GLM Odds and Sods

STATS 762 – Lecture Slides 8

07.04.2019

Negative Binomial Regression

One way to deal with count data that is overdispersed (relative to a Poisson model) is to use the negative binomial distribution.

- The probability density function (pdf) of the negative binomial distribution can be written as:

$$\Pr(X = x) = \frac{\Gamma(x+\theta)}{x!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{x} \quad \text{for } x = 0,\ 1,\ 2,\ 3,\ \ldots$$

where $\mu > 0$ and $\theta > 0$.

Properties of the Negative Binomial Distribution

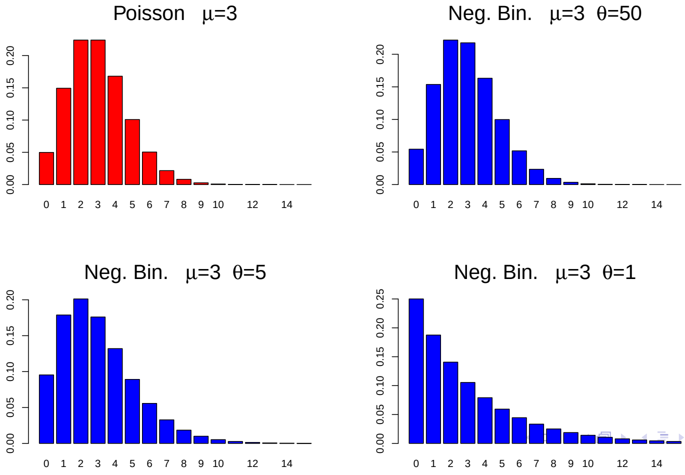
The mean and variance are given by:

$$\mathsf{E}(X) = \mu \quad \text{and} \quad \mathsf{Var}(X) = \mu + \frac{\mu^2}{\theta}$$

- The variance is always greater than the mean but approaches the mean as θ goes to infinity.
- In fact, as θ goes to infinity the negative binomial distribution goes to the Poisson distribution.

Negative Binomial compared with Poisson

Plots of the pdf's of the Poisson distribution ($\mu = 3$) and negative binomial distributions ($\mu = 3$; $\theta = 50$, 5 and 1).



Negative Binomial Regression using *R*

The negative binomial is a member of the exponential family only if θ is fixed.

- However we will almost always wish to estimate θ from the data. As a consequence we cannot use the `glm` function in *R* to fit a negative binomial regression model.
- The `MASS` package contains a function called `glm.nb` which fits negative binomial regression models.
  - This function works by fixing θ and fitting the model using the same algorithm as used in `glm`. An updated value of θ is obtained using information from the fitted model. The process is repeated until convergence.

Crab Example Revisited

Recall the female horseshoe crab example.

weight: Weight of the crab (grams),

colour: colour of the crab (1=light medium , 2=medium, 3=dark medium, 4=dark)

satellites: number of satellite crabs

First, we will revisit the analyses we did before using the Poisson and quasi-Poisson models.

The Poisson Model

```
> crab.glm=glm(sats˜weight+colour,family=poisson,data=crab.df)
> summary(crab.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.04961    0.23311  -0.213   0.8315
weight       0.54608    0.06809   8.020 1.06e-15 ***
colour2     -0.20508    0.15371  -1.334   0.1821
colour3     -0.44966    0.17574  -2.559   0.0105 *
colour4     -0.45228    0.20843  -2.170   0.0300 *
---
    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 551.78  on 168  degrees of freedom
AIC: 917.08

> anova(crab.glm,test="Chisq")
       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84  < 2e-16 ***
colour  3    9.062       168     551.78  0.02848 *

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
colour  3   23.653       169     609.14 2.952e-05 ***
weight  1   57.358       168     551.78 3.633e-14 ***
```

The Quasipoisson Model

```
> crabB.glm=glm(sats˜weight+colour,family=quasipoisson,data=crab.df)
> summary(crabB.glm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04961    0.41593  -0.119    0.905
weight       0.54608    0.12148   4.495 1.29e-05 ***
colour2     -0.20508    0.27426  -0.748    0.456
colour3     -0.44966    0.31356  -1.434    0.153
colour4     -0.45228    0.37189  -1.216    0.226
---
(Dispersion parameter for quasipoisson taken to be 3.183475)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
weight  1   71.949       171     560.84 1.994e-06 ***
colour  3    9.062       168     551.78    0.4159


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     632.79
colour  3   23.653       169     609.14   0.05939 .
weight  1   57.358       168     551.78 2.189e-05 ***
```

Summary

The quasi-Poisson model gave clear evidence of overdispersion.

