Timothy Pratt

Jim Scott

SC321

11 December 2023

Data Analysis Report 3

Introduction:

One may not expect an association between bird-owning and lung cancer; however, previous studies have suggested a relationship may exist. Our researcher collected data from an observational study to further investigate the relationship. There were 147 participants in this study, and the researcher kept track of each participant's lung cancer status, sex, socioeconomic status, bird-keeping status, age, number of years smoked, and cigarettes smoked per day. The researcher's goal is to find the association between lung cancer and each of the other variables, with a focus on bird keeping. Specifically, the researcher wants to know the crude and adjusted odds ratios for each explanatory variable with lung cancer. Additionally, the researcher wants to build the best model using these variables to predict lung cancer status.

Methods:

The researcher provided us with seven variables to conduct our analysis, which are shown in the table below. To meet the conditions for logistic regression, we had to apply a square root transformation to one of the given predictor variables, cigarettes per day (CD). All of our analyses will use the transformed variable (CDRoot) instead of the original.

Variable Name	**Definition**
LC	Binary Response Variable: Whether or not a participant has been diagnosed with lung cancer.

	(Yes or No)
SEX	Binary Explanatory Variable: Sex of a participant. (Female or Male)
SS	Binary Explanatory Variable: Socioeconomic status of a participant. (High or Low)
BK	Binary Explanatory Variable: Whether or not a participant is a bird owner. (Yes or No)
AG	Numeric Explanatory Variable: The age of a participant.
YR	Numeric Explanatory Variable: The number of years a participant has been smoking.
CD	Numeric Explanatory Variable: The number of cigarettes a participant smokes per day.
CDRoot (Modified Variable)	Numeric Explanatory Variable: Square root value of the number of cigarettes a participant smokes per day.

Table 1: Names and definitions for each variable collected.

All of our analyses will use RStudio. To find the crude odds ratio for each explanatory variable with lung cancer (LC), we will build six simple logistic regression models using one explanatory variable as a predictor, and LC as the response. We will then use the RStudio output to calculate and find a confidence interval for each crude odds ratio. Next, we will build one model using all six explanatory variables as predictors. Similarly, we will use the RStudio output to calculate and find a confidence interval for each adjusted odds ratio. To build the best possible main-effects model predicting LC, we will use backward stepwise regression to eliminate the least effective predictors. Lastly, we will conduct a Wald test with bird-keeping status (BK) and our final model to check for a relationship between bird-keeping and lung cancer, while controlling for possible confounders.

Results:

Tables 2 and 3 below outline summary statistics for all variables considered for our model. Figure 1 displays a mosaic plot comparing our main variables of interest, bird keeping (BK) and lung cancer (LC).

Lung Cancer (LC)	Count	Percent	Sex	Count	Percent
Yes	49	33.33333%	**Female**	36	24.4898%
No	98	66.66667%	**Male**	111	75.5102%
Socioeconomic Status (SS)	**Count**	**Percent**	**Bird-Keeping (BK)**	**Count**	**Percent**
High	45	30.61224%	**Yes**	67	45.57823%
Low	102	69.38776%	**No**	80	54.42177%

Table 2: The counts and percentages for each binary variable.

Variable	Mean	Standard Deviation	Median	IQR
Age (AG)	56.96599	7.348856	59	11
Years Smoking (YR)	27.85034	13.97569	30	19
Cigarettes Per Day (CDRoot)	3.613737	1.645485	3.872983	1.309858

Table 3: Summary statistics (mean standard deviation, median, and IQR) for each numeric variable.

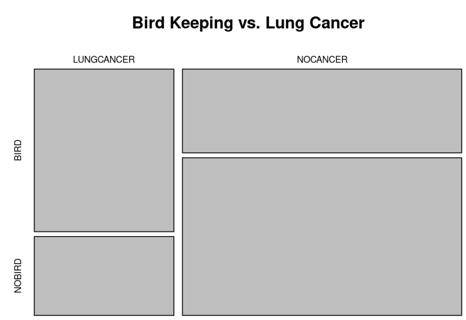

Figure 1: The mosaic plot comparing bird-keeping to lung cancer.

