STU33011 Lab 10

Make sure that you have completed the previous lab sessions (https://www.scss.tcd.ie/~arwhite/Teaching/STU33011.html) before advancing to this one.

In this session, we will

- Implement logistic regression models and interpret the output;
- Investigate how to implement models with interaction terms;
- Use the model output to make classification predictions;
- Briefly compare models.

Logistic Regression

In this lab we will use the pulse data. This can be found at https://www.scss.tcd.ie/~arwhite/Teaching/STU33011/pulse.txt. Download this file and load it into R using the `read.table` command, with option `header=TRUE`. Name the data frame `pulse`.

The `glm` function performs logistic regression. This function is a much used feature of R for when we are interested in fitting what is known as a generalized linear model.

As usual, spend some time examining the help file to see the arguments it takes and the output it returns.

If the pulse data has been saved and given the name `pulse`, then you can perform a simple logistic regression by entering the following:

```r
lreg <- glm(RestingPulse ~ Smokes + Weight, data = pulse, family = binomial(logit))
summary(lreg)
```

```r
## Call:
## glm(formula = RestingPulse ~ Smokes + Weight, family = binomial(logit),
##     data = pulse)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0159  0.3107  0.5851  0.7668  1.3858
## Coefficients:
##                  Estimate Std. Error     z value Pr(>|z|)
## (Intercept) -1.98717     1.67931    -1.183 0.23670
## SmokesYes -1.19297     0.55298    -2.157 0.03102 *
## Weight      0.02502     0.01226     2.042 0.04121 *
## ---
## Signif. codes:  <none>
## (Dispersion parameter for binomial family taken to be 1)
## Null deviance: 101.21  on 91  degrees of freedom
## Residual deviance: 93.64  on 89  degrees of freedom
## AIC: 99.64
## Number of Fisher Scoring iterations: 4
```
Exercise

• What interpretation do you gain from the summary output for the regression?

In the above code, the first argument specifies a linear model for predicting RestingPulse from the Smokes values and the Weight values. The second argument specifies that the pulse data is to be analysed. The final argument specifies that the linear model is applied to the logit of the probability for RestingPulse.

Now load the MASS package and load the data set birthwt. There is a help file for this data set that can also be called.

Unfortunately there are a few changes that need to be made to this data set before applying it within a regression.

For example, we should omit the actual birth weight data and ensure that certain variables are treated as categorical (factor), rather than numerical:

```r
newbirthdata <- birthwt[, -10]

is.factor(newbirthdata$race)
## [1] FALSE

newbirthdata$race <- as.factor(newbirthdata$race)

is.factor(newbirthdata$race)
## [1] TRUE
```

```r
for(i in 1:length(newbirthdata$ptl)){
    if(newbirthdata$ptl[i]>1){
        newbirthdata$ptl[i] <- 1
    }
}
newbirthdata$ptl <- as.factor(newbirthdata$ptl)
```

```r
for(i in 1:length(newbirthdata$ftv)){
    if(newbirthdata$ftv[i]>2){
        newbirthdata$ftv[i] <- 2
    }
}
newbirthdata$ftv <- as.factor(newbirthdata$ftv)
```

This code has ensured all variables that should be treated as categorical (factors) are indeed being treated as such. To ensure the results are the same as given in lectures we have also changed the ptl values so that they are indicators of whether or not there was a previous premature labour (rather than how many), and placed all ftv values that were greater than or equal to 2 into the same factor.

To perform a logistic regression we can now enter the following:

```r
lreg2 <- glm(low ~ ., data = newbirthdata, family = binomial(logit))
```

In the above the expression `low ~ .` is used as short hand to specify a linear model featuring all other covariates within the data set.

Including Interactions

Returning to the pulse data set, the interaction between smoking and weight can be included within the model by entering the following:
lreg3 <- glm(RestingPulse ~ Smokes + Weight + Smokes:Weight,
data = pulse, family = binomial(logit))

If there are many covariates and all interactions are required, then an easy way to specify the model is to use code resembling response~.*. as the first argument in glm.

Task

- Perform a logistic regression including interactions for newbirthdata.
- Comment on any warnings received and on any conclusions you can draw from the output. If you are unsure about this then please ask.

