b
UNIVERSITÄT
BERN

Faculty of Business, Economics and
Social Sciences

Department of Social Sciences

University of Bern Social Sciences Working Paper No. 44

Estimation of marginal odds ratios

Ben Jann and Kristian Bernt Karlson

January 6, 2023

http://ideas.repec.org/p/bss/wpaper/44.html
http://econpapers.repec.org/paper/bsswpaper/44.htm

University of Bern
Department of Social Sciences
Fabrikstrasse 8
CH-3012 Bern

Tel. +41 (0)31 684 48 11
Fax +41 (0)31 684 48 17
info@sowi.unibe.ch
www.sowi.unibe.ch

Estimation of marginal odds ratios

Ben Jann
Institute of Sociology
University of Bern
ben.jann@unibe.ch

Kristian Bernt Karlson
Department of Sociology
University of Copenhagen
kbk@soc.ku.dk

January 6, 2023

Abstract. Coefficients from logistic regression are affected by noncollapsibility, which means that the comparison of coefficients across models may be misleading. Several strategies have been proposed in the literature to respond to these difficulties, the most popular of which is to report average marginal effects (on the probability scale) rather than odds ratios. Average marginal effects (AMEs) have many desirable properties but at least in part they throw the baby out with the bathwater. The size of an AME strongly depends on the marginal distribution of the dependent variable; for events that are very likely or very unlikely the AME necessarily has to be small because the probability space is bounded. Logistic regression, in contrast, estimates odds ratios which are free from such flooring and ceiling effects. Hence, odds ratios may be more appropriate than AMEs for comparison of effect sizes in many applications. Yet, logistic regression estimates conditional odds ratios, which are not comparable across different specifications.

In this paper, we aim to remedy the declining popularity of the odds ratio by introducing an estimand that we term the "marginal odds ratio"; that is, logit coefficients that have properties similar to AMEs, but which retain the odds ratio interpretation. We define the marginal odds ratio theoretically in terms of potential outcomes, both for binary and continuous treatments, we develop estimation methods using three different approaches (G-computation, inverse probability weighting, RIF regression), and we present an example that illustrates the usefulness and interpretation of the marginal odds ratio.

Keywords: Stata, `lnmor`, `ipwlogit`, `riflogit`, marginal odds ratio, noncollapsibility, logistic regression, G-computation, inverse probability weighting, recentered influence functions

Contents

1 Introduction

Logistic response models form the backbone of much applied quantitative research. However, recent methodological literature highlights difficulties in interpreting odds ratios, particularly in a multivariate modeling setting (e.g., Allison 1999; Mood 2010; Karlson et al. 2012; Breen et al. 2018). These difficulties arise from the fact that coefficients from nonlinear probability models such as the logistic response model (i.e., log odds ratios) depend on covariates in ways that differ from the linear model. In short, coefficients from nonlinear probability models are affected by so-called noncollapsibility, which means that conditional coefficients have a different inherent scaling than unconditional (marginal) coefficients, even in the absence of confounding, and hence that coefficients cannot be compared across different model specifications because they correspond to different estimands (e.g., Pang et al. 2016; Daniel et al. 2021; Schuster et al. 2021). Applied researchers have responded to this situation in different ways, but a very popular recommendation is to report average marginal effects on the probability scale implied by the nonlinear probability model or approximated by the linear probability model (Breen et al. 2018; Williams and Jorgensen 2023). Main arguments for using marginal effects are that they are not scaled arbitrarily (Cramer 2007) and that they yield readily interpretable effects on the probability scale, which to many is more intuitive than (log) odds ratios.

Although average marginal effects (AMEs) have many desirable properties, they do not align with research in which relative effects are of interest. This is because the magnitude of an AME depends on the marginal distribution of the dependent variable: the more uneven the distribution, the smaller the AME tends to be. Odds ratios, in contrast, quantify relative effect sizes, such that results can be compared across situations characterized by different baseline probabilities. However, as mentioned above, conventionally used "conditional" odds ratios are affected by noncollapsibility, a property that limits their usefulness for comparative purposes. In this paper, we aim to remedy the declining popularity of odds ratios by introducing *marginal odds ratios*; that is, estimands that are not affected by noncollapsibility and have similar properties as marginal effects on the probability scale, but which retain the odds ratio interpretation.[1]

Drawing on existing literature (Zhang 2008; Daniel et al. 2021) we first define the marginal odds ratio theoretically in terms of potential outcomes and illustrate its relation to logistic regression (Section 2). In contrast to most existing literature, we do not only focus on binary treatments; we also cover continuous predictors. We then discuss different estimation approaches (Section 3) and present corresponding software implementations (Section 4), again covering both categorical as well as continuous predictors, and including consistent variance estimation based on influence functions. We conclude the paper with an example application (Section 5) and some final remarks (Section 6). An appendix provides a brief comparison of the proposed estimators on simulated data (Section 8).

2 Marginal odds ratios

2.1 Definition

Following Zhang (2008) and Daniel et al. (2021), we define marginal odds ratios in terms of potential outcomes (Neyman 1990[1923]; Rubin 1974). Let Y_t be the potential outcome that would realize if treatment T was set to level t by manipulation (i.e., without changing anything else). Comparison of Y_t for different levels of T informs, by definition, about the causal effect of T on Y. In this article we are only interested in binary outcomes $Y_t \in \{0, 1\}$ (e.g. failure and success). $\Pr(Y_t = 1) = E[Y_t]$ is the (marginal) probability that Y_t will be equal to 1 (probability of success).

We first consider the case in which T is *binary*, with $T = 0$ as a standard treatment and $T = 1$ as an alternative treatment. The marginal odds ratio of the alternative treatment versus the standard treatment is defined as

$$\textsc{or} = \frac{\upsilon[\Pr(Y_1 = 1)]}{\upsilon[\Pr(Y_0 = 1)]} = \exp\{\ln \upsilon[\Pr(Y_1 = 1)] - \ln \upsilon[\Pr(Y_0 = 1)]\} \tag{1}$$

where $\upsilon(p) = p/(1-p)$ (odds) and $\ln \upsilon(p) = \ln(p/(1-p))$ (log odds). We may interpret this as the ratio of the odds of success if everyone would receive the alternative treatment versus the odds of success if everyone would receive the standard treatment (provided SUTVA holds).

The probability of success may not only depend on T, but also on other factors $\mathbf{X}$. Assume that $\mathbf{X}$ has a specific distribution in the population and let $\Pr(Y_t = 1|\mathbf{X} =$

1. We use the term "marginal odds ratios" because the quantity of interest refers to how a predictor affects the "marginal" distribution of the outcome. An alternative would be to use the term "unconditional odds ratio", which might lead to less confusion because "marginal effect" is sometimes also understood in the sense of an effect of a marginal change in a predictor. We adopt the term "marginal odds ratio" because it is established in the literature (e.g., Stampf et al. 2010; Karlson et al. 2021). On the difference between average marginal effects and odds ratios, particularly the "flipped-signs phenomenon" related to interaction effects, also see Bloome and Ang (2022). While Bloome and Ang (2022) advise against using odds ratios, their critique pertains to *conditional* odds ratios, not *marginal* odds ratios.

$\mathbf{x}) = E[Y_t|\mathbf{X} = \mathbf{x}]$ be the conditional success probability given $\mathbf{X} = \mathbf{x}$. The law of iterated expectations implies that $\Pr(Y_t = 1) = E_{\mathbf{X}}[\Pr(Y_t = 1|\mathbf{X} = \mathbf{x})]$, where $E_{\mathbf{X}}$ is the expectation over the distribution of $\mathbf{X}$. Equation (1) can thus be rewritten as

$$\begin{aligned} \textsc{or} &= \frac{\upsilon\{E_{\mathbf{X}}[\Pr(Y_1 = 1|\mathbf{X} = \mathbf{x})]\}}{\upsilon\{E_{\mathbf{X}}[\Pr(Y_0 = 1|\mathbf{X} = \mathbf{x})]\}} \\ &= \exp(\ln \upsilon\{E_{\mathbf{X}}[\Pr(Y_1 = 1|\mathbf{X} = \mathbf{x})]\} - \ln \upsilon\{E_{\mathbf{X}}[\Pr(Y_0 = 1|\mathbf{X} = \mathbf{x})]\}) \end{aligned} \tag{2}$$

We call (2) the "adjusted marginal odds ratio", although by definition it is identical to the (unadjusted) marginal odds ratio given in (1). The usefulness of (2) will become evident once we estimate the marginal OR from data. Most importantly, estimation based on formulation (2) can be used to address confounding bias in observational data.

Now consider the case in which treatment T is *continuous*. For such a treatment, the marginal (log) odds ratio can be defined as the derivative of the marginal log odds by the treatment, that is

$$\begin{aligned} \ln \textsc{or}(t) &= \lim_{\epsilon \to 0} \frac{\ln \upsilon[\Pr(Y_{t+\epsilon} = 1)] - \ln \upsilon[\Pr(Y_t = 1)]}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\ln \upsilon\{E_{\mathbf{X}}[\Pr(Y_{t+\epsilon} = 1|\mathbf{X} = \mathbf{x})]\} - \ln \upsilon\{E_{\mathbf{X}}[\Pr(Y_t = 1|\mathbf{X} = \mathbf{x})]\}}{\epsilon} \end{aligned} \tag{3}$$

Likewise, we could define the marginal (log) odds ratio as the difference in marginal log odds induced by a discrete change in the treatment, say, an increase by one unit (unit change effect).

In any case, it is evident that the marginal OR for a continuous predictor is a function of t. That is, results will, in general, depend on the level of t at which we evaluate the marginal OR. We may thus want to apply some kind of averaging. Assume that T has a specific distribution in the population. To obtain an "overall" or "average" marginal OR we can either evaluate the OR at the population average of T, that is,

$$\textsc{or}^* = \textsc{or}(t = E[T]) \tag{4}$$

or integrate $\textsc{or}(t)$ over the distribution of T, that is

$$\overline{\textsc{or}} = \exp\{E_T[\ln \textsc{or}(t)]\} \tag{5}$$

Yet another possibility is to integrate over T (or the joint distribution of T and X) when obtaining the population-averaged probabilities on which the marginal OR is based, that is,

$$\begin{aligned} \ln \textsc{or}' &= \lim_{\epsilon \to 0} \frac{\ln \upsilon\{E_T[\Pr(Y_{t+\epsilon} = 1)]\} - \ln \upsilon\{E_T[\Pr(Y_t = 1)]\}}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\ln \upsilon\{E_{T,\mathbf{X}}[\Pr(Y_{t+\epsilon}|\mathbf{X} = \mathbf{x})]\} - \ln \upsilon\{E_{T,\mathbf{X}}[\Pr(Y_t|\mathbf{X} = \mathbf{x})]\}}{\epsilon} \end{aligned} \tag{6}$$

Results from equations (4), (5), and (6) will generally be different. Equation (4) quantifies the marginal OR at average treatment; (5) is the average marginal OR over the treatment distribution; (6) corresponds to the marginal OR that is obtained if treatment is slightly increased for each population member, given each member's existing values for T and $\mathbf{X}$.

2.2 Relation to the logistic model

Assume that $\Pr(Y_t = 1)$ comes about through a logistic model defined as

$$\Pr(Y_t = 1) = \text{logit}(\alpha + \delta t) \quad \text{where} \quad \text{logit}(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{7}$$

which implies

$$\ln \upsilon\{\Pr(Y_t = 1)\} = \alpha + \delta t$$

It is easy to see that in this case the (exponent of) slope parameter δ has a marginal OR interpretation. If T is binary, we get

$$\text{OR} = \exp\{(\alpha + \delta) - (\alpha)\} = \exp(\delta)$$

Likewise, if T is continuous, we get

$$\ln \text{OR}(t) = \lim_{\epsilon \to 0} \frac{\{\alpha + \delta(t + \epsilon)\} - (\alpha + \delta t)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\delta \epsilon}{\epsilon} = \delta$$

Note that $\text{OR}^* = \overline{\text{OR}} = \text{OR}' = \text{OR}(t) = \exp(\delta)$ in case of the simple logistic model, because $\text{OR}(t)$ is constant.

