Homework 7

Gitanshu Munjal

December 4, 2014

1 Poisson regression

1.1

[10] Explain why a Poisson family would be appropriate for the model.

The response variable in the current study consists of **counts** (number of individuals of a relatively rare mammal) from **randomly (haphazardly) chosen locations**. The expected distribution for such a variable is likely a Poisson distribution as by definiton a Poisson distribution is a **discrete probability distribution for counts** of occurence of rare events (number of individuals of a relatively rare mammal) over a time/space interval.

1.2

[20] Follow the reading (Chapter 9 of Zuur et al. 2009) pages 221-224 and build a suitable model for the data. Start with a full model that includes:

river, road, cover, tveg, river:road, river:cover and road:cover. Report each step in the variable selection. In order to keep an interaction, both interacting factors should be included in the model. Report the final model.

1.2.1 Code and Relevant Output

```
setwd("C:\\Users\\Gitanshu\\Desktop")
exotia <- read.table("hw7.txt",header=T,sep=",")

m1 <- glm(nrrm~river+tveg+road+cover+river:road+river:cover+road:cover,
          family=poisson,data=exotia)

drop1(m1,test="Chi")
```

```
Single term deletions

Model:
nrrm ~ river + tveg + road + cover + river:road + river:cover +
    road:cover
            Df Deviance    AIC    LRT Pr(>Chi)
<none>           113.03 391.06
tveg         4   115.84 385.87 2.8104  0.59004
river:road   1   117.42 393.45 4.3962  0.03602 *
river:cover  1   115.75 391.78 2.7223  0.09895 .
road:cover   1   113.12 389.15 0.0979  0.75440
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1
```

The **significance (not-significant) and single term deletion (lower AIC without tveg)** results shown above make the case for **removing tveg from the overall model** to improve model fit. The above presented cycle was repeated by a stepwise removal of model terms and the model was reduced to reach a final model with the lowest AIC. The same result is automated and done faster by using the step() command in the following section.

```
step(m1)
```

```
Start:  AIC=391.06
nrrm ~ river + tveg + road + cover + river:road + river:cover +
    road:cover

              Df Deviance    AIC
- tveg         4   115.84 385.87
- road:cover   1   113.12 389.15
<none>             113.03 391.06
- river:cover  1   115.75 391.78
- river:road   1   117.42 393.45

Step:  AIC=385.87
nrrm ~ river + road + cover + river:road + river:cover + road:cover

              Df Deviance    AIC
- road:cover   1   116.00 384.03
<none>             115.84 385.87
- river:cover  1   118.69 386.71
- river:road   1   119.13 387.15

Step:  AIC=384.03
nrrm ~ river + road + cover + river:road + river:cover

              Df Deviance    AIC
<none>             116.00 384.03
- river:cover  1   118.82 384.85
- river:road   1   119.36 385.39

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

Coefficients:
(Intercept)        river         road        cover   river:road  river:cover
  -0.359746    -0.023844     0.060278     2.575014     0.006085    -0.104628

Degrees of Freedom: 99 Total (i.e. Null);  94 Residual
Null Deviance:      247.7
Residual Deviance: 116   AIC: 384
```

The final(reduced) model thus contains river + road + cover + river:road + river:cover.

1.3

[10] Use the dispersiontest() function of the AER package to determine if there is significant overdispersion.

1.3.1 Code and Relevant Output

```
m3 <- glm(nrrm~river+road+cover+river:road+river:cover,family=poisson,data=exotia)

library(AER)
dispersiontest(m3)
```

```
Overdispersion test

data:  m3
z = 0.1683, p-value = 0.4332
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.020997
```

The non-significant p-value for the above presented Overdispersion test tells us that there was **no significant overdispersion** in our final model

1.4

[10] In one or two sentences, what is overdispersion?

