Michael Leibert

Part I: Chronic respiratory disease [25 points]

Table 1 summarizes the data from an epidemiological study of chronic respiratory disease. Researchers collected information on subjects' exposure to general pollution (low or high), exposure to pollution in their jobs (yes or no), and their smoking status (current smoker, ex-smoker, non-smoker). The measured response is chronic respiratory disease status classified into four categories:

- 1 – no symptoms
- 2 – cough or phlegm less than 3 months a year
- 3 – cough or phlegm more than 3 months a year
- 4 – cough and phlegm plus shortness of breath more than 3 months a year

			Response level				
Air pollution	Job exposure	Smoking status	1	2	3	4	Total
Low	No	Non	158	9	5	0	172
		Ex	167	19	5	3	194
		Current	307	102	83	68	560
	Yes	Non	26	5	5	1	37
		Ex	38	12	4	4	58
		Current	94	48	46	60	248
High	No	Non	94	7	5	1	107
		Ex	67	8	4	3	82
		Current	184	65	33	36	318
	Yes	Non	32	3	6	1	42
		Ex	39	11	4	2	56
		Current	77	48	39	51	215

Table 1: Chronic respiratory disease data

1. Fit a proportional odds cumulative logit model with pairwise interaction effects for all covariates and assess its goodness of fit. Use low air pollution, no job exposure and non-smoker as reference group.

```
rm(list = ls());setwd("G:\\math\\661"); options(scipen=999);  library(VGAM,quietly=T)
tempp<-data.frame( rep( "low" ,6 ) );names(tempp)<-"pollution"
tempp$exposure<- c( rep( "no" ,3 ),  rep( "yes" ,3 ))
tempp$smoker<-  rep( c("non","ex","current") ,2 )
mat<-matrix(NA,6,4)
mat[1,]<-c(158,9,5,0 );mat[2,]<-c(167 ,19 , 5, 3)
mat[3,]<-c(307,102,83 ,68);mat[4,]<-c(26, 5 ,5, 1)
mat[5,]<-c(38, 12, 4, 4 );mat[6,]<-c(94, 48, 46, 60 )
cough<-cbind(tempp,mat); names(cough)[ 4:ncol(cough)]<-paste0("Y",1:4)

tempp<-data.frame( rep( "high" ,6 ) );names(tempp)<-"pollution"
```

```
tempp$exposure<- c( rep( "no" ,3 ),  rep( "yes" ,3 ))
tempp$smoker<-  rep( c("non","ex","current") ,2 )
mat<-matrix(NA,6,4)
mat[1,]<-c(94, 7, 5 ,1 ); mat[2,]<-c(67, 8, 4 ,3)
mat[3,]<-c(184, 65, 33, 36); mat[4,]<-c(32, 3, 6, 1)
mat[5,]<-c(39, 11, 4 ,2); mat[6,]<-c(77, 48, 39, 51)
tempp<-cbind(tempp,mat); names(tempp)[ 4:ncol(tempp)]<-paste0("Y",1:4)
cough<-rbind(cough,tempp)
cough$smoker<-relevel(as.factor(cough$smoker), ref="non")
cough$exposure<-relevel(as.factor(cough$exposure), ref="no")
cough$pollution<-relevel(as.factor(cough$pollution), ref="low");rm(tempp)

cough
```

```
##    pollution exposure  smoker  Y1  Y2 Y3 Y4
## 1        low       no     non 158   9  5  0
## 2        low       no      ex 167  19  5  3
## 3        low       no current 307 102 83 68
## 4        low      yes     non  26   5  5  1
## 5        low      yes      ex  38  12  4  4
## 6        low      yes current  94  48 46 60
## 7       high       no     non  94   7  5  1
## 8       high       no      ex  67   8  4  3
## 9       high       no current 184  65 33 36
## 10      high      yes     non  32   3  6  1
## 11      high      yes      ex  39  11  4  2
## 12      high      yes current  77  48 39 51
```

```
str(cough)
```

```
## 'data.frame':    12 obs. of  7 variables:
##  $ pollution: Factor w/ 2 levels "low","high": 1 1 1 1 1 1 2 2 2 2 ...
##  $ exposure : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 1 1 2 ...
##  $ smoker   : Factor w/ 3 levels "non","current",..: 1 3 2 1 3 2 1 3 2 1 ...
##  $ Y1       : num  158 167 307 26 38 94 94 67 184 32 ...
##  $ Y2       : num  9 19 102 5 12 48 7 8 65 3 ...
##  $ Y3       : num  5 5 83 5 4 46 5 4 33 6 ...
##  $ Y4       : num  0 3 68 1 4 60 1 3 36 1 ...
```

```
inter.fit<-vglm(cbind(Y1,Y2,Y3,Y4) ~ pollution + exposure + smoker + smoker*exposure +
    smoker*pollution+pollution*exposure,    family=cumulative(parallel=T),data=cough)
summary(inter.fit)
```

```
##
## Call:
## vglm(formula = cbind(Y1, Y2, Y3, Y4) ~ pollution + exposure +
##     smoker + smoker * exposure + smoker * pollution + pollution *
##     exposure, family = cumulative(parallel = T), data = cough)
##
##
## Pearson residuals:
##                      Min      1Q   Median      3Q   Max
## logit(P[Y<=1]) -1.044 -0.5076 -0.30835 0.08112 1.079
## logit(P[Y<=2]) -1.398 -0.9158  0.02686 0.79555 1.389
## logit(P[Y<=3]) -0.708 -0.3029  0.41987 1.05726 1.739
```

```
##
## Coefficients:
##                              Estimate Std. Error z value
## (Intercept):1                2.311609   0.244777   9.444
## (Intercept):2                3.191984   0.248520  12.844
## (Intercept):3                4.114690   0.254066  16.195
## pollutionhigh               -0.172089   0.327883  -0.525
## exposureyes                 -1.217546   0.339051  -3.591
## smokercurrent               -2.113864   0.254365  -8.310
## smokerex                    -0.547229   0.306770  -1.784
## exposureyes:smokercurrent    0.405182   0.342715   1.182
## exposureyes:smokerex         0.296951   0.419860   0.707
## pollutionhigh:smokercurrent  0.260242   0.335750   0.775
## pollutionhigh:smokerex       0.075304   0.416457   0.181
## pollutionhigh:exposureyes   -0.003072   0.191241  -0.016
##                                        Pr(>|z|)
## (Intercept):1               < 0.0000000000000002 ***
## (Intercept):2               < 0.0000000000000002 ***
## (Intercept):3               < 0.0000000000000002 ***
## pollutionhigh                          0.599688
## exposureyes                            0.000329 ***
## smokercurrent               < 0.0000000000000002 ***
## smokerex                               0.074450 .
## exposureyes:smokercurrent              0.237098
## exposureyes:smokerex                   0.479404
## pollutionhigh:smokercurrent            0.438277
## pollutionhigh:smokerex                 0.856509
## pollutionhigh:exposureyes              0.987185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Residual deviance: 27.2964 on 24 degrees of freedom
##
## Log-likelihood: -84.4742 on 24 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##               pollutionhigh                 exposureyes
##                   0.8419039                   0.2959556
##               smokercurrent                    smokerex
##                   0.1207705                   0.5785508
##   exposureyes:smokercurrent        exposureyes:smokerex
##                   1.4995756                   1.3457487
## pollutionhigh:smokercurrent      pollutionhigh:smokerex
##                   1.2972434                   1.0782114
##   pollutionhigh:exposureyes
```

```
##                   0.9969330
```