Now assume a more complicated data-generating process that also involves covariates $\mathbf{X}$. The model is given as

$$\Pr(Y_t = 1 | \mathbf{X} = \mathbf{x}) = \text{logit}(\alpha + \delta t + \mathbf{x}\beta) \tag{8}$$

which implies

$$\ln \upsilon\{\Pr(Y_t = 1 | \mathbf{X} = \mathbf{x})\} = \alpha + \delta t + \mathbf{x}\beta$$

(where $\mathbf{x}$ is a row vector and β is a column vector). In this case, $\exp(\delta)$ describes the *conditional* odds ratio, that is, the odds ratio given a specific value of $\mathbf{X}$. A property of the model is that the conditional odds ratio is constant (i.e., does not depend on $\mathbf{X}$). For example, if T is binary, we have

$$\text{OR}_{\mathbf{X}} = \frac{\upsilon[\Pr(Y_1 = 1 | \mathbf{X} = \mathbf{x})]}{\upsilon[\Pr(Y_0 = 1 | \mathbf{X} = \mathbf{x})]} = \exp\{(\alpha + \delta + \mathbf{x}\beta) - (\alpha + \mathbf{x}\beta)\} = \exp(\delta)$$

The *marginal* odds ratio has a more complicated form. For binary T it is given as

$$\text{OR} = \exp(\ln \upsilon\{E_{\mathbf{X}}[\text{logit}(\alpha + \delta + \mathbf{x}\beta)]\} - \ln \upsilon\{E_{\mathbf{X}}[\text{logit}(\alpha + \mathbf{x}\beta)]\})$$

which is different from the conditional odds ratio whenever $\beta \neq \mathbf{0}$. This is what is meant by "noncollapsibility" of nonlinear models. "Noncollapsibility of the OR derives from the fact that when the expected probability of outcome is modeled as a nonlinear function of the exposure, the marginal effect cannot be expressed as a weighted average of the conditional effects" (Pang et al. 2016, 1926). As a result of the sigmoid functional form of the logistic model, the marginal OR will be attenuated compared to the conditional OR. Likewise, the conditional OR of a model with fewer covariates will be attenuated compared to the conditional OR of a model with more covariates, and these differences in scaling mean that comparing coefficients from logistic regressions across models is problematic.

Stated differently, the models correspond to different estimands: the marginal OR is conceptually different from the conditional OR, and the conditional OR given $\mathbf{X}_1$ is conceptually different from the conditional OR given $\mathbf{X}_2$, when $\mathbf{X}_1$ and $\mathbf{X}_2$ are two different sets of covariates. In the words of Pang et al. (2016, 1926): "In the absence of confounding or when confounding is adjusted appropriately, both the marginal OR and conditional OR are valid measures. They are unbiased estimators for two different parameters, and the choice of reporting the marginal or conditional OR should depend on the research question. One should report the marginal OR if the average effect at the population level is of interest, while one should report the conditional OR if the conditional effect at the individual or subgroup level is of interest."

3 Estimation

For the following discussion assume that we have data from a random sample of size n, including a binary dependent variable Y, a treatment T, k covariates $\mathbf{X}$, as well as sampling weights w. That is, the data is given as $(y_i, t_i, x_{1i}, \dots, x_{ki}, w_i)$, $i = 1, \dots, n$ (in a simple random sample, $w_i = 1$ for all i). To keep notation concise, we typically assume that one element in $\mathbf{X}$ is a constant (so that models can be written without intercept). Treatment variable T can be categorical or continuous.

As detailed above, we are interested in estimating the marginal OR, that is, how a change in T affects the unconditional odds of $Y = 1$. Controlling for $\mathbf{X}$ should not change this goal (that is, the estimand does not change). The data could come from a randomized experiment in which treatment status is independent from potential outcomes, such that consistent estimation of the treatment effect is possible by simple analysis of Y by T, ignoring $\mathbf{X}$. In such a case, adjusting for $\mathbf{X}$ is not necessary for unbiased results, but it can make the estimation more efficient (smaller standard error). The data may also originate from an observational study with nonrandom selection into treatment, such that naïve estimation is biased. In this case we can try to remove confounding bias by adjusting for covariates $\mathbf{X}$, which will be successful if the conditional independence assumption holds (i.e., if treatment assignment is independent from potential outcomes given $\mathbf{X}$). Hence, methods for covariate-adjusted estimation of the marginal OR can be useful both in experimental data and in observational data.

Below we present three different estimation strategies. The first strategy, G-computation, is based on model predictions and closely mirrors the theoretical formulas above. The second approach is based on inverse probability weighting (IPW). The third approach employs RIF regression.[2] Note that we will discuss the estimation of the log

2. A fourth approach may be to regress T on X using linear regression, compute the residuals, and then regress Y on the residuals using logistic regression. Yet another possibility would be to construct a doubly-robust estimator by combining IPW and G-computation. We leave it to future research to explore these additional estimation strategies.

odds ratio (rather than the odds ratio), because normality is more likely to hold for the log odds ratio.[3]

3.1 G-computation

Binary treatment

First consider the case in which the treatment is binary, that is, $T \in \{0, 1\}$. As outlined by Zhang (2008), the marginal odds ratio can be estimated by comparing counterfactual predictions from an outcome model fit to the data (also see Daniel et al. 2021 or Section 2.1 in Stampf et al. 2010). In some literature this approach is called "G-computation," a term coined by Robins (1986; also see, e.g., Snowden et al. 2011 or Chatton et al. 2020). In our case, the procedure is to first regress Y on T and $\mathbf{X}$, for example, using logistic regression, and then use the model fit to generate two predictions of $\Pr(Y = 1)$ for each observation, one with T set to 0 and one with T set to 1 (and X as observed). The two sets of predictions are then averaged across the sample to obtain counterfactual estimates of the population-averaged success probability for the two treatment levels. These estimates can then be plugged into the formula for the marginal OR. That is, the marginal (log) OR is estimated as

$$\ln \widehat{\mathrm{OR}} = \ln \upsilon(\overline{p}^1) - \ln \upsilon(\overline{p}^0) \tag{9}$$

with $\ln \upsilon(p) = \ln(p/(1-p))$, where

$$\overline{p}^t = \frac{1}{W} \sum_{i=1}^{n} w_i \hat{p}_i^t \quad \text{and} \quad W = \sum_{i=1}^{n} w_i \tag{10}$$

The counterfactual predictions $\hat{p}_i^t$ are obtained as follows. Assume a parametric outcome model defined as

$$\Pr(Y = 1|T = t, \mathbf{X} = \mathbf{x}) = g\{\mathbf{z}(t, \mathbf{x})\theta\} \tag{11}$$

where $g(z)$ is a nonlinear transformation, $\mathbf{z}(t, \mathbf{x})$ is a (row) vector composed of t and elements from $\mathbf{x}$ (typically including a constant), and θ is a (column) vector of parameters to be estimated. For example, in case of logistic regression, $g(z) = \mathrm{logit}(z) = e^z/(1+e^z)$; in case of probit, $g(z) = \Phi(z)$, where Φ is the standard normal distribution function.[4] In the simplest case, $\mathbf{z}$ is defined as $\mathbf{z}(t, \mathbf{x}) = (t, x_1, \ldots, x_k, 1)$. However, $\mathbf{z}$ could, for example, only include a selection of elements form $\mathbf{x}$, or it could include products between elements, possibly including t, to model interactions.

3. The standard error of the odds ratio can easily be computed from the standard error of the log odds ratio by the delta method, in particular, $\mathrm{SE}(\widehat{\mathrm{OR}}) = \widehat{\mathrm{OR}} \times \mathrm{SE}(\ln \widehat{\mathrm{OR}})$. Confidence intervals can be obtained by endpoint transformation, that is: first compute the confidence limits for $\ln \mathrm{OR}$ and then take the exponent of these limits.
4. The model for generating predictions does not necessarily have to be a logit or probit model. Any model appropriate for a binary dependent variable will do. Some of the expressions below depend on the specific model, but the general approach remains the same. We could also use a flexible nonparametric model, although derivation of standard errors then would be more challenging.

Given parameter estimate $\hat{\theta}$, predictions from the outcome model are obtained as $\hat{p}_i = g\{\mathbf{z}(t_i, \mathbf{x}_i)\hat{\theta}\}$. For the counterfactual predictions $\hat{p}_i^0$ and $\hat{p}_i^1$ we replace t_i by 0 or 1, respectively. That is, the counterfactual predictions are obtained as

$$\hat{p}_i^0 = g\{\mathbf{z}(0, \mathbf{x}_i)\hat{\theta}\} \quad \text{and} \quad \hat{p}_i^1 = g\{\mathbf{z}(1, \mathbf{x}_i)\hat{\theta}\} \tag{12}$$

Standard errors

We make use of influence functions to derive the standard errors of marginal ORs. Once the influence function of a statistic is known, the standard error of the statistic can be obtained by taking a mean estimate of the influence function (or a total estimate, depending on the scaling of the influence function); the standard error of this mean (or total) provides an estimate of the standard error of the statistic. One of the advantages of this approach is that the form of the influence function does not depend on the survey design, and aspects such as clustering or stratification can easily be taken into account when estimating the mean (or total), using textbook formulas. Furthermore, in many cases, influence functions are fairly easy to derive even for complex statistics because they can be pieced together recursively from the influence functions of the single components that are part of the statistic; see Jann (2020) for an extensive treatment. In terms of resulting estimates, the influence function approach is equivalent to what is known as linearization in survey estimation.

For the marginal OR of a binary treatment, we can derive the influence function in the following three steps.

1. At the uppermost level, estimator (9) is defined as a function of two estimates, $\overline{p}^0$ and $\overline{p}^1$. Taking the derivatives of (9) by $\overline{p}^0$ and $\overline{p}^1$ leads to the following expression for the influence function:

$$\lambda_i(\ln \widehat{\mathrm{OR}}) = \frac{\lambda_i(\overline{p}^1)}{\overline{p}^1(1-\overline{p}^1)} - \frac{\lambda_i(\overline{p}^0)}{\overline{p}^0(1-\overline{p}^0)} \tag{13}$$

2. Expression (13) depends on the influence functions for $\overline{p}^0$ and $\overline{p}^1$. Both will have the same general form. Probability $\overline{p}^t$ is defined as

$$\overline{p}^t = \frac{1}{W}\sum_{i=1}^{n} w_i \hat{p}_i^t \quad \text{with} \quad \hat{p}_i^t = g(\hat{z}_i^t) \quad \text{and} \quad \hat{z}_i^t = \mathbf{z}_i^t\hat{\theta} \quad \text{and} \quad \mathbf{z}_i^t = \mathbf{z}(t, \mathbf{x}_i) \tag{14}$$

   and is thus a function of $\hat{\theta}$. Working through the equations leads to the following expression:

$$\lambda_i(\overline{p}^t) = (\hat{p}_i^t - \overline{p}^t) + \left[\frac{1}{W}\sum_{j=1}^{n} w_j \frac{\partial g(\hat{z}_j^t)}{\partial \hat{z}_j^t}\mathbf{z}_j^t\right]\lambda_i(\hat{\theta}) \tag{15}$$

   The derivative within the sum depends on the type of model (i.e., on the definition of g); for example, $\partial g(z)/\partial z = p(1-p)$ with $p = g(z)$ in case of logistic regression and $\partial g(z)/\partial z = \phi(z)$ (standard normal density) in case of probit.

3. Expression (15) depends on the influence function for $\hat{\theta}$, which will be specific to the type of outcome model. For logistic regression, as shown by Jann (2020), the influence function can be written as

$$\lambda_i(\hat{\theta}) = \left[\frac{1}{W} \sum_{j=1}^{n} w_j \mathbf{z}_j' \hat{p}_j (1 - \hat{p}_j) \mathbf{z}_j \right]^{-1} \mathbf{z}_i' (y_i - \hat{p}_i) \tag{16}$$

   For probit, the expression is more complicated, but see Jann (2020) for a simple general approach to obtain influence functions for maximum-likelihood models, including probit.

