EstefanyArgueta-Week3LabReport

E

2/19/2021

Take Home Exercise

We will use a dataset studying population fluctuations of birds in an urban rural study system in OH. The project includes a spot map survey to measure bird territory density at 17 2ha sites from 2005-2010.

```
colrip <- read.csv(file ="Data/colrip.csv", header = TRUE)
str(colrip)
```

```
## 'data.frame':    306 obs. of  5 variables:
##  $ site    : chr  "casto" "cherry" "creeks" "elkrun" ...
##  $ year    : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
##  $ spp     : chr  "ACFL" "ACFL" "ACFL" "ACFL" ...
##  $ mig.stat: chr  "long" "long" "long" "long" ...
##  $ n.terr  : num  1.5 1 1.25 0 0.75 0 0 1 0 1.5 ...
```

```
# Change year from integer to factor
colrip$year <- as.factor(colrip$year)
# Change spp to factor
colrip$spp <- as.factor(colrip$spp)
str(colrip)
```

```
## 'data.frame':    306 obs. of  5 variables:
##  $ site    : chr  "casto" "cherry" "creeks" "elkrun" ...
##  $ year    : Factor w/ 6 levels "2005","2006",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ spp     : Factor w/ 3 levels "ACFL","AMRO",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ mig.stat: chr  "long" "long" "long" "long" ...
##  $ n.terr  : num  1.5 1 1.25 0 0.75 0 0 1 0 1.5 ...
```

```
levels(colrip$year)
```

```
## [1] "2005" "2006" "2007" "2008" "2009" "2010"
```

```
levels(colrip$spp)
```

```
## [1] "ACFL" "AMRO" "NOCA"
```

```
# The references are going to 2005 and ACFL
```

Looking at the data, we have a continuous variable - the mean number of territories over all years for each species `n terr` and two factors, three migratory categories (`mig.stat`) and three species of birds (`spp`).

> We are interested in whether the number of territories at a site differs between our three species.

The three species are: Acadian flycatcher (ACFL), Northern cardinal (NOCA) and American Robin (AMRO).

We want to see if the populations of these birds differ in sites throughout the county.

Alternately, the differences might be explained by the regional climate differences between years (2005-2010).

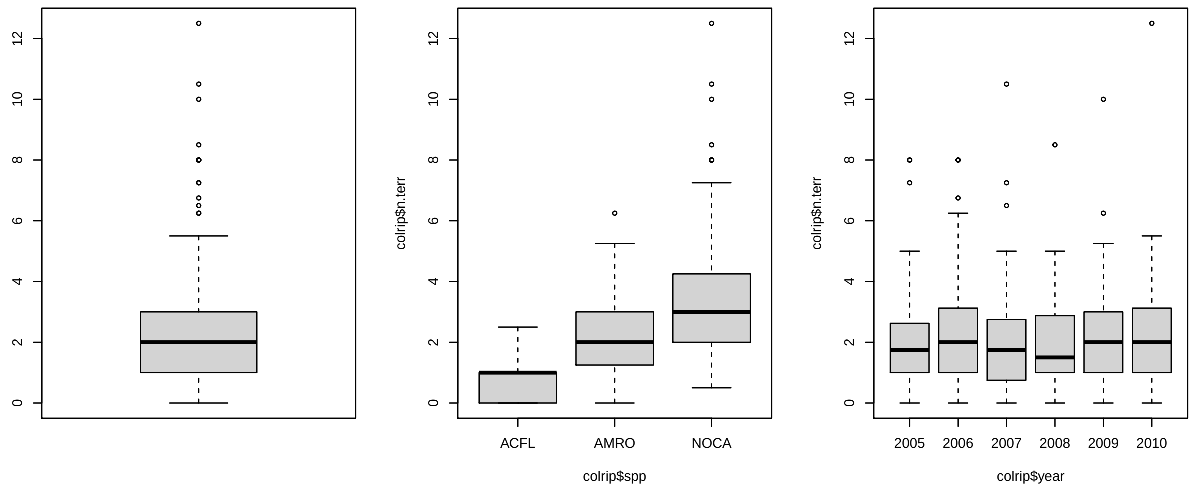
Our *null* hypotheses are:

- H_0: There are no differences in numbers of territories between species, or year
- H_{spp}: The number of breeding territories at a site differs significantly between species
- H_{yr}: The number of breeding territories at a site differ significantly between years

