Owning a Bird

Nico Hillison

May 2022

1 Introduction

I was given a dataset containing information about 147 patients, 49 with lung cancer, and 98 were controlled. Each patient had descriptive information about their sex, if they had lung cancer, if they owned a bird, socio-economic status, age, years smoked, and cigarettes per day. I was tasked to find if there was a relationship between lung cancer and bird keeping.

Using this data I created a model that predicted if a patient owns a bird: The model consists of one extra variable and no secondary terms.

2 Data Cleaning

I examined the data and no individual data point seemed invalid, but I did create two additional variables. As my goal was to predict the relationship with those who have lung cancer, I created a variable called LC_ind which takes in the patients who have lung cancer and not those who do not. I also did the same thing for bird-keeping. Because we were interested in the patients who owned a bird, I created a variable called BK_ind which takes in the patients who own a bird and not those who do not. See *Table 1* for the final table of variables.

Variable	Type	Description
LC_ind	Factor	Patients who have lung cancer
BK_ind	Factor	Patients who own a bird
Sex	Factor	Sex patients identify as (Male or Female in this study)
Socio-Economic Status (SS)	Factor	Patient's economic status (Low or High)
Age (AG)	Integer	Patients age
Years Smoked (YR)	Integer	Years a patient has been smoking
Cigarettes per Day (CD)	Integer	Cigarettes consumed per day by a patient

Table 1: Variables

3 Methods

First, I calculated the crude odds ratio of those who had lung cancer and owned a bird by hand.

For my model, I decided to start by adding every variable into a multivariate logistic regression in R. Then I used the step method respectively. I validated that the simpler model was different than the full model by checking on the AIC.

In order to find potential interaction terms, I used the rpart library to generate a regression tree. From there, I started from the simplified model, added the secondary term from the regression tree, and used the same procedure described above. After using the step function again with the new interaction terms, it showed that the secondary term was useless and unnecessary. Afterward, I conducted the odds ratio with the simple (better) model to find the adjusted odds ratio.

Lastly, I found the 95% confidence interval of both the crude odds ratio and the adjusted odds ratio. All analyses were performed using the software package used (R) and the add-on packages: lmtest, Stat2Data, and rpart.

4 Results

	LungCancer	**NoCancer**
Bird	33	34
NoBird	16	64

Table 2: 2x2 Table

To find the crude odds ratio, I had to make some calculations by hand with the variables shown in *Table 2*.

Then, I decided to create the logistic regression model with all the variables. This model can be seen in *Table 3,* This came out with an AIC of 168.2, and the null deviance is 187.4 on 146 degrees of freedom while the residual deviance was 154.20 on 140 degrees of freedom.

	Estimate	**Std. Error**	**z - value**	**Pr(>|z|)**
(Intercept)	-1.2706	1.8253	-0.70	0.48635
bk.df$BK_ind	1.3626	0.4113	3.31	0.00092
as.factor(bk.df$SEX)MALE	-0.5613	0.5312	-1.06	0.29065
as.factor(bk.df$SS)LOW	-0.1054	0.4688	-0.22	0.82205
bk.df$AG	-0.0398	0.0355	-1.12	0.26250
bk.df$YR	0.0729	0.0265	2.75	0.00594
bk.df$CD	0.0260	0.0255	1.02	0.30806

Table 3: Full Regression Model Output

The reduced regression model had an AIC of 164.1. Because the AIC was lowered, this showed that the reduced regression model is better. The model output is displayed in *Table 4.* The null deviance is 187.14 on 146 degrees of freedom while the residual deviance is 158.11 on 144 degrees of freedom.

	Estimate	**Std. Error**	**z - value**	**Pr(>|z|)**
(Intercept)	-3.1802	0.6364	-5.00	5.8e-07
bk.df$BK_ind	1.4756	0.3959	3.73	0.00019
bk.df$YR	0.0582	0.0168	3.46	0.00054

Table 4: Reduced Regression Model Output

Looking at the regression tree in *Figure 1*, I see that YR may be interacting with itself.