All necessary components to compute the influence function of the marginal OR are now complete (i.e., plug 16 into 15, and 15 into 13). As indicated, the standard error of the mean of the influence function provides the standard error of the marginal OR. Confidence intervals and tests can then be computed in the usual way.

Categorical treatment

For a categorical treatment with more than two levels, the procedure is analogous, but the treatment needs to be included as a factor variable in the outcome model (i.e., a series of indicator variables for the different levels). For each combination of levels, a marginal OR can then be computed using counterfactual predictions as above. Typically, one of the levels is chosen as the base levels, to which all other levels are compared. One such set of contrast is sufficient to describe the whole system, as the remaining contrasts directly follow as differences between contrasts with respect to the base level.

Continuous treatment

In Section 2, we defined the marginal OR of a continuous treatment as the derivative of the population-averaged success probability at treatment level t. A natural estimate for the marginal OR thus is

$$\ln \widehat{\mathrm{OR}}(t) = \frac{\partial \ln \upsilon(\overline{p}^t)}{\partial t} = \frac{\overline{q}^t}{\overline{p}^t(1 - \overline{p}^t)} \tag{17}$$

with

$$\overline{q}^t = \frac{1}{W} \sum_{i=1}^{n} w_i \hat{q}_i^t \quad \text{and} \quad \hat{q}_i^t = \frac{\partial \hat{p}_i^t}{\partial t} = \frac{\partial g(\hat{z}_i^t)}{\partial \hat{z}_i^t} \frac{\partial \hat{z}_i^t}{\partial t} \tag{18}$$

and with $\overline{p}^t$ as in (14). The influence function of (17) can be obtained as

$$\lambda_i\{\ln \widehat{\mathrm{OR}}(t)\} = \frac{1}{\overline{p}^t(1 - \overline{p}^t)} \left[\lambda_i(\overline{q}^t) + \frac{\overline{q}^t(2\overline{p}^t - 1)}{\overline{p}^t(1 - \overline{p}^t)} \lambda_i(\overline{p}^t) \right] \tag{19}$$

with $\lambda_i(\overline{p}^t)$ as in (15) and where

$$\lambda_i(\overline{q}^t) = (\hat{q}_i^t - \overline{q}^t) + \left[\frac{1}{W}\sum_{j=1}^{n} w_j \frac{\partial \hat{q}_j^t}{\partial \hat{\theta}'}\right]\lambda_i(\hat{\theta}) \tag{20}$$

with

$$\frac{\partial \hat{q}_j^t}{\partial \hat{\theta}} = \frac{\partial g(\hat{z}_j^t)}{\partial \hat{z}_j^t}\left\{\hat{u}_j^t \frac{\partial \hat{z}_j^t}{\partial t}\mathbf{z}_j^t + \frac{\partial^2 \hat{z}_j^t}{\partial \hat{\theta}' \partial t}\right\} \tag{21}$$

For logistic regression, $\hat{u}_j^t = 1 - 2\hat{p}_j^t$; for probit, $\hat{u}_j^t = -\hat{z}_j^t$.[5]

In general, $\text{OR}(t)$ will not be constant across t. We may thus want to report an overall measure such as OR^* (equation 4), $\overline{\text{OR}}$ (equation 5), or OR' (equation 6):

- To estimate OR^* (marginal odds ratio at the mean), simply evaluate (17) with t set to $\hat{\mu}_T$, the mean of T. That is,

  $$\ln \widehat{\text{OR}}^* = \ln \widehat{\text{OR}}(t = \hat{\mu}_T) \quad \text{with} \quad \hat{\mu}_T = \frac{1}{W}\sum_{i=1}^{n} w_i t_i \tag{22}$$

  The influence function of (22) can be obtained as described above, including a correction for the fact that $\hat{\mu}_T$ is an estimate. In particular, addend

  $$\left[\frac{1}{W}\sum_{j=1}^{n} w_j \hat{q}_j^t\right]\lambda_i(\hat{\mu}_T) \tag{23}$$

  needs to be added to the influence function of $\overline{p}^t$, and addend

  $$\left[\frac{1}{W}\sum_{j=1}^{n} w_j \frac{\partial \hat{q}_j^t}{\partial t}\right]\lambda_i(\hat{\mu}_T) \quad \text{with} \quad \frac{\partial \hat{q}_j^t}{\partial t} = \frac{\partial g(\hat{z}_j^t)}{\partial \hat{z}_j^t}\left\{\hat{u}_j^t\left(\frac{\partial \hat{z}_j^t}{\partial t}\right)^2 + \frac{\partial^2 \hat{z}_j^t}{\partial t^2}\right\} \tag{24}$$

  needs to be added to the influence function of $\overline{q}^t$, where $\lambda_i(\hat{\mu}_T) = t_i - \hat{\mu}_T$.

- To estimate $\overline{\text{OR}}$, take the average of the marginal odds ratio across the distribution of T, that is,

  $$\ln \widehat{\overline{\text{OR}}} = \frac{1}{W}\sum_{i=1}^{n} w_i \ln \widehat{\text{OR}}(t = t_i) \tag{25}$$

  Evaluation of (25) can be computationally burdensome in large datasets. If there are ties, speed improvements are achieved by evaluating $\text{OR}(t)$ only at unique

5. The result of $\partial \hat{z}/\partial t$ depends on the definition of $\mathbf{z}$. For example, if $\mathbf{z} = (t, x_1, \ldots, x_k, 1)$, then $\partial \hat{z}/\partial t = \hat{\theta}_1$. Likewise, if $\mathbf{z}$ includes an interaction between t and x_1, that is, $\mathbf{z} = (t, t \times x_1, x_1, \ldots, x_k, 1)$, then $\partial \hat{z}/\partial t = \hat{\theta}_1 + \hat{\theta}_2 x_1$. The second derivative $\partial^2 \hat{z}/\partial \hat{\theta}' \partial t$ is obtained by taking partial derivatives of $\partial \hat{z}/\partial t$ by elements of $\hat{\theta}$. In the two examples this leads to $(1, 0, \ldots)$ and $(1, x_1, 0, \ldots)$, respectively.

treatment levels. Likewise, if the treatment has a large number of unique levels in the data, as one would expect for a truly continuous treatment, an approximation of $\overline{\text{OR}}$ can be obtained by applying (25) based on linearly binned treatment levels. Let κ_ℓ, $\ell = 1, \ldots, L$, be a series of cut points with $\kappa_{\ell-1} < \kappa_\ell$ and $\kappa_0 = -\infty$. We define the treatment levels as

$$\tau_\ell = \frac{1}{W_\ell} \sum_{i=1}^{n} w_i t_i \mathbb{1}(\kappa_{\ell-1} < t_i \leq \kappa_\ell) \quad \text{with} \quad W_\ell = \sum_{i=1}^{n} w_i \mathbb{1}(\kappa_{\ell-1} < t_i \leq \kappa_\ell) \tag{26}$$

where $\mathbb{1}(x)$ is the indicator function (equal to 1 if x is true, 0 else). An estimate of $\overline{\text{OR}}$ is then obtained as

$$\ln \widehat{\overline{\text{OR}}} = \sum_{\ell=1}^{L} \hat{\omega}_\ell \ln \widehat{\text{OR}}(t = \tau_\ell) \quad \text{with} \quad \hat{\omega}_\ell = \widehat{\Pr}(\kappa_{\ell-1} < T \leq \kappa_\ell) = \frac{W_\ell}{W} \tag{27}$$

The influence function of (27) is

$$\lambda_i(\ln \widehat{\overline{\text{OR}}}) = \sum_{\ell=1}^{L} \hat{\omega}_\ell \lambda_i[\ln \widehat{\text{OR}}(t = \tau_\ell)] + \ln \widehat{\text{OR}}(t = \tau_\ell) \lambda_i(\hat{\omega}_\ell) \tag{28}$$

with

$$\lambda_i(\hat{\omega}_\ell) = \mathbb{1}(\kappa_{\ell-1} < t_i \leq \kappa_\ell) - \hat{\omega}_\ell \tag{29}$$

In case of linear binning, a regular grid is used for κ_ℓ such that the created intervals span the observed range of T (using half intervals at the bottom and top). Treatment level τ_ℓ is then the average of T within the relevant interval. As long as the size of the grid (i.e., L, the number of levels) is not too small, (27) should provide a fairly accurate approximation of (25). If no binning is applied, the cut points (and hence the treatment levels) are set to the observed levels of T, and (27) is exact.

- Finally, for OR$'$, use the same equations as for OR(t), but replace t by the observed treatment value of the relevant observation in all expressions.

Discrete change effects

Rather than obtaining the marginal OR of a continuous treatment in terms of derivatives, we can also compute discrete change effects, of which the approach discussed above for a binary treatment is a special case. In general, discrete change marginal ORs are computed by comparing averaged counterfactual predictions for two treatment levels. In particular, define

$$\ln \widehat{\text{OR}}(t) = \frac{1}{h} \left\{ \ln \upsilon\left(\overline{p}^{\text{up}(t)}\right) - \ln \upsilon\left(\overline{p}^{\text{lo}(t)}\right) \right\} \tag{30}$$

where lo(t) and up(t) are the two treatment levels. For non-centered discrete change effects, lo$(t) = t$ and up$(t) = t + e$, where $e > 0$ is the size of the discrete change. For centered discrete change effects, $\mathrm{lo}(t) = t - e/2$ and $\mathrm{up}(t) = t + e/2$. Furthermore, h is a normalizing constant either set to 1 (no normalization) or set to e. If normalization is applied, (30) converges (17) as e approaches zero. The influence function of (30) is analogous to the influence function of (9), divided by h if relevant. Discrete-change variants of $\textsc{or}^*$, $\overline{\textsc{or}}$, and $\textsc{or}'$ follow in a similar way as discussed above.

Unified approach using fractional logit

A unified approach to obtain marginal ORs for categorical and continuous predictors is to apply fractional logit to counterfactual predictions across treatment levels. Let τ_ℓ, $\ell = 1, \dots, L$, be the unique (possibly binned) treatment levels. Furthermore, let τ_{ℓ_i} denote the treatment level realized for observation i. Then obtain

$$\overline{p}^{\tau_{\ell_i}} = \frac{1}{W} \sum_{j=1}^{n} w_j \hat{p}_j^{\tau_{\ell_i}} \quad \text{with} \quad \hat{p}_j^{\tau_{\ell_i}} = g\{\mathbf{z}(\tau_{\ell_i}, \mathbf{x}_j)\hat{\theta}\} \tag{31}$$

for all i and regress these averaged predictions on the treatment using a fractional logit model. Define $\mathbf{t}_i$ as the treatment covariate vector to be included in the model; typically $\mathbf{t}_i = (\tau_{\ell_i}, 1)$. The log-likelihood of the fractional logit can then be written as

$$\ln L = \frac{1}{n} \sum_{i=1} w_i \{\overline{p}^{\tau_{\ell_i}} \ln(\pi_i) + (1 - \overline{p}^{\tau_{\ell_i}}) \ln(1 - \pi_i)\} \quad \text{with} \quad \pi_i = \mathrm{logit}(\mathbf{t}_i \delta) \tag{32}$$

Taking the derivative of the log-likelihood by δ we obtain the model's main moment condition as

$$\mathbf{h}_i(\delta) = \mathbf{t}_i'(\overline{p}^{\tau_{\ell_i}} - \pi_i) \tag{33}$$

based on which the influence function of $\hat{\delta}$ can be derived as

$$\lambda_i(\hat{\delta}) = G^{-1} \left\{ \mathbf{t}_i'(\overline{p}^{\tau_{\ell_i}} - \hat{\pi}_i) + \sum_{\ell=1}^{L} \left[\frac{1}{W} \sum_{j=1}^{n} w_j \mathbf{t}_j' \mathbb{1}(\ell = \ell_j) \right] \lambda_i(\overline{p}^{\tau_\ell}) \right\} \tag{34}$$

with

$$G = \frac{1}{W} \sum_{i=1}^{n} w_i \mathbf{t}_i' \overline{p}^{\tau_{\ell_i}} (1 - \overline{p}^{\tau_{\ell_i}}) \mathbf{t}_i \tag{35}$$

and $\lambda_i(\overline{p}^{\tau_\ell})$ as in (15).

