Original Manuscript

Supplementary material for "A general regression framework for a secondary outcome in case-control studies"

Eric J. Tchetgen Tchetgen

Departments of Biostatistics and Epidemiology, Harvard University

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

1

2 Regression with a Log link function

Here we give a generalization of the methods by considering a nonnegative outcome $Y \geq 0$ which we model with the log-link function. In order to account for the retrospective sampling design, we again condition on case-control status in the regression model, while obtaining simultaneously, inferences about a marginal regression model with respect to disease status, for the underlying population. To proceed, we now give a key re-parametrization of $E(Y|\mathbf{X}, D) = \widetilde{\mu}(\mathbf{X}, D)$ on the multiplicative scale. Note that:

$$\begin{aligned} E(Y|\mathbf{X}, D) &= \frac{E(Y|\mathbf{X}, D)}{E(Y|\mathbf{X})} \times E(Y|\mathbf{X}) \\ &= \frac{E(Y|\mathbf{X}, D)}{E(Y|\mathbf{X}, D=0)} \times \left\{ \sum_{d^*=0}^{1} \frac{E(Y|\mathbf{X}, D=d^*)}{E(Y|\mathbf{X}, D=0)} \Pr(D=d^*|\mathbf{X}) \right\}^{-1} \times E(Y|\mathbf{X}) \\ &= \exp\left[\log \mu(\mathbf{X}) + \nu(\mathbf{X}, D) - \overline{\nu}(\mathbf{X})\right] \end{aligned}$$

where $\nu(\mathbf{X}, D) = \log E(Y|\mathbf{X}, D)/E(Y|\mathbf{X}, D=0)$ measures the multiplicative association between D and Y within levels of $\mathbf{X}$, and accounts for selection bias possibly induced by the retrospective sampling design. The term $\overline{\nu}(\mathbf{X}) = \log\left\{\exp\left(\nu(\mathbf{X}, D=1)\right)\Pr(D=1|\mathbf{X}) + \Pr(D=0|\mathbf{X})\right\}$ ensures that, as one would hope would be the case, upon marginalization over D in the target population, the conditional mean function $E(Y|\mathbf{X}, D)$ reduces exactly to $E(Y|\mathbf{X})$. Similar to the identity link, the current re-parametrization is nonparametric and also variation independent, which here implies that, except for the restriction that $E(Y|\mathbf{X}, D) \geq 0$, it does not a priori rule out any data generating mechanism.

A simplification occurs when D is rare in the population, such that $E(Y|\mathbf{X}, D=0) \approx E(Y|\mathbf{X})$ and therefore $\overline{\nu}(\mathbf{X}) \approx 1$, which gives $E(Y|\mathbf{X}, D=1) \approx \exp\left[\log \mu(\mathbf{X}) + \nu(\mathbf{X}, 1)\right]$. Therefore $E(Y|\mathbf{X}, D) \approx \exp\left\{\log \mu(\mathbf{X}) + \nu(\mathbf{X}, 1)D\right\}$. Note here again, that only the first term on the exponential scale can be interpreted as an association measure between $\mathbf{X}$ and Y in the target population, any interaction between $\mathbf{X}$ and D encoded in the second term of the right hand side of the expression does not have such an interpretation. In the simple case where $\log \mu(\mathbf{X}) = (1, \mathbf{X}^T)\beta_0$, and where the multiplicative association between D and Y is constant across levels of $\mathbf{X}$, simply adding the main effect for D to the population model of interest to obtain $E(Y|\mathbf{X}, D) \approx \exp\left\{(1, \mathbf{X}^T)\beta_0 + \nu D\right\}$ is approximately correct.

