STATS 201B Project

Predicting number of publications by biochemistry PhD students

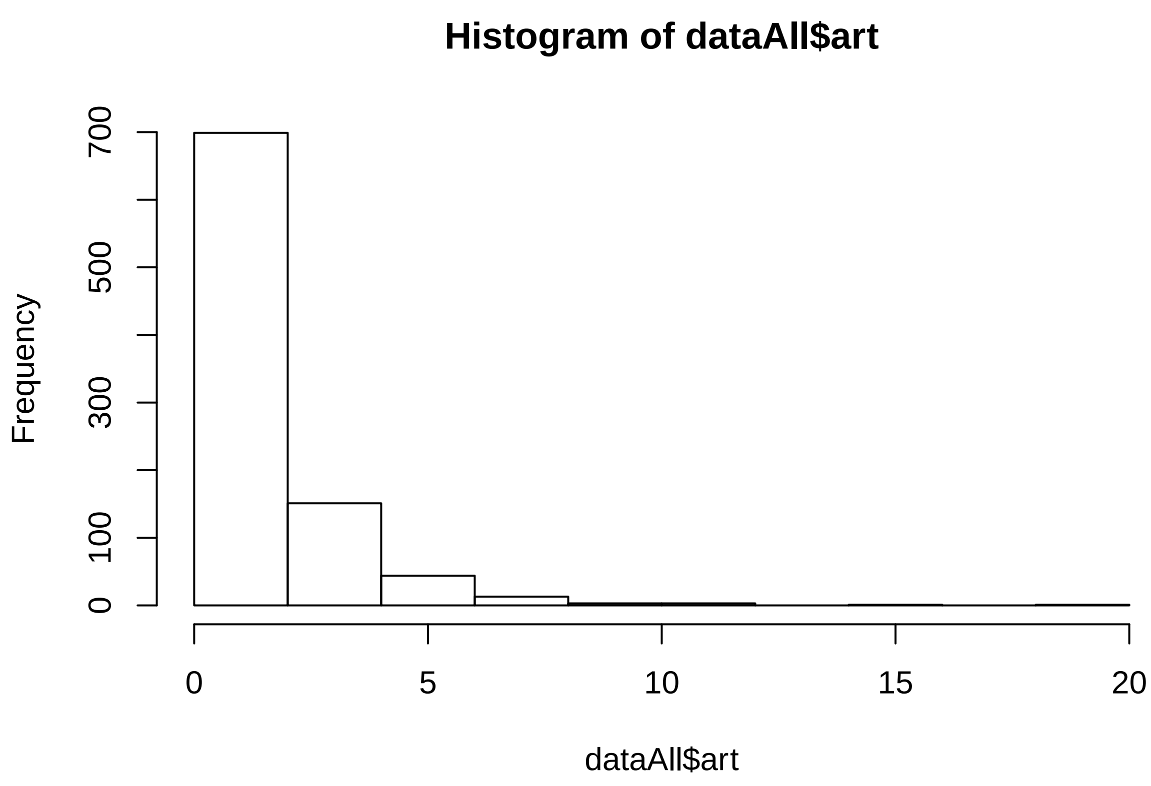
Shuchi Goyal, Suoyi Yang, Heather Zhou

2/23/2019

```r
library(pscl) #for the bioChemists data set
library(MASS) #for NB regression
library(countreg) #for hurdle and ZI

set.seed(1)
```

```r
set.seed(1)
#Load data set
dataAll<-bioChemists
?bioChemists
nTotal<-dim(dataAll)[1] #915 observations total
hist(dataAll$art) #about 700 students had 0 publications
```




```r
#Set the reference levels
dataAll$fem<-relevel(dataAll$fem,ref="Men")
dataAll$mar<-relevel(dataAll$mar,ref="Single")

#Split the data into training and testing
#Will not look at the testing data until shortly before the presentation
trainingInd<-sample(1:nTotal,size=615)
dataTrain<-dataAll[trainingInd,]
```

```r
dataTest<-dataAll[-trainingInd,]

xTrain <- dataTrain[,-1]
yTrain<-dataTrain$art
```

