MA678 HW5

Wendy liang

10/14/2020

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?

Firstly, I make a sumary.

To summarize:

- `sex` is the sex of the person, recorded as "man" or "woman" here
- `couples` is an indicator for if the couple was counseled together
- `women_alone` is an indicator for if the woman went to counseling by herself
- `bs_hiv` indicates if the individual is HIV positive
- `bupacts` is the number of unprotected sex acts reported as a baseline (before treamtnet)
- `fupacts` is the number of unprotected sex acts reported at the end of the study

```
#summary
risk <- read.csv("risky.csv",header=T)
risk$fupacts = round(risk$fupacts)
risk$couples=factor(risk$couples)
risk$women_alone=factor(risk$women_alone)
summary(risk)
```

```
##      sex             couples women_alone    bs_hiv             bupacts
##  Length:434         0:272   0:288       Length:434         Min.   :  0.00
##  Class :character   1:162   1:146       Class :character   1st Qu.:  5.00
##  Mode  :character                       Mode  :character   Median : 15.00
##                                                            Mean   : 25.91
##                                                            3rd Qu.: 36.00
##                                                            Max.   :300.00
##     fupacts
##  Min.   :  0.00
##  1st Qu.:  0.00
##  Median :  5.00
##  Mean   : 16.49
##  3rd Qu.: 21.00
##  Max.   :200.00
```

Then, I use poisson model to fit.

```r
#possion regresison
fit1=glm(fupacts~women_alone,family = poisson,data=risk)
summary(fit1)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone, family = poisson, data = risk)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.093  -4.979  -3.304   1.237  27.150
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.92114    0.01368  213.58   <2e-16 ***
## women_alone1 -0.40367    0.02719  -14.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 13064  on 432  degrees of freedom
## AIC: 14393
##
## Number of Fisher Scoring iterations: 6
```

```r
#check over
portion_factor = 13064/432
portion_factor
```

```
## [1] 30.24074
```

```r
#plot(risk$women_alone,risk$fupacts)
#curve(exp(coef(fit1)[1]+coef(fit1)[2]*x),add=TRUE)
```

Although the `woman_alone` is statistically significant, the [poisson model is poor. Since the portion factor = Residual/degrees » 1, so this model may has overdisperision.

b)