The goodness-of-fit test:

$$H_0\text{: model fits the data well} \quad \text{vs.} \quad H_1\text{: model does not fit the data well.}$$

The current (interaction) model has $df = 36 - 12 = 24$, $\Big((4-1)\cdot 12 = 36\Big)$ and the residual deviance is 27.2964, which leads to a p-value $= 0.2908274$. The model provides an adequate fit.

```
1-pchisq(27.29641,24)
```

```
## [1] 0.2908274
```

2. Use a likelihood ratio test to check the proportional odds assumption in the model above.

   We use a likelihood ratio test to check the proportional odds assumption in the model above. We can test the proportional odds assumption:

$$H_0\text{: same slope for all cumulative logits} \quad \text{vs.} \quad H_1\text{: different slopes}$$

```
noprop.fit<-vglm(cbind(Y1,Y2,Y3,Y4) ~ pollution + exposure + smoker +   smoker*exposure +
  smoker*pollution+pollution*exposure , family=cumulative(parallel=F),data=cough)

1-pchisq(-2*(logLik(inter.fit)-logLik(noprop.fit)), df=df.residual(inter.fit)-df.residual((noprop.fit))
```

```
## [1] 0.2369979
```

   We fail to reject the null and can assume a proportional odds structure.

3. Use a likelihood ratio test to determine whether to include or not the interaction terms in the proportional odds cumulative logit model.

$$H_0\text{: model with main effects fits as well as model with interaction effects}$$

   We can use a likelihood ratio test to compare this model to the main effects model, which will have an approximate chi-square distribution with $df = 29 - 24 = 5$. This is equivalent to evaluating the change in deviance between the two models:

```
main.fit<-vglm(cbind(Y1,Y2,Y3,Y4) ~ pollution + exposure + smoker,
               family=cumulative(parallel=T),data=cough)
deviance(main.fit)-deviance(inter.fit)
```

```
## [1] 2.700513
```

```
df.residual(main.fit); df.residual(inter.fit)
```

```
## [1] 29
```

```
## [1] 24
```

```
1-pchisq(deviance(main.fit)-deviance(inter.fit), 5)
```

```
## [1] 0.7460399
```

We fail to reject H_0 at $\alpha = 0.05$, thus the model with interaction effects doesn't provide a better fit compared to the main effects model.

4. In the following questions, use the main effects cumulative logit proportional odds model:

```
summary(main.fit)
```

```
##
## Call:
## vglm(formula = cbind(Y1, Y2, Y3, Y4) ~ pollution + exposure +
##     smoker, family = cumulative(parallel = T), data = cough)
##
##
## Pearson residuals:
##                       Min      1Q   Median      3Q   Max
## logit(P[Y<=1]) -0.9786 -0.6485 -0.06329 0.09382 1.164
## logit(P[Y<=2]) -2.0789 -0.9668  0.23074 0.77460 1.502
## logit(P[Y<=3]) -0.4941 -0.3384  0.32979 1.04577 1.773
##
## Coefficients:
##                Estimate Std. Error z value             Pr(>|z|)
## (Intercept):1  2.08844    0.16329  12.790 <0.0000000000000002 ***
## (Intercept):2  2.96964    0.16927  17.544 <0.0000000000000002 ***
## (Intercept):3  3.89385    0.17786  21.893 <0.0000000000000002 ***
## pollutionhigh  0.03929    0.09370   0.419               0.6750
## exposureyes   -0.86476    0.09546  -9.059 <0.0000000000000002 ***
## smokercurrent -1.85271    0.16503 -11.227 <0.0000000000000002 ***
## smokerex      -0.40003    0.20187  -1.982               0.0475 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  3
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])
##
## Residual deviance: 29.9969 on 29 degrees of freedom
##
## Log-likelihood: -85.8245 on 29 degrees of freedom
##
## Number of iterations: 4
```

```
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
## pollutionhigh   exposureyes smokercurrent      smokerex
##     1.0400744     0.4211522     0.1568115     0.6703009
```

(a) Interpret each of the three intercepts.

$\hat{\alpha}_1 = 2.08844$: Is the estimated baseline log odds of falling into category 1 (no symptoms) versus all other categories for individuals with exposure to low air pollution, exposure to low job pollution, and who are non-smokers.

$\hat{\alpha}_2 = 2.96964$: Is the estimated baseline log odds of falling into either category 1 (no symptoms) or category 2 (cough or phlegm less than 3 months a year) versus all other categories for individuals with exposure to low air pollution, exposure to low job pollution, and who are non-smokers.

$\hat{\alpha}_3 = 3.89385$: Is the estimated baseline log odds of falling into either category 1 (no symptoms), category 2 (cough or phlegm less than 3 months a year), category 3 (cough or phlegm more than 3 months a year) versus all other categories for individuals with exposure to low air pollution, exposure to low job pollution, and who are non-smokers.

(b) Which variables appear to be associated with chronic respiratory disease? Interpret the regression coefficients for the covariates with significant association.

The variables that appear to be associate with chronic respiratory disease are **exposure to pollution in their jobs**, **smoking status: current**, and **smoking status: ex-smoker**.

Controlling for air pollution and smoking status, the estimated log-odds of a response in the category 1 direction rather than the category 4 direction for individuals who are exposed to pollution in their jobs is 0.86476 lower than individuals who are not exposed to pollution in their jobs. (i.e., the estimated odds ratio for individuals who are exposed to pollution in their jobs versus individuals who are not exposed to pollution in their jobs is $\exp(-0.86476) = 0.4211526$). This indicates, controlling for pollution and smoking status, individuals who are exposed to pollution are more likely to have worse symptom outcomes.

Controlling for air pollution and job pollution, the estimated log-odds of a response in the category 1 direction rather than the category 4 direction for current smokers is 1.85271 lower than non-smokers. (i.e., the estimated odds ratio for smokers versus non-smokers is $\exp(-1.85271) = 0.1568116$). This indicates, controlling for pollution and job pollution, individuals who are smokers are much more likely to have worse symptom outcomes.