Table 4 below shows the crude and adjusted odds ratios that we calculated for each explanatory variable. As stated earlier, the crude odds ratios were calculated from the six simple models, and the adjusted odds ratios were calculated from the one "full" model. Our main explanatory variable of interest, BK has the highest crude odds ratio and the highest adjusted odds ratio.

Variable	Crude Odds Ratio	95% Confidence Interval	Adjusted Odds Ratio	95% Confidence Interval
Sex	≈1	(0.4504789, 2.21986)	0.5657121	(0.1999158, 1.600839)
Socioeconomic Status (SS)	0.6388408	(0.2945301, 1.385569)	1.123187	(0.4491999, 2.808449)
Bird Keeping (BK)	3.882192	(1.875302, 8.037458)	3.920856	(1.753515, 8.767116)
Age (AG)	0.9981028	(0.9525296, 1.045857)	0.9658565	(0.8992949, 1.03735)
Years Smoking (YR)	1.054714	(1.022729, 1.087703)	1.068782	(1.010648, 1.130253)
Cigarettes Per Day (CDRoot)	1.507722	(1.149108, 1.978225)	1.251471	(0.8478277, 1.847287)

Table 4: The crude and adjusted odds ratios for each explanatory variable, each with a 95% confidence interval.

After running backward stepwise regression on our full model, we find that we can drop all variables from the model except for BK and years of smoking (YR). These two variables are the most effective at predicting LC. This leaves us with our final, best model for predicting LC, with the equation $log(odds(LC)) = -3.18 + 1.476(BK) + 0.058(YR)$. The summary of this model is shown in Table 5 below.

Variable	Coefficient	Standard Error	z-value	p-value
Intercept	-3.18016	0.63640	-4.997	$5.82 * 10^{-7}$
Bird Keeping (BK)	1.47555	0.39588	3.727	0.000194
Years Smoking (YR)	0.05825	0.01685	3.458	0.000544

Table 5: Summary of our best model for predicting lung cancer.

We conducted a Wald test using the summary of our final model (see Table 5) and found that BK is an effective predictor of LC after controlling for cigarettes smoked per day. We can tell from the positive slope coefficient that this is a positive relationship. (In other words, bird owners are more likely to have lung cancer.)

Discussion:

We found that there was a strong relationship between owning a bird and developing lung cancer, even after controlling for possible confounders. The table of odds ratios revealed that patients who owned a bird had higher odds of developing lung cancer. This result did not change much after controlling for all other variables that the researcher collected. Our final, best model for predicting lung cancer used bird-owning and the number of years someone has been smoking as the only two predictors. The Wald test we conducted using this model revealed that the bird-owning variable was an effective predictor. Using the odds ratio from our final model, we can say that when holding years smoked constant, the odds of a participant having lung cancer is 4.373441 times higher if they own a bird than if they do not.

While bird-owning and lung cancer seem unrelated, this study suggests that this relationship may need to be investigated further. Since this is an observational study and not an experiment, we cannot draw any conclusions of causality, which is a limitation. To gain a better understanding of this relationship, and conclude a cause-and-effect relationship, the researcher could conduct an experiment by randomly assigning participants to take care of a bird and watch their health over time. Additionally, if all the participants are from the same area, it is harder to draw generalizations to the population, so this type of study could be repeated in multiple cities or countries to see if the results change.

Appendix:

Creating The Mosaic Plot:

```
df <- read.csv(file = "~/birdkeeping.csv")
table = table(df$LC, df$BK)
mosaicplot(table, main = "Bird Keeping vs. Lung Cancer")
```

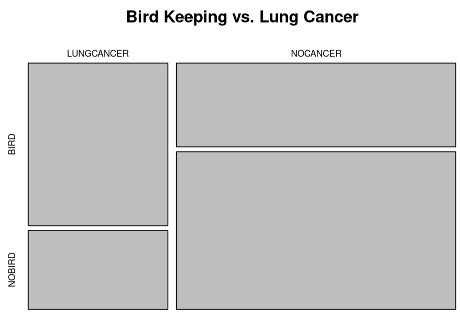

Figure 1: The mosaic plot comparing bird-keeping to lung cancer.

