

Smoking & Birds - which model??

Jo Hardin

October 27, 2014

Data are from a study on birdkeeping and lung cancer (an observational study!). (Holst et al., *British Medical Journal* 1988)

1972-1981 study in the Netherlands discovered an association between keeping pet birds and an increased risk of lung cancer. The study is done with a case-control design including 49 cases of lung cancer and 98 controls. Variables measured include:

measured age (ag), sex, socioeconomic status (ss), years of smoking (yr), cig per day (cd), bird or not (bk)

note: lung.c = 0: no cancer / lung.c = 1: cancer

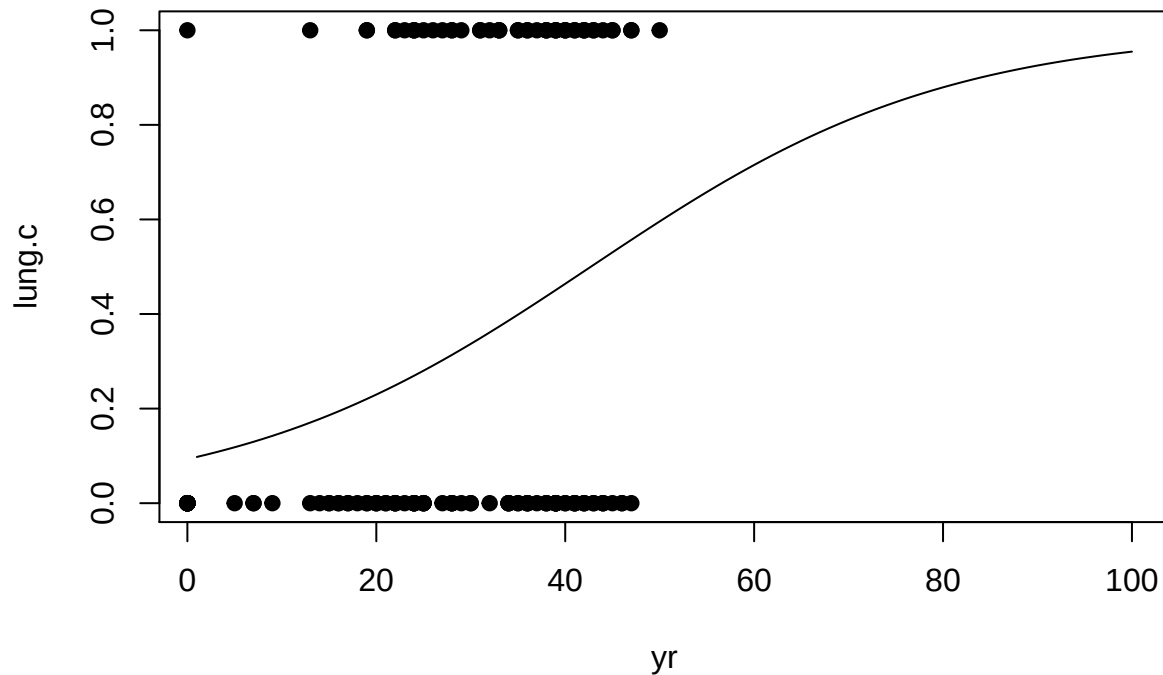
```
smoking <- read.table("H:/teaching/MA150/smoking.txt", header=TRUE, sep="\t")
attach(smoking)
lung.c <- ifelse(lc=="LUNGANCER",1,0)
smoking <- cbind(smoking, lung.c)
smok.log <- glm(lung.c ~ yr, family="binomial", data=smoking)
summary(smok.log)
```

```
##
## Call:
## glm(formula = lung.c ~ yr, family = "binomial", data = smoking)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.276  -0.988  -0.657   1.216   2.179
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2755     0.5282  -4.31  1.6e-05 ***
## yr              0.0533     0.0157   3.39   7e-04 ***
## ---
## 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: 173.17  on 145  degrees of freedom
## AIC: 177.2
##
## Number of Fisher Scoring iterations: 4
```