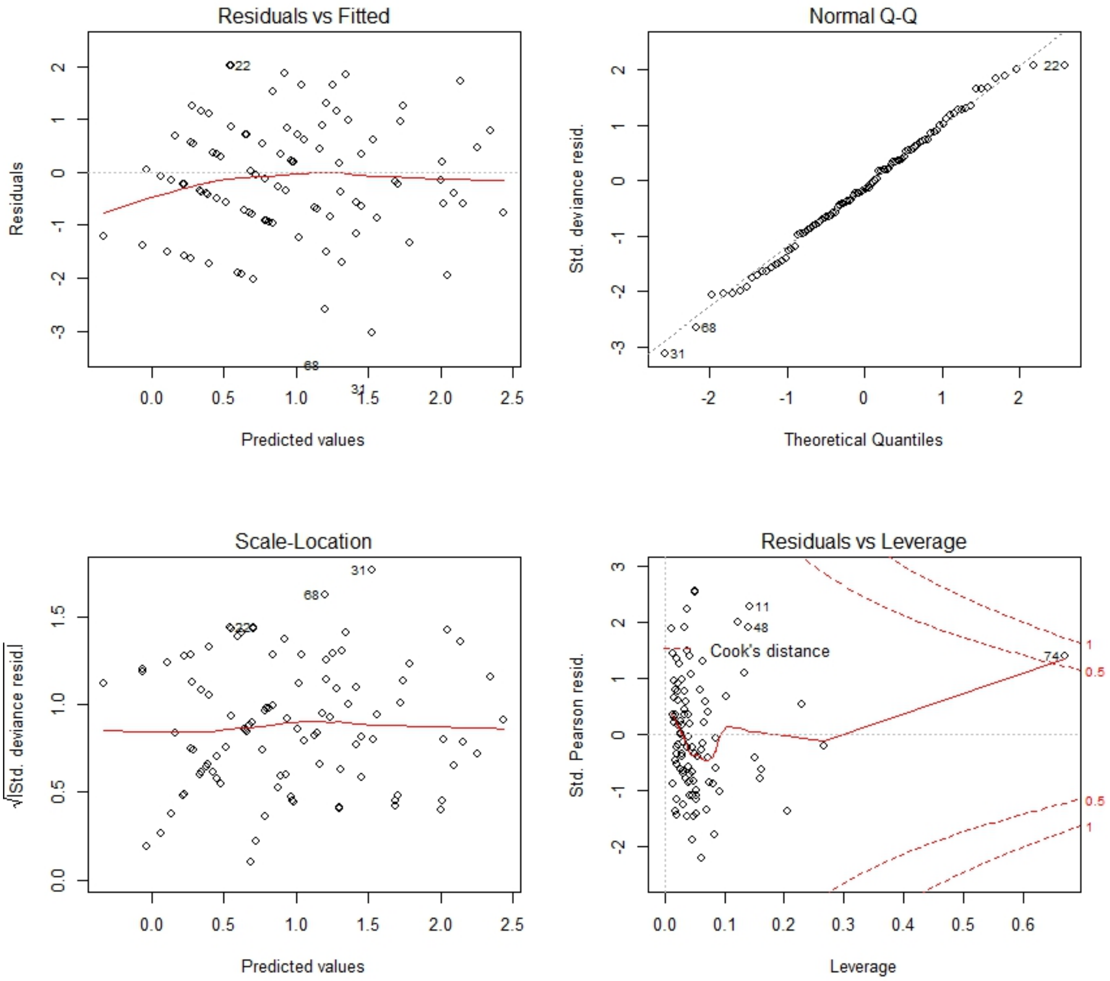
Overdispersion describes the situation where the **variance for a given model is larger than expected**. In our case, we're assuming our counts are modeled like a Poisson distribution. A Poisson regression assumes that the mean and variance of the data are equal. Therefore, **in our case overdisperion is described as the situation where variance is greater than the mean (the mean is the expectation for variance based on the Poisson assumption).**

1.5

[15] Follow the reading (Chapter 9 of Zuur et al. 2009) pages 228-231 to check for deviations of assumptions (what Zuur et al. refer to as "validation"). Report the graphs and one sentence interpretation for each graph.