- For the data the variance appears to be greater than the mean and thus the Poisson model is not appropriate.

- The quasi-Poisson model compensates by estimating a "fudge factor" to adjust for overdispersion. In this case, it was estimated that the observed variance was about 3.2 times what the Poisson model indicated.

- The quasi-Poisson model adjusted the estimated standard errors for the fitted coefficients and the p-values for the chi-squared test produced by `anova` using this fudge factor.

Fitting the Negative Binomial Model

To fit the negative binomial model for the crab data using *R*:

```
library(MASS)
crabNB.glm=glm.nb(sats ~ weight + colour,data = crab.df)
```

- Note that the `MASS` library needs to be loaded.
- The `glm.nb` function has the same form as the `glm` function – the only difference is we don't need to use `family` to specify a distribution.
- The default is to use the log link (same as for Poisson regression). The link can also be set to the identity, `link = identity` or to the square root, `link = sqrt`, if desired.

The Fitted Model

The output from `summary` is similar to that for GLM's.

```
> summary(crabNB.glm)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4253     0.5380  -0.790    0.429
weight        0.7117     0.1614   4.409 1.04e-05 ***
colour2      -0.2527     0.3486  -0.725    0.468
colour3      -0.5217     0.3799  -1.373    0.170
colour4      -0.4807     0.4282  -1.123    0.262

(Dispersion parameter for Negative Binomial(0.9597) family taken to be 1)
    Null deviance: 220.02  on 172  degrees of freedom
Residual deviance: 196.57  on 168  degrees of freedom
AIC: 757.93
              Theta:  0.960
          Std. Err.:  0.175
 2 x log-likelihood:  -745.934
```

The Fitted Model (cont.)

The output from anova is also similar to that for GLM's.

```
> anova(crabNB.glm,test="Chisq")
Analysis of Deviance Table
Model: Negative Binomial(0.9597), link: log
Response: sats
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     220.02
weight  1  20.7135       171     199.30 5.334e-06 ***
colour  3   2.7373       168     196.56    0.4339
---
Warning message:
In anova.negbin(crabNB.glm, test = "Chisq") :
  tests made without re-estimating 'theta'
```

- Note that we get a similar result to the corresponding test using the quasi-Poisson model.

Reduced Model

Model that just has `weight` as a regressor:

```
> crabNB2.glm=glm.nb(sats ~ weight, data = crab.df)
> summary(crabNB2.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8637     0.4046  -2.135   0.0328 *
weight        0.7599     0.1578   4.817 1.46e-06 ***
---
(Dispersion parameter for Negative Binomial(0.9311) family taken to be 1)

    Null deviance: 216.44  on 172  degrees of freedom
Residual deviance: 196.16  on 171  degrees of freedom
AIC: 754.64
              Theta:  0.931
          Std. Err.:  0.168
 2 x log-likelihood:  -748.643
```

- Note that the estimate for θ has changed and as a result the value of the "Null deviance" has also changed.

Lack of Fit Test

We can also do a lack of fit test using the residual deviance.

```
> 1-pchisq(196.16,171)
[1] 0.09100232
```

- This suggests weak evidence of lack of fit.

