AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

The Stata Journal

Editors

H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, Texas
editors@stata-journal.com

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
editors@stata-journal.com

Associate Editors

Christopher F. Baum, Boston College
Nathaniel Beck, New York University
Rino Bellocco, Karolinska Institutet, Sweden, and University of Milano-Bicocca, Italy
Maarten L. Buis, WZB, Germany
A. Colin Cameron, University of California–Davis
Mario A. Cleves, University of Arkansas for Medical Sciences
William D. Dupont, Vanderbilt University
Philip Ender, University of California–Los Angeles
David Epstein, Columbia University
Allan Gregory, Queen's University
James Hardin, University of South Carolina
Ben Jann, University of Bern, Switzerland
Stephen Jenkins, London School of Economics and Political Science
Ulrich Kohler, University of Potsdam, Germany
Frauke Kreuter, Univ. of Maryland–College Park
Peter A. Lachenbruch, Oregon State University
Jens Lauritsen, Odense University Hospital
Stanley Lemeshow, Ohio State University
J. Scott Long, Indiana University
Roger Newson, Imperial College, London
Austin Nichols, Urban Institute, Washington DC
Marcello Pagano, Harvard School of Public Health
Sophia Rabe-Hesketh, Univ. of California–Berkeley
J. Patrick Royston, MRC Clinical Trials Unit, London
Philip Ryan, University of Adelaide
Mark E. Schaffer, Heriot-Watt Univ., Edinburgh
Jeroen Weesie, Utrecht University
Nicholas J. G. Winter, University of Virginia
Jeffrey Wooldridge, Michigan State University

Stata Press Editorial Manager

Lisa Gilmore

Stata Press Copy Editors

David Culwell and Deirdre Skaggs

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.

Copyright © 2013 by StataCorp LP

Copyright Statement: The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, fileservers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 1, pp. 185–205

Doubly robust estimation in generalized linear models

Nicola Orsini
Unit of Biostatistics and Unit of Nutritional Epidemiology
Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Rino Bellocco
Department of Statistics and Quantitative Methods
University of Milano–Bicocca
Milan, Italy
and
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
rino.bellocco@ki.se

Arvid Sjölander
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
arvid.sjolander@ki.se

Abstract. A common aim of epidemiological research is to assess the association between a particular exposure and a particular outcome, controlling for a set of additional covariates. This is often done by using a regression model for the outcome, conditional on exposure and covariates. A commonly used class of models is the generalized linear models. The model parameters are typically estimated through maximum likelihood. If the model is correct, then the maximum likelihood estimator is consistent but may otherwise be inconsistent. Recently, a new class of estimators known as doubly robust estimators has been proposed. These estimators use two regression models, one for the outcome and one for the exposure, and are consistent if either model is correct, not necessarily both. Thus doubly robust estimators give the analyst two chances instead of only one to make valid inference. In this article, we describe a new Stata command, `drglm`, that implements the most common doubly robust estimators for generalized linear models.

Keywords: st0290, drglm, doubly robust, generalized linear model

© 2013 StataCorp LP

1 Introduction

A common aim of epidemiological research is to assess the association between a particular exposure and a particular outcome, controlling for a set of additional covariates. This is often done by fitting a regression model for the outcome, conditional on exposure and covariates. A commonly used class of models is the generalized linear models (GLMs). The model parameters are typically estimated through maximum likelihood (ML). If the model is correct, then the ML estimator is consistent but may otherwise be inconsistent.

When the mechanisms that bring about the outcome are well understood, the outcome is a natural target for regression modeling. Sometimes, the researcher may have a better understanding of the exposure mechanisms, in which case the exposure may be a more natural target. For example, this could be the case when the exposure is a treatment or a medical drug, which are typically assigned to patients according to reasonably well-defined protocols. Robins, Mark, and Newey (1992) showed that exposure regression models, like outcome regression models, can be used to estimate the conditional exposure–outcome association, given covariates.

Often the researcher may not have a strong preference for either modeling strategy, in which case a doubly robust (DR) estimator is attractive. A DR estimator requires one model for the outcome and one model for the exposure but is consistent if either model is correct, not necessarily both. Thus a DR estimator gives the researcher two chances instead of only one to make valid inference. Over the last decade, DR estimators have been developed for various parameters (see Bang and Robins [2005] and the references therein).

In this article, we describe a new Stata command, `drglm`, that implements DR estimators for GLMs. The article is organized as follows: In section 2, we establish notation and definitions and define the target estimand. In section 3, we review estimators that use outcome regression models, estimators that use exposure regression models, and DR estimators. The DR estimators that we review in section 3 are special cases of more general estimators developed in Robins (2000) and Tchetgen Tchetgen and Robins (2010). In section 4, we present the `drglm` command with syntax and options. In section 5, we carry out a simulation study to investigate the performance of the DR estimators, and in section 6, we describe a practical example.

2 Target parameter

Let A and Y denote the exposure and outcome of interest, respectively. Let L denote a vector of covariates that we wish to control for. We use $p(\cdot)$ generically for both population probabilities and densities, and we assume that data consist of n independent and identically distributed observations from $p(Y, A, L)$. We use $E(\cdot)$ for population means and $\widetilde{E}(\cdot)$ for sample means; that is, $E(R) = \int rp(r)dr$, and $\widetilde{E}(R) = \sum_{i=1}^{n} R_i/n$ for any random variable R.

A standard way to assess the conditional association between A and Y, given L, is to use a GLM on the form

$$g\{E(Y|A,L;\beta,\gamma)\} = \beta A + \gamma^T L \tag{1}$$

where β quantifies the conditional A-Y association, given L, and $g(\cdot)$ is a suitable link function. Typical link functions are the identity link (for continuous Y), the log link (for "counts"), and the logit link (for binary Y), for which β is a mean difference, a log risk-ratio, and a log odds-ratio, respectively. Typically, a constant term ("intercept") is included in the model. This can be achieved without changing notation by defining the first component of L to be the constant 1. The model in (1) has no interaction term between A and L; thus it assumes a constant strength of A-Y association on the scale defined by $g(\cdot)$ across levels of L. To allow for interactions between A and L and between separate components of L, we consider GLMs on the form

$$g\{E(Y|A,L;\beta,\gamma)\} = \beta^T AX + \gamma^T V \tag{2}$$

where X is a $(p \times 1)$-dimensional function of L, and V is a $(q \times 1)$-dimensional function of L. For instance, if $L = (L_1, L_2)$, $X = (1, L_1)$, and $V = (1, L_1, L_2, L_1L_2)$, then (2) reduces to

$$g\{E(Y|A,L;\beta,\gamma)\} = \beta_0 A + \beta_1 AL_1 + \gamma_0 + \gamma_1 L_1 + \gamma_2 L_2 + \gamma_{12} L_1 L_2$$

The model in (2) consists of two parts. The part

$$m(A,L;\beta) = g\{E(Y|A,L)\} - g\{E(Y|A=0,L)\} = \beta^T AX \tag{3}$$

quantifies the conditional A-Y association, given L, and is typically of main interest; we refer to it as the "main model". The parameter β in the main model (3) is our target parameter. The part

$$g\{E(Y|A=0,L;\gamma)\} = \gamma^T V \tag{4}$$

is primarily included to control for L; we refer to it as the "outcome nuisance model".