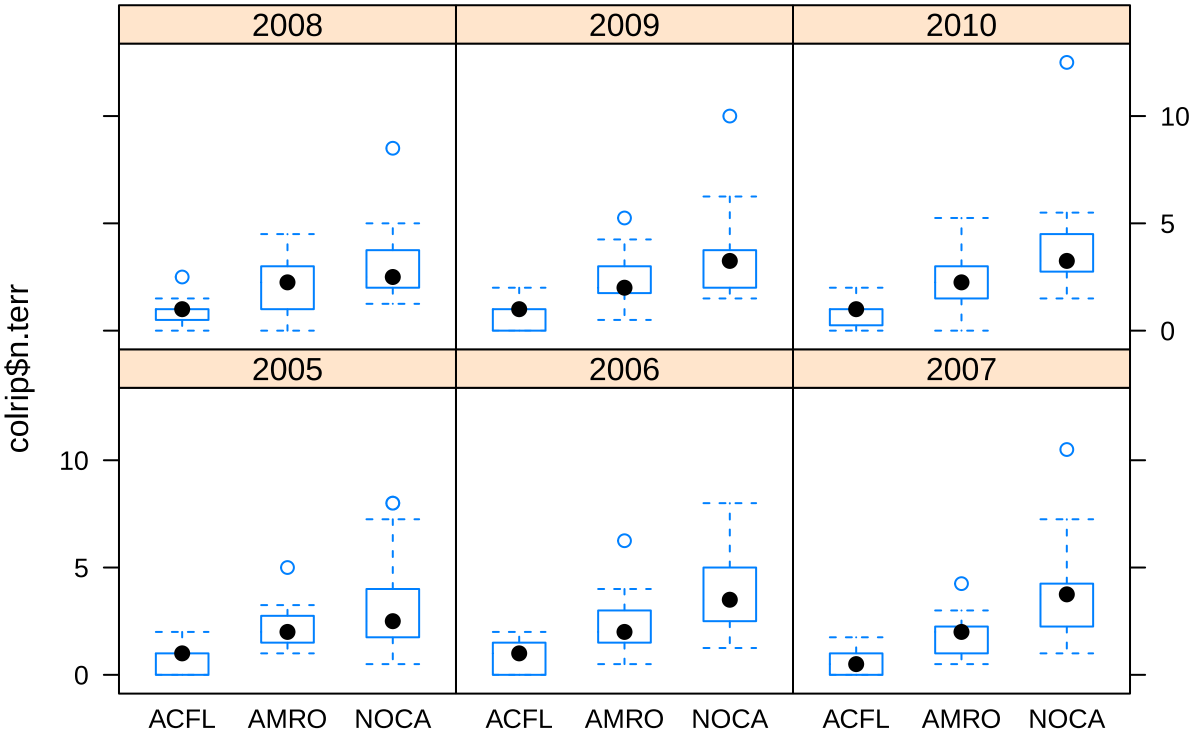
> **Using the process we went thorugh in class: Formulate a statisical model for each hypothesis, fit the model, evaluate the fit of those models and interpret the model.**

Before anything, lets start with data exploration. Data Exploration

```
par(mfrow=c(1,3))
boxplot(colrip$n.terr)
boxplot(colrip$n.terr ~ colrip$spp)
boxplot(colrip$n.terr ~ colrip$year)
```



```
par(mfrow=c(1,1))
bwplot(colrip$n.terr ~ colrip$spp | colrip$year)
```



We can see potential outliers in the plots. Lets go ahead with our models and then check with CooksD

Formulate and fit a statistical model for each hypothesis

# of Territories between Spp or Year

```
mSppYear <- glm(n.terr ~ spp + year, data = colrip, family = "gaussian")
coef(mSppYear)
```

```
## (Intercept)     sppAMRO     sppNOCA    year2006    year2007    year2008
##  0.62418301  1.46078431  2.81372549  0.39705882  0.04411765 -0.01960784
##    year2009    year2010
##  0.13235294  0.27450980
```

```
tapply(colrip$n.terr, list(colrip$spp, colrip$year), mean, na.rm = T)
```

```
##           2005      2006      2007      2008      2009      2010
## ACFL 0.7058824 0.8235294 0.5882353 0.8382353 0.7647059 0.8529412
## AMRO 2.2205882 2.3676471 1.9411765 2.1470588 2.3235294 2.3382353
## NOCA 3.2205882 4.1470588 3.7500000 3.1029412 3.4558824 3.7794118
```

# of Territories between Spp

```
mSpp <- glm(n.terr ~ spp, data = colrip, family = "gaussian")
coef(mSpp)
```

```
## (Intercept)     sppAMRO     sppNOCA
##   0.7622549   1.4607843   2.8137255
```

```
tapply(colrip$n.terr, list(colrip$spp), mean, na.rm = T)
```

```
##      ACFL      AMRO      NOCA
## 0.7622549 2.2230392 3.5759804
```

# of Territories between years

```
mYr <- lm(n.terr ~ year, data = colrip)
coef(mYr)
```

```
## (Intercept)    year2006    year2007    year2008    year2009    year2010
##  2.04901961  0.39705882  0.04411765 -0.01960784  0.13235294  0.27450980
```

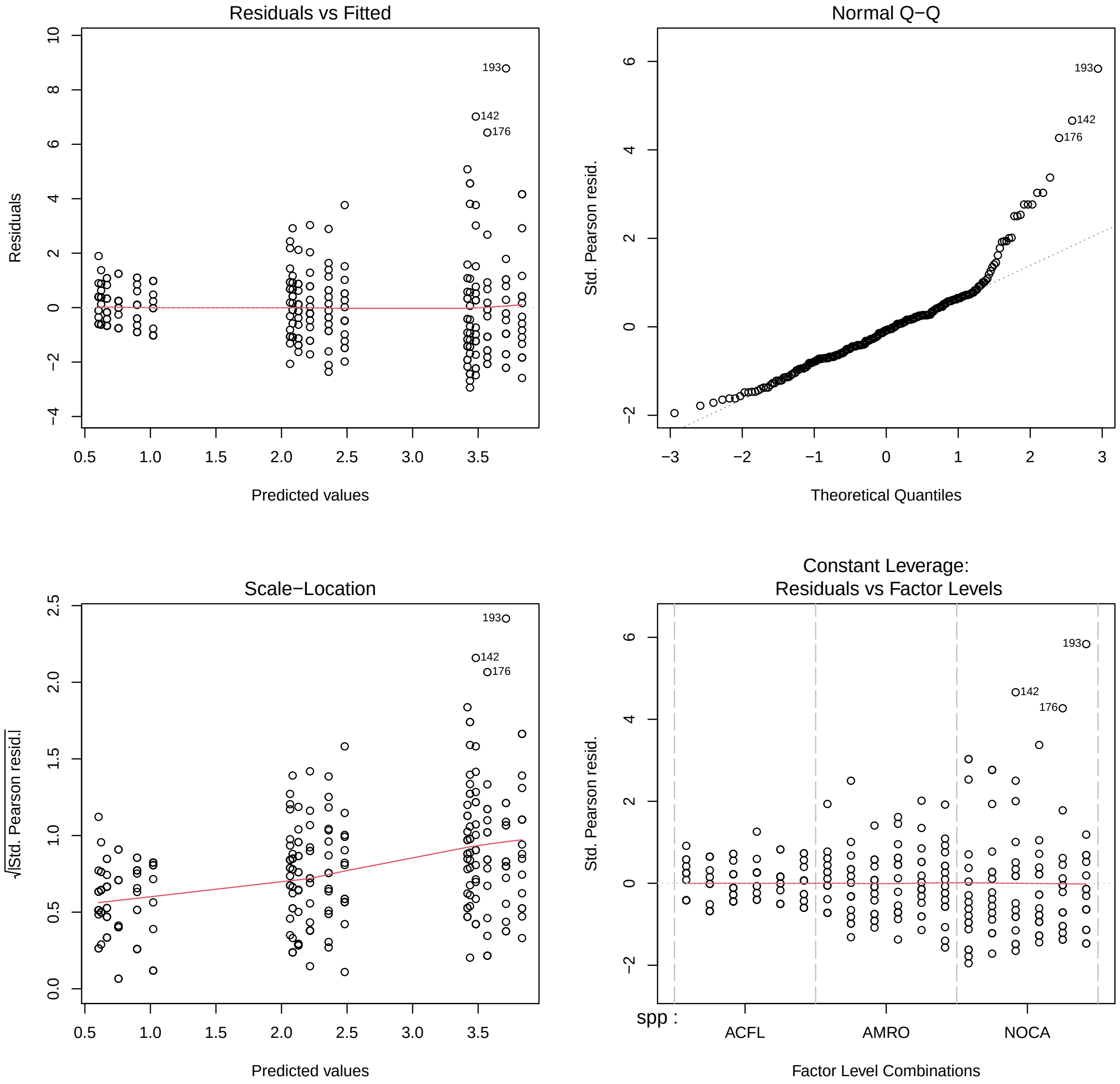
```
tapply(colrip$n.terr, list(colrip$year), mean, na.rm = T)
```

```
##     2005     2006     2007     2008     2009     2010
## 2.049020 2.446078 2.093137 2.029412 2.181373 2.323529
```

