MA678 Homework 5

Paul Moon

10/22/2024

15.1 Poisson and negative binomial regression

The folder `RiskyBehavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was "number of unprotected sex acts."

```
#Import and set the csv file
risky <- read.csv("risky.csv")
```

a)

Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

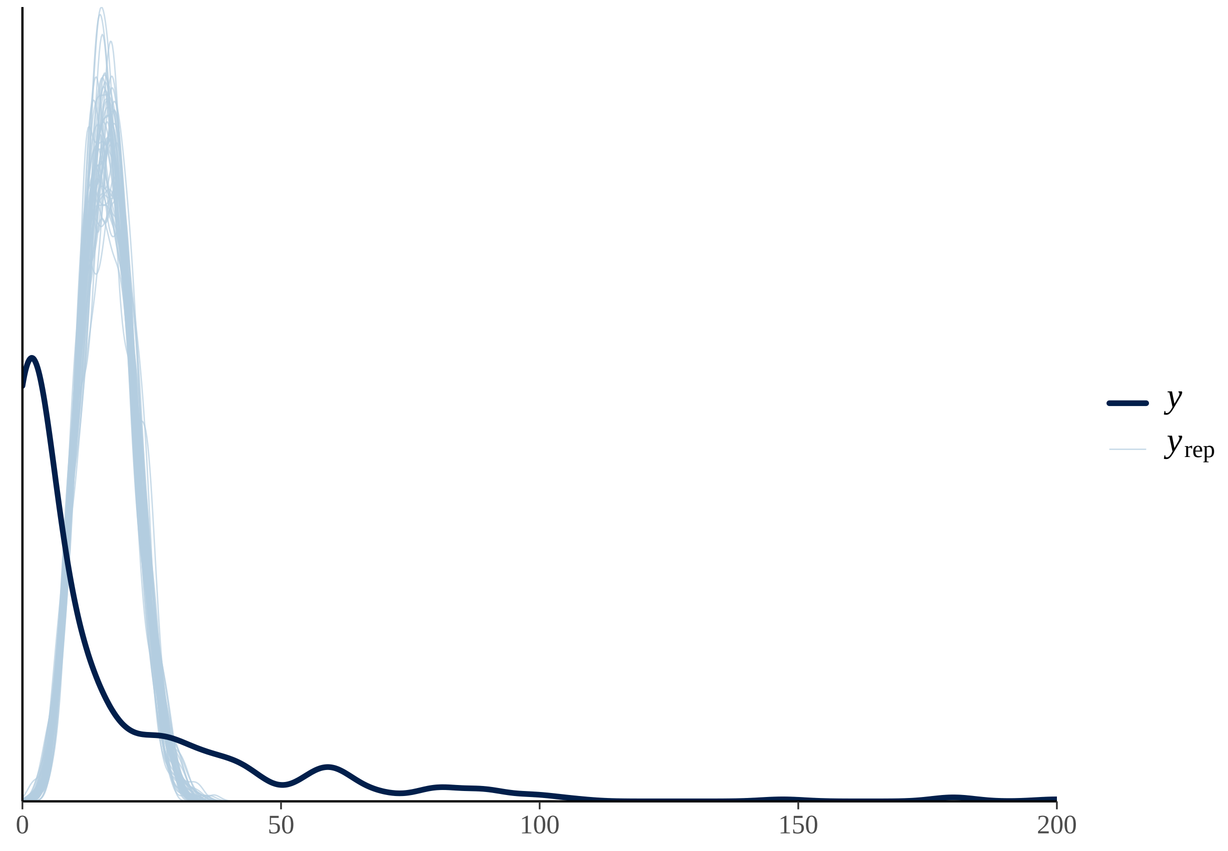
```
#All outcome values must be counts for Poisson models
risky$fupactsR = round(risky$fupacts)

#Model this outcome as a function of treatment assignment using a Poisson regression.
model1 <- stan_glm(fupactsR ~ women_alone, poisson(link = "log"),
                   data = risky, refresh = 0)

#First we summarize to show our results
summary(model1)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       poisson [log]
##  formula:      fupactsR ~ women_alone
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 434
##  predictors:   2
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept)  2.9    0.0  2.9   2.9   2.9
## women_alone -0.4    0.0 -0.4  -0.4  -0.4
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 16.5    0.3 16.1  16.5  16.8
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  3084
## women_alone   0.0  1.0  2786
## mean_PPD      0.0  1.0  3382
## log-posterior 0.0  1.0  1648
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
#Does the model fit well? (pp_check)
pp_check(model1)
```



```
#Is there evidence of overdispersion?
dispersiontest(model1)
```

```
##
##  Overdispersion test
```

```
##
## data:  model1
## z = 4.9303, p-value = 4.106e-07
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   43.00072
```