Making Logistic Plots

```
smok.log <- glm(lung.c~yr,family="binomial", data=smoking)
smok.b <- coef(smok.log)[2]
smok.a <- coef(smok.log)[1]
```

```
plot(yr, lung.c, pch=19, xlim=c(1, 100))
lines(c(1:100), exp(smok.a + smok.b*c(1:100))/(1+exp(smok.a+smok.b*c(1:100))))
```



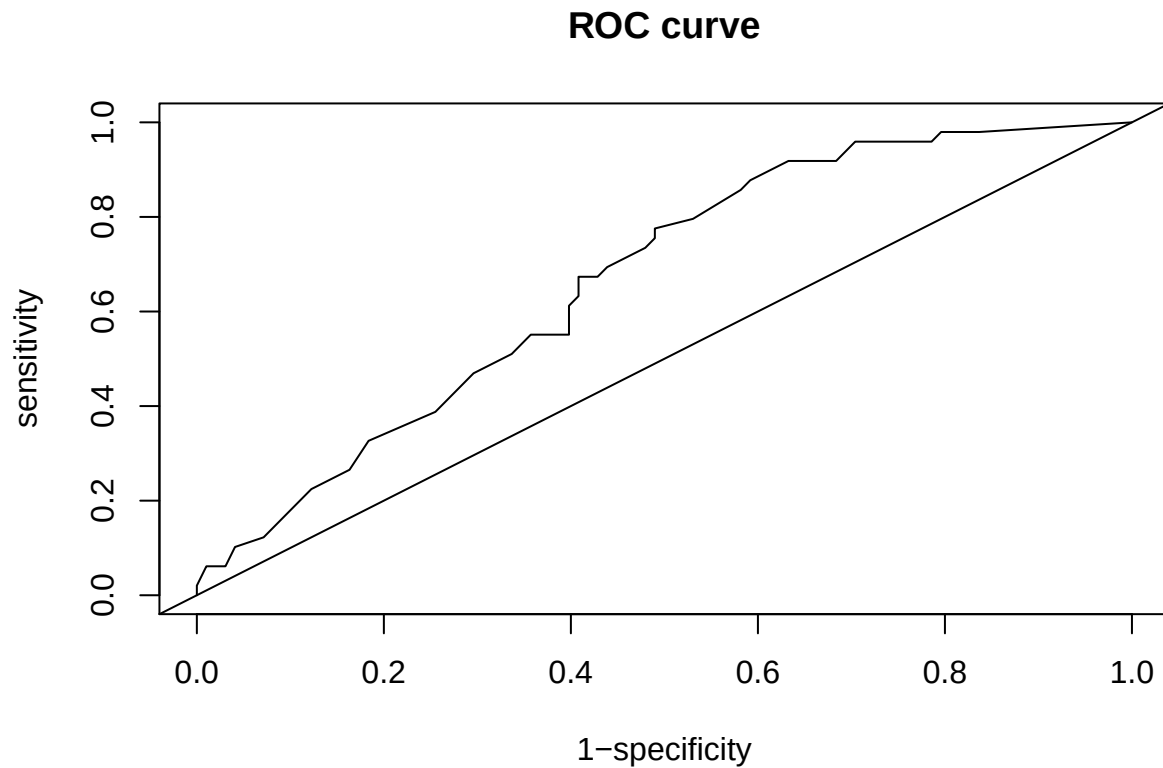
Making ROC Curves

```
#fitted.values the fitted mean values, obtained by transforming
# the linear predictors by the inverse of the link function.
rbind(fitted(smok.log), lung.c)[, order(fitted(smok.log))]
```

```
##          47      88      93      98      101      121      123      125
##      0.09317 0.09317 0.09317 0.09317 0.09317 0.09317 0.09317 0.09317
## lung.c 1.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
##          127      129      133      134      139      141      142      144
##      0.09317 0.09317 0.09317 0.09317 0.09317 0.09317 0.09317 0.09317
## lung.c 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
##          146      60      84      147      136      28      52      94      59      80
##      0.09317 0.1182 0.1298 0.1298 0.1423 0.1704 0.1704 0.178 0.186 0.186
## lung.c 0.00000 0.0000 0.0000 0.0000 0.0000 1.0000 0.0000 0.000 0.000 0.000
##          50      128      54      61      68      1      3      65      73
##      0.1942 0.1942 0.2026 0.2026 0.2114 0.2204 0.2204 0.2204 0.2204
## lung.c 0.0000 0.0000 0.0000 0.0000 0.0000 1.0000 1.0000 0.0000 0.0000
##          51      103      143      53      124      2      38      71      79
```

```
##      0.2297 0.2297 0.2297 0.2392 0.2392 0.2491 0.2491 0.2491 0.2491
## lung.c 0.0000 0.0000 0.0000 0.0000 0.0000 1.0000 1.0000 0.0000 0.0000
##      99    100    42    135    4    6    39    56    75
##      0.2491 0.2491 0.2592 0.2592 0.2695 0.2695 0.2695 0.2695 0.2695
## lung.c 0.0000 0.0000 1.0000 0.0000 1.0000 1.0000 1.0000 0.0000 0.0000
##      76    126    137    40    55    89    130    132    10
##      0.2695 0.2695 0.2695 0.2801 0.2801 0.2801 0.2801 0.2801 0.291
## lung.c 0.0000 0.0000 0.0000 1.0000 0.0000 0.0000 0.0000 0.0000 1.000
##      41    131    21    43    57    58    63    97    48    64
##      