SOS3003
Applied data analysis for social science
Lecture note 09a-2009

Erling Berge
Department of sociology and political science
NTNU

Sample size in logistic regression

Literature cited:

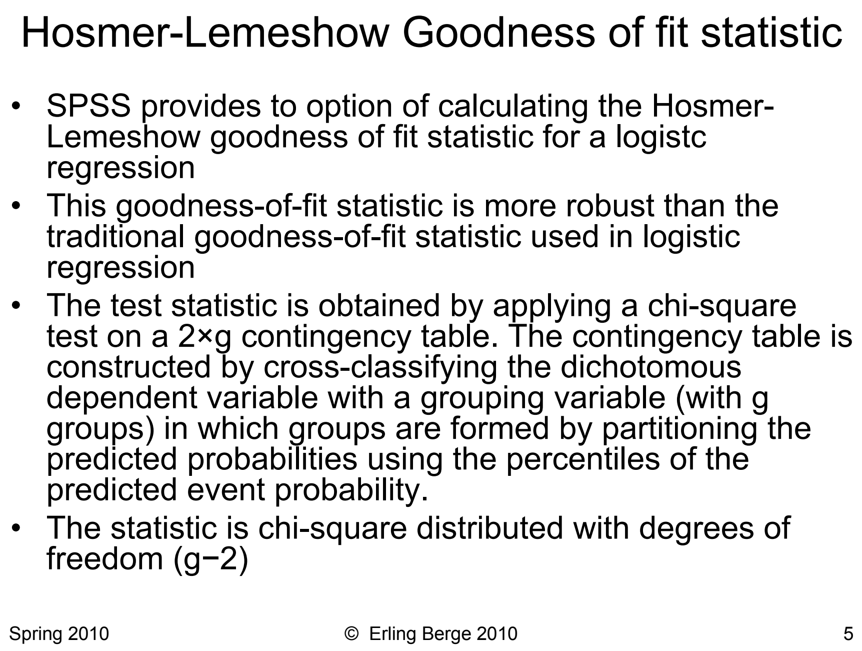
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. London: Sage.
- Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49 (12):1373-1379.

Large sample properties

- The good properties of ML estimates of binary logistic regression models are large sample properties that obtain as sample size goes towards infinity. Hence
- A sample needs to be "large enough"
- What "large enough" means is not clear
- What happens when you have too small a sample is largely unknown
- Long (1997) puts 100 cases as an absolute lower bound

Calculation of lower bounds

- A lower bound of 100 must be adjust according to number of variables in the model and the distribution of cases on the dependent variable.
- Peduzzi et al. (1996) suggest:
- Let p be the smallest of the proportions of negative or positive cases in the population and k the number of covariates (the number of independent variables), then the minimum number of cases to include is:
- $N = 10\ k\ /\ p$
- If the resulting number is less than 100 you should increase it to 100



Hosmer-Lemeshow Goodness of fit statistic

- SPSS provides to option of calculating the Hosmer-Lemeshow goodness of fit statistic for a logistc regression
- This goodness-of-fit statistic is more robust than the traditional goodness-of-fit statistic used in logistic regression
- The test statistic is obtained by applying a chi-square test on a 2×g contingency table. The contingency table is constructed by cross-classifying the dichotomous dependent variable with a grouping variable (with g groups) in which groups are formed by partitioning the predicted probabilities using the percentiles of the predicted event probability.
- The statistic is chi-square distributed with degrees of freedom (g−2)