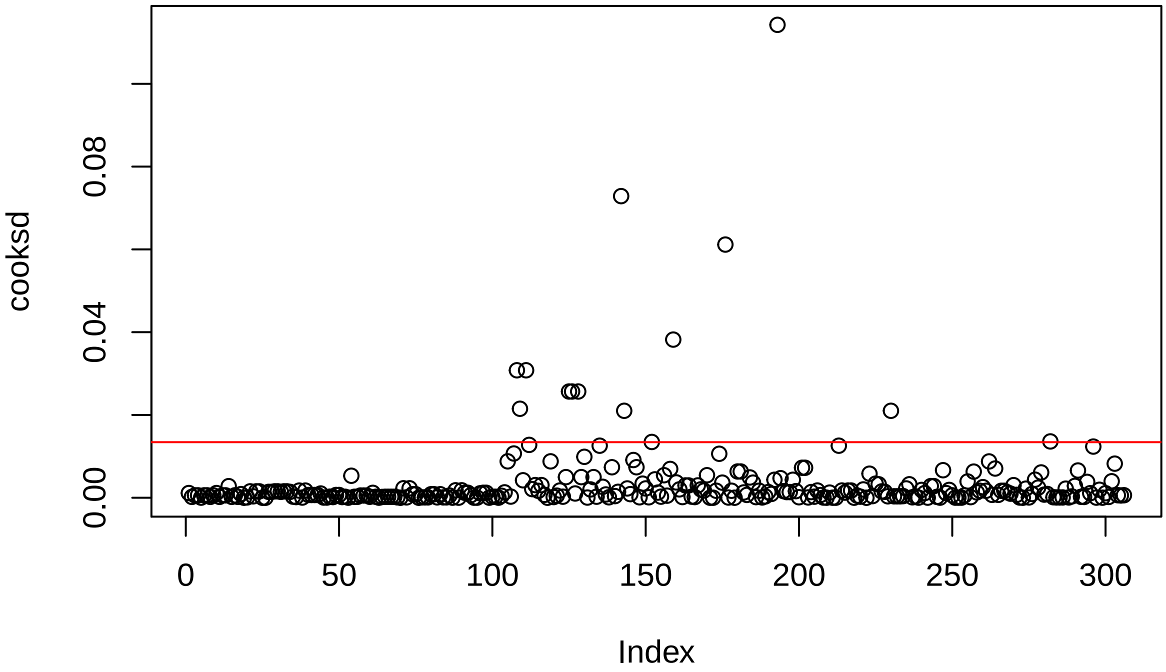
> Evaluate the fit of these models. Do they fulfill the assumptions of a linear model? If not, what can we do about it?

