Philip Booth

Prof. Ning

SC321

December 15, 2024

Data Analysis Report

Introduction:

Cancer is a disease that affects many lives, whether it's a family member or a friend. Many factors contribute to the likelihood of developing this illness, including lifestyle choices and environmental exposures. Anecdotal data suggests that birdkeeping could be specifically linked to lung cancer. In this paper, I seek to create a model that incorporates various factors, including birdkeeping, to predict the likelihood of developing lung cancer. The data used for this analysis consists of observational information from 147 subjects, 49 of whom have lung cancer and 98 of whom do not. This dataset also contains additional variables potentially associated with lung cancer, such as socioeconomic status and years of smoking. The objective of this paper is to construct a model using these predictors to assess the relationship between birdkeeping and lung cancer and to determine whether birdkeeping remains a significant predictor when other factors are accounted for.

Methods:

The key variables in this dataset are described in Table 1. I did not modify any variables in the dataset, except for converting categorical variables to binary indicators (1 or 0) for logistic regression.

Table 1: Key variables in the dataset and their descriptions.

Variable	Data Type	Description
LC	Categorical	Indicator of whether or not a subject has lung cancer.
Sex	Categorical	Indicator of whether a subject is male or female.
SS	Categorical	Subjects socioeconomic status. (Low, High)
BK	Categorical	Indicator of bird-keeping status.

AG	Quantitative	Subject's Age.
YR	Quantitative	Years that a subject has smoked.
CD	Quantitative	Cigarettes per day smoked.

Table 2 includes the crude and adjusted odds ratios each predictor with lung cancer. To measure crude odds ratios, I ran a single logistic regression for each predictor and lung cancer. These initial crude ratios do not account for any covariance or confounding overlap between variables. Then to get adjusted odds ratios, I ran a multiple logistic regression including all variables and calculated odds ratios for the next slope coefficients of each predictor. Table 2 include's the results of calculating both sets of odds ratios. To determine if there is a relationship between birdkeeping and lung cancer, I performed multiple likelihood ratio tests to remove unnecessary predictors from the logistic regression, including all predictors. All conditions for multiple logistical regression were met, so no further changes to the variables were necessary. All these tests and analyses were performed in R.

Table 2: Crude and Adjusted Odds Ratios for All Predictors

Lung Cancer vs.	Sex	socio-economic status(SS)	bird keeping status(BK)	Age (AG)	years smoked(YR)	cigarettes per day(CD)
Odds Ratio Crude	1.00	0.638	3.882	1.001	0.948	0.950
AdjustedOdds Ratio	1.30	1.094	1.087	1.780	2.162	1.222

Results (Part 1):

Descriptive statistics for the quantitative variables in the dataset are provided in Table 3. As I mentioned in the introduction, 49 subjects (33.33%) have lung cancer and 98 (66.67%) do not. For the primary categorical variable of interest BK, 67 subjects (45.58%) have birds and 80 subjects (55.42%) do not have birds.

Table 3: Descriptive statistics for the key categorical variables in the dataset.

Category	Percentage Distribution
Lung Cancer (Yes)	33.33%
Lung Cancer (No)	66.67%
Sex (Female)	24.49%
Sex (Male)	75.51%
Socio-economic Status (High)	69.39%
Socio-economic Status (Low)	30.61%
Bird-Keeping (Yes)	45.58%
Bird-Keeping (No)	54.42%

Table 4: Descriptive statistics for the key quantitative variables in the dataset.

