The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2013)
13, Number 2, pp. 356–365

Goodness-of-fit tests for categorical data

Rino Bellocco
University of Milano–Bicocca
Milan, Italy
rino.bellocco@ki.se
and
Karolinska Institutet
Stockholm, Sweden
rino.bellocco@ki.se

Sara Algeri
Texas A&M University
College Station, TX
s.algeri@stat.tamu.edu

Abstract. A significant aspect of data modeling with categorical predictors is the definition of a saturated model. In fact, there are different ways of specifying it—the casewise, the contingency table, and the collapsing approaches—and they strictly depend on the unit of analysis considered.

The analytical units of reference could be the subjects or, alternatively, groups of subjects that have the same covariate pattern. In the first case, the goal is to predict the probability of success (failure) for each individual; in the second case, the goal is to predict the proportion of successes (failures) in each group. The analytical unit adopted does not affect the estimation process; however, it does affect the definition of a saturated model. Consequently, measures and tests of goodness of fit can lead to different results and interpretations. Thus one must carefully consider which approach to choose.

In this article, we focus on the deviance test for logistic regression models. However, the results and the conclusions are easily applicable to other linear models involving categorical regressors.

We show how Stata 12.1 performs when implementing goodness of fit. In this situation, it is important to clarify which one of the three approaches is implemented as default. Furthermore, a prominent role is played by the shape of the dataset considered (individual format or events–trials format) in accordance with the analytical unit choice. In fact, the same procedure applied to different data structures leads to different approaches to a saturated model. Thus one must attend to practical and theoretical statistical issues to avoid inappropriate analyses.

Keywords: st0299, saturated models, categorical data, deviance, goodness-of-fit tests

1 Deviance test for goodness of fit

It is common to find applications of logistic regression models in categorical data analysis. In particular, considering the simplest case of a binary outcome Y, the logistic regression model for the probability of success π $\{P(Y=1)\}$ is defined as

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

where $\pi(\mathbf{x})$ is the probability of success given the set of covariates $\mathbf{x} = (x_1, \ldots, x_p)$. Considering $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, the vector containing the unknown parameters in (1), under the assumption of independent outcomes, we can obtain the corresponding maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ by maximizing the following log-likelihood function:

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

where n is the total number of observations and y_i is the observed outcome for the ith subject. This situation is based on having subjects as analytical units; thus the data layout presents one record for each individual considered in the dataset (individual format).

When one works with categorical data, it is possible (and frequently more useful) to consider some groups of subjects as units of analysis. These groups correspond to the covariate patterns (that is, the specific combinations of predictor values $\mathbf{x}_j$). Thus it is possible to reshape the dataset so that each record will correspond to a particular covariate pattern or profile (events–trials format), including the total number of individuals and total number of successes (deaths, recoveries, etc.). In this case, the goal is to predict the proportion of successes for each group. The quantity π will be the same for any individual in the same group (Kleinbaum and Klein 2010), and we adopt the binomial distribution as reference to model this probability. So if we rewrite the log-likelihood function (2) in terms of covariate patterns, we obtain

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

where K is the total number of possible (observed) covariate patterns, s_j represents the number of successes, m_j is the number of total individuals, and $\pi(\mathbf{x}_j)$ is the proportion of successes corresponding to the jth covariate pattern. Therefore, in spite of different structures, because the information contained is exactly the same, the parameter estimates from (2) and (3) are exactly the same.

Having defined the log-likelihood function, we can perform the assessment of goodness of fit with different methods. In this article, we focus our attention on the likelihood-ratio test (LRT) based on the deviance statistics. The deviance statistic compares, in terms of likelihood, the model being fit with the saturated model. The deviance statistic for a generalized linear model (see Agresti [2007]) is defined as

$$G^2 = 2\left[\ln\left\{L_s\left(\widehat{\boldsymbol{\beta}}\right)\right\} - \ln\left\{L_m\left(\widehat{\boldsymbol{\beta}}\right)\right\}\right] \tag{4}$$

where $\ln\{L_m(\widehat{\boldsymbol{\beta}})\}$ is the maximized log likelihood of the model of interest and $\ln\{L_s(\widehat{\boldsymbol{\beta}})\}$ is the maximized log likelihood of the saturated model. This quantity can also be interpreted as a comparison between the values predicted by the fitted model and those predicted by the most complete model. Evidence for model lack-of-fit occurs when the value of G^2 is large (see Hosmer et al. [1997]).