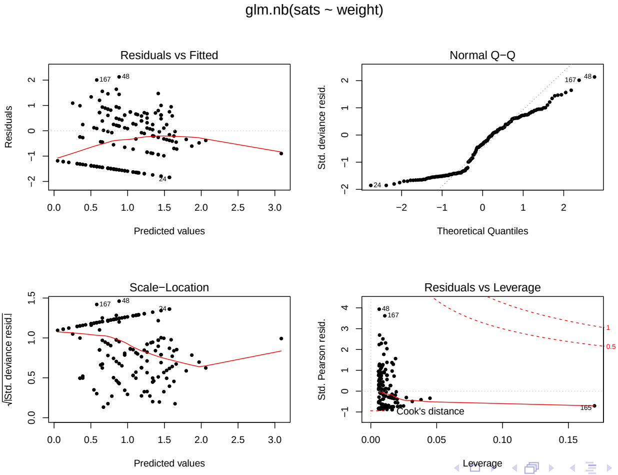
Compare this to the lack of fit test for the Poisson model.

```
> 1-pchisq(551.78,168)
[1] 0
```

- Extremely strong evidence of lack of fit.

Diagnostic Plots

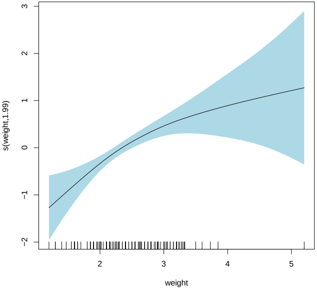
Standard diagnostic plots from `plot(crabNB2.glm)`:



A GAM plot

To get a GAM plot for a negative binomial model:

```
> library(mgcv)
> crab.gam=gam(sats~s(weight),family=nb,data=crab.df)
> plot(crab.gam,shade=TRUE,shade.col="lightblue")
```



- `family=nb` is used for a negative binomial where θ is estimated.

The Fitted Model

The GAM plot suggested that either adding a quadratic term for `weight` or using `log(weight)` might be useful. However, neither of these modifications results in a significant improvement to the model.

As we used the log link for the negative binomial regression, the fitted model can be written as:

$$\log(\widehat{\mu}) = -.86 + .76 \times \texttt{weight}$$

or

$$\widehat{\mu} = \exp(-.86 + .76 \times \texttt{weight})$$

Interpreting the Impact of `weight`

For our fitted model, we can describe the impact of weight in two ways:

1. For each increase of 1 unit in weight the log of the expected number of satellites increases by .76.

2. For each increase of 1 unit in weight the expected number of satellites is multiplied by $\exp(.76) = 2.14$.

Which is exactly how we would interpret a fitted Poisson regression model (or any fitted GLM that uses a log link).

Inference using a Negative Binomial Model

The same techniques (and the same commands in *R*) can be used to do inference using a negative binomial regression model as would be done for a GLM.

For, example to create a 95% CI for the mean number of satellites when `weight`=3.3:

```
> predict(crabNB2.glm,data.frame(weight=3.3),se.fit=TRUE)
$fit
1.643869

$se.fit
[1] 0.1564291

> vals=c(1.644-1.96*.156,1.644+1.96*.156)
> exp(vals)
[1] 3.812328 7.027001
```

Summary

- A negative binomial regression model is often useful for count data where the variability is more than can adequately modelled using a Poisson regression model (overdispersion).

- The negative binomial regression model is not a GLM if the θ parameter is to be estimated. In this case the `glm` function in *R* cannot be used.

- To fit a negative binomial regression model use the `glm.nb` function from the MASS package.

- Diagnostics, model building and inference all work the same as for a GLM.

Offsets

- Often count data are concerned with rates, such as deaths per 100,000 or accidents per million vehicle miles.
- Sometimes the counts in a data set may have been taken over different population sizes or different time periods etc. and the analysis needs to take this into account.
- One common approach is to use **offsets**.

Cancer Example

Deaths from childhood cancers 1951-1960 in Northumberland and Durham, classified by

- Cytology (Lymphoblastic/Myeloblastic)
- Residence (Rural/Urban)
- Age (0-5, 6-14)

Cancer Data

```
Cytology Residence   Age   n     pop
       L         R   0-5  38  103857
       L         R  6-14  13  155786
       L         U   0-5  51  135943
       L         U  6-14  37  203914
       M         R   0-5   5  103857
       M         R  6-14   8  155786
       M         U   0-5  13  135943
       M         U  6-14  20  203914
```

Rates

- For this data, the counts were taken for different population sizes. As a result, we expect counts to be higher when the value of pop is higher. We are really interested in how the rate of cancer deaths is related to `Cytology`, `Residence` and `Age`.

