Lab 5

Kent Codding

2023-10-05

library(MASS)  
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.3

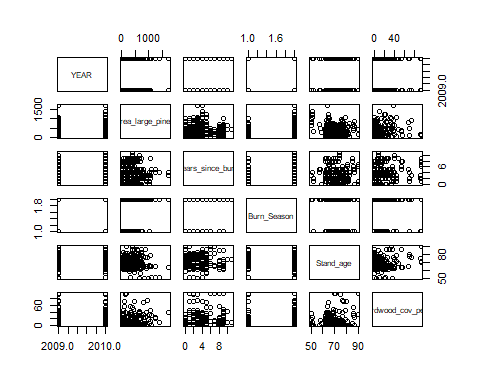
##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

setwd("C:/Users/Kent Codding/Desktop/Adv Biostat/Lab 5")  
df <- read.csv("BACS.csv", row.names = 1)

# Question 1

plot(df[-1]) #plots all potential explanatory variables (not including response var - BACS)

 no obvious collinearity, but considering biologicical phenomena, it would make sense that hardwood cover percentage increases with years after burn. This also may confound with Burn\_Season.

# use pearson's r to check  
cor(df$hardwood\_cov\_perc, df$years\_since\_burn) #not collinear

## [1] 0.2281264

# cor(df$hardwood\_cov\_perc, df$Burn\_Season) not numeric

My hypothesis was incorrect. The hardwood cover percentage does not correlate enough with years after burn for it to be an issue # Question 2 justification for error distribution: since we are looking at count data, which is a poisson process, I will start with a poisson distribution to determine whether the residual deviance displays overdispersion.

mod1p <- glm(BACS ~ ., data = df, family = "poisson")   
summary(mod1p)

##   
## Call:  
## glm(formula = BACS ~ ., family = "poisson", data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.6048 -1.4531 -0.7464 0.4029 5.8810   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.061e+02 1.538e+02 -1.991 0.0465 \*   
## YEAR 1.518e-01 7.653e-02 1.983 0.0473 \*   
## area\_large\_pines 6.325e-04 1.533e-04 4.126 3.70e-05 \*\*\*  
## years\_since\_burn -1.414e-01 1.905e-02 -7.424 1.14e-13 \*\*\*  
## Burn\_SeasonNGS -4.950e-01 7.951e-02 -6.226 4.78e-10 \*\*\*  
## Stand\_age 3.879e-02 5.660e-03 6.854 7.19e-12 \*\*\*  
## hardwood\_cov\_perc -5.946e-02 5.661e-03 -10.502 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1264.99 on 286 degrees of freedom  
## Residual deviance: 788.05 on 280 degrees of freedom  
## AIC: 1272.9  
##   
## Number of Fisher Scoring iterations: 6

mod1p$deviance/mod1p$df.residual > 1.5 #if TRUE, then overdispersion is apparent

## [1] TRUE

Yes, there is overdispersion :( now try negative binomial.

mod1nb <- glm.nb(BACS ~ ., data = df)   
summary(mod1nb)

##   
## Call:  
## glm.nb(formula = BACS ~ ., data = df, init.theta = 0.9038291142,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0593 -0.9502 -0.5161 0.1470 3.1330   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.141e+02 3.249e+02 -0.659 0.510013   
## YEAR 1.060e-01 1.617e-01 0.655 0.512216   
## area\_large\_pines 4.645e-04 3.064e-04 1.516 0.129481   
## years\_since\_burn -2.010e-01 3.614e-02 -5.561 2.69e-08 \*\*\*  
## Burn\_SeasonNGS -6.504e-01 1.683e-01 -3.864 0.000111 \*\*\*  
## Stand\_age 4.257e-02 1.149e-02 3.704 0.000212 \*\*\*  
## hardwood\_cov\_perc -5.567e-02 7.836e-03 -7.105 1.21e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.9038) family taken to be 1)  
##   
## Null deviance: 441.35 on 286 degrees of freedom  
## Residual deviance: 277.63 on 280 degrees of freedom  
## AIC: 1028.4  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.904   
## Std. Err.: 0.138   
##   
## 2 x log-likelihood: -1012.437

mod1nb$deviance / mod1nb$df.residual

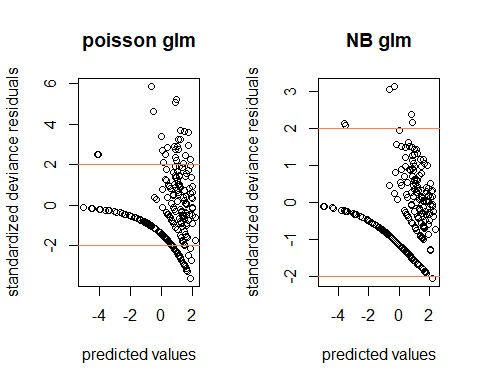
## [1] 0.9915485

lrtest(mod1p,mod1nb)

## Likelihood ratio test  
##   
## Model 1: BACS ~ YEAR + area\_large\_pines + years\_since\_burn + Burn\_Season +   
## Stand\_age + hardwood\_cov\_perc  
## Model 2: BACS ~ YEAR + area\_large\_pines + years\_since\_burn + Burn\_Season +   
## Stand\_age + hardwood\_cov\_perc  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 7 -629.45   
## 2 8 -506.22 1 246.46 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The residual deviance divided by degrees of freedom is much closer to 1. Now plot the residuals for each model. Log likelihood is also higher for the negative binomial model

#poisson  
par(mfrow = c(1,2))  
plot(predict(mod1p),residuals.glm(mod1p),   
 main = "poisson glm",  
 xlab = "predicted values",  
 ylab = "standardized deviance residuals")  
#ideally deviance residuals will fall between -2 and 2  
abline(h = c(2,-2), col = "coral")  
  
#now the NB model  
plot(predict(mod1nb),residuals.glm(mod1nb), main = "NB glm",  
 xlab = "predicted values",  
 ylab = "standardized deviance residuals")  
abline(h = c(2,-2), col = "coral")

 Now that I have explored the residual distributions of the two models, I have selected the negative binomial model as it performed better on each of the three tests: deviance/degrees of freedom, log likelihood, and deviance falling between (mostly) +/- 2 for predicted values.

# Question 3

#drop 1 method  
drop1(mod1nb, test = "F")

## Single term deletions  
##   
## Model:  
## BACS ~ YEAR + area\_large\_pines + years\_since\_burn + Burn\_Season +   
## Stand\_age + hardwood\_cov\_perc  
## Df Deviance AIC F value Pr(>F)   
## <none> 277.63 1026.4   
## YEAR 1 278.06 1024.9 0.4327 0.5111864   
## area\_large\_pines 1 279.56 1026.4 1.9471 0.1640067   
## years\_since\_burn 1 305.28 1052.1 27.8839 2.589e-07 \*\*\*  
## Burn\_Season 1 291.18 1038.0 13.6593 0.0002636 \*\*\*  
## Stand\_age 1 290.43 1037.2 12.9089 0.0003862 \*\*\*  
## hardwood\_cov\_perc 1 354.41 1101.2 77.4356 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod2 <- glm.nb(BACS ~ area\_large\_pines + years\_since\_burn + Burn\_Season +   
 Stand\_age + hardwood\_cov\_perc, data = df)  
summary(mod2)

##   
## Call:  
## glm.nb(formula = BACS ~ area\_large\_pines + years\_since\_burn +   
## Burn\_Season + Stand\_age + hardwood\_cov\_perc, data = df, init.theta = 0.898220356,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0383 -0.9519 -0.5068 0.1632 3.0510   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.1221376 0.8650820 -1.297 0.194581   
## area\_large\_pines 0.0004661 0.0003074 1.516 0.129429   
## years\_since\_burn -0.2005907 0.0361936 -5.542 2.99e-08 \*\*\*  
## Burn\_SeasonNGS -0.6615595 0.1685991 -3.924 8.71e-05 \*\*\*  
## Stand\_age 0.0429774 0.0115124 3.733 0.000189 \*\*\*  
## hardwood\_cov\_perc -0.0559441 0.0078581 -7.119 1.08e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.8982) family taken to be 1)  
##   
## Null deviance: 439.89 on 286 degrees of freedom  
## Residual deviance: 277.20 on 281 degrees of freedom  
## AIC: 1026.9  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.898   
## Std. Err.: 0.137   
##   
## 2 x log-likelihood: -1012.865

drop1(mod2, test = "F")

## Single term deletions  
##   
## Model:  
## BACS ~ area\_large\_pines + years\_since\_burn + Burn\_Season + Stand\_age +   
## hardwood\_cov\_perc  
## Df Deviance AIC F value Pr(>F)   
## <none> 277.20 1024.9   
## area\_large\_pines 1 279.13 1024.8 1.9568 0.1629621   
## years\_since\_burn 1 304.74 1050.4 27.9108 2.551e-07 \*\*\*  
## Burn\_Season 1 291.28 1036.9 14.2691 0.0001934 \*\*\*  
## Stand\_age 1 290.24 1035.9 13.2211 0.0003292 \*\*\*  
## hardwood\_cov\_perc 1 354.47 1100.1 78.3228 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod3 <- glm.nb(BACS ~ years\_since\_burn + Burn\_Season +   
 Stand\_age +   
 hardwood\_cov\_perc,  
 data = df)  
summary(mod3)

##   
## Call:  
## glm.nb(formula = BACS ~ years\_since\_burn + Burn\_Season + Stand\_age +   
## hardwood\_cov\_perc, data = df, init.theta = 0.8723621238,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0680 -0.9500 -0.5231 0.1751 2.9999   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.946845 0.863857 -1.096 0.273050   
## years\_since\_burn -0.206292 0.036496 -5.652 1.58e-08 \*\*\*  
## Burn\_SeasonNGS -0.569869 0.165211 -3.449 0.000562 \*\*\*  
## Stand\_age 0.042285 0.011662 3.626 0.000288 \*\*\*  
## hardwood\_cov\_perc -0.054164 0.007611 -7.116 1.11e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.8724) family taken to be 1)  
##   
## Null deviance: 433.06 on 286 degrees of freedom  
## Residual deviance: 275.04 on 282 degrees of freedom  
## AIC: 1026.8  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.872   
## Std. Err.: 0.131   
##   
## 2 x log-likelihood: -1014.758

# Question 4

calculate pseudo R-squared for GLM

# residual deviance/ null deviance  
mod3$deviance / mod3$null.deviance

## [1] 0.6351174

The final model is useful as the R^2 value is above 0.4, which is typically “good enough for ecology” even though the model is not perfect.

Based on the z-scores, the two most significant predictors are hardwood\_cover\_perc and years\_since\_burn in that order. Moreover, both are negatively correlated, suggesting that Bachman’s sparrows dislike luxurious, unburned forests with a high hardwood tree density and like less dense and more recently burned forests.

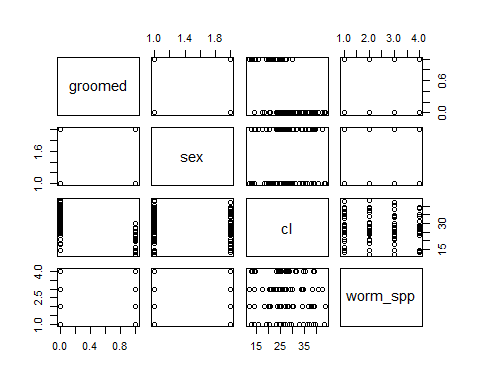
# Question 5 - Crayfish!

read in VT crayfish data

cray <- read.csv("cray\_groom.csv", row.names = 1)

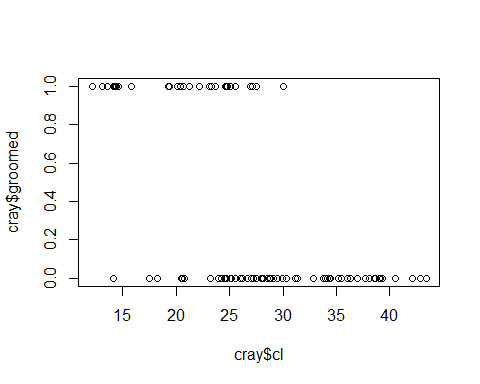
visualize data

plot(cray)



It looks like experimental design purposefully avoided collinearity among predictors. CL appears to correlate most with grooming based on this visualization.

plot(cray$groomed ~ cray$cl)

 Evidently, the response variable only has two levels, so a binomial error distribution is most appropriate.

cray.mod1 <- glm(groomed ~ sex + cl + worm\_spp,  
 data = cray,  
 family = "binomial")  
summary(cray.mod1)

##   
## Call:  
## glm(formula = groomed ~ sex + cl + worm\_spp, family = "binomial",   
## data = cray)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2915 -0.6170 -0.1607 0.4185 1.7067   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.8068 1.5914 3.649 0.000263 \*\*\*  
## sexM 0.5392 0.6260 0.861 0.389003   
## cl -0.2986 0.0667 -4.476 7.59e-06 \*\*\*  
## worm\_sppfallax 0.5509 0.8909 0.618 0.536350   
## worm\_sppingens 1.4296 0.8785 1.627 0.103674   
## worm\_sppptero 0.8570 0.9214 0.930 0.352329   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 118.392 on 92 degrees of freedom  
## Residual deviance: 70.413 on 87 degrees of freedom  
## AIC: 82.413  
##   
## Number of Fisher Scoring iterations: 6

drop1(cray.mod1, test = "Chisq")

## Single term deletions  
##   
## Model:  
## groomed ~ sex + cl + worm\_spp  
## Df Deviance AIC LRT Pr(>Chi)   
## <none> 70.413 82.413   
## sex 1 71.170 81.170 0.757 0.3844   
## cl 1 116.543 126.543 46.130 1.107e-11 \*\*\*  
## worm\_spp 3 73.416 79.416 3.003 0.3912   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

worm\_spp has the least significant effect based on the Chi-squared test and the AIC, so worm\_spp should be removed. This will remove the 3 levels displayed in the previous summary.

cray.mod2 <- glm(groomed ~ sex + cl,  
 data = cray,  
 family = "binomial")  
summary(cray.mod2)

##   
## Call:  
## glm(formula = groomed ~ sex + cl, family = "binomial", data = cray)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1445 -0.6687 -0.2145 0.4574 1.9818   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.17143 1.50617 4.097 4.18e-05 \*\*\*  
## sexM 0.48181 0.59302 0.812 0.417   
## cl -0.28191 0.06207 -4.542 5.57e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 118.392 on 92 degrees of freedom  
## Residual deviance: 73.416 on 90 degrees of freedom  
## AIC: 79.416  
##   
## Number of Fisher Scoring iterations: 6

drop1(cray.mod2, test = "Chisq")

## Single term deletions  
##   
## Model:  
## groomed ~ sex + cl  
## Df Deviance AIC LRT Pr(>Chi)   
## <none> 73.416 79.416   
## sex 1 74.086 78.086 0.670 0.413   
## cl 1 118.370 122.370 44.954 2.017e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

sex has an insignificant effect based on the Chi-squared test, Likelihood ratio test, and the AIC, so sex should be removed.

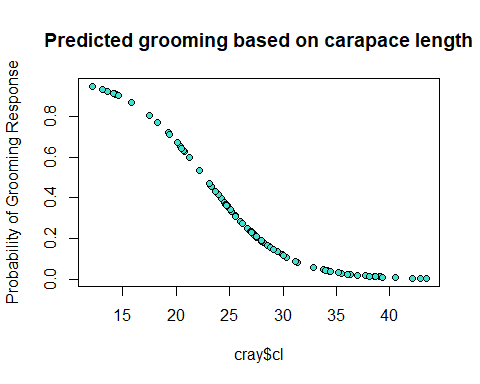
## Final Model

ultimate.cray.mod <- glm(groomed ~ cl,  
 data = cray,  
 family = "binomial")  
summary(ultimate.cray.mod)

##   
## Call:  
## glm(formula = groomed ~ cl, family = "binomial", data = cray)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2085 -0.6463 -0.2143 0.4387 2.0736   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.22349 1.49545 4.162 3.16e-05 \*\*\*  
## cl -0.27471 0.06035 -4.552 5.31e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 118.392 on 92 degrees of freedom  
## Residual deviance: 74.086 on 91 degrees of freedom  
## AIC: 78.086  
##   
## Number of Fisher Scoring iterations: 5

## plot predictor and grooming response

y\_pred <- predict(ultimate.cray.mod, type = "response")  
plot(y\_pred ~ cray$cl,  
 main = "Predicted grooming based on carapace length",  
 ylab = "Probability of Grooming Response",  
 pch = 21,  
 bg = "turquoise")



## for fun

I hypothesize that larger crayfish do not groom as much due to the previous relationship that we observed between carapace length and log number of worms. For instance, a larger crayfish is already expected to be free real estate for worms, so natural selection likely selected for larger crayfish that decided to gamble and let their body become a haven for worms of all kinds, *sometimes* living longer lives depending on if the large crayfish gets a royal flush with all mutualistic worms or a 2 and 7 off-suit with all parasitic worms. This would make biological sense as larger crayfish would safe energy for more important tasks like foraging rather than wasting time by trying to groom hundreds of worms.

## more fun

plot(cray$groomed ~ cray$cl,  
 main = "Observed vs Predicted Grooming Based on Carapace Length",  
 ylab = "Probability of Grooming Response",  
 pch = 21,  
 bg = "firebrick1")  
lines(y\_pred ~ cray$cl,  
 col = "turquoise")  
legend(35, 1, c("observed", "predicted"),  
 pch = 21,  
 pt.bg = c("firebrick1", "turquoise"))

