000000e+00      BlMale      65+,HS
## 3961   1.460600e-04      BlMale    18-29,SC
## 3962   6.572701e-04      BlMale    30-44,SC
## 3963   1.460600e-04      BlMale    45-64,SC
## 3964   3.651501e-05      BlMale      65+,SC
## 3965   1.095450e-04      BlMale    18-29,CG
## 3966   1.095450e-04      BlMale    30-44,CG
## 3967   0.000000e+00      BlMale    45-64,CG
## 3968   0.000000e+00      BlMale      65+,CG
## 3969   3.286351e-04       HMale   18-29,<HS
## 3970   2.921201e-04       HMale   30-44,<HS
## 3971   7.303002e-05       HMale   45-64,<HS
## 3972   1.095450e-04       HMale     65+,<HS
## 3973   4.016651e-04       HMale    18-29,HS
## 3974   2.921201e-04       HMale    30-44,HS
## 3975   2.190900e-04       HMale    45-64,HS
## 3976   3.651501e-05       HMale      65+,HS
## 3977   3.651501e-04       HMale    18-29,SC
## 3978   2.921201e-04       HMale    30-44,SC
## 3979   2.190900e-04       HMale    45-64,SC
## 3980   0.000000e+00       HMale      65+,SC
## 3981   0.000000e+00       HMale    18-29,CG
## 3982   2.190900e-04       HMale    30-44,CG
## 3983   1.095450e-04       HMale    45-64,CG
## 3984   1.095450e-04       HMale      65+,CG
## 3985   1.566494e-02       WhFem   18-29,<HS
## 3986   1.018769e-02       WhFem   30-44,<HS
## 3987   1.515373e-02       WhFem   45-64,<HS
## 3988   4.078726e-02       WhFem     65+,<HS
## 3989   2.143431e-02       WhFem    18-29,HS
## 3990   4.297816e-02       WhFem    30-44,HS
## 3991   5.557584e-02       WhFem    45-64,HS
## 3992   4.575330e-02       WhFem      65+,HS
```

STA 610L

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.

