Lab 09: Logistic Regression

OBJECTIVES

1. Fit logistic regression models in R.

2. Familiarize yourselves with some options of logistic regression in R.

Logistic Regression is a type of predictive model that can be used when the dependent variable is a categorical variable with two categories –for example male/female, fail/pass, live/die, disease/non-disease, wins/doesn't win, etc. Thus, the dependent variable can take the value 1 with a probability of success (p), or the value 0 with a probability of failure (1-p). The independent or predictor variable can take any form (continuous, dichotomous and/or dummy variable with more than two categories). That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the independent and dependent variables is not a linear function as shown below:

$$\boldsymbol{p} = e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}/\{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}\}$$

where α= the constant of the equation and, β_i= the coefficients of the independent variables. The computed value, **p**, is a probability in the range of 0 to 1.

Odds are an expression of the likelihood of some event happens compared to the likelihood that it does not happen. Much of the interpretation of logistic regression model centers on the odds as follows:

$$\textbf{Odds} = \mathrm{p}/(1-\mathrm{p}) = \mathrm{e}^{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

where Odds can take on values between zero and infinity.

The logarithm of odds, **logit,** results in a linear model, **the logistic regression:**

$$\textbf{Logit} = \textbf{log (odds)} = \log\{p/(1\text{-}p)\} = \alpha+\beta_1 X_1+ \beta_2 X_2 +\ldots+\beta_i X_i$$

Interpretation of the parameters differs in logistic regression, as parameter estimates need to be back-transformed to be meaningful. To estimate a predicted probability, you must calculate $\alpha+\beta_1 X_1+ \beta_2 X_2 +\ldots+\beta_i X_i$ at the desired X_is to get the predicted **logit** first, and then exponentiate it to get the predicted **Odds** that can be back-transformed as:

The predicted probability (p) = predicted odds/(1+predicted odds)

The important tests generated by logistic regression are the "Tests of Global Null Hypothesis: Beta=0" and the "Analysis of Maximum Likelihood Estimates". The "Analysis of Maximum Likelihood Estimates" uses Wald statistics to test the null hypothesis H_0 that the associated parameter estimates are not equal zero. The "Tests of Global Null Hypothesis" are essentially tests of model significance, much like the model F-test for linear regression. Typically, the best test to use is the likelihood ratio test, which uses a Chi-square test of significance to test whether the slope parameter βs are significant different from zero. If this test is not significant, it indicates that the logistic regression is not an appropriate model for the experimental data.

LABORATORY INSTRUCTIONS

Part I.

Housekeeping Statements

Before we dive into the main part of the code it is good to create a pre-amble in which we will load all the necessary packages for R to execute the following tasks. If you have them installed already great. If not, you can install it using the "packages" tab on the bottom right panel. Click install and put the name of the package you want on the "install packages" window that pops up. The default setting is installing the packages from the CRAN repository where most "mainstream" packages can be found.

To activate the packages, use the following commands:

```
library(easypackages)
libraries("ggplot2","MASS","aod","ResourceSelection")
```

Dataset

The data set to be used is taken from a collection of data sets included in the textbook An Introduction to Categorical Data Analysis, written by Alan Agresti (John Wiley & Sons, Inc., New York, NY, 1996). The data consists only of data for the response (dependent) variable (a binary response of whether or not thermal distress was detected in a given O-ring seal on a space shuttle), and the explanatory (independent) variable (outside air temperature at time of shuttle launch), for a randomly and independently selected set of 23 shuttle launches. Two additional observations have been included in this dataset for purposes of estimating predicted responses to given values of the independent variable.

First, we need some codes to read the dataset, view it and perform some summary statistics. The following commands will do the trick:

```
ONE=read.table("data_lab9_r.txt", header = TRUE, sep = " ", dec = ".") #Reads the dataset and save it as the dataframe asphalt

View(ONE) #Views the dataset ONE

summary(ONE) # Gives the summary statistics of the dataset

sd(ONE$temp) # Computes standard deviation for the variable temp
```

The name of the Data frame is ONE.

Part II

Fitting a Logistic Model with glm

The following code will train our logistic model.

```
mylogit <- glm(distress ~ temp, data = ONE, family = "binomial") # Training of the logistic model

summary(mylogit) # Gives a summary of the model
```

In order to compute confidence intervals, we use the following commands:

```
confint(mylogit,level=0.99) # CIs using profiled log-likelihood
```

or the traditional one:

```
exp(confint.default(mylogit,level=0.99)) # Wald Confidence Intervals for Coefficients
```

To compute the corresponding Odds Ratios, we use:

```
exp(coef(mylogit,level=0.99)) # Odds Ratios
```

To get the Odds ratios and the corresponding confidence intervals:

```
exp(cbind(OR = coef(mylogit,level=0.99), confint(mylogit,level=0.99))) # Odds ratio and CI
```

In order to compare the actual results with the logit predictions we need to create a new dataframe as follows:

```
TWO=ONE
TWO$predictions=predict(mylogit,TWO,type = "response")
```

Basically we are copying dataframe ONE and we are adding the new column "predictions" by applying our predictive model to the original temp data using the command "predict"

Then in order for us to plot the predicted values vs the actual ones we need to use a simple scatterplot command from ggplot2 as follows:

```
plot1=ggplot(TWO, aes(x = distress, y = predictions)) +
  geom_point()+
  theme_classic()
plot1
ggsave("Scatter1.pdf",plot1)
```

Finally, to visualize our model, lets create a dummy dataset, called newdata with x values mimicking the range of the values in the original dataset, namely from 5 to 81. We will use increments of 0.01 steps and get the following code:

```
newdata=data.frame(temp=seq(5,81,0.01))
newdata$distress=predict(mylogit,newdata,type = "response")
```

The new distress column in the newdataset, comes from the predictions of our logit model. Let's go ahead and graph that using a simple ggplot code as follows:

```
plot2=ggplot(newdata, aes(x = temp, y = distress)) +
  geom_point()+
  theme_classic()
ggsave("Scatter2.pdf",plot2)
```

LAB ASSIGNMENT

Your assignment is to perform necessary analysis using R and answer the following questions (Please do not print all the output. Only print the graphs and tables that you think are relevant to your answers).

1.Write the logistic regression equation to model the odds of distress as a function of temperature.

2.Perform a logistic regression and report the regression parameters and their 99%confidence intervals.

3.Does temperature affect the odds of distress? Explain the reason for your answer.

4.What is the probability of distress at 66 degrees? How about at 35 degrees?

5.Plot the probability curve and describe it.

*Remember to attach your R code with your lab report.