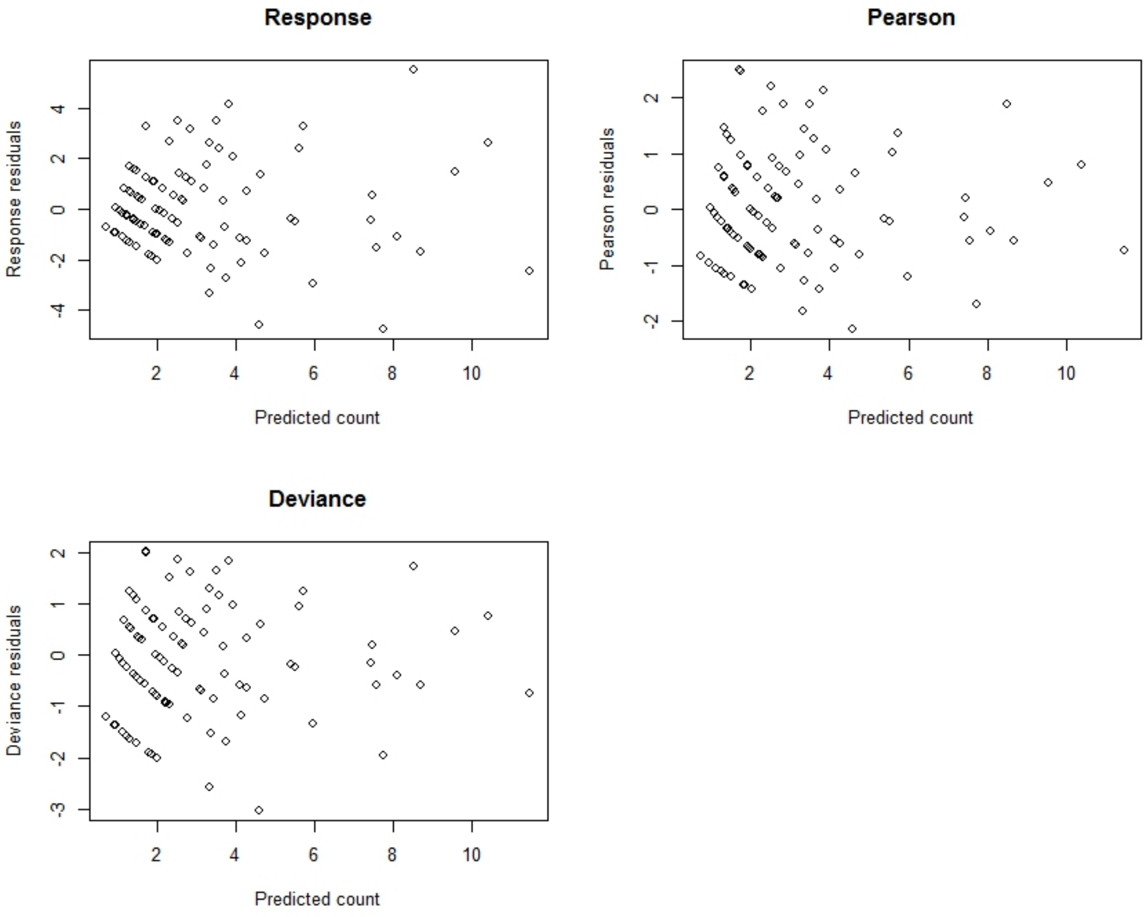
1.5.1 Code And Relevant Output

```
par(mfrow=c(2,2))
plot(m3)
```



The two plots on the left (Residuals Vs Fitted and Scale-Location) both show a definite pattern indicating that our assumptions regarding independence might be violated and/or that the current model is not good. The normal Q-Q plot does not show any signs of concern about normality except for a few outliers (which are also higlighted in the earlier mentioned two plots). The Residuals vs Leverage plot shows that there are a few data points that are influencing our model more than other data points.

```
j1 <-m3
ep <- resid(j1, type = "pearson")
ed <- resid(j1, type = "deviance")
mu <- predict(j1, type = "response")
e  <- exotia$nrrm - mu
par(mfrow = c(2,2))
plot(x = mu, y = e, ylab = "Response residuals",xlab="Predicted count",main="Response")
plot(x = mu, y = ep, ylab = "Pearson residuals",xlab="Predicted count",main="Pearson")
plot(x = mu, y = ed, ylab = "Deviance residuals",xlab="Predicted count",main="Deviance")
```



The pattern mentioned for residuals remains consistent through our analysis indicating an insufficient model and maybe a possibly missing covariate.

1.6

[15] What is the effect of cover on the abundance of the mammal? Answer this by interpreting the corresponding coefficient in the model.

