Problem Set 4, 10/11/2017

Group 1	Dan	dbuonaiuto@g.harvard.edu
	Nick	nherrmann@g.harvard.edu
	Zane	rzwolf@g.harvard.edu
	Lydia	lakrasilnikova@g.harvard.edu

Section 10.10 Problem 1

Part a and b: fitting models and evaluating and comparing models

We first try fitting a model with all the variables and no interactions:

```
stan_glm(formula = presvote_2party ~ income + gender + race + educ1 + partyid3_b + ideo,
       family = binomial(link = "logit"), data = d)
```

	Median	MAD_SD
(Intercept)	-4.0	0.5
income	-0.1	0.1
gender2. female	0.5	0.2
race2. black	-1.9	0.5
race3. asian	0.0	0.8
race4. native american	0.5	0.6
race5. hispanic	0.7	0.4
educ1	0.2	0.1
partyid3_b2. indpendents and apolitical (1966 only	1.8	0.3
partyid3_b3. republicans (including leaners)	4.1	0.2
ideo3. moderate ('middle of the road')	0.9	0.4
ideo5. conservative	1.8	0.2

We have a lot of predictors and not all of them helpful, so let us try paring them down. `income` has an extremely small coefficient compared to its standard deviation. `partyid3_b`, in contrast, has large coefficients compared to its standard deviation. `gender` and `educ1` have small coefficients compared to their standard deviations, `race` has a large coefficient compared to its standard deviation only for `black` voters, and `ideo` also has a large coefficient compared to its standard deviation for `conservative` voters.

Since they have extremely small coefficients compared to their standard deviations, `income` and `educ1` are good candidates for removal from the model:

```
stan_glm(formula = presvote_2party ~ gender + race + partyid3_b + ideo,
       family = binomial(link = "logit"), data = d)
```

	Median	MAD_SD
(Intercept)	-3.3	0.3
gender2. female	0.3	0.2

```
race2. black                                           -2.0    0.4
race3. asian                                            0.4    0.7
race4. native american                                  0.3    0.6
race5. hispanic                                         0.4    0.4
partyid3_b2. indpendents and apolitical (1966 only      1.7    0.3
partyid3_b3. republicans (including leaners)            4.0    0.2
ideo3. moderate ('middle of the road')                  0.6    0.4
ideo5. conservative                                     1.6    0.2
```

In this new model, we see that `gender` has a very low coefficient compared to its standard deviation; therefore we also remove `gender`. We are left with this model including only important variables:

```
stan_glm(formula = presvote_2party ~ race + partyid3_b + ideo,
       family = binomial(link = "logit"), data = d)

                                                       Median MAD_SD
(Intercept)                                            -3.1    0.2
race2. black                                           -1.9    0.4
race3. asian                                            0.3    0.8
race4. native american                                  0.3    0.6
race5. hispanic                                         0.4    0.4
partyid3_b2. indpendents and apolitical (1966 only      1.6    0.3
partyid3_b3. republicans (including leaners)            3.9    0.2
ideo3. moderate ('middle of the road')                  0.6    0.4
ideo5. conservative                                     1.6    0.2
```

Since it has an especially large coefficient compared to its standard deviations, `partyid3_b` is a good candidate for including interactions. We try interacting `partyid3_b` with `race`:

```
stan_glm(formula = presvote_2party ~ race + partyid3_b + ideo + partyid3_b:race,
       family = binomial(link = "logit"), data = d)

                                                                        Median MAD_SD
(Intercept)                                                             -3.1    0.2
race2. black                                                            -2.0    0.6
race3. asian                                                            -0.4    1.2
race4. native american                                                   0.4    0.7
race5. hispanic                                                          0.3    0.5
partyid3_b2. indpendents and apolitical (1966 only                       1.5    0.3
partyid3_b3. republicans (including leaners)                             3.9    0.2
ideo3. moderate ('middle of the road')                                   0.6    0.4
ideo5. conservative                                                      1.6    0.2
race2. black:partyid3_b2. indpendents and apolitical (1966 only          0.5    1.0
race3. asian:partyid3_b2. indpendents and apolitical (1966 only          1.4    1.6
race4. native american:partyid3_b2. indpendents and apolitical (1966 only -2.0    1.8
race5. hispanic:partyid3_b2. indpendents and apolitical (1966 only       1.3    1.3
race2. black:partyid3_b3. republicans (including leaners)               -0.3    0.9
race3. asian:partyid3_b3. republicans (including leaners)                0.9    1.5
race4. native american:partyid3_b3. republicans (including leaners)      1.6    1.8
race5. hispanic:partyid3_b3. republicans (including leaners)            -0.1    1.1
```

We also try interacting `partyid3_b` with `ideo`:

```
stan_glm(formula = presvote_2party ~ race + partyid3_b + ideo + partyid3_b:ideo,
       family = binomial(link = "logit"), data = d)
```

	Median	MAD_SD
(Intercept)	-3.3	0.3
race2. black	-2.0	0.4
race3. asian	0.3	0.7
race4. native american	0.3	0.6
race5. hispanic	0.4	0.4
partyid3_b2. indpendents and apolitical (1966 only	1.7	0.6
partyid3_b3. republicans (including leaners)	4.5	0.4
ideo3. moderate ('middle of the road')	0.9	0.6
ideo5. conservative	1.9	0.3
partyid3_b2. indpendents and apolitical (1966 only:ideo3. moderate ('middle of the road')	0.4	0.9
partyid3_b3. republicans (including leaners):ideo3. moderate ('middle of the road')	-1.0	0.7
partyid3_b2. indpendents and apolitical (1966 only:ideo5. conservative	-0.4	0.7
partyid3_b3. republicans (including leaners):ideo5. conservative	-0.7	0.5

In both cases, the coefficients of the interactions are overwhelmed by the standard deviations. Therefore, the simpler model without interactions (`presvote_2party ~ race + partyid3_b + ideo`) seems to be best.

