15-1-15-2

Nikhil Gopal

2/13/2022

```
rm(list = ls())
library(rstanarm)
library(rosdata)
library(bayesplot)
library(AER)
```

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."

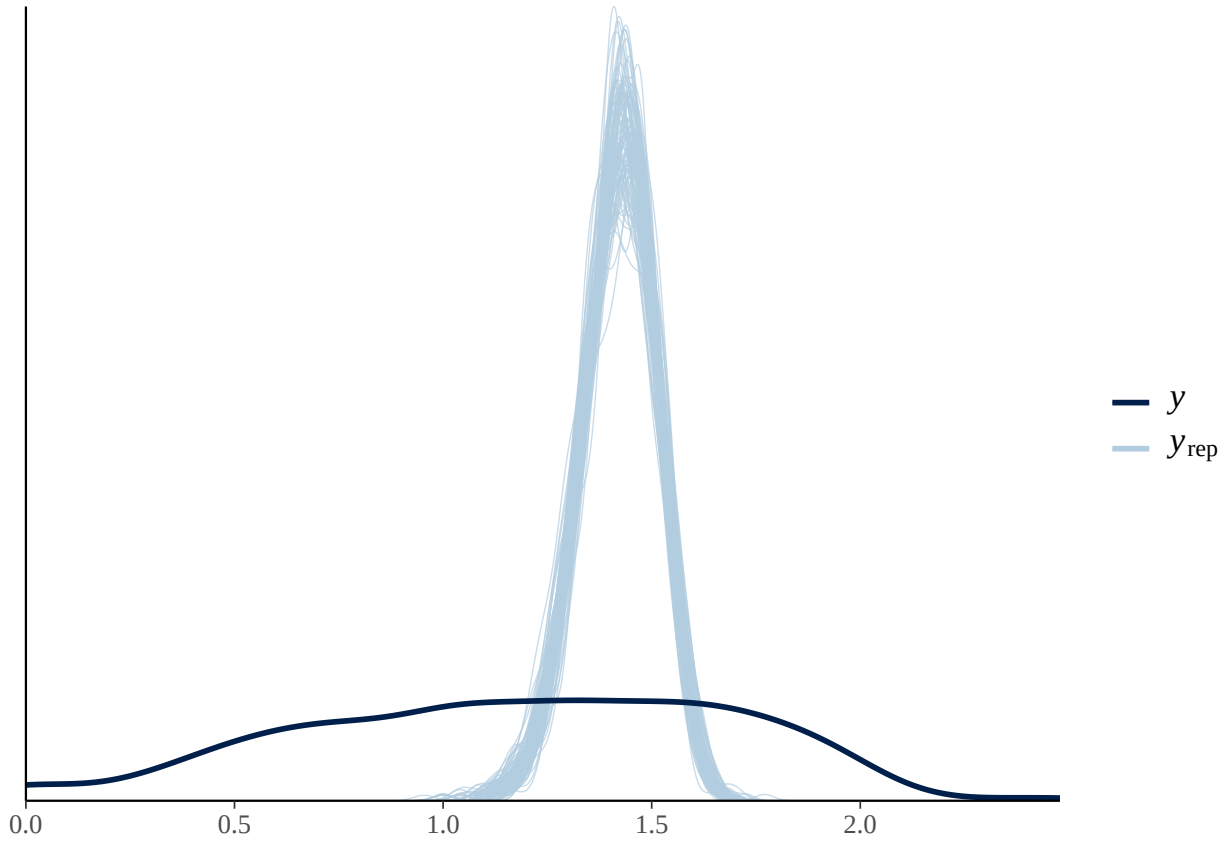
- a: Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

```
#need rosdata package
data("risky")

fit_1 <- stan_glm(bupacts ~ couples+women_alone, family=poisson(link="log"), data=risky, refresh = 0)

print(fit_1)
```

```
## stan_glm
##  family:       poisson [log]
##  formula:      bupacts ~ couples + women_alone
##  observations: 434
##  predictors:   3
## ------
##             Median MAD_SD
## (Intercept) 3.2    0.0
## couples     0.1    0.0
## women_alone 0.0    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Let us evaluate model fit:

```
#from page 269
yrep_1 <- posterior_predict(fit_1)

n_sims <- nrow(yrep_1)
subset <- sample(n_sims, 100)

ppc_dens_overlay(log10(risky$bupacts+1), log10(yrep_1[subset,]+1))
```




```
test <- function (y){
  mean(y==0)
}

test_rep_1 <- apply(yrep_1, 1, test)
```

We observe in the density overlay plot that the simulated values as predicted by the model do not appear to overlap well with the actual data. The blue line represents the replicated values and the black line represents the actual data.