# of Territories between Spp or Year

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

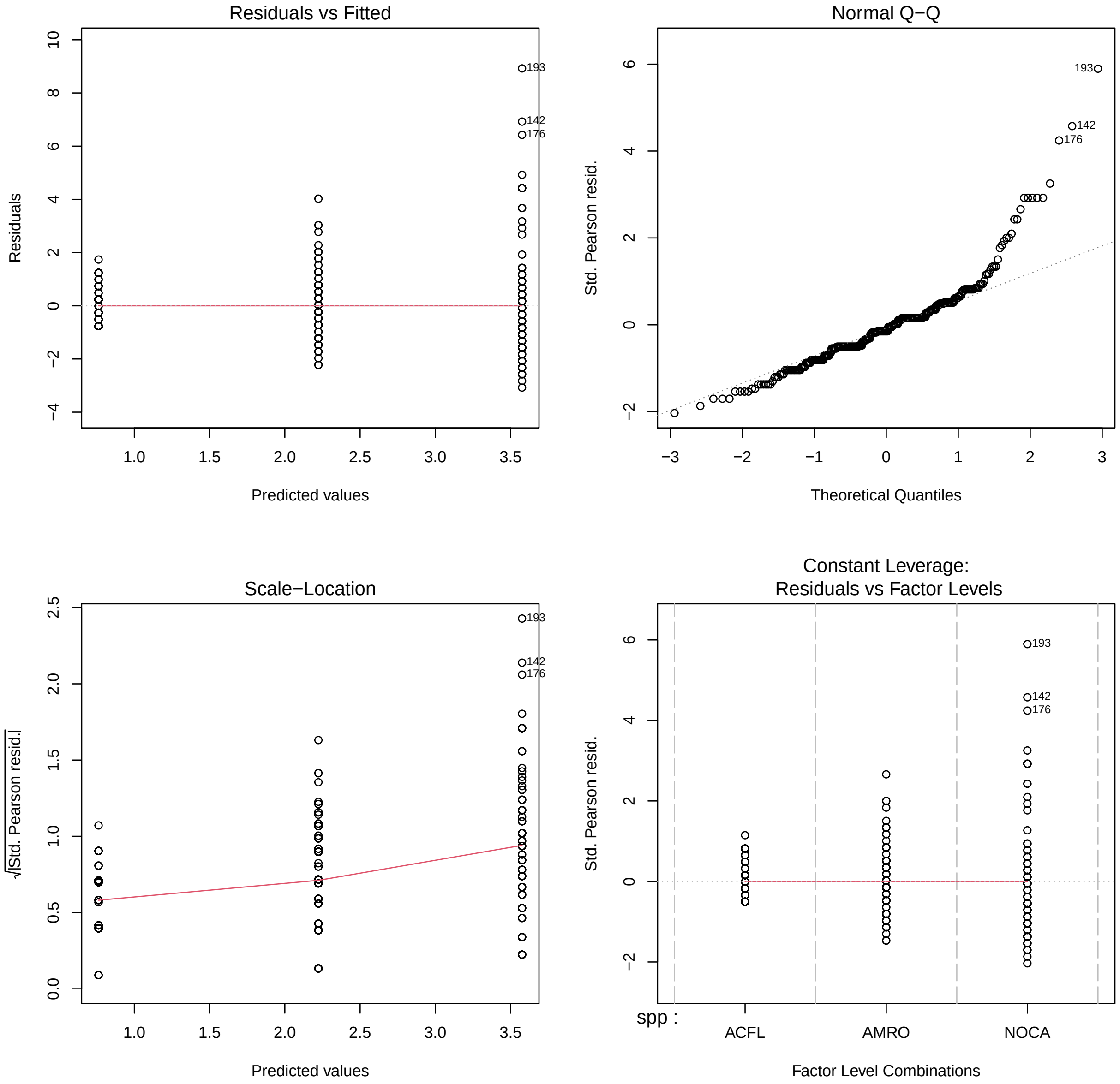


```
par(mfrow = c(1,1))
cooksd <- cooks.distance(mSppYear)
plot(cooksd)
abline(h = 4*mean(cooksd, na.rm = T), col = "red")
```