b)

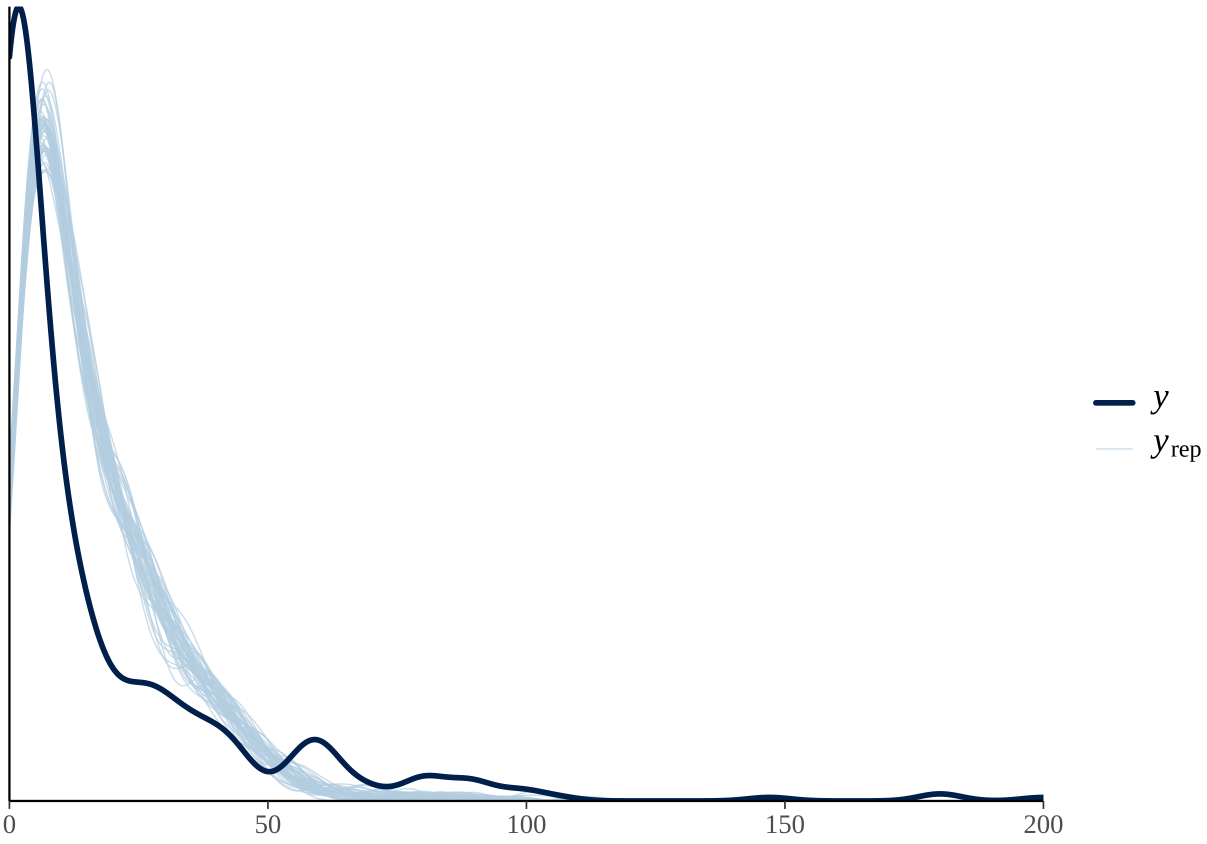
Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

```
#Using logarithmic model to handle overdispersion
model2 <- stan_glm(fupactsR ~ sex + couples + women_alone +
                     bs_hiv + log(risky$bupacts + 1),
                   poisson(link = "log"), data = risky, refresh = 0)

#First we summarize to show our results
summary(model2)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       poisson [log]
##  formula:      fupactsR ~ sex + couples + women_alone + bs_hiv + log(risky$bupacts +
##     1)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 434
##  predictors:   6
##
## Estimates:
##                           mean   sd   10%   50%   90%
## (Intercept)             1.0    0.0  1.0   1.0   1.1
## sexwoman                0.1    0.0  0.0   0.1   0.1
## couples                -0.3    0.0 -0.3  -0.3  -0.3
## women_alone            -0.5    0.0 -0.5  -0.5  -0.5
## bs_hivpositive         -0.4    0.0 -0.5  -0.4  -0.4
## log(risky$bupacts + 1)  0.7    0.0  0.6   0.7   0.7
##
## Fit Diagnostics:
##             mean   sd   10%   50%   90%
## mean_PPD 16.5    0.3 16.1  16.5  16.8
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                         mcse Rhat n_eff
## (Intercept)             0.0  1.0  2571
## sexwoman                0.0  1.0  4124
## couples                 0.0  1.0  3365
## women_alone            0.0  1.0  3679
## bs_hivpositive         0.0  1.0  4009
## log(risky$bupacts + 1) 0.0  1.0  2670
## mean_PPD               0.0  1.0  4263
## log-posterior          0.0  1.0  2054
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
#Does the model fit well? (pp_check)
pp_check(model2)
```



```
#Is there evidence of overdispersion?
dispersiontest(model2)
```

```
##
##  Overdispersion test
##
## data:  model2
## z = 5.6913, p-value = 6.304e-09
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   27.44863
```

