Environ 710 Lab 12

Logistic Regression

Natalia Neal-Walthall

11.26.18

#### Introduction:

This analysis investigates whether increased lung cancer is associated with birdkeeping. The variables of interest are age (AG), years the individual has smoked (YR), and the indicator variable for birdkeeping (BK). The data frame was modified to include the variable (SV) where a value of 1 indicates that the individual had lung cancer and a value of 0 indicates that the individual did not have lung cancer. The null hypothesis of this study is that birdkeeping has no significant effect on lung cancer after accounting for the effects of age and years the individual smoked. The alternative hypothesis is that birdkeeping has a significant effect on lung cancer rates as determined by a generalized linear model relationship. To test this hypothesis, data from 147 individuals was collected. These data include some who had lung cancer and (negative controls) that did not.

#### Methods

Age and years smoking are both continuous data points, while birdkeeping was a factor variable. The lung cancer data were converted into a binary response variable: 1 = lung cancer and 0 = no lung cancer. Relationships between factors were first examined with ggplots to assess corellations between variables. Based on these results, it does not appear that colinearity is a concern for this data set. The data were then analyzed with a bionomial logistic regression, first creating a saturated model with all factors and interactions included. Models were then created with each interaction removed and compared with the Likelihood Ratio Test (LRT) and AIC. The likelihood ratio test uses a chi squared test statistic to assess the difference in the deviance of the simpler model to the more complex model; if the statistic is significant then the more complex model better predicts the response data. Similarly, models with AIC values within (2) of one another are not significantly different. In all cases, the interactions did not appear to be significant based on minimal change to residual model deviance and AIC. As such, all interactions were removed before examining main effects. As before, models were created by removing one main effect at a time and results were compared with the LRT and AIC. Lastly, the final model was assessed with chi squared test of deviance and McFadden’s pseudo R2 to assess the goodness of fit and predictive ability of the model.

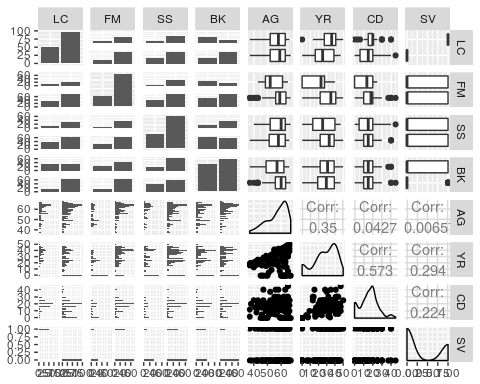
#### Results

The results indicate that years smoking and bird keeping were both significant, but age was not. Therefore, the minimum adequate model is: LungCancer ~ YR + BK With the coefficients provided by the model: LungCancer = -1.70 +.058(YR) + 1.48 The coefficients provided by glm() are log-odds. By taking the anti-log of the coefficients, we can interpret them as odds ratios: Odds(Lung Cancer) = .182 + 1.06(YR) + .228 (If there is a bird) Last, the inverse logit function may be used to interpret coefficients in terms of probabilities: Probability(Lung Cancer) = .153 + .514(Years Smoked) + .186 (If there is a bird)

Examining the model with coefficients in terms of odds, the intercept implies that the odds of an individual who is 0 years old and does not have birds developing lung cancer is .182. This is not meaningful, in part because this age value is well outside the range of data used to create the model. Intercept is therefor not a useful parameter. The model also implies that odds of developing lung cancer increase by 6% per year smoked and 23% if the individual keeps birds. For a bird keeper who has smoked for 32 years, the probablity of developing lung cancer is 54% based on this model. This decreases to 21% if the same individual does not keep birds. Thus, we can reject the null hypothesis and say there is a strong binomial logistic regression relationship of lung cancer to bird keeping. However, it should be noted that the results of the chi squared test of deviance and McFadden’s pseudo R2 indicate that while the model fits the data reasonable well, it does not explain a large portion of the variance. It is therefore likely that an important predictive factor has not been included in this analysis. However, even when all other factors found in the data set were added to analysis, the model was not improved.

## Loading required package: Sleuth3

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK)  
## Model 3: SV ~ AG + YR + factor(BK) + AG:YR + YR:factor(BK)  
## Model 4: SV ~ AG + YR + factor(BK) + AG:factor(BK) + YR:factor(BK)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 140 154.53   
## 2 141 154.53 -1 -0.00164 0.9676  
## 3 141 154.60 0 -0.07221   
## 4 141 156.04 0 -1.44396

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ AG + YR + factor(BK) + AG:YR  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 140 154.53   
## 2 142 154.65 -2 -0.1254 0.9392

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ AG + YR + factor(BK) + AG:YR  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 140 154.53   
## 2 142 154.65 -2 -0.1254 0.9392

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ YR + factor(BK) + AG:YR  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 140 154.53   
## 2 143 154.99 -3 -0.46308 0.9269

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ AG + factor(BK) + AG:YR  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 140 154.53   
## 2 143 157.47 -3 -2.9382 0.4013

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ AG + YR + AG:YR  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 140 154.53   
## 2 143 167.20 -3 -12.675 0.005396 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## df AIC  
## lr1 7 168.5268  
## lr5 5 164.6522  
## lr6 4 162.9899  
## lr7 4 165.4650  
## lr8 4 175.2014

## Analysis of Deviance Table  
##   
## Model 1: SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK)  
## Model 2: SV ~ YR + factor(BK)  
## Resid. Df Resid. Dev Df Deviance  
## 1 140 154.53   
## 2 144 158.11 -4 -3.5876

##   
## Call:  
## glm(formula = SV ~ YR + factor(BK), family = binomial, data = sdat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6093 -0.8644 -0.5283 0.9479 2.0937   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.70460 0.56267 -3.030 0.002450 \*\*   
## YR 0.05825 0.01685 3.458 0.000544 \*\*\*  
## factor(BK)NoBird -1.47555 0.39588 -3.727 0.000194 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 187.14 on 146 degrees of freedom  
## Residual deviance: 158.11 on 144 degrees of freedom  
## AIC: 164.11  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) +   
## YR:factor(BK), family = binomial, data = sdat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6565 -0.8115 -0.4661 0.9619 2.1580   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -5.557938 5.076262 -1.095 0.274  
## AG 0.054623 0.089646 0.609 0.542  
## YR 0.301177 0.193975 1.553 0.121  
## factor(BK)NoBird -0.254966 3.351246 -0.076 0.939  
## AG:YR -0.003782 0.003193 -1.185 0.236  
## AG:factor(BK)NoBird -0.018583 0.068333 -0.272 0.786  
## YR:factor(BK)NoBird -0.001832 0.045137 -0.041 0.968  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 187.14 on 146 degrees of freedom  
## Residual deviance: 154.53 on 140 degrees of freedom  
## AIC: 168.53  
##   
## Number of Fisher Scoring iterations: 5

##   
## Call:  
## glm(formula = SV ~ YR + factor(BK) + factor(FM) + factor(SS) +   
## CD, family = binomial, data = sdat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6380 -0.8431 -0.4956 0.9778 2.1543   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.74265 0.72823 -2.393 0.016711 \*   
## YR 0.05677 0.02033 2.793 0.005229 \*\*   
## factor(BK)NoBird -1.43560 0.40655 -3.531 0.000414 \*\*\*  
## factor(FM)Male -0.58361 0.52725 -1.107 0.268340   
## factor(SS)Low -0.03958 0.46451 -0.085 0.932102   
## CD 0.03107 0.02476 1.255 0.209555   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 187.14 on 146 degrees of freedom  
## Residual deviance: 155.49 on 141 degrees of freedom  
## AIC: 167.49  
##   
## Number of Fisher Scoring iterations: 5



## (Intercept) YR factor(BK)NoBird   
## 0.1818444 1.0599797 0.2286526

## [1] 1.098017

## [1] 0.1990647

## Loading required package: pscl

## Classes and Methods for R developed in the  
## Political Science Computational Laboratory  
## Department of Political Science  
## Stanford University  
## Simon Jackman  
## hurdle and zeroinfl functions by Achim Zeileis

## llh llhNull G2 McFadden r2ML r2CU   
## -79.0572043 -93.5675827 29.0207568 0.1550791 0.1791543 0.2488193

require(Sleuth3)  
sdat<- case2002  
sdat$SV <- ifelse(sdat$LC == "LungCancer", 1, 0)  
ggpairs(sdat)  
#Does not appear that any of the variables are highly corellated  
#First, examine the saturated model vs. models with one interaction removed.  
lr1 <- glm(SV ~ AG + YR + factor(BK) + AG:YR + AG:factor(BK) + YR:factor(BK), family=binomial, data=sdat)   
lr2 <- update(lr1, ~.-YR:factor(BK))  
lr3 <- update(lr1, ~.-AG:factor(BK))  
lr4 <- update(lr1, ~.-AG:YR)  
anova(lr1, lr2, lr3, lr4, test = "Chisq")  
#Based on the results, it does not appear that interactions have an effect. These will be removed.   
lr5 <- update(lr2, ~.-AG:factor(BK))  
anova(lr1, lr5, test = "Chisq")  
#Great. OK. That's definitely a good start. What if we remove main effects?  
lr6 <- update(lr5, ~.-AG)  
lr7 <- update(lr5, ~.-YR)  
lr8 <- update(lr5, ~.-factor(BK))  
anova(lr1, lr5, test = "Chisq")  
anova(lr1, lr6, test = "Chisq")  
anova(lr1, lr7, test = "Chisq")  
anova(lr1, lr8, test = "Chisq")  
AIC(lr1, lr5, lr6, lr7, lr8)  
#It appears that BK and YR are significant. AG is not  
lr9 <- glm(SV ~ YR + factor(BK), family=binomial, data=sdat)   
anova(lr1, lr9)  
summary(lr9)  
summary(lr1)  
lr10 <- glm(SV ~ YR + factor(BK) +factor(FM) + factor(SS) + CD, family = binomial, data = sdat)  
summary(lr10)  
plot(jitter(sdat$YR), sdat$SV, las=1, pch=21, cex=1.2, bg="grey", xlab = "Years Smoking", ylab="Probability of Lung Cancer")  
legend(x= "topleft", y = "topleft", legend=c("Bird Owner", "Non-Bird Owner"), fill=c("blue", "red"), cex=0.8)  
x <- seq(min(sdat$YR), max(sdat$YR), length = 100)   
#curve(expr = inv.logit(lr9$coef[1] + lr9$coef[2]\*x +lr9$coef[3]), add=T, lwd=2, col="red")  
#curve(expr = inv.logit(lr9$coef[1] + lr9$coef[2]\*x), add=T, lwd=2, col="blue")  
exp(coefficients(lr9))  
#inv.logit(coefficients(lr9))  
 cf <- (coefficients(lr9))   
 #We can then calculate survival odds for a specific age.   
#inv.logit(cf[1] + cf[2]\*32 + cf[3]\*1)  
#inv.logit(cf[1] + cf[2]\*32)  
 lr9$deviance/lr9$df.residual  
pchisq(lr9$deviance, lr9$df.residual, lower=F)  
require(pscl)  
lr0 <- glm(SV ~ 1, family=binomial, data=sdat)  
pseudoR2 <- 1-logLik(lr9)/logLik(lr0)  
pR2(lr9)