Part 1, GLM

```r
#Poisson regression
modPoi<-glm(art~.,family=poisson,data=dataTrain)
summary(modPoi)
```

```
##
## Call:
## glm(formula = art ~ ., family = poisson, data = dataTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3231  -1.5015  -0.3526   0.5520   5.5300
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.113945   0.129048   0.883  0.37725
## femWomen    -0.218555   0.067290  -3.248  0.00116 **
## marMarried   0.249098   0.077815   3.201  0.00137 **
## kid5        -0.249715   0.048895  -5.107 3.27e-07 ***
## phd          0.049859   0.032465   1.536  0.12460
## ment         0.025445   0.002364  10.763  < 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: 1203.5  on 614  degrees of freedom
## Residual deviance: 1060.1  on 609  degrees of freedom
## AIC: 2172.9
##
## Number of Fisher Scoring iterations: 5
```

```r
logLikPoi<-logLik(modPoi)[1] #-1080.441 on 6 Df. p=6, no additional parameters
print(logLikPoi)
```

```
## [1] -1080.441
```

```r
yhatPoi <- predict(modPoi,newdata = xTrain)
MSEtPoi<-mean((yTrain-yhatPoi)^2)
print(MSEtPoi)
```

```
## [1] 4.768253
```

```r
#Negative binomial regression
modNB<-glm.nb(art~.,data=dataTrain)
summary(modNB)
```

```
##
```

```
## Call:
## glm.nb(formula = art ~ ., data = dataTrain, init.theta = 2.420262947,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1139  -1.3480  -0.2732   0.4390   3.1239
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.049412   0.169173   0.292 0.770224
## femWomen    -0.207823   0.087893  -2.365 0.018054 *
## marMarried   0.242771   0.101686   2.387 0.016965 *
## kid5        -0.240235   0.062948  -3.816 0.000135 ***
## phd          0.058172   0.043480   1.338 0.180921
## ment         0.028621   0.003682   7.773 7.66e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.4203) family taken to be 1)
##
##     Null deviance: 756.05  on 614  degrees of freedom
## Residual deviance: 670.78  on 609  degrees of freedom
## AIC: 2065.6
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.420
##           Std. Err.:  0.370
##
##  2 x log-likelihood:  -2051.616
```

```r
logLikNB<-logLik(modNB)[1] #-1025.808 on 7 Df. p=6, plus there is psi
print(logLikNB)
```

```
## [1] -1025.808
```

```r
yhatNB <- predict(modNB,newdata = xTrain)
MSEtNB<-mean((yTrain-yhatNB)^2)
print(MSEtNB)
```

```
## [1] 4.761233
```

```r
#Hurdle Poisson
mod_H_Poi<-hurdle(art~.|1, data=dataTrain,dist="poisson")
summary(mod_H_Poi) #-1077.468 on 7 Df. p=6, plus there is pi
```

```
##
## Call:
## hurdle(formula = art ~ . | 1, data = dataTrain, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.1253 -1.0696 -0.3006  0.5217  6.7443
##
```

```
## Count model coefficients (truncated poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.497123   0.156007   3.187  0.00144 **
## femWomen    -0.252330   0.081290  -3.104  0.00191 **
## marMarried   0.172008   0.093413   1.841  0.06557 .
## kid5        -0.249675   0.061102  -4.086 4.39e-05 ***
## phd          0.026021   0.038926   0.668  0.50384
## ment         0.019820   0.002706   7.323 2.42e-13 ***
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8127     0.0874   9.299   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -1077 on 7 Df
```

```r
logLik_H_Poi<-logLik(mod_H_Poi)[1]
print(logLik_H_Poi)
```

```
## [1] -1077.468
```

```r
yhatmod_H_Poi <- predict(mod_H_Poi,newdata = xTrain)
MSEt_H_Poi<-mean((yTrain-yhatmod_H_Poi)^2)
print(MSEt_H_Poi)
```

```
## [1] 3.317057
```

```r
#Hurdle NB