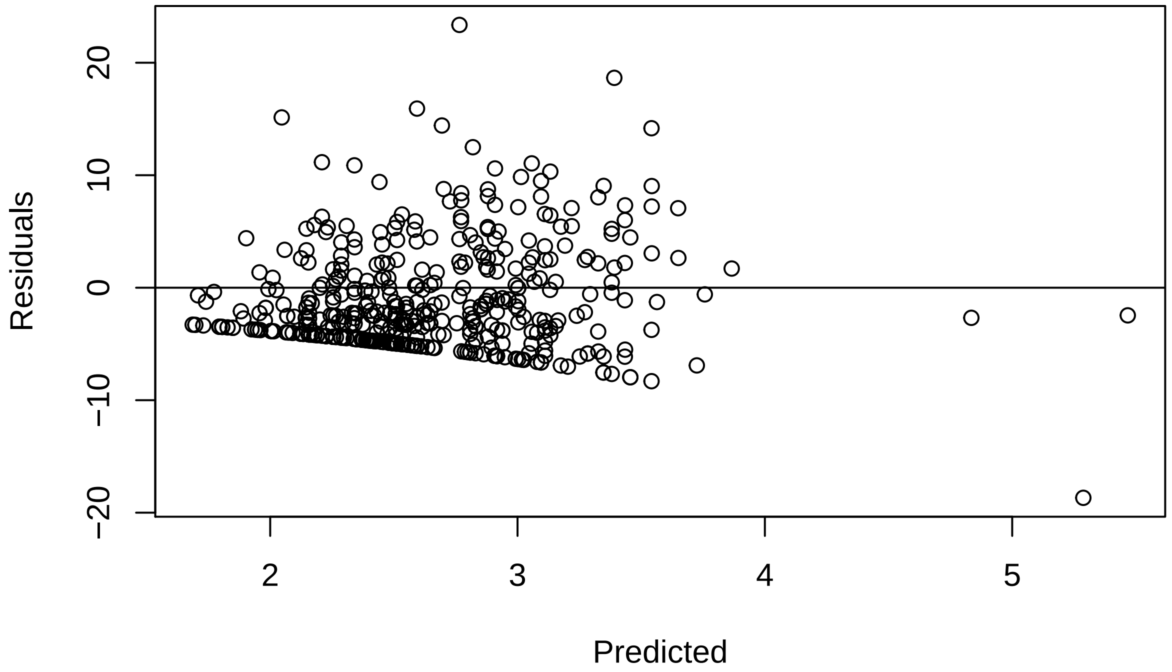
Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

```r
fit2=glm(fupacts~women_alone+sex+bupacts+couples+bs_hiv,family=poisson,data=risk)
summary(fit2)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + sex + bupacts + couples +
##     bs_hiv, family = poisson, data = risk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.679   -4.305   -2.511    1.368   23.361
```

```
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     2.7871257  0.0235599 118.300  < 2e-16 ***
## women_alone1   -0.6622159  0.0308962 -21.434  < 2e-16 ***
## sexwoman        0.1086694  0.0237301   4.579 4.66e-06 ***
## bupacts         0.0107789  0.0001738  62.013  < 2e-16 ***
## couples1       -0.4099761  0.0282298 -14.523  < 2e-16 ***
## bs_hivpositive -0.4383170  0.0353804 -12.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 13299  on 433  degrees of freedom
## Residual deviance: 10200  on 428  degrees of freedom
## AIC: 11537
## 
## Number of Fisher Scoring iterations: 6
```

```r
#residual plot
plot(predict(fit2),residuals(fit2),xlab="Predicted",ylab="Residuals")
abline(a=0,b=0)
```



```r
#overdispersion test
#[link]https://www.sciencedirect.com/science/article/abs/pii/030440769090014K
dispersiontest(fit2)
```

```
## 
##  Overdispersion test
## 
## data:  fit2
## z = 5.5689, p-value = 1.282e-08
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   29.65146
```

- The residual increases as the predicted values increase.
- This model has improved but it's still not good.
- I use function dispersiontest() and find this model has high overdispersion of 29.65.

c)

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

```
fit3=glm(fupacts~women_alone+sex+bupacts+couples+bs_hiv,family=quasipoisson,data=risk)
display(fit3)
```

```
## glm(formula = fupacts ~ women_alone + sex + bupacts + couples +
##     bs_hiv, family = quasipoisson, data = risk)
##                 coef.est coef.se
## (Intercept)      2.79     0.13
## women_alone1    -0.66     0.17
## sexwoman         0.11     0.13
## bupacts          0.01     0.00
## couples1        -0.41     0.15
## bs_hivpositive -0.44     0.19
## ---
##   n = 434, k = 6
##   residual deviance = 10200.4, null deviance = 13298.6 (difference = 3098.2)
##   overdispersion parameter = 30.0
```

The intervention decreased the number of unprotected sex acts fupacts. When only women participate in, the decrease is $1-e^{-0.66} = 1-0.51685=48.31% $. When both men and women participate in, the decrease is $1-e^{-0.41} = 1-0.66365=33.63% $

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?

In fact, the predictor women_alone and couples are not independent.

15.3 Binomial regression:

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

(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.

```
N <- 100
height <- rnorm(N, 72, 3)
p <- 0.4 + 0.1*(height - 72)/3
n <- round(runif(N,10,30)) #I forget round()
y <- rbinom(N, n, p)
data <- data.frame(n=n, y=y, height=height)
```

```
fit3_a <- stan_glm(cbind(y, n-y) ~ height, family=binomial(link="logit"),data=data,refresh=0)
print(fit3_a)
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      cbind(y, n - y) ~ height
##  observations: 100
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) -13.0    1.3
## height        0.2    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

(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.

