Homework 8

Gitanshu Munjal

December 13, 2014

Zero Inflated Poisson model

This exercise is based on the same scenario as HW07 but the data have been change to include zero inflation. The main purpose of the exercise is to model data with zero inflation. All the code is provided. You are asked to run and document each chunk of code as grouped below.

- For each chunk, write a comment describing the main purposes of the chunk. One to two sentences should be sufficient.
- Interpret the result of each chunk as appropriate. For example, when the code produces a residual plot you would write whether there is any violation of assumptions suggested by the plot or not. For the final plots, how is abundance of the rare mammal affected by river, road and cover?
- How does road affect the probability that the mammal found a particular location that might have suitable habitat? For this you need to interpret the zero inflation part of the model.

Chunk 1: Code and Relevant Output

```
setwd("C:\\Users\\uglysweaters\\Desktop")
d1 <- read.csv("zeroinfl.txt", header=TRUE)

library(car)
library(ggplot2)
head(d1)
```

```
  X river road      cover nrrm is.zero
1 1   6.7  7.1 0.56713729    6       1
2 2  10.5  8.2 0.04918115    3       1
3 3   9.9  2.1 0.03200358    0       0
4 4   8.8  5.5 0.01629375    0       0
5 5   4.0  1.3 0.32789126    0       0
6 6   4.7  4.9 0.11083271    1       1
```

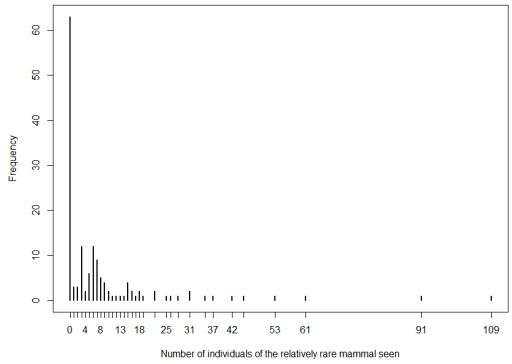
Chunk 1 sets the working directory, reads data (into "d1"), loads packages that we plan to use for our analysis, and looks at the top part of the data to make sure things were read correctly. We found that the first column contained redundant serial numbers so these were removed as follows:

```
d1 <- d1[2:6]
```

Chunk 2: Code and Relevant Output

```
par(mfrow=c(1,1))
plot(table(d1$nrrm),ylab="Frequency",xlab="Number of individuals of the relatively rare
 mammal seen")
```

Chunk 2 sets the graphic parameter "mfrow" to 1 row and 1 column since we have a single plot coming next. The second part of the chunk plots the contingency table for our count data.

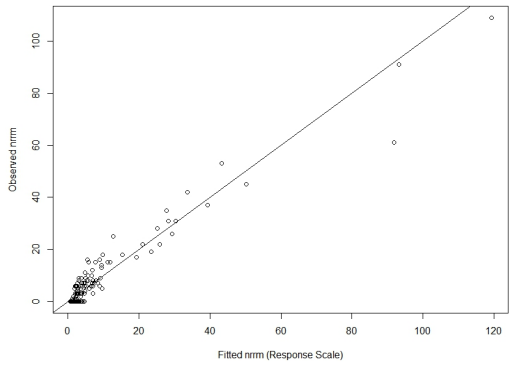


The above figure shows that our data possibly has zero inflation since the frequency of zero counts is much higher in comparison to the other counts.

Chunk 3: Code and Relevant Output

```
glm1 <- glm(nrrm ~ river + road + cover + river:road + river:cover + road:cover,
                        family=poisson, data=d1)
plot(d1$nrrm~fitted(glm1, type="response"),xlab="Fitted nrrm (Response Scale)",
         ylab="Observed nrrm")
abline(0,1)
```

The above code fits a glm model with a poisson error distribution to our data using nrrm as the response variable and accounting for all two way interactions between explanatory variables. The second part of the code produces a plot of actual observations for counts versus fitted values from our model.