Computing Summary Statistics (Binary):

We used R to find the counts and proportions for each of our binary variables.

```
summary(df$LC)
## LUNGCANCER   NOCANCER
##         49         98
49/147
## [1] 0.3333333
98/147
## [1] 0.6666667

summary(df$SEX)
## FEMALE   MALE
##     36    111
36/147
## [1] 0.244898
111/147
## [1] 0.755102

summary(df$SS)
## HIGH  LOW
##   45  102
45/147
## [1] 0.3061224
102/147
## [1] 0.6938776

summary(df$BK)
##   BIRD NOBIRD
##     67     80
67/147
## [1] 0.4557823
80/147
```

```
## [1] 0.5442177
```

Setting Baselines:

Before moving on, we made sure that we knew that R would use for the baseline for all binary variables.

```
#making sure NOLUNGCANCER = 0 and LUNGCANCER = 1
df$LC = (df$LC == "LUNGCANCER") * 1

#making sure FEMALE = 0 and MALE = 1
df$SEX = (df$SEX == "MALE") * 1

#making sure LOW = 0 and HIGH = 1
df$SS = (df$SS == "HIGH") * 1

#making sure NOBIRD = 0 and BIRD = 1
df$BK = (df$BK == "BIRD") * 1
```

Checking Conditions and Applying Transformations:

Before building any logistic regression models, we have to check 4 conditions: Linearity, Independence, Bernoulli Distribution, and Simple Random Sample. We know that for all models, the last 3 conditions are met. We can assume that one participant's lung cancer status does not affect any others, so Independence is met. The Bernoulli Distribution condition is met because it is reasonable to treat LC as random. Lastly, the researcher did not provide us with information about how the participants were selected, so we can only assume that they were chosen by SRS, or are representative of the population as a whole, meaning that our final SRS condition is met. Linearity is automatically met for binomial variables, but we need to check the numeric variables by looking at their emplogit plots. AG and YR seemed to show a linear trend, but CD seemed a bit non-linear, so we tried a square root transformation. This made a much more linear-looking trend, so we will proceed using this new transformed variable.

```
library(Stat2Data)
emplogitplot1(df$LC ~ df$AG, ngroups = 6)
```

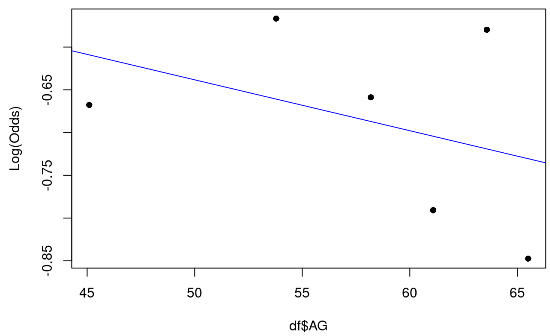

Figure 2: The emplogit plot for AG

```
emplogitplot1(df$LC ~ df$YR, ngroups = 6)
```

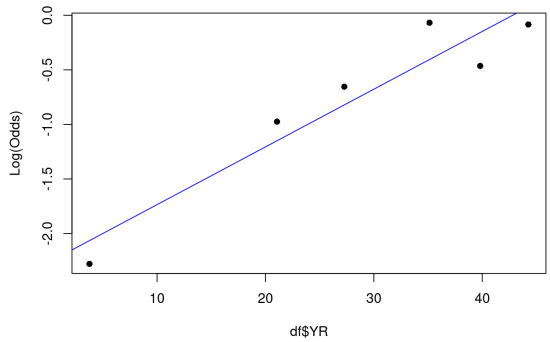

Figure 3: The emplogit plot for YR

```
emplogitplot1(df$LC ~ df$CD, ngroups = 6)
```

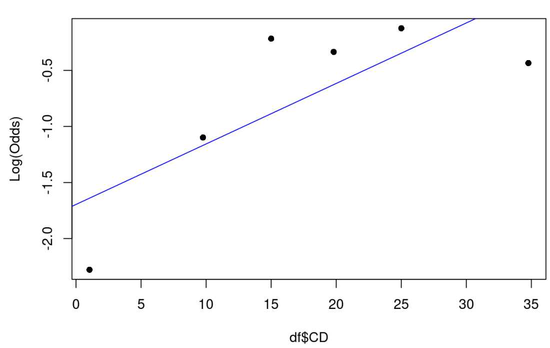

Figure 4: The emplogit plot for CD

```
emplogitplot1(df$LC ~ sqrt(df$CD), ngroups = 6)
```