- Suppose we wish to model deaths per 100,000 population using a Poisson regression model. Let μ_x and $\mu_{100,000}$ be the means for a populations of size x and 100,000 respectively.

$$\mu_x = \frac{x}{100,000} \times \mu_{100,000}$$

Taking logs we get:

$$\log(\mu_x) = \log\left(\frac{x}{100,000}\right) + \log \mu_{100,000}$$

Offsets Definition

- The term $\log(x/100000)$ is called an offset. An offset is a regressor in a regression model whose coefficient is fixed at 1.
- By including this offset in a Poisson regression model for the cancer data, it means that the other regressors are now modelling $\log \mu_{100{,}000}$.

Fitting Poisson Model with Offset

```
> model1<-glm(n ~ Cytology*Age*Residence+offset(log(pop/100000)),
                                family=poisson, data=cancer.df)
> anova(model1,test="Chisq")
Analysis of Deviance Table
                       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                       7     92.452
Cytology                1   48.952         6     43.500 2.624e-12 ***
Age                     1   23.875         5     19.625 1.028e-06 ***
Residence               1    5.848         4     13.777  0.015597 *
Cytology:Age            1    8.717         3      5.060  0.003152 **
Cytology:Residence      1    1.110         2      3.950  0.292032
Age:Residence           1    2.895         1      1.054  0.088849 .
Cytology:Age:Residence  1    1.054         0      0.000  0.304491
```

- `Cytology:Age` is the only interaction term that is clearly significant.

Fitting the New Model

```
> model2<-glm(n ~ Cytology*Age + Residence, family=poisson,
    offset=log(pop/100000), data=cancer.df)
> summary(model2)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.3893     0.1465  23.139  < 2e-16 ***
CytologyM          -1.5983     0.2584  -6.184 6.24e-10 ***
Age6-14            -0.9821     0.1767  -5.557 2.75e-08 ***
ResidenceU          0.3677     0.1546   2.379  0.01736 *
CytologyM:Age6-14   1.0184     0.3500   2.910  0.00362 **

Null deviance: 92.4517  on 7  degrees of freedom
Residual deviance:  5.0598  on 3  degrees of freedom
AIC: 52.858
```

```
> 1-pchisq(5.0598, 3)
[1] 0.1674703
```

Interpreting the cytology/age interaction

For rural residence, fitted cancer rates per 100,000 population:

	Age = 0-5	Age=6:14
Cytology=L	exp(3.3893) =29.6	exp(3.3893-0.9821) =11.1
Cytology=M	exp(3.3893-1.598) =5.9	exp(3.3893-0.9821-1.5983 +1.0184) =6.2

- For urban residence, multiply each of these rates by exp(0.3677)=1.44

West Nile Virus Example

WNV is a potentially fatal disease that attacks the central nervous system.

- Birds are the most commonly affected animal but WNV also affects mammals including horses and humans.
- WNV is spread mainly by mosquitos with birds being the primary hosts.
- Typically bird mortality due to WNV precedes human and equine infection.
- Dead bird surveillance is used to monitor the risk to human and horse populations – people are asked to report cases of dead birds which are then collected and tested.

West Nile Virus Data

Roberts and Foppa (*Vector-Borne and Zoonotic Diseases*, 2006, **6**, 1–6) presented data collected for 46 counties in South Carolina in 2003 which included the following variables:

equine The number of cases of WNV in horses (response variable).

farms The number of farms in the county (offset variable).

pbr The positive bird rate (regressor of main interest).

density The population density recorded as people per square mile (regressor).

Notes about the Data

- The number of farms is being used as a surrogate for the number of horses which is much harder to determine.

- As population increases it is expected that more dead birds will be reported and thus the number of positive tests will be higher. To compensate, the “positive bird rate” (the number of dead birds that tested positive for WNV divided by the population of the county) is used as the regressor of main interest.