```
mean(risky$bupacts)
```

```
## [1] 25.91014
```

```r
var(risky$bupacts)
```

```
## [1] 1018.756
```

```r
dispersiontest(fit_1, trafo = 1)
```

```
##
##  Overdispersion test
##
## data:  fit_1
## z = 4.093, p-value = 2.129e-05
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 37.61402
```

Above I perferformed a hypothesis test for overdispersion. In poisson regression, we assume that:

$$E(Y) = \mu / Var(y) = \mu$$

The above tests the alternative hypothesis that

$$Var(y) = \mu + c * f(\mu), c \neq 0$$

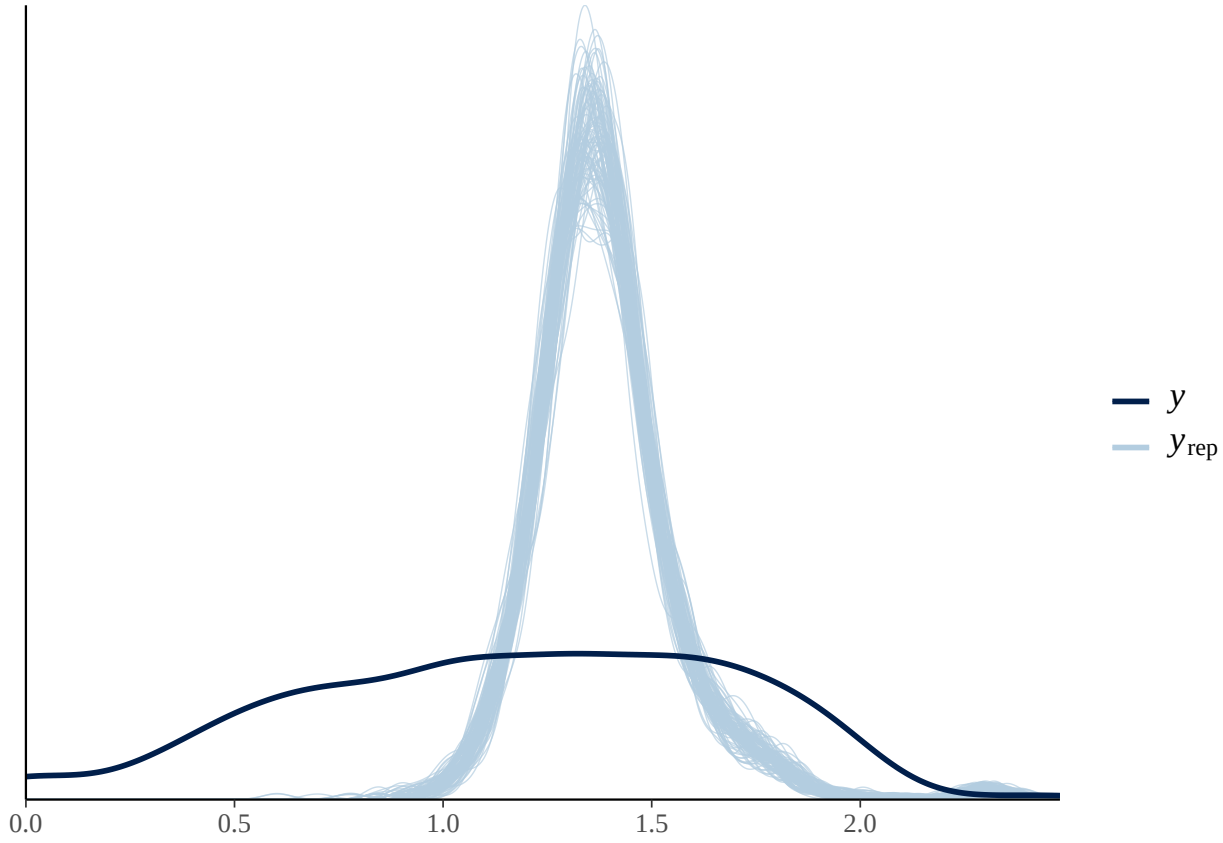
In this case, the p value was less than 0.05 and near 5, meaning that we can reject the null hypothesis and assume that the above case is true, showing strong evidence of over dispersion. This aligns with the poor model fit that we observed in the graph above.

- b: Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the data set. Does the model fit well? Is there evidence of over dispersion?

```r
fit_2 <- stan_glm(bupacts ~., family=poisson(link="log"), data=risky, refresh = 0)

print(fit_2)
```

```
## stan_glm
##  family:       poisson [log]
##  formula:      bupacts ~ .
##  observations: 434
##  predictors:   6
## ------
##                Median MAD_SD
## (Intercept)     2.9    0.0
## sexwoman       -0.1    0.0
## couples         0.2    0.0
## women_alone     0.1    0.0
## bs_hivpositive -0.1    0.0
## fupacts         0.0    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```r
#from page 269
yrep_1 <- posterior_predict(fit_2)

n_sims <- nrow(yrep_1)
subset <- sample(n_sims, 100)

ppc_dens_overlay(log10(risky$bupacts+1), log10(yrep_1[subset,]+1))
```




```r
test <- function (y){
  mean(y==0)
}

test_rep_1 <- apply(yrep_1, 1, test)
```

```r
dispersiontest(fit_2, trafo = 1)
```

```
##
##  Overdispersion test
##
## data:  fit_2
## z = 4.3902, p-value = 5.661e-06
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 26.62353
```

```r
hist(fit_2$residuals)
```

Like the previous model, this model also showed poor model fit and evidence of over disperion. The replicated data predicted by the model does not overlap well with the real data. The overdispersion test also provided strong evidence for over dispersion.

- c: Fit a negative binomial (over dispersed Poisson) model. What do you conclude regarding effectiveness of the intervention?

Fit model:

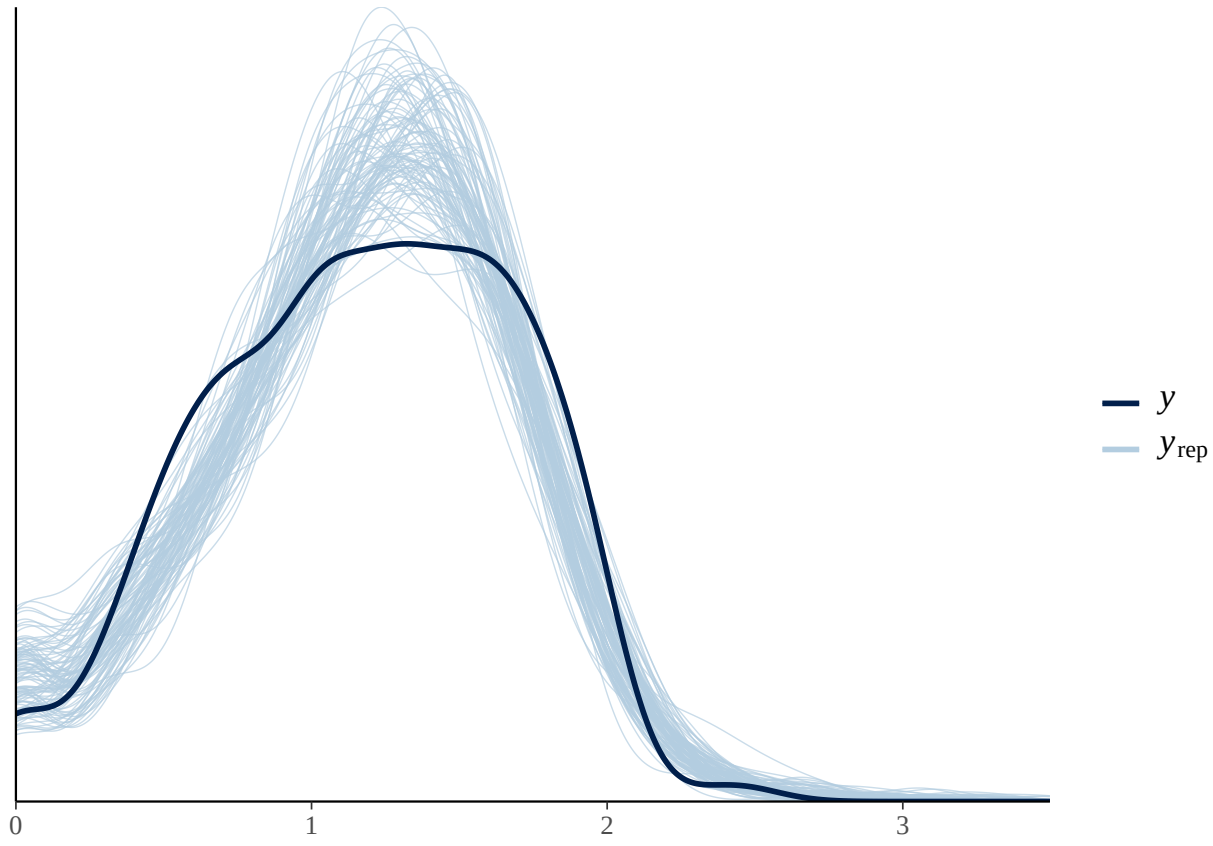
```r
fit_nb <- stan_glm(bupacts ~., family=neg_binomial_2(link="log"), data=risky, refresh = 0)

print(fit_nb)
```

```
## stan_glm
##  family:       neg_binomial_2 [log]
##  formula:      bupacts ~ .
##  observations: 434
##  predictors:   6
## ------
##                Median MAD_SD
## (Intercept)     2.8    0.1
## sexwoman       -0.1    0.1
```

```
## couples          0.3    0.1
## women_alone      0.1    0.1
## bs_hivpositive -0.1    0.1
## fupacts          0.0    0.0
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 1.0    0.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Evaluate fit:

```
#from page 269
yrep_1 <- posterior_predict(fit_nb)

n_sims <- nrow(yrep_1)
subset <- sample(n_sims, 100)

ppc_dens_overlay(log10(risky$bupacts+1), log10(yrep_1[subset,]+1))
```




```
test <- function (y){
  mean(y==0)
```

```
}

test_rep_1 <- apply(yrep_1, 1, test)
```

As shown in the graph above, the replicated data y values are much closer to the actual data values. This model is much more effective and fits the data much better.

- 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?

```
exp(coef(fit_nb))
```

```
##    (Intercept)       sexwoman        couples    women_alone bs_hivpositive
##     17.1864385      0.9018252      1.3182366      1.1089623      0.8884002
##        fupacts
##      1.0153722
```

Data where solely men were provided the treatment was not collected. Exponentiating coefficients allows us to interpret their effects on the treatment variable as multiplicative. Groups where women alone were provided treatment were shown to increase the number of sex acts by a factor of 1.1 whereas being a woman was shown to decrease the number of sex acts by a factor of 0.9. This is contradictory, and casts doubt on the efficacy of the model. This does give me concern and it would have been better to have included instances where only men were provided the treatment in the sample.

5.2

Offset in a Poisson or negative binomial regression: Explain why putting the logarithm of the exposure into a Poisson or negative binomial model as an offset, is equivalent to including it as a regression predictor, but with its coefficient fixed to the value 1.

Offset variables allow one to adjust for time spent surveyed in a study. In the above example with the HIV data, it was not documented how long each couple was measured for, and using an offset variable allows one to adjust for cases where some couples were surveyed for longer than others. In these cases, couples that had a longer survey period would naturally be expected to have more sex, and thus higher counts of the outcome variable (unprotected sex acts). The offset variable coefficient is fixed to a value of 1 to allow for the effects of time to be counted in predicting the outcome. If time spent surveyed were less than the baseline exposure, this would reduce the predicted number of unprotected sex acts, whereas baseline time above average would increase the predicted count.