```r
# 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.

```r
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.

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

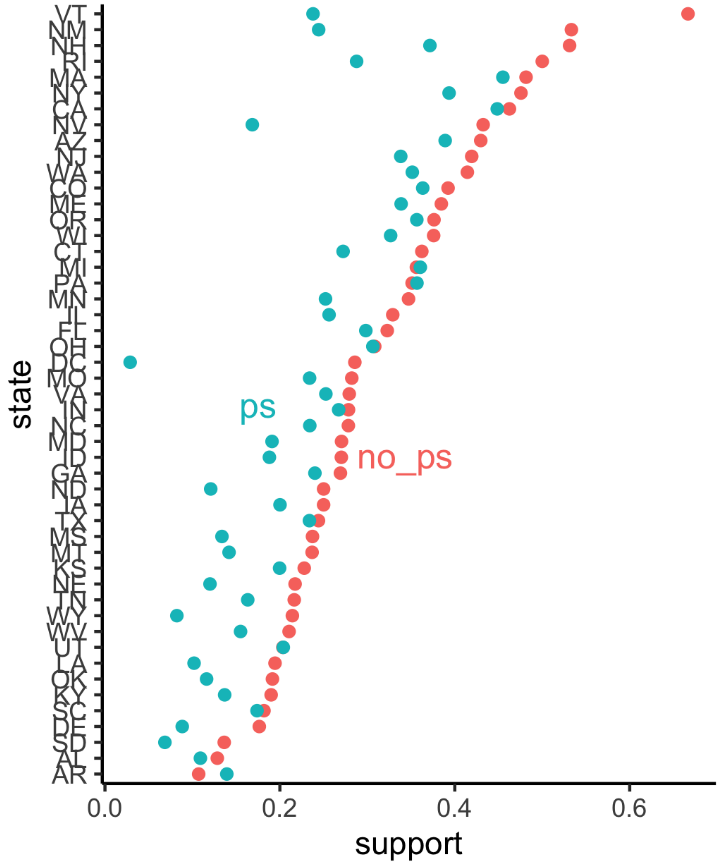
Define function for plotting (may want to use again)

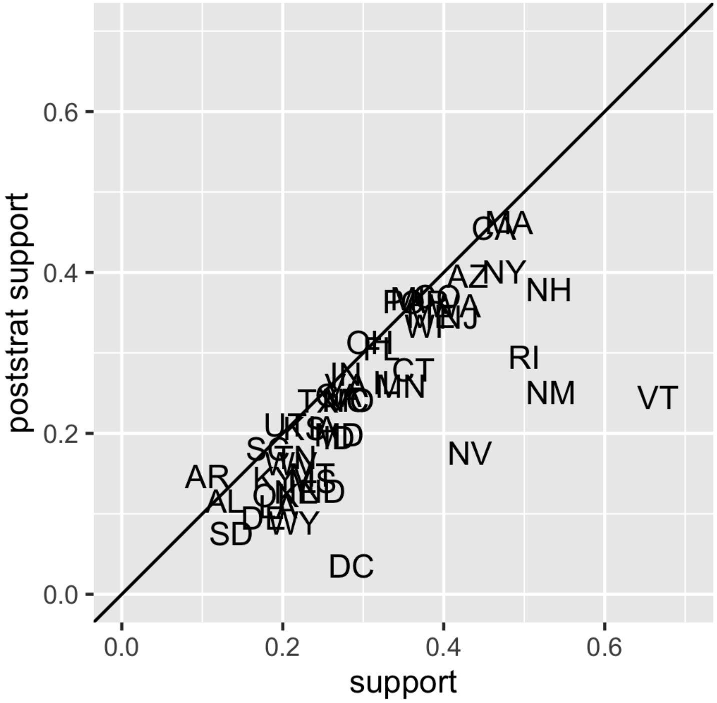
```r
#make a function so we don't have to type over and over
compare_scat <- function(d){
  return(
    ggplot(data=d, aes(x=support...2 ,y=support...3))+
      geom_text(aes(label=state),hjust=0.5,vjust=0.25) +
      geom_abline(slope=1,intercept=0) +
      xlim(c(0,0.7)) + ylim(c(0,0.7)) +
      xlab("support")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

Plotting empirical and poststratified means

```r
#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





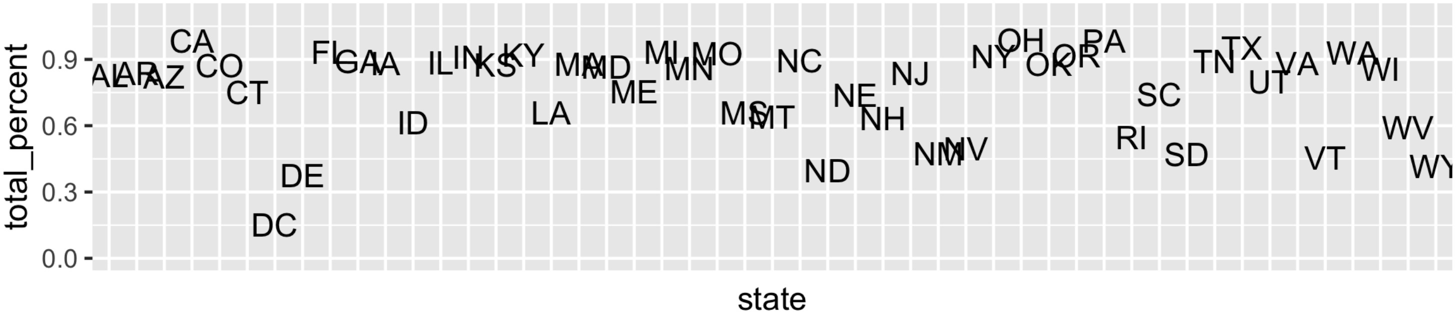
Variation in poststratified estimates is pretty large. Also, the poststratified estimates appear closer to zero -- what is going on?

Demographic representation by state

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

```r
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 therefore 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.

We will try to do this in the next module.

What's next?

Move on to the readings for the next module!