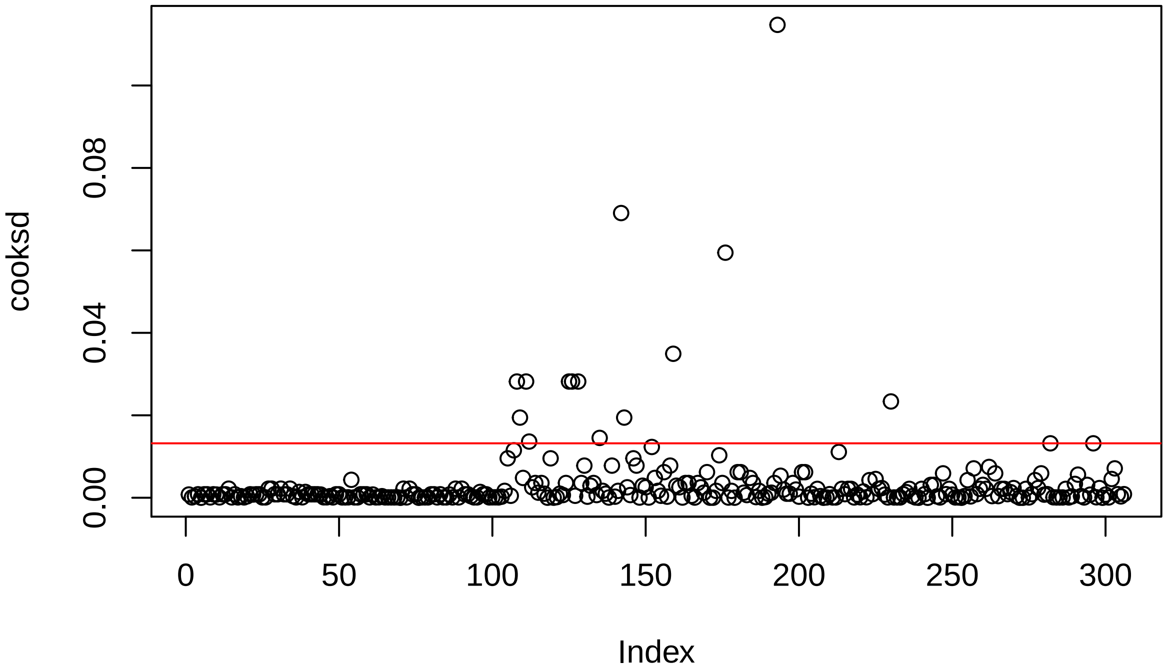


# of Territories between Spp

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

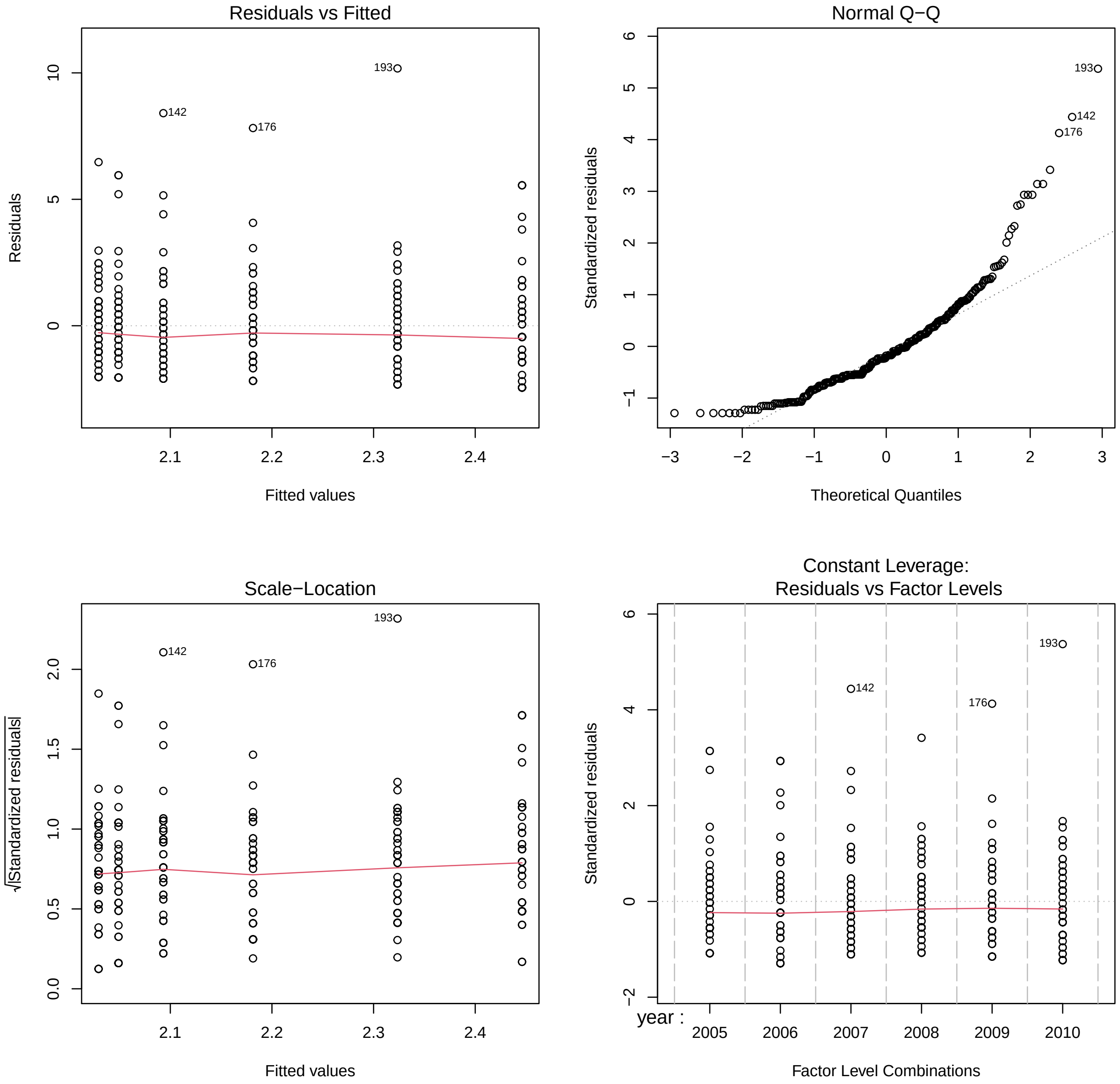


```
par(mfrow = c(1,1))
cooksd <- cooks.distance(mSpp)
plot(cooksd)
abline(h = 4*mean(cooksd, na.rm = T), col = "red")
```

