Logistic Regression Exercises – VHM-812/802

The data for the logistic regression exercises come from a retrospective analysis of the medical records from all diarrheic calves which were presented to the Atlantic Veterinary College between 1989 and 1993. The ultimate objective of the study was to develop a logistic model which would predict whether or not the calf was septic at the time of admission (septic calves have a much poorer prognosis than non-septic calves and are not usually worth treating from an economic point of view). The principal investigator in this study was Dr Jeanne Lofstedt (Lofstedt et al, 1999).

There are 254 observations (records) and 14 variables in the dataset (**calf.dta – summarized below**). The original dataset had far more variables (including a lot of laboratory data) but we will restrict ourselves to using a subset of the demographic data and the physical examination data collected.

```
. describe

Contains data from H:\VHM\VHM802\Data_Stata\calf.dta
  obs:           254
 vars:            14                          1 Jun 2004 04:44
 size:         7,874
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
case            int     %5.0f                 hospital case number
age             byte    %5.0f                 age at admision (in days)
breed           byte    %10.0f     breedlbl   breed (coded 1-9)
sex             byte    %6.0f      sexlbl     sex (0=female, 1=male)
attd            byte    %9.0f      attdlbl    attitude (0=bright, 1=depr.,
                                               2=comatose)
dehy            double  %5.1f                 % dehydration
eye             byte    %5.0f      ynlbl      uveitis or hypopyon (0=no, 1=yes)
jnts            byte    %5.0f                 swollen joints (# affected)
post            byte    %8.0f      postlbl    posture (0=standing, 1=sternal,
                                               2=lateral)
pulse           int     %5.0f                 pulse (beats per min.)
resp            int     %5.0f                 resp. rate (breaths per min)
temp            double  %6.1f                 rectal temperature (C)
umb             byte    %5.0f      ynlbl      swollen umbilicus (0=no, 1=yes)
sepsis          byte    %5.0f      ynlbl      sepsis (0=no, 1=yes)
-------------------------------------------------------------------------------
```

Exercise #1
Introduction to Logistic Regression

1. Familiarise yourself with the data. Look at descriptive statistics to check that all the values look reasonable. How many calves were ultimately diagnosed as septic?

2. Next, look into unconditional ("univariate") associations between the predictor variables and the outcome for sepsis (-sepsis-). Use simple statistics (e.g. t-tests or chi-square tests) to do this.

3. Identify all variables with a significant ($P \leq 0.1$) association with sepsis.

4. Build a simple logistic model using only posture (-post-) and swollen umbilicus (-umb-) as predictors. Remember, that -post- is not a dichotomous variable so you will have to convert it (implicitly or explicitly) to a series of indicator variables. Based on this model explain
   a) the relationship between a swollen umbilicus and the risk of being septic;
   b) the relationship between posture and the risk of being septic;
   c) how does the predicted probability of sepsis change as posture changes from standing to sternal to lateral?

5. What is the predicted probability of sepsis in calf #1294?

Exercise #2
Model-building for Logistic Regression

We want to build a logistic model for -sepsis- using, as a starting point, the following predictors which were found to have significant ($P \leq 0.1$) association with sepsis.

Categorical	Continuous
-attd-	-age-
-eye-	-dehy-
-jnts-	-resp-
-post-	-temp-
-umb-	

1. First, consider what type of causal structure might exist among the predictors and with the outcome.

2. One of the assumptions in a logistic regression model is that the relationship between the log-odds of the outcome and a continuous predictor variable is linear. Evaluate this assumption for each of the continuous predictor variables using the following two approaches:
   a) Categorise the predictor and assess the linearity of its categorical coefficients in a logistic regression model, and
   b) Categorise the predictor and plot the log-odds of disease against the predictor.

   Create quadratic or ordinal variables from continuous variables, which do not have an approximately linear relationship with the log-odds of sepsis.

3. Build a logistic model to predict sepsis using $P \leq 0.05$ as the criterion for statistical significance when considering terms to keep in the model. Approach this in two ways.
   a) First build one 'manually' by looking at various combinations of terms to include in the model. Use likelihood-ratio tests to evaluate the significance of groups of terms (e.g. categorical variables). Also, subjectively assess the impact of term addition/removal on the coefficients (and SEs) of other terms in the model.
   b) Once you have settled on what you feel is a reasonable model, try using an automated stepwise selection procedure to build the model. Do you get the same result?

4. Using a model including only -age- (both linear and quadratic terms), -post-, and -umb-,
   a) Investigate the association between swollen umbilicus and posture further by seeing if there is evidence of confounding or interaction between those two variables.
   b) Is -age- a confounder for posture or swollen umbilicus?

Exercise #3
Diagnostics for Logistic Regression

Evaluate the model specified in Exercise 2 (Question 4), i.e. including -age- (linear and quadratic terms), -post- and -umb-. Specifically:

1. Assess the fit of the model based on the Hosmer-Lemeshow goodness-of-fit test. Are goodness-of-fit tests based on deviance or the Pearson chi-square statistic applicable here?

2. Examine residuals, leverages, and delta-betas. Are there any individual calves that have an unduly large influence on the model?

3. How well does the model predict sepsis? Evaluate an ROC curve for the model.

4. What would be an appropriate value to use as a cutpoint or threshold for the model if we wanted to predict sepsis? What factors should you consider in making that choice? Use your chosen value to estimate the sensitivity and specificity of prediction from the model.