0.3021 0.3021 0.3135 0.3135 0.3135 0.3135 0.3135 0.3135 0.325 0.325
## lung.c 1.0000 0.0000 1.0000 1.0000 0.0000 0.0000 0.0000 0.0000 1.000 0.000
##      62    116    5    7    32    72    8    9    44    86
##      0.3368 0.3368 0.3488 0.3488 0.361 0.361 0.3734 0.3734 0.3734 0.386
## lung.c 0.0000 0.0000 1.0000 1.0000 1.000 0.000 1.0000 1.0000 1.0000 0.000
##      87    119    138    11    16    74    85    12    13    70
##      0.386 0.386 0.386 0.3987 0.3987 0.3987 0.3987 0.4115 0.4115 0.4115
## lung.c 0.000 0.000 0.000 1.0000 1.0000 0.0000 0.0000 1.0000 1.0000 0.0000
##      81    111    140    45    83    15    20    46    77
##      0.4115 0.4115 0.4115 0.4245 0.4245 0.4375 0.4375 0.4375 0.4375
## lung.c 0.0000 0.0000 0.0000 1.0000 0.0000 1.0000 1.0000 1.0000 0.0000
##      96    122    14    18    22    67    69    78    90
##      0.4375 0.4375 0.4507 0.4507 0.4507 0.4507 0.4507 0.4507 0.4507
## lung.c 0.0000 0.0000 1.0000 1.0000 1.0000 0.0000 0.0000 0.0000 0.0000
##      107    108    145    19    24    49    109    112    25
##      0.4507 0.4507 0.4507 0.4639 0.4639 0.4639 0.4639 0.4639 0.4772
## lung.c 0.0000 0.0000 0.0000 1.0000 1.0000 1.0000 0.0000 0.0000 1.0000
##      27    66    82    102    106    17    29    33    91
##      0.4772 0.4772 0.4772 0.4772 0.4772 0.4905 0.4905 0.4905 0.4905
## lung.c 1.0000 0.0000 0.0000 0.0000 0.0000 1.0000 1.0000 1.0000 0.0000
##      105    120    23    35    92    104    30    95    114
##      0.4905 0.4905 0.5038 0.5038 0.5038 0.5038 0.5171 0.5171 0.5171
## lung.c 0.0000 0.0000 1.0000 1.0000 0.0000 0.0000 1.0000 0.0000 0.0000
##      115    26    34    110    113    118    31    37    117
##      0.5171 0.5304 0.5304 0.5304 0.5436 0.5436 0.5568 0.5568 0.5568
## lung.c 0.0000 1.0000 1.0000 0.0000 0.0000 0.0000 1.0000 1.0000 0.0000
##      36
##      0.5958
## lung.c 1.0000
```

```
# install.packages("ROCR") # you need this
# library(ROCR)
smok.pred <- prediction(fitted(smok.log),lung.c)
smok.perf <- performance(smok.pred,measure="tpr",x.measure="fpr")
plot(smok.perf,xlab="1-specificity",ylab="sensitivity",main="ROC curve")
abline(a=0,b=1)
```