Part c: importance of input variables

The paragraph below describes the importance of our predictors for the model:

```
stan_glm(formula = presvote_2party ~ race + partyid3_b + ideo,
       family = binomial(link = "logit"), data = d)
```

	Median	MAD_SD
(Intercept)	-3.1	0.2
race2. black	-1.9	0.4
race3. asian	0.3	0.8
race4. native american	0.3	0.6
race5. hispanic	0.4	0.4
partyid3_b2. indpendents and apolitical (1966 only	1.6	0.3
partyid3_b3. republicans (including leaners)	3.9	0.2
ideo3. moderate ('middle of the road')	0.6	0.4
ideo5. conservative	1.6	0.2

First we start with the intercept. Here we see that if you were a white, democratic, and liberal person in 1992, you were 86% less likely to vote for Bush than your compatriots (or put another way, only a 14% probability of voting for Bush). If we used the "Divide by 4" rule (as we do from here on out), we achieve a relatively close approximation of 78%.

Of the five racial categories, only black is estimated within a certain degree of confidence. This factor tells us that, regardless of party id or ideological leanings, a black voter is 47% less likely to vote for a republican than a white voter. Other ethnicity switches do not drastically change the probability, only 7-10% with large error margins.

A white member of a the republican party is 98% more likely to vote republican than if they are a democrat, while a white independent is only 40% more likely to vote for Bush than a white democrat. We can see that party affiliation is a strong predictor of voting. Conservatives are also 40% more likely to vote republican than liberals. Moderates are only 15% more likely to vote republican than liberals.

Overall, party affiliation is the most important predictor of republican voting.

R Code

```
rm(list=ls())
library(arm)
library(rstanarm)
library(foreign)

# section 10.10 Problem 1

# import and prep data for regression
# data is in the NES folder at http://www.stat.columbia.edu/~gelman/arm/
d = read.dta(file="nes5200_processed_voters_realideo.dta")
colnames(d)

d = subset(d,year=="1992") #removes all years except 1992

# curate variables
d$income = as.numeric(factor(d$income), levels = unique(d$income)) #income as numeric rank
(1-5)
d$gender = as.factor(d$gender) #gender as factor with 2 levels
d$race = as.factor(d$race) #race as factor with 5 levels
d$educ1 = as.numeric(factor(d$educ1), levels = unique(d$educ1)) #education as numeric rank
(1-4)
d$partyid3_b = as.factor(d$partyid3_b) #party ID as factor with 3 levels
d$ideo = as.factor(d$ideo) #political ideaology as factor with 3 levels

# part a - logistic regression predicting support for Bush in 1992 given
# income, sex, ethnicity, education, party ID, and political ideology

# include all variables
fit1=stan_glm(presvote_2party~income + gender + race + educ1 + partyid3_b + ideo,
              family=binomial(link="logit"),data=d)
print(fit1)

# remove variables with low coefficients compared to standard deviations
fit2=stan_glm(presvote_2party~gender + race + partyid3_b + ideo,
              family=binomial(link="logit"),data=d)
print(fit2)

# further remove variables with low coefficients compared to standard deviations,
# leaving only important variables
fit3=stan_glm(presvote_2party~race + partyid3_b + ideo,
              family=binomial(link="logit"),data=d)
print(fit3)

# try interactions with partyid3_b
fit4=stan_glm(presvote_2party~race + partyid3_b + ideo + partyid3_b:race,
              family=binomial(link="logit"),data=d)
print(fit4)

fit5=stan_glm(presvote_2party~race + partyid3_b + ideo + partyid3_b:ideo,
              family=binomial(link="logit"),data=d)
print(fit5)
```

```
###Dan’s failed attempt to zscore:
d2<-dplyr::select(d,presvote_2party,income,gender,race,educ1,partyid3,ideo)
d2$presvote_2party<-as.numeric(d2$presvote_2party)-2
d2$income<-as.numeric(d2$income)-2
d2$gender<-as.numeric(d2$gender)-2
d2$race<-as.numeric(d2$race)-1
d2$educ1<-as.numeric(d2$educ1)-1
d2$partyid3<-as.numeric(d2$partyid3)-2
d2$ideo<-as.numeric(d2$ideo)-2

d2$z_income<-(d2$income-mean(d2$income))/(2*sd(d2$income))
d2$z_gender<-(d$gender-mean(d$gender))/(2*sd(d$gender))
d2$z_race<-  (d$race-mean(d$race))/(2*sd(d$race))
d2$z_educ1<-(d2$educ1-mean(d2$educ1))/2*sd(d2$educ1)
d2$z_partyid3<-(d2$partyid3-mean(d2$partyid3))/(2*sd(d2$partyid3))
d2$z_ideo<-(d2$ideo-mean(d2$ideo))/(2*sd(d2$ideo))
```