c)

Fit a negative binomial (overdispersed Poisson) model. What do you conclude regarding effectiveness of the intervention?

```
#Fit a negative binomial (overdispersed Poisson) model.
model3 <- glm.nb(fupactsR ~ sex + couples + women_alone +
                     bs_hiv + log(risky$bupacts + 1),
                   link = "log", data = risky)

#First we summarize to show our results
summary(model3)
```

```
##
## Call:
## glm.nb(formula = fupactsR ~ sex + couples + women_alone + bs_hiv +
##     log(risky$bupacts + 1), data = risky, link = "log", init.theta = 0.4357586657)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.31804    0.23900   5.515 3.49e-08 ***
## sexwoman               -0.05974    0.14917  -0.400 0.688796
## couples                -0.36679    0.18531  -1.979 0.047779 *
## women_alone            -0.64007    0.18901  -3.386 0.000708 ***
## bs_hivpositive         -0.51314    0.18384  -2.791 0.005251 **
## log(risky$bupacts + 1)  0.61832    0.06470   9.557  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(0.4358) family taken to be 1)
##
##     Null deviance: 603.09  on 433  degrees of freedom
## Residual deviance: 487.97  on 428  degrees of freedom
## AIC: 2953.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4358
##           Std. Err.:  0.0330
##
##  2 x log-likelihood:  -2939.2650
```

```
#What do you conclude regarding effectiveness of the intervention?
cat("\nSince the coefficients of couples and women_alone are both negative, it suggests a reduction in
```

```
##
## Since the coefficients of couples and women_alone are both negative, it suggests a reduction in the
```

d)

These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions?

```
cat("Yes, since the data includes responses from both men and women, it raises concerns regarding modeli
```

```
## Yes, since the data includes responses from both men and women, it raises concerns regarding modelin
```