```
N <- 100
height <- rnorm(N, 72, 3)
p<-ifelse(height>72,0.4,0.3)
n <- round(runif(N,10,30))
y <- rbinom(N, n, p)
data <- data.frame(n=n, y=y, height=height)
fit3_b <- stan_glm(cbind(y, n-y) ~ height, family=binomial(link="logit"),data=data,refresh=0)
print(fit3_b)
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      cbind(y, n - y) ~ height
##  observations: 100
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) -5.7    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.

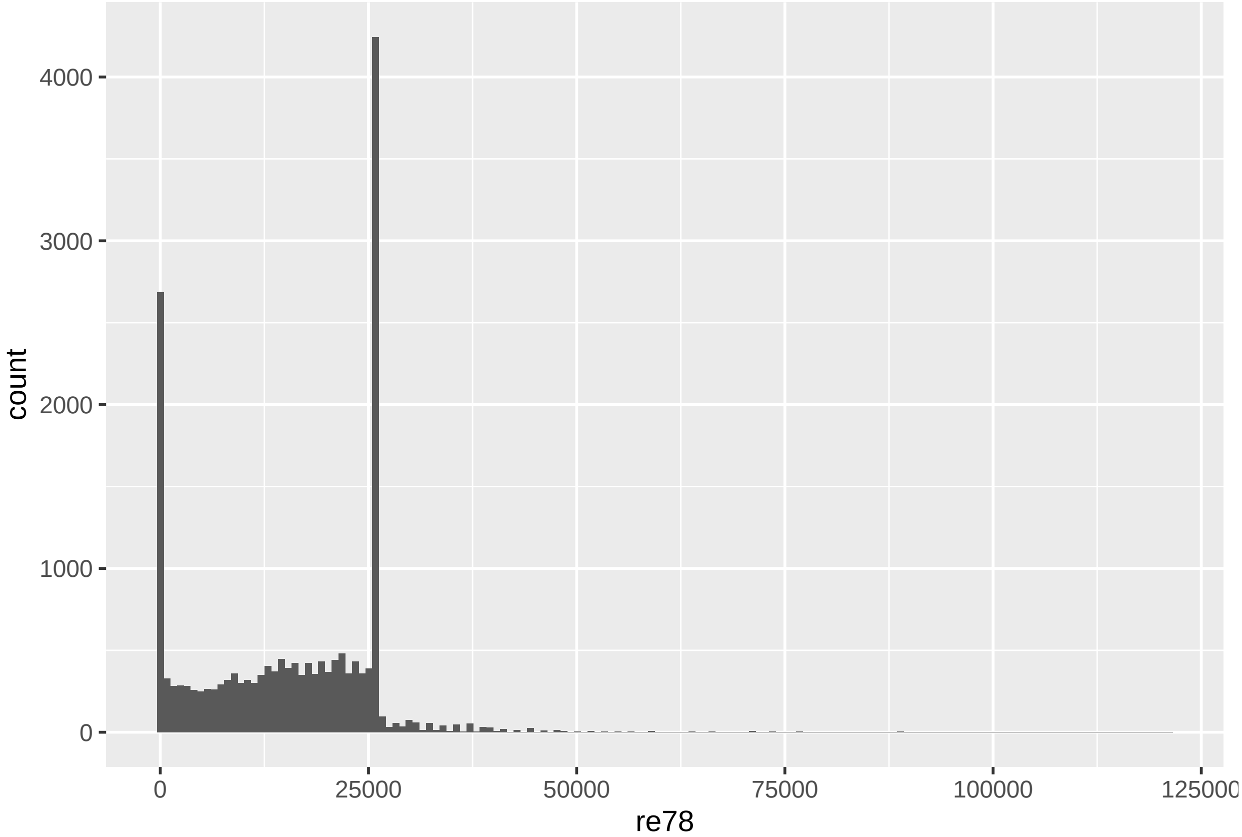
This dataset description is [link]http://www.stat.columbia.edu/~gelman/arm/examples/lalonde/NSW.vars.final.doc.

