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.

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

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2011)
11, Number 3, pp. 386–402

The impact of different sources of body mass index assessment on smoking onset: An application of multiple-source information models

Maria Paola Caria
Karolinska Institutet, Sweden
Avogadro University, Novara, Italy
maria.paola.caria@ki.se

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

Maria Rosaria Galanti
Karolinska Institutet, Sweden
rosaria.galanti@ki.se

Nicholas J. Horton
Smith College, Northampton, MA
nhorton@smith.edu

Abstract. Multiple-source data are often collected to provide better information of some underlying construct that is difficult to measure or likely to be missing. In this article, we describe regression-based methods for analyzing multiple-source data in Stata. We use data from the BROMS Cohort Study, a cohort of Swedish adolescents who collected data on body mass index that was self-reported and that was measured by nurses. We draw together into a single frame of reference both source reports and relate these to smoking onset. This unified method has two advantages over traditional approaches: 1) the relative predictiveness of each source can be assessed and 2) all subjects contribute to the analysis. The methods are applicable to other areas of epidemiology where multiple-source reports are used.

Keywords: st0234, multiple informants, multiple-source predictors, regression analysis, generalized estimating equations, missing data

1 Introduction

One of the fundamental tasks of modern epidemiology is to quantify the association between a risk factor and the outcome of interest (Rothman, Greenland, and Lash 2008) while taking account of possible biases, such as confounding, measurement error, or misclassification. While confounding can be addressed by appropriate regression methods, and many articles in the *Stata Journal* have already tackled this topic (Fewell et al. 2004; Wang 2007; Cummings 2009), the issue of measurement error is often more delicate and requires ad hoc methods (Hardin, Schmiediche, and Carroll 2003).

Body mass index (BMI; weight in kilograms per square meter of height) is frequently used in epidemiological studies to assess prevalence of overweight and obese people in populations. Because of constraints on time, money, location, and personnel, obtaining

measures of the two components of BMI (height and weight) is not always feasible. Therefore, many epidemiologic studies use a single source of information, typically by asking subjects to self-report their weight and height.

The BROMS Cohort Study, a seven-year cohort of Swedish pupils (Galanti et al. 2001), has used information collected from two sources: one is the typical self-report information and the other is a measurement taken on the same students by school nurses. By collecting reports from multiple sources, one expects that BMI can be more accurately and reliably determined.

In recent years, multiple-source reports (also known as multiple-informant data, proxy reports, and coinformants) have been used in a variety of different fields of study. Multiple-source data can be used to better define both the exposure and the outcome of interest. However, many of the traditional methods for analyzing multiple-source data are not completely satisfactory.

In response to the shortcomings of existing analytic methods for multiple-source data, Horton and colleagues (Horton, Laird, and Zahner 1999; Horton and Fitzmaurice 2004) have proposed regression methodology for simultaneously analyzing information from multiple-source predictors. These models allow separate regressions to be analyzed together, tested for differences, and simplified if appropriate to determine a final overall regression. Partially observed multiple-source reports may be incorporated into these regression models to account for differential missingness.

Our goal with this article is to illustrate how to implement these models in Stata using data from a cohort study (the BROMS Cohort Study) where the multiple reports of BMI (self-reported and measured by nurses) are used as predictors in a regression model to understand the association between being overweight and subsequent smoking onset in Swedish adolescents.

2 Example: Multiple reports of BMI and smoking onset

2.1 Study sample

In this article, we analyze data from the BROMS (Swedish acronym for Children's Smoking and Environment in the Stockholm County) cohort, established at the Stockholm Centre of Public Health to study the uptake of smoking in Swedish adolescents over time (Galanti et al. 2001). The cohort was selected in 1998 through a random sample of all schools in the Stockholm region. The data consist of observations on 3,020 children of both sexes recruited in the fifth grade of compulsory school (at the age of 11 years) with follow up until age 18. Cigarette smoking was self-reported by the adolescents in a yearly paper-and-pencil questionnaire. Parental cigarette smoking was reported at baseline and categorized dichotomously as "at least one parent" versus "neither parent" currently smoking cigarettes.