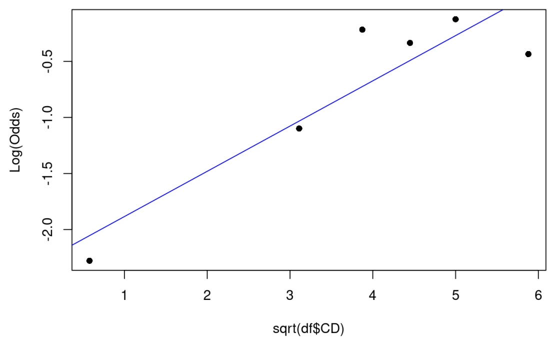

Figure 5: The emplogit plot for CDRoot

```
df$CDRoot = sqrt(df$CD)
```

Computing Summary Statistics (Numeric):

For our numeric variables, we used R to compute each's mean, standard deviation, median, and interquartile range.

```
mean(df$AG)
## [1] 56.96599
sd(df$AG)
## [1] 7.348856
median(df$AG)
## [1] 59
IQR(df$AG)
## [1] 11
mean(df$YR)
## [1] 27.85034
sd(df$YR)
## [1] 13.97569
median(df$YR)
## [1] 30
IQR(df$YR)
## [1] 19
mean(df$CDRoot)
## [1] 3.613737
sd(df$CDRoot)
## [1] 1.645485
median(df$CDRoot)
## [1] 3.872983
IQR(df$CDRoot)
## [1] 1.309858
```

Crude Odds Ratios:

To calculate the crude odds ratio for each variable, we created a simple logistic regression model using only that variable to predict LC. We then exponentiated the slope coefficient to find the crude odds ratio. For the confidence interval, we exponentiated the lower and upper bounds of the confidence interval for the slope.

```
model1 = glm(df$LC ~ df$SEX, family = binomial)
summary(model1)
##
## Call:
## glm(formula = df$LC ~ df$SEX, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9005  -0.9005  -0.9005   1.4823   1.4823
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.931e-01  3.536e-01  -1.961   0.0499 *
## df$SEX      -7.227e-16  4.069e-01   0.000   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 187.14  on 145  degrees of freedom
## AIC: 191.14
##
## Number of Fisher Scoring iterations: 4
exp(-7.227e-16)
## [1] 1
confint.default(model1)
##                 2.5 %        97.5 %
## (Intercept) -1.386099 -0.0001953956
## df$SEX      -0.797444  0.7974440019
exp(-0.797444)
## [1] 0.4504789
exp(0.7974440019)
## [1] 2.21986
model2 = glm(df$LC ~ df$SS, family = binomial)
summary(model2)
##
## Call:
## glm(formula = df$LC ~ df$SS, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9493  -0.9493  -0.7876   1.4241   1.6259
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5635     0.2059  -2.736  0.00622 **
## df$SS        -0.4481     0.3950  -1.134  0.25662
## ---
## 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: 185.81  on 145  degrees of freedom
## AIC: 189.81
##
## Number of Fisher Scoring iterations: 4
```

```
exp(-0.4481)
## [1] 0.6388408
confint.default(model2)
##                 2.5 %     97.5 %
## (Intercept) -0.967106 -0.1598327
## df$SS       -1.222374  0.3261108
exp(-1.222374)
## [1] 0.2945301
exp(0.3261108)
## [1] 1.385569
model3 = glm(df$LC ~ df$BK, family = binomial)
summary(model3)
##
## Call:
## glm(formula = df$LC ~ df$BK, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1648  -0.6681  -0.6681   1.1901   1.7941
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3863     0.2795  -4.960 7.06e-07 ***
## df$BK         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
exp(1.3564)
## [1] 3.882192
confint.default(model3)
##                  2.5 %     97.5 %
## (Intercept) -1.9341208 -0.8384679
## df$BK        0.6287699  2.0841129
exp(0.6287699)
## [1] 1.875302
exp(2.0841129)
## [1] 8.037458
model4 = glm(df$LC ~ df$AG, family = binomial)
summary(model4)
##
## Call:
## glm(formula = df$LC ~ df$AG, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9146  -0.9012  -0.8963   1.4776   1.4900
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.585013   1.368778  -0.427    0.669
## df$AG       -0.001899   0.023845  -0.080    0.937
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 187.13  on 145  degrees of freedom
## AIC: 191.13
##
## Number of Fisher Scoring iterations: 4
```

```
exp(-0.001899)
## [1] 0.9981028
confint.default(model4)
##                   2.5 %     97.5 %
## (Intercept) -3.26776876 2.09774294
## df$AG       -0.04863413 0.04483656
exp(-0.04863413)
## [1] 0.9525296
exp(0.04483656)
## [1] 1.045857
model5 = glm(df$LC ~ df$YR, family = binomial)
summary(model5)
##
## Call:
## glm(formula = df$LC ~ df$YR, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2757  -0.9876  -0.6571   1.2165   2.1787
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.27554    0.52825  -4.308 1.65e-05 ***
## df$YR        0.05327    0.01571   3.390 0.000698 ***
## ---
## 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: 177.17
##
## Number of Fisher Scoring iterations: 4
exp(0.05327)
## [1] 1.054714
confint.default(model5)
##                   2.5 %      97.5 %
## (Intercept) -3.31088804 -1.24019972
## df$YR        0.02247447  0.08406855
exp(0.02247447)
## [1] 1.022729
exp(0.08406855)
## [1] 1.087703
model6 = glm(df$LC ~ df$CDRoot, family = binomial)
summary(model6)
##
## Call:
## glm(formula = df$LC ~ df$CDRoot, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3935  -1.0038  -0.7087   1.3139   2.1720
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2594     0.5844  -3.866 0.000111 ***
## df$CDRoot     0.4106     0.1386   2.963 0.003047 **
## ---
## 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: 176.14  on 145  degrees of freedom
## AIC: 180.14
```

```
##
## Number of Fisher Scoring iterations: 4
exp(0.4106)
## [1] 1.507722
confint.default(model6)
##                  2.5 %     97.5 %
## (Intercept) -3.4048272 -1.1139945
## df$CDRoot    0.1389864  0.6822002
exp(0.1389864)
## [1] 1.149108
exp(0.6822002)
## [1] 1.978225
```