I gain help from [link]https://github.com/IamGianluca/arm/blob/master/ch6/arm_ch6p5.ipynb

```
nsw = read.dta("NSW_dw_obs.dta")

# variables as factor
nsw$sample = factor(nsw$sample, labels=c("NSW", "CPS", "PSID"))
nsw$black = factor(nsw$black)
nsw$hisp = factor(nsw$hisp)
nsw$nodegree = factor(nsw$nodegree)
nsw$married = factor(nsw$married)
nsw$treat = factor(nsw$treat)
nsw$educ_cat4 = factor(nsw$educ_cat4, labels=c("less than high school", "high school", "sm college", "c
```

When I observe the data file, I find out many same values. So I make a histogram to see data distribution.



From this plot, I find two peak count concentrated on value 0 and 25564.67. This data `re78` may be censored data for some reasons. So there are two subsets: 1) 0 to 25564.67; 2) 25564.67 to the max value.

```
#split indicator
nsw$ind=ifelse(nsw$re78>=25564.669921875,1,0)
nsw$ind=factor(nsw$ind,labels=c("up","low"))

#regression model to predict ind
fit4=glm(ind~age+educ+re75+black+married,family=binomial(link="logit"),data=nsw)
#summary(fit4)
pre.ind=predict(fit4,nsw,type = "response")
real.ind=ifelse(nsw$re78>=25564.669921875, 1, 0)
error.rate=mean((pre.ind>0.5 & real.ind==0)|(pre.ind<0.5 & real.ind==1))
print(error.rate)
```

```
## [1] 0.1587829
```

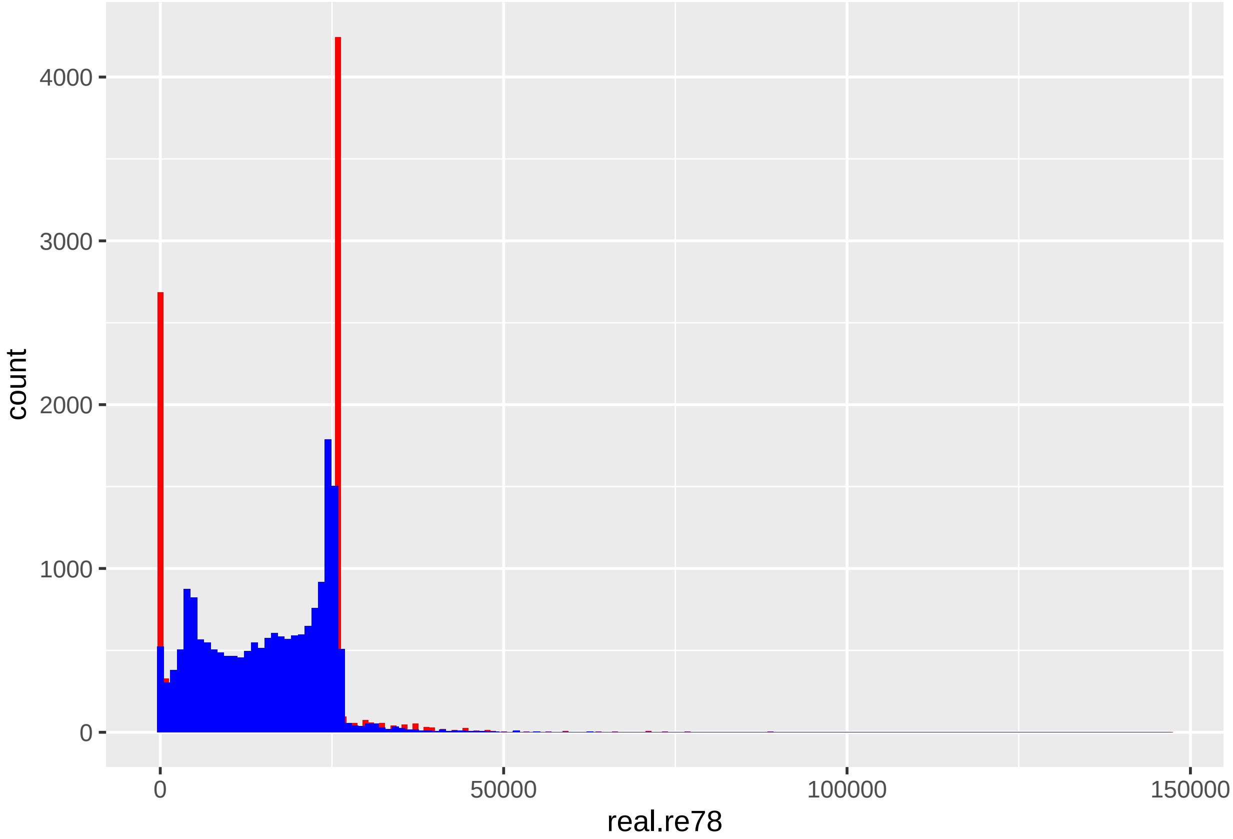
The error rate of the prediction of indicator is 15.88%, which means this ind regression model `fit4` fit well.So I will use predicted indicator `pre.ind` to replace the `real.ind`.

```
library(VGAM)
#create two subsets
nsw$pre.ind=pre.ind
sub1=nsw[nsw$pre.ind<0.5, ]
sub2=nsw[nsw$pre.ind>=0.5, ]

# two tobit regression
fit5_1=vglm(re78 ~ age + educ + re75,tobit(Lower=0, Upper=25563),data=sub1)
#summary(fit5_1)
pre_1=predict(fit5_1,nsw)

fit5_2=vglm(re78 ~ age + educ + re75,tobit(Lower=25564, Upper=max(nsw$re78)),data=sub2)
#summary(fit5_2)
pre_2=predict(fit5_2,nsw)

pre.re78=ifelse(pre.ind<0.5,
                ifelse(pre_1>0,pre_1,0),
                pre_2)
real.re78=nsw$re78
y=data.frame(cbind(real.re78,pre.re78))
ggplot(data=y)+geom_histogram(aes(x=real.re78),fill="red",binwidth = (range(real.re78)[2]-range(real.re
```



