Estimation of propensity score using spatial logistic regression

Hilwin Nisa¹,¹ Corresp., Maria B T Mitakda¹ and Suci Astutik¹
¹Dept. of Statistics, Brawijaya University, Malang

Abstract. Propensity score is a method used to reduce bias due to confounding factors in the estimation of the treatment impact on observational data. Propensity score is the conditional probability to get certain treatments involving the observed covariates. In general, propensity score can be calculated using two methods, they are logistic regression and Classification and Regression Tree Analysis (CART). Logistic regression model is the most common method used. In which, logistic regression model is a model used to estimate the probability of an event. In other side, collecting data by observing many subjects in different place will be influenced spatial effect. Thus, this paper will estimate propensity score using spatial logistic regression.

Keywords: propensity score, logistic regression, spatial logistic regression.

1. Introduction

Propensity Score is a method that can be used to reduce bias in research when examples cannot be taken randomly. Propensity scores can be used to evaluate the effects of treatment given in observational and non-experimental studies [1].

There are several methods used in estimating the propensity score. In general, the methods are logistic regression and Classification and Regression Tree Analysis (CART). To predict the probability of an event, we can use logistic regression methods.

In observation research, the data generated can be cross section data. The cross section data involves several locations. Therefore, it is possible that spatial effect influences the model. In the presence of spatial effects, the usual logistic regression model is not enough to model the data. The spatial logistic regression model will be better to model data that contains spatial effects.

Estimating parameters in the spatial logistic regression model can be estimated using maximum likelihood. However, estimating parameters using maximum likelihood has not produced predictive parameter values that are closed form. So, iteration is needed by using numerical methods. This paper will estimate propensity score using spatial logistic regression.

2. Bernoulli Distribution

If density function \( g(z) \) of a discrete random variable \( Z \) with support \( 0, 1, 2, ..., n \) is

\[
g(z) = \binom{n}{z} p^z (1 - p)^{n-z},
\]

then \( Z \) is said to be distributed according to the binomial distribution. The equation (1) can be expressed as \( Z \sim B(n, p) \).

For \( n = 1 \), the binomial distribution describes a trial which is distributed according to bernoulli distribution with density function
\[ g(y) = p^y(1 - p)^{1-y} = \begin{cases} 1 - p, & \text{for } y = 0 \\ p, & \text{for } y = 1 \end{cases} \]  

The notation \( Y \sim \text{Bern}(p) \) or \( Y \sim \text{B}(1,p) \) is short for the above expression. Where, \( p \) is probability of success in each trial [3].

3. Logistic Regression Model

Logistic regression is a mathematical modelling approach to describe the relationship of several predictors to a dichotomous response. The logistic function \( f(z) \) is:

\[ f(z) = \frac{1}{1 + e^{-z}} \]  

where \( z \) varies from \(-\infty \) to \(+\infty \). Logistic function is designed to describe a probability, so that logistic function ranges between 0 and 1.

Logistic function is used to obtain logistic model. Define \( z \) as:

\[ z = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_kX_k \]  

where \( X \)'s are predictor variables and \( \beta_0 \) and \( \beta_i \) are constant terms representing unknown parameters. Based on (1) and (2), we can write:

\[ f(z) = \frac{1}{1 + \exp(-\beta_0 + \sum \beta_iX_i)} \]  

To estimate the parameters in logistic model, we can use maximum likelihood [4].

In a general case with \( n \) trials and \( y \) successes, the likelihood function is, [3]:

\[ L(\theta) = p(Y = y) = \left( \frac{n}{y} \right) \theta^y (1 - \theta)^{n-y}. \]  

Then, the log-likelihood function is:

\[ \log L(\theta) = y \log \theta + (n - y) \log(1 - \theta) + \text{const} \]

where \( \text{const} \) indicates a term that does not depend on \( \theta \). We can use \( \frac{\partial \log L(\theta)}{\partial \theta} = 0 \) to see that maximum likelihood estimate (MLE) for \( \theta \) is \( \hat{\theta}(y) = \frac{y}{n} \).

