On the omission of continuous covariates in logistic regression

Matteo Gasparin[1], Bruno Scarpa[1], Elena Stanghellini[2]

[1]Università degli Studi di Padova, Italy [2]Università degli Studi di Perugia, Italy

Introduction

In linear models, the well-known Cochran's formula (Cochran, 1938) allows to determine the exact relationship between marginal and conditional parameters. When the assumption of linearity is not met, the formula does not carry over, and rather complex formulations arise. One of the most interesting cases appears when the outcome is **binary**.

Problems:

- ✗ non collapsibility of the link function;
- ✗ the marginal model may not be linear even if the conditional model is.

However, due to their elegance and interpretability, in the applied world several instances exist where investigators still use Cochran's formula or difference method.

This problem plays a key role in **mediation analysis**, where the goal is to investigate the mechanism that underlie a relationship between an outcome, a treatment and a third intermediate variable, called mediator. In particular, in this work we explore the framework in which the outcome is binary while the covariates present continuous nature.

Parametric Mediation Analysis

The relationship between marginal and conditional parameters is widely used in parametric **mediation analysis**. In this context, the aim is to decompose the **total effect** of a continuous treatment into a **direct** effect and an **indirect** one, this second transmitted through a continuous mediator.

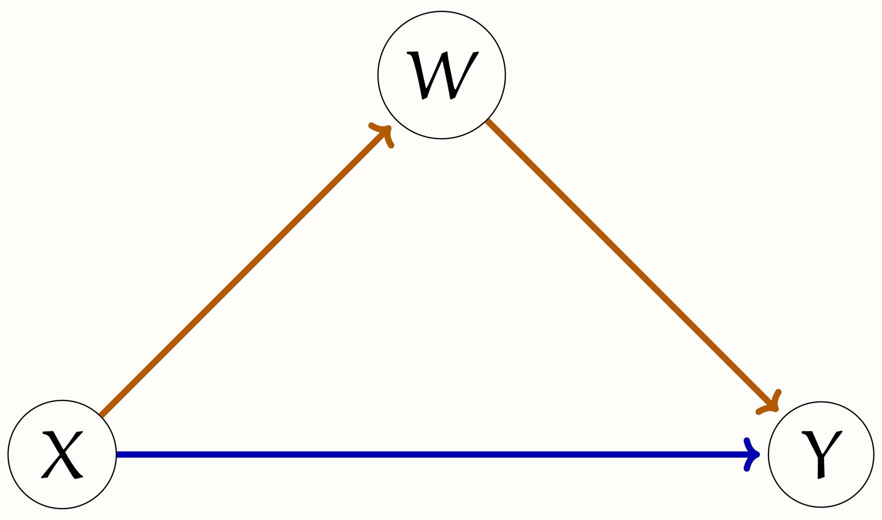

Figure 1: Relationship between variables

In particular, as shown in Figure 1, we define:

- Y the outcome variable;
- X the treatment of interest;
- W the mediator.

Under the rare outcome assumption, VanderWeele & Vansteelandt (2010) show that the relationship between conditional and marginal parameters mimics the Cochran's one in an approximated way. This assumption is quite stringent, and other works try to solve this problem by using approximations of the Cochran's method or of the difference method (MacKinnon et al., 2007).

Description of our DGP

Our postulated models are a bivariate normal distribution (with standard marginal for simplicity) for the two regressors, that is,

$$\mathbf{Z} = (X, W) \sim N_2(0, \Omega), \quad \Omega = \begin{pmatrix} 1 & \omega_{xw} \\ \omega_{xw} & 1 \end{pmatrix},$$

while the continuous version of the outcome variable is defined by a linear combination between $\mathbf{Z}$ and $T \sim \mathrm{Lo}(0,1)$, that is,

$$Y^* = \beta_x X + \beta_w W - T,$$

where the term T can be interpreted as an additive error.

The binary outcome is defined by the following dichotomization

$$Y = \begin{cases} 1 & \text{if } Y^* > -\beta_0, \\ 0 & \text{otherwise.} \end{cases}$$

The probability of the first class given the values of the explanatory variables is

$$\mathbb{P}(Y = 1 \mid X = x, W = w) = \frac{\exp(\beta_0 + \beta_x x + \beta_w w)}{1 + \exp(\beta_0 + \beta_x x + \beta_w w)},$$

that is a simple logistic regression.

Using the results on the **skew-symmetric distributions** described in Azzalini & Capitanio (2013), it is possible to obtain the density function of the vector $\mathbf{Z}$ given the value of the outcome.

