Lab 12: Logistic Regression

November 11, 2018

Generalized linear models (GLM's) are an extension of regular linear models (i.e., linear regression and ANOVA) to situations where the probability model is not a normal distribution. In the last lab, we studied Poisson regression and log-linear models. In this lab, we focus on two cases of logistic regression. The first is the case where the response variable is binary (presence/absence or 1/0). The second is the case where the responses are counts, in which we model the number of "successes" out of a specified number of possibilities.

The models are used to answer questions such as: "Which indicators of health describe whether humans are infected (infection = yes/no) with malaria?" Or, "what predictors determine the proportion of death penalty verdicts that were overturned in each of 50 states".

The learning goals of the lab are to:

- Understand the situations in which logistic regression is used
- Learn to implement both models of logistic regression in R
- Understand how to interpret the GLM coefficients and make predictions from the models.

At the end of the lab, there are a few problems to answer. *Submit your R markdown document to the class Sakai site under the folder Assignments before 11:55 pm on Mon., Nov. 26 (all sections).*

More functions in R

- `inv.logit()` - given a numeric object, this function from the `boot` package returns the inverse logit of the values.

Logistic Regression for Binary Responses

Let's try a logistic regression model on the Bumpus sparrow data. House sparrows were found on the ground after a severe winter storm in 1898, some of which survived and others of which perished. We are going to analyze the explanatory variables that determine whether a sparrow survived or not. Although there are lots of different morphometric data, we will limit our analysis to bird weight, `WT`, and age, `AG`. Age is coded as adult (1) or young (2). If you are interested in trying the other variables, they include: total length, `TL`, alar extent, `AE`, length of beak and head, `BH`, length of humerus, `HL`, length of femur, `FL`, length of tibio-tarsus, `TT`, width of skull, `SK`, and length of keel of sternum, `KL`.

After loading the data, we have to transform the `Status` column into a numerical variable, `SV`, where 1 = survived and 0 = died.

```r
require(boot)
 require(Sleuth3)

 sdat<- ex2016
  sdat$SV <- ifelse(sdat$Status == "Survived", 1, 0)
   attach(sdat)
```

Fit the full model

Because we are only going to assess two variables, one of which is a factor, it isn't necessary to do the pairwise scatter plots. However, if you do plot all the data, you will notice that many of the variables are highly correlated – which is expected with morphological measurements. Here we run three models - a full model with an interaction, a main effects model, and a model with just one main effect. Note that we set the probability distribution of the model to binomial for logistic regression (`family=binomial`).

```
lr1 <- glm(SV ~ factor(AG)*WT, family=binomial, data=sdat)
 lr2 <- glm(SV ~ factor(AG)+WT, family=binomial, data=sdat)
  lr3 <- glm(SV ~ WT, family=binomial, data=sdat)
```

Compare nested models

As with Poisson Regression, we will use the Likelihood Ratio Test to compare nested models. What is your conclusion from this test?

```
anova(lr1, lr2, lr3, test="Chisq")
```

We could also use AIC to select the best model.

```
AIC(lr1, lr2, lr3)
```

The most parsimonious model, `lr3`, is the minimum adequate model. Removing the interaction term `AG x WT` and then `AG` did not make statistically significant differences to the model. The AIC comparison also shows `lr3` to have the lowest AIC. `lr2` is not more than 2 AIC points from `lr3`, indicating that they fit the data equally well. We prefer the most parsimonious model.

Diagnose model fit

The χ^2 test of deviance shows evidence for lack-of-fit of the model to the data. In other words, the fitted values are significantly different from the observed values. (Goodness-of-fit tests are based on the premise that the data will be divided into subsets and within each subset the predicted number of outcomes will be computed and compared to the observed number of outcomes.)

```
lr3$deviance/lr3$df.residual
pchisq(lr3$deviance, lr3$df.residual, lower=F)
```

We could also calculate the Hosmer-Lemeshow goodness of fit test. The test assesses whether or not the observed rates of survival match the expected rates of survival in subgroups of the model population. The result is similar to the test of deviance.

```
HLgof.test(fit = fitted(lr3), obs = SV)
```

We are going to ignore the lack of fit for now, mostly because we don't have many options with a model that only has one variable. The result indicates that there is still extra variation to be explained, and so perhaps we are missing an important predictor (one of the many others in the dataset).