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.

```r
#logistic
nsw$earningind1=ifelse(nsw$re78==0,0,1)
fit9=glm(earningind1~age+educ+re75+married+black,data=nsw,family = binomial(link = "logit"))
display(fit9)
```

```
## glm(formula = earningind1 ~ age + educ + re75 + married + black,
##     family = binomial(link = "logit"), data = nsw)
##             coef.est coef.se
## (Intercept)  3.25     0.13
## age         -0.05     0.00
## educ        -0.08     0.01
## re75         0.00     0.00
## married1    -0.26     0.06
## black1      -0.02     0.07
## ---
##   n = 18667, k = 6
##   residual deviance = 11778.6, null deviance = 14712.7 (difference = 2934.1)
```

```r
#linear regression
nsw=filter(nsw,re78!=0)
nsw$level=ifelse(nsw$re78>=25564,1,0)
fit10=glm(level~age+educ+re75+married+black,data=nsw)
display(fit10)
```

```
## glm(formula = level ~ age + educ + re75 + married + black, data = nsw)
##             coef.est coef.se
## (Intercept) -0.31     0.02
## age          0.00     0.00
## educ         0.02     0.00
## re75         0.00     0.00
## married1     0.02     0.01
## black1      -0.05     0.01
## ---
##   n = 16164, k = 6
##   residual deviance = 2385.3, null deviance = 3473.8 (difference = 1088.5)
##   overdispersion parameter = 0.1
##   residual sd is sqrt(overdispersion) = 0.38
```

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.