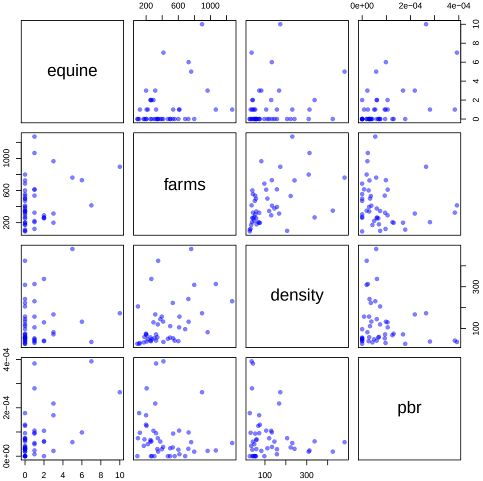
- Population density was included as it was thought to be a potential confounder and/or effect modifier of the association between dead bird counts and WNV risk.

A First Look at the Data

```
> head(wnv.df)
     county equine farms density       pbr
1 Abbeville      0   471   51.50 0.0000000
2     Aiken      6   729  132.86 0.0000982
3 Allendale      0   114   27.51 0.0001782
4  Anderson      1  1271  230.81 0.0000543
5   Bamberg      2   254   42.43 0.0000600
6  Barnwell      1   325   42.83 0.0003834
```

```
> summary(wnv.df[,-1])
     equine              farms              density               pbr
 Min.   : 0.000   Min.   :  92.0   Min.   : 27.51   Min.   :0.000e+00
 1st Qu.: 0.000   1st Qu.: 254.2   1st Qu.: 49.00   1st Qu.:2.225e-05
 Median : 0.000   Median : 348.0   Median : 74.20   Median :5.615e-05
 Mean   : 1.174   Mean   : 437.8   Mean   :123.13   Mean   :8.086e-05
 3rd Qu.: 1.000   3rd Qu.: 590.5   3rd Qu.:157.34   3rd Qu.:9.785e-05
 Max.   :10.000   Max.   :1271.0   Max.   :480.55   Max.   :3.923e-04
```

Pairs Plot


equine
farms
density
pbr
200
600
1000
0e+00
2e-04
4e-04
0
2
4
6
8
10
100
300

Poisson Regression Model

To model the rate of equine WNV per 100 farms.

```
> wnv.glm<-glm(equine~density*pbr+offset(log(farms/100)),
                                    family=poisson,data=wnv.df)
> summary(wnv.glm)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.377e+00  3.664e-01  -6.488 8.72e-11 ***
density      5.903e-04  1.673e-03   0.353 0.724188
pbr          5.330e+03  1.596e+03   3.340 0.000837 ***
density:pbr  2.600e+01  1.238e+01   2.100 0.035706 *
---
    Null deviance: 109.324  on 45  degrees of freedom
Residual deviance:  62.007  on 42  degrees of freedom
AIC: 122.54

> 1-pchisq(62.007,42)
[1] 0.02388447
```

Quasi-Poisson Model

Try the quasi-Poisson model.

```
> wnvA.glm<-glm(equine~density*pbr+offset(log(farms)),
                     family=quasipoisson,data=wnv.df)
> summary(wnvA.glm)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.982e+00  4.800e-01 -14.546   <2e-16 ***
density      5.903e-04  2.191e-03   0.269   0.7890
pbr          5.330e+03  2.090e+03   2.550   0.0145 *
density:pbr  2.600e+01  1.622e+01   1.603   0.1164
---
(Dispersion parameter for quasipoisson family taken to be 1.716101)

    Null deviance: 109.324  on 45  degrees of freedom
Residual deviance:  62.007  on 42  degrees of freedom
```

