FW663 -- Laboratory Exercise

Individual Covariates with Radio-tracking Data

Logistic regression is a useful procedure to examine the relationship between a dependent, categorical variable such as dead or alive from a radio-tracking experiment and an independent, continuous variable such as weight or a categorical variable such as gender. A good, readable reference on logistic regression procedures is Hosmer and Lemeshow (1989). Today's exercise is to determine if the size of a mule deer (*Odocoileus hemionus*) fawn at the start of the winter affects its over-winter survival rate. The data were collected in the Piceance Basin during the winter of 1982-83. Fawns were radio collared on 2 areas, a control and a treatment area. Body size was measured with 2 variables; mass (kg) and total body length (cm), a measurement taken from the tip of the nose to the base of the tail. We are going to use Program MARK to analyze these data, but want to emphasize that the analyses you will do are equivalent to logistic regression, and can be done in standard statistical packages. Hence details of how to conduct the analysis in SAS are provided as an appendix.

To analyze these data, you must specify the individual covariates to Program MARK. The input file with the raw data is provided in J:\CLASSES\FW663\EXERCISE.14\FAWNS.INP. The order of the 4 individual covariates is Area, Sex, Mass, and Length. The encounter history is coded as a single LD history, i.e., only one encounter occasion. Thus, only one survival rate is being estimated. Each record in the file represents a single animal. When creating the new MARK file, you should specify that 4 individual covariates are included in the input file, and then give them names to help you remember them when building models. These names are used in the Design Matrix to build models. Each design matrix will consist of a single row (only one survival rate), but with an intercept term, and 0 to 4 covariates. Individual covariates are specified in the design matrix via the name you specified for them. With 4 individual covariates, $2^4 = 16$ models are possible if no interactions are included.

Check for over-dispersion in the data by examining the Deviance/df value for the most general model. If necessary, correct the AICc value with $\hat{c}$.

Questions for Discussion

1. Which model seems to fit the data the best? Does it fit, i.e., what is the chi-square goodness-of-fit value? Is there evidence for excess variation compared to the binomial distribution?

2. How do you interpret extra-binomial variation in models which include continuous variables like length or weight?

3. What happens when you have missing values for individual covariates in the analysis?

4. What different attributes of fawn body size are measured by mass and length?

5. What graphical plot helps display the relationship between survival and body size? Survival, sex, and body size?

6. Would any information about the effect of length on survival be obtained if all the animals lived or died?

7. How would interactions between individual covariates be specified in Program MARK?

8. Could discrete individual covariates, such as Area and Sex, be treated as groups rather than individual covariates in Program MARK?

Program MARK Results

Model	AICc	Delta AICc	AICc Weight	#Par	Deviance
{S(Sex+Length)}	149.798	0.00	0.43599	3	143.582
{S(Sex+Length+Mass)}	151.717	1.92	0.16702	4	143.353
{S(Area+Sex+Length)}	151.875	2.08	0.15434	4	143.512
{S(Area+Sex+Length+Mass)}	153.824	4.03	0.05824	5	143.273
{S(Sex+Mass)}	154.642	4.84	0.03869	3	148.426
{S(Length)}	154.719	4.92	0.03723	2	150.612
{S(Sex)}	154.956	5.16	0.03307	2	150.849
{S(Length+Mass)}	156.578	6.78	0.01470	3	150.362
{S(Area+Length)}	156.776	6.98	0.01331	3	150.560
{S(Area+Sex+Mass)}	156.782	6.98	0.01327	4	148.418
{S(Area+Sex)}	157.064	7.27	0.01153	3	150.848
{S(.)}	157.603	7.81	0.00880	1	155.567
{S(Area+Length+Mass)}	158.678	8.88	0.00514	4	150.314
{S(Mass)}	159.283	9.48	0.00380	2	155.176
{S(Area)}	159.496	9.70	0.00342	2	155.389
{S(Area+Mass)}	161.220	11.42	0.00144	3	155.004

Literature Cited

Hosmer, D. W. and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, New York, N.Y. 307pp.

McCullagh, P. and J. A. Nelder. 1989. Generalized linear models. 2nd ed. Chapman and Hall, London. 511pp.

SAS Institute Inc. 1993. SAS® Technical Report P-243, SAS/STAT® Software: The GENMOD Procedure, Release 6.09. SAS Institute Inc., Cary, NC. 88pp.

Appendix – Logistic Regression in SAS

SAS/STAT for Windows contains 3 procedures to compute logistic regression estimates. PROC LOGISTIC is the easiest to use, especially if all independent variables are continuous, i.e.,

none are categorical. However, you must create your own dummy variables for categorical variables if you want to use them in LOGISTIC. LOGISTIC provides various step-wise procedures for model selection and does provide the AIC value in the output. Unfortunately, the AIC is not used for model selection in any of the step-wise procedures. An advantage of LOGISTIC is that the model and data are easily plotted to assess the fit. PROC CATMOD is a more general procedure than LOGISTIC, making it more difficult to understand and use. Both categorical and continuous variables work fine as independent variables, and interaction terms can be specified in the MODEL statement. However, the procedure has no model selection procedure. The -2log Likelihood value is printed, so you have to manually select the most appropriate model by computing AIC by hand. Also, getting graphical output is a real programming nightmare. PROC GENMOD (SAS Institute Inc. 1993) is a relatively new procedure that fits generalized linear models. In GENMOD, you specify what link function is to be used to link the dependent and independent variables, and what the error distribution should be assumed. The error distribution specifies the probability density function that will be used to construct the likelihood function. Refer to McCullagh and Nelder (1989) for a thorough treatment of generalized linear models. GENMOD does not compute AIC directly, but provides the log likelihood. Thus, you have to compute the AIC value by hand.

