VTM Survey Final Paper

Averie St.Germaine

2024-05-18

## Introduction

In the fall of 2019 the MassInc Polling Group and UMB’s Sustainable Solutions Lab distributed a survey to gather Massachusetts residents’ views on climate change (Estrada-Martínez et al. 2020). Along with the collection of personal demographic information, such as race, age, gender, familial income, education level, and political leaning, participants were asked five questions regarding their level of preparedness in the event of extreme weather due to climate change. Multiple analyses have been done with this survey data through the lens of how racial disparities may impact individual emergency preparedness, but had failed to incorporate other demographic variables. In my analysis of the survey data I aimed to determine which variables had the greatest impact on an individual's preparedness for extreme weather and if individuals that have a higher income are more likely to be better prepared for these events.

## Methods

### Survey Design

The 2019 Climate Change Poll conducted by MassInc Polling Group was done in effort to gather Massachusetts residents opinions on climate change. The survey was distributed through phone call and online in English and Spanish and consisted of 35 questions, 14 of which were in regards to participant demographics. MassInc Polling Group ensured oversampling of underrepresented minority groups (Black, Hispanic, and Asian peoples) in an effort to standardize the sampling pool. The Umass Boston Sustainable Solutions Lab created this survey with the intention of analyzing how different racial groups in the Boston area view climate change (Estrada-Martínez et al. 2020). The 2020 Views that Matter report highlights that an overwhelming amount of Massachusetts residents (86%) believe that climate change is happening and that an individuals preparedness for extreme weather events was similar across all races. (Estrada-Martínez et al. 2020).

### Data Cleaning

This survey recorded extensive demographic and opinion data from over 2,000 Massachusetts residents. To hone in on the variables that impact an individuals preparedness, it was necessary to remove all partially incomplete or unknown survey responses. I also wanted to focus on the 4 most prevalent racial groups, White, Black, Hispanic, and Asian, and removed all others from my analysis due to low sampling quantity. The education demographic category contained nine different response types which I decided to simplify into five categories for computational purposes. The five new education categories consisted of “no high school diploma,” “high school diploma,” “some college,” “bachelors degree,” and “graduate school.”

library(tidybayes)  
library(tidyverse)

library(rethinking)

library(ggplot2)  
library(haven)  
library(ordinal)

library(likert)

library(performance)

library(brms)

require(foreign)

require(ggplot2)  
require(MASS)

require(Hmisc)

require(reshape2)

library(patchwork)

library(GGally)

data <- read\_sav("~/Grad/4 Semester/Climate Change Survey Full Sample 01312020.sav")  
  
View(data)  
  
sub\_data <- data |>  
 dplyr::select(ID, RACE, CAGE, PARTYID, EDUC, INCOME, GEN, EVAC, EFDM, ETCH, EVAL, EINS) |>  
 filter(!RACE == 99, !RACE == 5, !RACE == 6, !RACE == 7, !RACE == 8, !CAGE == 99, !PARTYID == 99, !EDUC == 99, !INCOME == 99, !GEN == 3, !GEN == 4, !GEN == 99, !EVAC == 99, !EFDM == 99, !ETCH == 99, !EVAL == 99, !EINS == 99) |>  
 drop\_na() |>  
 mutate(EDUC = as.character(EDUC)  
 ,EDUC = dplyr::recode(EDUC,  
 "1" = 1,  
 "2" = 2,  
 "3" = 3,  
 "4" = 3,  
 "5" = 4,  
 "6" = 4,  
 "7" = 5,  
 "8" = 5,  
 "9" = 5))

To standardize the data across all races, I took a random sample of 168 individuals (the lowest count) from each race category, narrowing my data down to 672 individuals.

set.seed(47)  
  
white <- sub\_data |>  
 filter(RACE == 1) |>  
 sample\_n(168)  
  
hispanic <- sub\_data |>  
 filter(RACE == 3) |>  
 sample\_n(168)  
  
black <- sub\_data |>  
 filter(RACE == 2) |>  
 sample\_n(168)  
  
asian <- sub\_data |>  
 filter(RACE == 4) |>  
 sample\_n(168)   
  
race\_sub <- rbind(white, hispanic, black, asian)

### What factors influence individual preparedness?

Participants were asked five yes or no questions on whether they had an evacuation plan, family communication plan, home insurance, food and medicine stockpile, and protection plan for valuables. These five factors were used to create a Likert-type response index of preparedness, ranging from 1 (very unprepared) to 6 (very prepared).