Quantitative Variables (5 Number Summary)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age (AG)	37.00	52.00	59.00	56.97	63.00	67.00
Years Smoked (YG)	0.00	20.00	30.00	27.85	39.00	50.00
Cigarettes per Day (CD)	0.00	10.00	15.00	15.75	20.00	45.00

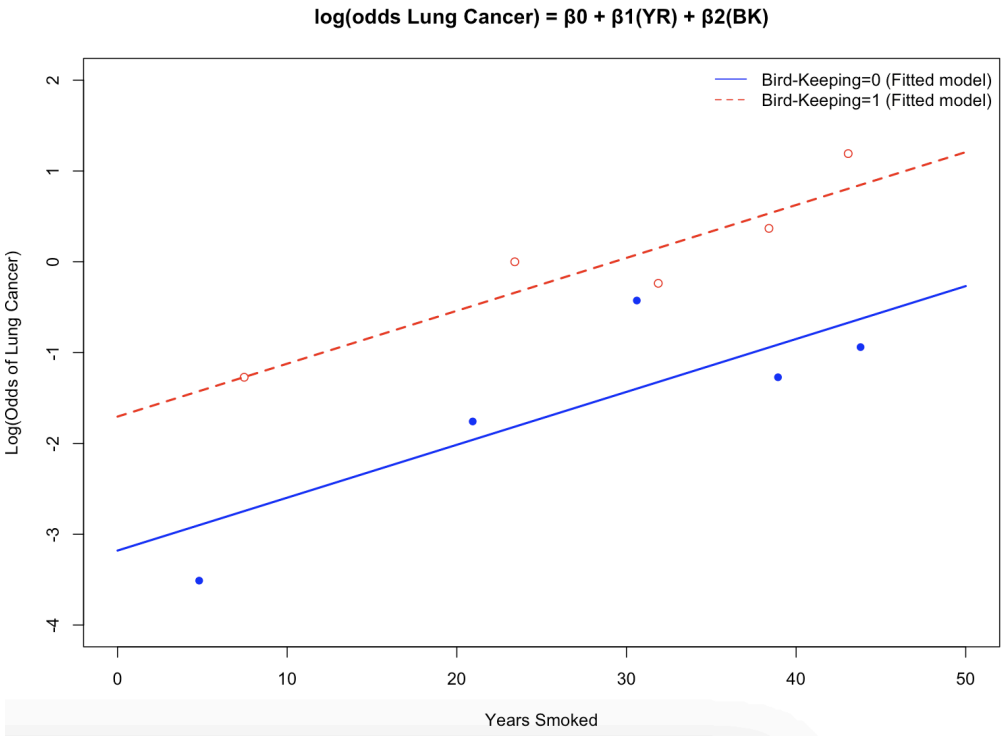
Results (Part 2):

The multiple logistic regression model including all predictors is given as:

$log(odds\ Lung\ Cancer = 0.0354 - 0.54(SexMale) + 1.35212(BKNOBIRD) + 0.03884(AG) - 0.07187(YR) - 0.02632(CD)$. There is no detected multicollinearity in each predictor (VIFs < 5). However, only BK (p-value = 0.000929) and YR (p-value = 0.005887) pass wald z-tests for individual predictors, which tells us that other predictors could be confounding. As a result, I used multiple rounds of likelihood ratio tests to drop unnecessary predictors from the model. The final model I arrived at is

$log(odds\ Lung\ Cancer = 1.70460 + 1.47555(BKNOBIRD) - 0.05825(YR)$. Figure 1 displays the graph of this model. From running a all nested likelihood ratio test from the full model to the final model, I can conrfirm from my p-value = 0.27 that he two predictor model including BK and YR are preferable to the full model.

Figure 1: Fitted Multiple Logistic Regression with YR in 5 groups
Equation of MLR: $log(odds\ Lung\ Cancer = 1.70460 + 1.47555(BKNOBIRD) - 0.05825(YR)$



Discussion:

I found that only two of the five predictors in the dataset were necessary in explaining the model. These two being BK (Bird-Keeping Status) and YR (Years Smoked). By converting the fitted slopes of our best model to odds ratios we can infer that being a bird-keeper decreases the odds of having Lung Cancer by ~77% and for each additional year you smoke the odds of having lung cancer decreases by ~5.7%. Once again the final prediction equation is:

$$log(odds\ Lung\ Cancer = 1.70460 + 1.47555(BKNOBIRD) - 0.05825(YR)$$

This is purely observational data so one cannot claim from these statistics that the longer you smoke for the less likely you are to have lung cancer.