mod_H_NB<-hurdle(art~.|1, data=dataTrain,dist="negbin")
summary(mod_H_NB) #-1043.847 on 8 Df. p=6, plus there is psi and pi
```

```
##
## Call:
## hurdle(formula = art ~ . | 1, data = dataTrain, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.0558 -0.9381 -0.2657  0.4653  5.6817
##
## Count model coefficients (truncated negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.138751   0.249258   0.557  0.57776
## femWomen    -0.264171   0.118585  -2.228  0.02590 *
## marMarried   0.189091   0.138180   1.368  0.17118
## kid5        -0.272276   0.088042  -3.093  0.00198 **
## phd          0.048532   0.059369   0.817  0.41367
## ment         0.024885   0.004964   5.013 5.35e-07 ***
## Log(theta)   0.646487   0.281114   2.300  0.02146 *
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8127     0.0874   9.299   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta: count = 1.9088
```

```
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -1044 on 8 Df
```

```r
logLik_H_NB<-logLik(mod_H_NB)[1]
print(logLik_H_NB)
```

```
## [1] -1043.847
```

```r
yhatmod_H_NB <- predict(mod_H_NB,newdata = xTrain)
MSEt_H_NB<-mean((yTrain-yhatmod_H_NB)^2)
print(MSEt_H_NB)
```

```
## [1] 3.311042
```

```r
#Zero-inflated Poisson
mod_ZI_Poi<-zeroinfl(art~.|1, data=dataTrain,dist="poisson")
summary(mod_ZI_Poi) #-1063.258 on 7 Df. p=6, plus there is pi
```

```
##
## Call:
## zeroinfl(formula = art ~ . | 1, data = dataTrain, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.5307 -0.9789 -0.2921  0.5333  6.9213
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.354025   0.142755   2.480  0.01314 *
## femWomen    -0.236946   0.072061  -3.288  0.00101 **
## marMarried   0.218083   0.084207   2.590  0.00960 **
## kid5        -0.250927   0.051930  -4.832 1.35e-06 ***
## phd          0.042350   0.034885   1.214  0.22476
## ment         0.022205   0.002539   8.746  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7643     0.2065  -8.542   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -1063 on 7 Df
```

```r
logLik_ZI_Poi<-logLik(mod_ZI_Poi)[1]
print(logLik_ZI_Poi)
```

```
## [1] -1063.258
```

```r
yhatmod_ZI_Poi <- predict(mod_ZI_Poi,newdata = xTrain)
MSEt_ZI_Poi<-mean((yTrain-yhatmod_ZI_Poi)^2)
print(MSEt_ZI_Poi)
```

```
## [1] 3.244258
```

```r
#Zero-inflated NB
mod_ZI_NB<-zeroinfl(art~.|1, data=dataTrain,dist="negbin")
summary(mod_ZI_NB) #-1025.808 on 8 Df. p=6, plus there is psi and pi
```

```
##
## Call:
## zeroinfl(formula = art ~ . | 1, data = dataTrain, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2709 -0.8738 -0.2552  0.4853  5.5584
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.049413   0.170934   0.289 0.772524
## femWomen    -0.207823   0.087697  -2.370 0.017799 *
## marMarried   0.242772   0.101874   2.383 0.017169 *
## kid5        -0.240234   0.063021  -3.812 0.000138 ***
## phd          0.058172   0.043727   1.330 0.183405
## ment         0.028622   0.003943   7.259 3.91e-13 ***
## Log(theta)   0.883877   0.153200   5.769 7.95e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -12.25      81.00  -0.151     0.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 2.4203
## Number of iterations in BFGS optimization: 35
## Log-likelihood: -1026 on 8 Df
```

```r
logLik_ZI_NB<-logLik(mod_ZI_NB)[1]
print(logLik_ZI_NB)
```

```
## [1] -1025.808
```

```r
yhatmod_ZI_NB <- predict(mod_ZI_NB,newdata = xTrain)
MSEt_ZI_NB<-mean((yTrain-yhatmod_ZI_NB)^2)
print(MSEt_ZI_NB)
```

```
## [1] 3.336789
```

