Different types of "goodness-of-fit tests" for logistic regression

Hui Lin

January 10, 2010

The following are what I found so far.

1. $Pearson\ \chi^2$ and $Deviance$

   (a) Suppose for sake of discussion there are J covariate patterns, for a particular covariate pattern j, m_j is the number of subjects, O_j stands for the observed value. $E_j = m_j\hat{\pi}_j$ stands for the fitted value. $Pearson\ \chi^2$ is defined as the following χ^2.

   $$r(O_j,\ \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

   $$\chi^2 = \sum_{j=1}^{J} r(O_j,\ \hat{\pi}_j)^2$$

   Note: $r(O_j,\ \hat{\pi}_j)$ can be considered as the standardization of y_i so it approximates to $Normal\ distribution$. (I think this way ^_^, just help me understand) When χ^2 is significant, we need to examine the model.

   (b) Another way to compare observed value and fitted value is accoreding to **log likelihood function (**$\mathcal{L}_r$—— stnads for reduced model**)** which describes the probability of the sample we observed under various parameters. [For more about the **"likelihood"**, can see **Chapter 6** of **"Statistical Inference"** by **Casella & Berger**] However, only $\mathcal{L}_r$ is not sufficient to study "goodness of fit" for it depends on the number of parameters included. Here, $\frac{\mathcal{L}_r}{\mathcal{L}_f}$ (**likelihood ratio**) is then used to show the sufficience of the model and $\mathcal{L}_f$ here stands for full model (saturated model). The statisitc **D (deviance)** is defined as follows which is approximate to χ^2 staistics.

   $$D = -2ln(\frac{\mathcal{L}_r}{\mathcal{L}_f}) = -2(ln\mathcal{L}_r - ln\mathcal{L}_f)$$

   (c) Remark on $Pearson\ \chi^2$ and $Deviance$:

      i. They are both approximate to χ^2

      ii. When "maximun likehood" is used to estimate logistic regression model, $Deviance$ is better than $Pearson\ \chi^2$

      iii. Sometimes there is big difference between the two statisitcs for the same sample, it may because they are not approximate to χ^2 distribution enough. (sample size is too small......)

      - *[More detail see: Clogg. C.C. & S.R. Eliason. 1988. Some common problems in log-linear analysis. PP.226~257 in J.S. Long (ed.) Common Problems/ Proper Solutions: Avoiding Error in Quantitative Research. Newbury Park, CA: Sage]*
      - *About the request on sample size see: Stokes, M.E., C. S. Davis, & G. G. Koch, 1995. Categorical Data Analysis Using the SAS System, Cary, NC: SAS Institute, Inc. p.169.*

   (d) SAS CODE:
   If the logistic model is : $ln\frac{\pi}{1-\pi} = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

```
#########SAS CODE EXAMPLE#########

proc logistic descending;
model Y= X1 X2/ scale = none aggregate;
run;

~n_n~~ EXAMPLE STOPS HERE ~~n_n~
```

In above SAS program, "aggregate" is equal to "aggregate=(X1 X2)" , ie: list all the independent variables. "scale" can be used to remedy over dispersion. Here "scale = none" means no adjust to over-dispersion. It gives too goodness-of-fit statisitcs: $Pearson\ \chi^2$ and $Deviance$.

2. Hosmer-Lemeshow

   (a) Why Hosmer-Lemeshow ?
   When the number of independent variables is large, the number of covariate pattern will be large too. m_j (mentioned in 1(a)) may be small. In this case, $Pearson\ \chi^2$ and $Deviance$ are not appropriate. Hosmer and Lemshow proposed another way to estimate goodness of fit.

   (b) What is Hosmer-Lemeshow?