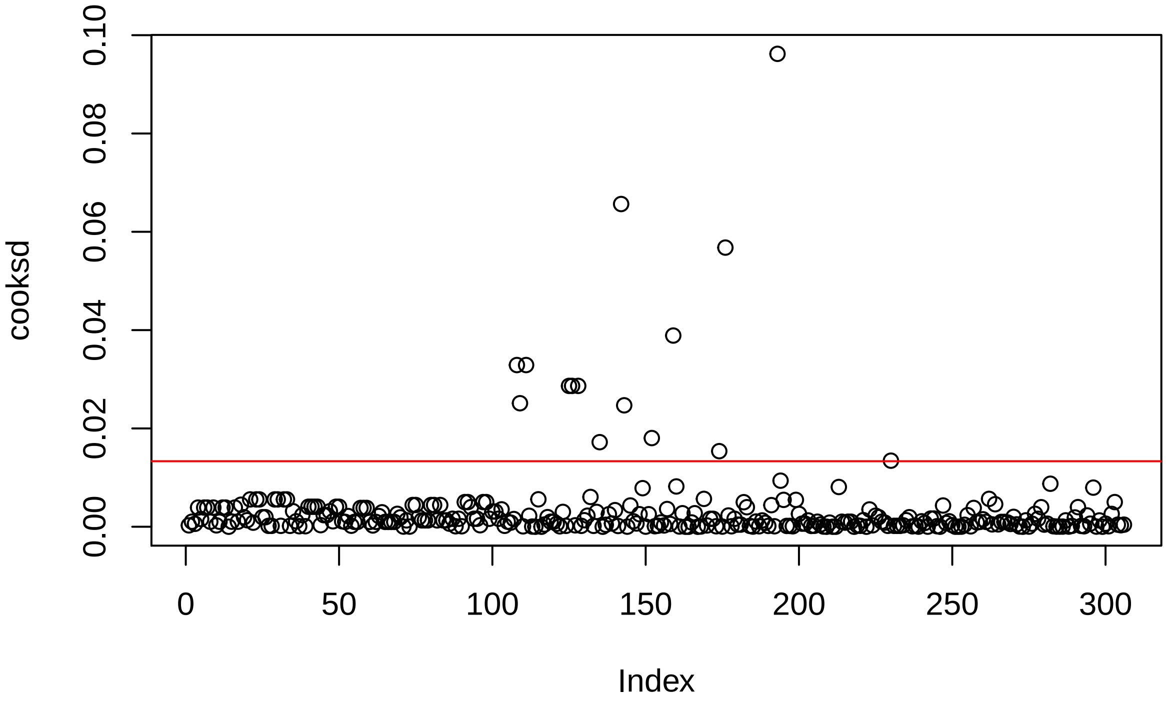


# of Territories between Years

```r
par(mfrow = c(2,2))
plot(mYr)
```



```
par(mfrow = c(1,1))
cooksd <- cooks.distance(mYr)
plot(cooksd)
abline(h = 4*mean(cooksd, na.rm = T), col = "red")
```



By examining the fit of the three models, it is clear by the Q-Q plot that our data is not normally distributed.

By examining the `Residuals vs Fitted` plots, we see that there is an increase in the spread of the residuals for the larger values of the fitted values. This indicates a violation of the homogeneity assumption

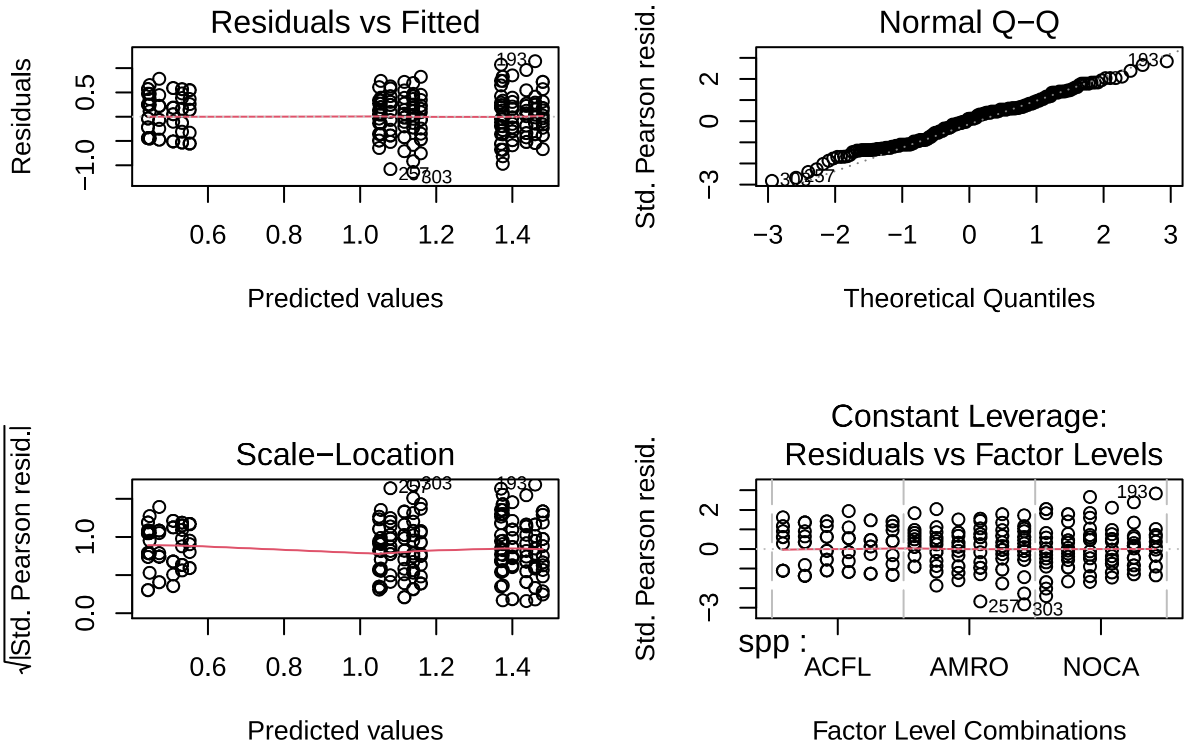
We would also be able to look for violations by creating a histogram of the residuals and a conditional boxplot.

We can transform the data to look more normal. We will create a new column for the log transformed `n.terr`

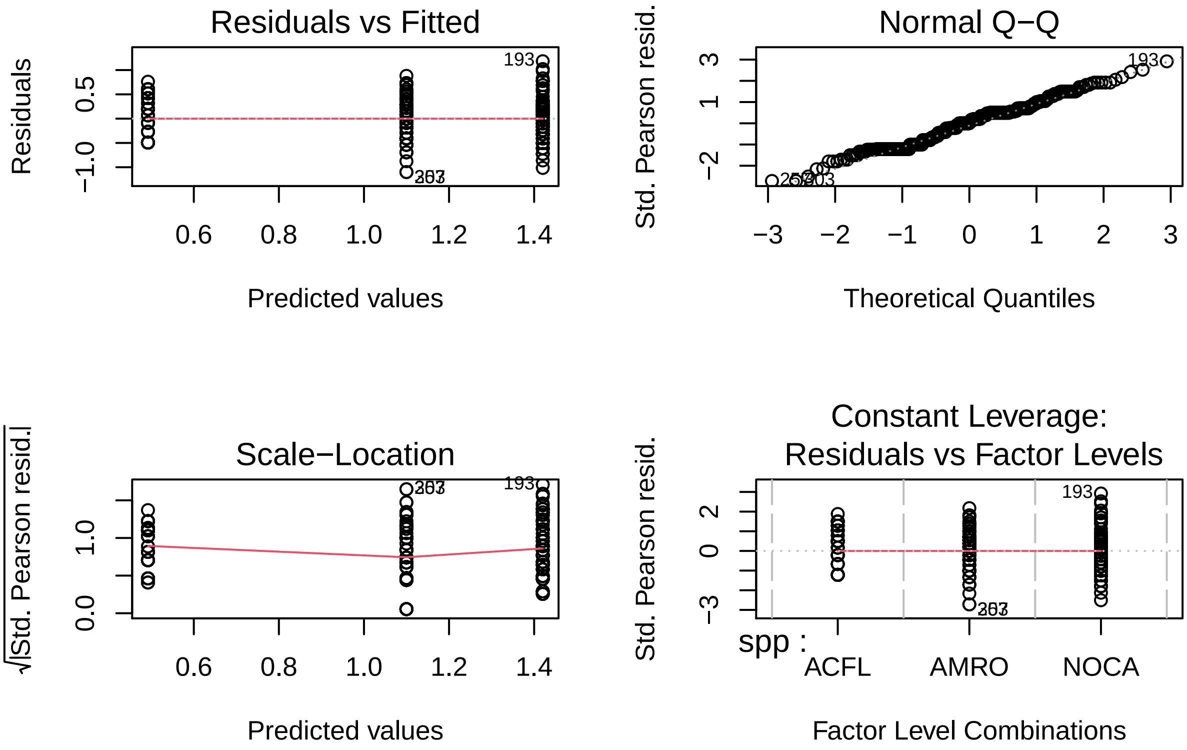
```
range(colrip$n.terr)
```

```
## [1]  0.0 12.5
```