Measures of Association

```
#library(rms) # you need this
smoke.glm <- lrm(lung.c ~ yr)
print(smoke.glm)
```

```
##
## Logistic Regression Model
##
## lrm(formula = lung.c ~ yr)
##
##               Model Likelihood      Discrimination      Rank Discrim.
##               Ratio Test           Indexes           Indexes
## Obs           147      LR chi2      13.97      R2           0.126      C           0.666
## 0              98      d.f.          1         g           0.832      Dxy          0.333
## 1              49      Pr(> chi2) 0.0002      gr          2.298      gamma         0.343
## max |deriv| 2e-07                                gp          0.162      tau-a         0.149
##
##                               Brier          0.204
##
##               Coef      S.E.    Wald Z Pr(>|Z|)
## Intercept -2.2755 0.5282 -4.31 <0.0001
## yr         0.0533 0.0157  3.39  0.0007
```

Multiple Logistic Regression

Forward Drop in Deviance

```
smok.mlog <- glm(lung.c~yr + ag + sex + ss + yr + cd +bk,family="binomial", data=smoking)
summary(smok.mlog)
```

```
##
## Call:
## glm(formula = lung.c ~ yr + ag + sex + ss + yr + cd + bk, family = "binomial",
##      data = smoking)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.564  -0.833  -0.461   0.981   2.246
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0920     1.7546   0.05  0.95820
## yr            0.0729     0.0265   2.75  0.00594 **
## ag           -0.0398     0.0355  -1.12  0.26250
## sexMALE       -0.5613     0.5312  -1.06  0.29065
## ssLOW         -0.1054     0.4688  -0.22  0.82205
## cd            0.0260     0.0255   1.02  0.30806
## bkNOBIRD     -1.3626     0.4113  -3.31  0.00092 ***
## ---
## 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
```

```
smok.f1 <- glm(lung.c ~ cd, family="binomial", data=smoking)
summary(smok.mlog)
```

```
##
## Call:
## glm(formula = lung.c ~ yr + ag + sex + ss + yr + cd + bk, family = "binomial",
##      data = smoking)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.564  -0.833  -0.461   0.981   2.246
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0920     1.7546   0.05  0.95820
```

```

## yr          0.0729      0.0265      2.75  0.00594 **
## ag          -0.0398      0.0355     -1.12  0.26250
## sexMALE     -0.5613      0.5312     -1.06  0.29065
## ssLOW       -0.1054      0.4688     -0.22  0.82205
## cd           0.0260      0.0255      1.02  0.30806
## bkNOBIRD    -1.3626      0.4113     -3.31  0.00092 ***
## ---
## 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

add1(smok.f1, ~yr + ag + sex + ss + yr + cd + bk, family="binomial", data=smoking, test="Chisq" )

## Single term additions
##
## Model:
## lung.c ~ cd
##      Df Deviance AIC    LRT Pr(>Chi)
## <none>      180 184
## yr      1      172 178  7.55    0.006 **
## ag      1      180 186  0.05    0.826
## sex     1      179 185  0.68    0.408
## ss      1      178 184  1.16    0.282
## bk      1      164 170 15.26   9.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

smok.f2 <- glm(lung.c ~ cd + bk, family="binomial", data=smoking)
add1(smok.f2, ~yr + ag + sex + ss + yr + cd + bk, family="binomial", data=smoking, test="Chisq" )

## Single term additions
##
## Model:
## lung.c ~ cd + bk
##      Df Deviance AIC    LRT Pr(>Chi)
## <none>      164 170
## yr      1      157 165  7.61    0.0058 **
## ag      1      164 172  0.18    0.6706
## sex     1      164 172  0.03    0.8525
## ss      1      164 172  0.18    0.6745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

smok.f3 <- glm(lung.c ~ cd + bk + yr, family="binomial", data=smoking)
add1(smok.f3, ~yr + ag + sex + ss + yr + cd + bk, family="binomial", data=smoking, test="Chisq" )