The above plot shows that our current model does not do a very good job of explaining the data and needs improvement to account for the zero inflation which is leading to the clustering observed around (0,0) on the above plot.

```
summary(glm1)
```

```
Call:
glm(formula = nrrm ~ river + road + cover + river:road + river:cover +
    road:cover, family = poisson, data = d1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4400  -1.7052  -0.6972   0.8537   3.6625

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.68728    0.31990  -2.148  0.03168 *
river       -0.03273    0.03044  -1.075  0.28225
```

```
road          0.10379    0.02047   5.072 3.94e-07 ***
cover         1.25168    0.46870   2.671  0.00757 **
river:road    0.01531    0.00220   6.961 3.37e-12 ***
river:cover   0.01423    0.03622   0.393  0.69432
road:cover    0.05432    0.03229   1.682  0.09256 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2546.64  on 149  degrees of freedom
Residual deviance:  395.63  on 143  degrees of freedom
AIC: 755.57

Number of Fisher Scoring iterations: 5
```

```
anova(glm1, test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: nrrm

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          149    2546.64
river        1   351.94       148    2194.70 < 2.2e-16 ***
road         1  1390.02       147     804.68 < 2.2e-16 ***
cover        1   347.14       146     457.53 < 2.2e-16 ***
river:road   1    58.77       145     398.76 1.768e-14 ***
river:cover  1     0.28       144     398.48   0.59527
road:cover   1     2.85       143     395.63   0.09164 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above presented anova (using the chi-squared test) tells us about the significance of the overall model as explanatory terms are added sequentially (from NULL to road:cover) to the overall model. Important to note here is the decrease in residual deviance as terms are added indicating improved explanatory power at each step up until river:road. Subsequent, interaction terms do not reduce the residual deviance much indicating a need for removing these (one at a time) to free up some degrees of freedom.

```
1-pchisq(deviance(glm1),glm1$df.residual)
```

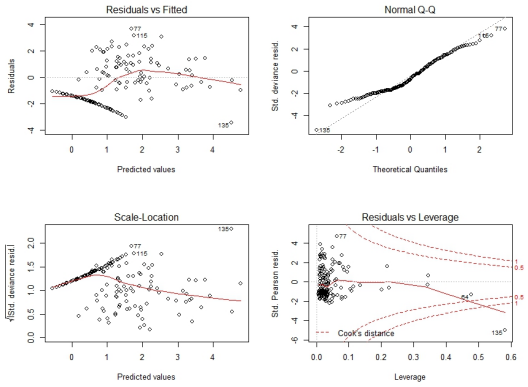
The above code tests the residual deviance from our current model against a chi-squared distribution with 143 degrees of freedom (residual deviance dF in summary above) to see if there is more deviance than expected at random.

```
[1] 0
```

The significant p-value (assuming a 95% confidence level) above tells us that indeed there is excessive deviance and we need to improve our model.

```
par(mfrow=c(2,2))
plot(glm1)
```

We ask the graphic device to plot the forthcoming results in 2 rows and 2 columns (so 4 plots on 1). Next, we plot graphs to test for the assumptions of our model.



The above graphs show that there are two points of concern for our current model: (1) the effect of zero inflation (seen as the streak of points on the residual vs. fitted graaphs, the cluster of black around y=-1 on the residual vs leverage plot, and the departure from normality at the lower end of the q-q plot); the residual value for all 0 observations was fixed at "-1" leading to the violations and departures from normality observed here. (2) three outliers to our assupmtions (outside of the zero value observations) were also identified but these are less of a concern in comparison to the zero-inflation.

Chunk 4: Code And Relevant Output

```
library(AER)
dispersiontest(glm1, trafo=1)
```

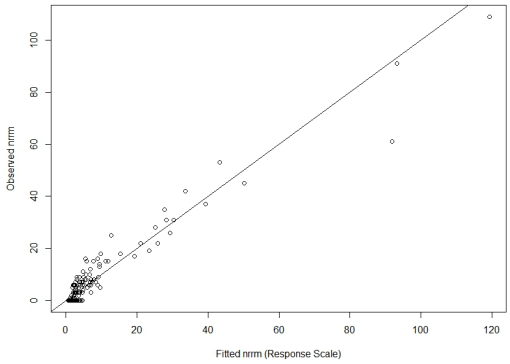
```
        Overdispersion test

data:  glm1
z = 6.5148, p-value = 3.641e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
1.364721
```

The above code and output tests our data for overdispersion and finds that indeed there is significant overdispersion at the 95% confidence level meaning that the observed variance is greater than the variance expected (=mean) on basis of our poisson assumption. We should proceed by using a quasi-GLM model where the variance is given by the mean multiplied by the dispersion parameter.

Chunk 5: Code and Relevant Output