Controlling for air pollution and job pollution, the estimated log-odds of a response in the category 1 direction rather than the category 4 direction for ex-smokers is 0.40003 lower than non-smokers. (i.e., the estimated odds ratio for smokers versus non-smokers is $\exp(-0.40003) = 0.6702999$). This indicates,

controlling for pollution and job pollution, individuals who are ex-smokers are more likely to have worse symptom outcomes.

(c) What are the estimated probabilities of falling in each of the different response categories for a current smoker with job exposure to pollution and high general air pollution exposure? Show the details of your calculations manually.

$$\log\left(\frac{\hat{P}(y_i \leq 1)}{1-\hat{P}(y_i \leq 1)}\right) = 2.08843944 + 0.03929225 - 0.86476105 - 1.85271076 = -0.5897401$$

$$\hat{P}(y_i \leq 1) = \frac{\exp(-0.5897401)}{1+\exp(-0.5897401)} = 0.3566945$$

$$\log\left(\frac{\hat{P}(y_i \leq 2)}{1-\hat{P}(y_i \leq 2)}\right) = 2.96964165 + 0.03929225 - 0.86476105 - 1.85271076 = 0.2914621$$

$$\hat{P}(y_i \leq 2) = \frac{\exp(0.291462)}{1+\exp(0.291462)} = 0.572354$$

$$\hat{P}(y_i = 2) = \hat{P}(y_i \leq 2) - \hat{P}(y_i \leq 1) = 0.572354 - 0.3566945 = 0.2156595$$

$$\log\left(\frac{\hat{P}(y_i \leq 3)}{1-\hat{P}(y_i \leq 3)}\right) = 3.89385000 + 0.03929225 - 0.86476105 - 1.85271076 = 1.215671$$

$$\hat{P}(y_i \leq 3) = \frac{\exp(1.21567)}{1+\exp(1.21567)} = 0.7713007$$

$$\hat{P}(y_i = 3) = \hat{P}(y_i \leq 3) - \hat{P}(y_i \leq 2) = 0.7713007 - 0.572354 = 0.1989467$$

$$\hat{P}(y_i = 4) = 1 - \hat{P}(y_i \leq 3) = 1 - 0.7713007 = 0.2286993$$

(d) For each covariate pattern, provide the predicted number of people falling in each of the response levels.

```
coughing<-cough;coughing$tY<-rowSums(coughing[,4:7]);
coughing<-cbind(coughing,round(predict(main.fit,type="response"),3))
coughing; names(coughing)[(ncol(coughing)-3):ncol(coughing)]<-paste0("pi",1:4)
```

```
##    pollution exposure  smoker  Y1  Y2 Y3 Y4  tY    Y1    Y2    Y3    Y4
## 1        low       no     non 158   9  5  0 172 0.890 0.061 0.029 0.020
## 2        low       no      ex 167  19  5  3 194 0.844 0.085 0.042 0.029
## 3        low       no current 307 102 83 68 560 0.559 0.195 0.132 0.115
## 4        low      yes     non  26   5  5  1  37 0.773 0.119 0.062 0.046
## 5        low      yes      ex  38  12  4  4  58 0.695 0.151 0.087 0.067
## 6        low      yes current  94  48 46 60 248 0.348 0.215 0.202 0.236
## 7       high       no     non  94   7  5  1 107 0.894 0.059 0.028 0.019
## 8       high       no      ex  67   8  4  3  82 0.849 0.082 0.040 0.028
## 9       high       no current 184  65 33 36 318 0.568 0.192 0.128 0.111
```

```
## 10      high      yes     non  32   3  6  1  42 0.780 0.116 0.060 0.044
## 11      high      yes      ex  39  11  4  2  56 0.703 0.148 0.084 0.065
## 12      high      yes current  77  48 39 51 215 0.357 0.216 0.199 0.229
```

```
print( cbind(coughing, coughing[,(ncol(coughing)-3):ncol(coughing)]*
    coughing[,(ncol(coughing)-4)] ) )
```

```
##    pollution exposure  smoker  Y1  Y2 Y3 Y4  tY   pi1   pi2   pi3   pi4
## 1        low       no     non 158   9  5  0 172 0.890 0.061 0.029 0.020
## 2        low       no      ex 167  19  5  3 194 0.844 0.085 0.042 0.029
## 3        low       no current 307 102 83 68 560 0.559 0.195 0.132 0.115
## 4        low      yes     non  26   5  5  1  37 0.773 0.119 0.062 0.046
## 5        low      yes      ex  38  12  4  4  58 0.695 0.151 0.087 0.067
## 6        low      yes current  94  48 46 60 248 0.348 0.215 0.202 0.236
## 7       high       no     non  94   7  5  1 107 0.894 0.059 0.028 0.019
## 8       high       no      ex  67   8  4  3  82 0.849 0.082 0.040 0.028
## 9       high       no current 184  65 33 36 318 0.568 0.192 0.128 0.111
## 10      high      yes     non  32   3  6  1  42 0.780 0.116 0.060 0.044
## 11      high      yes      ex  39  11  4  2  56 0.703 0.148 0.084 0.065
## 12      high      yes current  77  48 39 51 215 0.357 0.216 0.199 0.229
##        pi1     pi2    pi3    pi4
## 1  153.080  10.492  4.988  3.440
## 2  163.736  16.490  8.148  5.626
## 3  313.040 109.200 73.920 64.400
## 4   28.601   4.403  2.294  1.702
## 5   40.310   8.758  5.046  3.886
## 6   86.304  53.320 50.096 58.528
## 7   95.658   6.313  2.996  2.033
## 8   69.618   6.724  3.280  2.296
## 9  180.624  61.056 40.704 35.298
## 10  32.760   4.872  2.520  1.848
## 11  39.368   8.288  4.704  3.640
## 12  76.755  46.440 42.785 49.235
```

Part II: Number of plant species in the Galápagos [25 points]

The 30 islands in the Galápagos archipelago have long been studied by botanists, zoologists and biologists to learn about species survival and the process of natural selection in an almost experimental setting. The islands are essentially uninhabited by humans and all experience the same surrounding climate. Yet some species of birds, plants and mammals thrive on only a few or even just one of the islands. In addition, some islands have a wide variety of species, while others are not nearly as biodiverse. We are interested in investigating which variables may be related to the number of plant species in the archipelago islands. \

The data `Galapagos.txt` posted on Canvas contain information on plant species on the Galápagos islands. The variables in the data correspond to