We can calculate a pseudo R^2 for logistic regression, using the equation $1 - \mathcal{L}_M/\mathcal{L}_0$, where $\mathcal{L}_M$ is the log-likelihood of the model of interest and $\mathcal{L}_0$ is the model with just the intercept. This is McFadden's pseudo R^2, and it and others can be calculated with the `pR2` function from the `pscl` package. It supports our conclusion that this model is not very informative.

```
require(pscl)
 lr0 <- glm(SV ~ 1, family=binomial, data=sdat)
  pseudoR2 <- 1-logLik(lr3)/logLik(lr0)
   pR2(lr3)
```

Interpret parameters

The coefficients of the fitted model can be interpreted in two different ways – as log-odds and odds. The coefficients provided by `glm()` are log-odds. By taking the anti-log of the coefficients, we can interpret them as odds ratios.[1]

In our example, a 1-unit difference in weight corresponds to a multiplicative change of $e^{-0.42} = 0.657$ in the odds of survival. So the coefficient of `WT` can be interpreted as: "for every gram of additional weight the odds of survival change by a factor of 0.65 (or decrease by 35%)". Another way of saying this is that for every additional gram of weight, the odds of failure increase by 1.5 times.

```
1/exp(lr3$coef[2])
```

We can take a look at the above answer by substituting in values for weight. The equation from our model is:

$$log(\frac{p}{1-p}) = logit(p) = 11.3201 - 0.4244 \cdot WT$$

The mean weight is approximately 25 grams, so let's look at the effect of a change from 25 to 26 grams on the odds of survival.

```
cf <- coefficients(lr3)
 (cf[1]+cf[2]*26)-(cf[1]+cf[2]*25)
```

For a one-unit increase in weight, the expected change in log-odds is -0.424. By taking the anti-log, we get a change in odds of 0.65, or a 35% decrease in odds of survival for each additional gram of weight.

We can then calculate the probability of survival for a specific weight or weights. To convert log-odds to probabilities, we use the inverse logit: $1/(1+e^{-x})$. In R, the `inv.logit` function will calculate the probabilities for us. For example, the probability of survival for a bird weighing 30 grams is 19.6%. Note that to predict the response for a specific value of the covariates, the inverse logit is performed on the entire equation.

```
coef.prob1 <- c(1/(1+exp(-cf[1])), 1/(1+exp(-cf[2])))
 coef.prob1 <- inv.logit(coef(lr3))

 bird.30 <- inv.logit(cf[1] + cf[2]*30)
 bird.31 <- inv.logit(cf[1] + cf[2]*31)

 bird.31 - bird.30
```

We can evaluate how a difference in 1 gram of weight would affect the probability of survival between birds that are 30 and 31 grams. The probability of survival of a bird weighing 31 grams is 13.7%, compared to 19.6% for a 30 gram bird. This is a change of 5.8% change in survival.

We could quickly find the change in probability of survival over the possible range of bird weights with the following code. What do you notice?

[1] If an outcome has a probability p, then $p/(1-p)$ is the odds of the outcome.

```r
newdat <- data.frame(WT = c(23:31))
newdat$predSV <- predict(lr3, newdat, "response")
newdat$changeSV <- c(0, newdat$predSV[1:8] - newdat$predSV[2:9])
```

Finally, we can measure model accuracy by determining the proportion of observations that have been correctly classified as 'survived' from the model. Line `problr3` below extracts the probability of survival over all the observed weights in the raw data based on our simple model. If the probability is greater than 0.5, then it signifies that the bird survived. `pred.classes` codes the probabilities as 1 (survived) or 0 (perished) based on the probabilities, and then we evalute the number of times predicted survival was equal to the observed survival. The proportion of correctly classified observations is ~59.8%, which is not great.

```r
problr3 <- lr3 %>% predict(sdat, type = "response")
pred.classes <- ifelse(problr3 > 0.5, 1, 0)

mean(pred.classes == sdat$SV)
```

Just to see if a more complicated model would improve model goodness of fit, pseudo-R^2 and accuracy, let's build a more complicated model and run all of those assessments again. What do you find?

```r
lr4 <- glm(SV ~ WT + TL + AE + TT + SK, family = binomial)
 pR2(lr4)
  HLgof.test(fit = fitted(lr4), obs = sdat$SV)

problr4 <- lr4 %>% predict(sdat, type = "response")
pred.classes <- ifelse(problr4 > 0.5, 1, 0)

mean(pred.classes == sdat$SV)
```

In summary:

- As with linear regression, the intercept is estimated assuming zero values for the other predictors. The intercept is not usually interpreted in a logistic regression model.
- A difference of 1 gram in weight corresponds to a negative difference of 0.424 in log-odds of bird survival. By taking the antilog of -0.424, the change can be interpreted as a change in odds.
- We can predict the probability of survival for different scenarios (e.g. birds at different weights). This is done by fitting our logistic regression equation with different values of the predictors and then taking the inverse logit. Note that it is difficult to interpret the slope coefficients (e.g. weight) in terms of the change in probability with each one-unit change in the predictor variable because the relationship is non-linear (see below figures).
- Model fit can be assessed in several ways, including the test of deviance, McFadden's pseudo-R^2, and the Hosmer-Lemeshow statistic.
- Model accuracy can be assessed by predicting the proportion of times the model currently predicts observed survival.

Plotting Results

For most people, probabilities are more intuitive than odds, so let's plot the data and the probability of survival over weight of the birds. The fitted function provides the probabilities from the model by calculating the inverse of the logit function (the logistic function). If $logit(\pi) = \eta$, then $\pi = e^{\eta}/(1 + e^{\eta})$. These are the expected probabilities ($\hat{y}$) of the observed data given the model.

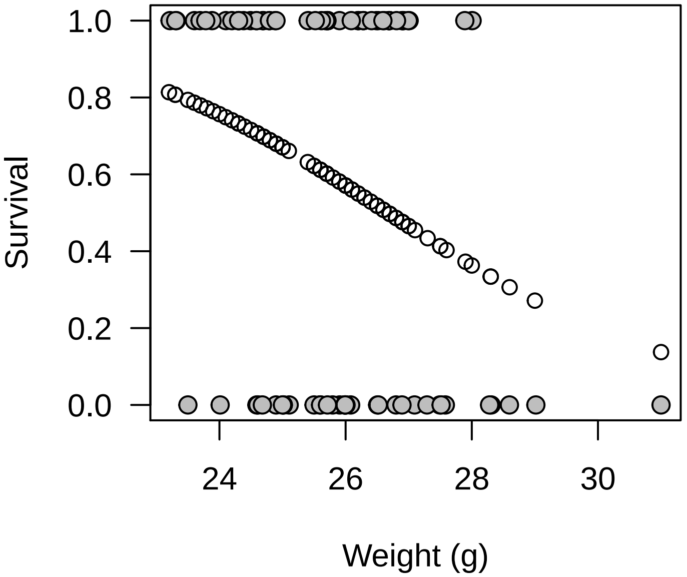
```r
fitted(lr3)
```

The `predict()` function returns predictions from a new set of predictor variables. If you don't specify a new set of predictor variables, then it will use the original data giving the same results as fitted for some models.[2] `predict` returns the fitted values *before* the inverse of the link function is applied (to return the data to the same scale as the response variable), and `fitted` shows it *after* it is applied. To get results on the original (response) scale, you must use `predict(model, type = "response")`.

```
predict(lr3, type = "response")
```

First, we will plot the raw data and the fitted values from the model.

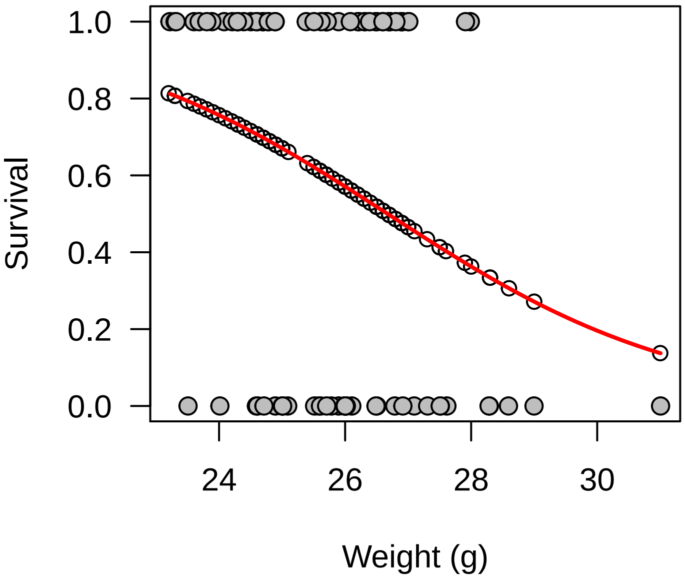
```
plot(jitter(WT), SV, las=1, pch=21, cex=1.2, bg="grey",
     xlab = "Weight (g)", ylab="Survival")
 points(WT, fitted(lr3))
```



To add the fitted curve to the plot we first define x as a range of values from the minimum bird weight to the maximum weight. Then we fit the curve, using `inv.logit()` to transform log-odds to proportions to fit our plot.