For categorical treatments, results from (32) will be identical to the results from the more direct estimation approach described earlier (this also holds for the standard errors). For continuous predictors, fractional logit provides an approximation of $\ln \overline{\textsc{or}}$, the average marginal OR across the treatment distribution.[6]

6. The difference between fractional logit and explicit averaging as in (27) should be negligible in most cases. Technically, fractional logit assumes effect homogeneity across treatment levels and thus employs a slightly different implicit weighting of levels.

Interaction effects

Vector $\mathbf{z}$ may include interactions between treatment T and the covariates. The formulas above will take account of such terms, but they will not be informative about the interaction effects per se. To explore interaction effects we can estimate the marginal OR while keeping selected covariates at fixed values. For example, assume a model with $\mathbf{z}(t, \mathbf{x}) = (t, t \times x_1, x_1, \ldots, x_k, 1)$ where both T and X_1 are binary. We could then apply the above formulas with x_1 in $\mathbf{z}$ set to 0 or 1, respectively, to obtain the OR by level of X_1 (still using the full sample in all computations). Comparing these results will illustrate how the marginal OR of T depends on X_1. Similar exercises are possible if T and X_1 are continuous.

Subpopulation effects

All estimates of marginal ORs discussed so far are obtained by averaging over the whole sample. They thus quantify an odds-ratio equivalent to an Average Treatment Effect (ATE). In case of a binary treatment, to estimate an odds-ratio equivalent of an Average Treatment Effect on the Treated (ATET), one could only include the treated when generating counterfactual predictions and taking averages. Subpopulations across which to evaluate the marginal OR could also be defined in other ways. In other words, the outcome model may cover the whole sample, but the implied marginal OR may only be evaluated across a specific subsample. Typically, such an exercise makes most sense if the outcome model is flexible enough to capture subpopulation-specific structures (e.g. through interaction terms). Furthermore, note that there is a fundamental difference between such subpopulation-restricted estimates and estimates that are obtained by fixing covariates at specific values (as in the preceding section on interaction effects). The estimand of the former is at the level of the subpopulation, the estimand of the later is at the level of the population. This means that the former is conditional on the subpopulation-specific distribution of treatment and covariates, while the later is based on the overall distribution. Naturally, the two procedures can also be combined; for example; we may explore interactions within a subpopulation by fixing selected covariates at specific values while restricting evaluation to the selected subpopulation.

3.2 Inverse probability weighting

The basic idea of inverse probability weighting (IPW) is that covariate distributions between treatment levels can be balanced by reweighting observations by the inverse probability of treatment.[7] In this way, for each treatment level, a situation is created in which the corresponding subsample's covariate distribution approximates the covariate distribution in the overall population. We can then apply a simple logistic regression of Y on T to recover the marginal OR, because in the reweighted data the treatment is independent from the covariates. In practice, the difficulty is that the treatment probabilities are not known and need to be estimated from the data.

7. See, e.g., Stampf et al. (2010), who also discuss some other propensity-score based approaches.

Binary treatment

First consider the case of a binary treatment $T \in \{0, 1\}$. We can, for example, fit a logit model defined as

$$\Pr(T = 1|\mathbf{X} = \mathbf{x}) = \operatorname{logit}(\mathbf{x}\gamma) \tag{36}$$

to the observed data (taking account of sampling weights) and obtain observation-specific propensity scores using model predictions, that is

$$\hat{q}_i = \widehat{\Pr}(T = t_i|\mathbf{X} = \mathbf{x}_i) = \begin{cases} \operatorname{logit}(\mathbf{x}_i\hat{\gamma}) & \text{if } t_i = 1 \\ 1 - \operatorname{logit}(\mathbf{x}_i\hat{\gamma}) & \text{if } t_i = 0 \end{cases} \tag{37}$$

We then define inverse-probability weights as

$$\hat{\omega}_i = 1/\hat{q}_i \tag{38}$$

and compute the marginal OR by fitting a simple logistic regression of Y on T, that is,

$$\Pr(Y = 1|T = t) = \operatorname{logit}(\alpha + \delta t) \tag{39}$$

while applying weights $w_i\hat{\omega}_i$. Coefficient $\hat{\delta}$ provides an estimate of the marginal OR.

Stabilized weights

The literature sometimes suggests using "stabilized" weights defined as

$$\hat{\omega}_i^s = \hat{\pi}_i\hat{\omega}_i = \hat{\pi}_i/\hat{q}_i \quad \text{with} \quad \hat{\pi}_i = \widehat{\Pr}(T = t_i) = \frac{1}{W}\sum_{j=1}^{n} w_j \mathbb{1}(t_j = t_i) \tag{40}$$

(e.g. Naimi et al. 2014), but this does not change the resulting estimate in case of a binary treatment (nor its standard error; $\hat{\pi}_i$ is constant within treatment group and thus cancels out). The conceptual difference between $\hat{\omega}_i$ and $\hat{\omega}_i^s$ is that for the former the sum of weights within each group approximates the overall population size; for the latter, the sum of weights within each group approximates the corresponding subpopulation size. $\hat{\omega}_i$ thus mimics a balanced design in which each treatment level has the same overall probability, whereas $\hat{\omega}_i^s$ corresponds to a design in which the treatment distribution is as observed.

Categorical treatment

For an (unordered) categorical treatment with levels τ_ℓ, $\ell = 1, \ldots, L$, we can employ the same approach as outlined above, but use a series of logistic regressions, one for each treatment level against all other treatment levels, to estimate the propensity scores. That is, we fit

$$\Pr(T = \tau_\ell|\mathbf{X} = \mathbf{x}) = \operatorname{logit}(\mathbf{x}\gamma_\ell) \tag{41}$$

for each $\ell = 1, \ldots, L$ and then obtain the propensity scores as

$$\hat{q}_i = \widehat{\Pr}(T = t_i | \mathbf{X} = \mathbf{x}_i) = \begin{cases} \text{logit}(\mathbf{x}_i\hat{\gamma}_1) & \text{if } t_i = \tau_1 \\ \quad\vdots & \\ \text{logit}(\mathbf{x}_i\hat{\gamma}_L) & \text{if } t_i = \tau_L \end{cases} \tag{42}$$

We then fit the outcome model using a logistic regression of Y on a series of indicators for the different treatment levels, omitting one indicator which represents the base level, while applying weights $w_i\hat{\omega}_i = w_i/\hat{q}_i$ or $w_i\hat{\omega}_i^s = w_i\hat{\pi}_i/\hat{q}_i$ (again, the choice of type of weights does not matter for the resulting estimate). The slope coefficients of this model can be interpreted as (log) marginal ORs, comparing each treatment level to the base level.

A series of binary logit models as described above may not represent the structure of the data very well, yielding poor balancing of covariate distributions across treatment levels. An improved approach is to model the propensity score using multinomial logistic regression. That is, estimate a system of equations

$$\Pr(T = \tau_\ell | \mathbf{X} = \mathbf{x}) = \frac{\text{logit}(\mathbf{x}\gamma_\ell)}{\sum_{j=1}^{L} \text{logit}(\mathbf{x}\gamma_j)}, \quad \ell = 1, \ldots, L, \tag{43}$$

where one of the levels, say τ_b, is declared as the base level with its coefficient vector $\hat{\gamma}_b$ set to zero to identify the model. The propensity scores are then obtained as

$$\hat{q}_i = \begin{cases} 1/D_i & \text{if } t_i = \tau_b \\ \exp(\mathbf{x}_i\gamma_\ell)/D_i & \text{if } t_i = \tau_\ell,\ \ell \neq b \end{cases} \quad \text{with} \quad D_i = 1 + \sum_{\ell \neq b} \exp(\mathbf{x}_i\hat{\gamma}_\ell) \tag{44}$$

Note that the base level in the treatment-assignment model does not necessarily have to be the same as the base level in the outcome model. That is, the choice of the base level in the multinomial logit is irrelevant; the results of the outcome model will always be the same.

Ordered treatment

For a categorical treatment whose levels have an ordered interpretation (e.g., low, medium, and high treatment intensity) the procedure is similar as above, but one might want to use a treatment-assignment model that takes account of the qualitative order of the levels. An obvious candidate is standard ordered (i.e. cumulative) logistic regression, that is, to model the treatment assignment as

$$\Pr(T > \tau_\ell | \mathbf{X} = \mathbf{x}) = \text{logit}(\tilde{\mathbf{x}}\gamma - \kappa_\ell), \quad \ell = 1, \ldots, L-1 \tag{45}$$

where $\tilde{\mathbf{x}}$ is a copy of $\mathbf{x}$ without the constant, and compute the propensity scores as

$$\hat{q}_i = \widehat{\Pr}(T = t_i | \mathbf{X} = \mathbf{x}_i) = \hat{c}_i^{\ell_i - 1} - \hat{c}_i^{\ell_i} \tag{46}$$

where ℓ_i is set such that $t_i = \tau_{\ell_i}$ and where

$$\hat{c}_i^\ell = \begin{cases} 1 & \text{if } \ell = 0 \\ 0 & \text{if } \ell = L \\ \text{logit}(\tilde{\mathbf{x}}_i \hat{\gamma} - \hat{\kappa}_\ell) & \text{else} \end{cases} \tag{47}$$

The standard ordered logit model relies on the proportional odds assumption and may be too restrictive to fit the data well. A more flexible approach is to use so-called generalized ordered logistic regression,[8] which relaxes the proportional odds assumption and can be written as a system of equations given as

$$\Pr(T > \tau_\ell | \mathbf{X} = \mathbf{x}) = \text{logit}(\mathbf{x}\gamma_\ell), \quad \ell = 1, \ldots, L-1 \tag{48}$$

The propensity scores are obtained in the same way as for the standard ordered logit, with $\hat{c}_i^\ell = \text{logit}(\mathbf{x}_i \gamma_\ell)$, $1 \leq \ell < L$. Simultaneous estimation of all parameters of the generalized ordered logit can be computationally demanding; an asymptotically equivalent but computationally more efficient procedure is to estimate the parameters by separate logistic regressions, one for each equation.

For the outcome model, instead of using dummy-coding for the treatment levels (which results in ORs with respect to a chosen base level), a coding that leads to ORs between adjacent levels (split-coding) might be preferable. These ORs, however, can also be recovered easily from the results obtained via dummy-coding by taking contrasts.

Continuous treatment

For a continuous treatment, we define $\pi = f(t)$ and $q = f(t|\mathbf{X} = \mathbf{x})$ as the marginal and the conditional density of $T = t$, respectively, and then reweight the data by $w\hat{\omega}$ with $\hat{\omega} = 1/\hat{q}$, or by $w\hat{\omega}^s$ with $\hat{\omega}^s = \hat{\pi}/\hat{q}$, when applying a logit regression of Y on T. We prefer stabilized weights $\hat{\omega}^s$ here because results will generally depend on the choice of the type of weights in case of a continuous treatment. When using stabilized weights each treatment level will receive an overall weight equivalent to its proportion in the population. Use $\hat{\omega}$ instead of $\hat{\omega}^s$ if you are interested in results that reflect a balanced design.