To do:

We will compare the 6 models (Poisson, NB, Hurdle Poisson, Hurdle NB, zero-inflated Poisson, zero-inflated NB) and pick one.

TESTING RESULTS:

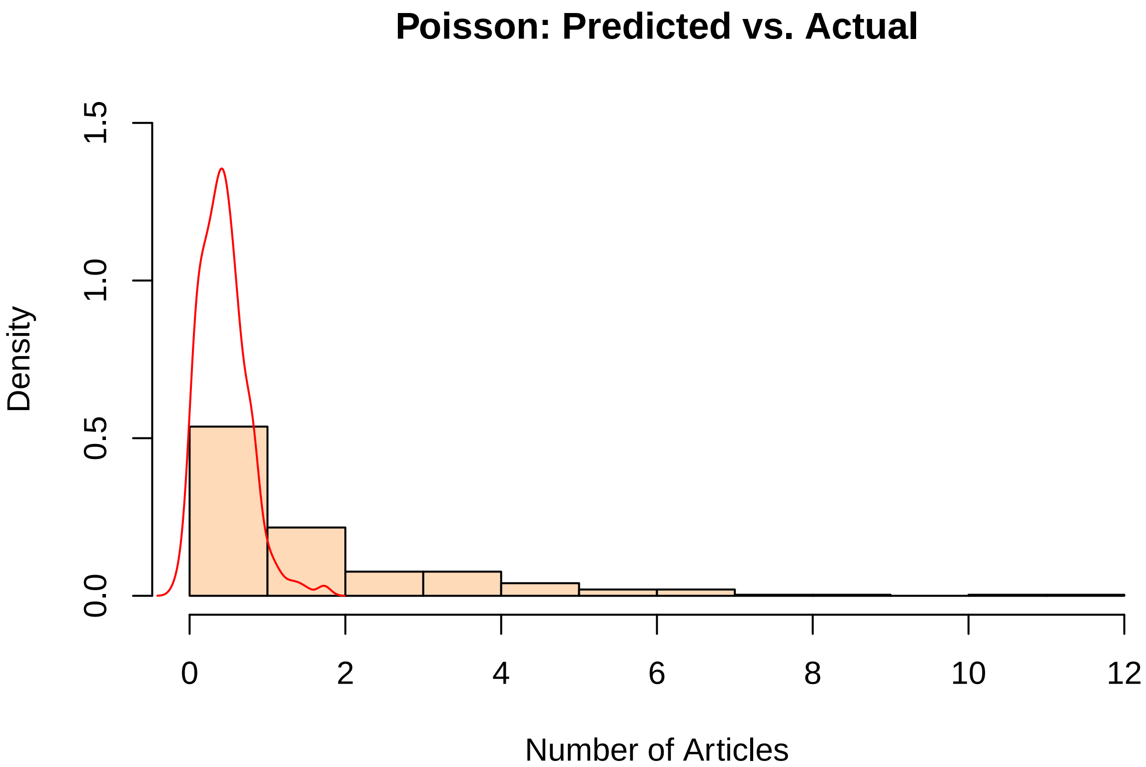
```r
xTest <- dataTest[,-1]
yTest <- dataTest[,1]
```

Poisson regression TEST MLE

```
yhatTestPoi <- predict(modPoi,newdata = as.data.frame(xTest))
MSETestPoi<-mean((yTest-yhatTestPoi)^2)
print(sqrt(MSETestPoi))
```

```
## [1] 2.343024
```

```
hist(yTest,freq = F, ylim = c(0,1.5),col="peachpuff", main = "Poisson: Predicted vs. Actual", xlab = "Nu
lines(density(yhatTestPoi), col = "red")
```