3 Estimators

3.1 Estimators that use the nuisance model for the outcome

We first consider an estimator of β that uses the outcome nuisance model for $E(Y|A = 0, L)$ in (4). This estimator is obtained by solving the estimating equation

$$\widetilde{E}\left[\begin{pmatrix} AX \\ V \end{pmatrix} \{Y - E(Y|A,L;\beta,\gamma)\}\right] = 0 \tag{5}$$

for $(\beta^T, \gamma^T)^T$. We use $\widehat{\beta}_{\text{OBE}}$ to denote the first p elements of the solution to (5), where OBE stands for outcome-based estimation. Using the law of iterated expectations, we have that

$$\begin{aligned} & E \left[\begin{pmatrix} AX \\ V \end{pmatrix} \{Y - E(Y|A, L; \beta, \gamma)\} \right] \\ = \; & E \left[\begin{pmatrix} AX \\ V \end{pmatrix} E\{Y - E(Y|A, L; \beta, \gamma)|A, L\} \right] \end{aligned}$$

which equals 0, so the estimating equation in (5) is unbiased when both (3) and (4) are correct. It follows from standard theory (Newey and McFadden 1994) that $\widehat{\beta}_{\text{OBE}}$ is consistent and asymptotically normal (CAN) when both (3) and (4) are correct.

In the standard use of GLMs, Y is assumed to follow a distribution in the exponential family, conditional on A and L. If $g(\cdot)$ is the canonical link function (for example, the identity link in the normal distribution, the log link in the Poisson distribution, and the logit link in the Bernoulli distribution), then $\widehat{\beta}_{\text{OBE}}$ is an ML estimator. $\widehat{\beta}_{\text{OBE}}$ is the default estimator produced by the `glm` command. We emphasize that $\widehat{\beta}_{\text{OBE}}$ is CAN even when it is not an ML estimator. The default standard errors produced by the `glm` command are consistent under the distributional assumption, but are generally inconsistent when the distributional assumption is incorrect. Consistent standard errors that do not rely on any distributional assumptions can be obtained through the "sandwich" formula by specifying the `vce(robust)` option in the `glm` command.

3.2 Estimators that use the nuisance model for the exposure

We next consider estimators of β that use the nuisance model for the exposure. We first give a heuristic argument for the case when $g(\cdot)$ is the identity link. Suppose that the true value of β was known. We could then construct residuals on the form $Y - m(A, L; \beta)$. These residuals unbiasedly predict $E(Y|A = 0, L)$. Conditionally on L, $E(Y|A = 0, L)$ is a constant and therefore uncorrelated with A. This argument suggests the following estimation strategy: find the value of β for which the residual $Y - m(A, L; \beta)$ becomes conditionally uncorrelated with A, given L, in the sample. In terms of an estimating equation, we find the value of β that solves

$$\widetilde{E}\left[X\{A - E(A|L)\}\{Y - m(A, L; \beta)\}\right] = 0 \tag{6}$$

Equation (6) involves $E(A|L)$, which typically is unknown. Therefore, we predict $E(A|L)$ by using the exposure nuisance model in the form

$$h\{E(A|L; \alpha)\} = \alpha^T Z \tag{7}$$

where $h(\cdot)$ is a smooth link function not necessarily equal to $g(\cdot)$ used in the main model (3) and in the outcome model (4). Z is an $(r \times 1)$-dimensional function of L, with the first element typically being the constant 1. We will allow for the identity link, the log link, and the logit link in the exposure nuisance (7). We fit the model in (7) by solving the unbiased estimating equation for α,

$$\widetilde{E}\left[Z\{A - E(A|L;\alpha)\}\right] = 0$$

and we replace the true value of $E(A|L)$ in (6) with the model-based prediction.

Combining these steps into one estimating equation for $(\beta^T, \alpha^T)^T$ gives

$$\widetilde{E}\left[\begin{array}{c} X\{A - E(A|L;\alpha)\}\{Y - m(A,L;\beta)\} \\ Z\{A - E(A|L;\alpha)\} \end{array}\right] = 0 \tag{8}$$

We use $\widehat{\beta}_{\text{EBE}}$ to denote the first p elements of the solution to (8), where EBE stands for exposure-based estimation. Using the law of iterated expectations, we have that

$$\begin{aligned} &E\left[\begin{array}{c} X\{A - E(A|L;\alpha)\}\{Y - m(A,L;\beta)\} \\ Z\{A - E(A|L;\alpha)\} \end{array}\right] \\ &\qquad = E\left[\begin{array}{c} XE\{A - E(A|L;\alpha)|L\}E(Y|A=0,L) \\ ZE\{A - E(A|L;\alpha)|L\} \end{array}\right] \end{aligned} \tag{9}$$

if (3) with the identity link is correct. If (7) is also correct, then the right-hand side of (9) equals 0, so the estimating equation in (8) is unbiased when both (3) with the identity link and (7) are correct. Thus $\widehat{\beta}_{\text{EBE}}$ is CAN when both (3) with the identity link and (7) are correct.

A minor modification is required when $g(\cdot)$ in (3) is the log link. For this link function, we replace $Y - m(A,L;\beta)$ on the first p rows in (8) with $Ye^{-m(A,L;\beta)}$. Using the law of iterated expectations, we can easily show that this modified estimating equation is unbiased when both (3) with the log link and (7) are correct.

We now consider the case when $g(\cdot)$ is the logit link. For this link, we assume that both A and Y are binary (0/1). We use the nuisance model in the form

$$\text{logit}\{E(A|Y=0,L;\delta)\} = \delta^T W \tag{10}$$

where W is an $(s \times 1)$-dimensional function of L, with the first element typically being the constant 1. Because of the symmetry of the odds ratio, (3) with the logit link and (10) together define the joint model

$$\text{logit}\{E(A|Y,L;\beta,\delta)\} = \beta^T YX + \delta^T W$$

Under (3) with the logit link and (10), an ML estimator of $(\beta^T, \delta^T)^T$ is obtained by solving the estimating equation

$$\widetilde{E}\left[\begin{pmatrix} YX \\ W \end{pmatrix}\{A - E(A|Y,L;\beta,\delta)\}\right] = 0 \tag{11}$$

Using the law of iterated expectations, we can show that the estimating equation in (11) is unbiased when both (3) with the logit link and (10) are correct. For simplicity, we use $\widehat{\beta}_{\text{EBE}}$ to denote the first p elements of the solution to either (8) or (11).