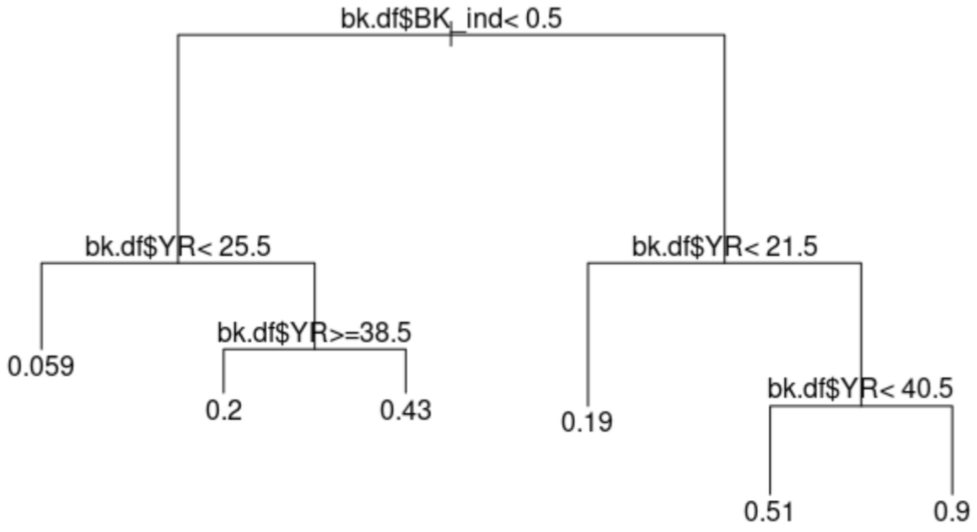


Figure 1: Regression Tree

After Adding the new interaction term to the reduced model, It came back with an AIC of 165.1. Though it might not be a lot, it is still worse than our basic reduced model. After using the step function on the reduced model with the new secondary term, It showed that the secondary term was useless. Lastly, I decided to conduct the adjusted odds ratio.

5 Conclusion/Discussion

I was tasked with trying to find a relationship between those who have lung cancer and those who own a bird. This was accomplished by finding the crude odds ratio and then the adjusted odds ratio from a logistic regression model. The crude odds ratio found that those who have lung cancer are 3.88 times more likely to have a bird than those who do not have lung cancer, and the adjusted odds ratio showed that those who have lung cancer are 4.37 times more likely to have a bird than those who do not have lung cancer. To find the adjusted odds ratio, I had to test different regression models and came to the conclusion that the simplest and the most reduced was the best one. This model can be seen again in *Table 4.*

In conclusion, I believe that those who have lung cancer have a higher chance of owning a bird than those who do not have lung cancer.

A Appendix

```
#load data
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

```
library(leaps)
library(Stat2Data)
bk.data <- read.csv(file = "~/SC321/Reports/Report 4/birdkeeping.csv")
attach(bk.data)
bk.df <- data.frame(bk.data)
#creating table to see how many have lung cancer and are bird keepers
my.table <- table(bk.df$BK, bk.df$LC)
my.table
```

```
## 
##          LUNGCANCER NOCANCER
##   BIRD           33       34
##   NOBIRD         16       64
```

observations:

What are the odds of having a bird? 67/80 What are the odds of having a bird if you have lung cancer? 33/16 What are the odds of having a bird if you don’t have lung cancer? 34/64 What are the odds ratio for having a bird comparing those who have and don’t have lung cancer? (33/16)/(34/64)

```
#do the math
yesbird <- 67/80
birdcancer <- 33/16
birdnocancer <- 34/64
oddsRatio <- (33/34)/(16/64)
yesbird
```

```
## [1] 0.8375
```

```
birdcancer
```

```
## [1] 2.0625
```

```
birdnocancer
```

```
## [1] 0.53125
```

```
oddsRatio
```

```
## [1] 3.882353
```

This means that those who have lung cancer are 3.88 times more likely to have a bird than those who do not have lung cancer.

Linear models

```r
bk.df$LC_ind <- as.numeric(bk.df$LC == "LUNGCANCER")
bk.df$BK_ind <- as.numeric(bk.df$BK == "BIRD")
main.glm <- glm(bk.df$LC_ind ~ bk.df$BK_ind + as.factor(bk.df$SEX) + as.factor(bk.df$SS) + bk.df$AG +
bk.df$YR + bk.df$CD, family="binomial")
#str(bk.df)
summary(main.glm)
```