Today, we are going to use PROC GENMOD, as this procedure provides output most closely related to the type of analysis we have performed previously in this course.

Part of the goal of this course is to improve your computer literacy. The data are in the file J:\CLASSES\FW663\EXERCISE.14\FAWNS.DAT. You are to construct a SAS code to read the data and produce a logistic regression. Lets break the problem down into the parts.

First, the data must be read into a SAS data set. Hence you need a DATA statement. Next, the INFILE statement will specify the location of the data file, i.e., that the data are to be read from J:\CLASSES\FW663\EXERCISE.14\FAWNS.DAT. Then comes an INPUT statement to read the variables from this INFILE. The variables are the following:

Variable	Type	Columns
Area	Character	1 - 9
Radio Frequency	Character	11 - 17
Fawn Sex	Character	19 - 24
Fawn Weight	Numeric	26 - 29
Fawn Length	Numeric	31 - 35
Days Lived	Numeric	37 - 39

Examine the input file to see what it looks like. Remember that a $ following a variable name on the INPUT statement tells SAS that a variable is a character variable. Note that if a fawn lived through the winter, the number of days lived is coded as missing, i.e., as a single period (.). Thus, to create a variable that specifies the binary outcome of the fawn's survival experiment, you need to recode the number of days lived with something like the following:

```
IF DAYS = . THEN STATUS = 1;
ELSE STATUS = 0;
```

where STATUS is the fawn's status at the end of the winter. To complete the DATA step, you may want to provide some LABELs for each of your variables so that you can remember their contents in 2 weeks. Using SAS LABELs is good programming technique.

All of the above creates the SAS dataset. You may want to include a PROC PRINT statement to verify that the data were read correctly, and that variables you created contain the correct results.

PROC GENMOD is relatively easy to use. First, specify the class variables, i.e., variables in the model that are categorical. In our example, these variables are SEX and AREA. Next, specify the dependent variable on the left side of a MODEL statement, and the independent variables on the right side. In addition, you need to specify the link function, the error distribution, the TYPE3 option to obtain LR tests of the parameters for the null hypothesis that each equals zero, and LRCI to obtain profile likelihood confidence intervals on the parameter estimates. The following SAS code is needed.

```
PROC GENMOD;
    CLASS AREA SEX;
    MODEL STATUS=AREA SEX WEIGHT LENGTH /
                LINK=LOGIT
                DIST=BINOMIAL
                TYPE3
                LRCI    ;
    RUN;
```

If you suspect that extra-binomial variation is present (any evidence of this with these data?), you can add the DSCALE option to the MODEL statement to construct *F*-ratios instead of χ^2 tests.

You can read the manual to figure out additional options to plot your data and visually evaluate the fit of the model. The OBSTATS option must be added to the above MODEL statement, and the following statement added to generate an output file with the predicted values.

```
MAKE 'OBSTATS' OUT=OBSTATS;
```

Then, this dataset must be merged with the original dataset to obtain the values of the independent variables needed for plotting. Alternatively, I have supplied SAS code in the file

J:\CLASSES\FW663\EXERCISE.14\FAWNS.SAS

that will plot the data. You may want to resort to this file after you've fought with SAS for a while in generating an initial analysis.

After you have fit the above model, consider interactions between AREA and SEX, because on the Treatment area, the main source of mortality was predation, while on the Control area, starvation dominated. What about interactions between WEIGHT and LENGTH? What about all possible interactions? You have to compute the AIC value manually for each model, but this model selection criterion is useful for this analysis because a number of models should be considered, and not all the models are nested.

GENMOD defaults to tests of the significance of a variable as Wald statistics, constructed as $\chi^2_{(1\,df)} = (\hat{\theta}/\hat{se}(\hat{\theta}))^2$. These test statistics are not generally too different from the equivalent likelihood ratio test, although not quit as powerful (Hosmer and Lemeshow 1989). A major disadvantage of the Wald statistic is that it is undefined if the parameter estimate is ∞ or $-\infty$. This situation occurs when the observed value of the binomial trial, i.e., the dependent variable, is 0 or 1. That is, all the animals lived or they all died. In the fawn data example, the interaction between gender and length would have a parameter estimate of ∞ or $-\infty$ if all the males died or they all lived. To obtain LR tests of the significance of each effect, we have specified the TYPE3 option on the MODEL statement of GENMOD. To obtain profile likelihood confidence intervals, we have specified the LRCI option on the MODEL statement.

The deviance can be considered as a χ^2 statistic for goodness-of-fit of the model. However, McCullagh and Nelder (1989) caution that the deviance is not always distributed as a χ^2 statistic, and so interpretation may be impossible. This may be the reason that PROC GENMOD prints the deviance and its degrees of freedom, but does not report the probability level.