During the school survey at age 14, the adolescents' weight and height were measured by the school nurses using a standardized protocol (standing, without clothes and shoes)

as well as self-reported by the adolescents with a questionnaire, under the same specified conditions of measurement. Among students with complete data on smoking habits, 2,052 students provided information on their anthropometric measures, and 2,349 were visited by the school nurse, who was able to provide the same measurement. Although 1,743 subjects had both types of information available, 915 had only one report available (309 with only a self report and 606 with only a nurse report). These 2,658 subjects constitute the sample of interest in this article.

2.2 Variables

Smoking onset at age 18 (`beginsmo18`) will be the binary outcome variable in these analyses, with predictor variables `bmi14_nurse` (BMI at age 14 measured by the nurses) and `bmi14_self` (BMI at age 14 measured by the students). In addition, we will also use a covariate that records a student's gender (`female`) and a binary covariate that records the parental smoking status (`famsmoke`).

```
. describe beginsmo18 bmi14_nurse bmi14_self female famsmoke
variable      storage display  value
name          type    format   label   variable label
------------------------------------------------------------------------------
beginsmo18    float   %9.0g    yesno   Smoking onset at age 18
bmi14_nurse   float   %9.0g            BMI at age 14 measured by the nurses
bmi14_self    float   %9.0g            BMI at age 14 self reported by the students
female        float   %9.0g    yesno   Gender: 1 = female, 0 = male
famsmoke      float   %9.0g    yesno   Parental smoking status
```

3 Methods

We first establish some notation. We assume that there are N independent subjects. Let Y denote a univariate outcome for a given subject (smoking onset in our example). Let X_j denote the jth multiple-source predictor. In the BROMS Cohort Study, we have two sources ($J = 2$), where X_1 denotes the first source report (BMI self-reported by students) and X_2 denotes the second source report (BMI measured by nurses). The latent variable Q represents the unobserved true value of BMI. Let Z denote a vector of other covariates of interest for the subject (gender and parental smoking status). The general regression model of interest is $f(Y|X, Z)$.

3.1 Analytic approaches

Consensus decision

A first approach when the multiple sources are categorical could be to force a consensus decision. The model would be $f(Y|Q, Z)$. This forced decision generally needs to be done at the data-collection stage and as a result may not always be possible.

Separate analyses for each source

Another simple but grossly inefficient (and too common!) approach is to use only one source and fit either the model $f(Y|X_1, Z)$ or the model $f(Y|X_2, Z)$.

This approach, however, addresses sensitivity of choice of the source: 1) separate analyses yield multiple (and often differing) sets of results for the different sources, which may be difficult to interpret; 2) separate analyses provide no formal means of evaluating how similar or different the results are across the various sources (or of summarizing them in a single set of results, if they are sufficiently similar); and 3) separate analyses may be based on different subsets of the data if some subjects are missing data from one source and others are missing data from another source.

Combining (pooling) sources

The "pooling" strategy, where information from multiple sources is combined into a single summary number for each subject, has been a common alternative to separate analyses in the past. A variety of strategies and algorithms for pooling multiple-source data have been introduced (Horton, Laird, and Zahner 1999). For example, a strategy that is appealing when the source data are quantitative is to take the arithmetic average of the multiple-source data:

$$\text{MEAN} = \frac{(X_1 + X_2)}{2}$$

The model of interest will then be $f(Y|\text{MEAN}, Z)$.

When the multiple sources are dichotomous, variants such as "OR" rules or "AND" rules may be useful.

Although this approach simplifies the analysis, there are many reasons why the pooling of data from multiple sources is not very desirable: 1) the optimal algorithm for combining multiple-source data depends on the type of measurement error present; 2) pooling does not permit the examination of differences in risk-factor effects across sources; and 3) many pooling algorithms are not clearly defined in the presence of missing data from one or more sources.

Including both source reports

Another standard approach is to fit a regression model that includes both source reports:

$$\begin{aligned} f(Y|X_1, X_2, Z) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 (X_1 \times X_2) \\ + \beta_5 (X_1 \times Z) + \beta_6 (X_2 \times Z) \end{aligned} \tag{1}$$

Here the regression parameters are interpreted in terms of the effect of a report from one source, conditional on the report of the other source. This may be the appropriate model if prediction is of primary interest. However, in many settings, the marginal association of each source report with the outcome may be of greater scientific interest.

In addition, the association between the risk factor and the outcome will generally be attenuated in this model because of the conditioning on all source reports.

