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

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 \( \Pr(Y_{x=1}) \) and \( \Pr(Y_{x=0}) \) 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 \( \Pr(Y_{x=1} | X = x) \) and \( \Pr(Y_{x=0} | X = x) \).

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

\[
\Pr(X = 1 | Z = z) = \frac{\exp(\beta z)}{1 + \exp(\beta z)},
\]

where \( \beta \) 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 \( p_0 \) is explained by the following logistic regression model with a random coefficient:

\[
p_0 = \frac{\exp(\alpha + \beta z)}{1 + \exp(\alpha + \beta z)},
\]

where \( \log(\alpha) \) is a random coefficient. Using Equation (1), \( p_0 \) can be expressed as:

\[
p_0 = \frac{\Pr(X = 1 | Z = z)}{\Pr(X = 0 | Z = z) / \alpha + \Pr(X = 1 | Z = z)}.
\]

Using the inverse-probability-weighting (IPW) method,
\[ \Pr(Y_{x=1}|X=0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{1-p_{u_i}}{p_{u_i}} y_{x_i}, \] 
\[ \Pr(Y_{x=0}|X=0) = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{x_i}(1-x_i), \]
where \( n_0 = n \Pr(X=0) \) \([9,10]\). In the framework of MSMs, the causal effects are estimated using the weighted regression analyses of \( X \) on \( Y \) with the weights \( 1 / p_{u_i} \) for exposed subjects and \( 1 / (1 - p_{u_i}) \) for unexposed subjects.

Algebra similar to the above subsection yields:
\[ \Pr(Y_{x=1}|X=1) = \sum_{i=1}^{n} \beta_i \Pr(Y=1|X=0,Z=z_i) \Pr(Z=z_i|X=1). \]
Because
\[ \Pr(Y_{x=0}|X=1) = \sum_{i=1}^{n} \Pr(Y_{x=0}=1|X=1,Z=z_i) \Pr(Z=z_i|X=1), \]
\( \beta_i \) can be expressed as:
\[ \beta_i = \frac{\Pr(Y_{x=0}=1|X=1,Z=z_i) \Pr(Y_{x=1}=1|X=0,Z=z_i)}{\Pr(Y_{x=1}=1|X=1,Z=z_i)/\Pr(Y_{x=1}=1|X=0,Z=z_i)}. \]
This \( \beta_i \) 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 \( \log(\gamma) \) is a random coefficient. Then, by the IPW method, \( \Pr(Y_{x=1}) \) is estimated as:
\[ \Pr(Y_{x=1}) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_{x_i}}{p_i}, \] 
\[ \Pr(Y_{x=0}) = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_{x_i}(1-x_i)}{1-p_i}. \]
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:
\[ \Pr(Y_{x=1}) = \sum_{i=1}^{n} \Pr(X=0|Z=z_i) [\gamma_i + \Pr(X=1|Z=z_i)] \Pr(Z=z_i). \]
Because \( \Pr(Y_{x=1}) = \sum_{i=1}^{n} \Pr(Y_{x=1}=1|Z=z_i) \Pr(Z=z_i) \), the left- and right-hand sides of this equation are equal when:
\[ \Pr(Y_{x=1}=1|Z=z_i) = \left[ \frac{\Pr(X=0|Z=z_i) + \Pr(X=1|Z=z_i)}{\gamma_i + \Pr(X=1|Z=z_i)} \right], \]
which derives:
In the stratum with \(Z = z_i\), we let \(\text{RR}_{Y=1} = \text{Pr}(Y=1 \mid X=1, Z = z_i) / \text{Pr}(Y=1 \mid X=0, Z = z_i)\) denote the crude RR, \(\text{RR}_{Y=0} = \text{Pr}(Y=0 \mid X=1, Z = z_i) / \text{Pr}(Y=0 \mid X=0, Z = z_i)\) denote the crude RR with the exposed group as the target population, and \(\text{RR}_{Y=1} = \text{Pr}(Y=1 \mid X=0, Z = z_i) / \text{Pr}(Y=0 \mid X=1, Z = z_i)\) denote the crude RR with the unexposed group as the target population. Then, the CRR with the total group as the target population can be expressed as:

\[
\gamma_i = \frac{\sum_{x \in \{0, 1\}} \text{Pr}(Y=1 \mid X=x, Z = z_i) \text{Pr}(X=x \mid Z = z_i)}{\sum_{x \in \{0, 1\}} \text{Pr}(Y=0 \mid X=x, Z = z_i) \text{Pr}(X=x \mid Z = z_i)} \times \frac{\text{Pr}(X=1 \mid Z = z_i)}{\text{Pr}(X=0 \mid Z = z_i)} \times \frac{\text{RR}_{Y=1}}{\text{RR}_{Y=0}}.
\]

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