It is generally accepted that this statistic, under specific conditions of regularity, converges asymptotically to a χ^2 distribution with h degrees of freedom, where h is the difference between the parameters in the saturated model and the parameters in the model being fit:

$$G^2 \sim {\chi^2}_{(h)}$$

Therefore, we use the deviance test to assess the following hypothesis

$$H_0 : \boldsymbol{\beta}_h = 0$$

where $\boldsymbol{\beta}_h$ is the vector containing the additional parameters of the saturated model compared with the model considered. So H_0 is rejected when

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

where α is the level of significance. 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 (see section 2). We must clarify that the LRT can always be used to compare two nested models in terms of differences of deviances.

2 Definition of saturated model

A particular issue that is not carefully considered in categorical data analysis is the definition of a saturated model. In fact, according to Simonoff (1998), three different specifications are available and depend on the unit of analysis. In general, we can think of the saturated model as the model that leads to the perfect prediction of the outcome of interest and represents the largest model we can fit. Thus it is used as a reference for the assessment of the fitting of any other model of interest.

The traditional approach is the one that considers the saturated model as the model that gives a perfect fit of the data. So it assumes the subjects to be the analytical unit and is identified with the casewise approach. This model contains the intercept and $n-1$ covariates (where n is the total number of available observations as specified above). Consequently, the maximum likelihood function in (2) is always equal to 0 (see Kleinbaum and Klein [2010]). So the deviance statistic shown in (4) result in

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

This approach is generally followed in the case of continuous covariates whose values cannot be grouped into categorical values. In fact, in this situation, each covariate pattern will most likely correspond to one subject ($n = K$), and obviously, the most reasonable analytical unit is the subject. However, in this case, the G^2 goodness-of-fit statistics cannot be approximated to a χ^2 distribution (see Kuss [2002] and Kleinbaum and Klein

[2010]). Thus even if statistical packages provide a p-value from a χ^2 distribution, we recommend using the second or third approach—the contingency table or collapsing approach—if all the regressors are or can be reduced to categorical variables.

These two approaches are based on groups of subjects as analytical units (and the data layout will be in events–trials format). These groups correspond to the covariate patterns, and individuals sharing the same covariate pattern are members of the same group. The saturated model is identified as the one with the intercept and $K-1$ regressors, where K is the number of all the possible covariate patterns; in other words, the model includes all possible main effects and all possible interaction effects (two-way, three-way, etc., until the maximum possible interaction order). The difference of the two situations is based on the covariate pattern specification. In the first one, the covariate patterns are built by considering all the covariates available in the dataset. In the second one, the covariate patterns are based only on the variables specified in the model of interest.

Clearly, if the model of interest includes all the variables available in the dataset, the two approaches coincide. Under these situations (if $n \neq K$), the log likelihood of the saturated model is not equal to 0, and the G^2 statistic is

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

where $\widehat{\pi}_s(\mathbf{x}_j)$ is the proportion of successes for the jth covariate pattern predicted by the saturated model and $\widehat{\pi}_m(\mathbf{x}_j)$ is the one predicted by the fitted model.

The collapsing approach has a main drawback: it uses different saturated models corresponding to different models of interest, complicating the comparison of their results in terms of goodness of fit. On the other hand, the contingency approach may require the listing of a high number of covariates. In this case, we could have many covariate patterns with a small number of subjects, making the use of the χ^2 approximation in the LRT for goodness of fit difficult once again. A possible remedy could be that the hypothetical saturated model in the contingency approach should be based on variables identified through the corresponding directed acyclic graphs. In a causal inference framework, we could then use only the variables suggested by the d-separation algorithm applied to the directed acyclic graph, which imposes the researcher to specify the interrelationship among the variables (Greenland, Pearl, and Robins 1999).