Inc means the incumbency status in district i in election t, coded as 1 for Democratic incumbents, 0 for open seats, −1 for Republican incumbents)

```r
congress = read.csv("congress.csv")
congress88 <- data.frame(vote=congress$v88_adj,pastvote=congress$v86_adj,inc=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.

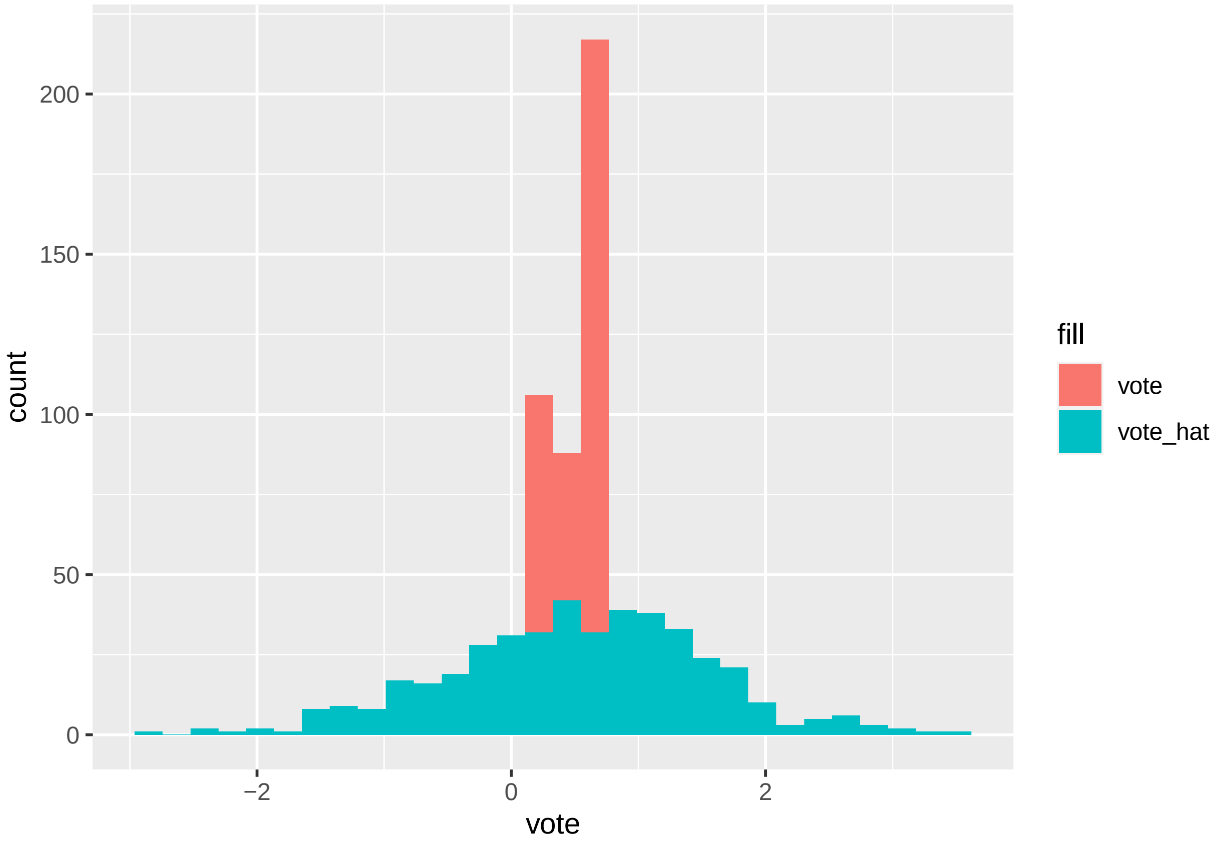
```r
congress88$inc=factor(congress88$inc)
fit6=stan_glm(vote~pastvote+inc,data=congress88,refresh=0)
print(fit6)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      vote ~ pastvote + inc
##  observations: 435
##  predictors:   4
## ------
##             Median MAD_SD
## (Intercept) 0.1    0.0
## pastvote    0.5    0.0
## inc0        0.1    0.0
## inc1        0.2    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.1    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```r
#predict
congress88$vote_hat=coef(fit6)[1]+coef(fit6)[2]*congress88$pastvote+coef(fit6)[3]*ifelse(congress88$inc

#predict vs real plot
ggplot(data=congress88)+geom_histogram(aes(x=vote,fill="vote"))+geom_histogram(aes(x=vote_hat,fill="vot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#error
error_vote=mean(congress88$vote-congress88$vote_hat)
print(error_vote)
```

```
## [1] 0.02265284
```

This model fit well since the mad_sd is low and the error of 0.08485 is small.

(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.

