SC321 Data Analysis Report 4

Julian Zhu

May 15th 2022

Contents

1 Introduction

Our data comes from an observational study analyzing 49 lung cancer patients and 98 control patients. We attempt to investigate the potential relationship between lung cancer and bird keeping after we control for potential confounders.

2 Variable Information

Our data contains 147 subjects. There are seven variables: 1) whether one has lung cancer (LC), 2) sex, 3) socio-economic status(SS), 4) bird keeping status(BK), 5) age(AG), 6) years smoked(YR), and 7) cigarettes per day(CD). The first four variables are categorical, and the remaining three are quantitative.

3 Methods

We use forward selection to fit a logistic regression model that can best predict the status of lung cancer. The response is LC, and the remaining six variables are potential predictors. We start with a null model without any predictors. Then we select variables by the drop-in-deviance that they give to the null model according to results from likelihood ratio tests. We repeat the process for additional variables, second-order terms, or interaction terms by comparing new models to reduced ones. We use R to perform all analyses.

4 Results

Here is a summary table of descriptive statistics that we are interested in (See Table 1).

Variables	Mean	Median	1st, 3rd Quantile	Min, Max	Standard Deviation
AG	56.97	59.00	52.00, 63.00	37.00, 67.00	7.348856
YR	27.85	30.00	20.00, 39.00	0.00, 50.00	13.97569
CD	15.75	15.00	10.00, 20.00	0.00, 45.00	9.703723
LC\BK	BIRD	NOBIRD			
LUNGCANCER	33	16			
NOCANCER	34	64			
Sex	FEMALE: 36	MALE: 111			
SS	HIGH: 45	LOW: 102			

Table 1. Descriptive statistics summary table.

We include a two-way table for LC and BK in the summary table. The odds ratio comparing the odds of not having cancer for those who do not keep birds to the odds of not having cancer for those who keep bird is

$$\frac{\frac{64}{16}}{\frac{34}{33}} = 3.8824.$$

We expect the status of not keeping birds is associated with an 3.8824 increase in the odds of not having a lung cancer.

The results of likelihood ratio tests for simple logistic regressions show that BK gives the largest drop-in-deviance 14.204 with a reasonably small p-value 0.000164 (see Table 2).

Likelihood Ratio Test					
Model 1: LC ~ None					
Model 2: LC ~ Potential Variable					
Variables	Df	LogLik	Df	Chisq	P-value
SEX	2	-93.568	1	0	1
SS	2	-92.904	1	1.3266	0.2494
BK	2	-86.466	1	14.204	0.000164 ***
AG	2	-93.564	1	0.0063	0.9366
YR	2	-86.584	1	13.968	0.0001859 ***
CD	2	-89.812	1	7.5103	0.006135 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 2. Test results for the first term.

So we add BK to our model. The results of likelihood ratio tests for 2-term multiple logistic regressions show that YR gives the largest drop-in-deviance 14.817 with a reasonably small p-value 0.0001185 (see Table 3).

Likelihood Ratio Test					
Model 1: LC ~ BK					
Model 2: LC ~ BK+ Potential Variable					
Variables	Df	LogLik	Df	Chisq	P-value
SEX	3	-86.221	1	0.49	0.4839
SS	3	-86.362	1	0.2078	0.6485
AG	3	-86.265	1	0.4006	0.5268
YR	3	-79.057	1	14.817	0.0001185 ***
CD	3	-82.181	1	8.5697	0.003418 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 3. Test results for the second term.

So we add YR to our model. The results of likelihood ratio tests for an additional term show that no term would give much drop-in-deviance, and the p-values are all larger than 0.1 (see Table 4).

Likelihood Ratio Test					
Model 1: LC ~ BK + YR					
Model 2: LC ~ BK + YR + Potential Variable					
Variables	Df	LogLik	Df	Chisq	P-value
YR^2	4	-78.534	1	1.0461	0.3064
BK*YR	4	-79.020	1	0.0744	0.7851
SEX	4	-78.549	1	1.017	0.3132
SS	4	-79.054	1	0.0068	0.9343
AG	4	-78.108	1	1.8974	0.1684
CD	4	-78.374	1	1.367	0.2423
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 4. Test results for the third term.

So we do not need to add an additional term. Our model is

$$\widehat{LCNOCANCER} = 1.70460 + 1.47555 \cdot \text{BKNOBIRD} - 0.05825 \cdot \text{YR}. \tag{1}$$

The p-values of all coefficients are smaller than 0.01 (see Table 5). For BK in particular, the p-value is 0.000194, which is smaller than 0.01 (see Table 5). So there is strong evidence of associations between LC and BK after controlling for YR.

	Coefficient	Standard Error	t-value	p-value
Intercept	1.70460	0.56267	3.030	0.002450 **
BKNOBIRD	1.47555	0.39588	3.727	0.000194 ***
YR	-0.05825	0.01685	-3.458	0.000544 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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				

Table 5. Results of multiple logistic regression.

We also provide the 95 percent confidence intervals of coefficients in the following table (see Table 6).

	2.5%	97.5%
Intercept	0.67001776	2.89540560
BKNOBIRD	0.71821092	2.27708008
YR	-0.09374514	-0.02719882

Table 6. Confidence intervals for coefficients.