```
glm2 <- glm(nrrm ~ river + road + cover + river:road + river:cover + road:cover,
                family=quasipoisson, data=d1)
par(mfrow=c(1,1))
plot(d1$nrrm~fitted(glm2, type="response"),xlab="Fitted nrrm (Response Scale)",
         ylab="Observed nrrm")
abline(0,1)
```

Since we detected overdispersion, we corrected our standard errors by fitting a quasi-GLM model here (variance is given by mean multiplied by dispersion parameter). Just as before, we plot the Observed vs. Fitted values.



Just as before, the above plot shows that our current model does not do a very good job of explaining the data and needs improvement to account for the zero inflation which is leading to the clustering observed around (0,0).

```
summary(glm2)
```

```
Call:
glm(formula = nrrm ~ river + road + cover + river:road + river:cover +
    road:cover, family = quasipoisson, data = d1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4400  -1.7052  -0.6972   0.8537   3.6625

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.687277   0.482041  -1.426  0.15612
river       -0.032729   0.045864  -0.714  0.47663
road         0.103789   0.030837   3.366  0.00098 ***
cover        1.251678   0.706244   1.772  0.07847 .
river:road   0.015314   0.003315   4.620 8.49e-06 ***
river:cover  0.014233   0.054571   0.261  0.79461
road:cover   0.054315   0.048657   1.116  0.26618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for quasipoisson family taken to be 2.270525)

    Null deviance: 2546.64  on 149  degrees of freedom
Residual deviance:  395.63  on 143  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

```
anova(glm2, test="Chisq")
```

```
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: nrrm

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          149    2546.64
river        1   351.94       148    2194.70 < 2.2e-16 ***
road         1  1390.02       147     804.68 < 2.2e-16 ***
cover        1   347.14       146     457.53 < 2.2e-16 ***
river:road   1    58.77       145     398.76 3.622e-07 ***
river:cover  1     0.28       144     398.48    0.7244
```

```
road:cover    1     2.85        143     395.63    0.2629
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1
```

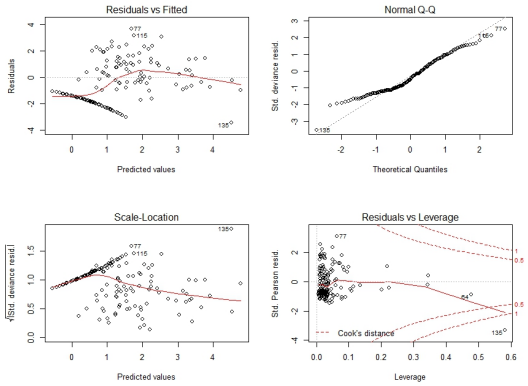
Just as before, the above presented anova (using the chi-squared test) tells us about the significance of the overall model as explanatory terms are added sequentially (from NULL to road:cover) to the overall model. Important to note here is the decrease in residual deviance as terms are added indicating improved explanatory power at each step up until river:road. Subsequent, interaction terms do not reduce the residual deviance much indicating a need for removing these (one at a time) to free up some degrees of freedom.

```
1-pchisq(deviance(glm2),143) # excessive deviance; model needs to be improved
```

```
[1] 0
```

The above code tests the residual deviance from our current model against a chi-squared distribution with 143 degrees of freedom (residual deviance dF in summary above) to see if there is more deviance than expected at random. The significant p-value (assuming a 95% confidence level) above tells us that indeed there is excessive deviance and we need to improve our model.

```
par(mfrow=c(2,2))
plot(glm2)
```



The story for our model assumptions remains unchanged too. We still need to account for zero-inflation and reduce our model further.