```
library(brms)

fit7=brm(vote~pastvote+inc,data=congress88,family = student,refresh=0)
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## clang -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I"/Library/Frameworks/R.frame
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
## ^
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
##                ^
##                 ;
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/Core:96:10: f
## #include <complex>
##          ^~~~~~~~~
## 3 errors generated.
## make: *** [foo.o] Error 1
```

```
summary(fit7)
```

```
##  Family: student
##   Links: mu = identity; sigma = identity; nu = identity
## Formula: vote ~ pastvote + inc
##    Data: congress88 (Number of observations: 435)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.13      0.01     0.10     0.15 1.00     2748     2876
## pastvote      0.55      0.04     0.48     0.62 1.00     2354     2032
## inc0          0.11      0.02     0.08     0.15 1.00     2883     2725
## inc1          0.19      0.02     0.16     0.22 1.00     2406     1950
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.05      0.00     0.05     0.06 1.00     2680     2226
## nu        6.30      2.57     3.42    12.28 1.00     2775     2372
##
## Samples were drawn 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).
```

```
#predict
congress88$vote_hat_brm=predict(fit7)[1]

#error
error_vote=mean(congress88$vote_hat_brm-congress88$vote)
print(error_vote)
```

```
## [1] 0.1852253
```

(c)

Which model do you prefer?

Comparing the error rate, there are not too much difference between these two model.

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.

```
congress88$win=ifelse(congress88$vote>0.5,1,0)
fit159_a=glm(win~pastvote+inc,data = congress88,family = binomial(link = "probit"))
display(fit159_a)
```

```
## glm(formula = win ~ pastvote + inc, family = binomial(link = "probit"),
##     data = congress88)
##             coef.est coef.se
## (Intercept) -4.22     0.58
## pastvote     5.76     1.25
## inc0         1.15     0.39
## inc1         2.84     0.43
## ---
##   n = 435, k = 4
##   residual deviance = 74.6, null deviance = 587.1 (difference = 512.5)
```

(b)

Fit a robit regression and assess model fit.

```
fit159_b=glm(win~pastvote+inc,data = congress88,family = binomial(link = gosset(2)))
display(fit159_b)
```

(c)

Which model do you prefer?

Comparing their null deviance and residual deviance, there are not too much difference between these two model. are close to each other.

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.

```
fita <- glm(fupacts ~ women_alone+sex+couples+bs_hiv, family=poisson, data=risk)

set.seed(12)
women_alone=factor(sample(c(0,1),1000,replace = T))
bs_hiv=sample(c("positive","negative"),1000,replace = T)
sex=sample(c("woman","man"),1000,replace = T)
couples=factor(sample(c(1,0),1000,replace = T))
newdata=data.frame(women_alone,sex,couples,bs_hiv)

y=predict(fita,newdata)
equal0=(sum(exp(y)==0)/1000)
great10=(sum(exp(y)>10)/1000)
```

```
equal0
```

```
## [1] 0
```

```
great10
```

```
## [1] 0.701
```

(b)

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

```
fitb=glm.nb(fupacts~women_alone+sex+couples+bs_hiv,data=risk)

set.seed(12)
women_alone=factor(sample(c(0,1),1000,replace = T))
bs_hiv=sample(c("positive","negative"),1000,replace = T)
sex=sample(c("woman","man"),1000,replace = T)
couples=factor(sample(c(1,0),1000,replace = T))
newdata=data.frame(women_alone,sex,couples,bs_hiv)

y=predict(fitb,newdata)
equal0=(sum(exp(y)==0)/1000)
great10=(sum(exp(y)>10)/1000)

equal0
```

```
## [1] 0
```

```
great10
```

```
## [1] 0.701
```

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

```
fitc = glm.nb(fupacts ~ women_alone+sex+couples+bs_hiv+bupacts, data=risk)
newdata$bupacts=sample(c(0,max(risk$bupacts)),1000,replace=T)
y=-predict(fitc,newdata)

y=predict(fitc,newdata)
equal0=(sum(exp(y)==0)/1000)
great10=(sum(exp(y)>10)/1000)

equal0
```

```
## [1] 0
```

```
great10
```

```
## [1] 0.549
```