Otherwise, if the disease is not necessarily rare, one may model $\nu(\mathbf{X}, D)$, using say $\nu(\mathbf{X}, D; \alpha) = D\left(1, \mathbf{X}^T\right)\alpha$, resulting in the following parametric model for $E(Y|\mathbf{X}, D) =$

$$\begin{aligned} \widetilde{\mu}(\mathbf{X}, D; \theta) &= \exp\left[(1, \mathbf{X}^T)\beta + D\left(1, \mathbf{X}^T\right)\alpha - \overline{\nu}(\mathbf{X}; \psi, \eta, \alpha)\right] \\ \overline{\nu}(\mathbf{X}; \psi, \eta, \alpha) &= \log\left[\exp\left\{\left(1, \mathbf{X}^T\right)\alpha\right\} p(\mathbf{X}; \psi, \eta) + 1 - p(\mathbf{X}; \psi, \eta)\right] \\ \theta &= (\beta', \eta, \psi', \alpha')' \end{aligned} \tag{1}$$

Estimation and inference about θ_0 the true value of θ, then follows as in the case of an identity link function, by solving the estimating equation $\mathbf{W}\left(\widehat{\theta}\right) = \sum_i \mathbf{U}_i\left(\widehat{\theta}\right) = 0$ given in the main text, upon substituting in (1) for $\widetilde{\mu}(\mathbf{X}, D; \theta)$, and where $\left(\widehat{\psi}, \widehat{\eta}\right)$ is the mle defined in the text. The asymptotic distribution of $\widehat{\theta}$ is then as described in the main text, upon making the foregoing substitutions.

3 Semiparametric efficiency for logit link

To derive the semiparametric efficiency bound for the logistic case, define the semiparametric model $\mathcal{M}_2$ with sole restriction the parametric models

$$\text{logit}\pi(\mathbf{X};\psi_0,\eta_0) = \text{logit}\Pr(D=d^*|\mathbf{X},Y=0,S=1;\psi_0,\eta_0) = \eta_0 + m\left(\mathbf{X};\psi_0\right), \tag{2}$$

and

$$\begin{aligned}\text{logit}\Pr\left(Y=1|D,\mathbf{X};\theta_0\right) &= \mu^{\dagger}(\mathbf{X};\beta_0) + \nu(\mathbf{X},D;\alpha_0) - \overline{\nu}(\mathbf{X};\psi_0,\eta_0,\alpha_0) \\ \theta_0 &= \left(\beta_0',\eta_0,\psi_0',\alpha_0'\right)'\end{aligned} \tag{3}$$

and the model is otherwise unrestricted in $f(\mathbf{X})$ and therefore in $f^*(\mathbf{X})$. Note that, whereas $\mathcal{M}_1$ parametrizes $\Pr(D=d^*|\mathbf{X},S=1)$, $\mathcal{M}_2$ places a model for the density $\Pr(D=d^*|\mathbf{X},Y=0,S=1)$. Nonetheless, as we show next, model (2) together with model (3) yield a parametric model for the conditional density $f(Y,D|\mathbf{X},S=1)$, under the following nonparametric characterization of a joint density (see for example Chen, 2007, Tchetgen Tchetgen et al, 2010 and Tchetgen Tchetgen and Rotnitzky, 2012):

$$\begin{aligned} f(Y,D|\mathbf{X},S=1) &= \frac{f(Y|D=0,\mathbf{X},S=1)OR(Y,D|\mathbf{X},S=1)f(D|Y=0,\mathbf{X},S=1)}{\sum_{d,y} f(Y|D=0,\mathbf{X},S=1)OR(Y,D|\mathbf{X},S=1)f(D|Y=0,\mathbf{X},S=1)} \\ &= \frac{f(Y|D=0,\mathbf{X})OR(Y,D|\mathbf{X})f(D|Y=0,\mathbf{X},S=1)}{\sum_{d,y} f(y|D=0,\mathbf{X})OR(y,d|\mathbf{X})f(d|Y=0,\mathbf{X},S=1)} \\ &= \frac{f(Y|D=0,\mathbf{X})OR(Y,D|\mathbf{X})f(D|Y=0,\mathbf{X})\left\{\overline{p}\left(1-\overline{\pi}\right)/\overline{\pi}\left(1-\overline{p}\right)\right\}^D}{\sum_{d,y} f(y|D=0,\mathbf{X})OR(y,d|\mathbf{X})f(d|Y=0,\mathbf{X},S=1)\left\{\overline{p}\left(1-\overline{\pi}\right)/\overline{\pi}\left(1-\overline{p}\right)\right\}^d} \end{aligned} \tag{4}$$

