Goodness of Fit Tests for Categorical Data: Comparing Stata, R and SAS

Rino Bellocco[1,2], Sc.D.
Sara Algeri[1], MS

[1]University of Milano-Bicocca, Milan, Italy & [2]Karolinska Institutet, Stockholm, Sweden

San Servolo, Venice, Italy

November 17-18, 2011

Content

Outline

Steps

- Definition of exposure, confounders, interaction
- Model Building
- Likelihood based theory: estimation, confidence intervals and testing
- Goodness of Fit

Outline

Unit of Analysis

In Generalized Linear Model when the covariates involved are/can be restricted to categorical data, the units of analysis could be:

- subjects

  or

- groups of subjects

Unit of Analysis: The Subjects

Let us consider the case of a Logistic Regression Model for a binary outcome Y:

$$ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (1)$$

- The units of analysis are subjects
- The data layout is based on one record for each subject (**individual format**)
- The goal is to predict $\pi(\mathbf{x})$ which is the probability of success $P(Y=1)$ given the set of covariates $\mathbf{x} = (x_1, \ldots, x_p)$
- Thus, the log-likelihood function will be:

$$\sum_{i=1}^{n}\{y_i ln[\pi(\mathbf{x}_i)] + (1-y_i) ln[1-\pi(\mathbf{x}_i)]\} \quad (2)$$

n = The total number of observations.

Unit of Analysis: Group of Subjects

- The analytical units are groups of subjects with the same covariate patterns
- The quantity $\pi(\mathbf{x})$ is now referred to the proportion of successes or each group to be estimated
- The data layout in this case will be one record for each covariate pattern (**events-trials format**)
- The log-likelihood function (2) can be written as:

$$\sum_{j=1}^{K}\left\{ s_j ln[\pi(\mathbf{x}_j)] + (m_j - s_j) ln[1 - \pi(\mathbf{x}_j)]\right\} \tag{3}$$

K = total number of possible (observed) covariate patterns
s_j = number of successes for the j^{th} covariate pattern
m_j = number of total individuals for the j^{th} covariate pattern

In spite of different structures, estimation will be the same

Outline

The Deviance Test Statistics

One methods for goodness-of-fit assessment is to use the deviance statistics (D^2)

$$D^2 = 2\{ln[L_s(\widehat{\beta})] - ln[L_m(\widehat{\beta})]\} \quad (4)$$

- $ln[L_m(\widehat{\beta})]$ = maximized log-likelihood of the fitted model
- $ln[L_s(\widehat{\beta})]$ = maximized log-likelihood of the saturated model
- This quantity compares the values predicted by the fitted model and those predicted by "the most complete model we could fit".
- Evidence for model lack-of-fit occurs when the value of D^2 is large

The Deviance Test Statistics

Asymptotic Distribution

- Under specific regularity conditions D^2 converges asymptotically to a χ^2 distribution with h degrees of freedom
- h is the difference between the number parameters in the saturated model and the number of parameters in the model being considered:

$$D^2 \sim \chi^2_{(h)} \tag{5}$$

- Thus, we can test the null hypothesis:

$$H_0 : \beta_h = 0$$

- So H_0 is rejected when:

$$D^2 \geq \chi^2_{1-\alpha}$$

α= fixed level of significance.

The Deviance Test Statistics

Asymptotic Distribution-cont

- If H_0 cannot be rejected we can safely conclude that the fitting of the model of interest is substantially similar to that of the most completed model that can be built
- The deviance test is to all intents and purposes a Likelihood Ratio Test which compares two nested models in terms of log-likelihood. In fact, all the possible models we can built are nested into the saturated model

Saturated Model

- The saturated model represents the largest model we can fit and leads to perfect prediction of the outcome of interest
- This definition does not lead to an unique specification but we can identify three different approaches for its specification
  - Casewise approach
  - Contingency table approach
  - Collapsing approach

Saturated Model

Casewise Approach

- When the unit of analysis is the subject the saturated model has as many parameters as the number of observations (n=number of subjects)
- "Perfect Fit" of the data
- the log-likelihood (2), is always equal to zero
- Consequently, the deviance statistics ((4)) results to be:

$$
\begin{aligned}
G^2 &= -2\left[lnL_m(\widehat{\beta})\right] \\
&= -2\sum_{i=1}^{n}\left\{\widehat{\pi} ln\left[\frac{\widehat{\pi}(\mathbf{x}_i)}{1-\widehat{\pi}(\mathbf{x}_i)}\right] + ln[1-\widehat{\pi}(\mathbf{x}_i)]\right\}
\end{aligned}
$$

Saturated Model

Casewise Approach - cont

- This approach is used in with continuous covariates, here the number of covariate patterns is quite similar to the number of subjects (n=K).
- D^2 cannot be approximated to a χ^2 distribution.