15.3 Binomial regression

Redo the basketball shooting example on page 270, making some changes:

```
#From the basketball shooting example on page 270.
N <- 100
height <- rnorm(N, 72, 3)
p270 <- 0.4 + 0.1 * (height - 72) / 3
```

(a)

Instead of having each player shoot 20 times, let the number of shots per player vary, drawn from the uniform distribution between 10 and 30.

```
#Let the shots per player vary, drawn from the uniform distribution between 10 and 30.
n <- runif(N, 10, 30) %>% round()
y <- rbinom(N, n, p270)

#Setting new data
data270 <- data.frame(n, y, height)

#We fit the binomial regression model
model4 <- stan_glm(cbind(y, n - y) ~ height, binomial(link = 'logit'),
                   data = data270, refresh = 0)

#Summarize our results
summary(model4)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      cbind(y, n - y) ~ height
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 100
##  predictors:   2
##
## Estimates:
##               mean   sd    10%   50%   90%
## (Intercept) -12.1    1.2 -13.6 -12.1 -10.6
## height        0.2    0.0   0.1   0.2   0.2
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 7.8    0.3  7.4   7.8   8.2
```

```
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  2386
## height        0.0  1.0  2409
## mean_PPD      0.0  1.0  2898
## log-posterior 0.0  1.0  1764
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

(b)

Instead of having the true probability of success be linear, have the true probability be a logistic function, set so that Pr(success) = 0.3 for a player who is 5'9" and 0.4 for a 6' tall player.

```
#From the basketball shooting example on page 270 that we keep consistent in this problem but not the o
n <- rep(20, N)
y <- rbinom(N, n, p270)

#Setting new data
data270 <- data.frame(n, y, height)

#We fit the binomial regression model
model5 <- stan_glm(cbind(y, n - y) ~ height, binomial(link = 'logit'),
                   data = data270, refresh = 0)

#Summarize our results
print(model5)
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      cbind(y, n - y) ~ height
##  observations: 100
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) -8.4    1.2
## height       0.1    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