Chunk 6: Code And Relevant Output

```
install.packages("pscl")
library(pscl)
zip1 <- zeroinfl(nrrm ~ river + road + cover + river:road + river:cover + road:cover
                 | river + road + cover, data=d1)
summary(zip1)
```

We install and load the required package for further analysis accounting for zero inflation. The zeroinfl() command specifies the count component (before ”|”) and the zero inflation component (after ”|”)

```
Call:
zeroinfl(formula = nrrm ~ river + road + cover + river:road + river:cover + road:cover |
    river + road + cover, data = d1)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.1497 -0.5164 -0.0130  0.1960  2.2691

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.181298   0.374330   0.484  0.62815
river       -0.082744   0.035701  -2.318  0.02047 *
road         0.065150   0.024446   2.665  0.00770 **
```

```
cover        1.655472   0.535259   3.093  0.00198 **
river:road   0.018065   0.002681   6.738 1.61e-11 ***
river:cover  0.033397   0.035732   0.935  0.34997
road:cover  -0.021370   0.038732  -0.552  0.58113

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 23.06043   10.12254   2.278   0.0227 *
river       -0.05453    0.14603  -0.373   0.7088
road        -4.22083    1.90981  -2.210   0.0271 *
cover       -3.06806    2.32337  -1.321   0.1867
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 24
Log-likelihood: -239.2 on 11 Df
```

The above presented summary for our zero inflated model shows that we can further reduce our model by a stepwise removal of non-significant terms from the count componenet (remove river:cover) and the zero inflation component (remove river) of our overall model.

Chunk 7: Code and Relevant Output

```
zip2 <- zeroinfl(nrrm ~ river + road + cover + river:road + road:cover | road + cover, data=d1)
summary(zip2)
```

```
Call:
zeroinfl(formula = nrrm ~ river + road + cover + river:road + road:cover | road + cover,
    data = d1)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.1488 -0.5221 -0.0112  0.2394  2.2615

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.025171   0.317156  -0.079  0.93674
river       -0.061745   0.028298  -2.182  0.02911 *
road         0.067010   0.024638   2.720  0.00653 **
cover        1.954520   0.437924   4.463 8.08e-06 ***
river:road   0.018062   0.002667   6.772 1.27e-11 ***
road:cover  -0.020157   0.038690  -0.521  0.60238

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   24.088     12.251   1.966   0.0493 *
road          -4.522      2.329  -1.942   0.0522 .
cover         -3.188      2.560  -1.246   0.2129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 20
Log-likelihood: -239.7 on 9 Df
```

The above presented summary for our new zero inflated model shows that we can further reduce our model by a stepwise removal of non-significant terms from the count componenet (remove road:cover) and the zero inflation component (remove cover) of our overall model.

Chunk 8: Code and Relevant Output

```
zip3 <- zeroinfl(nrrm ~ river + road + cover + river:road | road, data=d1)
summary(zip3)
```

```
Call:
zeroinfl(formula = nrrm ~ river + road + cover + river:road | road, data = d1)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-2.09269 -0.46801 -0.02856  0.24762  2.27802

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.072247   0.266103   0.272  0.78601
river       -0.060458   0.027769  -2.177  0.02947 *
road         0.059896   0.021308   2.811  0.00494 **
cover        1.741514   0.130537  13.341  < 2e-16 ***
river:road   0.017758   0.002553   6.955 3.54e-12 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   17.942      5.885   3.049  0.00230 **
road          -3.574      1.183  -3.020  0.00252 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 16
Log-likelihood:  -241 on 7 Df
```

Our final model (zip3) accounts for zero inflation and has a higher Log-likelihood than our initial model (zip1) indicating it is a better fit.

Chunk 9: Code And Relevant Output

```
vuong(glm1,zip3)
```

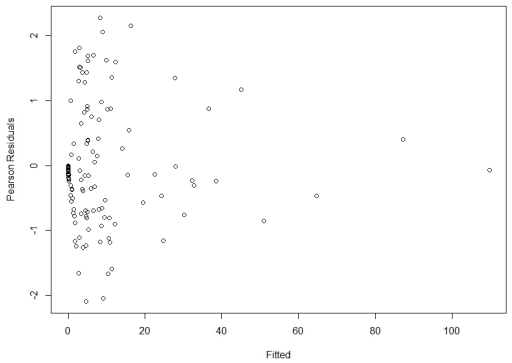
The purpose of this test is to check if our final model (zip3) is significantly better than our initial full model that did not account for zero inflation.

```
Vuong Non-Nested Hypothesis Test-Statistic: -8.076983
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 3.3194e-16
```

Indeed, our final model is significantly better than the initial full model (glm1) at the 95% confidence level.

```
par(mfrow=c(1,1))
plot(residuals(zip3, type="pearson") ~ fitted(zip3),xlab="Fitted",ylab="Pearson Residuals")
```

The purpose of this chunk is to check our model assumptions for our final model and then determine if any improvements were made over the earlier models by comparing to earlier presented figures.