3.3 DR estimators

We finally consider DR estimators of β. We first consider the case when $g(\cdot)$ is the identity link. For this case, a DR estimator of β can be obtained by "combining" the estimating equations (5) and (8) into

$$\widetilde{E}\begin{bmatrix} X\{A-E(A|L;\alpha)\}\{Y-E(Y|A,L;\beta,\gamma)\} \\ \begin{pmatrix} AX \\ V \end{pmatrix}\{Y-E(Y|A,L;\beta^{\dagger},\gamma)\} \\ Z\{A-E(A|L;\alpha)\} \end{bmatrix}=0 \qquad (12)$$

and solving for $(\beta^T,\beta^{\dagger T},\gamma^T,\alpha^T)^T$. We use $\widehat{\beta}_{\mathrm{DR}}$ to denote the first p elements of the solution to (12). It follows from a more general result in Robins (2000) that the estimating equation in (12) is unbiased if either (4) with the identity link or (7) is correct, together with the main model (3) with the identity link.[1] Thus $\widehat{\beta}_{\mathrm{DR}}$ is CAN if either of the nuisance models is correct, not necessarily both.

A minor modification is required when $g(\cdot)$ is the log link. For this link function, we replace $Y-E(Y|A,L;\beta,\gamma)=Y-m(A,L;\beta)-E(Y|A=0,L;\gamma)$ on rows 1 through p in (12) with $Ye^{-m(A,L;\beta)}-E(Y|A=0,L;\gamma)$; and replace $Y-E(Y|A,L;\beta^{\dagger},\gamma)=Y-m(A,L;\beta^{\dagger})-E(Y|A=0,L;\gamma)$ on rows $p+q+1$ through $2p+q+1$ in (12) with $Ye^{-m(A,L;\beta^{\dagger})}-E(Y|A=0,L;\gamma)$. Following Robins (2000), we can show that this modified estimating equation system is unbiased if either (4) with the log link or (7) is correct, together with the main model (3) with the log link.

We now consider the case when $g(\cdot)$ is the logit link. For this case, a DR estimator of β can be obtained by solving the estimating equation

$$\widetilde{E}\begin{bmatrix} X\{A-E^*(A|L;\beta,\gamma,\delta)\}\{Y-E(Y|A,L;\beta,\gamma)\} \\ \begin{pmatrix} AX \\ V \end{pmatrix}\{Y-E(Y|A,L;\beta^{\dagger},\gamma)\} \\ \begin{pmatrix} YX \\ W \end{pmatrix}\{A-E(A|Y,L;\beta^{\ddagger},\delta)\} \end{bmatrix}=0 \qquad (13)$$

for $(\beta^T,\beta^{\dagger T},\gamma^T,\beta^{\ddagger T},\delta^T)^T$, where

$$E^*(A|L;\beta,\gamma,\delta)=\left[1+\frac{\{1-E(A|Y=0,L;\delta)\}E(Y|A=0,L;\gamma)}{E(A|Y=0,L;\delta)E(Y|A=1,L;\beta,\gamma)}\right]^{-1}$$

For simplicity, we use $\widehat{\beta}_{\mathrm{DR}}$ to denote the first p elements of the solution to either (12) or (13). It follows from a more general result in Tchetgen Tchetgen and Robins (2010) that the estimating equation in (13) is unbiased if either (4) with the logit link or (10) is correct, together with the main model (3) with the logit link.[2]

1. Here we define $(\beta^{\dagger T},\gamma^T,\alpha^T)^T$ as the asymptotic solution to the last $p+q+r$ rows in (12) whether (4) and (7) are misspecified or not. It follows that the last $p+q+r$ rows in (12) are unbiased by definition.
2. Here we define $(\beta^{\dagger T},\gamma^T,\beta^{\ddagger T},\delta^T)^T$ as the asymptotic solution to the last $p+q+p+s$ rows in (13) whether (4) and (10) are misspecified or not. It follows that the last $p+q+p+s$ rows in (13) are unbiased by definition.

3.4 Standard errors

All estimators of β that we have considered in section 3 are generalized method of moments estimators, also referred to as Z-estimators (van der Vaart 1998). Specifically, they are the first p elements of the solution to an unbiased estimating equation on the form $\widetilde{E}\{U(\theta)\} = 0$, where $\theta = (\beta^T, \eta^T)^T$, and η is a nuisance parameter. It follows from general results on generalized method of moments estimators (Newey and McFadden 1994) that $n^{1/2}(\widehat{\theta} - \theta)$ is asymptotically normal with mean 0 and variance–covariance matrix

$$\Sigma = \left[E\left\{\frac{\partial U(\theta)}{\partial \theta^T}\right\}\right]^{-1} \mathrm{Var}\{U(\theta)\} \left(\left[E\left\{\frac{\partial U(\theta)}{\partial \theta^T}\right\}\right]^{-1}\right)^T \tag{14}$$

A consistent estimator of Σ is obtained by replacing θ in (14) with the estimator $\widehat{\theta}$ and the population moments in (14) with their sample counterparts.

3.5 A note on the possible combinations of link functions

The DR estimators that we have considered in section 3.3 only apply to main models on the parametric form in (3) and to the combination of link functions listed in table 1. In principle, it would be desirable to implement DR estimators that do not suffer from this limitation. In practice, though, such DR estimators typically require stronger modeling assumptions, or they may not even exist. For instance, when the outcome is binary and the exposure is continuous, it would be desirable to have a DR estimator that uses a logit link for the outcome and an identity link for the exposure. However, such an estimator requires not only a mean model for the exposure but also a fully specified model for the exposure distribution (Tchetgen Tchetgen and Robins 2010). This makes the estimator less robust and more computationally intensive. For binary outcomes and exposures, it would also be desirable to implement a DR estimator that uses probit links. However, to the best of our knowledge, no such DR estimator exists.

Table 1. Possible combinations of link functions

main/outcome link	exposure link
identity	identity
identity	log
identity	logit
log	identity
log	log
log	logit
logit	logit

4 The drglm command

`drglm` provides DR estimates for the main model (3) in GLMs.