Thus, it might be useful to use one of the following approaches.

Saturated Model

Contingency Table and Collapsing approach

- The units of analysis are the group of subjects defined by the covariate pattern
- The saturated model corresponds to a with K parameters, where K is the number of the possible covariate patterns.
- In these situations ($n \neq K$) and the log-likelihood of the saturated model is not equal to zero, and D^2 will be:

$$
\begin{aligned}
D^2 &= 2\left\{ln\left[L_s(\widehat{\beta})\right] - ln\left[L_m(\widehat{\beta})\right]\right\} \\
&= 2\left(\sum_{j=1}^{K}\left\{s_j ln\left[\frac{\widehat{\pi}_s(\mathbf{x}_j)}{\widehat{\pi}_m(\mathbf{x}_j)}\right] + (m_j - s_j) ln\left[\frac{1-\widehat{\pi}_s(\mathbf{x}_j)}{1-\widehat{\pi}_m(\mathbf{x}_j)}\right]\right\}\right)
\end{aligned}
$$

Saturated Model

Contingency table and Collapsing approach - Cont

$\widehat{\pi}_s(\mathbf{x}_j)$=proportion of successes for the j^{th} covariate pattern predicted by the saturated model

$\widehat{\pi}_m(\mathbf{x}_j)$=proportion of successes for the j^{th} covariate pattern predicted by the fitted model.

Saturated Model

Contingency Table and Collapsing approach - What is the difference

- If the covariate patterns are based on all covariates available in the data set, we are following the *contingency table approach*;
- If the covariate patterns are based only on the variables in the model of interest we are fitting the *collapsing approach*

As shown in the following application, different covariate pattern specifications lead to different results both in terms of likelihood and deviance.

Outline

Setup

- We implement the deviance test considering the three different approaches presented above using three of the most common software Stata (version 12.0), R (version 2.13.1) and SAS (version 9.2)
- In all the softwares considered the default method considers as saturated model the model which contains as many parameters as the number of records available in the data set. Thus, the data structure (individual format or events-trials format) cannot be neglected.

Data

The data used are based the famous Titanic disaster occurred on April 15, 1912.

- 2201 subjects;
- outcome: survival (1=survivor, 0=deceased, or number of survivors);
- covariates:
  - sex (male or female);
  - economic status (first class passenger, second class passenger, third class passenger or crew);
  - age (adult or child).

These covariates define 16 different covariate patterns and 14 observed.

Casewise approach

- In all the softwares considered the default method adopted assumes as saturated model the one which contains as many covariates as the number of records available in the data set
- Applying the most common procedures for logistic regression on the *individ* data set (analytical units=subjects), it is easy to obtain the deviance test considering the casewise definition of saturated model

Likelihood Ratio Test implementation-casewise approach with Stata

```
. xi:glm survival i.sex i.status, family(binomial) link(logit)

Generalized linear models                         No. of obs      =      2201
Optimization     : ML                             Residual df     =      2196
                                                  Scale parameter =         1
Deviance         =   2228.91282                   (1/df) Deviance =  1.014988
Pearson          =  2228.798854                   (1/df) Pearson  =  1.014936

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  1.017225
Log likelihood   =  -1114.45641                   BIC             = -14672.97

------------------------------------------------------------------------------
             |                 OIM
    survival |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
  _Istatus_2 |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
  _Istatus_3 |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
  _Istatus_4 |  -.7774228   .1423145    -5.46   0.000    -1.056354   -.4984916
       _cons |   1.187396   .1574664     7.54   0.000      .878767    1.496024
------------------------------------------------------------------------------
```

Likelihood Ratio Test implementation-casewise approach with Stata - cont

```
. scalar dev=e(deviance)

. scalar df=e(df)

. di "GOF casewise "" D^2="dev " df="df " \\\
  p-value= " chiprob(df, dev)

GOF casewise  D^2=2228.9128 df=2196 p-value= .30705384
```

Likelihood Ratio Test implementation-casewise approach with R

```
> Model<-glm(survival~sex+status,data=individ,
+ family=binomial(link=logit))
> Model

Call:  glm(formula = survival ~ sex + status, family = binomial(link = logit),
    data = individ)

Coefficients:
 (Intercept)       sexMale   statusFirst  statusSecond   statusThird
     1.18740      -2.42133       0.88081      -0.07178      -0.77742

Degrees of Freedom: 2200 Total (i.e. Null);  2196 Residual
Null Deviance:      2769
Residual Deviance: 2229         AIC: 2239
```

Likelihood Ratio Test implementation-casewise approach with R - cont

```
> dev<-deviance(Model)
> df<-df.residual(Model)
> p_value<-1-pchisq(dev,df)
> print(matrix(c("GOF casewise approach"," ","G^2",round(dev,4),"df",df,
+"p-value", round(p_value,4)),nrow=4,ncol=2,byrow=T))
     [,1]                    [,2]
[1,] "GOF casewise approach" " "
[2,] "G^2"                   "2228.9128"
[3,] "df"                    "2196"
[4,] "p-value"               "0.3071"
```