## Preparedness index (Q 25) (1 = yes, 2 = no, 99 = idk)  
## 6 = very prepared   
## 1 = very unprepared   
  
## make preparedness index  
prep <- race\_sub |>  
 mutate(PREP = (EVAC + EFDM + ETCH + EVAL + EINS)) |>  
 mutate(PREP = as.factor(PREP)  
 ,PREP = dplyr::recode(PREP,  
 "10" = 1,  
 "5" = 6,  
 "6" = 5,  
 "7" = 4,  
 "8" = 3,  
 "9" = 2))

Due to the categorical Likert-type response of preparedness and demographic data, an ordinal regression was the most appropriate analysis for this data. I created 5 models using the polr() function for ordinal regression and compared AIC scores to determine which demographic variables significantly impacted an individuals preparedness.

prep <- as.factor(prep)

mod1 <- polr(PREP ~ RACE + INCOME + CAGE, data = prep, Hess=TRUE)

mod2 <- polr(PREP ~ INCOME + CAGE , data = prep, Hess=TRUE)

mod3 <- polr(PREP ~ RACE + INCOME + CAGE + PARTYID + EDUC + GEN, data = prep, Hess=TRUE)

mod4 <- polr(PREP ~ RACE + INCOME + GEN, data = prep, Hess=TRUE)

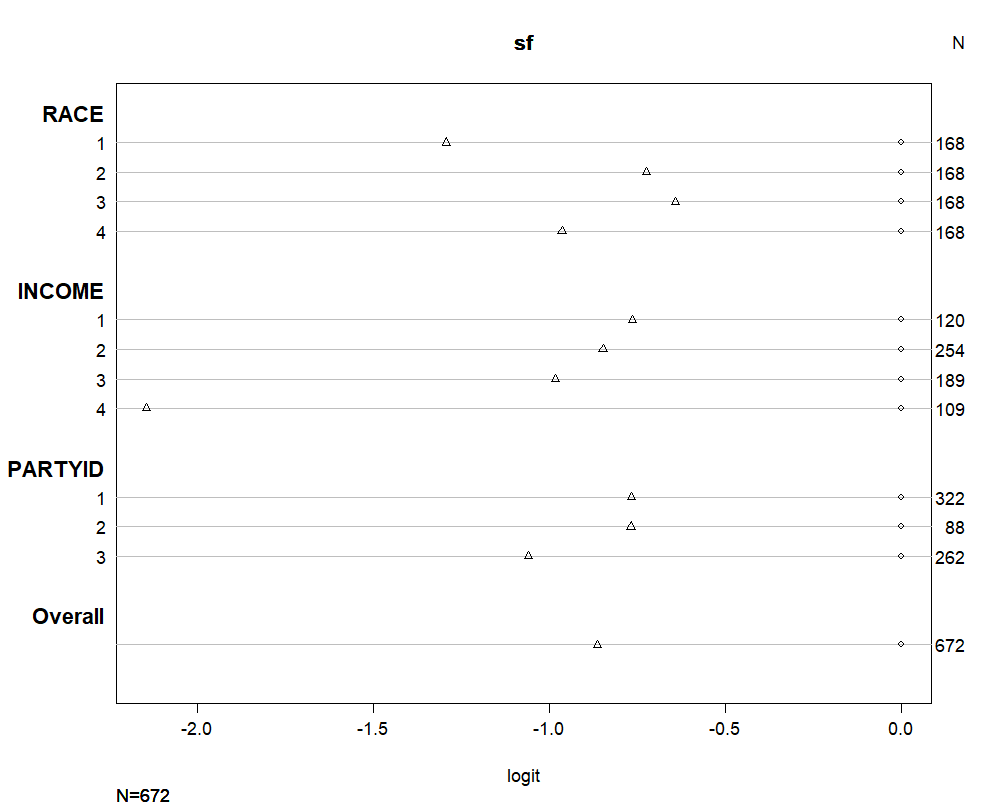
mod5 <- polr(PREP ~ RACE + INCOME + PARTYID, data = prep, Hess=TRUE)

mod6 <- polr(PREP ~ INCOME + CAGE + PARTYID, data = prep, Hess=TRUE)

## mod 1 AIC = 2297.36  
## mod 2 AIC = 2300.807  
## mod 3 AIC = 2296.95  
## mod 4 AIC = 2299.207  
## mod 5 AIC = 2292.68  
## mod 6 AIC = 2300.172