5 Conclusion

We attempt to investigate the relationship between lung cancer and bird keeping. We find a multiple logistic model with bird keeping status and years smoked. The model shows that, after controlling the number of years smoked, a person not keeping a bird is $e^1.47555 = 4.3734$ more likely to not get lung cancer. In other words, a person keeping a bird is more likely to get lung cancer after controlling the number of years smoked. The major constraint of this study is that it is an observational study, so we cannot administer treatment and make inference about the general population. I think this would be a constraint for many disease-related studies, and we hope to explore methods to address this constraint.

6 Appendix

Condition Check: *Multiple Logistic Regression*

The regression model is

$$\widehat{logit(\pi_{LCNOCANCER})} = \beta_0 + \beta_1 \cdot \text{BKNOBIRD} + \beta_2 \cdot \text{YR}.$$

Since LC and BK are categorical, and the empirical logit plot for LC versus YR seems linear (see Figure 1), then we do not need to worry about linearity. We can reasonably assume that the data points are independent though not necessarily random because our study is observational. But this won't be too much of a problem as long as we do not make generalization to the larger population.

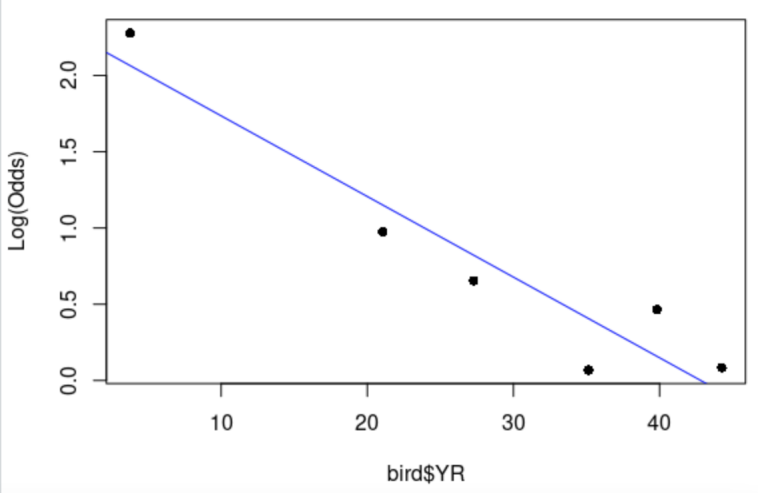

Figure 1. Empirical logit plot for log(LC) versus YR.

R Codes:

```
library(readr)
library(psych)
library(ggplot2)
library(gridExtra)
library(car)
library(rpart)
library(leaps)
library(lmtest)
library(Stat2Data)

bird <- read_csv("SC321/birdkeeping.csv")
str(bird)
summary(bird)

bird$LC <- as.factor(bird$LC)
bird$SEX <- as.factor(bird$SEX)
bird$SS <- as.factor(bird$SS)
bird$BK <- as.factor(bird$BK)
str(bird)
table(bird$LC, bird$BK)
sd(bird$AG)
sd(bird$YR)
sd(bird$CD)

emplogitplot1(bird$LC ~ bird$YR, ngroups = 6)

attach(bird)

# mlr
glm.SEX <- glm(LC ~ SEX, family = binomial)
glm.SS <- glm(LC ~ SS, family = binomial)
glm.BK <- glm(LC ~ BK, family = binomial)
glm.AG <- glm(LC ~ AG, family = binomial)
glm.YR <- glm(LC ~ YR, family = binomial)
glm.CD <- glm(LC ~ CD, family = binomial)
lrtest(glm(LC~1, family = binomial), glm.SEX)
lrtest(glm(LC~1, family = binomial), glm.SS)
lrtest(glm(LC~1, family = binomial), glm.BK)
lrtest(glm(LC~1, family = binomial), glm.AG)
lrtest(glm(LC~1, family = binomial), glm.YR)
lrtest(glm(LC~1, family = binomial), glm.CD)

glm.BK.SEX <- glm(LC ~ BK + SEX, family = binomial)
glm.BK.SS <- glm(LC ~ BK + SS, family = binomial)
glm.BK.AG <- glm(LC ~ BK + AG, family = binomial)
glm.BK.YR <- glm(LC ~ BK + YR, family = binomial)
glm.BK.CD <- glm(LC ~ BK + CD, family = binomial)
lrtest(glm.BK, glm.BK.SEX)
lrtest(glm.BK, glm.BK.SS)
lrtest(glm.BK, glm.BK.AG)
```

```
lrtest(glm.BK, glm.BK.YR) # choose glm.BK.YR
lrtest(glm.BK, glm.BK.CD)

glm.BK.YR.YR2 <- glm(LC ~ BK + YR + I(YR^2), family = binomial)
glm.BK.YR.BKYR <- glm(LC ~ BK + YR + BK:YR, family = binomial)
glm.BK.YR.SEX <- glm(LC ~ BK + YR + SEX, family = binomial)
glm.BK.YR.SS <- glm(LC ~ BK + YR + SS, family = binomial)
glm.BK.YR.AG <- glm(LC ~ BK + YR + AG, family = binomial)
glm.BK.YR.CD <- glm(LC ~ BK + YR + CD, family = binomial)
lrtest(glm.BK.YR, glm.BK.YR.YR2)
lrtest(glm.BK.YR, glm.BK.YR.BKYR)
lrtest(glm.BK.YR, glm.BK.YR.SEX)
lrtest(glm.BK.YR, glm.BK.YR.SS)
lrtest(glm.BK.YR, glm.BK.YR.AG)
lrtest(glm.BK.YR, glm.BK.YR.CD)
# choose none of them

# so the model is glm.BK.YR
summary(glm.BK.YR)
confint(glm.BK.YR)


detach(bird)
```