The likelihood function for bernoulli trial is [5]:

\[ L(Y, X; \beta) = \prod_{i=1}^{n} [P(Y = 1|X = x)]^{y_i} [1 - P(Y = 1|X = x)]^{1-y_i}. \]

The MLEs of \( \beta \) satisfies the usual consistency and asymptotic normality properties. Since first derivative of \( \ln \text{ likelihood} \) function is nonlinear in \( \hat{\beta} \) and has no simple analytical solution for \( \hat{\beta} \), the optimization of \( \ln \text{likelihood} \) function with respect to the unknown function \( \beta \) requires iterative techniques.

The equation,

\[ \pi_i = \frac{\exp(\sum_{j=1}^{n} \beta_jX_{ij})}{1 + \exp(\sum_{j=1}^{n} \beta_jX_{ij})} \]  

or

\[ \pi_i = \frac{1}{1 + \exp(-\sum_{j=1}^{n} \beta_jX_{ij})} \]

is strictly increasing function of \( \beta_j \).
4. Spatial Binary Regression Models

Regression model for an observed dichotomous response $Y_i$ is the latent response model with dependent variable the continuous variable $Y_i^*$, where [6]:

$$Y_i = \begin{cases} 1, & Y_i^* \geq 0 \\ 0, & \text{else} \end{cases}$$

with $i = 1, 2, ..., n$.

The linear model which is specified for this latent response is (Calabrese & Elkink, 2013):

$$Y^* = \rho W Y^* + X\beta + u$$
$$u = \lambda W u + \epsilon$$

where:
- $\epsilon$ is the error term which can follow a multivariate normal distribution in a probit model or a multivariate logistic distribution in a logit model.
- $Y^*$ is a continuous random vector $(n \times 1)$.
- $X$ is an $(n \times k)$ matrix of explanatory variables.
- $\beta$ is a $(k+1) \times 1$ vector of regression parameter coefficient.
- $\rho$ is parameter coefficient of spatial lag.
- $\lambda$ is parameter coefficient of spatial error.
- $u$ is an $(n \times 1)$ vector of error term in regression which has spatial effect.
- $W$ is an $(n \times n)$ spatial weights matrix.
- $n$ is the number of cross-section unit.

Anselin (2002) said that the latent variable only can be used for the spatial lag [8]. This is because of both the models $Y^* = \rho W Y^* + X\beta + \epsilon$ and $Y = \rho W Y + X\beta + \epsilon$ are infeasible.

Spatial lag model in binary variable is called the Binary Spatial AutoRegressive model (BSAR) (Calabrese & Elkink, 2013):

$$y^* = (I - \rho W)^{-1}(X\beta + \epsilon) = (I - \rho W)^{-1}X\beta + \epsilon = HX\beta + \epsilon$$

where

$H = (I - \rho W)^{-1}$ and $\epsilon = (I - \rho W)^{-1}\epsilon$.

5. Spatial Logistic Regression Model

Spatial logistic regression model is obtained by logistic regression model and spatial binary regression model. The model of spatial logistic regression is

$$y^* = \rho W y^* + X\beta + \epsilon; \epsilon \sim N(0, \sigma^2 I_n)$$

where
- $y^*$ is an $(n \times 1)$ vector of latent variable.
- $X$ is an $n \times (k+1)$ matrix of explanatory variables.
- $\beta$ is a $(k+1) \times 1$ vector of parameter coefficient.
- $W$ is an $(n \times n)$ spatial weights matrix.
- $\rho$ is parameter of spatial lag.

The model (16) can be written as

$$y^* = (I - \rho W)^{-1}(X\beta + \epsilon)$$

$$= (I - \rho W)^{-1}X\beta + (I - \rho W)^{-1}\epsilon$$

$$= HX\beta + \epsilon, \epsilon \sim MVN(0, \Omega)$$

with $\epsilon = (I - \rho W)^{-1}\epsilon$, where $\epsilon$ is an $(n \times 1)$ vector and $H = (I - \rho W)^{-1}$, where $H$ is an $(n \times n)$ matrix.

