Biostats VHM 802 Course
Winter 2011/2023, Atlantic Veterinary College, PEI
Henrik Stryhn

Additional exercises for Session 4: Logistic regression

Exercise 4.1

Logistic regression for binary data

Hosmer & Lemeshow (1989) analyze data on coronary heart disease (CHD) of 100 patients in a study. The only information given about the patients is their age and the presence/absence of CHD. Analyze the data, with particular focus on the following points:

1) Estimate the parameters of a logistic regression model, compute approximate 95% confidence intervals and give their interpretations. Does CHD seem to be associated with age?

2) Discuss methods to assess the fit of a logistic regression model for binary data (ungrouped). Try at least one of these.

3) Construct an age group variable by defining (arbitrarily) age groups by the intervals: $20-29$, $30-34$, $35-39$, $40-44$, $45-49$, $50-54$, $55-59$, $60-69$. Compute the mean age in each group, and compare the fits of models with age groups as a factor and the mean age group as a regression variable. Does there seem to be lack of fit in the linear relation (on logistic scale)?

4) For the final model, discuss ways of presenting the results. Produce a graph of the estimated probability of CHD as a function of age.

Exercise 4.2

Logistic regression for binomial data

In a bioassay a total of 481 beetles were exposed to different doses of the gas CS_2 during five hours, and afterwards the survival of each beetle was recorded. In the table the dose is given as the logarithm (base 10) of the concentration of CS_2 in mg/l. (Data from Bliss, C. L. (1935): The calculation of the dosage-mortality curve, *Ann. Appl. Biol.* **22**, 134–67.)

Dose	Dead beetles	Total number of beetles
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Formulate a statistical model for the data, and analyze along similar lines as in the previous exercise. In particular, compare the fit of linear and quadratic models. Present the estimated relation between dose and mortality graphically.

Exercise 4.3

Factorial design in logistic regression

In a hospital in New York a sample of size 100 was taken among alcoholics, and another sample among non-alcoholics of size 500. For each patient is was recorded whether he/she suffered from cirrhosis of the liver. A similar investigation was carried out in Philadelphia with samples of 228 alcoholics and 3772 non-alcoholics. (Data from Jellinek, E. M. (1942): *Alcohol addiction and chronic alcoholism*, Yale University Press, New Haven.)

	New York		Philadelphia	
Patient group	sick	not sick	sick	not sick
alcoholic	35	65	45	183
non-alcoholic	25	475	105	3667

Use a logistic regression model with two factors to analyse the dependence of disease prevalence on site and patient status. Present your results using odds-ratios of factors of importance.

Exercise 4.4

Logistic regression with factors and covariates

Consider the following dataset on the dependence of low birthweight on a number of factors related to the birth. These factors were recorded for a total of 189 women giving birth. Formulate a logistic regression model for the risk of low birthweight, and carry out a careful statistical analysis of these data. Points to be considered include model check, choice of quantitative or categorical modelling for the variables, tests for the impact of each variable, possible interactions between variables, and interpretation of the final model.

List of variables in the datafile:

No.	Name	Levels
1	Patient id	numerical
2	Low birth weight	0 (bw $\geq 2500g$), 1 (bw $< 2500g$)
3	Age of mother	years
4	Weight of mother	pounds (at last menstrual period)
5	Race	1 (white), 2 (black), 3 (other)
6	Smoking during pregnancy	0 (no), 1 (yes)
7	Premature labour	number of times
8	Hypertension	0 (no), 1 (yes)
9	Uterine irritability	0 (no), 1 (yes)
10	No. of physician visits	number (during first trimester)