One can conclude, however that birdkeeping has a significant relationship with lung cancer. Originally, I would have thought that having subjects in possession of a bird would be at a higher risk of lung cancer. This statistical report concludes that this is not the case and is in fact the reverse of my assumption. Perhaps there is a correlation between health and taking care of pets, but we do not have enough data to extract such a conclusion from this report.

One possible limitation for this analysis is presence of confounders. Since the data is observational, other factors that were not measured, might influence the relationship between birdkeeping, smoking, and lung cancer. Without accounting for these potential confounders, the model's results must be interpreted with caution. In addition, the conclusion that the longer one smokes the less likely they are to have lung cancer is likely confounding as smoking presents no health benefits and is a clearly destructive activity for ones lungs. For future research perhaps investigating whether birdkeeping is associated with other health behaviors or environmental factors could help provide a clearer understanding of this relationship between birdkeeping and lung cancer.

From this analysis, I can conclude that while a simple multiple logistic regression using birdkeeping and years smoked as predictors is adequate for predicting the odds of having lung cancer, the results should be used with moderate to high degree of caution as the observational data used for this analysis has limitations into the insight it can bring towards birdkeeping directly correlating to lung cancer.

Appendix:

To develop an understanding of the variables in the birdkeeping.csv dataset, I took summary statistics of the two primary quantitative variables. Then for categorical data, I took the percentage distribution from their respective frequency tables.

```
#Descriptive statistics of the data
summary(df$AG)
summary(df$YR)
summary(df$CD)

prop.table(table(df$LC)*100)
prop.table(table(df$SEX)*100)
prop.table(table(df$SS)*100)
prop.table(table(df$BK)*100)
```

For the crude odds ratio, I calculated individual logistic regressions and computed the inverse of the natural log of respective fitted slope coefficients of each model.

```
```{r}
model_SEX <- glm(LC ~ SEX, family = binomial, data = df)
model_SS <- glm(LC ~ SS, family = binomial, data = df)
model_BK <- glm(LC ~ BK, family = binomial, data = df)
model_AG <- glm(LC ~ AG, family = binomial, data = df)
model_YR <- glm(LC ~ YR, family = binomial, data = df)
model_CD <- glm(LC ~ CD, family = binomial, data = df)
or_SEX <- exp(coef(model_SEX)[2])
or_SS <- exp(coef(model_SS)[2])
or_BK <- exp(coef(model_BK)[2])
or_AG <- exp(coef(model_AG)[2])
or_YR <- exp(coef(model_YR)[2])
or_CD <- exp(coef(model_CD)[2])
crude_ORs <- do.call(rbind, list(or_SEX,or_SS, or_BK, or_AG, or_YR, or_CD))
crude_ORs


```
```

As for adjusted odds ratios, I set up the multiple logistic regression including all predictor and then I calculated the log odds of each predictor again by taking the inverse natural log each slope coefficient in the model. In the full model, I also calculated the VIF to detect any multicollinearity, of which I found none.

```
Call:
glm(formula = LC ~ SEX + SS + BK + AG + YR + CD, family = binomial,
    data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.27064    1.82531   0.696 0.486351
SEXMALE      0.56127    0.53116   1.057 0.290653
SSLOW        0.10545    0.46885   0.225 0.822050
BKBIRD      -1.36259    0.41128  -3.313 0.000923 ***
AG           0.03976    0.03548   1.120 0.262503
YR          -0.07287    0.02649  -2.751 0.005940 **
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
```

```
Lmodel_full <- glm(LC ~ SEX + SS + BK + AG + YR + CD, family=binomial, data=df)

vif_vals <- vif(Lmodel_reduced4)

summary(Lmodel_full)
exp(0.56127)
exp(0.10545)
exp(1.36259)
exp(0.03976)
exp(-0.07287)
exp(-0.02602)

vif_vals
```

```
[1] 1.752897
[1] 1.111211
[1] 3.906298
[1] 1.040561
[1] 0.9297217
[1] 0.9743156
```

Results: To determine the best model use, I conducted rounds of drop in deviance tests to take out unnecessary predictors from the equation. I iterated through this process until I only had two predictors left in my model.