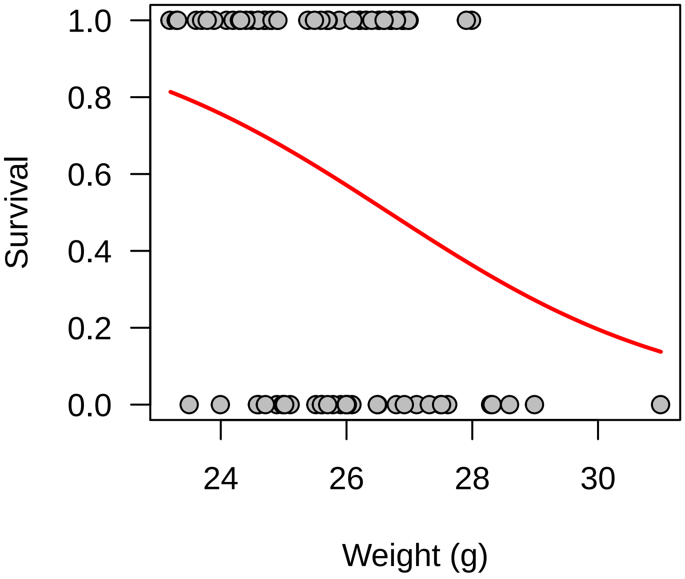
```
plot(jitter(WT), SV, las=1, pch=21, cex=1.2, bg="grey",
     xlab = "Weight (g)", ylab="Survival")

 x <- seq(min(WT), max(WT), length = 100)
  points(WT, fitted(lr3))

 curve(expr = inv.logit(lr3$coef[1] + lr3$coef[2]*x), add=T,
       lwd=2, col="red")
```

[2] Note that by default predict() calculates the log-odds, $logit(\pi)$, by sticking in values of the predictor variables, e.g. X_1, into the equation: $\beta_0 + \beta_1 X_1$.



We could make the same plot using the `predict()` function.

```
plot(jitter(WT), SV, las=1, pch=21, cex=1.2, bg="grey",
     xlab = "Weight (g)", ylab="Survival")

 x <- seq(min(WT), max(WT), length = 100)

  lines(x, predict(lr3, data.frame(WT=x),
                type = "response"), col = "red", lwd=2)
```



Logistic Regression for Count Data

Sometimes, we may have count data that need to be treated as proportions rather than simple counts, such as when we have a number of “successes” out of a total count (or number of trials). Examples include:

- Proportion of planted seeds that recruited into seedlings
- Proportion of death penalty verdicts overturned
- Proportion of population that survived a disease.

The model, called a logistic-binomial model, is used in settings where each data point represents the number of successes in some number of tries.

Let's look at an example where randomly chosen people were asked to respond to the following statement regarding the role of women in society: "Women should take care of running their homes and leave the running of the country up to men". Install and load the `HSAUR` package to get the data.

```r
require(HSAUR)
```

The observations have been grouped into counts of number of respondents who `agree` with the statement and number who `disagree` with the statement. The `sex` and years of `education` of the respondents are the predictor variables. To fit a logistic regression model to grouped data using the `glm` function, we need to specify the number of agreements and disagreements as a two-column matrix on the left hand side of the model formula with `cbind()`.

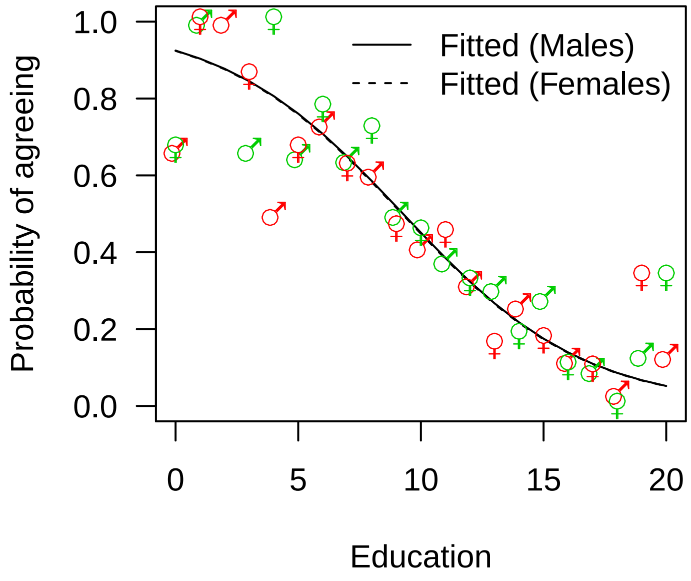
```r
logreg1 <- glm(cbind(agree, disagree) ~ factor(sex) + education,
               data = womensrole, family = binomial)
```