In comparison to the equivalent earlier presented plots, indeed it looks like our final model is much better as the prominent earlier cluster of values around y=-1 on the earlier graph seems to be broken apart and scattered. Not reported here, but the assupmtions for normality are also met by our final model (unlike our initial model).

Chunk 10: Code And Relevant Output

The purpose of this chunk is to look at some descriptive statistics for our data and generate a table of predictor values accordingly. This table is then used to make predictions for the response variable using our final model. The output is then graphed to maximize the information presented in a single plot. The predicted response is plotted on the y axis and the most significant predictor is plotted on the y axes. Different levels of the second predictor (river) are represented by different colors and different values of the third predictor (and only significant zero inflated component) are used as a wrapper.

```
mean(d1$road)
```

```
[1] 6.444667
```

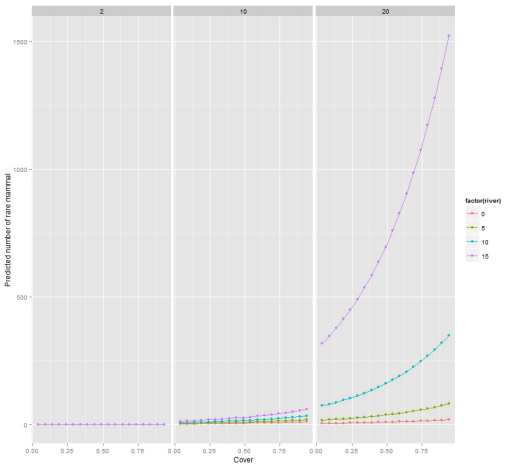
```
range(d1$river)
```

```
[1]  0.5 14.7
```

```
range(d1$cover)
```

```
[1] 0.008065092 0.921794636
```

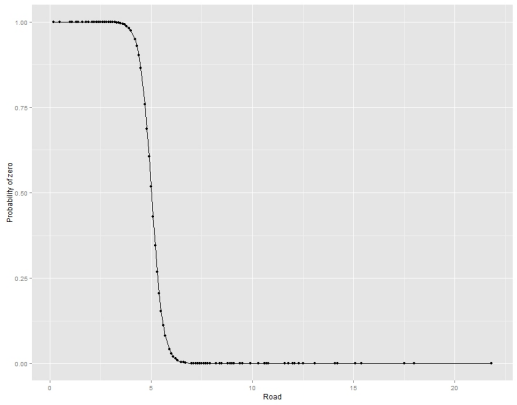
```
p.data <- expand.grid(cover=seq(0.04,0.94,by=0.05), river=c(0,5,10,15),road=c(2,10,20))

p.data$nrrm <- predict(zip3,newdata=p.data, type="response")

ggplot(p.data, aes(x = cover, y = nrrm, colour = factor(river))) +
  geom_point() +
  geom_line() +
  facet_wrap(~road) +
  labs(x = "Cover", y = "Predicted number of rare mammal")
```



Conclusions

The effect of cover on mammal abundance: From our significant results for the final model and the above presented graph, it is easy to conclude that there is a significant exponential effect of cover on the abundance of the rare mammal.
The effect of river and road on mammal abundance: From our significant results for the river:road interaction term in the final model and the above presented graph, it is easy to conclude that there the effect of river on mammal abundance is different at different levels of road. The effect of river is more profound at higher levels of road and vice-versa.

```
ggplot(d1,aes(x=d1$road,y=predict(zip3, type="zero")))+
  geom_point() +
  geom_line() +
  labs(x = "Road", y = "Probability of zero")
```



Conclusion

Based on the results for the zero inflation component of our final model and the above graph, it is easy to conclude that road has a signficant effect on the probability that the mammal found a particular location that might have suitable habitat. The graph suggest that around values for road near 5 there is a steep gradient along which the probability for suitability increases. Of course, this result comes with a grain of salt as some of the zeroes we encountered in the original data set might not have been true zeroes.