Latent variable $y^*$ has binary category which is defined as variable $y$:

$$y_i = \begin{cases} 1, & \text{for } y_i^* > 0 \\ 0, & \text{for } y^* \leq 0 \end{cases}$$
The probability for \( P(\gamma_i = 1) \) and \( P(\gamma_i = 0) \) is:
\[
P(\gamma_i = 1) = P(\gamma_i^* > 0) \\
= P([HX\beta]^\top + e > 0) \\
= P(-e \leq [HX\beta]^\top) \\
= \frac{1}{1 + \exp(-[HX\beta]^\top)}
\]
\[
P(\gamma_i = 0) = P(\gamma_i^* \leq 0) \\
= P([HX\beta]^\top + e \leq 0) \\
= P(-e > [HX\beta]^\top) \\
= 1 - P(-e \leq [HX\beta]^\top) \\
= 1 - \frac{1}{1 + \exp(-[HX\beta]^\top)}
\]
(19)
(20)

Based on (14) and (15), we get the limit of \( \gamma_i \):
\[
\gamma_i = \begin{cases} 
1, & \text{for } -e_i \leq [HX\beta]^\top_i \\
0, & \text{for } -e_i > [HX\beta]^\top_i
\end{cases}
\]
(21)

Since \( e \) is assumed to be distributed normal multivariate with 0 means and \( \Omega \) variance, the probability for \( P(\gamma_i = 0) \) and \( P(\gamma_i = 1) \) is:
\[
P(\gamma_i = 1) = \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)}
\]
(22)
\[
P(\gamma_i = 0) = 1 - \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)}
\]
(23)

where \( \Omega_{ii} \) mis element of diagonal \( \Omega \) which is formed as:
\[
\Omega = [(I - \rho W)'(I - \rho W)]^{-1}.
\]
(24)

6. Estimation of Spatial Logistic Regression Models

Estimation of spatial logistic regression parameters can be obtained by Maximum Likelihood Estimation (MLE). In this method, the parameter is estimated by maximizing likelihood function of random variable \( y_i \) which follow bernoulli distribution (1, p):
\[
L(\beta, \rho) = \prod_{i=1}^{n} \left( \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)} \right)^{y_i} \left( 1 - \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)} \right)^{1-y_i}
\]

Then, the likelihood function above is transformed by natural log (ln):
\[
\ln[L(\beta, \rho)] = \sum_{i=1}^{n} y_i \ln \left( \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)} \right) \\
+ \sum_{i=1}^{n} (1-y_i) \ln \left( 1 - \frac{1}{1 + \exp\left(-\frac{[HX\beta]^\top}{\Omega_{ii}}\right)} \right)
\]
\[
= \beta'Xy - \sum_{i=1}^{n} \ln \left( 1 + \exp\left(\frac{[K\beta]^\top}{\Omega_{ii}}\right) \right)
\]
(25)

Then, define the complete form of (25)
\[
\ln[L_c(\beta, \rho)] = \beta'Xy - \sum_{i=1}^{n} \ln \left[ 1 + \exp \left( \frac{\beta'Xi}{\Omega_{ii}} \right) \right] + \beta'Xy 
\]  
\[
- \sum_{i=n+1}^{m} \ln \left[ 1 + \exp \left( \frac{\beta'Xi}{\Omega_{ii}} \right) \right].
\]  

To obtain \( \beta \) estimated, find derivation of (25) to \( \beta \). Since the first derivation is not closed form, we can solve it by some methods. In this paper, we use Expectation-Maximization (EM) Algorithm to solve it.

Using an initial value of \( \beta \), for example \( \beta^{(0)} \), we will calculate \( H(\beta, \beta^{(0)}) = E\{\log L_c(\beta)|X, \beta^{(0)}\} \), an expectation of (26). This step replaces each of observed responses by \( y_j, j = n+1, \ldots, n+m \), using following conditional expectation (Anderson & Hardin, 2013):

\[
E(Y_j|X_j, \beta^{(0)}) = \frac{1}{1 + \exp \left( -\frac{\beta'X_j}{\Omega_{ii}} \right)}
\]  

To find the value of \( \beta \), that maximizes (26), we will use following iteratively reweighted least squares (IRLS):

1) Choose an initial estimation of the regression coefficients, like \( \beta \).