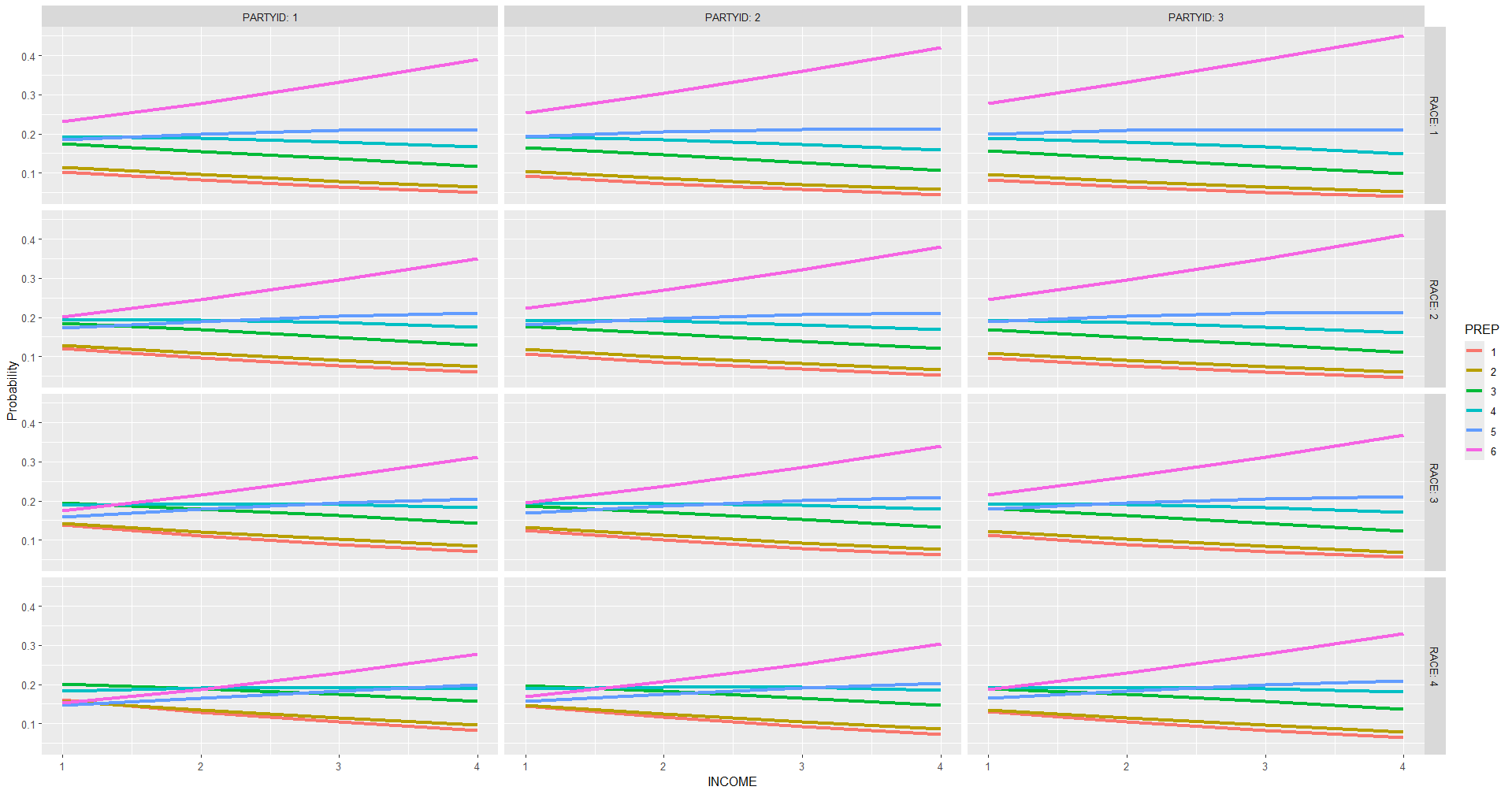
I found the odds ratio of model 5, the race, income, and political leaning model, as it had the lowest AIC score and only contained 3 dependent variables, simplifying the model. I then found the proportional odds assumption for model five by creating a quantile function to give the predicted regressed values of preparedness on the predictor variables of race, income, and political leaning one at a time. Creating a common reference point by normalizing the first set of coefficients allows for graphical comparison of the predictor variables.

exp(coef(mod5))  
(ci <- confint(mod5))  
exp(cbind(OR = coef(mod5), ci))  
  