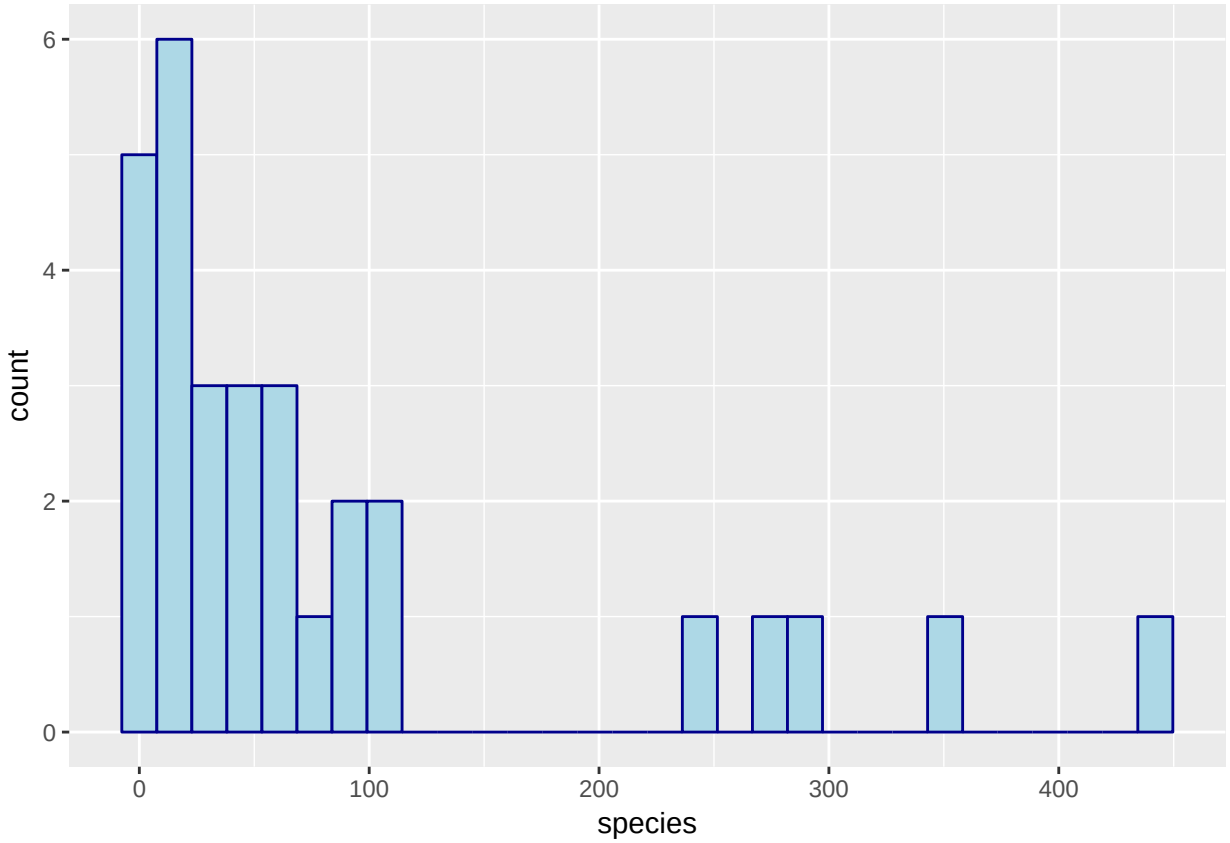
- `island` – name of island
- `species` – island total observed plant species count
- `endemics` – island endemic plant species count
- `area` – island area (km^2)
- `elevation` – island elevation (meters)
- `nearest` – distance in km from the island to its nearest neighbor (adjacent island)
- `scruz` – distance in km from the island to the largest island (Santa Cruz)
- `adjacent` – area of the adjacent island

(1) **Exploratory data analysis**

(a) Provide a histogram and summary statistics for the observed counts of total plant species. Discuss the distribution.

```
rm(list = ls()); setwd("G:\\math\\661"); options(scipen=999);require(ggplot2,quietly=T)
plant<-read.table("Galapagos.txt",header=T);source("multiplot.r")

ggplot(plant, aes(x=species))+  geom_histogram(color="darkblue", fill="lightblue",bins=30)
```




```
summary(plant$species);IQR(plant$species)
```

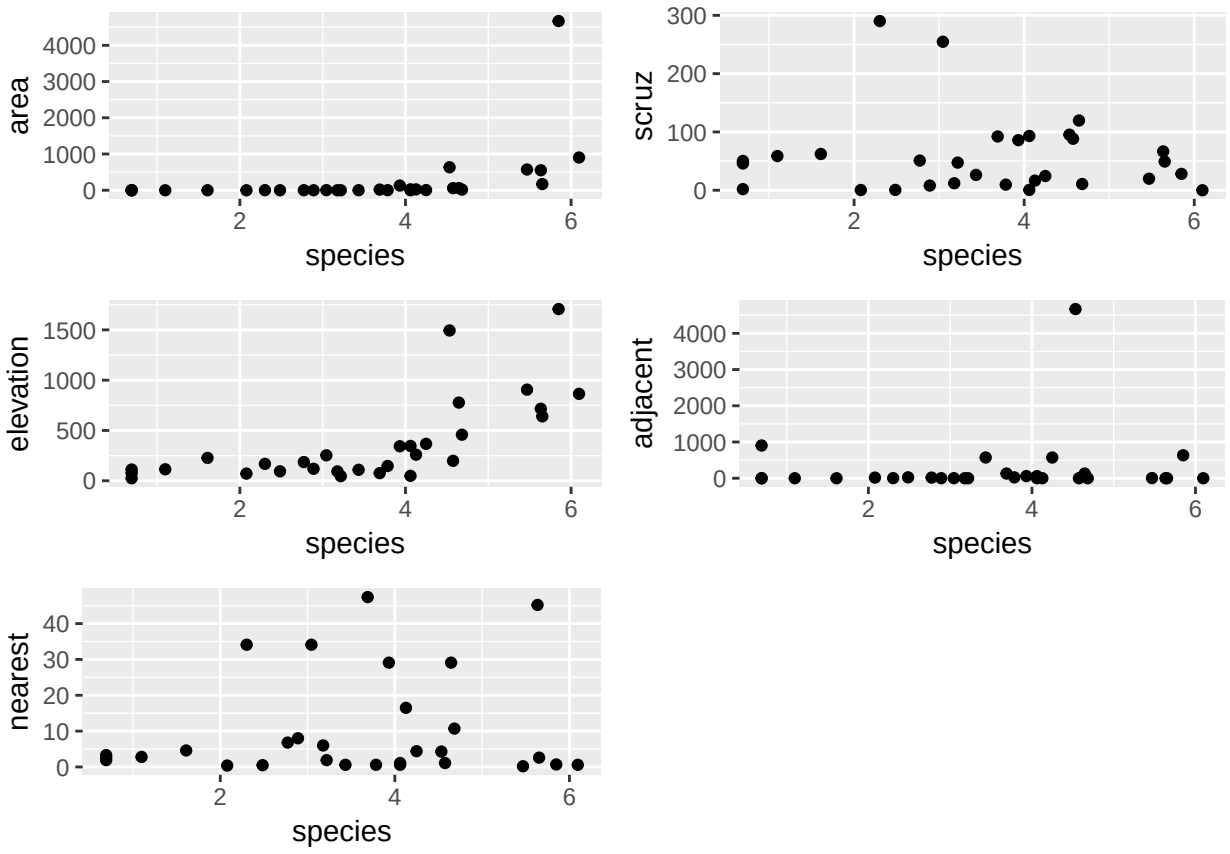
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   13.00   42.00   85.23   96.00  444.00
```

```
## [1] 83
```

The distribution of the counts is skewed to the right with a large proportion of islands having small counts. More than half of the islands have counts of species less than 50, with roughly 75% of the islands having less than 100. The median species count for the islands is 85.23333 (median = 42) and a standard deviation of 114.6331 (IQR = 83). The mean and the standard deviation are being heavily influenced by the islands that have disproportionately large counts.

(b) Create plots of the logarithm of the observed counts of total plant species, log(**species**), versus each of the five potential covariates: **area, elevation, nearest, scruz, adjacent**.