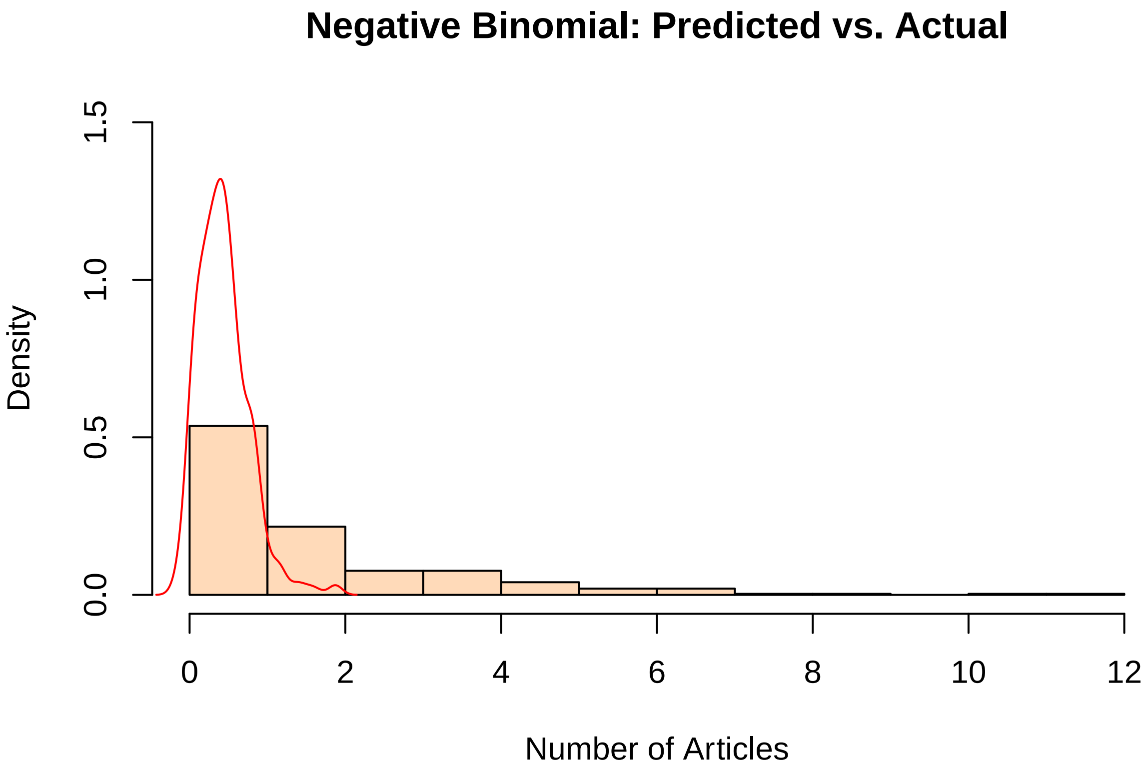


Negative binomial regression TEST MLE

```
yhatTestNB <- predict(modNB,newdata = as.data.frame(xTest))
MSETestNB<-mean((yTest-yhatTestNB )^2)
print(sqrt(MSETestNB))
```

```
## [1] 2.343141
```

```
hist(yTest,freq = F, ylim = c(0,1.5),col="peachpuff", main = "Negative Binomial: Predicted vs. Actual",
lines(density(yhatTestNB ), col = "red")
```