From the summary of `logreg1`, `education` significantly predicts whether a respondent will agree with the statement, but `sex` is unimportant. As a respondent's years of education increase, his/her probability of agreeing with the statement decreases. While using the z-value of the coefficient on `sex` is enough to determine that we don't need it in the model, we could also verify by running the model with and without `sex` and looking at the change in deviance.

```r
logreg2 <- glm(cbind(agree, disagree) ~ education,
               data = womensrole, family = binomial)
 anova(logreg1, logreg2, "Chisq")
```

The change in deviance from removing `sex` is very small (<1) confirming our previous conclusion. Let's look at how the probability of agreeing to the statement varies over education level for both men and women.

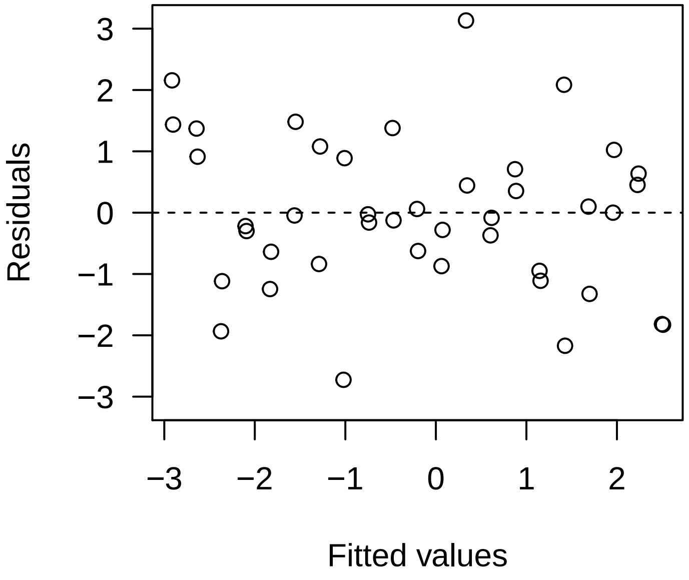
```r
lr1.fitted <- predict(logreg1, type = "response")
 f <- womensrole$sex == "Female"

plot(womensrole$education, lr1.fitted, type = "n",
     ylab = "Probability of agreeing",
      xlab = "Education", ylim = c(0,1), las = 1)

lines(womensrole$education[!f], lr1.fitted[!f], lty = 1)
 lines(womensrole$education[f], lr1.fitted[f], lty = 2)
  lgtxt <- c("Fitted (Males)", "Fitted (Females)")
   legend("topright", lgtxt, lty = 1:2, bty = "n")

   y <-  womensrole$agree / (womensrole$agree +
                            womensrole$disagree)
text(womensrole$education, y, ifelse(f, "\\VE", "\\MA"),
     family = "HersheySerif", cex = 1.25, col = c(2,3))
```



We can also check out the fit of `logreg1` by using the typical diagnostics. Try `plot(logreg1)`. Note that these diagnostics are not helpful for assessing the fit of the binary logistic regression (our first example) because the data were 0's and 1's, whereas here the data are counts.