```
dat<-plant[,c(which(names(plant) %in% c("species","area", "elevation",  "nearest",
  "scruz", "adjacent")))];  gglist<-list();dat[,1]<-log(dat[,1])
for(i in 2:ncol(dat)){  gglist[[i]] <- ggplot(dat, aes_string(x="species",y=colnames(dat)[i]
    )) +   geom_point()    }
multiplot(gglist[[2]], gglist[[3]], gglist[[4]], gglist[[5]],gglist[[6]],    cols=2)
```

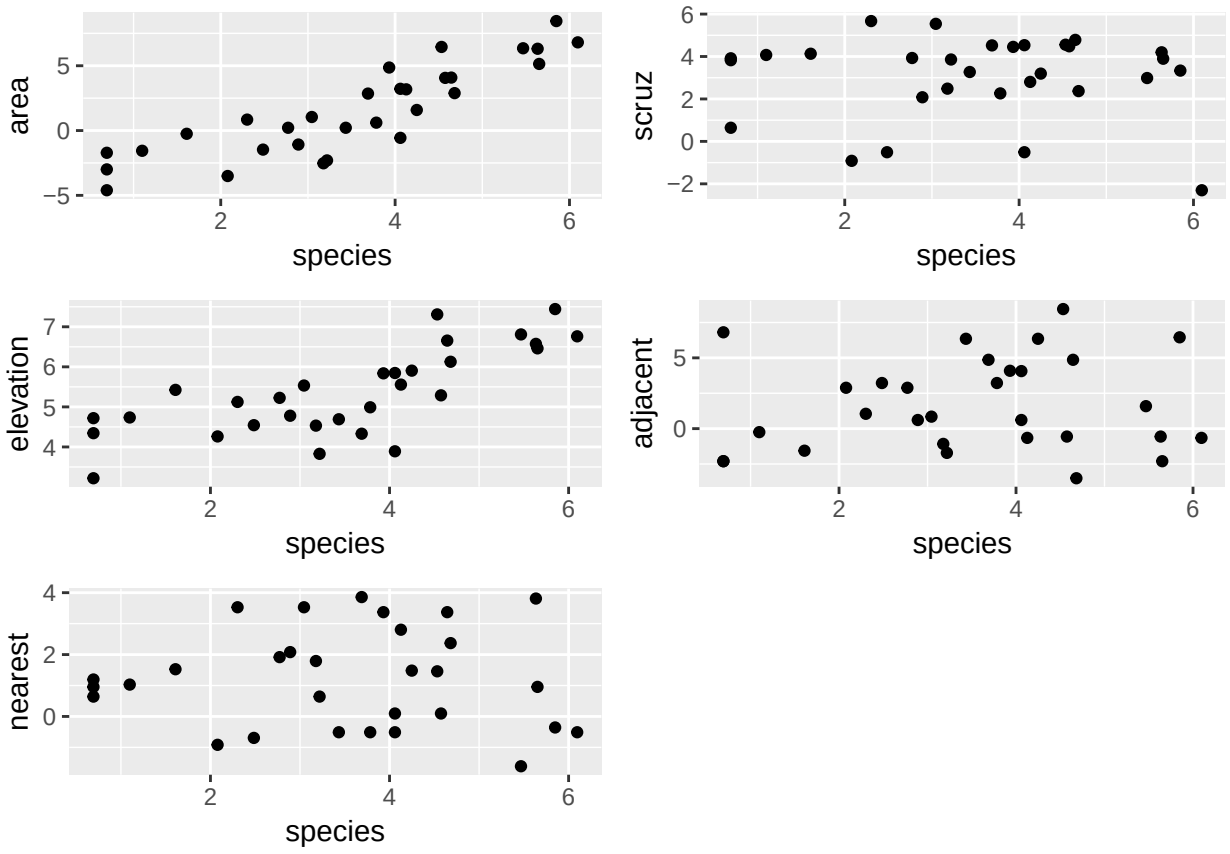


(c) Repeat the previous question using the logarithm of each of the covariates. Which variables appear to be related to log(`species`)?
Caution: Always check for any zero value before using a logarithm transformation. A quick fix is to add a small non-zero number, e.g., consider $x + 0.1$ instead of x.

```
dat[dat== 0] <- .1;dat[,-1]<-log(dat[,-1]);head(dat,3)
```

```
##    species       area elevation    nearest      scruz   adjacent
## 1 4.060443  3.2224694  5.846439 -0.5108256 -0.5108256  0.6097656
## 2 3.433987  0.2151114  4.691348 -0.5108256  3.2695689  6.3497157
## 3 1.098612 -1.5606477  4.736198  1.0296194  4.0724397 -0.2484614
```

```
for(i in 2:ncol(dat)){  gglist[[5+i]] <-
  ggplot(dat, aes_string(x="species",y=colnames(dat)[i]  )) +   geom_point()    }
multiplot(gglist[[2+5]], gglist[[3+5]], gglist[[4+5]], gglist[[5+5]],
    gglist[[5+6]],   cols=2)
```



The variables appear to be related to `log(species)` are `area` and `elevation`.

2. **Model building & diagnostics**

(a) Fit a Poisson model with all five covariates on the log scale. Which covariates appear to have a significant effect on species counts?