Adjusted Odds Ratios:

We used a simpler method to calculate the adjusted odds ratio for each variable. This time, we created one multiple logistic regression model using every variable to predict LC. Then we exponentiated the slope coefficient to find the adjusted odds ratio. For the confidence interval, we exponentiated the lower and upper bounds of the confidence interval for the slope.

```
modelfull = glm(df$LC ~ df$SEX + df$SS + df$BK + df$AG + df$YR + df$CDRoot, family = binomial)
summary(modelfull)
##
## Call:
## glm(formula = df$LC ~ df$SEX + df$SS + df$BK + df$AG + df$YR +
##     df$CDRoot, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5747  -0.8412  -0.4507   0.9853   2.3328
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.90019    1.90595  -0.997 0.318774
## df$SEX      -0.56967    0.53072  -1.073 0.283100
## df$SS        0.11617    0.46759   0.248 0.803786
## df$BK        1.36631    0.41057   3.328 0.000875 ***
## df$AG       -0.03474    0.03643  -0.953 0.340353
## df$YR        0.06652    0.02853   2.331 0.019745 *
## df$CDRoot    0.22432    0.19868   1.129 0.258866
## ---
## 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: 153.95  on 140  degrees of freedom
## AIC: 167.95
##
## Number of Fisher Scoring iterations: 5
exp(-0.56967)
## [1] 0.5657121
exp(0.11617)
## [1] 1.123187
exp(1.36631)
## [1] 3.920856
```

```
exp(-0.03474)
## [1] 0.9658565
exp(0.06652)
## [1] 1.068782
exp(0.22432)
## [1] 1.251471
confint.default(modelfull)
##                   2.5 %     97.5 %
## (Intercept) -5.63579037 1.83540169
## df$SEX      -1.60985923 0.47052764
## df$SS       -0.80028726 1.03263252
## df$BK        0.56162207 2.17100786
## df$AG       -0.10614428 0.03666943
## df$YR        0.01059137 0.12244119
## df$CDRoot   -0.16507781 0.61371789
exp(-1.60985923)
## [1] 0.1999158
exp(0.47052764)
## [1] 1.600839
exp(-0.80028726)
## [1] 0.4491999
exp(1.03263252)
## [1] 2.808449
exp(0.56162207)
## [1] 1.753515
exp(2.17100786)
## [1] 8.767116
exp(-0.10614428)
## [1] 0.8992949
exp(0.03666943)
## [1] 1.03735
exp(0.01059137)
## [1] 1.010648
exp(0.12244119)
## [1] 1.130253
exp(-0.16507781)
## [1] 0.8478277
exp(0.61371789)
## [1] 1.847287
```