The estimation of $q = f(t|\mathbf{X} = \mathbf{x})$ is challenging. Several procedures have been suggested in the literature (see, e.g., Naimi et al. 2014), but many of them make strong assumptions. Here we focus on a distribution-regression approach (Chernozhukov et al. 2013). The procedure is to first divide the domain of T into a number of (approximate) equal-probability bins using quantiles as cutoffs. Let T^c be such a categorized variant of T. We then run a cumulative odds model of T^c on $\mathbf{X}$ using one of the approaches discussed in the section on ordered treatments above, and recover the propensity scores $\hat{q}_i^c = \widehat{\Pr}(T^c = t_i^c | \mathbf{X} = \mathbf{x}_i)$ from the fitted model. As above, the weights are then defined

8. Similar to the multinomial logit, the generalized ordered logit maintains a full set of slope coefficients for each treatment level. The standard ordered logit only contains level-specific intercepts, and a set of slope coefficients common to all levels. For an overview see Williams (2006).

as $\hat{\omega}_i = 1/\hat{q}_i^c$ or $\hat{\omega}_i^s = \hat{\pi}_i^c/\hat{q}_i^c$ with $\hat{\pi}_i^c = \widehat{\Pr}(T^c = t_i^c)$. Note that categorized treatment T^c is only used for the computation of the weights. In the outcome model, we still simply regress Y on T using logistic regression, while applying the calculated weights.

Results from the distribution-regression approach will depend on the number of bins used to categorize the treatment. If only few bins are created, the treatment assignment model will not be very flexible and the achieved balance may be poor. In contrast, if many bins are used, the variance of the weights may get large and technical difficulties such as crossings in the predicted cumulative probabilities (implying negative propensity scores) may arise. Determining the optimal number of bins is a bias–variance tradeoff; the number of bins should grow with the sample size (to reduce bias), but at a slower rate (to improve efficiency). A simple approach may be to use a crude rule-of-thumb, such as Sturges' rule for the number of histogram bins that sets the number of bins to $L = \lceil \ln(n)/\ln(2) \rceil + 1$, where n is the sample size. More sophisticated approaches will, for example, also take the complexity of the treatment-assignment model into account or use cross-validation to determine the optimal number of bins.

Standard errors

Standard errors can again be estimated using influence functions. For a reweighted statistic, the general procedure is as follows (also see Jann 2021). Let θ be the statistic of interest and let $\tilde{\lambda}_i(\hat{\theta})$ be a preliminary influence function ignoring the fact that the weights $\hat{\omega}_i$ have been estimated. That is, $\tilde{\lambda}_i(\hat{\theta})$ is the influence function that we get if we treat $w_i\hat{\omega}_i$ as fixed sampling weights. In our case θ is estimated by logistic regression and $\tilde{\lambda}_i(\hat{\theta})$ is the influence function for logistic regression as given above (see equation 16). Furthermore, assume that the weights $\hat{\omega}$ have been constructed in a way such that they depend on a set of parameters $\hat{\gamma}$ from a treatment-assignment model (as in the cases discussed above). The final influence function for the reweighted statistic $\hat{\theta}$ can then be obtained as

$$\lambda_i(\hat{\theta}) = \hat{\omega}_i \tilde{\lambda}_i(\hat{\theta}) + \left[\frac{1}{W} \sum_{i=1} w_i \tilde{\lambda}_i(\hat{\theta}) \frac{\partial \hat{\omega}_i}{\partial \hat{\gamma}} \right] \lambda_i(\hat{\gamma}) \tag{49}$$

where $\lambda_i(\hat{\gamma})$ is the influence function of the parameters of the treatment-assignment model. For maximum-likelihood estimators, $\lambda_i(\hat{\gamma})$ can be obtained easily as shown in Jann (2020). Furthermore, Table 1 provides an overview $\partial\hat{\omega}_i/\partial\hat{\gamma}$ for the different models discussed above. For continuous treatments we can use analogous formulas based on the categorized treatment.[9]

3.3 Unconditional logistic regression

A simple approximate approach to estimate marginal odds ratios is to apply linear regression to the recentered influence function (RIF) of the marginal log odds. This is in

9. Some refinements could be applied because the categorization of the treatment relies on estimated quantiles. The effect of these refinements should be negligible.

Table 1: Derivatives for influence functions of IPW estimators

Treatment-assignment model	Derivatives
Binary treatment modeled by logistic regression	$\frac{\partial\hat{\omega}_i}{\partial\hat{\gamma}} = \begin{cases} (\hat{q}_i - 1)\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = 1 \\ (1-\hat{q}_i)\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = 0 \end{cases}$
Categorical treatment modeled by a series of logistic regressions, one for each treatment level	$\frac{\partial\hat{\omega}_i}{\partial\hat{\gamma}_\ell} = \begin{cases} (\hat{q}_i - 1)\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = \tau_\ell \\ \mathbf{0} & \text{else} \end{cases}$
Categorical treatment modeled by multinomial regression	$\frac{\partial\hat{\omega}_i}{\partial\hat{\gamma}_\ell} = \begin{cases} (\hat{q}_i - 1)\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = \tau_\ell \\ \frac{\exp(\mathbf{x}_i\gamma_\ell)}{D_i}\hat{\omega}_i\mathbf{x}_i & \text{else} \end{cases}$
Categorical treatment modeled by ordered logistic regression	$\frac{\partial\hat{\omega}_i}{\partial\hat{\gamma}} = \frac{\hat{c}_i^{\ell_i}(1-\hat{c}_i^{\ell_i}) - \hat{c}_i^{\ell_i-1}(1-\hat{c}_i^{\ell_i-1})}{\hat{q}_i}\hat{\omega}_i\mathbf{x}_i$ $\frac{\partial\hat{\omega}_i}{\partial\hat{\kappa}_\ell} = \begin{cases} -\hat{c}_i^\ell(1-\hat{c}_i^\ell)\frac{1}{\hat{q}_i}\hat{\omega}_i & \text{if } t_i = \tau_\ell \\ \hat{c}_i^\ell(1-\hat{c}_i^\ell)\frac{1}{\hat{q}_i}\hat{\omega}_i & \text{if } t_i = \tau_{\ell+1} \\ 0 & \text{else} \end{cases}$
Categorical treatment modeled by generalized ordered logistic regression (both variants)	$\frac{\partial\hat{\omega}_i}{\partial\hat{\gamma}_\ell} = \begin{cases} \hat{c}_i^\ell(1-\hat{c}_i^\ell)\frac{1}{\hat{q}_i}\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = \tau_\ell \\ -\hat{c}_i^\ell(1-\hat{c}_i^\ell)\frac{1}{\hat{q}_i}\hat{\omega}_i\mathbf{x}_i & \text{if } t_i = \tau_{\ell+1} \\ \mathbf{0} & \text{else} \end{cases}$

line with Firpo et al. (2009), who illustrated the approach for quantile regression (which, like the logit model, suffers from noncollapsibility). In analogy to the "unconditional quantile regression" by Firpo et al. (2009) we call this procedure the "unconditional logistic regression".

Intuitively, an influence function of a statistic (originally called "influence curve" by Hampel 1974) quantifies the degree to which the statistic changes if a small amount of data mass is added at a specific point in the distribution that underlies the statistic. If the distribution depends on covariates, then changing the covariate values will change the distribution, which then will lead to changes in the statistic. As discussed by Firpo et al. (2009), regressing the influence function on covariates can thus be used to approximate the partial effects of the covariates on the statistic. Results will only be approximate in most cases because the influence function is valid in the limit, that is, it provides a linear approximation to how a statistic changes if the underlying distribution is modified. Furthermore, because the influence function is centered around zero (i.e., has an expectation of zero), Firpo et al. (2009) suggest to use the RIF in such regressions. The RIF is a shifted variant of an influence function that is centered around the value of the statistic rather than around zero. This ensures that the intercept of the regression has a meaningful interpretation.

If the considered statistic in a RIF regression is unconditional (i.e. marginal), then also the regression coefficients have an unconditional interpretation. That is, the coefficients reflect the partial effects of the covariates on the unconditional statistic. For our purposes, we thus use the marginal log odds as the target statistic. In this case, the exponents of the regression coefficients can be interpreted as (possibly adjusted) marginal odds ratios (because the exponent of a difference in log odds is equivalent to an odds ratio).

The RIF of the marginal log odds can be derived as follows. Let $\pi = \Pr(Y = 1)$, such that the marginal log odds are given as

$$\alpha = \ln \upsilon(\pi) = \ln(\pi/(1-\pi)) \quad \text{which implies} \quad \pi = \frac{\exp(\alpha)}{1+\exp(\alpha)} \tag{50}$$

Based on moment equation

$$E[h(y;\alpha)] = 0 \quad \text{with} \quad h(y;\alpha) = y - \frac{\exp(\alpha)}{1+\exp(\alpha)} = y - \pi \tag{51}$$

the influence function of α can be derived as

$$\lambda(y;\alpha) = \frac{1}{-E[\partial h/\partial \alpha]} h(y;\alpha) = \frac{y-\pi}{\pi(1-\pi)} \tag{52}$$

(also see Jann 2020). To obtain the (empirical) RIF we replace π by its sample estimate (i.e. the sample mean of Y) and add the sample log odds to the equation, that is

$$\text{RIF}_i = \frac{y_i - \hat{\pi}}{\hat{\pi}(1-\hat{\pi})} + \ln \upsilon(\hat{\pi}) \tag{53}$$

To obtain the marginal odds ratio with respect to treatment T, we then regress the RIF on T using least-squares estimation. To obtain the adjusted marginal odds ratio, we regress RIF on T and covariates X. Robust standard errors from such a regression are consistent and no additional adjustments are needed. This is due to the fact that the moment conditions of least-squares coefficients have the same basic form as the moment condition of the mean and that the formulas behind robust standard errors are equivalent to the formulas one would use when obtaining the standard errors through influence functions.

4 Commands

Below we present three new commands implementing the estimation approaches discussed above. We focus on the main features of the commands and leave the details (e.g., on minor options and stored results) to the online documentation.

4.1 G-computation

G-computation is implemented by command `lnmor`, a post-estimation utility that can be applied after `logit` or `probit` to obtain marginal ORs. `lnmor` is also allowed after logit or probit models to which `svy` or `mi estimate` has been applied. Stata 15 or newer is required.[10] The syntax is

`lnmor` *termlist* [`,` *options*]

where *termlist* is

term [*term* ...]

and *term* may be a simple *varname*, a factor variable specification such as `i.`*varname*, or an interaction specification of a continuous variable with itself, such as `c.`*varname*`##c.`*varname*. Each *term* must refer to a distinct variable and all specified variables must appear among the covariates of the model after which `lnmor` is applied. Options are as follows.

`dx`[`(`*spec*`)`] requests derivative-based results for continuous terms (results for factor-variables and interaction terms will not be affected). By default, `lnmor` reports results obtained by fractional logit ([R] **fracreg**). *spec* may be one of the following.

`average`	report the average derivative across the distribution of the variable; this is the default if `dx` is specified without argument
`atmean`	report the derivative at the mean of the variable
`observed`	report a derivative based on a marginal shift in observed values
numlist	report derivative at each level
`levels`	report derivative at each observed level; not allowed if *termlist* contains multiple terms affected by `dx()`

`delta`[`(`#`)`] requests that `dx()` computes discrete change effects rather than derivatives. `delta` without argument is equivalent to `delta(1)` (unit change effect). `delta()` implies `dx()`.

Discrete change effects are not defined if # is 0. In this case, `lnmor` will report (log) odds rather than (log) odds ratios. That is, you can specify `delta(0)` to obtain levels rather than effects.

`centered` requests that discrete change effects are computed using predictions at $t+\#/2$ and $t-\#/2$ rather than $t+\#$ and t. `centered` is only relevant if `delta()` has been specified.

`normalize` divides discrete change effects by #. `normalize` is only relevant if `delta()` has been specified.

`at(`*spec*`)` reports results with covariates fixed at specific values. The syntax of *spec* is

varname `=` *numlist* [*varname* `=` *numlist* ...]

Computations will be repeated for each pattern of combinations of the specified

10. The `lnmor` command also requires `moremata` to be installed on the system (Jann 2005; type `ssc install moremata`).

covariate values. You can also type `at(`*varlist*`)` to use the levels found in the data for each variable instead of specifying custom values. In any case, the variables specified in `at()` must be different from the variables specified in *termlist*. Furthermore, only variables that appear as covariates in the original model are allowed.

`vce(`*vcetype*`)` specifies the variance estimation method. The default is to compute robust standard errors based on influence functions (taking account of clustering if the original model includes clustering). Use option `vce()` to request replication-based standard errors; *vcetype* may be `bootstrap` or `jackknife`; see [R] ***vce_option***. If replication-based standard errors are requested, `lnmor` will reestimate the original model within replications. Option `vce()` is not allowed after `svy` or `mi estimate`.