```
rm(list = ls()); setwd("G:\\math\\661"); options(scipen=999); library(MASS)

plants<-read.table("Galapagos.txt",header=T); plant<-read.table("Galapagos.txt",header=T)

plants<-plants[,c(which(names(plants) %in% c("species","area", "elevation",
    "nearest",  "scruz", "adjacent")))]
plants[plants== 0] <- .1;plants[,-1]<-log(plants[,-1])
plants.pois<-glm( species ~ . , family=poisson, data=plants)
summary(plants.pois)
```

```
##
## Call:
## glm(formula = species ~ ., family = poisson, data = plants)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -5.4488  -2.6730  -0.4513   2.5583   8.2983
##
## Coefficients:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  3.285301   0.284638  11.542 < 0.0000000000000002 ***
## area         0.348726   0.018026  19.346 < 0.0000000000000002 ***
## elevation    0.036338   0.056983   0.638              0.52368
## nearest     -0.041077   0.013789  -2.979              0.00289 **
## scruz       -0.029178   0.010447  -2.793              0.00522 **
## adjacent    -0.089277   0.006938 -12.867 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  359.52  on 24  degrees of freedom
## AIC: 532.35
##
## Number of Fisher Scoring iterations: 5
```

The covariates that appear to have a significant effect on species counts are **area**, **nearest**, **scruz**, and **adjacent**. The standard errors for some of these covariates may be too low and making them significant.

i. Evaluate the goodness-of-fit of this model.

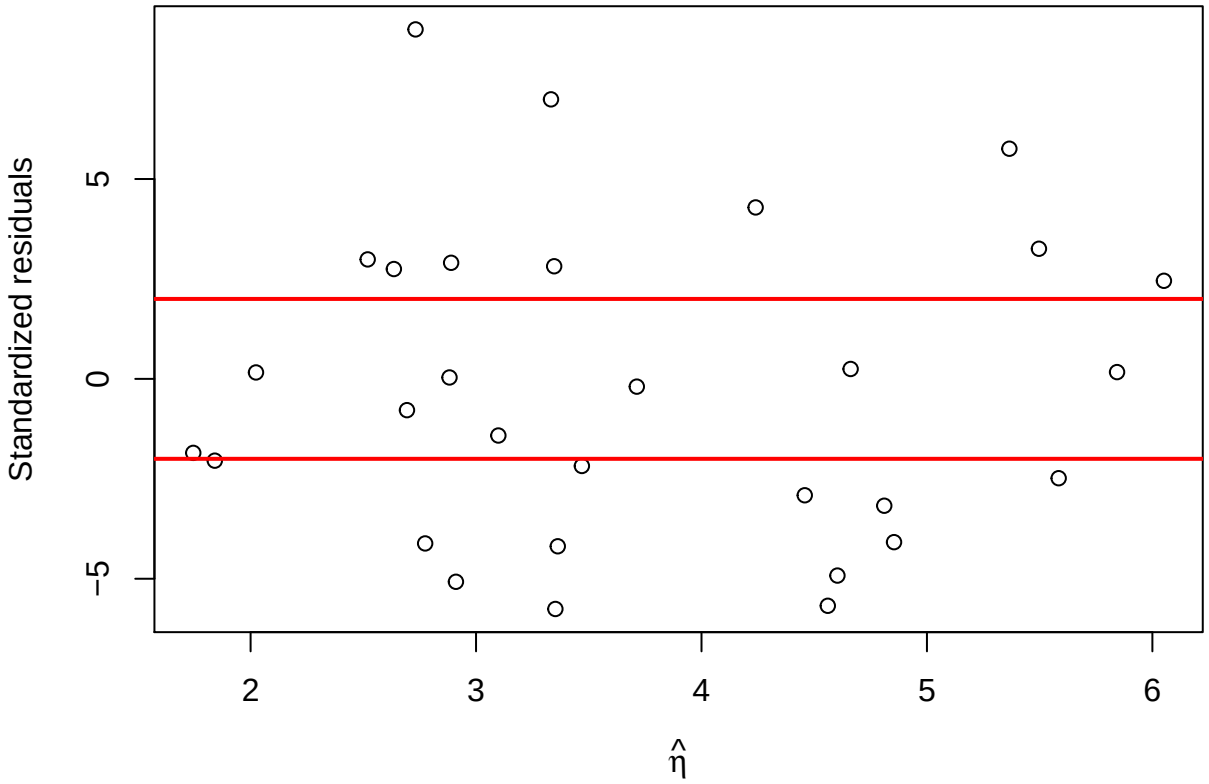
```
1-pchisq( 359.52  , 24  )
```

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

The model does not fit well.

ii. Examine the standardized residuals. Explain whether or not they suggest the presence of overdispersion?

```
par(mar=c(5.1,4.1,1.1, 1.1))
plot(plants.pois$linear.predictors, rstandard(plants.pois), xlab=expression(hat(eta)),
     ylab = "Standardized residuals")
abline(h=-2, col="red",lwd =2);abline(h=2, col="red",lwd =2)
```



The standardized residuals are distributed $r_i \sim N(0,1)$ so we expect about 95% of the points to be within the bounds ± 2 (the red lines). However, it is clear that many points are outside the two bounds which is suggesting the presence of overdispersion.

iii. Fit a negative binomial model with all five covariates on the log scale. Provide the point estimate and 95% confidence interval for the dispersion parameter. Which covariates appear to have a significant effect on species counts?

```
head(plants)
```

```
##   species       area elevation    nearest      scruz   adjacent
## 1      58  3.2224694  5.846439 -0.5108256 -0.5108256  0.6097656
## 2      31  0.2151114  4.691348 -0.5108256  3.2695689  6.3497157
## 3       3 -1.5606477  4.736198  1.0296194  4.0724397 -0.2484614
## 4      25 -2.3025851  3.828641  0.6418539  3.8586222 -1.7147984
## 5       2 -2.9957323  4.343805  0.6418539  0.6418539  6.8066302
## 6      18 -1.0788097  4.779123  2.0794415  2.0794415  0.6097656
```

```
plants.nb = glm.nb(species ~ ., data=plants)
summary(plants.nb)
```

```
##
## Call:
## glm.nb(formula = species ~ ., data = plants, init.theta = 2.944349644,
```

```
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1255  -0.8586  -0.3636   0.5552   1.7166
##
## Coefficients:
##             Estimate Std. Error z value     Pr(>|z|)
## (Intercept)  4.57865    1.26504   3.619     0.000295 ***
## area         0.42207    0.07690   5.488 0.0000000406 ***
## elevation   -0.23047    0.24810  -0.929     0.352911
## nearest     -0.09041    0.08749  -1.033     0.301422
## scruz       -0.02538    0.07254  -0.350     0.726470
## adjacent    -0.03752    0.03590  -1.045     0.295946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.9443) family taken to be 1)
##
##     Null deviance: 149.432  on 29  degrees of freedom
## Residual deviance:  32.801  on 24  degrees of freedom
## AIC: 287.86
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.944
##           Std. Err.:  0.867
##
##  2 x log-likelihood:  -273.861
```

Dispersion parameter:

$$\hat{\gamma} = \frac{1}{2.944} = 0.3396739$$

95% CI:

$$\frac{1}{2.944 \pm 1.96(0.867)} = (0.2153631, 0.8034194)$$

Now only the `area` appears to have a significant effect on species counts.

iv. Use a quasi-likelihood approach with an inflated quadratic function using all five covariates on the log scale. What is the estimated dispersion parameter? Which covariates appear to have a significant effect on species counts?