## proportional odds ratio  
sf <- function(y) {  
 c('Y>=1' = qlogis(mean(y >= 1)),  
 'Y>=2' = qlogis(mean(y >= 2)),  
 'Y>=3' = qlogis(mean(y >= 3)),  
 'Y>=4' = qlogis(mean(y >= 4)),  
 'Y>=5' = qlogis(mean(y >= 5)),  
 'Y>=6' = qlogis(mean(y >= 6)))  
}  
  
(s <- with(prep, summary(as.factor(PREP) ~ RACE + INCOME + PARTYID fun=sf)))  
  
## normalize first set of coefs to be 0, a common reference point  
s[, 6] <- s[, 6] - s[, 3]  
s[, 5] <- s[, 5] - s[, 3]  
s[, 4] <- s[, 4] - s[, 3]  
s[, 3] <- s[, 3] - s[, 3]  
  
s  
  
plot(s, which=1:3, xlab='logit', xlim=range(s[,3:4]))



Comparing the standard logistic distribution of preparedness, one can see that race 1 (white) and income 4 (75k-100k) are less clustered than other variables within each predictor, race and income respectively. This indicates that there may not be proportional odds of preparedness for white or high income individuals.

## new data frame   
newdat <- data.frame(  
 RACE = rep(1:4, 300),  
 INCOME = rep(1:4, each = 300),  
 PARTYID = rep(1:3, 400))  
  
  
newdat <- cbind(newdat, predict(mod5, newdat, type = "probs"))

lnewdat <- melt(newdat, id.vars = c("RACE", "INCOME", "PARTYID"),  
 variable.name = "PREP", value.name="Probability")  
  
## probability graphs   
race\_prob <- ggplot(lnewdat, aes(x = INCOME,   
 y = Probability,   
 color = PREP)) +  
 geom\_line(linewidth = 1.5) +  
 facet\_grid(RACE~PARTYID, labeller="label\_both")  
  
race\_prob

Creating a new data frame of 1,200 total predicted data points, 400 each for race, income, and political party, allows for easy visual comparison of the simulated data. Unfortunately, to run the predicted variables with this method it was necessary to run the model as numeric, although the data is categorical and should be factorial. Placing income on the x axis of the predicted probability graphs, as it is the only predictor that increases with categorical assignment, allowed for interpretation of the predicted probabilities. It was found that the probability of being very prepared (6) increased with increasing income across all race and political parties. Black people were less likely to be very prepared across all political parties, followed by Hispanic people, and then Asian people. Democrats were marginally less likely to be very prepared than republicans, with independent individuals having the highest probability of being very prepared across all incomes and races.

### Method 2.0

Due to the Likert-type preparedness response of this survey data and the plethora of predictor variables that contained varying numbers of categories, finding the proper predicted probabilities was complicated. I wanted to try out many methods of ordinal regression on this data to ensure I did my analysis properly. Using my pre-cleaned data and the mutate function, I created density, p(prep), and cumulative distribution function, Phi(prep) values for an individuals preparedness. I then created plots for the normal and cumulative normal densities of preparedness.

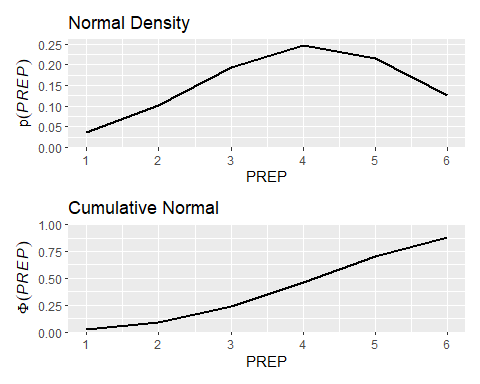
summary(prep)

## ID RACE CAGE PARTYID   
## Length:672 Min. :1.00 Min. :1.00 Min. :1.000   
## Class :character 1st Qu.:1.75 1st Qu.:1.00 1st Qu.:1.000   
## Mode :character Median :2.50 Median :2.00 Median :2.000   
## Mean :2.50 Mean :2.31 Mean :1.911   
## 3rd Qu.:3.25 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :4.00 Max. :4.00 Max. :3.000   
## EDUC INCOME GEN EVAC EFDM   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:1.000   
## Median :3.000 Median :2.000 Median :2.000 Median :1.00 Median :1.000   
## Mean :3.408 Mean :2.427 Mean :1.555 Mean :1.47 Mean :1.412   
## 3rd Qu.:4.000 3rd Qu.:3.000 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:2.000   
## Max. :5.000 Max. :4.000 Max. :2.000 Max. :2.00 Max. :2.000   
## ETCH EVAL EINS PREP   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:3.00   
## Median :1.000 Median :1.000 Median :1.000 Median :4.00   
## Mean :1.363 Mean :1.317 Mean :1.298 Mean :4.14   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:6.00   
## Max. :2.000 Max. :2.000 Max. :2.000 Max. :6.00

sd(prep$PREP)