Likelihood Ratio Test implementation-casewise approach with SAS

```
proc logistic data=individ;
class status sex;
model survival/n= status sex
          /scale=none;

                    (omitting the long estimation section)

Deviance and Pearson Goodness-of-Fit Statistics

Criterion          Value       DF     Value/DF     Pr > ChiSq

Deviance      2228.9128     2196        1.0150             0.3071
Pearson       2228.6553     2196        1.0149             0.3084

Number of events/trials observations: 2201

run;
```

Likelihood Ratio Test implementation - Contingency approach

- The contingency table approach can be obtained applying the procedures already introduced for the casewise approach on the *grouped* data set (with one record for each covariate pattern observed)
- Some attention must be spent on the outcome specification that now is expressed in terms of number of survivors.

Likelihood Ratio Test implementation - Contingency approach with Stata

```
. xi:glm survival i.sex i.status if n>0, family(binomial n) link(logit)

Generalized linear models                         No. of obs      =         14
Optimization     : ML                             Residual df     =          9
                                                  Scale parameter =          1
Deviance         =  131.4183066                   (1/df) Deviance =   14.60203
Pearson          =  127.8463371                   (1/df) Pearson  =   14.20515

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]
                                                  AIC             =   13.43138
Log likelihood   = -89.01967223                   BIC             =   107.6668
------------------------------------------------------------------------------
             |                 OIM
    survival |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
  _Istatus_2 |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
  _Istatus_3 |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
  _Istatus_4 |  -.7774228   .1423145    -5.46   0.000    -1.056354   -.4984916
       _cons |   1.187396   .1574664     7.54   0.000      .878767    1.496024
------------------------------------------------------------------------------

. scalar dev=e(deviance)

. scalar df=e(df)

. di "GOF contingency "" G^2="dev " df="df " p-value= " chiprob(df, dev)
GOF contingency  G^2=131.41831 df=9 p-value= 6.058e-24
```

Likelihood Ratio Test implementation - Contingency approach with R

```
> fail<-grouped$n-grouped$survival
> Model<-glm(cbind(survival,fail)~sex+status,data=grouped,
+ family=binomial(link=logit))
> Model

Call:  glm(formula = cbind(survival, fail) ~ sex + status,
    family = binomial(link = logit), data = grouped)

Coefficients:
 (Intercept)       sexMale   statusFirst  statusSecond   statusThird
     1.18740      -2.42133       0.88081      -0.07178      -0.77742

Degrees of Freedom: 13 Total (i.e. Null);  9 Residual
Null Deviance:      672
Residual Deviance: 131.4        AIC: 188
```

Likelihood Ratio Test implementation - Contingency approach with R

```
> dev<-deviance(Model)
> df<-df.residual(Model)
> p_value<-1-pchisq(dev,df)
> print("GOF contingency table approach")
[1] "GOF contingency table approach"
> print(matrix(c("GOF contingency approach"," ","G^2",round(dev,4),"df",df,
+"p-value", round(p_value,4)),nrow=4,ncol=2,byrow=T))
     [,1]                       [,2]
[1,] "GOF contingency approach" " "
[2,] "G^2"                      "131.4183"
[3,] "df"                       "9"
[4,] "p-value"                  "0"
```

Likelihood Ratio Test implementation - Contingency approach with SAS

```
proc logistic data=grouped;
class status sex;
model survival/n= status sex
          /scale=none;

run;
          (omitting the long estimation section)
Deviance and Pearson Goodness-of-Fit Statistics

Criterion         Value       DF     Value/DF     Pr > ChiSq

Deviance       131.4183        9       14.6020           <.0001
Pearson        127.8383        9       14.2043           <.0001

 Number of events/trials observations: 14
```

Likelihood Ratio Test implementation - Collapsing approach

- Now the covariate patterns are based only on the covariates involved in the fitted model
- So we cannot use the default options from the previous procedures
- Both in Stata and R, new programs or functions (not available in the standard version of these softwares) are available whereas in SAS, we can use the option `aggregate`

Likelihood Ratio Test implementation - Collapsing approach with Stata

```
. xi:logit survival i.sex i.status
i.sex              _Isex_1-2           (_Isex_1 for sex==Female omitted)
i.status           _Istatus_1-4        (_Istatus_1 for status==Crew omitted)
h
Logistic regression                               Number of obs   =       2201
                                                  LR chi2(4)      =     540.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -1114.4564                       Pseudo R2       =     0.1952

------------------------------------------------------------------------------
    survival |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
  _Istatus_2 |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
  _Istatus_3 |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
  _Istatus_4 |  -.7774228   .1423145    -5.46   0.000    -1.056354   -.4984916
       _cons |   1.187396   .1574664     7.54   0.000      .878767    1.496024
------------------------------------------------------------------------------
```