4.1 Syntax

`drglm` *depvar expvar* [*if*] [*in*] [, `main(`*varlist*`)` `outcome(`*varlist*`)` `exposure(`*varlist*`)` `olink(`*linkname*`)` `elink(`*linkname*`)` `level(`*#*`)` `obe` `ebe` `eform` `vce(`*vcetype*`)`]

The *expvar* (exposure, treatment, predictor, or covariate) must be numerical. After `drglm` estimation, one can use postestimation commands such as `test`, `testparm`, `lincom`, and `predictnl`.

Options

`main(`*varlist*`)` determines which variables are used in the main model part of the estimator. The constant 1 is always added to `main(`*varlist*`)`. Then each variable in `main(`*varlist*`)` is multiplied by *expvar* and saved in the current dataset.

`outcome(`*varlist*`)` determines which variables are used in the outcome model part of the estimator. The constant 1 is always added to `outcome(`*varlist*`)`.

`exposure(`*varlist*`)` determines which variables are used in the exposure model part of the estimator. The constant 1 is always added to `exposure(`*varlist*`)`.

`olink(`*linkname*`)` specifies the link function of the outcome model (`identity`, `logit`, `log`). The default is `olink(identity)`. If `olink(logit)` is specified, *expvar* can take on only two values (either 0 or 1).

`elink(`*linkname*`)` specifies the link function of the exposure model (`identity`, `logit`, `log`). The default is `elink(identity)`.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`.

`obe` specifies the outcome-based estimation.

`ebe` specifies the exposure-based estimation.

`eform` reports coefficient estimates as `exp(b)` rather than as `b`.

`vce(`*vcetype*`)` specifies the type of standard error reported. *vcetype* may be `robust`, `cluster` *clustvar*, `bootstrap`, or `jackknife`. The default is `vce(robust)`.

Saved results

`drglm` saves the following in `e()`:

Scalars			
`e(N)`	number of observations	`e(rank)`	rank of `e(V)`
Macros			
`e(cmd)`	`drglm`	`e(olink)`	link function of the outcome model
`e(cmdline)`	command as typed		
`e(depvar)`	name of dependent variable	`e(elink)`	link function of the exposure model
`e(vcetype)`	title used to label Std. Err.		
`e(properties)`	`b V`	`e(estimator)`	type of estimator (`dr`, `obe`, or `ebe`)
Matrices			
`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators
Functions			
`e(sample)`	marks estimation sample		

5 Simulation study

To demonstrate the doubly robustness of the implemented estimators, we present the results from two simulation studies.

5.1 Simulation 1

We generated 1,000 samples of 500 observations each from the model

$$\left.\begin{array}{rcl}
L & = & (L_1, L_2) \\
L_1 & \perp & L_2 \\
L_1 & \sim & N(0,1) \\
L_2 & \sim & N(0,1) \\
A|L & \sim & N\{E(A|L), 1\} \\
Y|A,L & \sim & N\{E(Y|A,L), 1\} \\
E(A|L) & = & \underbrace{\alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_{12} L_1 L_2}_{\text{Exposure nuisance model}} \\
E(Y|A=0,L) & = & \underbrace{\gamma_0 + \gamma_1 L_1 + \gamma_2 L_2 + \gamma_{12} L_1 L_2}_{\text{Outcome nuisance model}} \\
m(A,L) & = & E(Y|A,L) - E(Y|A=0,L) \\
 & = & \underbrace{\beta_0 A + \beta_1 A L_1}_{\text{Main model}}
\end{array}\right\}$$

with nuisance parameter $\eta = (\alpha_0, \alpha_1, \alpha_2, \alpha_{12}, \gamma_0, \gamma_1, \gamma_2, \gamma_{12}) = (0, 1, 1, -1.5, -1, -1, -1, 1.5)$ and target parameter $\beta = (\beta_0, \beta_1) = (1.5, 1)$. For each sample, we calculated $\widehat{\beta}_{\text{OBE}}$, $\widehat{\beta}_{\text{EBE}}$, and $\widehat{\beta}_{\text{DR}}$ by using correct models for $E(A|L)$, $E(Y|A=0,L)$, and $m(A,L)$. We calculated the mean estimates (over the 1,000 samples), the mean theoretical standard errors (as obtained from the sandwich formula), the empirical standard errors, and the empirical coverage probabilities of the corresponding 95% Wald confidence intervals (CIs). This procedure was repeated twice: we first used correct models for $E(Y|A=0,L)$ and $m(A,L)$ but the incorrect model $E(A|L)=\alpha_0+\alpha_1 L_1+\alpha_2 L_2$; we then used correct models for $E(A|L)$ and $m(A,L)$ but the incorrect model $E(Y|A=0,L)=\gamma_0+\gamma_1 L_1+\gamma_2 L_2$. Table 2 shows the results. All three estimators work well under correct model specifications. The mean estimates are close to the true value of β; the mean theoretical standard errors are close to the mean empirical standard errors; and the coverage probabilities of the CIs are very close to the nominal level of 95%. When the model for $E(A|L)$ is misspecified, $\widehat{\beta}_{\text{EBE}}$ is biased. Similarly, when the model for $E(Y|A=0)$ is misspecified, $\widehat{\beta}_{\text{OBE}}$ is biased. $\widehat{\beta}_{\text{DR}}$ is unbiased even if either of these models is misspecified. The differences in empirical standard error for the three estimators are minor.

Table 2. Simulation results for the estimate of β_0 and β_1. I: Correct models for $E(A|L)$, $E(Y|A=0,L)$, and $E(Y|A,L)-E(Y|A=0,L)$; II: Correct models for $E(Y|A=0,L)$ and $E(Y|A,L)-E(Y|A=0,L)$ and incorrect model for $E(A|L)$; III: Correct models for $E(A|L)$ and $E(Y|A,L)-E(Y|A=0,L)$ and incorrect model for $E(Y|A=0,L)$.

	mean estimate	mean theoretical standard error	empirical standard error	coverage probability
I				
$\widehat{\beta}_{0,\text{OBE}}$	1.50	0.04	0.05	94
$\widehat{\beta}_{1,\text{OBE}}$	1.00	0.02	0.02	93
$\widehat{\beta}_{0,\text{EBE}}$	1.52	0.06	0.06	96
$\widehat{\beta}_{1,\text{EBE}}$	0.99	0.14	0.14	97
$\widehat{\beta}_{0,\text{DR}}$	1.50	0.05	0.05	94
$\widehat{\beta}_{1,\text{DR}}$	1.00	0.05	0.05	95
II				
$\widehat{\beta}_{0,\text{OBE}}$	1.50	0.04	0.05	94
$\widehat{\beta}_{1,\text{OBE}}$	1.00	0.02	0.02	93
$\widehat{\beta}_{0,\text{EBE}}$	0.85	0.05	0.05	0
$\widehat{\beta}_{1,\text{EBE}}$	1.07	0.03	0.03	42
$\widehat{\beta}_{0,\text{DR}}$	1.50	0.04	0.05	94
$\widehat{\beta}_{1,\text{DR}}$	1.00	0.02	0.02	93
III				
$\widehat{\beta}_{0,\text{OBE}}$	0.84	0.04	0.05	0
$\widehat{\beta}_{1,\text{OBE}}$	1.06	0.03	0.03	40
$\widehat{\beta}_{0,\text{EBE}}$	1.52	0.06	0.06	96
$\widehat{\beta}_{1,\text{EBE}}$	0.99	0.14	0.14	97
$\widehat{\beta}_{0,\text{DR}}$	1.51	0.06	0.06	96
$\widehat{\beta}_{1,\text{DR}}$	0.98	0.13	0.13	96