## [1] 1.62362

d <- prep |>  
 mutate(`p(prep)` = dnorm(PREP, mean = 4.14, sd = 1.6),  
 `Phi(prep)` = pnorm(PREP, mean = 4.14, sd = 1.6))  
  
## plot   
p1 <-  
 d |>  
 ggplot(aes(x = PREP, y = `p(prep)`)) +  
 geom\_area(aes(fill = PREP <= 1),  
 show.legend = F) +  
 geom\_line(linewidth = 1) +  
 scale\_fill\_manual(values = c("transparent")) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.05))) +   
 labs(title = "Normal Density",  
 y = expression(p(italic(PREP))))  
  
p2 <-  
 d |>  
 ggplot(aes(x = PREP, y = `Phi(prep)`)) +  
 geom\_area(aes(fill = PREP <= 1),  
 show.legend = F) +  
 geom\_line(linewidth = 1) +  
 scale\_fill\_manual(values = c("transparent")) +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0)), limits = 0:1) +   
 labs(title = "Cumulative Normal",  
 y = expression(Phi(italic(PREP))))  
  
# combine and adjust with patchwork  
  
p1 / p2 &   
 scale\_x\_continuous(breaks = 1:6) &   
 coord\_cartesian(xlim = c(1, 6))



The density plot shows an almost Gaussain distribution of preparedness around 4. The cumulative plot lets us know that there is the greatest change in probability of preparedness from levels 3 to 4 and 4 to 5.

### Creating a brms model with priors

I used get\_priors to create a series of new models using the brms() function that included priors for intercept and the predictor variables from my data. All priors were found to be non-informative, rendering them unnecessary to include in the model. Using the mutate function, I created three new standardized predictor variables from the previous ones. I then compared the cumulative and acat families for this data both using the probit function.

get\_prior(PREP ~ RACE + INCOME + PARTYID + EDUC + GEN + CAGE,  
 data = prep)

## prior class coef group resp dpar nlpar lb ub  
## (flat) b   
## (flat) b CAGE   
## (flat) b EDUC   
## (flat) b GEN   
## (flat) b INCOME   
## (flat) b PARTYID   
## (flat) b RACE   
## student\_t(3, 4, 2.5) Intercept   
## student\_t(3, 0, 2.5) sigma 0   
## source  
## default  
## (vectorized)  
## (vectorized)  
## (vectorized)  
## (vectorized)  
## (vectorized)  
## (vectorized)  
## default  
## default

prep23 <- prep |>   
 mutate(INCOME\_s = (INCOME - mean(INCOME)) / sd(INCOME),  
 RACE\_s = (RACE - mean(RACE)) / sd(RACE),  
 PARTYID\_s = (PARTYID - mean(PARTYID)) / sd(PARTYID))  
  
fit23.333 <-  
 brm(data = prep23,  
 family = cumulative(probit),  
 PREP ~ 1 + RACE\_s + INCOME\_s + PARTYID\_s,  
 prior = c(prior(normal(4, 2.5), class = Intercept),  
 prior(normal(0, 2.5), class = b)),  
 iter = 3000, warmup = 1000, chains = 4, cores = 4,  
 seed = 23)

fit23.3333 <-  
 brm(data = prep23,  
 family = acat(probit),  
 PREP ~ 1 + RACE\_s + INCOME\_s + PARTYID\_s,  
 prior = c(prior(normal(4, 2.5), class = Intercept),  
 prior(normal(0, 2.5), class = b)),  
 iter = 3000, warmup = 1000, chains = 4, cores = 4,  
 seed = 23)

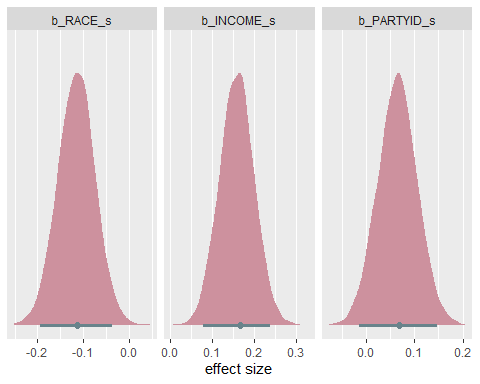
loo(fit23.333, fit23.3333)

