730 Group Project

Rebekah Kristal with collaborators Amani Chehimi & Shane Fitzgerald

2024-11-30

Rebekah's model: weighted ordinal regression w interactions

```
newdata <- read_csv("FreqCategories.csv") %>%  mutate(Weight = Freq / sum(Freq))
```

```
## New names:
## Rows: 5462 Columns: 9
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): AgeCat, EduCat dbl (7): ...1, y, REGION, SEX, RACENEW, POORYN, Freq
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
newdata<-mutate(newdata, weight.var=1/Freq) %>% mutate(REGION=as.factor(REGION)) %>% mutate(AgeCat=as.fa

#converting y's into factor variable, changing range from 0-8 to 1-9 to match with model output
newdata1<-mutate(newdata, y=y+1) %>% mutate(y, factor(y, ordered=TRUE))
modR<-brm(y|weights(Freq)~REGION + AgeCat + SEX + RACENEW + EduCat + POORYN + REGION*POORYN + REGION*RA
           chains = 4,
            iter = 2000, thin = 1)
```

```
## Compiling Stan program...
## Start sampling
```

```
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.006277 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 62.77 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
```

```
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 341.321 seconds (Warm-up)
## Chain 1:                272.521 seconds (Sampling)
## Chain 1:                613.842 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0.005025 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 50.25 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 364.68 seconds (Warm-up)
## Chain 2:                309.415 seconds (Sampling)
## Chain 2:                674.095 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 0.00558 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 55.8 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
```

```
## Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 314.618 seconds (Warm-up)
## Chain 3:                275.246 seconds (Sampling)
## Chain 3:                589.864 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL 'anon_model' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 0.004786 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 47.86 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 2073.86 seconds (Warm-up)
## Chain 4:                346.819 seconds (Sampling)
## Chain 4:                2420.68 seconds (Total)
## Chain 4:
```

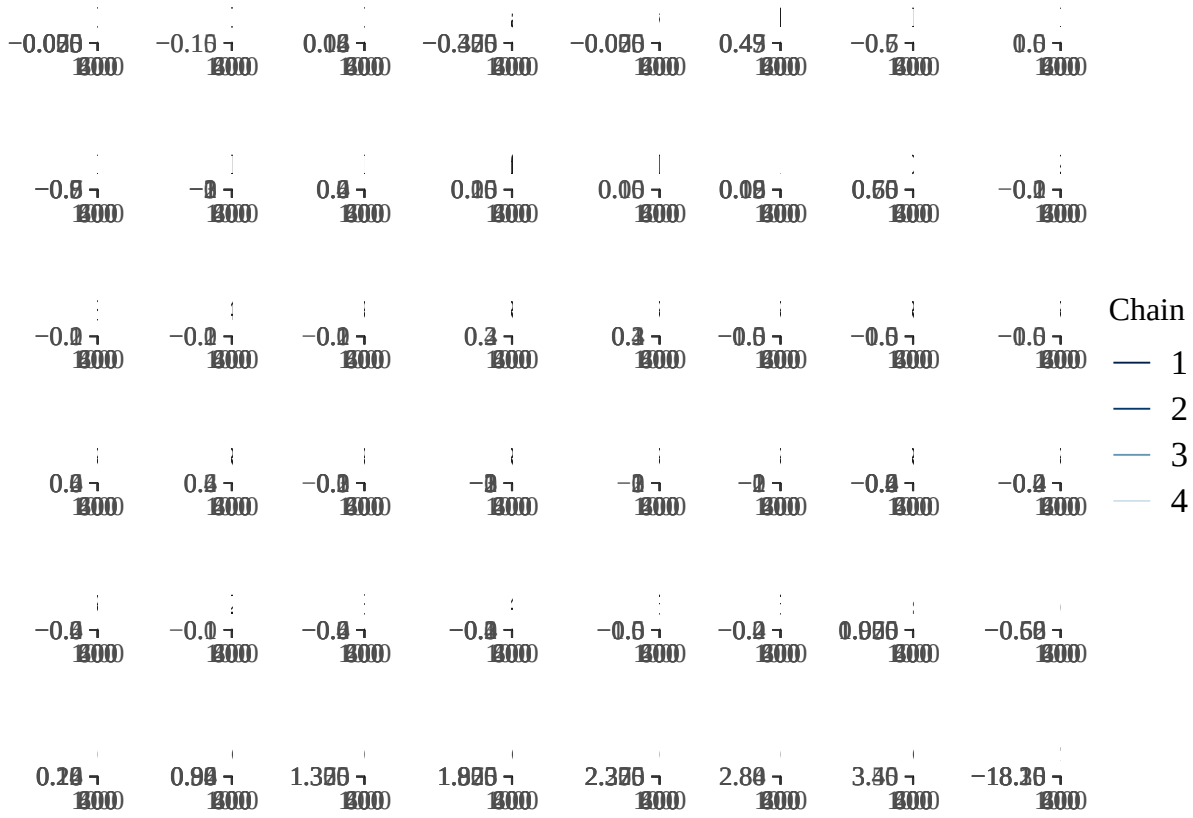
```
summary(modR)
```

```
##  Family: cumulative
##   Links: mu = logit; disc = identity
## Formula: y | weights(Freq) ~ REGION + AgeCat + SEX + RACENEW + EduCat + POORYN + REGION * POORYN + RI
##    Data: newdata1 (Number of observations: 5462)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                                       Estimate Est.Error l-95% CI u-95% CI
## Intercept[1]                             -0.37      0.02    -0.42    -0.33
## Intercept[2]                              0.41      0.02     0.37     0.45
## Intercept[3]                              1.13      0.02     1.09     1.18
## Intercept[4]                              1.54      0.02     1.50     1.59
## Intercept[5]                              2.08      0.02     2.04     2.13
## Intercept[6]                              2.53      0.02     2.48     2.57
## Intercept[7]                              3.05      0.02     3.00     3.10
## Intercept[8]                              3.67      0.03     3.61     3.72
## REGION2                                   0.01      0.02    -0.02     0.05
## REGION3                                  -0.12      0.02    -0.15    -0.09
```

```