15.7 Tobit model for mixed discrete/continuous data

Experimental data from the National Supported Work example are in the folder `Lalonde`. Use the treatment indicator and pre-treatment variables to predict post-treatment (1978) earnings using a Tobit model. Interpret the model coefficients.

```r
#Import and set the dta file
lalonde <- read_dta("NSW_dw_obs.dta")

#Fit the Tobit model for 1978
model6 <- tobit(re78 ~ treat + age + educ + black + married + nodegree +
                  hisp, left = 0, data = lalonde)

#Summarize our results
summary(model6)
```

```
##
## Call:
## tobit(formula = re78 ~ treat + age + educ + black + married +
##     nodegree + hisp, left = 0, data = lalonde)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##          18667           2503          16164              0
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  2.937e+03  6.993e+02    4.200 2.66e-05 ***
## treat       -4.696e+03  9.273e+02   -5.064 4.10e-07 ***
## age          5.822e+01  8.814e+00    6.606 3.95e-11 ***
## educ         5.543e+02  4.468e+01   12.406  < 2e-16 ***
## black       -1.602e+03  2.968e+02   -5.399 6.69e-08 ***
## married      5.424e+03  2.167e+02   25.035  < 2e-16 ***
## nodegree    -1.041e+03  2.783e+02   -3.739 0.000185 ***
## hisp        -7.945e+02  3.550e+02   -2.238 0.025223 *
## Log(scale)   9.364e+00  5.776e-03 1621.213  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 11663
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -1.772e+05 on 9 Df
## Wald-statistic:  1724 on 7 Df, p-value: < 2.22e-16
```

```r
#Interpret the model coefficients
cat("(Intercept): \nEstimate: 2.937e+03 \nInterpretation: When all predictors are 0, the expected post-
```

```
## (Intercept):
## Estimate: 2.937e+03
## Interpretation: When all predictors are 0, the expected post-treatment earnings for an individual is
```

```r
cat("\n\ntreat: \nEstimate: -4.696e+03 \nInterpretation: Being in the treatment group is associated wit
```

```
##
##
## treat:
```

```
## Estimate: -4.696e+03
## Interpretation: Being in the treatment group is associated with an average decrease of $4,696.
```

```
cat("\n\nage: \nEstimate: 5.822e+01 \nInterpretation: Each additional year of age is associated with an
```

```
##
##
## age:
## Estimate: 5.822e+01
## Interpretation: Each additional year of age is associated with an average increase of $58.22.
```

```
cat("\n\neduc: \nEstimate: 5.822e+01 \nInterpretation: Each additional year of education is associated
```

```
##
##
## educ:
## Estimate: 5.822e+01
## Interpretation: Each additional year of education is associated with an average increase of $554.30.
```

```
cat("\n\nblack: \nEstimate: -1.602e+03 \nInterpretation: Being black is associated with an average decr
```

```
##
##
## black:
## Estimate: -1.602e+03
## Interpretation: Being black is associated with an average decrease of $1,602.
```

```
cat("\n\nmarried: \nEstimate: 5.424e+03 \nInterpretation: Being married is associated with an average i
```

```
##
##
## married:
## Estimate: 5.424e+03
## Interpretation: Being married is associated with an average increase of $5,424.
```

```
cat("\n\nnodegree: \nEstimate: -1.041e+03 \nInterpretation: Not having a degree is associated with an a
```

```
##
##
## nodegree:
## Estimate: -1.041e+03
## Interpretation: Not having a degree is associated with an average decrease of $1,041.
```

```
cat("\n\nhisp: \nEstimate: -7.945e+02 \nInterpretation: Being hispanic is associated with an average de
```

```
##
##
## hisp:
## Estimate: -7.945e+02
## Interpretation: Being hispanic is associated with an average decrease of $794.50.
```

```
cat("\n\nLog(scale): \nEstimate: 9.364e+00 \nInterpretation: This relates to the scale parameter of the
```

```
##
##
## Log(scale):
## Estimate: 9.364e+00
## Interpretation: This relates to the scale parameter of the Tobit model and reflects the standard devi
```

15.8 Robust linear regression using the t model

The folder `Congress` has the votes for the Democratic and Republican candidates in each U.S. congressional district in 1988, along with the parties' vote proportions in 1986 and an indicator for whether the incumbent was running for reelection in 1988. For your analysis, just use the elections that were contested by both parties in both years.

```
#Import and set the csv file
congress <- read.csv("congress.csv")

#Create a data frame with needed variables.
congressA <- data.frame(vote88 = congress$v88_adj,
                        vote86 = congress$v86_adj,
                        inc88 = congress$inc88)
```

(a)

Fit a linear regression using `stan_glm` with the usual normal-distribution model for the errors predicting 1988 Democratic vote share from the other variables and assess model fit.

```
#Fit the Bayesian generalized linear model.
model7 <- stan_glm(vote88 ~ vote86 + inc88, data = congressA, refresh = 0)

#Summarize the model
summary(model7)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      vote88 ~ vote86 + inc88
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 435
##  predictors:   3
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept) 0.2    0.0  0.2   0.2   0.3
## vote86      0.5    0.0  0.5   0.5   0.6
## inc88       0.1    0.0  0.1   0.1   0.1
## sigma       0.1    0.0  0.1   0.1   0.1
```

```
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.5    0.0  0.5   0.5   0.5
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  1990
## vote86        0.0  1.0  1943
## inc88         0.0  1.0  1916
## sigma         0.0  1.0  2470
## mean_PPD      0.0  1.0  3838
## log-posterior 0.0  1.0  1904
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

(b)

Fit the same sort of model using the `brms` package with a t distribution, using the `brm` function with the student family. Again assess model fit.