## Output of model 'fit23.333':  
##   
## Computed from 8000 by 672 log-likelihood matrix.  
##   
## Estimate SE  
## elpd\_loo -1147.6 10.2  
## p\_loo 8.0 0.2  
## looic 2295.2 20.3  
## ------  
## MCSE of elpd\_loo is 0.0.  
## MCSE and ESS estimates assume MCMC draws (r\_eff in [0.9, 1.2]).  
##   
## All Pareto k estimates are good (k < 0.7).  
## See help('pareto-k-diagnostic') for details.  
##   
## Output of model 'fit23.3333':  
##   
## Computed from 8000 by 672 log-likelihood matrix.  
##   
## Estimate SE  
## elpd\_loo -1146.3 10.5  
## p\_loo 8.1 0.2  
## looic 2292.6 20.9  
## ------  
## MCSE of elpd\_loo is 0.0.  
## MCSE and ESS estimates assume MCMC draws (r\_eff in [0.7, 1.3]).  
##   
## All Pareto k estimates are good (k < 0.7).  
## See help('pareto-k-diagnostic') for details.  
##   
## Model comparisons:  
## elpd\_diff se\_diff  
## fit23.3333 0.0 0.0   
## fit23.333 -1.3 0.8

The regression coefficient estimates for race, income, and political party were all rather small in both models, with fit23.333 having greater estimates than fit23.3333. Using the loo() function to compare the models, I found that fit23.3333 had a smaller looic (2292.3) than fit 23.333 (2295.4), indicating that fit23.3333 using the acat family is a better fit.

Creating a data frame from the model fit23.333, I was able to create a graph of the distribution of effect sizes for each of the three predictor variables. All effect sizes were small, with the largest being income at 0.16, then race with -0.11, and lastly political leaning at 0.07.

draws <- as\_draws\_df(fit23.333)  
  
draws |>  
 pivot\_longer(ends\_with("\_s")) |>  
 mutate(name = factor(name,  
 levels = c("b\_RACE\_s","b\_INCOME\_s", "b\_PARTYID\_s"))) |>  
   
 ggplot(aes(x = value, y = 0)) +  
 stat\_halfeye(point\_interval = mode\_hdi, .width = .95,  
 fill = "pink3", color = "lightblue4", normalize = "panels") +  
 scale\_y\_continuous(NULL, breaks = NULL) +  
 xlab("effect size") +  
 facet\_wrap(~ name, scales = "free")



Conditional effects for model fit23.333 were used to create graphs showing the probability of preparedness for each predictor variable. Unfortunately these graphs don’t make a ton of sense as the data is categorical. Again, income is the only one with any sort of meaning.

ce <- conditional\_effects(fit23.333, categorical = T)  
  
## race prep  
plot(ce, plot = FALSE)[[1]] +  
 scale\_x\_continuous(expand = c(0, 0)) +  
 scale\_y\_continuous(expand = c(0, 0), limits = c(0, 1))  
  
## income prep  
plot(ce, plot = FALSE)[[2]] +  
 scale\_x\_continuous(expand = c(0, 0)) +  
 scale\_y\_continuous(expand = c(0, 0), limits = c(0, 1))  
  
## partyid prep  
plot(ce, plot = FALSE)[[5]] +  
 scale\_x\_continuous(expand = c(0, 0)) +  
 scale\_y\_continuous(expand = c(0, 0), limits = c(0, 1))

### Method 3.0 (aka please give me an A)

After attempting to follow multiple guides on ordinal regression of Likert-type data, I am still facing the struggle of my predictor variables being categorical and having a different number of categories within them. In a desperate attempt to get something meaningful I am going to create individual models for each predictor variable, race, income, and political leaning and compare them with the loo() function. Using the mutate function, I recoded every category to be their proper name, not a number representing them. I then created the individual models using brm() and the cumulative family with the probit function.