```

```
## Single term additions
##
## Model:
## lung.c ~ cd + bk + yr
##      Df Deviance AIC    LRT Pr(>Chi)
## <none>      157 165
## ag      1      155 165 1.431    0.23
## sex     1      156 166 1.246    0.26
## ss      1      157 167 0.023    0.88
```

```
summary(smok.f3)
```

```
##
## Call:
## glm(formula = lung.c ~ cd + bk + yr, family = "binomial", data = smoking)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.587  -0.826  -0.525   0.979   2.081
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9112     0.6104   -3.13  0.00174 **
## cd              0.0284     0.0243    1.17  0.24347
## bkNOBIRD     -1.4959     0.3982   -3.76  0.00017 ***
## yr              0.0493     0.0187    2.64  0.00836 **
## ---
## 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: 156.75  on 143  degrees of freedom
## AIC: 164.7
##
## Number of Fisher Scoring iterations: 4
```

Backward BIC

```
smoke.b = glm(lung.c~yr + ag + sex + ss + yr + cd +bk, family="binomial", data=smoking)
step(smoke.b, direction="backward", k=log(147))
```

```
## Start:  AIC=189.1
## lung.c ~ yr + ag + sex + ss + yr + cd + bk
##
##      Df Deviance AIC
## - ss    1      154 184
## - cd    1      155 185
## - sex   1      155 185
## - ag    1      156 185
## <none>   1      154 189
```

```

## - yr      1      164 194
## - bk      1      166 196
##
## Step: AIC=184.2
## lung.c ~ yr + ag + sex + cd + bk
##
##           Df Deviance AIC
## - sex     1      155 180
## - cd      1      155 180
## - ag      1      156 180
## <none>           154 184
## - yr      1      164 189
## - bk      1      166 191
##
## Step: AIC=180.3
## lung.c ~ yr + ag + cd + bk
##
##           Df Deviance AIC
## - cd      1      156 176
## - ag      1      157 177
## <none>           155 180
## - yr      1      164 184
## - bk      1      168 188
##
## Step: AIC=176.2
## lung.c ~ yr + ag + bk
##
##           Df Deviance AIC
## - ag      1      158 173
## <none>           156 176
## - bk      1      169 184
## - yr      1      172 188
##
## Step: AIC=173.1
## lung.c ~ yr + bk
##
##           Df Deviance AIC
## <none>           158 173
## - yr      1      173 183
## - bk      1      173 183
##
##
## Call: glm(formula = lung.c ~ yr + bk, family = "binomial", data = smoking)
##
## Coefficients:
## (Intercept)          yr      bkNOBIRD
##      -1.7046      0.0582      -1.4756
##
## Degrees of Freedom: 146 Total (i.e. Null);  144 Residual
## Null Deviance:      187
## Residual Deviance: 158   AIC: 164

```