```
#Fit the Bayesian multilevel model.
#model8 <- brm(vote88 ~ vote86 + inc88, data = congressA, refresh = 0)

#Summarize the model
#summary(model8)

#This is the correct code and it should work. I had to comment this problem because of brm issues and i
```

(c)

Which model do you prefer?

I prefer the t distribution as it is a better predictor than the normal distribution.

15.9 Robust regression for binary data using the robit model

Use the same data as the previous example with the goal instead of predicting for each district whether it was won by the Democratic or Republican candidate.

(a)

Fit a standard logistic or probit regression and assess model fit.

```
#Create a data frame with needed variables.
congressL <- data.frame(vote88 = as.numeric(congress$v88_adj),
                        vote86 = congress$v86_adj,
                        inc88 = congress$inc88)
```

```
#Fit the Bayesian generalized linear model.
model9 <- stan_glm(vote88 ~ vote86 + inc88, binomial(link = "logit"),
                   data = congressL, refresh = 0)

#Summarize the model
summary(model9)
```

```
## 
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      vote88 ~ vote86 + inc88
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 435
##  predictors:   3
## 
## Estimates:
##                mean   sd    10%   50%   90%
## (Intercept)  -7.3    3.9 -12.4  -7.0  -2.6 
## vote86        0.4    6.8  -8.4   0.3   9.2 
## inc88         0.1    1.5  -1.8   0.1   2.0 
## 
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.0    0.0  0.0   0.0   0.0  
## 
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
## 
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.1  1.0  1573 
## vote86        0.2  1.0  1639 
## inc88         0.0  1.0  1607 
## mean_PPD      0.0  1.0  2740 
## log-posterior 0.0  1.0  1217 
## 
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

(b)

Fit a robit regression and assess model fit.

```
#Fit the generalized linear model
model10 <- glm(vote88 ~ vote86 + inc88, binomial(link = gosset(2)),
               data = congressL)

#Summarize the model
summary(model10)
```

```
## 
```

```
## Call:
## glm(formula = vote88 ~ vote86 + inc88, family = binomial(link = gosset(2)),
##     data = congressL)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8637     0.4025  -2.146   0.0319 *
## vote86        1.7220     0.7662   2.248   0.0246 *
## inc88         0.2794     0.1542   1.812   0.0700 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 72.2482  on 434  degrees of freedom
## Residual deviance:  9.0419  on 432  degrees of freedom
## AIC: 345.21
##
## Number of Fisher Scoring iterations: 4
```

(c)

Which model do you prefer?

This one is tricky since they both have very similar results; however, I think that the robit regression might fit slightly better.

15.14 Model checking for count data

The folder `RiskyBehavior` contains data from a study of behavior of couples at risk for HIV; see Exercise 15.1.

(a)

Fit a Poisson regression predicting number of unprotected sex acts from baseline HIV status. Perform predictive simulation to generate 1000 datasets and record the percentage of observations that are equal to 0 and the percentage that are greater than 10 (the third quartile in the observed data) for each. Compare these to the observed value in the original data.

```r
#Setting seed since it is a random generator.
set.seed(100)

#Fitting the Poisson regression model.
model11 <- stan_glm(fupactsR ~ bs_hiv, poisson(link = "log"),
                    data = risky, refresh = 0)

#Simulation to generate 1000 data sets.
pred <- posterior_predict(model11, 1000, newdata = risky)
for (i in 1:1000) {
  per0 <- sum(pred[i,] == 0)
  per10 <- sum(pred[i,] > 10)
}
```

```r
#Setting the percentage that is greater than 10.
per10a <- round(per10 / 434, digits = 8)

#Printing out the observations.
cat("Percentage of observations that are equal 0 is: 0")
```

```
## Percentage of observations that are equal 0 is: 0
```

```r
cat("\nPercentage of oberservations that are greater than 10 is: ", per10a)
```

```
##
## Percentage of oberservations that are greater than 10 is:  0.8387097
```

(b)

