SciMedCentral | Annals of Biometrics & Biostatistics

Research Article

A Note on the Logistic Regression Model with a Random Coefficient to Predict Propensity Scores

Yasutaka Chiba*

Division of Biostatistics, Clinical Research Center, Kinki University School of Medicine, Osaka, Japan

Corresponding author

Yasutaka Chiba, Division of Biostatistics, Clinical Research Center, Kinki University School of Medicine, 377-2, Ohno-higashi, Osakasayama, Osaka 589-8511, Japan, E-mail: chibay@med.kindai.ac.jp

Submitted: 20 June 2013

Accepted: 01 July 2013

Published: 05 July 2013

Copyright

© 2013 Chiba

OPEN ACCESS

Keywords

- Confounding risk ratio
- Inverse-probability-weighting
- Marginal structural model
- Potential outcome

Abstract

In observational studies, marginal structural models are often used to adjust for confounding. When predicting propensity scores, some investigators may want to apply a logistic regression model with a random coefficient to take account of residual confounding. Here, we show that the random coefficient can be interpreted as the logarithm of the confounding risk ratio; *i.e.*, the ratio of crude risk ratio to causal risk ratio. Three target populations (the exposed, unexposed, and total groups) are discussed.

INTRODUCTION

Confounding is widely recognized as one of the principal problems faced by investigators conducting observational studies. In an analysis, some investigators may want to take account of residual confounding. In the situations, a random coefficient regression model (mixed effect model with random intercept) may be applied. The use of such a model has been discussed in the context of ordinal regression analysis [1-3]. However, this discussion has not been conducted in the context of marginal structural models (MSMs) [4,5], in which a logistic regression model is often used to predict propensity scores [6,7].

Here, we give an interpretation of a random coefficient when a logistic regression model with a random coefficient is used for predicting propensity scores. We discuss MSMs under the three target populations: the total, exposed, and unexposed groups.

MATERIALS AND METHODS

We use X as an exposure indicator and assume the now-standard deterministic potential outcome model [8], in which $Y_{X=1}$ and $Y_{X=0}$ are the potential outcome indicators under $X = 1$ and $X = 0$, respectively. The potential risks $\Pr(Y_{X=1} = 1)$ and $\Pr(Y_{X=0} = 1)$ are then the expectation of Y if everyone in the study population had been exposed and that if everyone had been not exposed, respectively. Causal effects with the total group as the target population are contrasts between these two risks. Those with $X = x$ as the target population are contrasts between $\Pr(Y_{X=1} = 1 \mid X = x)$ and $\Pr(Y_{X=0} = 1 \mid X = x)$.

Let i = 1, ..., n denote a subject and z_i denote a vector of measured confounders. The propensity score $\Pr(X = 1 \mid Z = z_i)$ is then predicted using a logistic regression model:

$$\Pr(X=1\mid Z=z_i)=\frac{\exp(\theta' z_i)}{1+\exp(\theta' z_i)}, \tag{1}$$

where θ is a vector of the regression coefficient. When residual confounding exists, however, Equation (1) derives the biased propensity scores. As a result, the MSM will derive biased estimates of causal effects. In the next section, we give an interpretation of a random coefficient when it is included in Equation (1).

RESULT AND DISCUSSION

Unexposed group as the target population

To take account of residual confounding, we assume that the propensity score p_{0i} is explained by the following logistic regression model with a random coefficient:

$$p_{0i}=\frac{\exp\{\log(\alpha_i)+\theta' z_i\}}{1+\exp\{\log(\alpha_i)+\theta' z_i\}},$$

where $\log(\alpha_i)$ is a random coefficient. Using Equation (1), p_{0i} can be expressed as:

$$\begin{aligned} p_{0i} &= \frac{\exp(\theta' z_i)}{1/\alpha_i+\exp(\theta' z_i)} \\ &= \frac{\Pr(X=1\mid Z=z_i)}{\Pr(X=0\mid Z=z_i)/\alpha_i+\Pr(X=1\mid Z=z_i)}. \end{aligned} \tag{2}$$

Using the inverse-probability-weighting (IPW) method,

$\Pr(Y_{X=x} = 1|X = 0)$ is estimated as follows:

$$\Pr(Y_{X=1}=1\,|\,X=0)=\frac{1}{n_0}\sum_{i=1}^{n}\frac{1-p_{0i}}{p_{0i}}y_ix_i\,, \tag{3}$$

$$\Pr(Y_{X=0}=1\,|\,X=0)=\frac{1}{n_0}\sum_{i=1}^{n}y_i(1-x_i)\,,$$

where $n_0 = n\Pr(X = 0)$ [9,10]. In the framework of MSMs, the causal risk difference (RD) is estimated using a weighted linear regression analysis of X on Y with the weights $(1 - p_{0i}) / p_{0i}$ for exposed subjects and 1 for unexposed subjects. The causal risk ratio (RR) is estimated using the weighted Poisson regression analysis of X on Y with the same weights.

By substituting Equation (2) into Equation (3) and replacing y_ix_i / n with $\Pr(Y = 1, X = 1, Z = z_i)$ in the summation, the following equation is derived:

$$\begin{aligned}&\Pr(Y_{X=1}=1\,|\,X=0)\\&=\frac{n}{n_0}\sum_{i=1}^{n}\frac{\Pr(X=0\,|\,Z=z_i)/\alpha_i}{\Pr(X=1\,|\,Z=z_i)}\frac{y_ix_i}{n}\\&=\frac{1}{\Pr(X=0)}\sum_{i=1}^{n}\frac{\Pr(X=0\,|\,Z=z_i)}{\alpha_i\Pr(X=1\,|\,Z=z_i)}\Pr(Y=1,X=1,Z=z_i)\\&=\sum_{i=1}^{n}\frac{\Pr(Y=1\,|\,X=1,Z=z_i)}{\alpha_i}\Pr(Z=z_i\,|\,X=0).\end{aligned}$$

The left- and right-hand sides of this equation are equal when:

$$\Pr(Y=1\,|\,X=1,Z=z_i)/\alpha_i=\Pr(Y_{X=1}=1\,|\,X=0,Z=z_i),$$

because

$$\begin{aligned}&\Pr(Y_{X=1}=1\,|\,X=0)\\&=\textstyle\sum_{i=1}^{n}\Pr(Y_{X=1}=1\,|\,X=0,Z=z_i)\Pr(Z=z_i\,|\,X=0).\end{aligned}$$

Therefore:

$$\begin{aligned}\alpha_i&=\frac{\Pr(Y_{X=1}=1\,|\,X=1,Z=z_i)}{\Pr(Y_{X=1}=1\,|\,X=0,Z=z_i)}\\&=\frac{\Pr(Y=1\,|\,X=1,Z=z_i)}{\Pr(Y=1\,|\,X=0,Z=z_i)}\bigg/\frac{\Pr(Y_{X=1}=1\,|\,X=0,Z=z_i)}{\Pr(Y_{X=0}=1\,|\,X=0,Z=z_i)}.\end{aligned}$$

This α_i is the confounding risk ratio (CRR) with the unexposed group as the target population [11], which is the ratio of crude RR to causal RR, for an individual with $Z = z_i$.

Exposed group as the target population

In the case of the exposed group as the target population, we can make an argument similar to the above subsection. We assume that the propensity score p_{1i} is explained by the following logistic regression model with a random coefficient:

$$p_{1i}=\frac{\exp\{\log(\beta_i)+\theta' z_i\}}{1+\exp\{\log(\beta_i)+\theta' z_i\}}\,,$$

where $\log(\beta_i)$ is a random coefficient. Then, by the IPW method, $\Pr(Y_{X=x} = 1|X = 1)$ is estimated as:

$$\Pr(Y_{X=1}=1\,|\,X=1)=\frac{1}{n_1}\sum_{i=1}^{n}y_ix_i\,,$$

$$\Pr(Y_{X=0}=1\,|\,X=1)=\frac{1}{n_1}\sum_{i=1}^{n}\frac{p_{1i}}{1-p_{1i}}y_i(1-x_i)\,,$$

where $n_1 = n\Pr(X = 1)$ [9,10]. In the framework of MSMs, the causal effects are estimated using the weighted regression analyses of X on Y with the weights 1 for exposed subjects and $p_{1i} / (1 - p_{1i})$ for unexposed subjects.

Algebra similar to the above subsection yields:

$$\begin{aligned}&\Pr(Y_{X=0}=1\,|\,X=1)\\&=\sum_{i=1}^{n}\beta_i\Pr(Y=1\,|\,X=0,Z=z_i)\Pr(Z=z_i\,|\,X=1).\end{aligned}$$

Because

$$\begin{aligned}&\Pr(Y_{X=0}=1\,|\,X=1)\\&=\textstyle\sum_{i=1}^{n}\Pr(Y_{X=0}=1\,|\,X=1,Z=z_i)\Pr(Z=z_i\,|\,X=1),\end{aligned}$$

β_i can be expressed as:

$$\begin{aligned}\beta_i&=\frac{\Pr(Y_{X=0}=1\,|\,X=1,Z=z_i)}{\Pr(Y_{X=0}=1\,|\,X=0,Z=z_i)}\\&=\frac{\Pr(Y=1\,|\,X=1,Z=z_i)}{\Pr(Y=1\,|\,X=0,Z=z_i)}\bigg/\frac{\Pr(Y_{X=1}=1\,|\,X=1,Z=z_i)}{\Pr(Y_{X=0}=1\,|\,X=1,Z=z_i)}.\end{aligned}$$

This β_i is the CRR with the exposed group as the target population [11], for an individual with $Z = z_i$.

Total group as the target population

We assume that the propensity score p_i is explained by the following logistic regression model with a random coefficient:

$$p_i=\frac{\exp\{\log(\gamma_i)+\theta' z_i\}}{1+\exp\{\log(\gamma_i)+\theta' z_i\}}\,,$$

where $\log(\gamma_i)$ is a random coefficient. Then, by the IPW method, $\Pr(Y_{X=x} = 1)$ is estimated as:

$$\Pr(Y_{X=1}=1)=\frac{1}{n}\sum_{i=1}^{n}\frac{y_ix_i}{p_i}\,, \tag{4}$$

$$\Pr(Y_{X=0}=1)=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i(1-x_i)}{1-p_i}\,. \tag{5}$$

In the framework of MSMs, the causal effects are estimated using the weighted regression models of X on Y with the weights $1 / p_i$ for the exposed subjects and $1 / (1 - p_i)$ for the unexposed subjects.

By a calculation similar to those in the above subsections, Equation (4) can be expressed as:

$$\Pr(Y_{X=1}=1)=\sum_{i=1}^{n}\begin{bmatrix}\{\Pr(X=0\,|\,Z=z_i)/\gamma_i+\Pr(X=1\,|\,Z=z_i)\}\\ \times\Pr(Y=1\,|\,X=1,Z=z_i)\Pr(Z=z_i)\end{bmatrix}.$$

Because $\Pr(Y_{X=1}=1)=\sum_{i=1}^{n}\Pr(Y_{X=1}=1\,|\,Z=z_i)\Pr(Z=z_i)$, the left- and right-hand sides of this equation are equal when:

$$\begin{aligned}&\Pr(Y_{X=1}=1\,|\,Z=z_i)\\&=\begin{bmatrix}\{\Pr(X=0\,|\,Z=z_i)/\gamma_i+\Pr(X=1\,|\,Z=z_i)\}\\ \times\Pr(Y=1\,|\,X=1,Z=z_i)\end{bmatrix},\end{aligned}$$

which derives:

$$\gamma_i = \alpha_i = \frac{\Pr(Y_{X=1} = 1 \mid X = 1, Z = z_i)}{\Pr(Y_{X=1} = 1 \mid X = 0, Z = z_i)}.$$

Likewise, Equation (5) derives:

$$\gamma_i = \beta_i = \frac{\Pr(Y_{X=0} = 1 \mid X = 1, Z = z_i)}{\Pr(Y_{X=0} = 1 \mid X = 0, Z = z_i)},$$

because Equation (5) can be expressed as:

$$\Pr(Y_{X=0} = 1) = \sum_{i=1}^{n} \left[\begin{array}{l} \{\Pr(X = 0 \mid Z = z_i) + \gamma_i \Pr(X = 1 \mid Z = z_i)\} \\ \times \Pr(Y = 1 \mid X = 0, Z = z_i)\Pr(Z = z_i) \end{array} \right].$$

This observation shows that γ_i cannot be interpreted until $\alpha_i = \beta_i$ holds (*i.e.*, the two CRRs with the exposed and unexposed groups as the target population are equal). When $\alpha_i = \beta_i$ holds, γ_i can be interpreted as the CRR with the total group as the target population, because this CRR is expressed as:

$$\frac{\Pr(Y = 1 \mid X = 1, Z = z_i)}{\Pr(Y = 1 \mid X = 0, Z = z_i)} \Bigg/ \frac{\Pr(Y_{X=1} = 1 \mid Z = z_i)}{\Pr(Y_{X=0} = 1 \mid Z = z_i)} = \frac{\Pr(X = 0 \mid Z = z_i) + \beta_i \Pr(X = 1 \mid Z = z_i)}{\dfrac{\Pr(X = 0 \mid Z = z_i)}{\alpha_i} + \Pr(X = 1 \mid Z = z_i)}, \qquad (6)$$

and then γ_i is equal to the CRR when $\alpha_i = \beta_i$ $(= \gamma_i)$. The derivation of Equation (6) is given in the Appendix.

CONCLUSION

Based on the formulas of the random coefficient model and the IPW approach, we have given an interpretation of a random coefficient when logistic regression with a random coefficient is used for predicting propensity scores. In conclusion, when the exposed or unexposed group is the target population, the random coefficient can be interpreted as the logarithm of CRR with its group as the target population. When the total group is the target population, however, the random coefficient cannot be interpreted in a straightforward manner. The random coefficient can be interpreted as the CRR only when the two CRRs with the exposed and unexposed groups as the target population are equal.

Although we have given an interpretation of a random coefficient in a logistic regression model, we have not discussed the predicted values of propensity scores or the estimates of causal effects themselves. We will need to research their characteristics through, for example, simulation studies.

APPENDIX: DERIVATION OF EQUATION (6)

In the stratum with $Z = z_i$, we let $\mathrm{RR}_{Ci} = \Pr(Y = 1 \mid X = 1, Z = z_i) / \Pr(Y = 1 \mid X = 0, Z = z_i)$ denote the crude RR, $\mathrm{RR}_{Ei} = \Pr(Y_{X=1} = 1 \mid X = 1, Z = z_i) / \Pr(Y_{X=0} = 1 \mid X = 1, Z = z_i)$ denote the causal RR with the exposed group as the target population, and $\mathrm{RR}_{Ui} = \Pr(Y_{X=1} = 1 \mid X = 0, Z = z_i) / \Pr(Y_{X=0} = 1 \mid X = 0, Z = z_i)$ denote the causal RR with the unexposed group as the target population. Then, the CRR with the total group as the target population can be expressed as:

$$\frac{\Pr(Y = 1 \mid X = 1, Z = z_i)}{\Pr(Y = 1 \mid X = 0, Z = z_i)} \Bigg/ \frac{\Pr(Y_{X=1} = 1 \mid Z = z_i)}{\Pr(Y_{X=0} = 1 \mid Z = z_i)}$$

$$= \frac{\sum_{x=0}^{1} \Pr(Y_{X=0} = 1 \mid X = x, Z = z_i)\Pr(X = x \mid Z = z_i)}{\sum_{x=0}^{1} \Pr(Y_{X=1} = 1 \mid X = x, Z = z_i)\Pr(X = x \mid Z = z_i)} \mathrm{RR}_{Ci}$$

$$= \frac{\left\{ \begin{array}{l} \Pr(Y_{X=1} = 1 \mid X = 1, Z = z_i) \\ \times \sum_{x=0}^{1} \dfrac{\Pr(Y_{X=0} = 1 \mid X = x, Z = z_i)}{\Pr(Y_{X=1} = 1 \mid X = 1, Z = z_i)} \Pr(X = x \mid Z = z_i) \end{array} \right\}}{\left\{ \begin{array}{l} \Pr(Y_{X=0} = 1 \mid X = 0, Z = z_i) \\ \times \sum_{x=0}^{1} \dfrac{\Pr(Y_{X=1} = 1 \mid X = x, Z = z_i)}{\Pr(Y_{X=0} = 1 \mid X = 0, Z = z_i)} \Pr(X = x \mid Z = z_i) \end{array} \right\}} \mathrm{RR}_{Ci}$$

$$= \frac{\left\{ \dfrac{\Pr(X = 0 \mid Z = z_i)}{\mathrm{RR}_{Ci}} + \dfrac{\Pr(X = 1 \mid Z = z_i)}{\mathrm{RR}_{Ei}} \right\} \mathrm{RR}_{Ci}}{\dfrac{\Pr(X = 0 \mid Z = z_i)\mathrm{RR}_{Ui} + \Pr(X = 1 \mid Z = z_i)\mathrm{RR}_{Ci}}{\mathrm{RR}_{Ci}}}$$

$$= \frac{\Pr(X = 0 \mid Z = z_i) + \beta_i \Pr(X = 1 \mid Z = z_i)}{\dfrac{\Pr(X = 0 \mid Z = z_i)}{\alpha_i} + \Pr(X = 1 \mid Z = z_i)}.$$

ACKNOWLEDGEMENTS

This work was supported partially by Grant-in-Aid for Scientific Research (No. 23700344) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

REFERENCES

1. Larsen K, Petersen JH, Budtz-Jørgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. Biometrics. 2000; 56: 909-914.
2. Greenland S. When should epidemiologic regressions use random coefficients? Biometrics. 2000; 56: 915-921.
3. Gustafson P, Greenland S. The performance of random coefficient regression in accounting for residual confounding. Biometrics. 2006; 62: 760-768.
4. Robins JM. Association, causation, and marginal structural models. Synthese. 1990; 121: 151-179.
5. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000; 11: 561-570.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70: 41-55.
7. Joffe MM, Rosenbaum PR. Propensity scores. Am J Epidemiol. 1999; 150: 327-333.
8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66: 688-701.
9. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003; 14: 680-686.
10. Chiba Y. A simple method for sensitivity analysis of unmeasured confounding. J Biomet Biostat. 2012; 3: e113.
11. Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Ann Epidemiol. 2008; 18: 637-646.

Cite this article

Chiba Y (2013) A Note on the Logistic Regression Model with a Random Coefficient to Predict Propensity Scores. Ann Biom Biostat 1: 1001.