Negative Binomial Regression Model

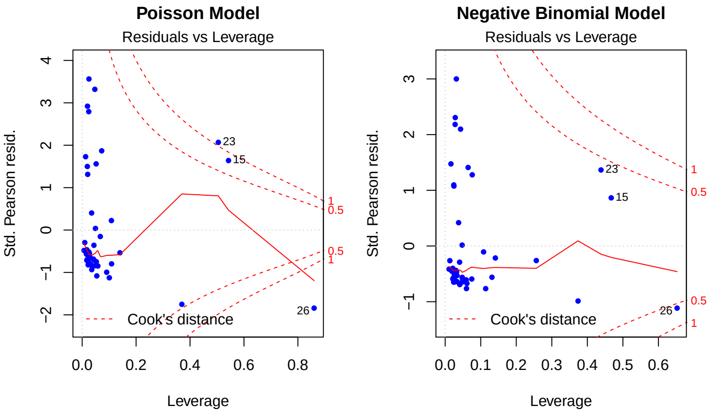
Offsets can also be used in negative binomial regression.

```
> wnvNB.glm<-glm.nb(equine˜density*pbr+offset(log(farms/100)),
                                         maxit=50,data=wnv.df)
> summary(wnvNB.glm)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.310e+00  4.341e-01  -5.322 1.03e-07 ***
density     -6.890e-04  2.299e-03  -0.300   0.7644
pbr          4.517e+03  2.403e+03   1.880   0.0602 .
density:pbr  4.464e+01  2.292e+01   1.948   0.0515 .
---
(Dispersion parameter for Negative Binomial(2.3909) family taken to be 1)

    Null deviance: 72.300  on 45  degrees of freedom
Residual deviance: 45.246  on 42  degrees of freedom
AIC: 121.82

> 1-pchisq(45.246,42)
[1] 0.3380468
```

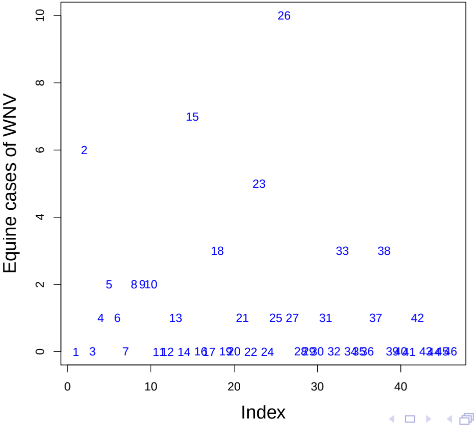
Residuals versus Leverage Plots



- The negative binomial model has reduced the influence of points 15, 23 and 26.

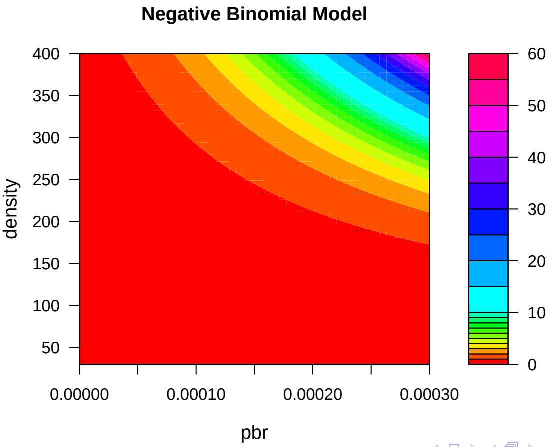
Points 15, 23 and 26

These observations have high values for `equine`:



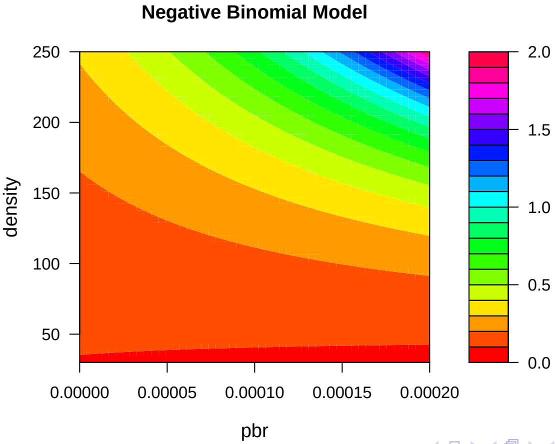
A Filled Contour Plot

A filled contour plots for the fitted negative binomial model.



Another Filled Contour Plot

A filled contour plots for the low values of density and `pbr`.


Negative Binomial Model
density
pbr