## REGION4                                    0.09      0.02     0.05     0.13
## AgeCatelderly                             -0.38      0.01    -0.41    -0.36
## AgeCatyoungAdult                          -0.03      0.01    -0.06    -0.00
## SEX2                                       0.45      0.01     0.43     0.47
## RACENEW200                                -0.60      0.04    -0.69    -0.52
## RACENEW300                                 0.51      0.22     0.07     0.93
## RACENEW400                                -0.69      0.06    -0.80    -0.58
## RACENEW530                                -0.32      0.61    -1.52     0.83
## RACENEW541                                 0.35      0.11     0.13     0.57
## EduCatGraduateProfessionalorotherDegree    0.12      0.02     0.08     0.17
## EduCatHighschoolDiplomaGEDgraduate         0.02      0.02    -0.01     0.06
## EduCatSomecollegeAAorBachelorsDegree       0.12      0.02     0.09     0.16
## POORYN2                                    0.60      0.04     0.52     0.67
## REGION2:POORYN2                            0.01      0.05    -0.08     0.11
## REGION3:POORYN2                           -0.07      0.04    -0.16     0.02
## REGION4:POORYN2                           -0.11      0.05    -0.21    -0.02
## REGION2:RACENEW200                         0.04      0.06    -0.07     0.15
## REGION3:RACENEW200                         0.30      0.05     0.21     0.40
## REGION4:RACENEW200                         0.26      0.06     0.13     0.39
## REGION2:RACENEW300                        -0.38      0.25    -0.87     0.11
## REGION3:RACENEW300                        -0.39      0.24    -0.87     0.08
## REGION4:RACENEW300                        -0.68      0.23    -1.12    -0.22
## REGION2:RACENEW400                         0.27      0.08     0.11     0.44
## REGION3:RACENEW400                         0.34      0.07     0.20     0.48
## REGION4:RACENEW400                         0.11      0.06    -0.01     0.24
## REGION2:RACENEW530                        -0.07      0.70    -1.40     1.31
## REGION3:RACENEW530                         0.29      0.65    -0.97     1.58
## REGION4:RACENEW530                        -0.31      0.61    -1.47     0.93
## REGION2:RACENEW541                         0.08      0.14    -0.19     0.36
## REGION3:RACENEW541                         0.10      0.13    -0.15     0.35
## REGION4:RACENEW541                        -0.31      0.12    -0.56    -0.06
## RACENEW200:POORYN2                        -0.05      0.04    -0.13     0.02
## RACENEW300:POORYN2                        -0.32      0.11    -0.55    -0.11
## RACENEW400:POORYN2                        -0.19      0.06    -0.31    -0.07
## RACENEW530:POORYN2                        -0.23      0.28    -0.77     0.30
## RACENEW541:POORYN2                        -0.07      0.09    -0.25     0.11
##                                        Rhat Bulk_ESS Tail_ESS
## Intercept[1]                           1.00     3197     2892
## Intercept[2]                           1.00     3118     2906
## Intercept[3]                           1.00     3031     2968
## Intercept[4]                           1.00     3146     2697
## Intercept[5]                           1.00     3254     2816
## Intercept[6]                           1.00     3348     2673
## Intercept[7]                           1.00     3475     2771
## Intercept[8]                           1.00     3605     3002
## REGION2                                1.00     3349     3308
## REGION3                                1.00     3613     3386
## REGION4                                1.00     3421     3307
## AgeCatelderly                          1.00     6076     2791
## AgeCatyoungAdult                       1.00     5001     2941
## SEX2                                   1.00     6131     3407
## RACENEW200                             1.00     2609     2804
## RACENEW300                             1.00     2110     2528
## RACENEW400                             1.00     2396     2760
```

```
## RACENEW530                                1.00     2356     2613
## RACENEW541                                1.00     2253     2793
## EduCatGraduateProfessionalorotherDegree 1.00     3568     2859
## EduCatHighschoolDiplomaGEDgraduate       1.00     3531     2961
## EduCatSomecollegeAAorBachelorsDegree     1.00     2778     2755
## POORYN2                                   1.00     2352     2508
## REGION2:POORYN2                           1.00     2927     2786
## REGION3:POORYN2                           1.00     2374     2833
## REGION4:POORYN2                           1.00     2816     3302
## REGION2:RACENEW200                        1.00     3065     3045
## REGION3:RACENEW200                        1.00     2754     2934
## REGION4:RACENEW200                        1.00     3090     3121
## REGION2:RACENEW300                        1.00     2248     2748
## REGION3:RACENEW300                        1.00     2286     2869
## REGION4:RACENEW300                        1.00     2130     2582
## REGION2:RACENEW400                        1.00     3059     3232
## REGION3:RACENEW400                        1.00     2925     3085
## REGION4:RACENEW400                        1.00     2728     2932
## REGION2:RACENEW530                        1.00     2528     2875
## REGION3:RACENEW530                        1.00     2446     2916
## REGION4:RACENEW530                        1.00     2388     2797
## REGION2:RACENEW541                        1.00     2921     3135
## REGION3:RACENEW541                        1.00     2445     2974
## REGION4:RACENEW541                        1.00     2358     2635
## RACENEW200:POORYN2                        1.00     5442     3033
## RACENEW300:POORYN2                        1.00     4989     3266
## RACENEW400:POORYN2                        1.00     5022     2978
## RACENEW530:POORYN2                        1.00     4900     3009
## RACENEW541:POORYN2                        1.00     5114     3113
##
## Further Distributional Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00   NA       NA       NA
##
## 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).
```

```
mcmc_trace(modR)
```



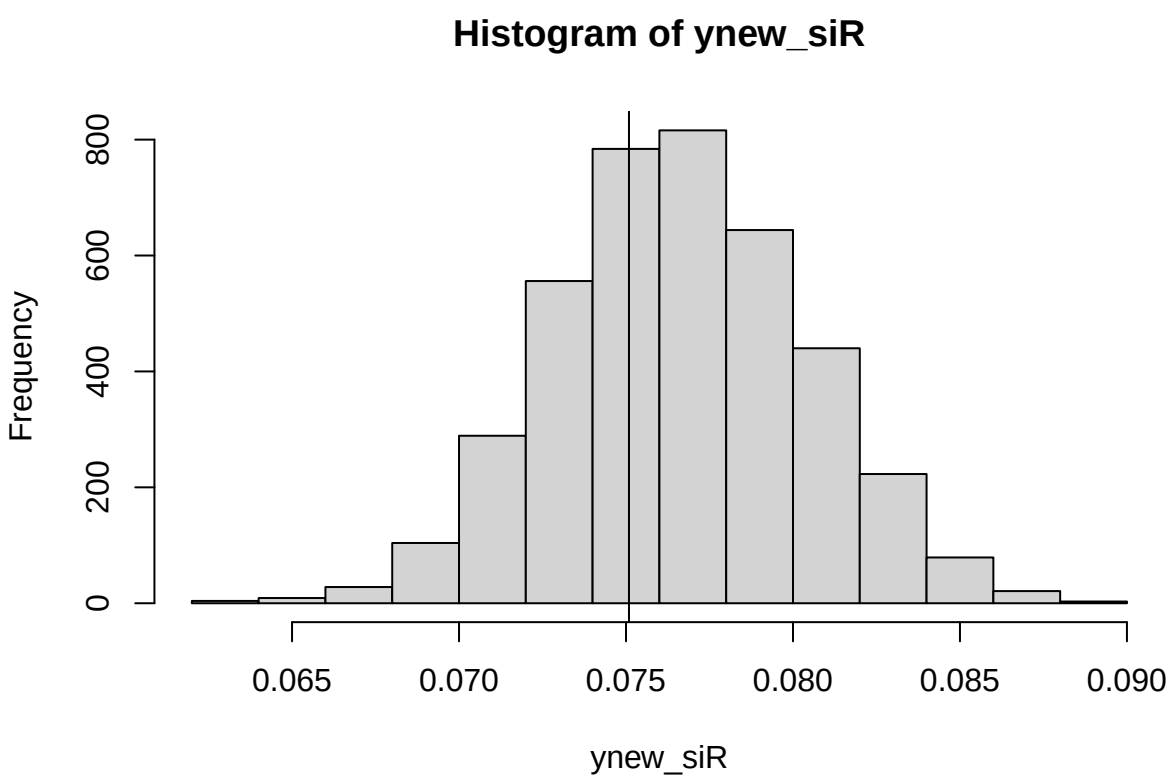
Analysis with Rebekah's Model

```
observed_counts <- select(newdata1, c(y, Freq))
total_freq<-group_by(observed_counts, y) %>% summarise(total=sum(Freq))
observed_props<-mutate(total_freq, observed=total/sum(total)) %>% mutate(y=as.factor(y))


get_sum_stat<-function(y, row){(sum(y==5))/nrow(row)}

tobs<-observed_props[5,3]

predicted_catsR<-as.data.frame(posterior_predict(modR))
ynew_siR<-apply(predicted_catsR, 1, get_sum_stat, newdata)
#ppc for proportion of observations in category 5
hist(ynew_siR)
abline(v = tobs)
```



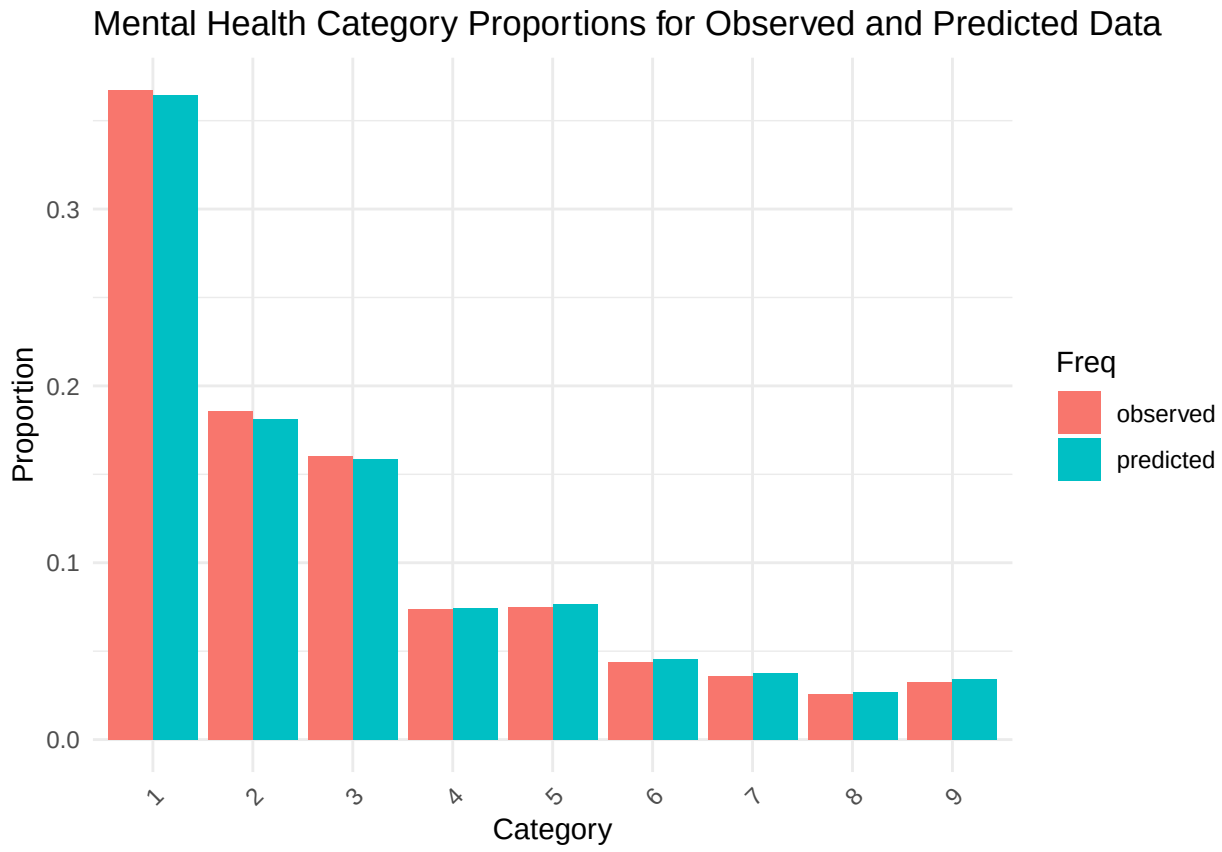

```
#ppc for all categories
#formatting for ggplot
posterior_preds_longR <- predicted_catsR %>%
  pivot_longer(cols = everything(), names_to = "chain", values_to = "predicted_category")

posterior_preds_longR$predicted_category <- as.factor(posterior_preds_longR$predicted_category)

category_countsR <- table(posterior_preds_longR$predicted_category)
category_counts_dfR <- as.data.frame(category_countsR)
colnames(category_counts_dfR) <- c("y", "Count")
category_counts_propR<-mutate(category_counts_dfR, predicted=Count/(4000*5462))

combinedR<-left_join(observed_props, category_counts_propR, by="y")
combined1R<-pivot_longer(combinedR, c(3,5), names_to = "Freq")

#plot of proportion of each category for observed and predicted data
ggplot(combined1R, mapping=aes(x=y, y=value, fill=Freq))+
  geom_bar(stat="identity", position="dodge")+
  labs(title = "Mental Health Category Proportions for Observed and Predicted Data",
       x = "Category",
       y = "Proportion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Mental Health Category Proportions for Observed and Predicted Data
Proportion
0.0
0.1
0.2
0.3
1
2
3
4
5
6
7
8
9
Category
Freq
observed
predicted