2) At each iteration \( k \), update the regression coefficients:

\[
\beta^{(k)} = \beta^{(k-1)} + (X'V^{(k-1)}X)^{-1}X'(y - p^{(k-1)})
\]

where \( X \) is a matrix of predictors, \( y \) is the observed response vector, \( p^{(k-1)} \) is the vector of fitted response probabilities for the previous iteration, which is

\[
p^{(k-1)} = \frac{1}{1 + \exp(-x^{(k-1)}\beta)}
\]

and \( V^{(k-1)} \) is a diagonal matrix, with diagonal entries \( p^{(k-1)}(1 - p^{(k-1)}) \).

3) Repeat step 2 until \( |\beta^{(k)} - \beta^{(k-1)}| \rightarrow 0 \).

\[
n \rightarrow \infty.
\]

7. Propensity Score Using Spatial Logistic Regression

The model of propensity score is [10]:

\[
e(X_i) = P(z_i = 1|X_i)
\]

where \( x_i = \) vector of observed covariate for subject-\( i \)

\[z_i = \{ \begin{array}{ll} 0, & \text{control} \\ 1, & \text{treatment} \end{array} \]

Since using spatial logistic regression, based on (22), propensity score using spatial logistic regression can be written as

\[
e(X_i) = \frac{1}{1 + \exp \left( -\frac{X'\beta}{\Omega_{ii}} \right)}
\]

where \( \beta^{(k)} = \beta^{(k-1)} + (X'V^{(k-1)}X)^{-1}X'(y - p^{(k-1)}) \).

8. Conclusion and Recommendation

In this paper, we know that estimation of propensity score using spatial logistic regression by maximum likelihood estimation in not closed form. So that, we use EM algorithm to solve it. Then, we find the
estimator of parameter $\beta^{(k)} = \beta^{(k-1)} + (X'V^{(k-1)}X)^{-1}X'(y - p^{(k-1)})$ and the estimation of propensity score using spatial logistic regression is

$$e(X_i) = \frac{1}{1 + \exp\left(-\frac{[HX\beta]_i}{\Omega_{ii}}\right)}$$

References

[1] Guo, S. & Fraser, M. W.. 2010. Propensity Score Analysis Statistical Methods and Applications. SAGE. United States of America.

[2] Calabrese, R. & Elkink, J. A.. 2014. “Estimators of Binary Spatial Autoregressive Models: a Monte Carlo Study.” Journal of Regional Science. 54 (4): 664-687.

[3] Madsen, H. & Thyregod, P.. 2010. Introduction to General and Generalized Linear Models. CRC Press. New York.

[4] Kleinbaum, D. G. & Klein, M.. 2002. Logistic Regression: A Self-Learning Text, Second Edition. Springer. New York.

[5] Antony, N., Otieno, R. O., & Mwita, P. N.. 2014. “On the Estimation and Properties of Logistic Regression Parameters.” IOSR Journal of Mathematics (IOSR-JM). 10 (4): 57-68.

[6] Verbeek, M. 2004. A Guide to Modern Econometrics 2nd Edition. John Wiley & Sons, Ltd. England.

[7] Calabrese, R. & Elkink, J. A.. 2014. “Estimating Binary Spatial Autoregressive Models for Rare Events.” Presented at the 2015 Asian Political Methodology Conference.

[8] Anselin, L. 2002. “Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models.” Agricultural Economics. 27 (3): 247-267.

[9] Anderson, B. S. & Hardin, J. M.. 2013. “Modified Logistic Regression Using the EM Algorithm for Reject Inference.” Accessed on 26 March 2019, at https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/Paper25-Paper.pdf

[10] Thavaneswaran, A. & Lix, L.. 2008. Propensity Score Matching in Observational Studies. University of Manitoba. Kanada.