Unified multiple-source regression model

Analytic methods of analysis for multiple-source predictor data have been described independently by Horton, Laird, and Zahner (1999) and Pepe, Whitaker, and Seidel (1999). They proposed the simultaneous estimation of separate regression equations, one for each source report. In the BROMS example, this could be represented by the following model:

$$\begin{array}{rcl} f(Y|X_1, Z) & = & \beta_0 + \beta_1 X_1 + \beta_2 Z \\ f(Y|X_2, Z) & = & \beta_0 + \gamma_0 + (\beta_1 + \gamma_1) X_2 + (\beta_2 + \gamma_2) Z \end{array} \tag{2}$$

One advantage of this approach is that it facilitates testing for source effects, that is, whether the regression models are sensitive to the choice of source. If the values γ are nonzero, then the models depend on the source. A test of $\gamma_1 = 0$ can be used to determine if the effects of BMI on the outcome differ by source. A test of $\gamma_2 = 0$ can be used to determine if the effects of other covariates on the outcome differ by source.

The following bivariate (two lines per subject) regression allows a single regression model to be fit to the multiple-source predictor data specified by (2):

$$f(Y|X) = \beta_0 + \gamma_0 \text{NURSE} + \beta_1 X + \gamma_1 (\text{NURSE} \times X) + \beta_2 Z + \gamma_2 (\text{NURSE} \times Z) \tag{3}$$

where NURSE is an indicator variable included in the model to indicate whether the BMI value was self-reported by the student (NURSE $= 0$) or measured by the nurse (NURSE $= 1$) and where

$$X = \begin{cases} X_1 & \text{if NURSE} = 0 \\ X_2 & \text{if NURSE} = 1 \end{cases}$$

Equation (3) assumes that the association between BMI and smoking onset as well as the association between the covariate and smoking onset may vary by source (nurse or student). Here each subject in the study contributes two lines to the dataset, with different values of the NURSE variable for the two lines. Additional predictor variables and interactions can be incorporated.

In general, source-related differences in the effect of BMI can be evaluated via tests of the γ parameters equaling zero. For example, the simplified bivariate regression model

$$f(Y|X) = \beta_0 + \gamma_0 \text{NURSE} + \beta_1 X + \beta_2 Z \tag{4}$$

assumes that neither the association between BMI and outcome ($\gamma_1 = 0$) nor the association between covariate and outcome ($\gamma_2 = 0$) vary by source.

This methodology is a special case of the generalized estimating equations (GEE) approach (Liang and Zeger 1986), in which the relationship between the outcome and

each predictor can be modeled separately (but estimated simultaneously). Unlike a traditional GEE, the outcomes are the two outcomes for the two lines per subject, but the value of the predictor differs. An independence working correlation matrix is specified, along with an empirical (`robust` in Stata parlance) variance estimator. The model can incorporate complex survey sampling designs (Särndal, Swensson, and Wretman 1992; Horton and Fitzmaurice 2004) and can easily be fit using Stata.

4 Results

We begin by reading in the dataset and creating the analytic set:

```
. use broms_source
. keep bmi14_nurse bmi14_self beginsmo18 female famsmoke
```

4.1 Summary statistics and distribution of variables

We then describe the variables of interest:

```
. tabulate beginsmo18
```

Smoking onset at age 18	Freq.	Percent	Cum.
No	1,565	58.88	58.88
Yes	1,093	41.12	100.00
Total	2,658	100.00	

```
. summarize bmi14_nurse bmi14_self
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bmi14_nurse	2349	20.55556	3.141448	13.97107	37.72291
bmi14_self	2052	19.92175	2.692268	13.38776	38.96455

```
. tabulate female
```

female	Freq.	Percent	Cum.
male	1,334	50.19	50.19
female	1,324	49.81	100.00
Total	2,658	100.00	

```
. tabulate famsmoke
```

famsmoke	Freq.	Percent	Cum.
No	1,654	62.23	62.23
Yes	1,004	37.77	100.00
Total	2,658	100.00	

4.2 Analytic approaches

Separate analyses for each source

Figure 1 displays a lowess (locally weighted smoothing spline) and straight-line fit for the association between age of onset of smoking and nurse-reported BMI.