5.2 Simulation 2

We generated 1,000 samples of 500 observations each from the model

$$\left.\begin{array}{rcl}
L & = & (L_1, L_2) \\
L_1 & \perp & L_2 \\
L_1 & \sim & N(0,1) \\
L_2 & \sim & N(0,1) \\
(A, Y) & = & \in (0,1) \\
\text{logit}\{E(A|Y=0,L)\} & = & \underbrace{\alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_{12} L_1 L_2}_{\text{Exposure nuisance model}} \\
\text{logit}\{E(Y|A=0,L)\} & = & \underbrace{\gamma_0 + \gamma_1 L_1 + \gamma_2 L_2 + \gamma_{12} L_1 L_2}_{\text{Outcome nuisance model}} \\
m(A,L) & = & \text{logit}\{E(Y|A,L)\} - \text{logit}\{E(Y|A=0,L)\} \\
 & = & \underbrace{\beta_0 A + \beta_1 A L_1}_{\text{Main model}}
\end{array}\right\}$$

with nuisance parameter $\eta = (\alpha_0, \alpha_1, \alpha_2, \alpha_{12}, \gamma_0, \gamma_1, \gamma_2, \gamma_{12}) = (-1, 1, 1, -1.5, -1, -1, -1, 1.5)$ and target parameter $\beta = (\beta_0, \beta_1) = (1.5, 1)$. For each sample, we calculated $\widehat{\beta}_{\text{OBE}}$, $\widehat{\beta}_{\text{EBE}}$, and $\widehat{\beta}_{\text{DR}}$ by using correct models for both logit$\{E(A|Y=0,L)\}$, logit$\{E(Y|A=0,L)\}$, and $m(A,L)$. We calculated the same summary measures as in simulation 1. This procedure was repeated twice: we first used correct models for logit$\{E(Y|A=0,L)\}$ and $m(A,L)$ but the incorrect model logit$\{E(A|Y=0,L)\} = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$; we then used correct models for logit$\{E(A|Y=0,L)\}$ and $m(A,L)$ but the incorrect model logit$\{E(Y|A=0,L)\} = \gamma_0 + \gamma_1 L_1 + \gamma_2 L_2$. Table 3 shows the results. All three estimators work well under correct model specifications. The mean estimates are close to the true value of β; the mean theoretical standard errors are close to the mean empirical standard errors; and the coverage probabilities of the CIs are very close to the nominal level of 95%. When the model for $E(A|L)$ is misspecified, $\widehat{\beta}_{\text{EBE}}$ is biased. Similarly, when the model for $E(Y|A=0)$ is misspecified, $\widehat{\beta}_{\text{OBE}}$ is biased. $\widehat{\beta}_{\text{DR}}$ is unbiased even if either of these models is misspecified. The differences in empirical standard error for the three estimators are minor.

Table 3. Simulation results for the estimate of β_0. I: Correct models for logit$\{E(A|Y=0,L)\}$, logit$\{E(Y|A=0,L)\}$, and logit $\{E(Y|A,L)\}-$logit $\{E(Y|A=0,L)\}$; II: Correct models for logit$\{E(Y|A=0,L)\}$ and logit $\{E(Y|A,L)\}$ − logit $\{E(Y|A=0,L)\}$ and incorrect model for logit$\{E(A|Y=0,L)\}$; III: Correct models for logit$\{E(A|Y=0,L)\}$ and logit $\{E(Y|A,L)\}$ − logit $\{E(Y|A=0,L)\}$ and incorrect model for logit$\{E(Y|A=0,L)\}$.

	mean estimate	mean theoretical standard error	empirical standard error	coverage probability
I				
$\widehat{\beta}_{0,\mathrm{OBE}}$	1.53	0.27	0.26	96
$\widehat{\beta}_{1,\mathrm{OBE}}$	1.03	0.30	0.28	95
$\widehat{\beta}_{0,\mathrm{EBE}}$	1.53	0.28	0.27	96
$\widehat{\beta}_{1,\mathrm{EBE}}$	1.04	0.35	0.33	96
$\widehat{\beta}_{0,\mathrm{DR}}$	1.54	0.28	0.28	95
$\widehat{\beta}_{1,\mathrm{DR}}$	1.05	0.41	0.39	94
II				
$\widehat{\beta}_{0,\mathrm{OBE}}$	1.53	0.27	0.26	96
$\widehat{\beta}_{1,\mathrm{OBE}}$	1.03	0.30	0.28	95
$\widehat{\beta}_{0,\mathrm{EBE}}$	0.73	0.25	0.25	13
$\widehat{\beta}_{1,\mathrm{EBE}}$	1.51	0.37	0.34	71
$\widehat{\beta}_{0,\mathrm{DR}}$	1.53	0.28	0.27	96
$\widehat{\beta}_{1,\mathrm{DR}}$	1.06	0.41	0.38	96
III				
$\widehat{\beta}_{0,\mathrm{OBE}}$	0.78	0.24	0.25	17
$\widehat{\beta}_{1,\mathrm{OBE}}$	1.28	0.26	0.25	80
$\widehat{\beta}_{0,\mathrm{EBE}}$	1.53	0.28	0.27	96
$\widehat{\beta}_{1,\mathrm{EBE}}$	1.04	0.35	0.33	96
$\widehat{\beta}_{0,\mathrm{DR}}$	1.54	0.28	0.27	96
$\widehat{\beta}_{1,\mathrm{DR}}$	1.06	0.40	0.37	94

6 Example

Sjölander and Vansteelandt (2011) used data from the National Match Cohort (NMC) (Bellocco et al. 2010) to illustrate the use of DR estimators of attributable fractions. We use the same dataset to illustrate the use of the `drglm` command. The NMC was established in 1997, when 300,000 Swedes participated in a national fund-raising event organized by the Swedish Cancer Society. Every participant was asked to fill out a questionnaire that included items on known or suspected risk factors for cardiovascular disease (CVD). Using the Swedish patient registry, the NMC followed participants until 2006, and each CVD event was recorded. Sjölander and Vansteelandt (2011) considered a binary outcome `cvd`, with `cvd` = 1 if a subject developed CVD before end of follow-up, and `cvd` = 0 otherwise. They considered a binary exposure `bmi`, with `bmi` = 0 for those subjects with baseline body mass index (BMI)—body weight in kilograms divided by height squared in meters—between 18.5 and 25 kg/m^2 and `bmi` = 1 for subjects with baseline BMI outside this range. The range $18.5 < \text{BMI} < 25$ kg/m^2 is considered normal weight by the World Health Organization (World Health Organization 1995). Based on self-reported history of physical activity, Sjölander and Vansteelandt (2011) constructed a continuous measure. They controlled for both age at baseline (`age`) and the constructed measure of physical activity (`pa`). The dataset `nmc_sj` of 41,295 individuals is a sample that can be requested from the authors; it can be used only to reproduce the current analysis.