```
fit.quasi = glm(species ~  area+ elevation +  nearest + scruz + adjacent ,  data=plants,
  family=quasi(link="log", variance="mu^2"))
summary(fit.quasi)
```

```
##
```

```
## Call:
## glm(formula = species ~ area + elevation + nearest + scruz +
##     adjacent, family = quasi(link = "log", variance = "mu^2"),
##     data = plants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4747  -0.5119  -0.2215   0.3158   1.0927
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  4.66229    1.51527   3.077   0.00517 **
## area         0.43177    0.09153   4.717 0.0000853 ***
## elevation   -0.24970    0.29665  -0.842   0.40825
## nearest     -0.08905    0.10560  -0.843   0.40740
## scruz       -0.02698    0.08600  -0.314   0.75649
## adjacent    -0.03670    0.04226  -0.869   0.39371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 0.5251162)
##
##     Null deviance: 56.266  on 29  degrees of freedom
## Residual deviance: 13.954  on 24  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 10
```

Dispersion parameter $\hat{\phi} = \frac{1}{N-p}\sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = 0.5251162$

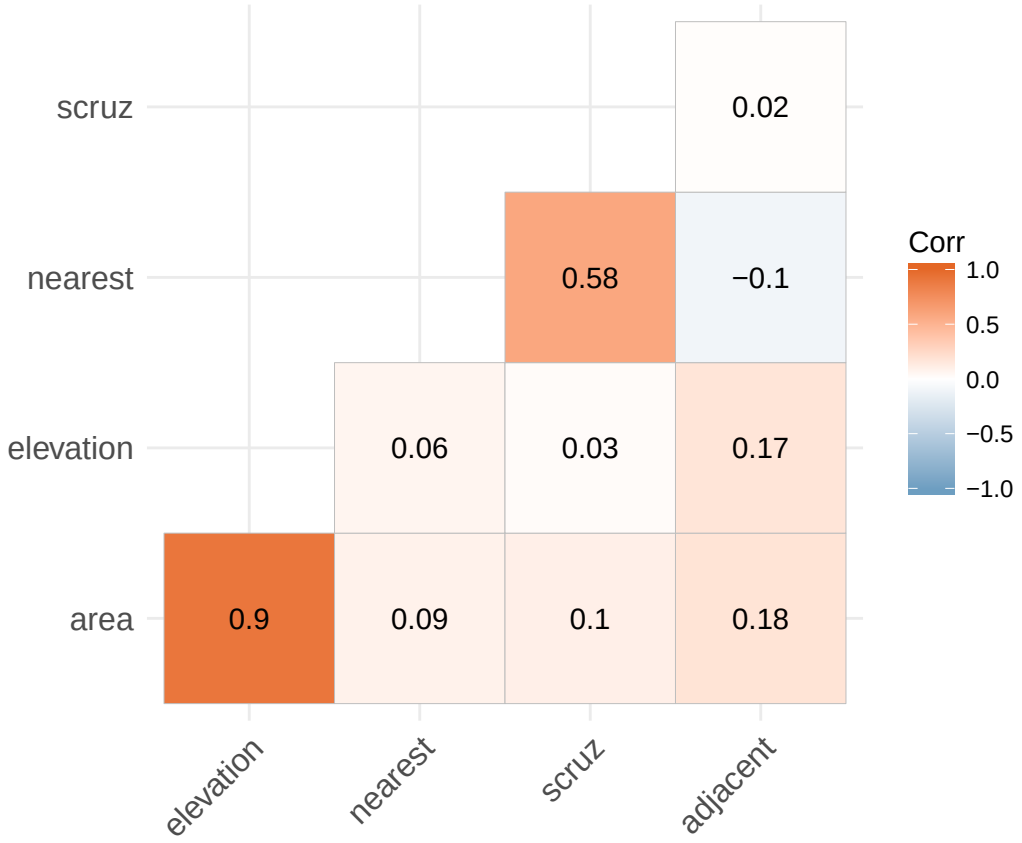
Still, only the area appears to have a significant effect on species counts.

(b) Calculate the pairwise sample correlation between the covariates on the log scale and comment on whether or not multicollinearity may be an issue.

```
library(ggcorrplot,quietly=T);plant.cor<-cor(plants[,-1]);cor(plants[,-1])
```

```
##                area  elevation     nearest      scruz    adjacent
## area      1.00000000 0.90371484  0.08639953 0.10300438  0.17941550
## elevation 0.90371484 1.00000000  0.06384615 0.02798327  0.17068418
## nearest   0.08639953 0.06384615  1.00000000 0.58361035 -0.09617790
## scruz     0.10300438 0.02798327  0.58361035 1.00000000  0.01596249
## adjacent  0.17941550 0.17068418 -0.09617790 0.01596249  1.00000000
```

```
ggcorrplot(plant.cor, type = "lower",lab = T,   colors = c("#6D9EC1", "white", "#E46726"))
```



There is very high positive correlation between elevation and area, and high positive correlations between nearest and scruz. The high degree of multicollinearity among the predictor variables may be responsbile for larger estimated regression coefficients. It may also be responsible for inflated variability of the estimated coefficients; however with an overdispersed model this would seem less of a problem than the larger coefficients. A penalized regression or centering may be worthy of investigation here.

(c) Perform stepwise selection for the Poisson model with all covariates on the log scale.

```
rm(plant);plant<-read.table("Galapagos.txt",header=T);dat<-plant
plant[plant== 0] <- .1; plant<-plant[,-1]; plant[,-1]<-log(plant[,-1]);head(plant,3)
```

```
##   species endemics       area elevation    nearest      scruz   adjacent
## 1      58 3.135494  3.2224694  5.846439 -0.5108256 -0.5108256  0.6097656
## 2      31 3.044522  0.2151114  4.691348 -0.5108256  3.2695689  6.3497157
## 3       3 1.098612 -1.5606477  4.736198  1.0296194  4.0724397 -0.2484614
```

```
pois.null<-glm(species ~ 1,family=poisson, data=plant)
pois.sat<-glm(species ~ . , family=poisson, data=plant)
step(pois.null, scope=list(lower=pois.null,     upper=pois.sat), direction="both",trace=0  )
```

```
##
## Call:  glm(formula = species ~ endemics + adjacent + area + scruz, family = poisson,
##     data = plant)
##
```

```
## Coefficients:
## (Intercept)     endemics     adjacent         area        scruz
##     1.69960      0.70629     -0.04829      0.15404     -0.02564
##
## Degrees of Freedom: 29 Total (i.e. Null);  25 Residual
## Null Deviance:       3511
## Residual Deviance: 239.6     AIC: 410.4
```

```
step.pois<-step(pois.null, scope=list(lower=pois.null,  upper=pois.sat), direction="both",trace=0  )
```

i. Evaluate the goodness-of-fit of the selected Poisson model.