`or` reports the results transformed to odds ratios (rather than log odds ratios). This option affects how results are displayed, not how they are estimated.

other_options are further options related to details of estimation as well as displaying and storing results; see the online documentation.

4.2 Inverse probability weighting

Inverse probability weighting is implemented by command `ipwlogit`. The syntax is

`ipwlogit` *depvar* *tvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

where *depvar* equal to nonzero and nonmissing (typically *depvar* equal to one) indicates a positive outcome and *depvar* equal to zero indicates a negative outcome. *indepvars* may include factor variables; `pweight`s, `fweight`s, and `iweight`s are allowed. Stata 14 or newer is required.

Treatment *tvar* can be categorical or continuous. A categorical treatment must be specified using factor variable notation, that is, as `i.`*varname*, where *varname* is the name of the treatment variable. The IPWs will then be based on the observed levels of the variable. A continuous treatment is specified as *varname* without factor variable operator. In this case, the IPWs will be based on a coarsened variable that divides the treatment into a series of equal probability bins (unless option `discrete` is specified; see below). A continuous treatment may also be specified, e.g., as `c.`*varname*`##c.`*varname* to model a nonlinear effect. Options are as follows.

`psmethod(`*method*`)` selects the propensity score estimation method. Supported methods are as follows.

`logit`	for each treatment level, fit a logistic regression of the level against all other levels (using command [R] **logit**)
`mlogit`	fit a multinomial logistic regression across all levels (using command [R] **mlogit**)
`ologit`	fit an ordered logistic regression across all levels (using command [R] **ologit**)
`gologit`	fit a generalized ordered logistic regression across all levels; this re-

quires command `gologit2` by Williams (2006) to be installed on the system (type `ssc install gologit2`)

`cologit` fit a series of cumulative odds models across treatment levels (using command [R] **logit**); this is asymptotically equivalent to `gologit2`, but imposes less computational burden

The default method depends on the type of the treatment variable. For a categorical treatment with two levels (dichotomous treatment), the default is `logit`; for a categorical treatment with more than two levels, the default is `mlogit`; for a continuous or discrete treatment, `cologit` is the default.

`truncate(#)`, with # in $[0, 0.5]$, applies truncation to the inverse probability weights. Weights smaller than quantile # of the overall distribution of weights will be replaced by the value of quantile # and weights larger than quantile $1 - \#$ will be replaced by the value of quantile $1 - \#$. For example, type `truncate(0.01)` to truncate the weights to the 1st and 99th percentile. Truncation will always be applied on the basis of stabilized weights; truncated non-stabilized weights will be obtained by rescaling the truncated stabilized weights.

`bins(#)` sets the number of quantile bins used to categorize a continuous treatment. The resulting number of bins may be less than # if there is heaping in the distribution. The default is to determine the number of bins as $\lceil \ln(n)/\ln(2) \rceil + 1$, where n is the number of obervations (Sturges' rule for the number of histogram bins).

`discrete` declares the treatment variable as discrete. In this case, the variable will not be categorized based on quantiles. Use this option for a quantitative treatment with relatively few distinct levels.

`asbalanced` scales the inverse probability weights in a way such that they correspond to a balanced design in which each treatment level has the same marginal probability. By default, `ipwlogit` uses so-called stabilized weights that reflect the observed distribution.

`vce(`*vcetype*`)` specifies the type of standard error reported. *vcetype* may be `robust` (robust standard errors), `cluster` *clustvar* (cluster-robust standard errors), `bootstrap` or `jackknife`; for bootstrap and jackknife see [R] ***vce_option***. The default is `vce(robust)`.

`or` reports the results transformed to odds ratios (rather than log odds ratios). This option affects how results are displayed, not how they are estimated.

other_options are further options related to details of estimation as well as displaying and storing results; see the online documentation.

4.3 Unconditional logistic regression

The RIF regression approach is implemented by command `riflogit`. The syntax is

`riflogit` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*] where *depvar* equal to nonzero and nonmissing (typically *depvar* equal to one) indicates a positive outcome and *depvar* equal to zero indicates a negative outcome. *indepvars* may include factor variables; `pweight`s, `fweight`s, and `iweight`s are allowed. Prefix commands `svy` and `mi estimate` are supported. Stata 11 or newer is required. Options are as follows.

`vce(`*vcetype*`)` specifies the type of standard error reported. *vcetype* may be `robust` (robust standard errors), `cluster` *clustvar* (cluster-robust standard errors), `bootstrap` or `jackknife`; for bootstrap and jackknife see [R] ***vce_option***. The default is `vce(robust)`.

`or` reports the results transformed to odds ratios (rather than log odds ratios). This option affects how results are displayed, not how they are estimated.

other_options are further options related to details of estimation and displaying results; see the online documentation.

5 Example application

The gender gap in STEM training

In Switzerland, like in many other countries, young men much more often than young women aspire to become a professional in the field of STEM (Science, Technology, Engineering, and Math). One reason for the difference may be that boys specialize more in math throughout their school career than girls, for example, due to gender stereotypes, such that a gender gap in math skills emerges over the school years. Such a gap may then lead to gender differences in occupational aspirations and choices of study fields. Based on such reasoning one would expect the gender STEM gap to decrease once math skills are controlled for. That is, at least part of the total effect of gender may be mediated by math skills. Mechanisms that suppress the gap are also possible. For example, boys may have lower academic motivation than girls, which would reduce their likelihood of becoming a STEM professional because it reduces the likelihood of becoming a professional at all. In such a case, we would expect the gender STEM gap to increase once we control for an indicator of low academic motivation such as grade repetition.

To disentangle the mechanisms we might be tempted to do a mediation analysis by running different logistic regressions, with and without controls, and comparing results across models. Unfortunately, however, such comparisons may be misleading due to the noncollapsibility property of logit coefficients. A solution to the problem is to look at adjusted marginal odds ratios.

For purpose of illustration, we use a data excerpt from the second cohort of the TREE study, a Swiss multi-cohort panel study on the transition from education to employment (TREE 2021). The data look as follows:

```
. use stem, clear
(Excerpt from TREE cohort 2)
. summarize
```

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        stem |      6,809     .262153    .4398377          0          1
        male |      6,809     .450727    .4976028          0          1
   mathscore |      6,809     .244739    1.343228      -5.36      5.214
      repeat |      6,809    .1631664    .3695446          0          1
       books |      6,809    4.354531    1.526026          1          7
-------------+---------------------------------------------------------
          wt |      6,809    11.54357    13.13314   .5645969   80.80351
         psu |      6,809    446.1088    235.8681          1        810
```

The dependent variable is `stem`, an indicator for whether a student is on an educational track that will eventually lead to a STEM profession (measured two years after leaving compulsory school; using a relatively wide definition of STEM that also includes some female dominated occupational fields). `male` is an indicator for the gender of the student, `mathscore` is the test score from a mathematics assessment at the end of compulsory school, `repeat` is an indicator of whether the student ever repeated a grade during compulsory school, and `books` is a measure of cultural capital (number of books at home) that is expected to have a positive effect on academic achievement.[11] Furthermore, `wt` contains sampling weights and `psu` contains the IDs of the primary sampling units.[12]

There is a strong association between `stem` and `male`. The gender difference in the probability of being in training for a STEM profession is about 11 percentage points and the (unadjusted) marginal odds ratio amounts to 1.94. That is, the odds of being in STEM training are almost twice as high for men than for women.

```
. mean stem [pw=wt], over(male) cluster(psu)

Mean estimation                          Number of obs = 6,809

                   (Std. err. adjusted for 800 clusters in psu)
---------------------------------------------------------------
             |             Robust
             |       Mean   std. err.     [95% conf. interval]
-------------+-------------------------------------------------
 c.stem@male |
          0  |    .163234   .0093646      .1448519    .1816161
          1  |   .2748687   .0145161      .2463745    .3033629
---------------------------------------------------------------

. lincom _b[c.stem@1.male] - _b[c.stem@0.male]

 ( 1)  - c.stem@0bn.male + c.stem@1.male = 0

------------------------------------------------------------------------------
        Mean | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |   .1116347   .0142959     7.81   0.000     .0835728    .1396966
------------------------------------------------------------------------------

. logit stem i.male [pw=wt], cluster(psu) nolog or

Logistic regression                                     Number of obs =  6,809
```

11. Variable `books` is a categorical measure with values from 1 "None" to 7 "More then 500 books". For simplicity, we treat the variable as quantitative in our analysis.
12. The sample design also includes stratification, but for simplicity we omit the strata in our analysis (the effect of stratification is negligible in our case).

```
                                                     Wald chi2(1)  =  67.37
                                                     Prob > chi2   = 0.0000
Log pseudolikelihood = -40949.278                    Pseudo R2     = 0.0172
                                  (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      1.male |   1.943131   .1572663     8.21   0.000     1.658099     2.27716
       _cons |   .1950773   .0133746   -23.84   0.000     .1705485    .2231338
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

We now control for `mathscore`, `repeat`, and `books`. The three controls do have the anticipated effects (positive effect of `mathscore`, negative effect of `repeat`, positive effect of `books`), but their addition to the model does not decrease the effect of gender. In fact, the odds ratio of gender even slightly increases to 1.96:

```
. logit stem i.male mathscore i.repeat books [pw=wt], cluster(psu) nolog or
Logistic regression                                  Number of obs =  6,809
                                                     Wald chi2(4)  = 596.03
                                                     Prob > chi2   = 0.0000
Log pseudolikelihood = -31905.554                    Pseudo R2     = 0.2343
                                  (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      1.male |   1.959295   .1675426     7.87   0.000      1.65696    2.316794
   mathscore |   2.606164   .1252437    19.93   0.000     2.371897     2.86357
    1.repeat |   .6563627   .0965248    -2.86   0.004     .4920011    .8756321
       books |   1.087051   .0341241     2.66   0.008     1.022185    1.156034
       _cons |   .1058314   .0166897   -14.24   0.000     .0776926    .1441616
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

From these results we would conclude that the gender STEM gap does not change when controlling for math skills and academic motivation, either because the gender effect is not mediated by these variables or because there are offsetting mechanisms. However, such a conclusion would be wrong, as indicated by an analysis of the adjusted marginal odds ratio using post-estimation command `lnmor`:

```
. lnmor i.male, or
Enumerating predictions: male..done
Marginal odds ratio                          Number of obs     =      6,809
                                             Command           =      logit
                                  (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds Ratio   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      1.male |   1.677032   .1103015     7.86   0.000     1.473911    1.908145
------------------------------------------------------------------------------
```

We see that controlling for `mathscore`, `repeat`, and `books` does reduce the marginal OR of gender to a level of about 1.68.

Note that `lnmor` can compute marginal ORs also for the other variables in the model. Simply list all covariates for which you want to obtain the marginal OR:

```
. lnmor i.male mathscore i.repeat books, or
(mathscore has 491 levels; using 82 binned levels)
Enumerating predictions: male..mathscore....................................
..........................................repeat..books.......done
Marginal odds ratio                              Number of obs     =      6,809
                                                 Command           =      logit
                                  (Std. err. adjusted for 800 clusters in psu)
```

stem	Odds Ratio	Robust std. err.	t	P>|t|	[95% conf.	interval]
1.male	1.677032	.1103015	7.86	0.000	1.473911	1.908145
mathscore	2.544257	.1227974	19.35	0.000	2.314279	2.797088
1.repeat	.7244256	.0839026	-2.78	0.006	.5771099	.9093458
books	1.065542	.025807	2.62	0.009	1.01607	1.117423

A test for confounding or mediation

A test for whether the unadjusted and adjusted estimates of the marginal OR are different is not directly included in the output of `lnmor`, but such a test can be constructed based on the influence functions from two calls to `lnmor`. Use option `rif()` to store the (recentered) influence functions (see the online documentation of `lnmor`). The procedure goes as follows.

Step 1: Obtain the RIF of the marginal OR based on the full model including the covariates.