A standard way to assess the association between `bmi` and `cvd`, controlling for `age` and `pa`, is to use the logistic regression model $\text{logit}\{E(\texttt{cvd}|\texttt{bmi},\texttt{age},\texttt{pa})\} = \beta\texttt{bmi} + \gamma_0 + \gamma_1\texttt{age} + \gamma_2\texttt{pa}$. Fitting this model with the `logit` command gives the output below. The option `vce(robust)` is used to allow a comparison of the standard errors with the `drglm` command.

```
. use nmc_sj
(National Match Cohort - SJ version)

. logit cvd bmi age pa, vce(robust) nolog

Logistic regression                               Number of obs   =      41295
                                                  Wald chi2(3)    =    1345.18
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -27190.223                 Pseudo R2       =     0.0253

------------------------------------------------------------------------------
             |               Robust
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .1464322    .044115     3.32   0.001     .0599684    .2328959
         age |   .0173548   .0006421    27.03   0.000     .0160964    .0186133
          pa |  -.1348361   .0067156   -20.08   0.000    -.1479983   -.1216738
       _cons |  -.7620794   .0434613   -17.53   0.000    -.8472621   -.6768968
------------------------------------------------------------------------------
```

If both the main model $\text{logit}\{E(\texttt{cvd}|\texttt{bmi},\texttt{age},\texttt{pa})\} - \text{logit}\{E(\texttt{cvd}|\texttt{bmi}=0,\texttt{age},\texttt{pa})\} = \beta\texttt{bmi}$ and the outcome nuisance model $\text{logit}\{E(\texttt{cvd}|\texttt{bmi}=0,\texttt{age},\texttt{pa})\} = \gamma_0+\gamma_1\texttt{age}+\gamma_2\texttt{pa}$ are correct, then the estimate of β is consistent. An identical analysis is performed by using the `drglm` command with the option `obe` (outcome-based estimator).

```
. drglm cvd bmi, outcome(age pa) olink(logit) elink(logit) obe
Generalized Linear Models                              Number of obs =     41295
Estimator: Outcome Based
Link functions: Outcome[logit]  Exposure[logit]
------------------------------------------------------------------------------
             |               Robust
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
         bmi |   .1464322    .044115     3.32   0.001     .0599684    .2328959
------------------------------------------------------------------------------
```

As argued in section 3.2, a consistent estimate of β can also be obtained through the model $\text{logit}\{E(\texttt{bmi}|\texttt{cvd},\texttt{age},\texttt{pa})\} = \beta\texttt{cvd}+\alpha_0+\alpha_1\texttt{age}+\alpha_2\texttt{pa}$. Fitting this model gives the output below.

```
. logit bmi cvd age pa, vce(robust) nolog
Logistic regression                               Number of obs   =      41295
                                                  Wald chi2(3)    =    2618.37
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -7552.0003                 Pseudo R2       =     0.1837
------------------------------------------------------------------------------
             |               Robust
         bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cvd |   .3012316   .0445681     6.76   0.000     .2138798    .3885834
         age |   .0975369   .0019346    50.42   0.000     .0937451    .1013287
          pa |  -.0545103   .0166798    -3.27   0.001    -.0872021   -.0218186
       _cons |   -8.53162   .1471068   -58.00   0.000    -8.819944   -8.243296
------------------------------------------------------------------------------
```

If both the main model $\text{logit}\{E(\texttt{bmi}|\texttt{cvd},\texttt{age},\texttt{pa})\} - \text{logit}\{E(\texttt{bmi}|\texttt{cvd}=0,\texttt{age},\texttt{pa})\} = \beta\texttt{cvd}$ and the exposure nuisance model $\text{logit}\{E(\texttt{bmi}|\texttt{cvd}=0,\texttt{age},\texttt{pa})\} = \alpha_0+\alpha_1\texttt{age}+\alpha_2\texttt{pa}$ are correct, then the estimate of β is consistent. An identical analysis is performed by using the `drglm` command with the option `ebe` (exposure-based estimator).

```
. drglm cvd bmi, exposure(age pa) olink(logit) elink(logit) ebe
Generalized Linear Models                              Number of obs =     41295
Estimator: Exposure Based
Link functions: Outcome[logit]  Exposure[logit]
------------------------------------------------------------------------------
             |               Robust
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
         bmi |   .3012316   .0445681     6.76   0.000     .2138798    .3885834
------------------------------------------------------------------------------
```

A DR estimate of β that uses both nuisance models is obtained as follows:

```
. drglm cvd bmi, outcome(age pa) exposure(age pa) olink(logit) elink(logit)
Generalized Linear Models                              Number of obs =     41295
Estimator: Double Robust
Link functions: Outcome[logit]  Exposure[logit]
```

cvd	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
main						
bmi	.2991997	.0442286	6.76	0.000	.2125133	.3858861

By not specifying the option `main()`, the main model becomes equal to

$$\text{logit}\left\{E(\texttt{cvd}|\texttt{bmi},\texttt{pa},\texttt{age})\right\} - \text{logit}\left\{E(\texttt{cvd}|\texttt{bmi}=0,\texttt{pa},\texttt{age})\right\} = \beta\texttt{bmi}$$

Interpretation of the regression coefficient is usually done on an exponential scale (odds ratios rather than log odds-ratios). One can use either the `drglm`'s option `eform` or the postestimation command `lincom`. Compared with subjects with $18.5 <$ BMI < 25 kg/m^2, the odds of CVD for subjects with BMI < 18.5 or BMI > 25 were 31% higher (95% CI: $[1.20, 1.43]$).

```
. lincom bmi, eform
 ( 1)  [main]bmi = 0
```

cvd	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.309507	.0579729	6.09	0.000	1.200672	1.428207

We observe that the DR estimate of β is very close to the estimate obtained through the exposure nuisance model (option `ebe`) but less close to the estimate obtained through the outcome nuisance model (option `obe`). This indicates that the exposure nuisance model may be reasonably correct, whereas the outcome nuisance model may suffer from more severe misspecifications.