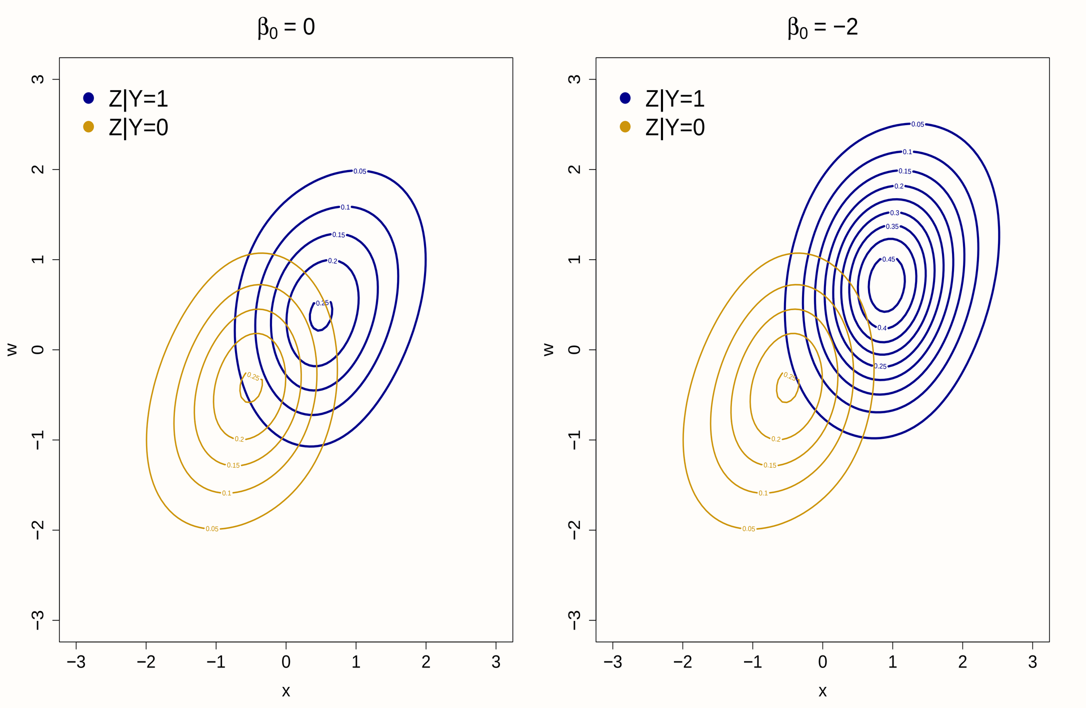

Figure 2: Density of $\mathbf{Z} \mid Y = y$ with $\beta_x = 2$ and $\beta_w = 1$

Main results

After marginalization with respect to the variable W, it is possible to demonstrate that the **probability** of the first class given the value of the only variable X is equal to

$$\mathbb{P}(\beta_w\sqrt{1-\omega_{xw}^2}Z_0 - T > -\beta_0 - (\beta_x + \beta_w\omega_{xw})x),$$

where $Z_0 \sim N(0,1), T \sim \mathrm{Lo}(0,1)$ with $T \perp\!\!\!\perp Z_0$.

Some **comments**:

- the marginal logit is linear with respect to x only if $\beta_w = 0$;
- the random variable $V = \beta_w\sqrt{1-\omega_{xw}^2}Z_0 - T$ is still symmetric with a bell shape.

Solution: in order to obtain a marginal linear logit we approximate the variable V with a logistic random variable with the same variance of V, that is,

$$V \overset{\text{appr}}{\sim} \mathrm{Lo}\left(0, \frac{\sqrt{3}}{\pi}\sqrt{\frac{\pi^2}{3} + \beta_w^2(1-\omega_{xw}^2)}\right).$$

Using this approximation and the properties of the logistic distribution, we obtain

$$\mathbb{P}(Y = 1 \mid X = x) \approx \frac{\exp(\eta_0 + \eta_x x)}{1 + \exp(\eta_0 + \eta_x x)},$$

where the marginal parameter η_x can be decomposed into the weighted sum of the conditional parameters, in particular,

$$\eta_x = \frac{\pi}{\sqrt{3}} \frac{\beta_x + \beta_w\omega_{xw}}{\sqrt{\frac{\pi^2}{3} + \beta_w^2(1-\omega_{xw}^2)}}.$$

This relationship is similar to Cochran's decomposition with the addition of a scale parameter, and it permits the disentangling of **direct and indirect effects**.

Some interesting cases

- If $W \perp\!\!\!\perp Y \mid X$ then
$$\eta_x = \beta_x;$$
- if $X \perp\!\!\!\perp W$ then
$$\eta_x \approx \frac{\pi}{\sqrt{3}} \frac{\beta_x}{\sqrt{\frac{\pi^2}{3} + \beta_w^2}};$$
- if $X \perp\!\!\!\perp Y \mid W$ then
$$\eta_x \approx \frac{\pi}{\sqrt{3}} \frac{\beta_w\omega_{xw}}{\sqrt{\frac{\pi^2}{3} + \beta_w^2(1-\omega_{xw}^2)}}.$$

Simulation studies

We generate $n = 500$ observations from our Data Generating Process with $\beta_0 = 0$ and $\beta_x = 0.6$. In order to compare our method with the others, it is necessary to point out that W can be interpreted as the linear regression

$$W = \theta_x x + \varepsilon_w, \quad \varepsilon_w \sim N(0, \sigma^2),$$

with $\theta_x = \omega_{xw}$, in fact $W \mid X = x \sim N(\omega_{xw}x, 1-\omega_{xw}^2)$.

We compare different methods to estimate the indirect effects:

- **difference method:** $\hat{\eta}_x - \hat{\beta}_x$;
- **product method:** $\hat{\beta}_w\hat{\theta}_x$;
- **difference method with standardized coefficients** as proposed in MacKinnon et al. (2007)**:**
$$\hat{\eta}_x\sqrt{\frac{3}{\pi^2}\left(\hat{\beta}_w\hat{\sigma}^2 + \frac{\pi^2}{3}\right)} - \hat{\beta}_x;$$
- **our method:**
$$\frac{\pi}{\sqrt{3}} \frac{\hat{\beta}_w\hat{\omega}_{xw}}{\sqrt{\frac{\pi^2}{3} + \hat{\beta}_w^2(1-\hat{\omega}_{xw}^2)}}.$$

The parameters are estimated using Maximum Likelihood and, for each setting, the number of Monte-Carlo replications is equal to 10 000. In the first scenario we fix $\omega_{xw} = 0.5$ and we estimate the indirect effect for different values of β_w (Figure 3(a)), in the second scenario we fix $\omega_{xw} = 0$ so $X \perp\!\!\!\perp W$ and mediated effects are not present (Figure 3(b)).

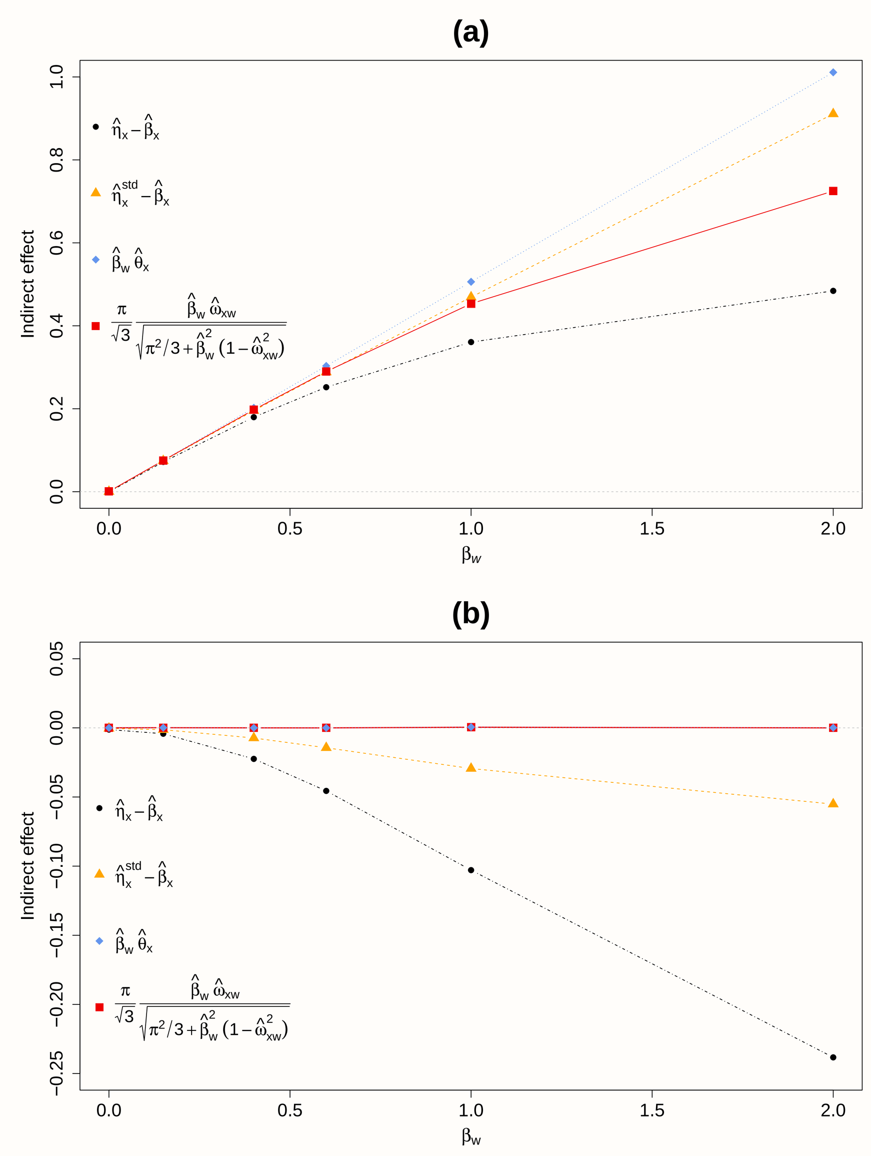

Figure 3: Indirect effects as β_w change for $\omega_{xw} = 0.5$(a) and $\omega_{xw} = 0$(b)

Our method performs well when indirect effects are absent and it seems to offer a right quantification of the mediated effect. In general, this procedure can be extended with other link functions.

References

Azzalini, A. & Capitanio, A. (2013). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, 5, 171–176.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, 4(5), 499–513.

VanderWeele, T. J. & Vansteelandt, S. (2010). Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*, 172(12), 1339–1348.