where $OR(Y,D|\mathbf{X},S=1) = OR(Y,D|\mathbf{X}) =$

$$\begin{aligned}&\frac{f(Y|D,\mathbf{X})f(Y=0|D=0,\mathbf{X})}{f(Y|D=0,\mathbf{X})f(Y=0|D,\mathbf{X})} \\ &= \nu(\mathbf{X},1)\end{aligned}$$

is the odds ratio function relating D and Y within levels of $\mathbf{X}$, which under our choice of parametrization, gives,

$$f(Y,D|\mathbf{X},S=1;\theta_0) = \frac{\exp\left\{Y\mu^{\dagger}(\mathbf{X};\beta_0) + Y\nu(\mathbf{X},D;\alpha_0) - Y\overline{\nu}(\mathbf{X};\alpha_0,\psi_0,\eta_0) + D\eta_0 + Dm\left(\mathbf{X};\psi_0\right)\right\}}{\sum_{d,y}\exp\left\{y\mu^{\dagger}(\mathbf{X};\beta_0) + y\nu(\mathbf{X},d;\alpha_0) - y\overline{\nu}(\mathbf{X};\alpha_0,\psi_0,\eta_0) + d\eta_0 + dm\left(\mathbf{X};\psi_0\right)\right\}} \tag{5}$$

This in turn implies a parametric model $f(D=1|\mathbf{X},S=1;\theta_0) = \sum_y f(y,D=1|\mathbf{X},S=1;\theta_0)$ for $\pi(\mathbf{X})$ in terms of θ_0. Note that in the target population, the analog to equation (4) is

$$f(Y,D|\mathbf{X}) = \frac{f(Y|D=0,\mathbf{X})OR(Y,D|\mathbf{X})f(D|Y=0,\mathbf{X})}{\sum_{d,y} f(y|D=0,\mathbf{X})OR(y,d|\mathbf{X})f(d|Y=0,\mathbf{X},S=1)},$$

which in turn can be used to verify that under the proposed parametrization,

$$\begin{aligned}\text{logit}\mathbb{E}\left\{\widetilde{\mu}(\mathbf{X},D)|\mathbf{X}\right\} &= \text{logit}\sum_d f(Y=1,D=d|\mathbf{X}) \\ &= \text{logit}\left[1+\exp\left\{-\log\frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})}\right\}\right]^{-1} \\ &= \text{logit}f(Y=1|\mathbf{X})\end{aligned}$$

formally justifying the earlier claim that our choice of parametrization is made to ensure such marginalization whether nonparametric, semiparametric or parametric models are used.

The following theorem gives the efficient score for θ_0 in models $\mathcal{M}_2$.

Proposition 1 *(Logit Link) The efficient score in model $\mathcal{M}_2$ is given by the score equation of θ corresponding to the log-likelihood $\mathbb{P}_n \log f(Y, D|\mathbf{X}, S=1;\theta)$ defined in equation* (5).

The result states that when Y is binary, the semiparametric efficiency bound is achieved by the maximum likelihood estimator that solves $\mathbb{P}_n \mathbf{R}_{bin}(\theta) = \mathbb{P}_n \partial \log f(Y, D|\mathbf{X}, S=1;\theta)/\partial\theta = 0$, with variance obtained by an empirical version of $\mathbb{E}\left\{\mathbf{R}_{bin}(\theta_0)\mathbf{R}_{bin}^T(\theta_0)\right\}^{-1}$. This results follows from standard maximum likelihood theory.