Removing SS from the model:

H0: $\beta 2 = 0$

Ha: Some $\beta i \neq 0$

```
Lmodel_full <- glm(LC ~ SEX + SS + BK + AG + YR + CD, family=binomial, data=df)
Lmodel_reduced1 <- glm(LC ~ SEX + BK + AG + YR + CD, family=binomial, data=df)

summary(Lmodel_reduced1)
#stats reduced
154.25
141

#full model
154.20
140

1 - pchisq(.05, 1)
0.8230633
```

```
Call:
glm(formula = LC ~ SEX + BK + AG + YR + CD, family = binomial,
    data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03540    1.66342   0.021 0.983020
SEXMALE      0.53657    0.51964   1.033 0.301796
BKNOBIRD     1.35212    0.40837   3.311 0.000929 ***
AG           0.03884    0.03528   1.101 0.270881
YR          -0.07187    0.02610  -2.754 0.005887 **
CD          -0.02632    0.02545  -1.034 0.301084
---
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: 154.25  on 141  degrees of freedom
AIC: 166.25
```

P-value is .82 so we fail to reject the null hypothesis therefore can conclude that the reduced model is preferable to the the full model

---

Removing Sex from the model:

H0: $\beta 2 = 0$

Ha: Some $\beta i \neq 0$

```
Lmodel_reduced1 <- glm(LC ~ SEX + BK + AG + YR + CD, family=binomial, data=df)
Lmodel_reduced2 <- glm(LC ~ BK + AG + YR + CD, family=binomial, data=df)
summary(Lmodel_reduced2)

#red2
155.32
142

# red1
154.25
141

diff <- 155.32-154.25

1 - pchisq(diff, 1)
0.3009454
```

```
Call:
glm(formula = LC ~ BK + AG + YR + CD, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.07408    1.64254   0.045 0.964025
BKNOBIRD     1.40754    0.40368   3.487 0.000489 ***
AG           0.04071    0.03470   1.173 0.240698
YR          -0.06561    0.02484  -2.642 0.008248 **
CD          -0.02375    0.02505  -0.948 0.343005
---
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: 155.32  on 142  degrees of freedom
AIC: 165.32
```

P-value is .300 so we fail to reject the null hypothesis therefore can conclude that the reduced model #2 is preferable to the original reduced model.

Removing CD from the model:

H0: $\beta 4 = 0$

Ha: Some $\beta i \neq 0$

```
Lmodel_reduced2 <- glm(LC ~ BK + AG + YR + CD, family=binomial, data=df)
Lmodel_reduced3 <- glm(LC ~ BK + AG + YR, family=binomial, data=df)

summary(Lmodel_reduced3)
#red3
155.32
142

# red2
156.22
143

diff <- 156.22-155.32

1 - pchisq(diff, 1)
0.3427817
```

```
Call:
glm(formula = LC ~ BK + AG + YR, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34296    1.58002  -0.217 0.828159
BKNOBIRD     1.37656    0.40073   3.435 0.000592 ***
AG           0.04610    0.03430   1.344 0.178952
YR          -0.07485    0.02296  -3.261 0.001111 **
---
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: 156.22  on 143  degrees of freedom
AIC: 164.22
```

P-value is .343 so we fail to reject the null hypothesis therefore can conclude that the reduced model #3 is preferable tol reduced model #2.

