STAT/BIOSTAT 534 Statistical Computing
Spring Quarter 2017
Homework 2

Adrian Dobra
adobra@uw.edu

This homework is due on Friday, April 14 at 11:00pm. You should use the dropbox to submit your code.

The file "534binarydata.txt" has $n = 148$ samples (rows) and 61 variables (columns). The last column is a binary variable y considered to be the response (outcome). The first 60 variables correspond to $p = 60$ explanatory variables. They are denoted by $x_1, x_2, \ldots, x_p$. This homework is focused on performing variable selection for logistic regression. Your handout "logisticregression.pdf" gives you all the background knowledge you need to successfully complete this assignment. Feel free to use the R code "bic-logistic.R" as much as you need in the code you will write.

We identify a logistic regression model by the indices of the explanatory variables it contains. Denote $V = \{1, 2, \ldots, p\}$. For a subset A of V, the logistic regression of y given $x_A = \{x_i : i \in A\}$ is denoted by $\mathcal{M}_A$ and it is written as

$$\mathcal{M}_A : \quad \log \frac{P(y = 1 \mid x_A)}{P(y = 0 \mid x_A)} = \beta_0 + \sum_{i \in A} \beta_i x_i$$

The full logistic regression model $\mathcal{M}_V$ contains all the explanatory variables, i.e.

$$\mathcal{M}_V : \quad \log \frac{P(y = 1 \mid x_1, \ldots, x_p)}{P(y = 0 \mid x_1, \ldots, x_p)} = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_p \cdot x_p.$$

The empty logistic regression model $\mathcal{M}_\emptyset$ does not contain any explanatory variables, i.e.

$$\mathcal{M}_\emptyset : \quad \log \frac{P(y = 1)}{P(y = 0)} = \beta_0$$

The number of all possible logistic regressions is $2^p = 2^{60}$ since this is the number of subsets of V. In this homework we will implement two algorithms that attempt to identify the

logistic regression model with the best (smallest) AIC score (see page 7 of your handout "logisticregression.pdf"). In other words, the two algorithms attempt to solve the optimization problem:

$$\min\{AIC(\mathcal{M}_{\mathcal{A}}) : A \subseteq V\} \tag{1}$$

These two algorithms are not guaranteed to reach the solution of this optimization problem, i.e. they might fail to output the logistic regression with the smallest AIC. We have already discussed in class (see the file "bic-logistic.R") an algorithm that generates all the possible logistic regressions and reports the logistic regression with the best BIC (the choice of AIC or BIC does not change how we visit regressions). For this particular dataset, the huge number of possible models (2^{60}) makes the generation of all possible regressions infeasible, hence the code "bic-logistic.R" is not applicable. Instead, we must rely on algorithms that give approximate solutions to problem (1). Examples of two such algorithms are given in Problems 2 and 3. If these two algorithms output the same logistic regression, we might be encouraged to believe we actually solved the optimization problem (1). Unfortunately this is not theoretically true. In most cases, however, we will need to rely on algorithms that give approximate solutions to optimization problems if the exact solutions are not available (think about calculating the MLEs with Newton-Raphson).

Problem 1 (10 points)

Consider the following R code that computes the BIC of a logistic regression with the binary outcomes associated with column "response" and explanatory variables associated with columns "explanatory" of the data matrix "data". Please note the special handling of the empty logistic regression $\mathcal{M}_{\emptyset}$.

```
#this function uses 'glm' to fit a logistic regression
#and returns the BIC = deviance + log(SampleSize)*NumberOfRegressionCoefficients
getLogisticBIC <- function(response,explanatory,data)
{
  #check if the regression has no explanatory variables
  if(0==length(explanatory))
  {
     #regression with no explanatory variables
     deviance = glm(data[,response] ~ 1,family=binomial(link=logit))$deviance;
  }
  else
  {
    #regression with at least one explanatory variable
    deviance = glm(data[,response] ~ as.matrix(data[,as.numeric(explanatory)]),
                                   family=binomial(link=logit))$deviance;
  }
```

```
  return(deviance+log(nrow(data))*(1+length(explanatory)));
}
```

From page 7 of your handout "logisticregression.pdf", you see that the formula for computing the BIC is

$$BIC(\mathcal{M}_A) = D(\mathcal{M}_A) + \log(n) \cdot |\mathcal{M}_A|,$$

where $D(\mathcal{M}_A)$ is the deviance of $\mathcal{M}_A$, $|\mathcal{M}_A|$ represents the number of regression coefficients in $\mathcal{M}_A$ and n is the sample size. Your task is to write a function `getLogisticAIC` that computes the AIC of a logistic regression $\mathcal{M}_A$:

$$AIC(\mathcal{M}_A) = D(\mathcal{M}_A) + 2 \cdot |\mathcal{M}_A|,$$

Problem 2 (40 points)

You are asked to implement the following *forward* greedy procedure that attempts to find logistic regressions with small AIC values. The procedure is called "greedy" because, at each iteration, we consider only the model that leads to the biggest decrease in the AIC of the current model if such a model exists.