PROOF OF PROPOSITION 1: Let L_2^0 denote the Hilbert space of mean zero functions of $O = (Y, D, \mathbf{X})$, with inner product given by the expectation wrt F_O the case-control distribution of O with density equivalently written $f(\varepsilon(\theta_0)|\mathbf{X}, D) f^*(D|\mathbf{X};\psi_0,\eta_0)\ f^*(\mathbf{X})$. The model is semiparametric in the sense that the conditional density of the residual $\varepsilon(\theta_0)$ given $(\mathbf{X}, D)$ and the case-control density of $\mathbf{X}$ are left unrestricted. Throughout, assume that the population disease prevalence is known. The nuisance tangent space Λ_{nuis} for the model is given by the closed linear span of all regular parametric scores for the conditional density of $\varepsilon(\theta_0)$ given $(\mathbf{X}, D)$ and of $f^*(\mathbf{X})$. Then, one can verify that

$$\Lambda_{nuis} = \left\{ \begin{array}{c} a_1(O) + a_2(\mathbf{X}) : \text{such that} \\ \mathbb{E}\{a_1(O)|\mathbf{X}, D\} = \mathbb{E}\{\varepsilon(\theta_0) a_1(O)|\mathbf{X}, D\} = \mathbb{E}\{a_2(\mathbf{X})\} = 0 \end{array} \right\} \cap L_2^0$$

It follows that the set of all influence functions is contained in the ortho-complement of Λ_{nuis} :

$$\Lambda_{nuis}^{\perp} = \{h_1(\mathbf{X}, D)\varepsilon(\theta_0) + h_2(\mathbf{X})\{D - \Pr(D=1|\mathbf{X}, S=1;\psi_0,\eta_0)\} :\ h_1, h_2\} \cap L_2^0$$

Next, let $\mathbf{S}_{\theta_0}(O;\theta_0) = \partial \log f(\varepsilon(\theta_0)|\mathbf{X}, D)/\partial\theta_0 + \partial \log f^*(D|\mathbf{X};\psi_0,\eta_0)/\partial\theta_0$ denote the score wrt $\theta_0 = (\beta_0', \alpha_0', \eta_0, \psi_0')'$, Then:

$$\begin{aligned} S_{\theta_0}(O;\theta_0) &= \mathbf{S}^1_{\theta_0}(O;\theta_0) + \mathbf{S}^2_{\theta_0}(O;\theta_0) \\ &= -\frac{\partial f(\varepsilon(\theta_0)|\mathbf{X}, D;\theta_0)}{\partial \varepsilon(\theta_0)} \times \frac{1}{f(\varepsilon(\theta_0)|\mathbf{X}, D;\theta_0)} \times \frac{\partial \widetilde{\mu}(\mathbf{X}, D;\theta)}{\partial\theta}|_{\theta_0} \\ &+ \left(\begin{array}{c} 0 \\ 1 \\ \frac{\partial m(\mathbf{X};\psi_0)}{\partial \psi_0} \end{array} \right) \{D - \Pr(D=1|\mathbf{X}, S=1;\psi_0,\eta_0)\} \end{aligned}$$

therefore, the efficient score of θ_0 is given by the orthogonal projection of $\mathbf{S}_{\theta_0}(O;\theta_0)$ onto $\Lambda_{nuis}^{\perp}$. Upon noting that $\mathbb{E}\left(\frac{\partial f(\varepsilon(\theta_0)|\mathbf{X}, D;\theta_0)}{\partial \varepsilon(\theta_0)} \times \frac{1}{f(\varepsilon(\theta_0)|\mathbf{X}, D;\theta_0)} \times \varepsilon(\theta_0)|\mathbf{X}\right) = -1$, it is straightforward to verify that this projection is given by $\mathbf{R}_{(\eta,\psi)}(\theta_0)$, with $\mathbf{S}^2_{\theta_0}(O;\theta_0) = \mathbf{S}(\psi_0,\eta_0)$.

□