Removing AG from the model

H0: $\beta 3 = 0$

Ha: Some $\beta i \neq 0$

```
Lmodel_reduced3 <- glm(LC ~ BK + AG + YR, family=binomial, data=df)
Lmodel_reduced4 <- glm(LC ~ BK + YR, family=binomial, data=df)

summary(Lmodel_reduced4)
#red4
158.11
144

# red3
156.22
143

diff <- 158.11-156.22
1 - pchisq(diff, 1)
0.1692019
```

```
Call:
glm(formula = LC ~ BK + YR, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(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

(Dispersion parameter for binomial family taken to be 1)

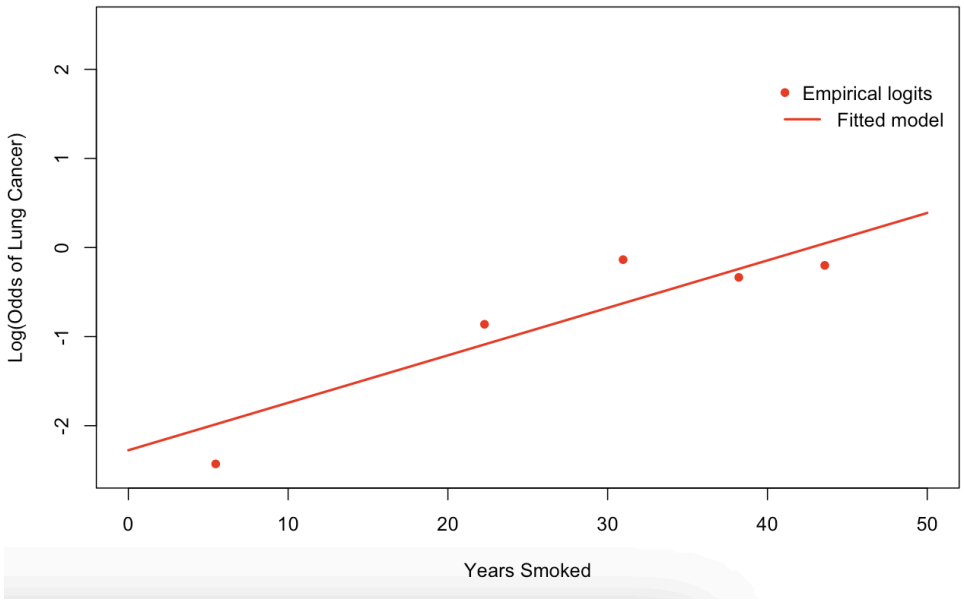
    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 158.11  on 144  degrees of freedom
AIC: 164.11
```

P-value is .169 so we fail to reject the null hypothesis therefore can conclude that the reduced model #4 is preferable tol reduced model #3. Since further reducing the model remove a significant predictor (YR), we can conclude running further nested likelihood ratio tests.

Logistic Regression Conditions:

Before concluding with the final model, I will check logistical regression conditions for the final model $log(odds\ Lung\ Cancer = 1.70460 + 1.47555(BKNOBIRD) - 0.05825(YR)$. First, I must check linearity conditions via an empirical logit plot. The linearity condition states that the logits (log odds) should have a linear relationship with the predictors. For multiple logistic model, we check for linearity of each continuous predictor, which in this model is years smoked (YR). The empirical logit plot from figure 4 is linear , so we can pass this condition.

Figure 4: Empirical logit plot of Years Smoked versus the Log Odds of Lung Cancer