Likelihood Ratio Test implementation - Collapsing approach with Stata

```
program define ldev
  version 12.0
  tempvar n d d2 j
  predict `d', de
  predict `n', n
  generate `d2' = (`d')^2
  sort `n'
  quietly by `n': generate `j' = _n
  quietly summarize `d2' if `j' == 1
  display
  display in green "Logistic model deviance goodness-of-fit test"
  display
  display in green "         number of observations = " in yellow %7.0f = e(N)
  display in green "   number of covariate patterns = " in yellow %7.0f = r(N)
  display in green "       deviance goodness-of-fit = " in yellow %10.2f = r(sum)
  display in green "                   degrees of freedom = " /*
       */ in yellow %7.0f = (r(N) - e(df_m) - 1)
  display in green "                         Prob > chi2 = " /*
       */ in yellow %12.4f = chiprob((r(N) - e(df_m) - 1),r(sum))
end
```

Likelihood Ratio Test implementation - Collapsing approach with Stata

```
. ldev

Logistic model deviance goodness-of-fit test

       number of observations =    2201
  number of covariate patterns =       8
      deviance goodness-of-fit =   65.18
            degrees of freedom =       3
                  Prob > chi2 =  0.0000
```

Likelihood Ratio Test implementation - Collapsing approach with R

```
> collapsing_approach<-function(Model){
+ y<-Model$y
+ x<-Model$model[,-1]
+ nx<-dim(x)[2]
+ toString(nx)
+ name<-(names(Model$model[,-1]))
+ fmla <- as.formula(paste("y~",paste("(",paste(name,collapse= "+"),")^", nx)))
+ m<-glm(fmla,data=Model$data,
+ family=binomial(link=logit))
+ ls<-logLik(m)
+ devS<--2*ls
+ dfS<-attr(ls,"df")
+ G2<-Model$deviance-devS
+ df<-dfS-attr(logLik(Model),"df")
+ p_value<-1-pchisq(G2,df)
+ print(matrix(c("GOF collapsing approach"," ","G^2",round(dev,4),"df",df,
+"p-value", round(p_value,4)),nrow=4,ncol=2,byrow=T))}
```

Likelihood Ratio Test implementation - Collapsing approach with R

```
> Model<-glm(survival~sex+status,data=individ,
+ family=binomial(link=logit))
> Model

Call:  glm(formula = survival ~ sex + status, family = binomial(link = logit),
    data = individ)

Coefficients:
 (Intercept)       sexMale   statusFirst  statusSecond   statusThird
     1.18740      -2.42133       0.88081      -0.07178      -0.77742

Degrees of Freedom: 2200 Total (i.e. Null);  2196 Residual
Null Deviance:      2769
Residual Deviance: 2229         AIC: 2239
> collapsing_approach(Model)
     [,1]                       [,2]
[1,] "GOF collapsing approach" " "
[2,] "G^2"                     "65.1798"
[3,] "df"                      "3"
[4,] "p-value"                 "0"
```

Likelihood Ratio Test implementation - Collapsing approach with SAS

```
proc logistic data=individ;
 class status sex;
model survival/n= status sex
          /scale=none aggregate;
run;
          (omitting the long estimation section)
Deviance and Pearson Goodness-of-Fit Statistics

Criterion         Value       DF     Value/DF     Pr > ChiSq

Deviance        65.1798        3       21.7266           <.0001
Pearson         60.8752        3       20.2917           <.0001

Number of unique profiles: 8
```

Outline

- Agresti, A. (2007), An Introduction to Categorical Data Analysis, John Wiley & Sons, Inc.
- Dickman, P. (1998), 'Evaluating the goodness-of-fit of logistic regression models with examples in SAS PROC LOGISTIC', Technical report, Karolinska Institute.
- Hosmer, D. W.; Hosmer, T.; Le Cessie, S. & Lemeshow, S. (1997), 'A comparison of Goodness-of-Fit Tests for the Logistic Regression Model'
- Hosmer, D. W.; Taber, S. & Lemeshow, S. (1991), 'The Importance of Assessing the Fit of Logistic Regression Models: A Case Study'
- Kleinbaum, D. G. & Klein, M. (2010), Logistic Regression: A Self-Learning Text, Springer
- Kuss, O. (2001), 'Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data', Technical report, University of Halle-Wittenberg
- Long, J. S. & Freese, J. (2000), 'Scalar Measure of Fit for Regression Models', Indiana University and University of Wisconsin-Madison
- Simonoff, J. S. (1998), 'Logistic Regression, Categorical Predictors, and Goodness-of-Fit: It Depends on Who You Ask'