## goodbye numeric.   
prep7 <- prep |>  
 mutate(RACE = recode\_factor(RACE, '1' = "White",  
 '2' = "Black",  
 '3' = "Hispanic",  
 '4' = "Asian")) |>  
 mutate(INCOME = recode\_factor(INCOME, '1' = "Under25k",  
 '2' = "25k-50k",  
 '3' = "50k-75k",  
 '4' = "75k-100k")) |>  
 mutate(CAGE = recode\_factor(CAGE, '1' = "18-29",  
 '2' = "30-44",  
 '3' = "45-59",  
 '4' = "60+")) |>  
 mutate(PARTYID = recode\_factor(PARTYID, '1' = "Dem",  
 '2' = "Rep",  
 '3' = "Ind")) |>  
 mutate(EDUC = recode\_factor(EDUC, '1' = "No HS",  
 '2' = "HS Diploma",  
 '3' = "Some College",  
 '4' = "Bachelors",  
 '5' = "Graduate Courses")) |>  
 mutate(GEN = recode\_factor(GEN, '1' = "Male",  
 '2' = "Female"))

## race model  
r\_mod <- brm(PREP ~ 1 + RACE,  
 data = prep7,  
 family = cumulative("probit"))

## income model  
i\_mod <- brm(PREP ~ 1 + INCOME,  
 data = prep7,  
 family = cumulative("probit"))

## politcal leaning model  
p\_mod <- brm(PREP ~ 1 + PARTYID,  
 data = prep7,  
 family = cumulative("probit"))

loo(r\_mod, i\_mod, p\_mod)

## Output of model 'r\_mod':  
##   
## Computed from 4000 by 672 log-likelihood matrix.  
##   
## Estimate SE  
## elpd\_loo -1156.2 10.2  
## p\_loo 8.0 0.2  
## looic 2312.2 20.4  
## ------  
## MCSE of elpd\_loo is 0.0.  
## MCSE and ESS estimates assume MCMC draws (r\_eff in [0.6, 1.5]).  
##   
## All Pareto k estimates are good (k < 0.7).  
## See help('pareto-k-diagnostic') for details.  
##   
## Output of model 'i\_mod':  
##   
## Computed from 4000 by 672 log-likelihood matrix.  
##   
## Estimate SE  
## elpd\_loo -1150.0 10.1  
## p\_loo 8.1 0.2  
## looic 2300.0 20.2  
## ------  
## MCSE of elpd\_loo is 0.0.  
## MCSE and ESS estimates assume MCMC draws (r\_eff in [0.6, 1.6]).  
##   
## All Pareto k estimates are good (k < 0.7).  
## See help('pareto-k-diagnostic') for details.

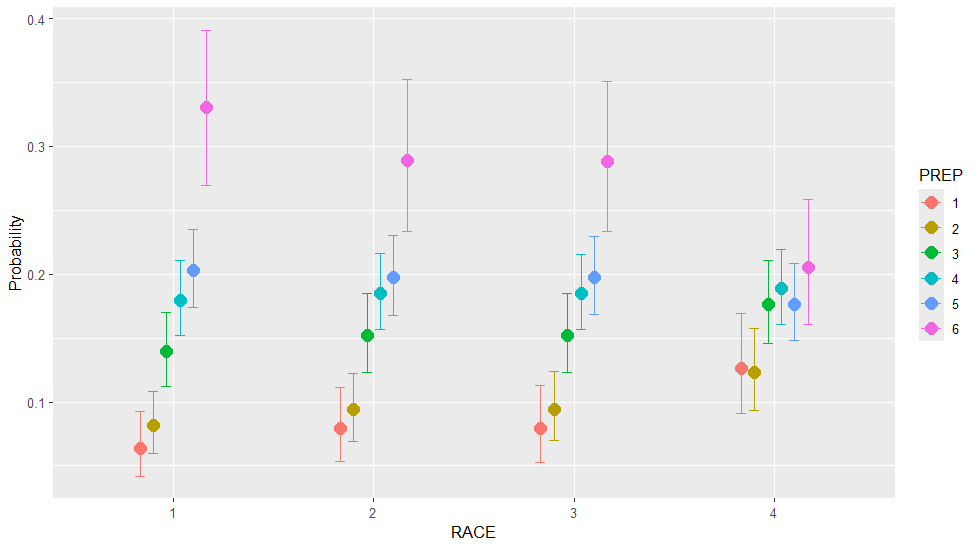
## Output of model 'p\_mod':  
##   
## Computed from 4000 by 672 log-likelihood matrix.  
##   
## Estimate SE  
## elpd\_loo -1156.2 10.2  
## p\_loo 7.0 0.2  
## looic 2312.5 20.4  
## ------  
## MCSE of elpd\_loo is 0.0.  
## MCSE and ESS estimates assume MCMC draws (r\_eff in [0.8, 1.3]).  
##   
## All Pareto k estimates are good (k < 0.7).  
## See help('pareto-k-diagnostic') for details.  
##   
## Model comparisons:  
## elpd\_diff se\_diff  
## i\_mod 0.0 0.0   
## r\_mod -6.2 6.0

## p\_mod -6.2 5.4

After comparison of the three above models with loo(), it was found that the model for income (looic = 2300.0) was a better indicator of preparedness than those for race (looic = 2312.3) and political leaning (looic = 2312.5).