```
. logit stem i.male mathscore i.repeat books [pw=wt], cluster(psu)
  (output omitted)
. lnmor i.male, nodots noheader notable rif(RIFadj*)
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
RIFadj1         double  %10.0g                RIF of 0b.male
RIFadj2         double  %10.0g                RIF of 1.male
```

Step 2: Obtain the RIF of the marginal OR based on the reduced model excluding the covariates.

```
. logit stem i.male [pw=wt] if e(sample), cluster(psu)
  (output omitted)
. lnmor i.male, nodots noheader notable rif(RIF*)
Variable      Storage   Display    Value
    name         type    format    label      Variable label
```

```
RIF1           double  %10.0g               RIF of 0b.male
RIF2           double  %10.0g               RIF of 1.male
```

Qualifier "`if e(sample)`" makes sure that observations with missing values in the covariates are excluded from the reduced model.

Step 3: Perform the difference test.

```
. total RIFadj2 RIF2 [pw=wt], cluster(psu)
Total estimation                         Number of obs = 6,809
                  (Std. err. adjusted for 800 clusters in psu)

                            Robust
                  Total   std. err.     [95% conf. interval]

     RIFadj2   .5170253   .0657719      .3879192    .6461313
        RIF2   .6643005   .0809345      .5054311    .8231699

. lincom RIFadj2 - RIF2
 ( 1)  RIFadj2 - RIF2 = 0

       Total   Coefficient  Std. err.      t    P>|t|     [95% conf. interval]

         (1)    -.1472752   .0420574   -3.50   0.000    -.2298313   -.0647192

. drop RIF*
```

The difference in the log odds ratio is about 0.15 and appears to be highly significant, as indicated by a t value of 3.5. That is, adding the covariates significantly reduces the remaining gender effect.

Interactions and nonlinear effects

We might be concerned that our original model is not flexible enough to fit the data sufficiently well. Possibly, some interaction terms or polynomials should be included in the model. Increasing the complexity of the specification does not change the definition of the marginal OR. That is, the marginal OR of a predictor can always be obtained in the same way, even if the predictor is involved in interactions or if a nonlinear effect has been modeled. For example, here is the marginal OR of gender from a model that includes interaction terms between all variables and a nonlinear effect of `mathscore`:

```
. logit stem i.male##c.mathscore##c.mathscore##i.repeat##c.books [pw=wt], ///
>     cluster(psu)
  (output omitted)
. lnmor i.male, nodots or
Marginal odds ratio                              Number of obs     =      6,809
                                                 Command           =      logit
                                   (Std. err. adjusted for 800 clusters in psu)

                            Robust
```

```
        stem │ Odds Ratio   std. err.      t    P>|t|     [95% conf. interval]
─────────────┼────────────────────────────────────────────────────────────────
      1.male │   1.676628   .1103018     7.86   0.000      1.47351    1.907746
─────────────┴────────────────────────────────────────────────────────────────
```

There is not much change in the gender STEM gap compared to the simpler model. However, since the model includes interactions it may be interesting to evaluate effect heterogeneity. The `at()` option can be used for this purpose; it computes results under different scenarios with covariates set to specific values. For example, here is how the gender effect differs by `mathscore`:

```
. lnmor i.male, nodots or at(mathscore = -2(2)2)
Marginal odds ratio                               Number of obs     =      6,809
                                                  Command           =      logit
Evaluated at:
    1: mathscore = -2
    2: mathscore = 0
    3: mathscore = 2
                                    (Std. err. adjusted for 800 clusters in psu)
─────────────┬────────────────────────────────────────────────────────────────
             │               Robust
        stem │ Odds Ratio   std. err.      t    P>|t|     [95% conf. interval]
─────────────┼────────────────────────────────────────────────────────────────
1            │
      1.male │   1.697829   .6740845     1.33   0.183     .7788088    3.701323
─────────────┼────────────────────────────────────────────────────────────────
2            │
      1.male │   1.890954   .2009018     6.00   0.000     1.535003    2.329448
─────────────┼────────────────────────────────────────────────────────────────
3            │
      1.male │   1.991302   .3565062     3.85   0.000     1.401245    2.829831
─────────────┴────────────────────────────────────────────────────────────────
```

It seems that the gender effect tends to increase with math score. Note, however, that these differences in effect sizes are too small to be statistically significant, as is confirmed by the following test:

```
. lnmor i.male, at(mathscore = -2(2)2) post
  (output omitted)
. test _b[1:1.male] = _b[2:1.male] = _b[3:1.male]
 ( 1)  [1]1.male - [2]1.male = 0
 ( 2)  [1]1.male - [3]1.male = 0
       F(  2,   799) =    0.06
            Prob > F =    0.9393
```

In the above model, a nonlinear effect has been included for `mathscore`. If we are interested evaluating the corresponding effect pattern, we can use option `dx()` to obtain level-specific marginal ORs of `mathscore`:

```
. logit stem i.male##c.mathscore##c.mathscore##i.repeat##c.books [pw=wt], ///
>     cluster(psu)
  (output omitted)
. lnmor mathscore, nodots or dx(-3(1)3)
```

```
Marginal odds ratio                            Number of obs     =      6,809
                                               Command           =      logit
                                               Type of dx()      =     levels
                                   (Std. err. adjusted for 800 clusters in psu)
```

stem	Odds Ratio	Robust std. err.	t	P>|t|	[95% conf.	interval]
mathscore@l1	2.746111	.5964997	4.65	0.000	1.792845	4.206233
mathscore@l2	2.944801	.4688244	6.78	0.000	2.154446	4.025097
mathscore@l3	2.935113	.3196357	9.89	0.000	2.370215	3.634644
mathscore@l4	2.811596	.1889376	15.38	0.000	2.464143	3.208042
mathscore@l5	2.609491	.157727	15.87	0.000	2.317545	2.938215
mathscore@l6	2.358956	.2203644	9.19	0.000	1.963737	2.833717
mathscore@l7	1.878049	.242352	4.88	0.000	1.457798	2.419449

```
Terms affected by dx(): mathscore
Levels of dx(): -3 -2 -1 0 1 2 3
```

The pattern suggests that the effect of `mathscore` decreases somewhat if the score is high, but still remains positive.

Comparison to ipwlogit and riflogit

The above analyses, at least some of them, could also be performed using `ipwlogit` or `riflogit`. We prefer `lnmor` because it directly quantifies the marginal OR that is *implied* by the chosen model and because it is fully flexible with respect to how the right-hand side of the model is specified. However, for the record, here are the marginal ORs for `male` estimated by `ipwlogit` or `riflogit`.

Unadjusted marginal OR by `ipwlogit`:

```
. ipwlogit stem i.male [pw=wt], or cluster(psu) nolog
(estimating balancing weights ... done)
Marginal logistic regression                   Number of obs     =      6,809
                                               Wald chi2(1)      =      67.37
                                               Prob > chi2       =     0.0000
                                               Pseudo R2         =     0.0172
                                               Treatment type    =     factor
                                               Number of levels  =          2
                                               PS method         =      logit
                                   (Std. err. adjusted for 800 clusters in psu)
```

stem	Odds ratio	Robust std. err.	z	P>|z|	[95% conf.	interval]
1.male	1.943131	.1572663	8.21	0.000	1.658099	2.27716
_cons	.1950773	.0133746	-23.84	0.000	.1705485	.2231338

```
Distribution of IPWs
      level          N       mean        sum        min        max         cv
          0       3740          1   36847.15          1          1          0
          1       3069          1   41753.04          1          1          0
```

Adjusted marginal OR by `ipwlogit`:

```
. ipwlogit stem i.male mathscore i.repeat books [pw=wt], or cluster(psu) nolog
(estimating balancing weights ... done)
Marginal logistic regression                    Number of obs     =      6,809
                                                Wald chi2(1)      =      70.25
                                                Prob > chi2       =     0.0000
                                                Pseudo R2         =     0.0128
                                                Treatment type    =     factor
                                                Number of levels  =          2
                                                PS method         =      logit
                                 (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds ratio   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      1.male |   1.769612   .1205055     8.38   0.000     1.548508    2.022285
       _cons |   .2038352    .013625   -23.79   0.000     .1788061    .2323679
------------------------------------------------------------------------------
(adjusted for mathscore i.repeat books)
Distribution of IPWs
       level          N       mean        sum        min        max         cv
           0       3740   .9988502   36804.78   .7358384   1.549773   .1029689
           1       3069   1.000906   41790.88   .7368664   1.551103   .1009686
```

Unadjusted marginal OR by `riflogit`:

```
. riflogit stem i.male [pw=wt], or cluster(psu)
Unconditional logistic regression                       Number of obs =  6,809
                                                        F(1, 799)     =  60.97
                                                        Prob > F      = 0.0000
                                                        R-squared     = 0.0179
                                                        Adj R-squared = 0.0178
                                                        Root MSE      = 2.3828
                                 (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds ratio   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      1.male |    1.90644   .1575382     7.81   0.000      1.62098    2.242171
       _cons |   .2031708   .0109977   -29.44   0.000     .1826904    .2259473
------------------------------------------------------------------------------
```

Adjusted marginal OR by `riflogit`:

```
. riflogit stem i.male mathscore i.repeat books [pw=wt], or cluster(psu)
Unconditional logistic regression                       Number of obs =  6,809
                                                        F(4, 799)     = 207.19
                                                        Prob > F      = 0.0000
                                                        R-squared     = 0.2145
                                                        Adj R-squared = 0.2140
                                                        Root MSE      = 2.1316
                                 (Std. err. adjusted for 800 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        stem | Odds ratio   std. err.      t    P>|t|     [95% conf. interval]
```

1.male	1.731805	.1189147	8.00	0.000	1.51343	1.981689
mathscore	2.019566	.0603115	23.54	0.000	1.904582	2.141493
1.repeat	.7736294	.0657626	-3.02	0.003	.6547365	.9141119
books	1.059187	.0257164	2.37	0.018	1.009891	1.110889
_cons	.1944999	.0214248	-14.86	0.000	.1566804	.2414483

Results are qualitatively similar to the results from `lnmor`, that is, the adjusted marginal OR is lower than the unadjusted marginal OR, although the reduction is somewhat less pronounced than with `lnmor`.

6 Conclusions

This article defines the marginal odds ratio as an estimand, reviews different estimation techniques, and describes the software implementation of these techniques. The main advantage of marginal odds ratio over conventionally used conditional odds ratios (typically obtained from logistic response models) is that it is unaffected by noncollapsibility: it's magnitude does not change if we adjust for a covariate orthogonal to the treatment variable of interest. Marginal odds ratios can thus be compared across different covariate adjustment sets and will be relevant to both experimental research (in which covariates are added to increase efficiency) and observational research (where confounding is ubiquitous).

7 Acknowledgements

We thank Jeff Pitblado from StataCorp for his advice on some technical details related to the Stata implementation of the estimators.

Kristian Bernt Karlson received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 851293).

8 Appendix: Simulation results

The purpose of the following simulation is to evaluate whether the discussed estimation approaches yield consistent estimates of the marginal OR and whether the computed standard errors are consistent.

We generate data where binary outcome Y depends on treatment T and control variable X through a logistic model. X has a standard normal distribution and the effects of T and X on Y (the conditional log odds ratios) are set to 1 in all simulations (the intercept is set to 0). Treatment T can either be binary or continuous. In the binary case, T is generated through a logistic model with X as predictor. In the continuous case, T is generated as a linear function of X plus a standard normal error. We look at two scenarios, a no-confounding scenario in which the effect of X on T is equal to zero, and a confounding scenario in which the effect is set to 0.5 (the intercept is always 0). We report results from 10'000 runs using a sample size of $n = 1000$.

Figure 1 shows violin plots (Jann 2022) of the distribution of estimates for the binary treatment by estimation method. The dashed line marks the conditional effect (equal to 1); the solid line marks the marginal log odds ratio, which we obtain as the average effect from the unadjusted logit model in the non-confounding scenario (the simulation is set up such that the true marginal odds ratio in the confounding scenario is the same as in the non-confounding scenario). The value of the marginal OR is about 0.84. In the graph, solid circles display the averages of estimates across simulations; the medians are displayed as hollow circles (not visible in this figure since covered by the solid circles for the means). The curves display kernel density estimates of the distributions, the horizontal spikes are box-plot whiskers, and the white space between the whiskers is equal to the inter-quartile range.

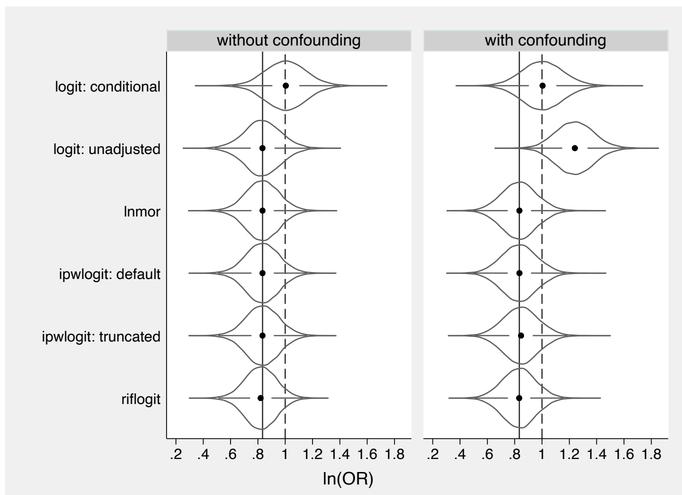

Figure 1: Distribution of effect estimates for binary treatment

We see, as expected, that the conditional logit model provides unbiased estimates of the conditional treatment effect in both the non-confounding as well as the confounding scenario (a log conditional OR of 1). Furthermore, we see that in the non-confounding scenario (left panel) all evaluated estimation techniques, `lnmor`, `ipwlogit` (without and with truncation at the 1st and 99th percentile), and `riflogit` provide estimates that are consistent with the effect estimated by the unadjusted logit. More importantly, the techniques also successfully uncover the marginal odds ratio in the confounding scenario, in which case the unadjusted logit is severely biased (right panel).

Figure 2 displays the simulation results for the continuous treatment. For `lnmor` we now report results for three different estimates, the default estimate based on fractional logit, the average derivative across the treatment distribution, and the derivative-at-observed-values estimate. We see that the "default" and "averaged" methods of `lnmor` provide unbiased estimates of the marginal OR in both the non-confounding and the confounding scenario. The same is true for `ipwlogit` in the non-confounding scenario, but `ipwlogit` does not seem to be fully successful in the confounding scenario. Compared to unadjusted logit, `ipwlogit` substantially reduces confounding, but a small bias seems to remain. Also note that the `ipwlogit` estimate has inflated variance (wider distribution) in the confounding scenario. The instability of `ipwlogit` is due to the fact that the IPWs can get very large if there is poor overlap (i.e., if the distribution of X strongly differs by treatment level), which is likely to happen if confounding is as strong as in the chosen setup. Applying truncation to the weights helps reducing the variance, but increases bias.

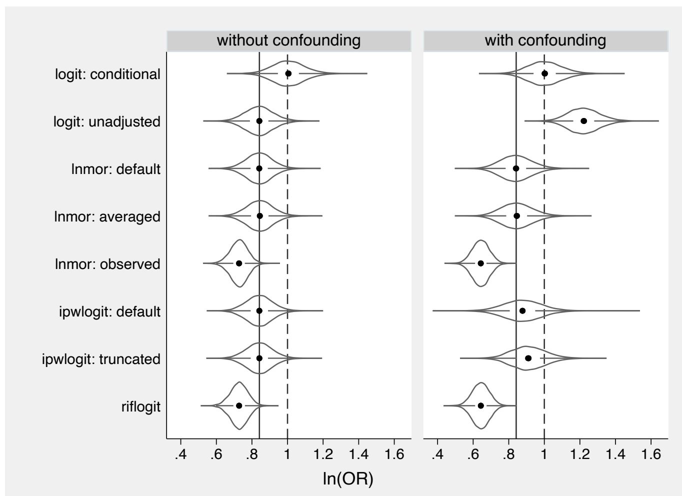

Figure 2: Distribution of effect estimates for continuous treatment

The derivative-at-observed-values estimate by `lnmor` does not recover the "average" marginal OR, but this was not expected, as the estimand is a different one. Interestingly, however, `riflogit` also does not recover the "average" marginal OR. From the results in Figure 2 we see that `riflogit` corresponds to the same estimand as `lnmor` with option `dx(observed)`. That is, for continuous treatments, results from `riflogit` appear to have a derivative-at-observed-values interpretation.

Figures 3 and 4 show the distributions of standard errors for the different estimators. Vertical spikes on these plots depict the observed standard deviations of estimates across simulations. In the case of a binary treatment (Figure 3), we see that standard errors are consistent and well-behaved for all methods. Furthermore, the left panel (non-confounding scenario) illustrates the efficiency gain achieved by the adjusted marginal OR over the unadjusted marginal OR (both are consistent, as shown above, but unadjusted logit has a larger standard deviation than the estimates from `lnmor`, `ipwlogit`, and `riflogit`).

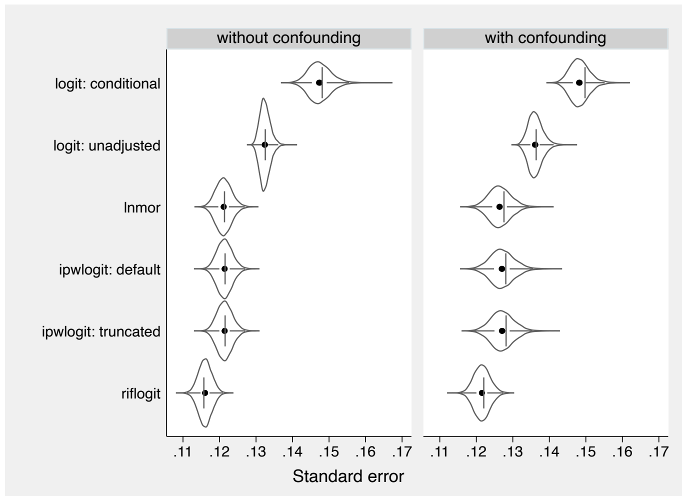

Figure 3: Distribution of standard errors for binary treatment

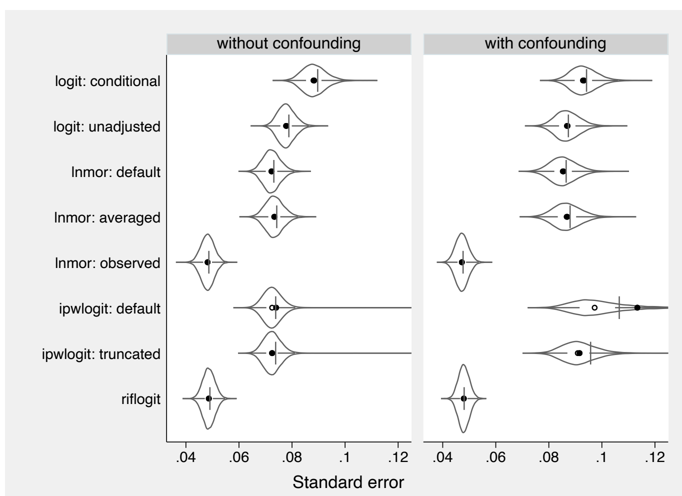

Figure 4: Distribution of standard errors for continuous treatment

In the continuous case (Figure 4), standard errors are again consistent in the non-confounding scenario for all estimators. However, we see that the distribution of standard errors from `ipwlogt` is skewed, with some strong outliers (the density curves have been truncated on the right for purpose of plotting). Applying truncation leads to a less skewed distribution (median and mean are closer together as without truncation), but does not completely remove the outliers. Furthermore, standard error estimates are very unstable for `ipwlogt` in the confounding scenario. The distribution is now considerably skewed. Again, truncation helps, but does not completely remove the problem. For the other estimators, standard errors are consistent and well-behaved also in the confounding scenario.

9 References

Allison, P. D. 1999. Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods & Research* 28(2): 186–208.

Bloome, D., and S. Ang. 2022. Is the Effect Larger in Group A or B? It Depends: Understanding Results From Nonlinear Probability Models. *Demography* 59(4): 1459–1488.

Breen, R., K. B. Karlson, and A. Holm. 2018. Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. *Annual Review of Sociology* 44: 39–54.

Chatton, A., F. Le Borgne, C. Leyrat, F. Gillaizeau, C. Rousseau, L. Barbin, D. Laplaud, M. Léger, B. Giraudeau, and Y. Foucher. 2020. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports* 10(9219).

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2013. Inference on Counterfactual Distributions. *Econometrica* 81(6): 2205–2268.

Cramer, J. S. 2007. Robustness of Logit Analysis: Unobserved Heterogeneity and Misspecified Disturbances. *Oxford Bulletin of Economics and Statistics* 69(4): 545–555.

Daniel, R., J. Zhang, and D. Farewell. 2021. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal* 63(3): 528–557.

Firpo, S., N. M. Fortin, and T. Lemieux. 2009. Unconditional Quantile Regressions. *Econometrica* 77(3): 953–973.

Hampel, F. R. 1974. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association* 69: 383–393.

Jann, B. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001. Boston College Department of Economics. Available from https://ideas.repec.org/c/boc/bocode/s455001.html.

———. 2020. Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment.

University of Bern Social Sciences Working Papers 35. Available from https://ideas.repec.org/p/bss/wpaper/35.html.

———. 2021. Entropy balancing as an estimation command. University of Bern Social Sciences Working Papers 39. Available from https://ideas.repec.org/p/bss/wpaper/39.html.

———. 2022. VIOLINPLOT: Stata module to draw violin plots. Statistical Software Components S459132, Boston College Department of Economics. Available from https://ideas.repec.org/c/boc/bocode/s459132.html.

Karlson, K. B., A. Holm, and R. Breen. 2012. Comparing Regression Coefficients Between Models using Logit and Probit: A New Method. *Sociological Methodology* 42: 286–313.

Karlson, K. B., F. Popham, and A. Holm. 2021. Marginal and Conditional Confounding Using Logits. *Sociological Methods & Research* (Online First, https://doi.org/10.1177/0049124121995548).

Mood, C. 2010. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26(1): 67–82.

Naimi, A. I., E. E. Moodie, N. Auger, and J. S. Kaufman. 2014. Constructing Inverse Probability Weights for Continuous Exposures. A Comparison of Methods. *Epidemiology* 25(2): 292–299.

Neyman, J. 1990[1923]. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (tranlsated and edited by D. M. Dabrowska and T. P. Speed). *Statistical Science* 5(4): 465–472.

Pang, M., J. S. Kaufman, and R. W. Platt. 2016. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research* 25(5): 1925–1937.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9-12): 1393–1512.

Rubin, D. B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Pdychology* 66(5): 688–701.

Schuster, N. A., J. W. R. Twisk, G. ter Riet, M. W. Heymans, and J. J. M. Rijnhart. 2021. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Medical Research Methodology* 21(136).

Snowden, J. M., S. Rose, and K. M. Mortimer. 2011. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *American Journal of Epidemiology* 173(7): 731–738.

Stampf, S., E. Graf, C. Schmoor, and M. Schumacher. 2010. Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. *Statistics in Medicine* 29: 760–769.

TREE. 2021. Transitions from Education to Employment, Cohort 2 (TREE2), Panel waves 0-2 (2016-2018) [Dataset]. University of Bern. Distributed by FORS, Lausanne. https://doi.org/10.23662/FORS-DS-1255-1.

Williams, R. 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal* 6(1): 58–82.

Williams, R., and A. Jorgensen. 2023. Comparing logit & probit coefficients between nested models. *Social Science Research* 109: 102802.

Zhang, Z. 2008. Estimating a Marginal Causal Odds Ratio Subject to Confounding. *Communications in Statistics – Theory and Methods* 38(3): 309–321.