Denote by $\mathcal{M}_{A^{(r)}}$ ($A^{(r)} \subseteq V$) the current model at iteration r of the algorithm. At iteration 0 we start with the empty logistic regression $\mathcal{M}_\emptyset$, i.e. $A^{(0)} = \emptyset$. Iteration r proceeds as follows:

1) If all the variables are currently in the model (i.e., if $A^{(r)} = V$), STOP. Since there are no variables that could be added to the current model, the algorithm must end.

2) The variables that do not belong to the current model $\mathcal{M}_{A^{(r)}}$ are $\{x_i : i \in V \setminus A^{(r)}\}$. Consider all the logistic regressions that are obtained by adding one variable x_i, $i \in V \setminus A^{(r)}$, to the current model:

$$\mathcal{F}(r) = \{\mathcal{M}_{A^{(r)} \cup \{i\}} : i \in V \setminus A^{(r)}\}.$$

Use the function `getLogisticAIC` to calculate the AIC of all the models in $\mathcal{F}(r)$. Note: the R functions `union`, `setdiff` or `is.element` might prove to be useful (use the function `help()` to get more information about them).

3) Find the logistic regression $\mathcal{M}_{A^{(r)} \cup \{i_0\}}$ in $\mathcal{F}(r)$ that has the smallest AIC.

4) If the AIC of $\mathcal{M}_{A^{(r)} \cup \{i_0\}}$ is smaller than the AIC of $\mathcal{M}_{A^{(r)}}$ (i.e., if we can improve the current model by including variable x_{i_0}), set $A^{(r+1)} = A^{(r)} \cup \{i_0\}$ and proceed to the next iteration. Otherwise, STOP (the AIC of the current model cannot be improved by the addition of any variable).

The above procedure always ends since we either added all possible variables to the model

or we cannot improve the current model by adding a variable. Nevertheless, there does not exist any theoretical guarantee that the final model that is obtained from this procedure is actually the logistic regression model with the smallest AIC.

Problem 3 (40 points)

You are asked to implement the following *backward* greedy procedure that also attempts to find logistic regressions with small AIC values. As in Problem 2, denote by $\mathcal{M}_{A^{(r)}}$ ($A^{(r)} \subseteq V$) the current model at iteration r of the algorithm. At iteration 0 we start with the full logistic regression $\mathcal{M}_V$, i.e. $A^{(0)} = V$. Iteration r proceeds as follows:

1) If the current model does not include any variables (i.e., if $A^{(r)} = \emptyset$), STOP. Since there are no variables that could be deleted from the current model, the algorithm must end.

2) Consider all the logistic regressions that are obtained by deleting one variable x_i, $i \in A^{(r)}$, from the current model:

$$\mathcal{B}(r) = \{\mathcal{M}_{A^{(r)} \setminus \{i\}} : i \in A^{(r)}\}.$$

Use the function `getLogisticAIC` to calculate the AIC of all the models in $\mathcal{B}(r)$.

3) Find the logistic regression $\mathcal{M}_{A^{(r)} \setminus \{i_0\}}$ in $\mathcal{B}(r)$ that has the smallest AIC.

4) If the AIC of $\mathcal{M}_{A^{(r)} \setminus \{i_0\}}$ is smaller than the AIC of $\mathcal{M}_{A^{(r)}}$ (i.e., if we can improve the current model by deleting variable x_{i_0}), set $A^{(r+1)} = A^{(r)} \setminus \{i_0\}$ and proceed to the next iteration. Otherwise, STOP (the AIC of the current model cannot be improved by the deletion of any variable).

The above procedure always ends since we either deleted all the variables from the logistic regression model or we cannot improve the current model by deleting a variable. Again, there does not exist any theoretical guarantee that the final model that is obtained from this procedure is actually the logistic regression model with the smallest AIC.

Problem 4 (10 points)

Comment on the logistic regression models you identified in Problems 2 and 3. Are they the same? Do they have the same AIC? Repeat the forward and backward greedy procedures with the function `getLogisticBIC` instead of the function `getLogisticAIC` (i.e., calculate the BIC instead of the AIC). What logistic regressions do you obtain?