We refined the nuisance models by taking into account nonlinearities for both `age` and `pa`. We modeled both quantitative covariates by using restricted cubic splines with three knots at fixed percentiles of the distribution.

```
. mkspline pas = pa, nk(3) cubic
. mkspline ages = age, nk(3) cubic
. drglm cvd bmi, outcome(ages1 ages2 pas1 pas2) exposure(ages1 ages2 pas1 pas2)
> olink(logit) elink(logit)
Generalized Linear Models                              Number of obs =    41295
Estimator: Double Robust
Link functions: Outcome[logit]  Exposure[logit]
```

cvd	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf. Interval]	
main						
bmi	.2696505	.0442708	6.09	0.000	.1828814	.3564196

With the refined outcome and exposure nuisance model, we obtained $\widehat{\beta}_{\mathrm{OBE}} = 0.25$ and $\widehat{\beta}_{\mathrm{EBE}} = 0.27$, respectively. Whereas the refinement resulted in a change in $\widehat{\beta}_{\mathrm{OBE}}$ with $(0.15 - 0.25)/0.15 = -67\%$, it only resulted in a change in $\widehat{\beta}_{\mathrm{EBE}}$ with $(0.30 - 0.27)/0.30 = 10\%$. This further indicates that the misspecification in the simple outcome nuisance model was more severe than the misspecification in the simple exposure nuisance model.

We next considered the hypothesis that the association between BMI and CVD may vary with physical activity. Therefore, we specify the main model of the form below by specifying the `main(pa)` option.

$$\operatorname{logit}\{E(\texttt{cvd}|\texttt{bmi},\texttt{pa},\texttt{age})\} - \operatorname{logit}\{E(\texttt{cvd}|\texttt{bmi}=0,\texttt{pa},\texttt{age})\} = \beta_0\texttt{bmi} + \beta_1\texttt{bmipa}$$

```
. drglm cvd bmi, main(pa) outcome(ages1 ages2 pas1 pas2)
> exposure(ages1 ages2 pas1 pas2) olink(logit) elink(logit)
Generalized Linear Models                              Number of obs =    41295
Estimator: Double Robust
Link functions: Outcome[logit]  Exposure[logit]
```

cvd	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf. Interval]	
main						
bmi	.2316385	.1228567	1.89	0.059	-.0091562	.4724332
bmipa	.0106238	.0319545	0.33	0.740	-.0520059	.0732535

The variable `bmipa` is the product of `bmi` and `pa` created internally by the `drglm` command. The coefficient of the interaction term, `bmipa` is not statistically significant ($p = 0.740$). A test for overall no association between BMI on CVD is obtained with the postestimation command `testparm`.

```
. testparm bmi bmipa
 ( 1)  [main]bmi = 0
 ( 2)  [main]bmipa = 0
           chi2(  2) =   37.31
         Prob > chi2 =    0.0000
```

Because of the interaction between BMI and physical activity in the main model, to quantify the association between BMI (1 versus 0) and CVD, we need to consider a specific value for physical activity. The coefficient of BMI depends on physical activity via ($\beta_0 + \beta_1$`pa`). For example, the odds ratios of BMI for the minimal (0), median (4), and maximal (8) physical activity level are calculated as follows:

```
. lincom _b[bmi] + _b[bmipa]*0, eform
 ( 1)  [main]bmi = 0
```

cvd	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.260664	.154881	1.89	0.059	.9908856	1.603892

```
. lincom _b[bmi] + _b[bmipa]*4, eform
 ( 1)  [main]bmi + 4*[main]bmipa = 0
```

cvd	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.315391	.0607068	5.94	0.000	1.20163	1.439921

```
. lincom _b[bmi] + _b[bmipa]*8, eform
 ( 1)  [main]bmi + 8*[main]bmipa = 0
```

cvd	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.372493	.2028368	2.14	0.032	1.027339	1.833609

To present graphically how the odds ratio for CVD associated with BMI varies with physical activity (figure 1), we can use the convenient postestimation command `predictnl`.

```
. predictnl logor =  _b[bmi] + _b[bmipa]*pa, ci(lo hi)
note: Confidence intervals calculated using Z critical values
. generate or = exp(logor)
. generate lb = exp(lo)
. generate ub = exp(hi)
. by pa, sort: generate flag = (_n == 1)
. twoway (line or lb ub pa, sort lp(l - -) lc(black black black)) if flag,
> yscale(log) ytitle("Odds Ratio of BMI") xtitle("Physical activity")
> legend(off) scheme(sj) ylabel(1(.2)1.8, angle(horiz) format(%3.2fc))
```

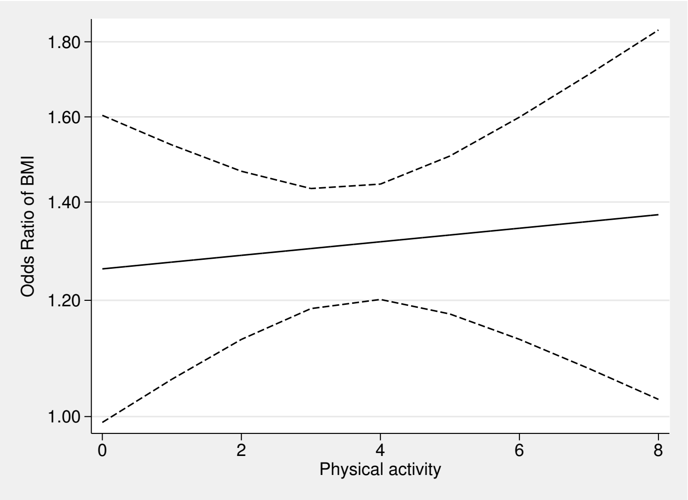

Figure 1. Odds ratio for CVD associated with BMI as function of physical activity

Although the logit link is by far the most common link for binary exposures and outcomes, all combinations listed in table 1 are possible. In table 4, we present $\widehat{\beta}_{\rm OBE}$, $\widehat{\beta}_{\rm EBE}$, and $\widehat{\beta}_{\rm DR}$ together with the corresponding 95% CIs, obtained by using the main model $g\{E(\texttt{cvd}|\texttt{bmi},\texttt{age},\texttt{pa})\} - g\{E(\texttt{cvd}|\texttt{bmi}=0,\texttt{age},\texttt{pa})\} = \beta$, the outcome nuisance model $g\{E(\texttt{cvd}|\texttt{bmi} = 0,\texttt{age},\texttt{pa})\} = \gamma_0 + \gamma_1\texttt{age} + \gamma_2\texttt{pa}$, and the exposure nuisance model $h\{E(\texttt{bmi}|\texttt{age},\texttt{pa})\} = \alpha_0 + \alpha_1\texttt{age} + \alpha_2\texttt{pa}$ for each of the first six link-function combinations in table 1. We remind the reader that the interpretation of β depends on the choice of link function in the main model.