```
##
## Call:
## glm(formula = bk.df$LC_ind ~ bk.df$BK_ind + as.factor(bk.df$SEX) +
##     as.factor(bk.df$SS) + bk.df$AG + bk.df$YR + bk.df$CD, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5642  -0.8333  -0.4605   0.9808   2.2460
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.27064    1.82531  -0.696 0.486351
## bk.df$BK_ind              1.36259    0.41128   3.313 0.000923 ***
## as.factor(bk.df$SEX)MALE -0.56127    0.53116  -1.057 0.290653
## as.factor(bk.df$SS)LOW   -0.10545    0.46885  -0.225 0.822050
## bk.df$AG                 -0.03976    0.03548  -1.120 0.262503
## bk.df$YR                  0.07287    0.02649   2.751 0.005940 **
## bk.df$CD                  0.02602    0.02552   1.019 0.308055
## ---
## 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: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```

```r
step(main.glm)
```

```
## Start:  AIC=168.2
## bk.df$LC_ind ~ bk.df$BK_ind + as.factor(bk.df$SEX) + as.factor(bk.df$SS) +
##     bk.df$AG + bk.df$YR + bk.df$CD
##
##                          Df Deviance    AIC
## - as.factor(bk.df$SS)    1   154.25 166.25
## - bk.df$CD               1   155.24 167.24
## - as.factor(bk.df$SEX)  1   155.32 167.32
## - bk.df$AG               1   155.49 167.49
## <none>                       154.20 168.20
## - bk.df$YR               1   163.93 175.93
## - bk.df$BK_ind           1   165.87 177.87
##
## Step:  AIC=166.25
## bk.df$LC_ind ~ bk.df$BK_ind + as.factor(bk.df$SEX) + bk.df$AG +
##     bk.df$YR + bk.df$CD
##
##                          Df Deviance    AIC
## - as.factor(bk.df$SEX)  1   155.32 165.32
## - bk.df$CD               1   155.32 165.32
## - bk.df$AG               1   155.50 165.50
## <none>                       154.25 166.25
## - bk.df$YR               1   164.09 174.09
## - bk.df$BK_ind           1   165.90 175.90
##
## Step:  AIC=165.32
## bk.df$LC_ind ~ bk.df$BK_ind + bk.df$AG + bk.df$YR + bk.df$CD
##
##                  Df Deviance    AIC
## - bk.df$CD      1   156.22 164.22
## - bk.df$AG      1   156.75 164.75
## <none>               155.32 165.32
## - bk.df$YR      1   164.18 172.18
## - bk.df$BK_ind  1   168.35 176.35
##
## Step:  AIC=164.22
## bk.df$LC_ind ~ bk.df$BK_ind + bk.df$AG + bk.df$YR
##
```

```
##                Df Deviance    AIC
## - bk.df$AG      1   158.11 164.11
## <none>              156.22 164.22
## - bk.df$BK_ind  1   168.83 174.83
## - bk.df$YR      1   172.53 178.53
##
## Step:  AIC=164.11
## bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR
##
##                Df Deviance    AIC
## <none>              158.11 164.11
## - bk.df$YR      1   172.93 176.93
## - bk.df$BK_ind  1   173.17 177.17
```

```
##
## Call:  glm(formula = bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR, family = "binomial")
##
## Coefficients:
##  (Intercept)  bk.df$BK_ind      bk.df$YR
##     -3.18016       1.47555       0.05825
##
## Degrees of Freedom: 146 Total (i.e. Null);  144 Residual
## Null Deviance:       187.1
## Residual Deviance: 158.1     AIC: 164.1
```

```
fitMain.glm <- glm(bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR, family="binomial")
#str(bk.df)
summary(fitMain.glm)
```

```
##
## Call:
## glm(formula = bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.18016    0.63640  -4.997 5.82e-07 ***
## bk.df$BK_ind  1.47555    0.39588   3.727 0.000194 ***
## bk.df$YR      0.05825    0.01685   3.458 0.000544 ***
## ---
## 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
```

```
lrtest(main.glm, fitMain.glm)
```

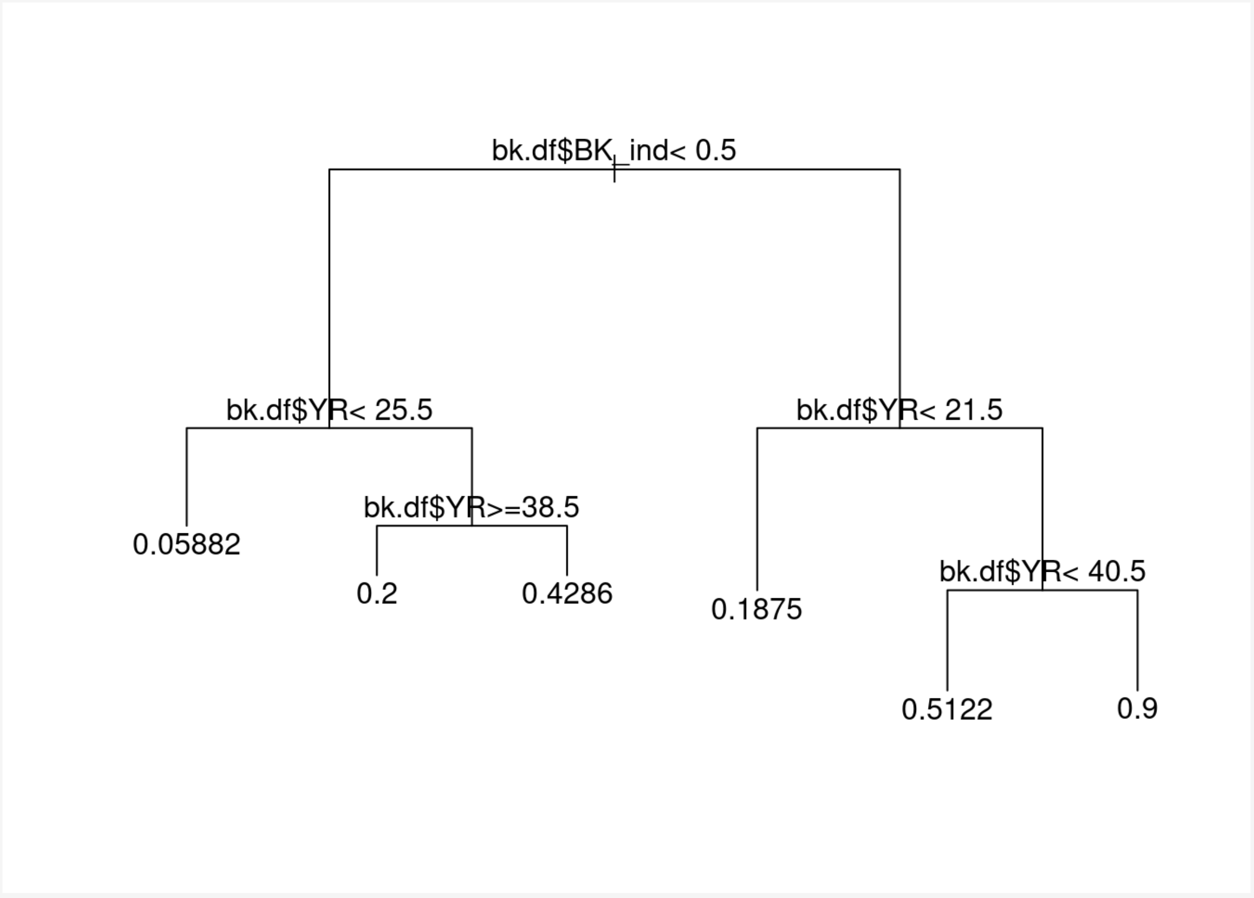
```
## Likelihood ratio test
##
## Model 1: bk.df$LC_ind ~ bk.df$BK_ind + as.factor(bk.df$SEX) + as.factor(bk.df$SS) +
##     bk.df$AG + bk.df$YR + bk.df$CD
## Model 2: bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   7 -77.099
## 2   3 -79.057 -4 3.916     0.4175
```

write description:

Since this p-value is not less than .05, we will fail to reject the null hypothesis.

This means the model and the fitModel fit the data equally well. Thus, we should use the fitModel because the additional predictor variables in the first model don’t offer a significant improvement in fit.

```
library(rpart)
bk.rt <- rpart(bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR)
par(xpd = TRUE)
plot(bk.rt)
text(bk.rt, pretty=0)
```



```
bk.rt
```

```
## n= 147
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 147 32.666670 0.33333330
##    2) bk.df$BK_ind< 0.5 80 12.800000 0.20000000
##      4) bk.df$YR< 25.5 34  1.882353 0.05882353 *
##      5) bk.df$YR>=25.5 46  9.739130 0.30434780
##       10) bk.df$YR>=38.5 25  4.000000 0.20000000 *
##       11) bk.df$YR< 38.5 21  5.142857 0.42857140 *
##    3) bk.df$BK_ind>=0.5 67 16.746270 0.49253730
##      6) bk.df$YR< 21.5 16  2.437500 0.18750000 *
##      7) bk.df$YR>=21.5 51 12.352940 0.58823530
##       14) bk.df$YR< 40.5 41 10.243900 0.51219510 *
##       15) bk.df$YR>=40.5 10  0.900000 0.90000000 *
```

```
printcp(bk.rt)
```

```
##
## Regression tree:
## rpart(formula = bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR)
##
## Variables actually used in tree construction:
## [1] bk.df$BK_ind bk.df$YR
##
## Root node error: 32.667/147 = 0.22222
##
## n= 147
##
##         CP nsplit rel error  xerror     xstd
## 1 0.095522      0   1.00000 1.01991 0.059784
## 2 0.059872      1   0.90448 1.03777 0.079900
## 3 0.037011      2   0.84461 0.96643 0.086257
## 4 0.036077      3   0.80759 0.96014 0.085379
## 5 0.018253      4   0.77152 0.93658 0.088225
## 6 0.010000      5   0.75326 0.98538 0.096229
```

```
better.glm <- glm(bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR + I(bk.df$YR^2), family="binomial")
#str(bk.df)
summary(better.glm)
```

```
##
## Call:
## glm(formula = bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR + I(bk.df$YR^2),
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4755  -0.8454  -0.4819   0.9544   2.2214
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.833264   1.011905  -3.788 0.000152 ***
## bk.df$BK_ind   1.454542   0.394154   3.690 0.000224 ***
## bk.df$YR       0.123289   0.070954   1.738 0.082282 .
## I(bk.df$YR^2) -0.001236   0.001261  -0.980 0.327052
## ---
## 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: 157.07  on 143  degrees of freedom
## AIC: 165.07
##
## Number of Fisher Scoring iterations: 5
```

```
step(better.glm)
```

```
## Start:  AIC=165.07
## bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR + I(bk.df$YR^2)
##
##                  Df Deviance    AIC
## - I(bk.df$YR^2)  1   158.11 164.11
## <none>               157.07 165.07
## - bk.df$YR       1   160.97 166.97
## - bk.df$BK_ind   1   171.74 177.74
##
## Step:  AIC=164.11
## bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR
##
##                Df Deviance    AIC
## <none>             158.11 164.11
## - bk.df$YR      1   172.93 176.93
## - bk.df$BK_ind  1   173.17 177.17
```

```
##
## Call:  glm(formula = bk.df$LC_ind ~ bk.df$BK_ind + bk.df$YR, family = "binomial")
##
## Coefficients:
##  (Intercept)  bk.df$BK_ind      bk.df$YR
##     -3.18016       1.47555       0.05825
##
## Degrees of Freedom: 146 Total (i.e. Null);  144 Residual
## Null Deviance:       187.1
## Residual Deviance: 158.1     AIC: 164.1
```

```
#creating table to see how many have lung cancer and are bird keepers with new model (odds ratio)
exp(fitMain.glm$coefficients[-1])
```

```
## bk.df$BK_ind     bk.df$YR
##     4.373447     1.059980
```

This means that those who do have lung cancer are 4.37 times more likely to have a bird than those who do not have lung cancer.

```
simple.glm <- glm(bk.df$LC_ind ~ bk.df$BK_ind, family = "binomial")
summary(simple.glm)
```

```
##
## Call:
## glm(formula = bk.df$LC_ind ~ bk.df$BK_ind, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1648  -0.6681  -0.6681   1.1901   1.7941
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.3863     0.2795  -4.960 7.06e-07 ***
## bk.df$BK_ind   1.3564     0.3713   3.654 0.000259 ***
## ---
## 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: 172.93  on 145  degrees of freedom
## AIC: 176.93
##
## Number of Fisher Scoring iterations: 4
```

I conducted the 95% confidence interval for the crude odds ratio, as well as the adjusted odds ratio.

	2.5%	**97.5%**
(Intercept)	*0.13971*	*0.42111*
bk.df$BK_ind	*1.90171*	*8.19884*

Table 5: 95% C.I. of Crude Odds Ratio

And the confidence interval for the adjusted odds ratio in *Table 6.*

	2.5%	**97.5%**
(Intercept)	*0.010631*	*0.1312*
bk.df$BK_ind	*2.050761*	*9.7482*
bk.df$YR	*1.027572*	*1.0983*

Table 6: 95% C.I. of Adjusted Odds Ratio