3 Implementation of the LRT

In this section, we implement the LRT, and we show the results by considering the three different saturated model specifications, using Stata 12.1. The data used in the analyses refer to the Titanic disaster on 15 April 1912. Information on 2,201 persons is available on three covariates: sex (male or female), economic status (first-class passenger, secondclass passenger, third-class passenger, or crew), and age (adult or child), which defines 16 different covariate patterns (among which 14 were observed). The outcome of interest is either passenger's survival (`1` = survivor, `0` = deceased) or the number of survivors and total number of passengers.

As anticipated above, two possible ways to represent these data can be considered with respect to the goal and the unit of analysis (Kleinbaum and Klein 2010):

- *Individual-record format:* One record for each subject considered with the information on survival (or death) contained in a binary variable (`individ.txt`).
- *Events–trials format:* One record for each covariate pattern with frequencies on survivors and total number of passengers (`grouped.txt`) available as follows:

	age	sex	status	survival	n
1.	Adult	Male	First	57	175
2.	Adult	Male	Second	14	168
3.	Adult	Male	Third	75	462
4.	Adult	Male	Crew	192	862
5.	Child	Male	First	5	5
6.	Child	Male	Second	11	11
7.	Child	Male	Third	13	48
8.	Child	Male	Crew	0	0
9.	Adult	Female	First	140	144
10.	Adult	Female	Second	80	93
11.	Adult	Female	Third	76	165
12.	Adult	Female	Crew	20	23
13.	Child	Female	First	1	1
14.	Child	Female	Second	13	13
15.	Child	Female	Third	14	31
16.	Child	Female	Crew	0	0

Clearly, with simple reshaping data procedures, available in Stata, it is possible to swap from one format to another. This will allow us to implement each of the three approaches summarized in the previous section.

3.1 The casewise approach

The `glm` procedures, available in Stata, use as the default saturated model definition the model with as many covariates as the number of records in the data file. Thus using `individ.txt` with subjects as analytical units, we can easily implement the deviance test by considering the casewise definition of the saturated model, shown below:

```
. insheet using individ.txt, tab clear
(5 vars, 2201 obs)

. generate male = sex=="Male"

. encode status, generate(econ_status)

. glm survival i.male i.econ_status, family(binomial) link(logit)

Iteration 0:   log likelihood = -1116.4813
Iteration 1:   log likelihood = -1114.4582
Iteration 2:   log likelihood = -1114.4564
Iteration 3:   log likelihood = -1114.4564

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]
-------------+----------------------------------------------------------------
      1.male |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
             |
 econ_status |
          2  |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
          3  |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
          4  |  -.7774228   .1423145    -5.46   0.000    -1.056354   -.4984916
             |
       _cons |   1.187396   .1574664     7.54   0.000      .878767    1.496024
------------------------------------------------------------------------------
```

The LRT for goodness of fit can be obtained with the following code:

```
. scalar dev=e(deviance)

. scalar df=e(df)

. display "GOF casewise "" G^2="dev " df="df " p-value= " chiprob(df, dev)
GOF casewise  G^2=2228.9128 df=2196 p-value= .30705384
```

Thus the deviance statistic G^2 is 2228.91 with 2196 $(= 2201-5)$ degrees of freedom, and the p-value referred to the deviance test is 0.3071. We notice that as expected, the G^2 corresponds to $-2\{\ln_m(\widehat{\boldsymbol{\beta}})\}$ $(= -2[-1114.46])$. So in this case, the null hypothesis cannot be rejected, and the fit of the model of interest is not different from the fit of the saturated model.