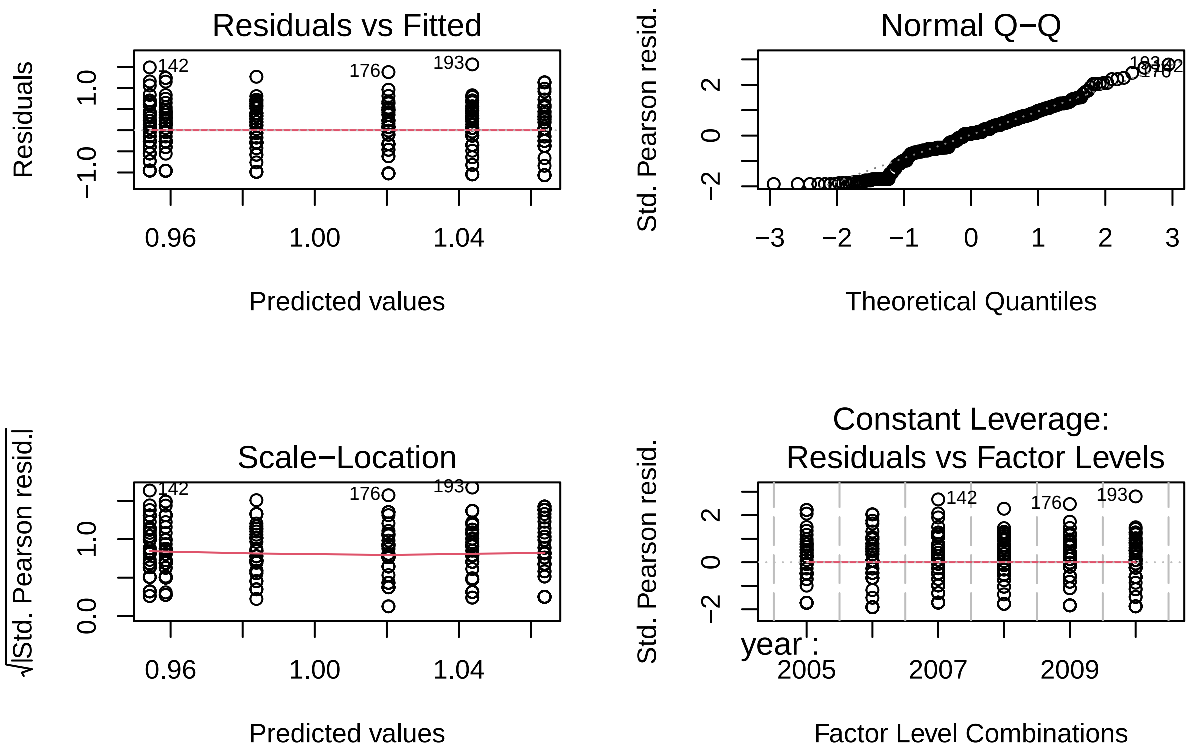
```
# We see that there are zeros in our data.
#We will have to add 1 to our dataset in order for the transformation to work.
# Lets Log Transform
colrip$logterr <- log(colrip$n.terr + 1)

# Updating models
log.MSppYear <- glm(logterr ~ spp + year, data = colrip, family = "gaussian")
log.MSpp <- glm(logterr ~ spp, data = colrip, family = "gaussian")
log.MYear <- glm(logterr ~ year, data = colrip, family = "gaussian")

par(mfrow = c(2,2))

plot(log.MSppYear)
```



```
plot(log.MSpp)
```



```
plot(log.MYear)
```



Now our data looks more normal and we can proceed with choosing the best model.

> Using model selection, with AIC, compare the three competing hypotheses and interpret the output, which model best explains the variation in your data?

```
library(AICcmodavg)
log.mod0 <- glm(logterr ~ 1, colrip, family = "gaussian")

models <- list()
models[[1]] <- log.mod0
models[[2]] <- log.MSppYear
models[[3]] <- log.MSpp
models[[4]] <- log.MYear

names(models) <- c("Null", "Species and Year", "Species", "Year")

aictab(models)
```

```
##
## Model selection based on AICc:
##
##                  K   AICc Delta_AICc AICcWt Cum.Wt      LL
## Species          4 323.17       0.00   0.97   0.97 -157.52
## Species and Year 9 330.40       7.23   0.03   1.00 -155.90
## Null             2 516.07     192.91   0.00   1.00 -256.02
## Year             7 524.71     201.54   0.00   1.00 -255.17
```

We notice that the best model is `log.MSpp`. The lowest AIC score indicates the most parsimonious model.