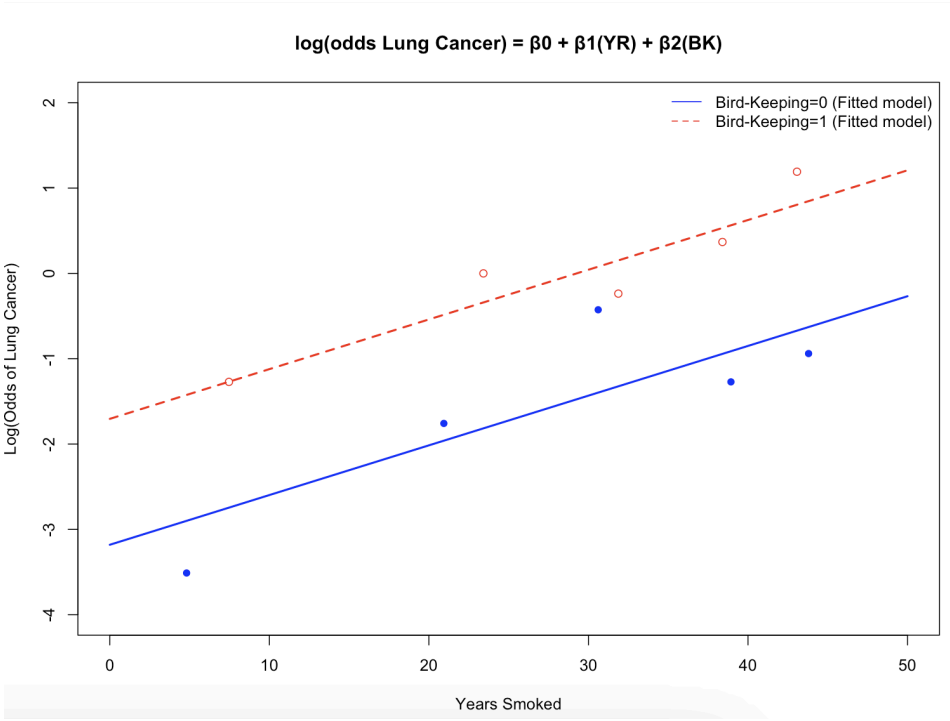



```
curve(predict(lmodCommon, data.frame(YR_numeric=x)), col="red",lwd=2, lty=1,
      xlab="Years Smoked", ylab = "Log(Odds of Lung Cancer)", xlim = c(0,50),
      ylim = c(-2.5, 2.5))
points(TableCommon$Logit~TableCommon$XMean,ylab="Log(Odds of Lung Cancer)",
       xlab="Years Smoked",pch=16,col="red",  ylim = c(2,6))
legend(40, 2, "Empirical logits", col = "red", pch = 16,  bty = "n")
legend(40, 1.7, "Fitted model", col = "red", lty=1,  bty = "n", lwd=2)

lmodParallel = glm(LC_numeric ~ YR_numeric + BK_numeric, family=binomial,data=df)
summary(lmodParallel)
```

The next two conditions are independence and randomness: For independence, there should be no pairing or clustering of data. Since this is contingent on the methods of data collection, we can assume that the independence condition passes. The randomness condition is also contingent on how the data was collected. However, since the sample size is comprised of two groups, those with cancer and those without cancer, then we should be hesitant about the population that these statistics can be assumed onto. If we are assuming that all conditions pass then our reduced model #4 is our final model and the likelihood of having lung cancer can be expressed in the equation: $log(odds\ Lung\ Cancer = 1.70460 + 1.47555(BKNOBIRD) - 0.05825(YR)$.

Here is the final fitted model:




```
Table1 = emplogitplot1(LC_numeric ~ YR_numeric, ngroups = 5,
                       data=subset(df, BK_numeric == 1), showplot=FALSE,out=TRUE)
Table0 = emplogitplot1(LC_numeric ~ YR_numeric, ngroups = 5,
                       data=subset(df, BK_numeric == 0), showplot=FALSE,out=TRUE)


plot(Table1$Logit~Table1$XMean,ylab="Log(Odds of Lung Cancer)",xlab="Years Smoked",
     pch=1,col="red",  xlim = c(0,50),ylim = c(-4, 2),
     main = "log(odds Lung Cancer) = β0 + β1(BK) + β2(YR)")
points(Table0$Logit~Table0$XMean,ylab="Log(Odds of Lung Cancer)",xlab="Years Smoked",
       pch=16,col="blue",  ylim = c(-4, 2))
curve(predict(lmodParallel, data.frame(YR_numeric=x, BK_numeric=1)), col="red",lwd=2,
      lty=2, xlab="YR", ylab = "Log(Odds of Lung Cancer)", xlim = c(0,50),  add=TRUE)
curve(predict(lmodParallel, data.frame(YR_numeric=x, BK_numeric=0)), col="blue",lwd=2,
      lty=1, xlab="YR", ylab = "Log(Odds of Lung Cancer)", xlim = c(0,50),   add=TRUE)
legend("topright", legend = c("Bird-Keeping=0 (Fitted model)",
                              "Bird-Keeping=1 (Fitted model)"), lty = c(1,2),
                              col=c("blue", "red"), bty="n")
```