### Do white people have a greater chance of being very prepared than minorities?

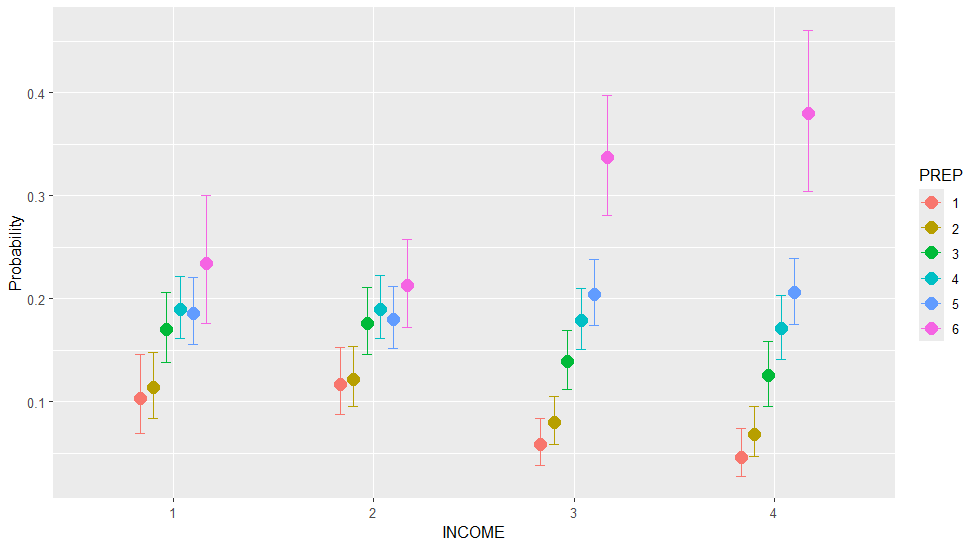
conditional\_effects(r\_mod, "RACE", categorical = TRUE)



The probability that a White person is very prepared for extreme weather events is 33%, followed by Black and Hispanic people with 29% being very prepared, and then by Asian people with 21% being very prepared. It can also be seen that the probability of being very unprepared is highest among Asian people and lowest among White people.

### Are individuals with higher income more likely to be very prepared?

conditional\_effects(i\_mod, "INCOME", categorical = TRUE)



There is a substantial increase in probability of being very prepared once an individuals income exceeds 50k. Surprisingly, individuals earning between 25k and 50k had a decreased probability of being very prepared compared to individuals making under 25k.

## Discussion

The 2019 Climate Change Poll surveyed over 2,000 Massachusetts residents on their views on climate change. Analysis of differing racial, income, gender, political leaning, age, and education demographics on an individuals preparedness for extreme climate events revealed that income was the greatest predictor of preparedness. An individuals race and political beliefs predicted preparedness to a less degree, and gender, age, and education had very little influence.  
It is important to note that survey data is not representative of the general public and is often skewed towards containing data from individuals that have the means to answer surveys and speak the language that the survey is given in. This can certainly be seen within the data collected, with 76% of participants having some college education or more and 48% identifying as democrats while only 13% identified as republicans. Creating the models using three different approaches allowed for an interesting comparison of ordinal regression methods. Unfortunately, the complexity of this data did not allow for me to do a very in depth analysis of the probability of an individuals preparedness when looking at multiple demographic factors. Parsing apart the three most significant predictor variables, race, income, and political leaning, allowed for model comparison and easily understood visualization of preparedness probability. The estimated error between categories of was substantial, rendering nearly everything besides income categories 3 and 4 as insignificant.

## Citations

Estrada-Martíne, L. M., Watanabe, P., & Rivera-Kientz, K. (2020, September). Views that Matter - Race and Opinions on Climate Change of Boston Area Residents. <https://www.umb.edu/media/umassboston/content-assets/iaas/SSL_Views_That_Matter_9-2020.pdf>