```
. twoway scatter beginsmo18 bmi14_nurse || (lowess beginsmo18 bmi14_nurse)
> || (lfit beginsmo18 bmi14_nurse), scheme(sj)
```

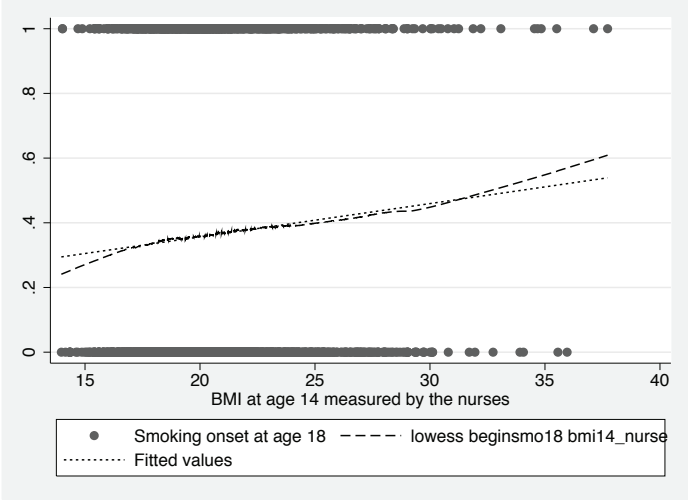

Figure 1. Lowess and straight-line fit for the association between nurse-reported BMI and age of onset of smoking

Figure 2 displays a lowess and straight-line fit for the association between age of onset of smoking and self-reported BMI.

```
. twoway scatter beginsmo18 bmi14_self || (lowess beginsmo18 bmi14_self)
> || (lfit beginsmo18 bmi14_self), scheme(sj)
```

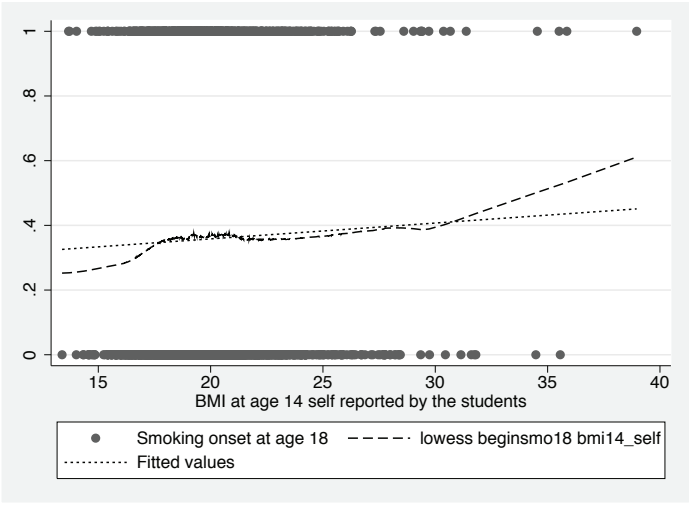

Figure 2. Lowess and straight-line fit for the association between self-reported BMI and age of onset of smoking

The assumption of a linear association between nurse-reported, as well as self-reported, BMI and the probability of smoking seems reasonable in the midrange of BMI values, though there is greater deviation between the straight line and the lowess with self-reported BMI.

We first fit a logistic regression model separately for each source, controlling for gender and parental smoking status. Nonsignificant interaction terms are dropped from the models.

```
. logistic beginsmo18 bmi14_nurse female famsmoke
Logistic regression                             Number of obs   =       2349
                                                LR chi2(3)      =      72.39
                                                Prob > chi2     =     0.0000
Log likelihood = -1552.9799                     Pseudo R2       =     0.0228

------------------------------------------------------------------------------
  beginsmo18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 bmi14_nurse |   1.035984   .014013     2.61   0.009      1.00888    1.063817
      female |   1.443318   .1230764    4.30   0.000     1.221174    1.705873
    famsmoke |   1.773314   .1548034    6.56   0.000     1.494442    2.104225
       _cons |   .2231046   .0637182   -5.25   0.000     .1274698    .3904899
------------------------------------------------------------------------------
```

```
. estimates store Separate_nurse
. logistic beginsmo18 bmi14_self female famsmoke
Logistic regression                                 Number of obs   =       2052
                                                    LR chi2(3)      =      52.00
                                                    Prob > chi2     =     0.0000
Log likelihood = -1356.7054                         Pseudo R2       =     0.0188

------------------------------------------------------------------------------
  beginsmo18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  bmi14_self |   1.023967    .017413     1.39   0.164     .9904009    1.058671
      female |   1.539319   .1416506     4.69   0.000     1.285286    1.843561
    famsmoke |   1.642746   .1542083     5.29   0.000     1.366678     1.97458
       _cons |   .2783781    .097847    -3.64   0.000     .1397804    .5544005
------------------------------------------------------------------------------
. estimates store Separate_self
```

When BMI is assessed using the self-report information, the odds ratio of smoking onset is 1.024 and not statistically significant (95% confidence interval [CI]: [0.99, 1.06]). When BMI measured by nurses is used, the odds ratio of smoking onset is 1.036 and statistically significant (95% CI: [1.01, 1.06]). The association between BMI and smoking appears to be stronger when using the nurse-reported values. An obvious limitation of this analysis is that there is no way to quantify this difference from the separate model.

Combining (pooling) sources

We generate a variable, `bmi14_mean1`, that is equal to the arithmetic average of the two BMI values when both are available and otherwise is equal to the only value available. We then fit a logistic regression, controlling for gender and parental smoking status:

```
. egen bmi14_mean1 = rowmean(bmi14_nurse bmi14_self)
. logistic beginsmo18 bmi14_mean1 female famsmoke
Logistic regression                                 Number of obs   =       2658
                                                    LR chi2(3)      =      74.47
                                                    Prob > chi2     =     0.0000
Log likelihood = -1763.0187                         Pseudo R2       =     0.0207

------------------------------------------------------------------------------
  beginsmo18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 bmi14_mean1 |   1.034027    .013712     2.52   0.012     1.007498    1.061254
      female |   1.464281   .1171575     4.77   0.000     1.251756     1.71289
    famsmoke |   1.706915   .1396021     6.54   0.000     1.454104     2.00368
       _cons |   .2369137   .0657249    -5.19   0.000     .1375462    .4080672
------------------------------------------------------------------------------
. estimates store Mean1
```

For every one-point increase in BMI, the odds of having started smoking by age 18 increases by 0.034. This odds ratio is statistically significant (95% CI: [1.01, 1.06]). Then we generate another variable, `bmi14_mean2`, which reports the arithmetic average of the two BMI values when both are available and reports a missing value if one of the sources is missing.

```
. generate bmi14_mean2 = (bmi14_nurse + bmi14_self) / 2
. logistic beginsmo18 bmi14_mean2 female famsmoke
(915 missing values generated)
Logistic regression                                 Number of obs   =       1743
                                                    LR chi2(3)      =      49.34
                                                    Prob > chi2     =     0.0000
Log likelihood = -1146.6655                         Pseudo R2       =     0.0211

------------------------------------------------------------------------------
  beginsmo18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 bmi14_mean2 |   1.029351   .0187437     1.59   0.112      .993262    1.066752
      female |    1.53314   .1531183     4.28   0.000     1.260579    1.864633
   famsmoke  |   1.712501   .1755853     5.25   0.000     1.400734    2.093658
       _cons |   .2422215   .0919753    -3.73   0.000     .1150795     .509832
------------------------------------------------------------------------------
. estimates store Mean2
```

Again for every one-point increase in BMI, the odds of having started smoking by age 18 increases by about 3%, but this effect is no longer statistically significant (95% CI: [0.99, 1.07]).

Including both source reports

We next fit a regression akin to (1) that includes both source reports (and potentially their interaction):

```
. logistic beginsmo18 bmi14_self bmi14_nurse female famsmoke
Logistic regression                                 Number of obs   =       1743
                                                    LR chi2(4)      =      49.44
                                                    Prob > chi2     =     0.0000
Log likelihood = -1146.6167                         Pseudo R2       =     0.0211

------------------------------------------------------------------------------
  beginsmo18 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  bmi14_self |   1.000548   .0454791     0.01   0.990     .9152663    1.093777
 bmi14_nurse |   1.027377   .0422885     0.66   0.512     .9477479    1.113696
      female |   1.525057   .1544594     4.17   0.000     1.250476    1.859931
   famsmoke  |   1.711882   .1755344     5.24   0.000     1.400208    2.092931
       _cons |   .2481597   .0961403    -3.60   0.000     .1161354    .5302709
------------------------------------------------------------------------------
. estimates store Adjusted
```

Neither report of BMI is statistically significant. However, while this model may be attractive if the primary goal is prediction of the outcome, the regression parameters now are interpreted in terms of the effect on the outcome of a one-point increase in the BMI report from one source, conditional on the report of the other source (and of the other covariates) being held fixed. In addition to being challenging to interpret, this model will tend to have an attenuated association if both sources have a positive correlation.

Unified multiple-source regression model

To fit the models of Horton, Laird, and Zahner (1999) as specified in (3), we need to reshape the dataset from wide to long format (that is, from one observation per subject to two observations per subject).

```
. generate id = _n
. rename bmi14_nurse bmi141
. rename bmi14_self bmi140
. list id female bmi141 bmi140 famsmoke beginsmo18 if id < 6
```

	id	female	bmi141	bmi140	famsmoke	beginsmo18
1.	1	No	18.80921	.	Yes	No
2.	2	No	20.98399	21.38594	No	No
3.	3	Yes	28.51563	.	No	No
4.	4	Yes	21.33821	20.57613	No	No
5.	5	Yes	16.09645	15.29291	No	No

```
. reshape long bmi14, i(id) j(nurse)
(note: j = 0 1)
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                     2658   ->    5316
Number of variables                  12   ->      12
j variable (2 values)                     ->   nurse
xij variables:
                         bmi140 bmi141   ->   bmi14
-----------------------------------------------------------------------------
```

```
. list id beginsmo18 bmi14 nurse female famsmoke if id < 6
```

	id	beginsmo18	bmi14	nurse	female	famsmoke
1.	1	No	.	0	No	Yes
2.	1	No	18.80921	1	No	Yes
3.	2	No	21.38594	0	No	No
4.	2	No	20.98399	1	No	No
5.	3	No	.	0	Yes	No
6.	3	No	28.51563	1	Yes	No
7.	4	No	20.57613	0	Yes	No
8.	4	No	21.33821	1	Yes	No
9.	5	No	15.29291	0	Yes	No
10.	5	No	16.09645	1	Yes	No

We fit a logistic regression model accounting for the clustering of multiple-source observations within subject. The regression model controls for the main effect of source, gender, and parental smoking status, as well as the interaction between each variable and source (the γ terms). We retained interactions if the overall p-value was less than or equal to 0.05. Other strategies may be used to find a balance between a parsimonious and flexible model.

```
. xtset id
      panel variable:  id (balanced)
. xtgee beginsmo18 nurse##(c.bmi14 female famsmoke), link(logit) corr(ind)
> family(binomial) vce(robust) eform nolog
GEE population-averaged model                   Number of obs      =      4401
Group variable:                         id      Number of groups   =      2658
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       1.7
Correlation:                   independent                     max =         2
                                               Wald chi2(7)       =     74.40
Scale parameter:                         1     Prob > chi2        =    0.0000

Pearson chi2(4401):                4404.44     Deviance           =   5819.37
Dispersion (Pearson):             1.000781     Dispersion         =  1.322284
                                          (Std. Err. adjusted for clustering on id)
```

beginsmo18	Odds Ratio	Semirobust Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.nurse	.8014445	.2124384	-0.84	0.404	.4767009	1.347414
bmi14	1.023967	.0172544	1.41	0.160	.9907017	1.05835
1.female	1.539319	.1417153	4.69	0.000	1.28518	1.843713
1.famsmoke	1.642746	.1540527	5.29	0.000	1.366932	1.974213
nurse#c.bmi14						
1	1.011736	.0129836	0.91	0.363	.9866057	1.037506
nurse#female						
1 1	.937634	.0552886	-1.09	0.275	.8352977	1.052508
nurse#famsmoke						
1 1	1.079481	.0651481	1.27	0.205	.9590562	1.215028
_cons	.2783781	.0970596	-3.67	0.000	.1405576	.5513353

There is little evidence for source effects (testing $\gamma_3 = \gamma_2 = \gamma_1 = 0$):

```
. testparm nurse#c.bmi14 nurse#female nurse#famsmoke
 ( 1)  1.nurse#c.bmi14 = 0
 ( 2)  1.nurse#1.female = 0
 ( 3)  1.nurse#1.famsmoke = 0

           chi2(  3) =    4.37
         Prob > chi2 =    0.2241
```

In addition, none of the individual CIs came close to excluding 0 (all γ p-values were greater than or equal to 0.205). We then refit a regression similar to (4) after dropping the nonsignificant interactions (and the source main effect):

```
. xtgee beginsmo18 bmi14 female famsmoke, link(logit) corr(ind)
> family(binomial) vce(robust) eform

Iteration 1: tolerance = 2.252e-10

GEE population-averaged model                   Number of obs      =      4401
Group variable:                         id      Number of groups   =      2658
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       1.7
Correlation:                   independent                     max =         2
                                                Wald chi2(3)       =     67.69
Scale parameter:                         1      Prob > chi2        =    0.0000

Pearson chi2(4401):                4404.37      Deviance           =   5820.38
Dispersion (Pearson):              1.000767     Dispersion         =  1.322513

                                    (Std. Err. adjusted for clustering on id)
```

beginsmo18	Odds Ratio	Semirobust Std. Err.	z	P>|z|	[95% Conf.	Interval]
bmi14	1.03116	.0137493	2.30	0.021	1.004561	1.058463
female	1.490927	.1242893	4.79	0.000	1.266185	1.755561
famsmoke	1.710896	.1455271	6.31	0.000	1.448175	2.021277
_cons	.2437989	.0683924	-5.03	0.000	.1406844	.4224912

```
. estimates store Unified
```

This yields a shared parameter model with a parameter estimate between that of self and nurse report. Table 1 summarizes the results from the different models by using the user-written `estout` command (Jann 2005) (values with an * do not include 1 in the associated 95% CI):

```
. estout Separate_nurse Separate_self Mean1 Mean2 Adjusted Unified,
> cells(b(star fmt(2)) ci(fmt(2)) se(par fmt(3))) legend eform
> title("Odds ratio, confidence interval and standard error for different models")
> order(bmi14_self bmi14_nurse bmi14_mean1 bmi14_mean2 bmi14 female famsmoke)
> mlabels("Separate nurse" "Separate self" "Pooled mean1" "Pooled mean2" "Adjusted"
> "Unified") style(smcl) starlevels(* 0.05) drop(_cons) collabels(none)
```

	Separate nurse	Separate self	Mean1	Mean2	Adjusted	Unified
main						
bmi14_self		1.02			1.00	
		0.99,1.06			0.92,1.09	
		(0.017)			(0.045)	
bmi14_nurse	1.04*				1.03	
	1.01,1.06				0.95,1.11	
	(0.014)				(0.042)	
bmi14_mean1			1.03*			
			1.01,1.06			
			(0.014)			
bmi14_mean2				1.03		
				0.99,1.07		
				(0.019)		
bmi14						1.03*
						1.00,1.06
						(0.014)
female	1.44*	1.54*	1.46*	1.53*	1.53*	1.49*
	1.22,1.71	1.29,1.84	1.25,1.71	1.26,1.86	1.25,1.86	1.27,1.76
	(0.123)	(0.142)	(0.117)	(0.153)	(0.154)	(0.124)
famsmoke	1.77*	1.64*	1.71*	1.71*	1.71*	1.71*
	1.49,2.10	1.37,1.97	1.45,2.00	1.40,2.09	1.40,2.09	1.45,2.02
	(0.155)	(0.154)	(0.140)	(0.176)	(0.176)	(0.146)

* p<0.05

Table 1. Odds ratio, CI, and standard error for different models

5 Conclusions

Researchers in epidemiology are often interested in the results of regression models based on multiple-source reports. Separate regression models for each source are straightforward to fit but difficult to interpret if they provide differing results. Also, interpretability of models where both sources are included can be problematic. In the BROMS Cohort Study, how should we interpret the effect on smoking onset of a one-unit increase of self-reported BMI while holding the nurse report constant? However, regression models for the combined reports have disadvantages in that they must often make a number of a priori assumptions, and they can yield biased estimates of the regression parameters and standard errors when there are missing source reports and the data are missing at random (Goldwasser and Fitzmaurice 2001).

We have illustrated methods using a single model that have several advantages over approaches that combine the reports. The proposed methods allow formal assessment of whether covariate (for example, risk factor) effects vary according to the source and allow for the pooling of information from different sources when appropriate. For example, in the analysis of the BROMS data, there are no significant source effects, so a single model that pooled information from nurse report and self report is fit to these data. This joint analysis of both source reports results in smaller standard errors than those obtained from separate analyses of each source report. As an example, the robust standard error for family smoking is 0.146, while for each of the separate source models it is at least 0.154.

Another appealing feature of the proposed methods is that they can be implemented using existing, general-purpose, statistical software. These methods are attractive because they can account for complex survey designs and can be generalized to other epidemiologic investigations that use multiple-source reports.

Our models assumed that the functional form of the relationship between BMI and smoking onset was approximately linear, though there was some indication in figures 1 and 2 of nonlinearity for extremely low and extremely high values. Additional analyses (not reported here) that allowed for quadratic form of the association yielded similar results. This is likely because of the relatively small number of subjects with extreme values.

A practical concern in analyzing multiple-source reports is the presence of missing data. While a full review of missing-data methods is beyond the scope of this article, missingness can induce bias and loss of efficiency (Little and Rubin 2002). This model allows partially observed subjects to contribute to the analysis and fully uses all available information. A limitation is that the GEE approach assumes that data are missing completely at random; that is, missingness does not depend on observed or unobserved measurements. Horton et al. (2001) described how to fit a weighted estimating equation model, which is unbiased when the missingness is missing at random in the sense of Little and Rubin (2002).

6 Acknowledgments

We thank the Stockholm Centre for Public Health—Tobacco Prevention and the Department of Public Health Sciences at the Karolinska Institutet, Stockholm, Sweden, for providing the data used in this article. We would like to thank Dr. Silvana Romio and Dr. Jens Lauritsen for their valuable suggestions. The BROMS Cohort Study was partially funded by the Swedish Research Council, grant number 345-2992-3515, while these analyses were also supported by NIH grant R01-MH54693.

7 References

Cummings, P. 2009. Methods for estimating adjusted risk ratios. *Stata Journal* 9: 175–196.

Fewell, Z., M. A. Hernan, F. Wolfe, K. Tilling, H. Choi, and J. A. C. Sterne. 2004. Controlling for time-dependent confounding using marginal structural models. *Stata Journal* 4: 402–420.

Galanti, M. R., I. Rosendahl, A. Post, and H. Gilljam. 2001. Early gender differences in adolescent tobacco use—the experience of a Swedish cohort. *Scandinavian Journal of Public Health* 29: 314–317.

Goldwasser, M. A., and G. Fitzmaurice. 2001. Multivariate linear regression of childhood psychopathology using multiple informant data. *International Journal of Methods in Psychiatric Research* 10: 1–11.

Hardin, J. W., H. Schmiediche, and R. J. Carroll. 2003. The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata Journal* 3: 373–385.

Horton, N. J., and G. M. Fitzmaurice. 2004. Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine* 23: 2911–2933.

Horton, N. J., N. M. Laird, J. M. Murphy, R. R. Monson, A. M. Sobol, and A. H. Leighton. 2001. Multiple informants: mortality associated with psychiatric disorders in the Stirling County Study. *American Journal of Epidemiology* 154: 649–656.

Horton, N. J., N. M. Laird, and G. E. P. Zahner. 1999. Use of multiple informant data as a predictor in psychiatric epidemiology. *International Journal of Methods in Psychiatric Research* 8: 6–18.

Jann, B. 2005. Making regression tables from stored estimates. *Stata Journal* 5: 288–308.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Pepe, M. S., R. C. Whitaker, and K. Seidel. 1999. Estimating and comparing univariate associations with application to the prediction of adult obesity. *Statistics in Medicine* 18: 163–173.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Wang, Z. 2007. Two postestimation commands for assessing confounding effects in epidemiological studies. *Stata Journal* 7: 183–196.

About the authors

Maria Paola Caria is a PhD student in the Department of Public Health Sciences at the Karolinska Institutet, Sweden, and a research fellow at Avogadro University, Italy. Her research concerns the evaluation of the effectiveness of school-based prevention for tobacco, alcohol, and drug use.

Rino Bellocco is an associate professor of biostatistics in the Department of Statistics at the University of Milano–Bicocca, Italy, and in the Department of Medical Epidemiology and Biostatistics at the Karolinska Institutet, Sweden. His research is currently on cancer epidemiology and the application of statistical methods in observational data.

Rosaria Galanti is an associate professor of epidemiology in the Department of Public Health Sciences at the Karolinska Institutet, Sweden. Her research encompasses longitudinal studies of determinants of tobacco use in youths.

Nicholas Horton is an associate professor in the Department of Mathematics and Statistics at Smith College, Northampton, MA. His research interests involve the development and dissemination of methods for the analysis of clustered and incomplete data.