```
1-pchisq(239.6,25)
```

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

It is a poor fitting model.

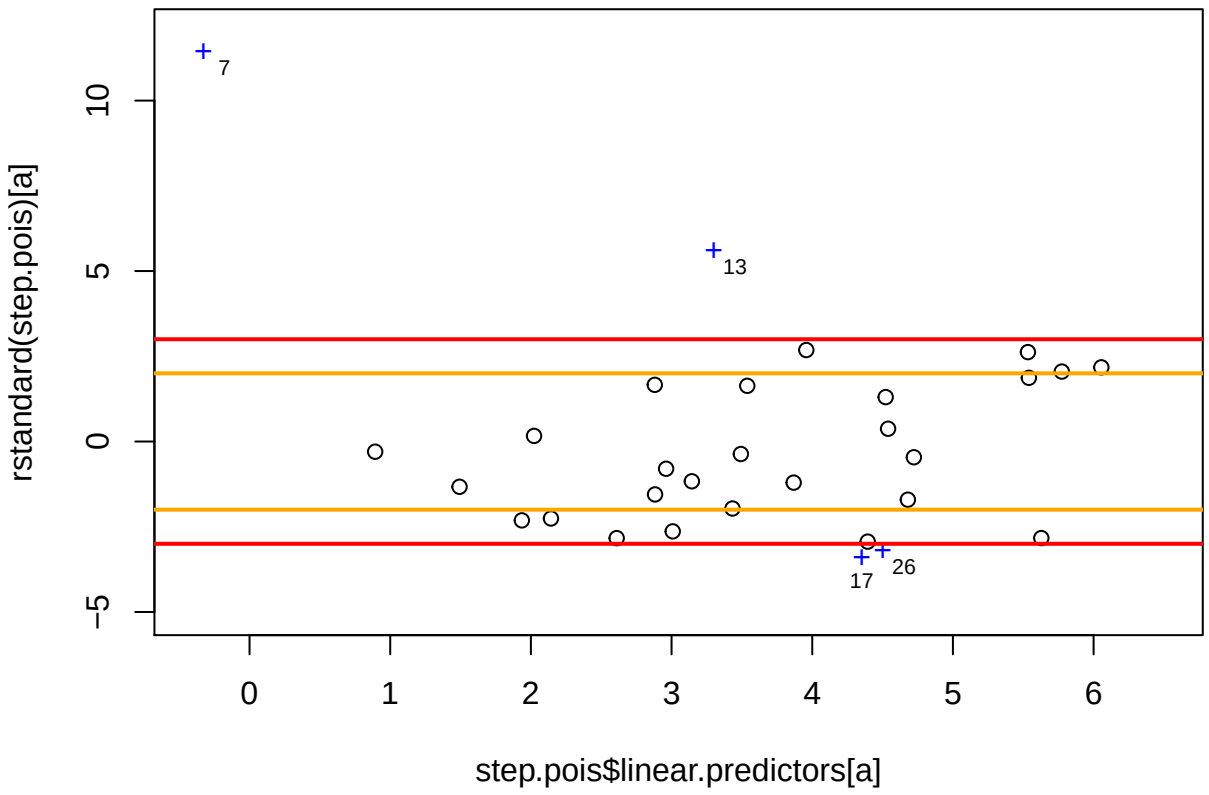
ii. Examine the standardized residuals and identify potential outliers.

Based on this model the outliers are `DaphneMinor`, `Gardner1`, `Marchena`, and `SantaFe`. They have standardized residuals outside the range (-3,3) and are identified as blue + on the plot. The orange lines are at ± 2 to show observations that close to being outliers. The red lines are at ± 3, the values outside this range would be considered outliers.

```
a<-which(abs(rstandard(step.pois))< 3);b<-which(abs(rstandard(step.pois))> 3);dat[b,-c(5:6)]
```

```
##         island species endemics   area scruz adjacent
## 7  DaphneMinor      24        0   0.08  12.0     0.34
## 13    Gardner1      58       17   0.57  93.1    58.27
## 17    Marchena      51       23 129.49  85.9    59.56
## 26     SantaFe      62       28  24.08  16.5     0.52
```

```
par(mar=c(5.1,4.1,1.1, 1.1))
plot(step.pois$linear.predictors[a], rstandard(step.pois)[a],ylim=c(-5,12),xlim=c(-.4,6.5))
points(step.pois$linear.predictors[b], rstandard(step.pois)[b],pch="+",col="blue")
abline(h=2, col="orange",lwd =2);abline(h=-2, col="orange",lwd =2)
abline(h=3, col="red",lwd =2);abline(h=-3, col="red",lwd =2)
text(step.pois$linear.predictors[b[-3]]+.15,  rstandard(step.pois)[b[-3]]-.4,
     labels=paste0(b[-3]), cex= 0.7)
text(step.pois$linear.predictors[b[3]],  rstandard(step.pois)[b[3]]-.6,
     labels=paste0(b[3]), cex= 0.7)
```



(d) Perform stepwise selection for the negative binomial model with all covariates on the log scale.

```
head(plant,3)
```

```
##   species endemics       area elevation    nearest      scruz   adjacent
## 1      58 3.135494  3.2224694  5.846439 -0.5108256 -0.5108256  0.6097656
## 2      31 3.044522  0.2151114  4.691348 -0.5108256  3.2695689  6.3497157
## 3       3 1.098612 -1.5606477  4.736198  1.0296194  4.0724397 -0.2484614
```

```
nb.null<-glm.nb(species ~ 1, data=plant)
nb.sat<-glm.nb(species ~ ., data=plant,maxit=100) ## Warning: glm.fit: algorithm did not converge
step(nb.null, scope=list(lower=nb.null,     upper=nb.sat), direction="both",trace=0  )
```

```
##
## Call:  glm.nb(formula = species ~ area, data = plant, init.theta = 2.533342913,
##     link = log)
##
## Coefficients:
## (Intercept)         area
##      3.2248       0.3499
##
## Degrees of Freedom: 29 Total (i.e. Null);  28 Residual
## Null Deviance:       130.2
```

```
## Residual Deviance: 32.6  AIC: 283.8
```

```
step.nb<-step(nb.null, scope=list(lower=nb.null,    upper=nb.sat), direction="both",trace=0  )
```

i. Evaluate the goodness-of-fit of the selected negative binomial model.

```
1-pchisq(32.604  , 28 )
```

```
## [1] 0.2506323
```

It is an adequate fitting model.

ii. Examine the standardized residuals and identify potential outliers.

There are a few standardized residuals with absolute values near two, with one greater than two (-2.04). But aside from those cases on the edges, there are not any outliers in this model.

```
par(mar=c(5.1,4.1,1.1, 1.1))
plot(step.nb$linear.predictors, rstandard(step.nb),ylim=c(-2.5,2.5),xlim=c(1.5,6.25))
abline(h=2, col="red",lwd =2);abline(h=-2, col="red",lwd =2)
```