Table 4. Estimated values of $\widehat{\beta}$ using three estimators (outcome based, exposure based, and DR) and various combinations of link functions

main/outcome link	exposure link	$\widehat{\beta}_{\mathrm{OBE}}$	95% CI	$\widehat{\beta}_{\mathrm{EBE}}$	95% CI	$\widehat{\beta}_{\mathrm{DR}}$	95% CI
identity	identity	0.04	[0.02, 0.06]	0.04	[0.02, 0.06]	0.04	[0.02, 0.06]
identity	log	0.04	[0.02, 0.06]	0.08	[0.06, 0.10]	0.08	[0.06, 0.10]
identity	logit	0.04	[0.02, 0.06]	0.07	[0.05, 0.10]	0.07	[0.05, 0.10]
log	identity	0.06	[0.02, 0.11]	0.08	[0.03, 0.12]	0.06	[0.02, 0.10]
log	log	0.06	[0.02, 0.11]	0.16	[0.12, 0.20]	0.16	[0.12, 0.21]
log	logit	0.06	[0.02, 0.11]	0.15	[0.11, 0.19]	0.15	[0.11, 0.20]

Let us consider two alternative DR measures of association with logit as exposure link. When the outcome link is identity, the regression coefficient is a difference in mean outcome.

```
. drglm cvd bmi, outcome(age pa) exposure(age pa) olink(identity) elink(logit)
Generalized Linear Models                              Number of obs =     41295
Estimator: Double Robust
Link functions: Outcome[identity]  Exposure[logit]

                             Robust
         cvd       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

main
         bmi    .0738476   .0109156     6.77   0.000     .0524533    .0952418
```

The CVD risk difference comparing subjects with $18.5 < \text{BMI} < 25$ kg/m^2 versus subjects with BMI < 18.5 or BMI > 25 was 7% (95% CI: [5%, 10%]). If the outcome link instead is log, the regression coefficient is a log risk-ratio.

```
. drglm cvd bmi, outcome(age pa) exposure(age pa) olink(log) elink(logit) eform
Generalized Linear Models                              Number of obs =     41295
Estimator: Double Robust
Link functions: Outcome[log]  Exposure[logit]

                             Robust
         cvd      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]

main
         bmi    1.165457   .0248518     7.18   0.000     1.117752    1.215198
```

Compared with subjects with $18.5 < \text{BMI} < 25$ kg/m^2, the risk of CVD for subjects with BMI < 18.5 or BMI > 25 was 17% higher (95% CI: $[1.12, 1.22]$).

7 Discussion

In this article, we have presented the new Stata command `drglm`, which carries out DR estimation in GLMs. The DR estimators use two regression models and are consistent if either model is correct, not necessarily both. In our simulated scenarios, the DR estimators were almost as efficient as the more "standard" estimators, which used only one regression model. Furthermore, in our simulated scenarios, the estimators that used only one regression model were severely biased whenever the model was incorrect. These results speak in favor of the DR estimators.

The target parameter β is a subpopulation parameter; it quantifies the conditional A-Y association, given covariates L (that is, the association in each subpopulation defined by a distinct level of L). In the special case when $g(\cdot)$ is the identity link or the log link, and there are no interactions between A and L in the main model, β may be interpreted as a population parameter because of the collapsibility of mean differences and log risk-ratios. In the general case (that is, for a link function other than the identity link and the log link and with interactions between A and Y), it is possible to construct DR estimators for population parameters through inverse probability weighting. These methods have been implemented in Stata by Emsley et al. (2008).

In practice, it is unlikely for any model to be exactly correct. Several authors have investigated the performance of DR estimators in various contexts when both working models are misspecified (Bang and Robins 2005; Davidian, Tsiatis, and Leon 2005; Kang and Schafer 2007). These authors have drawn somewhat different conclusions. Bang and Robins (2005) state: "In our opinion, a DR estimator has the following advantage that argues for its routine use: if either the [outcome] model or the [exposure] model is nearly correct, then the bias of a DR estimator ... will be small". In contrast, Kang and Schafer (2007) provided a simulated example where DR estimators were outperformed by estimators that rely on only one regression model; all involved models being moderately misspecified. They concluded that "two wrong models are not necessarily better than one".

8 Acknowledgments

Nicola Orsini was partly supported by a Young Scholar Award from the Karolinska Institutet's Strategic Program in Epidemiology. Arvid Sjölander acknowledges financial support from the Swedish Research Council (2008-5375). Rino Bellocco acknowledges financial support from the Italian Ministry of University and Research (PRIN 2009 X8YCBN).

9 References

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962–973.

Bellocco, R., C. Jia, W. Ye, and Y. T. Lagerros. 2010. Effects of physical activity, body mass index, waist-to-hip ratio and waist circumference on total mortality risk in the Swedish National March Cohort. *European Journal of Epidemiology* 25: 777–788.

Davidian, M., A. A. Tsiatis, and S. Leon. 2005. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science* 20: 261–301.

Emsley, R., M. Lunt, A. Pickles, and G. Dunn. 2008. Implementing double-robust estimators of causal effects. *Stata Journal* 8: 334–353.

Kang, J. D. Y., and J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22: 523–539.

Newey, W. K., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. R. F. Engle and D. L. McFadden, vol. 4, 2111–2245. Amsterdam: Elsevier.

Robins, J. M. 2000. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, 6–10. Alexandria, VA: American Statistical Association.

Robins, J. M., S. D. Mark, and W. K. Newey. 1992. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 48: 479–495.

Sjölander, A., and S. Vansteelandt. 2011. Doubly robust estimation of attributable fractions. *Biostatistics* 12: 112–121.

Tchetgen Tchetgen, E. J., and J. M. Robins. 2010. On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* 97: 171–180.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge University Press: Cambridge.

World Health Organization. 1995. Physical status: The use and interpretation of anthropometry. Report of a WHO Expert Committee. Technical Report Series 854, Geneva: World Health Organization.

About the authors

Nicola Orsini is an associate professor of medical statistics and an assistant professor of epidemiology in the Unit of Biostatistics and Unit of Nutritional Epidemiology at the Institute of Environmental Medicine, Karolinska Institutet, Sweden.

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

Arvid Sjölander is a postdoc at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden.