730 Group Project

Rebekah Kristal with collaborators Amani Chehimi & Shane Fitzgerald

2024-11-30

Amani's Model: weighted linear regression

```
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))



modA <- brm(
  y | weights(Weight) ~ (1 | REGION + AgeCat + SEX + RACENEW + EduCat + POORYN),
  family = gaussian(),
  data = newdata1,
  iter = 1000,
  chains = 4,
  cores = getOption("mc.cores", 4),
  seed = 12345
)
```

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

```
summary(modA)
```

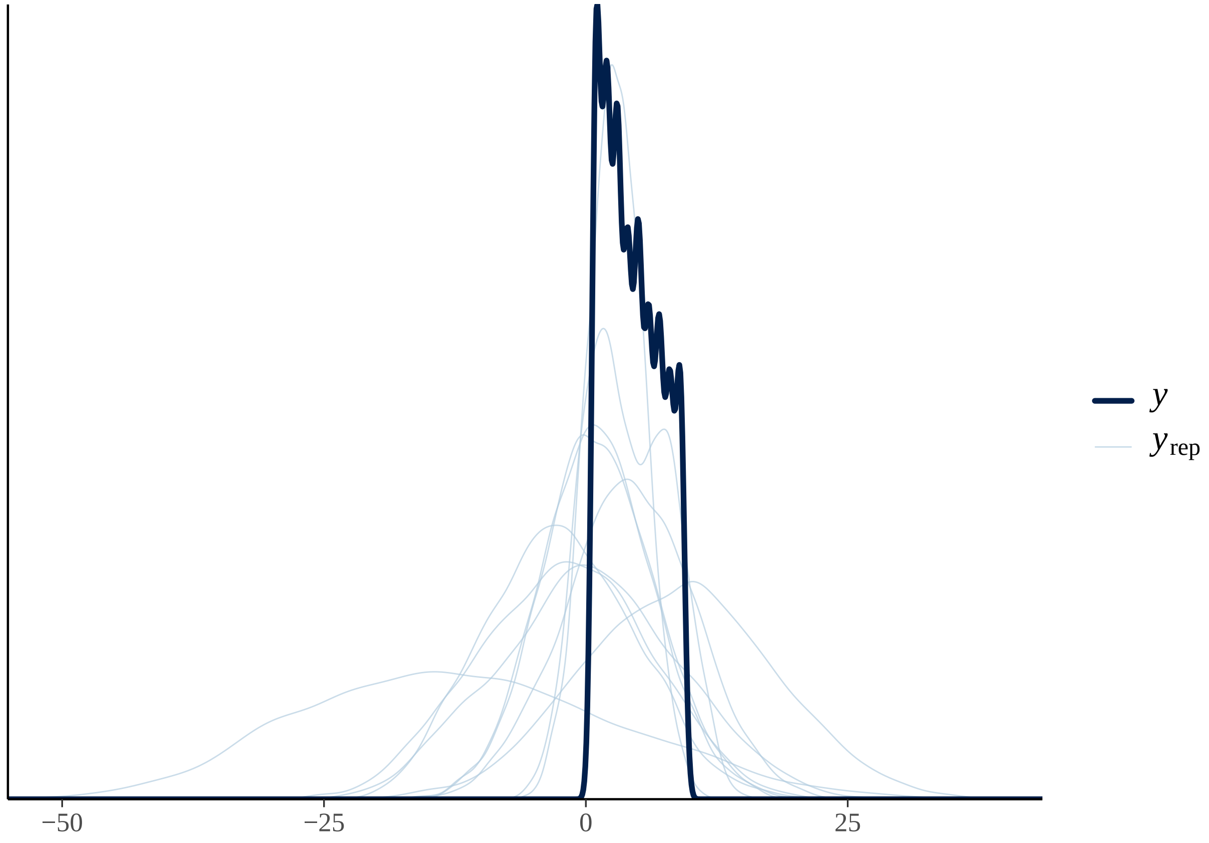
```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: y | weights(Weight) ~ (1 | REGION + AgeCat + SEX + RACENEW + EduCat + POORYN)
##    Data: newdata1 (Number of observations: 5462)
##   Draws: 4 chains, each with iter = 1000; warmup = 500; thin = 1;
##          total post-warmup draws = 2000
##
## Multilevel Hyperparameters:
## ~AgeCat (Number of levels: 3)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.46      2.35     0.08     8.39 1.00     2338      992
##
## ~EduCat (Number of levels: 4)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.39      2.12     0.09     7.93 1.00     1998     1179
##
## ~POORYN (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.55      2.36     0.12     8.83 1.00     1689     1088
##
## ~RACENEW (Number of levels: 6)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.48      2.31     0.08     8.55 1.00     2080      940
##
## ~REGION (Number of levels: 4)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.32      2.26     0.07     7.90 1.00     1739     1037
##
## ~SEX (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     2.50      2.46     0.10     8.54 1.00     2590     1320
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     3.67      3.67    -3.94    10.73 1.00     2520     1185
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     5.89      4.21     1.87    14.78 1.00     2839     1446
##
## 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).
```

```
prior_summary(modA)
```

```
##                prior     class      coef   group resp dpar nlpar lb ub
##  student_t(3, 4, 3) Intercept
##  student_t(3, 0, 3)        sd                                    0
##  student_t(3, 0, 3)        sd              AgeCat                  0
##  student_t(3, 0, 3)        sd Intercept  AgeCat                  0
##  student_t(3, 0, 3)        sd              EduCat                  0
##  student_t(3, 0, 3)        sd Intercept  EduCat                  0
##  student_t(3, 0, 3)        sd              POORYN                  0
##  student_t(3, 0, 3)        sd Intercept  POORYN                  0
##  student_t(3, 0, 3)        sd             RACENEW                  0
##  student_t(3, 0, 3)        sd Intercept RACENEW                  0
##  student_t(3, 0, 3)        sd            REGION                  0
##  student_t(3, 0, 3)        sd Intercept  REGION                  0
##  student_t(3, 0, 3)        sd               SEX                  0
##  student_t(3, 0, 3)        sd Intercept     SEX                  0
##  student_t(3, 0, 3)     sigma                                    0
##        source
##       default
##       default
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##  (vectorized)
##       default
```

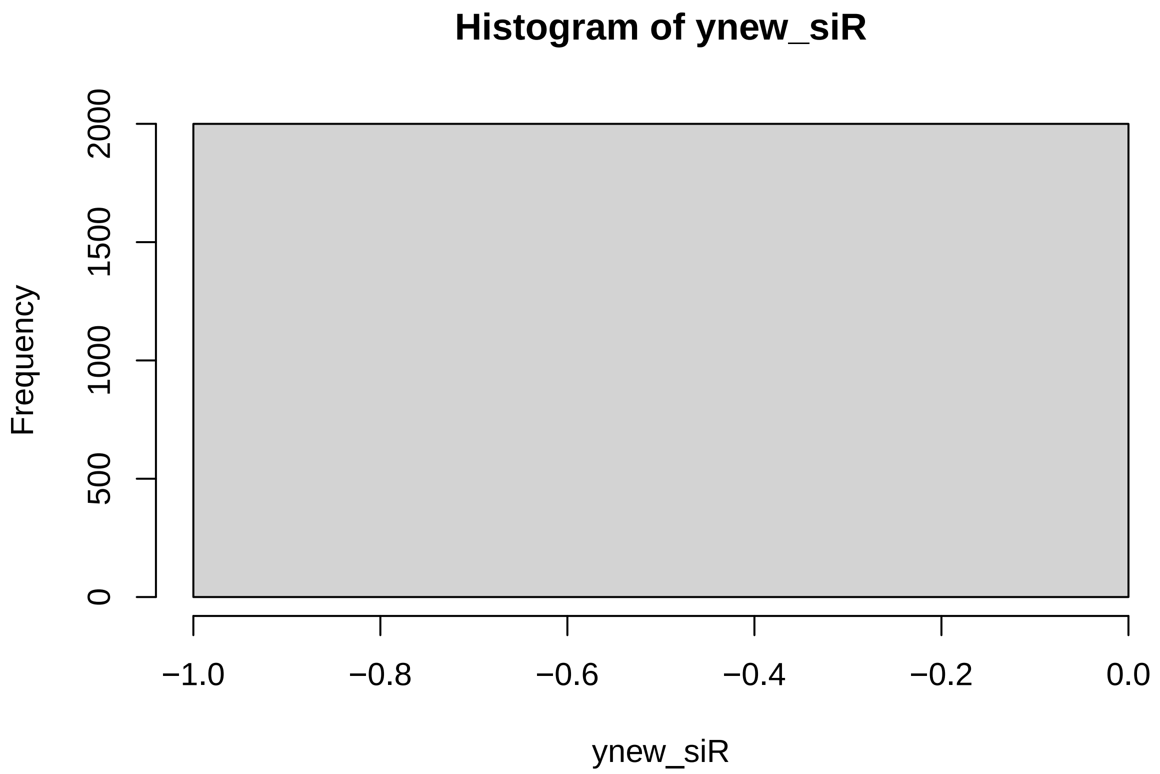
```
pp_check(modA)
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```



Analysis with Amani's Model

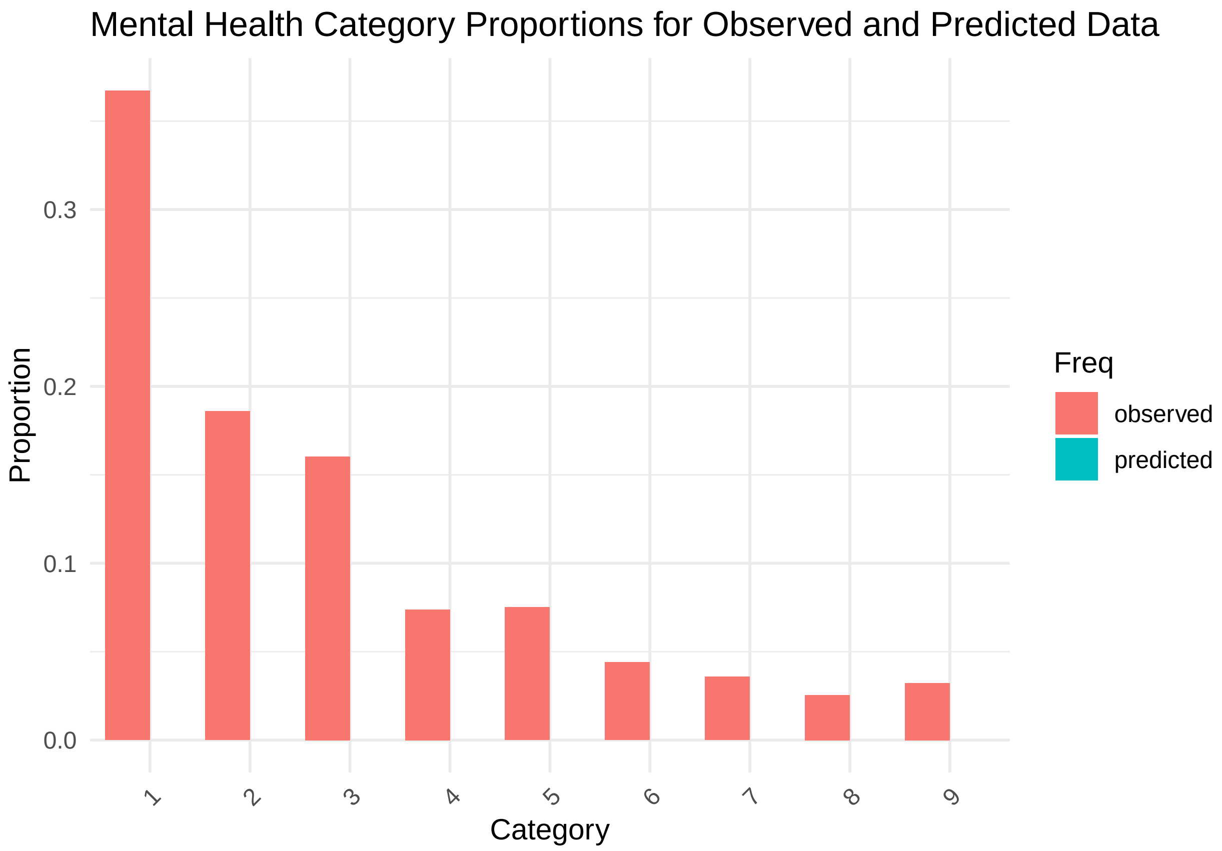
```
observed_counts <- select(newdata1, c(y, Freq))
total_freq<-group_by(observed_counts, y) %>% summarise(total=sum(Freq))
observed_props<-mutate(total_freq, observed=total/sum(total)) %>% mutate(y=as.factor(y))


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

tobs<-observed_props[5,3]

predicted_catsR<-as.data.frame(posterior_predict(modA))
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))
```

```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```


Mental Health Category Proportions for Observed and Predicted Data
Proportion
Category
Freq
🟥 observed
🟦 predicted