1.6.1 Code and Relevant Output

```
summary(m3)
```

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.359746   0.371802  -0.968   0.3333
river       -0.023844   0.049758  -0.479   0.6318
road         0.060278   0.024249   2.486   0.0129 *
cover        2.575014   0.446991   5.761 8.37e-09 ***
river:road   0.006085   0.003355   1.814   0.0697 .
river:cover -0.104628   0.061876  -1.691   0.0908 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1
```

The poisson model is given by

$$Y_i \sim (\mu_i)$$

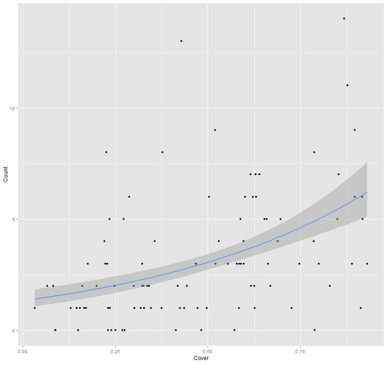
$$E(Y_i) = \mu_i$$

$$var(Y_i) = \mu_i$$

$$\mu_i = e^{\alpha+\beta_1 X_1+...+\beta_n X_n}$$

Where,
Y is the response variable
μ is the mean
β is the coefficient of X
X is the effect of an explanatory variable
α is the slope for the model
Based on this model and the estimate for the coefficient (2.57) for cover in the model, we can conclude that there is a significant exponential effect of cover on the abundance of the mammal at the 95% confidence level and that for every unit increasing in cover there is an $e^{2.6}$ increase in abundance. Indeed, this effect is evident in a plot of mammal counts against cover as shown below.

```
library(ggplot2)
ggplot(aes(x=cover,y=nrrm),data=exotia,xlab="Cover",ylab="Count")+
  geom_point()+
  geom_smooth(method=glm,family=poisson)+
  labs(x="Cover",y="Count")
```



1.7

[20] Interpret the effects of cover and river by making predictions for nrrm for a factorial combination of two reasonable values for each variable. Report a table with predictor and predicted values in both response and link scales. Use help(”predict.glm”) for details.

1.7.1 Code And Relevant Output

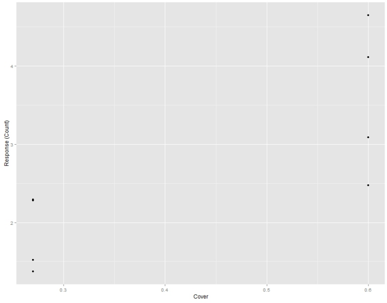
```
newexo <- data.frame(river= rep(c(4.1,7.5),c(4,4)),
                     road=rep(c(3.5,8.3),c(2,2)),cover= rep(c(0.27,0.6),c(1,1)))

link<-t(t(predict(m3,newdata=newexo,type = "link")))
response<-t(t(predict(m3,newdata=newexo,type = "response")))
newexo<-cbind(newexo,link,response)
newexo
```

```
river road cover         link response
1   4.1  3.5  0.27 0.4202169 1.522292
2   4.1  3.5  0.60 1.1284100 3.090738
3   4.1  8.3  0.27 0.8293054 2.291726
4   4.1  8.3  0.60 1.5374985 4.652936
5   7.5  3.5  0.27 0.3155104 1.370959
6   7.5  3.5  0.60 0.9063109 2.475175
7   7.5  8.3  0.27 0.8239077 2.279390
8   7.5  8.3  0.60 1.4147083 4.115286
```

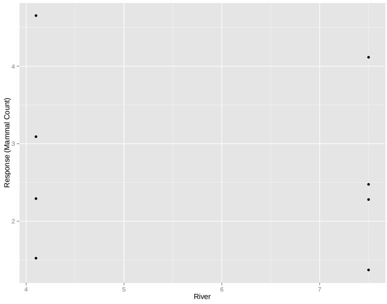
The above table shows predictor(first 3 columns) and predicted values(column name = scale type).

```
ggplot(aes(x=cover,y=response),data=newexo)+
  geom_point()+
  labs(x="Cover",y="Response (Mammal Count)")
```



The above presented scatterplot shows the positive effect of cover on mammal abundance. This finding supports our significant results for cover although the exponential nature cannot be visualized here due to the fact that we only used two values for cover. Next we look ath effects of river on mammal abundance

```
ggplot(aes(x=river,y=response),data=newexo)+
  geom_point()+
  labs(x="River",y="Response (Mammal Count)")
```



The above presented scatterplot shows that there is no easily observable effect of river on mammal abundance. This finding supports our non-significant results for river.