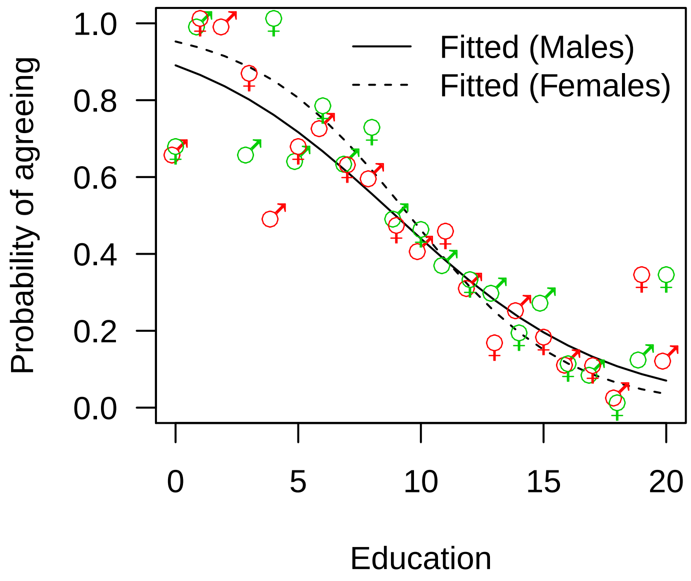
We can also create our own plot of deviance residuals plotted against fitted values using the below code. Most of the residuals fall into a horizontal band between -2 and 2 (standard errors). We expect 5% of the observations to fall outside of this band just by chance, so this pattern does not suggest a poor fit overall.

```
res <- residuals(logreg1, type = "deviance")
 plot(predict(logreg1), res, xlab="Fitted values", las = 1,
          ylab = "Residuals", ylim = max(abs(res)) * c(-1,1))
abline(h = 0, lty = 2)
```



As before with linear regression, we can also try to model the interaction between *sex* and *education*.

```
logreg2 <- glm(cbind(agree, disagree) ~ education*factor(sex),
               data = womensrole, family = binomial)
```



The interaction model leads to a deviance that is nearly 7 points lower than the main effects model. The interaction between `sex` and `education` is also statistically significant. We can plot this in the same way as above, but this time we see the effect of the interaction on the probability of agreeing to the statement.

To Do

How is this figure different from the above figure with the model without the interaction?

The interaction term can be understood in two ways. Looking from one direction, for a female, the value -0.08138 is added to the coefficient for `education` (-0.23403), and so we can see the interaction as saying that the importance of education as a predictor increases for respondents that are female (i.e., it has an even stronger negative effect on the probability of agreeing).

Looking at the interaction from the other way, for each year of education the value -0.08138 is added to `sex` (0.90474 for a female). So we can understand the interaction as saying that the importance of `sex` as a predictor decreases as respondents gain education.

We can interpret the coefficients as log-odds or odds as before.

Note that like Poisson Regression, when logistic regression is applied to count data it is possible for the data to have more variation than is explained by the model. (This was not the case for binary logistic regression, because the response variable was constrained to 0 and 1.) This overdispersion problem arises because the model does not have a variance parameter. As with the Poisson model, we can compute the estimated overdispersion and adjust our inference if the model is overdispersed. Here we implement the quasibinomial model.

```
logreg3 <- glm(cbind(agree, disagree) ~ education*factor(sex), data = womensrole,
               family = quasibinomial)
```

Then we can extract the dispersion estimate from the model and determine if the degree of overdispersion is significantly different from 1. Dispersion is 1.96, and the probability of getting that high of a dispersion is less than 0.001 givent the model. Therefore we should report the results from `logreg3`, rather than `logreg2`.

```r
pchisq(summary(logreg3)$dispersion * logreg3$df.residual,
       logreg3$df.residual, lower = F)
```

Problems

Your assignment is to conduct the below logistic regression (see descriptions). As always, turn in a R markdown document that includes your R code as an Appendix.

Write a 1-page description of your analysis, results, and inference. The write-up should include the following information:

- Null and alternative hypotheses of your test.
- Results of your model selection process. What is the minimum adequate model?
- Interpretation of your statistical test. Test whether your final model is a good fit to the data. Interpret the model coefficients, writing 2-3 sentences that include the appropriate reporting of the statistics.
- A description of how you checked the assumptions of your statistical test.
- A figure that demonstrates the results of your test/model. Specifically, the figure should show the (a) raw data, (b) curves that demonstrate the effect of bird-keeping over the range of one of the remaining significant predictors. Include a legend on the figure and make sure the axes are labeled appropriately.

Problem 1

Use the data from the package `Sleuth3`, `case2002`, to examine whether increased lung cancer is associated with birdkeeping, even after accounting for factors such as age. Develop a logistic model with the log-odds of getting lung cancer explained by age `AG`, years the individual has smoked `YR`, the indicator variable for birdkeeping `BK`, *and their interactions*. There are other variables in the database, which you do not need to consider. Specifically, test the hypothesis that birdkeepers have higher rates of lung cancer than non-birdkeepers. As requested above, plot the probability of having lung cancer for birdkeepers and non-birdkeepers. Finally, answer the following:

- What is the probability of having lung cancer if you are a bird keeper and have smoked for 32 years?
- What is the probability of the intercept? What does it represent exactly?