Constructing the Best Model:

We used backwards stepwise regression to eliminate ineffective variables from our full model.

We ended up with a final model using only BK and YR to predict LC.

```
step(modelfull)
## Start:  AIC=167.95
## df$LC ~ df$SEX + df$SS + df$BK + df$AG + df$YR + df$CDRoot
##
##             Df Deviance    AIC
## - df$SS      1   154.01 166.01
## - df$AG      1   154.88 166.88
## - df$SEX     1   155.10 167.10
## - df$CDRoot  1   155.24 167.24
## <none>           153.95 167.95
## - df$YR      1   160.33 172.33
## - df$BK      1   165.70 177.70
##
## Step:  AIC=166.01
## df$LC ~ df$SEX + df$BK + df$AG + df$YR + df$CDRoot
```

```
##
##             Df Deviance    AIC
## - df$AG      1   154.90 164.90
## - df$SEX     1   155.10 165.10
## - df$CDRoot  1   155.32 165.32
## <none>           154.01 166.01
## - df$YR      1   160.36 170.36
## - df$BK      1   165.72 175.72
##
## Step:  AIC=164.9
## df$LC ~ df$SEX + df$BK + df$YR + df$CDRoot
##
##             Df Deviance    AIC
## - df$SEX     1   156.16 164.16
## <none>           154.90 164.90
## - df$CDRoot  1   157.10 165.10
## - df$YR      1   160.69 168.69
## - df$BK      1   168.29 176.29
##
## Step:  AIC=164.16
## df$LC ~ df$BK + df$YR + df$CDRoot
##
##             Df Deviance    AIC
## - df$CDRoot  1   158.11 164.11
## <none>           156.16 164.16
## - df$YR      1   160.93 166.93
## - df$BK      1   171.36 177.36
##
## Step:  AIC=164.11
## df$LC ~ df$BK + df$YR
##
##         Df Deviance    AIC
## <none>       158.11 164.11
## - df$YR  1   172.93 176.93
## - df$BK  1   173.17 177.17
##
## Call:  glm(formula = df$LC ~ df$BK + df$YR, family = binomial)
##
## Coefficients:
## (Intercept)        df$BK        df$YR
##    -3.18016      1.47555      0.05825
##
## Degrees of Freedom: 146 Total (i.e. Null);  144 Residual
## Null Deviance:       187.1
## Residual Deviance: 158.1     AIC: 164.1
bestmodel = glm(df$LC ~ df$BK + df$YR, family = binomial)
summary(bestmodel)
##
## Call:
## glm(formula = df$LC ~ df$BK + df$YR, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.18016    0.63640  -4.997 5.82e-07 ***
## df$BK        1.47555    0.39588   3.727 0.000194 ***
## 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
```

Odds Ratio of BK in Best Model:

In the results section, we referenced the odds ratio of BK with our final model. We used R to calculate this by exponentiating the slope coefficient.

```
exp(1.47555)
## [1] 4.373441
```

Wald Test:

We used the output from our final model (see above) to run a Wald Test to check BK's significance in the final model.

Parameter of Interest: β_{BK} = BK's slope in the final model.

H_0: $\beta_{BK} = 0$ (BK is not an effective predictor in our final model)

H_A: $\beta_{BK} \neq 0$ (BK is an effective predictor in our final model)

Test statistic: $z = 3.727$

P-value: $p = 0.000194 < 0.05$

Conclusion: We reject the null hypothesis, there is strong evidence that BK is an effective predictor in our final model.