3.2 The contingency table approach

The intuitive way of implementing the contingency table approach is to apply this same procedure (`glm`) on `grouped.txt`. In this situation, we want to estimate the proportion of successes; thus we need to redefine the outcome by specifying two new variables: the first is the total number of subjects in each category `n`, and the second is the total number of events, `survival`.

In Stata, we also need to add the variable containing the number of trials, `n`, in the `family()` option:

```
. insheet using grouped.txt, tab clear
(5 vars, 16 obs)
. generate male = sex=="Male"
. encode status, generate(econ_status)
. glm survival i.male i.econ_status if n>0, family(binomial n) link(logit)
Iteration 0:   log likelihood = -91.841683
Iteration 1:   log likelihood = -89.026084
Iteration 2:   log likelihood = -89.019672
Iteration 3:   log likelihood = -89.019672
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]
-------------+----------------------------------------------------------------
      1.male |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
             |
 econ_status |
          2  |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
          3  |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
          4  |  -.7774228   .1423145    -5.46   0.000    -1.056354   -.4984916
             |
       _cons |   1.187396   .1574664     7.54   0.000      .878767    1.496024
------------------------------------------------------------------------------
```

Now we can obtain the deviance test statistic:

```
. scalar dev=e(deviance)
. scalar df=e(df)
. display "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
```

The parameter estimates do not change as they do in the casewise approach. But as expected, the deviance statistic (131.42) has significantly decreased; the degrees of freedom have changed ($9 = 14 - 5$); and the p-value for the deviance test will now let us reject the null hypothesis, implying that the model of interest is not as good as the saturated model.

3.3 The collapsing approach

Both the casewise and contingency table approaches can be applied very easily by using the procedures shown above, whereas the collapsing approach requires more effort.

Thus, concerning the `grouped` dataset and by using the `egen` command, we first generate a variable that allows us to identify all the possible covariate patterns referring just to the variables `male` and `econ_status`.

```
. insheet using grouped.txt, tab clear
(5 vars, 16 obs)
. generate male = sex=="Male"
. encode status, generate(econ_status)
. egen trtp=group(male econ_status)
. list
```

	age	sex	status	survival	n	male	econ_s~s	trtp
1.	Adult	Male	First	57	175	1	First	6
2.	Adult	Male	Second	14	168	1	Second	7
3.	Adult	Male	Third	75	462	1	Third	8
4.	Adult	Male	Crew	192	862	1	Crew	5
5.	Child	Male	First	5	5	1	First	6
6.	Child	Male	Second	11	11	1	Second	7
7.	Child	Male	Third	13	48	1	Third	8
8.	Child	Male	Crew	0	0	1	Crew	5
9.	Adult	Female	First	140	144	0	First	2
10.	Adult	Female	Second	80	93	0	Second	3
11.	Adult	Female	Third	76	165	0	Third	4
12.	Adult	Female	Crew	20	23	0	Crew	1
13.	Child	Female	First	1	1	0	First	2
14.	Child	Female	Second	13	13	0	Second	3
15.	Child	Female	Third	14	31	0	Third	4
16.	Child	Female	Crew	0	0	0	Crew	1

Second, we collapse the data by using the variable obtained in the previous step and applying it to the two variables introduced into the model of interest (`male` and `econ_status`). In this way, we obtain a dataset where each record corresponds to a covariate pattern identified by the combination of the covariates in the model.

```
. collapse (sum) survival n (first) male econ_status, by(trtp)
. list
```

	trtp	survival	n	male	econ_s~s
1.	1	20	23	0	1
2.	2	141	145	0	2
3.	3	93	106	0	3
4.	4	90	196	0	4
5.	5	192	862	1	1
6.	6	62	180	1	2
7.	7	25	179	1	3
8.	8	88	510	1	4