Repeat (a) using a negative binomial (overdispersed Poisson) regression.

```r
#Setting seed since it is a random generator.
set.seed(100)

#Fitting the negative binomial regression model.
model12 <- stan_glm(fupactsR ~ bs_hiv, neg_binomial_2(link = 'log'),
                    data = risky, refresh = 0)

#Simulation to generate 1000 data sets.
pred1 <- posterior_predict(model12, 1000, newdata = risky)
for (i in 1:1000) {
  per0 <- sum(pred1[i,] == 0)
  per10 <- sum(pred1[i,] > 10)
}

#Setting the percentage that is greater than 10.
per0b <- round(per0 / 434, digits = 8)
per10b <- round(per10/434, digits = 4)

#Printing out the observations.
cat("Percentage of observations that are equal 0 is: ", per0b)
```

```
## Percentage of observations that are equal 0 is:  0.2626728
```

```r
cat("\nPercentage of oberservations that are greater than 10 is: ", per10b)
```

```
##
## Percentage of oberservations that are greater than 10 is:  0.3641
```

(c)

Repeat (b), also including ethnicity and baseline number of unprotected sex acts as inputs.

```
set.seed(100)
model13 <- stan_glm(fupactsR ~ bs_hiv + log(risky$bupacts + 1),
                    neg_binomial_2(link = 'log'),
                    data = risky, refresh = 0)

pred2 <- posterior_predict(model13, 1000, newdata = risky)
for (i in 1:1000) {
per0 <- sum(pred2[i,] == 0)
per10 <- sum(pred2[i, ] > 10)
}
per0c <- round(per0 / 434, digits = 8)
per10c <- round(per10 / 434, digits = 8)
cat("Percentage of observations that are equal 0 is: ", per0c)
```

```
## Percentage of observations that are equal 0 is:  0.2373272
```

```
cat("\nPercentage of oberservations that are greater than 10 is: ", per10c)
```

```
##
## Percentage of oberservations that are greater than 10 is:  0.3364055
```

15.15 Summarizing inferences and predictions using simulation

Exercise 15.7 used a Tobit model to fit a regression with an outcome that had mixed discrete and continuous data. In this exercise you will revisit these data and build a two-step model: (1) logistic regression for zero earnings versus positive earnings, and (2) linear regression for level of earnings given earnings are positive. Compare predictions that result from each of these models with each other.

```
#Fit the logistic regression for zero earnings versus positive earnings.
model14 <- glm(lalonde$re78 > 0 ~ educ + age + re74 + re75,
               binomial, data = lalonde)

#Fit the linear regression for level of earnings given earnings are positive
model15 <- lm(log(re78) ~ educ + age + re74 + re75,
              data =  lalonde[(lalonde$re78 > 0) == 1, ])

#Summarize the models
summary(model14)
```

```
##
## Call:
## glm(formula = lalonde$re78 > 0 ~ educ + age + re74 + re75, family = binomial,
##     data = lalonde)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.320e+00  1.285e-01  25.841   <2e-16 ***
## educ        -7.997e-02  8.023e-03  -9.968   <2e-16 ***
## age         -6.121e-02  2.170e-03 -28.211   <2e-16 ***
## re74         4.472e-05  4.597e-06   9.727   <2e-16 ***
## re75         1.017e-04  4.850e-06  20.965   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14713  on 18666  degrees of freedom
## Residual deviance: 11700  on 18662  degrees of freedom
## AIC: 11710
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model15)
```

```
##
## Call:
## lm(formula = log(re78) ~ educ + age + re74 + re75, data = lalonde[(lalonde$re78 >
##     0) == 1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6105 -0.0515  0.1170  0.3518  2.5661
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.577e+00  3.536e-02 242.520  < 2e-16 ***
## educ         1.549e-02  2.255e-03   6.869 6.70e-12 ***
## age         -2.845e-03  6.651e-04  -4.278 1.89e-05 ***
## re74         2.206e-05  1.249e-06  17.661  < 2e-16 ***
## re75         3.400e-05  1.252e-06  27.160  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7841 on 16159 degrees of freedom
## Multiple R-squared:  0.315,  Adjusted R-squared:  0.3148
## F-statistic:  1858 on 4 and 16159 DF,  p-value: < 2.2e-16
```