4.7 Practical 4

Datasets required: `lbw (webuse) and insect.dta`

Introduction

In the first part of this session we will use a publicly available dataset on birthweights of babies, taken from the book *A*pplied Logistic Regression by David W Hosmer and Stanley Lemeshow.

> "Data were collected as part of a larger study at Baystate Medical Center in Springfield, Massachusetts. This dataset contains information on 189 births to women seen in the obstetrics clinic. Fifty nine of these births were defined as low birth weight [defined as less than 2500 grams]."
>
> ***H**osmer & Lemeshow, Applied Logistic Regression (second edition, 2000), Pub. by Wiley-Interscience*

In the second part of the session we will re-visit the insect dataset we used in Practical 3 and assess the goodness of fit of the logistic regression model we used then.

Aims

The aims of this session are:

1. understand how to compare logistic regression models (in Stata)
2. understand how to assess goodness of fit of logistic regression models (in Stata)

Part A: Investigation into low birthweights in Massachusetts, USA

We will now investigate factors which may be related to the chance of having a low birthweight baby, using the lbw dataset. This dataset is available directly from the Stata website, using the "webuse" command:

```
webuse lbw, clear
```

Explore the dataset, checking for missing values. Be sure to identify the outcome variable, and check how it is coded. Also note that the mother's weight is given in lbs, whilst the baby's weight is given in grams.

1. Fit a logistic model for the low birthweight variable with the mother's weight at last menstrual period (in lbs) as the single covariate. Run the model using the glm, logistic and logit commands, and compare the output.

   Discuss: What are the main differences between the commands and the output they produce?

2. To visually assess the reasonableness of the assumption of a linear lwt effect (on the logit scale), we can compare the fitted values from this model to a lowess smoother plot. First, we calculate the fitted probabilities, then construct the plot using the twoway command:

```
predict fitted_prob , pr

twoway (lowess low lwt) (scatter fitted_prob lwt) (scatter low lwt)
```

Discuss: What do you conclude regarding the appropriateness of the linearity assumption here?

3 Using pen & paper and a calculator, calculate the probability that a women who weighs 120lbs (about 54.4kg) has a baby with birth weight less than 2500g.

4 Use Stata to construct a confidence interval around your estimate from question 3. You will need to use either the lincom command, or re-parameterise the model to centre weight at 120lbs.

Using lincom:

```
lincom 120*lwt + _cons , eform
```

Re-parameterise the model:

```
gen lwt120 = lwt - 120

logit low lwt120, or
```

5 Fit an appropriate model (using logistic or logit), and perform an appropriate test, to examine whether race has an independent association with the probability of having a low birthweight baby, after adjusting for the mother's weight.

Discuss: What do you conclude?

6 When the logistic model is fitted using logistic or logit, a `LR chi2( )` and `Prob > chi2` output is given in the top right hand corner. To understand what this is, perform a profile log-likelihood ratio test comparing the model you fitted in the previous part to the null model. The null model is the model which contains no covariates. What can you deduce about the meaning of the `LR chi2( )` and `Prob > chi2` values in the model you fitted in the previous part? How can you interpret the result in this case?

7 Assess the goodness of fit of the model you fitted in question 5 using the Hosmer-Lemeshow goodness of fit test with ten groups.

```
estat gof, group(10)
```

Repeat the test with five groups, and compare the results. Try using groups of other sizes.

Discuss: What do you conclude about the goodness of fit of this model, and about the Hosmer-Lemeshow test?

Part B: Investigation into effect of CS_2 at different doses on insect survival

8 Reload the insect data used in the last practical, and use the `glm` command to fit the logistic model which assumes (on the logit scale) a linear effect of dose.

Based on the reported deviance, test whether there is any evidence that this model does not fit the data well. You will need to be make use of the `chi2tail` command, or otherwise your Neave tables.

Discuss: Why can the deviance be used to test the fit of this model, but not that of any of the models in Part A?

9 As we did in the last practical, add the square of the dose to the model, and use a profile log-likelihood ratio test to compare the fit of this model to the simpler one which assumes a linear effect. How does the result compare with the corresponding Wald test?