We continue as we did in the contingency table approach:

```
. glm survival i.male i.econ_status if n>0, family(binomial n) link(logit)
Iteration 0:   log likelihood =  -54.70349
Iteration 1:   log likelihood = -52.362699
Iteration 2:   log likelihood = -52.356281
Iteration 3:   log likelihood = -52.356281
Generalized linear models                          No. of obs      =         8
Optimization     : ML                              Residual df     =         3
                                                   Scale parameter =         1
Deviance         =  65.17983096                    (1/df) Deviance =  21.72661
Pearson          =  60.87983277                    (1/df) Pearson  =  20.29328
Variance function: V(u) = u*(1-u/n)                [Binomial]
Link function    : g(u) = ln(u/(n-u))              [Logit]
                                                   AIC             =  14.33907
Log likelihood   = -52.35628099                    BIC             =  58.94151

------------------------------------------------------------------------------
             |                 OIM
    survival |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.male |  -2.421328   .1390931   -17.41   0.000    -2.693946   -2.148711
             |
 econ_status |
          2  |   .8808128   .1569718     5.61   0.000     .5731537    1.188472
          3  |  -.0717844   .1709268    -0.42   0.675    -.4067948     .263226
          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)
. display "GOF contingency "" G^2="dev " df="df " p-value= " chiprob(df, dev)
GOF contingency  G^2=65.179831 df=3 p-value= 4.591e-14
```

By reshaping the data, we obtain the results according to the collapsing approach definition. As in the contingency table approach, we reject H_0, but now the value of the deviance statistic has changed to 65.18 with 3 ($= 8 - 5$) degrees of freedom. As expected, estimates do not change.

4 Discussion

The casewise approach is often considered the standard for defining the saturated model. The reason is that the analysis is focused on subjects, and the saturated model, instead of the fully parameterized model, is seen as the model that gives the "perfect fit" (see Kleinbaum and Klein [2010]). This fact does not affect the estimation process; however, it fatally compromises the inferential step in a goodness-of-fit evaluation where the χ^2 approximation becomes questionable. The consideration of the other approaches can lead to different and meaningful results in terms of both descriptive and inferential analysis, but the problem is how to implement them in the right way with the statistical package we are working on.

Considering Stata 12.1, we have noticed that in all cases, the default procedures for goodness of fit consider the saturated model to be the one with as many covariates as the number of records present in the dataset. Thus, using an individual data layout, we obtain results relative to the casewise saturated model, where the analytical units are subjects. However, when considering an events–trials data format, we assess the goodness of fit based on the contingency table approach, where the unit of analysis is the covariate pattern defined by the possible values of all the independent variables in the dataset. The less intuitive implementation is the one based on the collapsing approach, which uses the covariate patterns defined by the variables involved in the model. One simple solution could be to build a new dataset containing only these variables, like we did with the useful commands `egen` and `collapse`, which are very helpful in showing how the collapsing approach works.

5 References

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken, NJ: Wiley.

Greenland, S., J. Pearl, and J. M. Robins. 1999. Causal diagrams for epidemiologic research. *Epidemiology* 10: 37–48.

Hosmer, D. W., T. Hosmer, S. Le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 15: 965–980.

Kleinbaum, D. G., and M. Klein. 2010. *Logistic Regression: A Self-Learning Text*. 3rd ed. New York: Springer.

Kuss, O. 2002. Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine* 21: 3789–3801.

Simonoff, J. S. 1998. Logistic regression, categorical predictors, and goodness-of-fit: It depends on who you ask. *American Statistician* 52: 10–14.

About the authors

Rino Bellocco is an associate professor of biostatistics in the Department of Statistics and Quantitative Methods at the University of Milano–Bicocca, Italy, and in the Department of Medical Epidemiology and Biostatistics at the Karolinska Institutet, Sweden.

Sara Algeri is a statistician and currently a PhD student at Texas A&M University in College Station, TX. She obtained both her bachelor's and her master's degrees from the University of Milano–Bicocca, Italy. In the last year of her master's studies, she worked at Mount Sinai School of Medicine in New York as a visiting master's student. This experience has been crucial in developing her interest in biostatistics and clinical trials. Her current research mainly focuses on longitudinal data analysis, Bayesian statistics, and statistical applications in genetics.