> Is there a significant effect of species or year on number of breeding territories in this study system?

Lets examine our models further

```
summary(log.MSppYear)
```

```
##
## Call:
## glm(formula = logterr ~ spp + year, family = "gaussian", data = colrip)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1395  -0.3583   0.0442   0.2464   1.1427
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.446746   0.065987   6.770 6.84e-11 ***
## sppAMRO      0.607595   0.057146  10.632  < 2e-16 ***
## sppNOCA      0.928083   0.057146  16.241  < 2e-16 ***
## year2006     0.105155   0.080817   1.301    0.194
## year2007    -0.004431   0.080817  -0.055    0.956
## year2008     0.025190   0.080817   0.312    0.755
## year2009     0.061840   0.080817   0.765    0.445
## year2010     0.085110   0.080817   1.053    0.293
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.1665498)
##
##     Null deviance: 95.490  on 305  degrees of freedom
## Residual deviance: 49.632  on 298  degrees of freedom
## AIC: 329.79
##
## Number of Fisher Scoring iterations: 2
```

The reference level is species ACFL in year 2005. The only significant values are between the different species.

```
summary(log.MSpp)
```

```
##
## Call:
## glm(formula = logterr ~ spp, family = "gaussian", data = colrip)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0998  -0.3217   0.0127   0.2339   1.1824
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.49222    0.04029   12.22   <2e-16 ***
## sppAMRO      0.60760    0.05697   10.66   <2e-16 ***
## sppNOCA      0.92808    0.05697   16.29   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.165547)
##
##     Null deviance: 95.490  on 305  degrees of freedom
## Residual deviance: 50.161  on 303  degrees of freedom
## AIC: 323.03
##
## Number of Fisher Scoring iterations: 2
```

The reference level for this model is the ACFL species. There is a significant difference between the AMRO and NOCA species with the reference level.

```
summary(log.MYear)
```

```
##
## Call:
## glm(formula = logterr ~ year, family = "gaussian", data = colrip)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.06379  -0.32733   0.04389   0.36582   1.55894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.958639   0.078782  12.168   <2e-16 ***
## year2006     0.105155   0.111415   0.944    0.346
## year2007    -0.004431   0.111415  -0.040    0.968
## year2008     0.025190   0.111415   0.226    0.821
## year2009     0.061840   0.111415   0.555    0.579
## year2010     0.085110   0.111415   0.764    0.446
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.3165381)
##
##     Null deviance: 95.490  on 305  degrees of freedom
## Residual deviance: 94.961  on 300  degrees of freedom
## AIC: 524.34
##
## Number of Fisher Scoring iterations: 2
```

The reference level for this model is the year 2005. There are no significant differences between years 2006-2010 and year 2005.

Fully interpret your *best* model and “report” your findings.

Our best model was H_{spp}: The number of breeding territories at a site differs significantly between species.

Based on this output, we see that there are differences between the AMRO and NOCA species from the ACFL species. The two species are have a larger number of territories than the ACFL species. The differences are significant.

We also notice that the residual deviance is lower than our null deviance, indicating that we have a good global model.

```
aov.mSpp <- aov(log.MSpp)
tuk.mSpp <- TukeyHSD(aov.mSpp)
tuk.mSpp
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log.MSpp)
##
## $spp
##                diff       lwr       upr p adj
## AMRO-ACFL 0.6075952 0.4734053 0.7417851 0e+00
## NOCA-ACFL 0.9280834 0.7938935 1.0622733 0e+00
## NOCA-AMRO 0.3204882 0.1862983 0.4546782 1e-07
```

By using the Tukey HSD pairwise comparisons, we see that there are significant differences all three species.

> What about differences between species each year? That also seems like a reasonable hypothesis doesn't it. Do any of these models test that?

The `log.MSppYear` model tests the hypothesis between species each year.

If we wanted to explore this model further we can run a TUkey HSD pariwaise comparison:

```
aov.mSppYear <- aov(log.MSppYear)
tuk.mSppYear <- TukeyHSD(aov.mSppYear)
tuk.mSppYear
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log.MSppYear)
##
## $spp
##                diff       lwr       upr p adj
## AMRO-ACFL 0.6075952 0.4729883 0.7422021 0e+00
## NOCA-ACFL 0.9280834 0.7934765 1.0626903 0e+00
## NOCA-AMRO 0.3204882 0.1858814 0.4550951 1e-07
##
## $year
##                  diff        lwr       upr     p adj
## 2006-2005  0.10515520 -0.1266567 0.3369671 0.7842956
## 2007-2005 -0.00443077 -0.2362426 0.2273811 0.9999999
## 2008-2005  0.02518993 -0.2066219 0.2570018 0.9996051
## 2009-2005  0.06183960 -0.1699723 0.2936515 0.9730873
## 2010-2005  0.08511022 -0.1467017 0.3169221 0.8992747
## 2007-2006 -0.10958597 -0.3413978 0.1222259 0.7531981
## 2008-2006 -0.07996527 -0.3117771 0.1518466 0.9210813
## 2009-2006 -0.04331560 -0.2751275 0.1884963 0.9946588
## 2010-2006 -0.02004498 -0.2518569 0.2117669 0.9998714
## 2008-2007  0.02962070 -0.2021912 0.2614326 0.9991304
## 2009-2007  0.06627037 -0.1655415 0.2980822 0.9637009
## 2010-2007  0.08954099 -0.1422709 0.3213529 0.8778835
## 2009-2008  0.03664967 -0.1951622 0.2684615 0.9975766
## 2010-2008  0.05992029 -0.1718916 0.2917322 0.9765630
## 2010-2009  0.02327062 -0.2085412 0.2550825 0.9997321
```

We notice there are no differences between years. There is a significant difference between species.