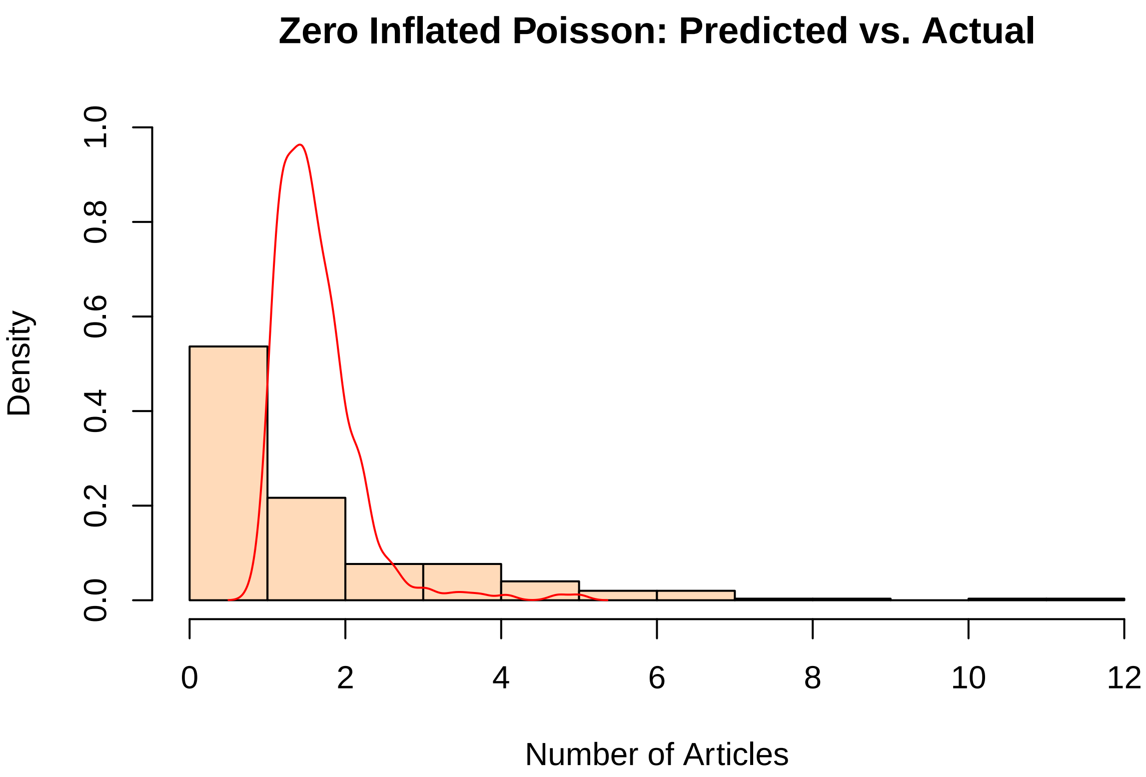


Zero-inflated Poisson TEST MLE

```
yhatTest_ZI_Poi <- predict(mod_ZI_Poi,newdata = as.data.frame(xTest))
MSETest_ZI_Poi<-mean((yTest-yhatTest_ZI_Poi)^2)
print(sqrt(MSETest_ZI_Poi))
```

```
## [1] 1.901107
```

```
hist(yTest,freq = F, ylim = c(0,1),col="peachpuff",main = "Zero Inflated Poisson: Predicted vs. Actual"
lines(density(yhatTest_ZI_Poi), col = "red")
```