Prediction & Classification

To use the output from a logistic regression fit, we can again use its predict function. Try the following:

\[
\text{predict}(\text{lreg, pulse[, c(2, 3)]})
\]

```
## # 1 2 3 4 5 6
## # 1.5159947 1.64110770 0.82347270 1.57415086 1.89133375 2.14155980
## # 7 8 9 10 11 12
## # 1.76622073 2.76712493 2.89223796 1.46594946 0.82347270 1.89133375
## # 13 14 15 16 17 18
## # 0.64831446 1.64110770 2.26667283 2.39178586 1.19881178 1.07369875
## # 19 20 21 22 23 24
## # 1.32392480 1.39088165 2.26667283 1.94137896 1.26576862 1.44903783
## # 25 26 27 28 29 30
## # 1.51599467 0.07279454 1.46594946 -0.15240890 1.14065560 1.14065560
## # 31 32 33 34 35 36
## # 0.91545215 1.64110770 0.57324665 -0.37761235 1.14065560 2.76712493
## # 37 38 39 40 41 42
## # 1.89133375 1.07369875 1.89133375 3.39269006 0.57324665 0.44813362
## # 43 44 45 46 47 48
## # 1.89133375 1.89133375 1.76622073 0.69835967 1.76622073 1.32392480
## # 49 50 51 52 53 54
## # 2.01644678 1.39088165 2.01644678 0.07279454 0.69835967 0.57324665
## # 55 56 57 58 59 60
## # 1.71617552 1.89133375 1.76622073 0.32302059 2.51689888 1.57415086
## # 61 62 63 64 65 66
## # 1.64110770 0.57324665 0.92356312 1.5159947 1.56603988 1.41590425
## # 67 68 69 70 71 72
## # 1.09061039 1.89133375 1.26576862 1.01554257 1.26576862 0.09781715
## # 73 74 75 76 77 78
## # 1.01554257 0.96549736 1.14065560 0.19790757 1.14065560 0.96549736
## # 79 80 81 82 83 84
## # 1.06558778 0.89042954 0.89042954 0.89042954 1.76622073 0.76531652
## # 85 86 87 88 89 90
## # 0.91545215 -0.47770277 0.38997744 -0.05231848 1.34083644 0.76531652
## # 91 92
## # 1.76622073 0.71527131
```

You should notice that the results are neither 0 or 1 valued, nor are they even probabilities. This is because the results are given on the logit scale. To instead return probabilities we can enter the following:
Exercise:

- Using the regression coefficients previously found, check by hand the result from entering:

```r
predict(lreg, data.frame(Smokes="No", Weight=140), type="response")
```

```
## 1
## 0.8199479
```

To use the logistic regression fit as a classifier we simply select those data points which have a probability greater than 0.5:

```r
which(as.vector(predict(lreg, pulse[, c(2, 3)], type = "response")) > 0.5)
```

```
## [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## [47] 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71
## [70] 72 73 74 75 76 77 78 79 80 81 82 83 84 85 87 89 90 91 92
```

There are of course many possibilities for plotting, tabulating, and comparing your results, many of which you will have seen in previous work sheets. Once you have finished this lab, it would be a good idea to return to this section and see what graphical and table summaries you can produce. For example,

```r
res1 <- round(as.vector(predict(lreg, pulse[, c(2, 3)], type = "response")))
```
Model Comparison

As far as using the deviance information is concerned, this is best for comparing a nested sequence of models to determine if the additional parameters are justified by the data.

For example, consider the pulse data with no interaction term \texttt{lreg} and the model with an interaction term \texttt{lreg3}. The return from the following output suggests that there is no additional benefit of including the interaction (a non-significant value is returned).

\begin{verbatim}
1-pchisq(lreg$deviance-lreg3$deviance,lreg$df.residual - lreg3$df.residual)
\end{verbatim}

\begin{verbatim}
## [1] 0.9942037
\end{verbatim}

Exercise

- Create a new logistic model \texttt{lreg0} that only takes weight as a coefficient and determine if the deviance criteria suggests moving to the more advanced model that includes both smoke and weight information.
- Load the Salmon data and perform a classification analysis using logistic regression. What results do you find? How does this compare to alternative classification techniques such as $k$-nearest neighbour, LDA or QDA?