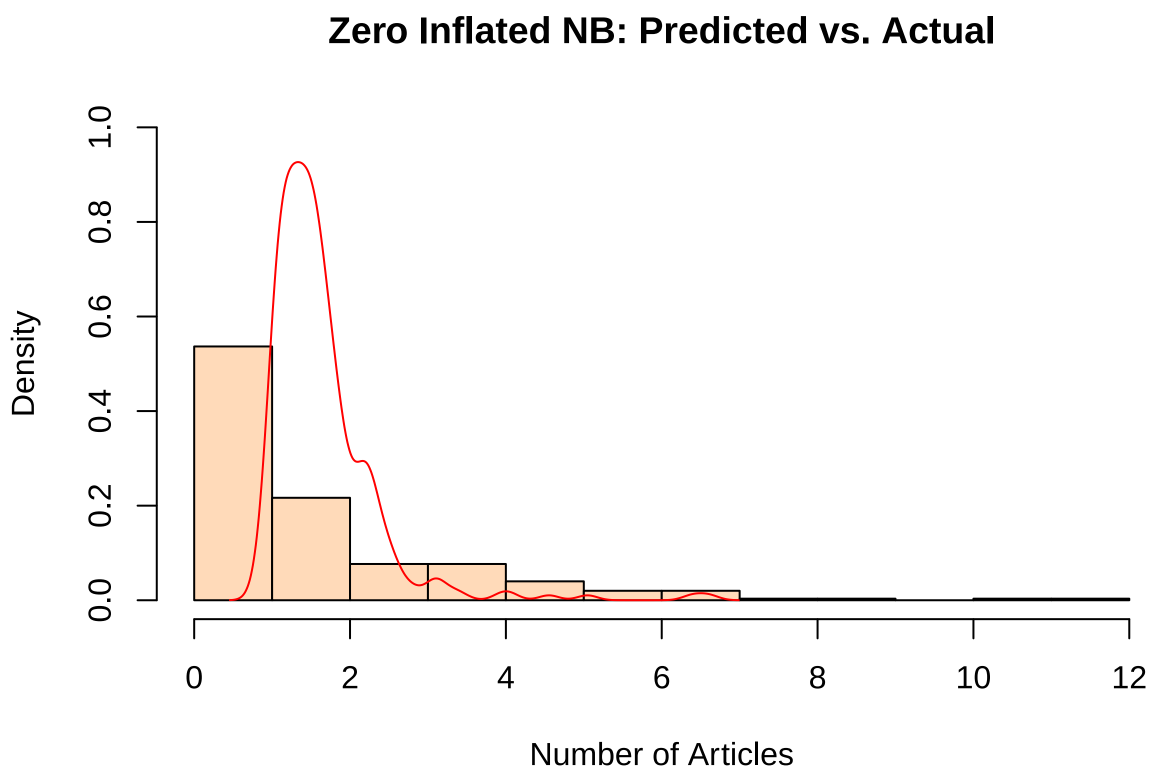


Zero-inflated NB TEST MLE

```
yhatTest_ZI_NB <- predict(mod_ZI_NB,newdata = as.data.frame(xTest))
MSETest_ZI_NB<-mean((yTest-yhatTest_ZI_NB)^2)
print(sqrt(MSETest_ZI_NB))
```

```
## [1] 1.909785
```

```
hist(yTest,freq = F, ylim = c(0,1),col="peachpuff", main = "Zero Inflated NB: Predicted vs. Actual", xla
lines(density(yhatTest_ZI_NB), col = "red")
```

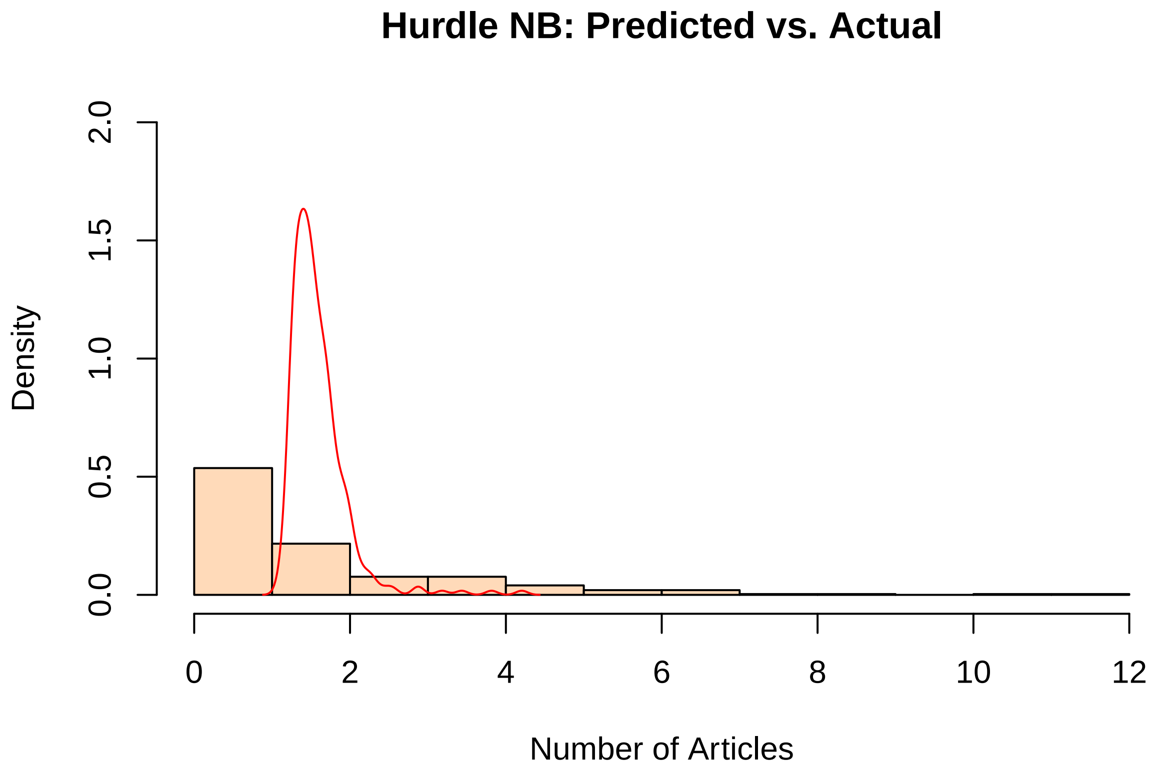


Hurdle NB

```
yhatTest_H_NB <- predict(mod_H_NB,newdata = as.data.frame(xTest))
MSETest_H_NB<-mean((yTest-yhatTest_H_NB)^2)
print(sqrt(MSETest_H_NB))
```

```
## [1] 1.90768
```

```
hist(yTest,freq = F, ylim = c(0,2),col="peachpuff", main = "Hurdle NB: Predicted vs. Actual", xlab = "Nu
lines(density(yhatTest_H_NB), col = "red")
```


Hurdle NB: Predicted vs. Actual
Density
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10
12
Number of Articles