Web Material

Matched Versus Unmatched Analysis of Matched Case-Control Studies

Fei Wan, Graham A. Colditz, and Siobhan Sutcliffe

Contents

Web Appendix 1: Derivation of logistic regression model in the matched case-control data.

For the simplicity of demonstration, we assume the confounder terms in the outcome model (1) and the exposure model (2) have linear main effect terms only. That is, $f(\boldsymbol{X}) = \beta_2'\boldsymbol{X}$, $g(\boldsymbol{X}) = \alpha_1'\boldsymbol{X}$

1.1) When the disease outcome is rare event

In a matched case control study, we match the cases and the same number of controls using confounder X. We let $x_1, x_2, \ldots, x_k$ be the unique matching values of $\boldsymbol{X}$. We assume there are n_{1j}cases and n_{0j} controls having the same value of x_j $(j = 1,2, \ldots, k)$. For simplicity, we assume $n_{1j} \ll n_{0j}$ for a rare disease outcome. Within this matching stratum, we let S denote the selection process with $S = 1$ for a subject being selected into the matched case control data and $S = 0$ for not being selected. Then, we can derive the conditional probability of having a disease outcome for a subject selected into the matched set.

$$P(Y = 1|\boldsymbol{X}, E, S = 1) = \frac{P(Y = 1, E, \boldsymbol{X} = x_j, S = 1)}{P(E, X = x_j, S = 1)}$$

$$= \frac{P(Y = 1, E, \boldsymbol{X} = x_j, S = 1)}{P\big(Y = 1, E, \boldsymbol{X} = x_j, S = 1\big) + P\big(Y = 0, E, \boldsymbol{X} = x_j, S = 1\big)}$$

$$= \frac{1}{1 + \dfrac{P\big(Y = 0, E, \boldsymbol{X} = x_j, S = 1\big)}{P(Y = 1, E, \boldsymbol{X} = x_j, S = 1)}}$$

$$= \frac{1}{1 + \frac{P(S=1|Y=0,E,\boldsymbol{X}=x_j)P(Y=0|E,\boldsymbol{X}=x_j)}{P(S=1|Y=1,E,\boldsymbol{X}=x_j)P(Y=1|E,\boldsymbol{X}=x_j)}}$$

$$= \frac{1}{1 + \frac{P(S=1|Y=0,E,\boldsymbol{X}=x_j)}{P(S=1|Y=1,E,\boldsymbol{X}=x_j)} e^{-\beta_0-\beta_1 E-\beta_2' x_j}}$$

Within the stratum formed by x_j, we will select all cases (selection probability=1) and select the equal number of controls. Of note, this selection probability does not depend on individual's exposure status. Thus, we have

$$P(S=1|Y=1,E,\boldsymbol{X}=x_j) = P(S=1|Y=1,\boldsymbol{X}=x_j) = 1$$

and

$$P(S=1|Y=0,E,\boldsymbol{X}=x_j) = P(S=1|Y=0,\boldsymbol{X}=x_j)$$

$$= \frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}$$

The last equation holds because $P(S=1|Y=0,\boldsymbol{X}=x_j)$ is estimated by $\frac{n_{1j}}{n_{0j}}$, which is equivalent to $\frac{n_{1j}}{n_{0j}+n_{1j}} / \frac{n_{0j}}{n_{0j}+n_{1j}}$. The proportion of cases in the stratum $\frac{n_{1j}}{n_{0j}+n_{1j}}$ converges to $P(Y=1|\boldsymbol{X}=x_j)$ and the proportion of controls $\frac{n_{0j}}{n_{0j}+n_{1j}}$ converges to $P(Y=0|\boldsymbol{X}=x_j)$ asymptotically as the sample size in this stratum increases. The other way is

$$\frac{P(S=1|Y=0,E,\boldsymbol{X}=x_j)}{P(S=1|Y=1,E,\boldsymbol{X}=x_j)} = \frac{P(S=1|Y=0,\boldsymbol{X}=x_j)}{P(S=1|Y=1,\boldsymbol{X}=x_j)}$$

$$= \frac{P(Y=0|S=1,\boldsymbol{X}=x_j)P(S=1,\boldsymbol{X}=x_j)/P(Y=0,\boldsymbol{X}=x_j)}{P(Y=1|S=1,\boldsymbol{X}=x_j)P(S=1,\boldsymbol{X}=x_j)/P(Y=1,\boldsymbol{X}=x_j)}$$

$$= \frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}$$

The last equation holds because $P(Y=0|S=1,\boldsymbol{X}=x_j) = P(Y=1|S=1,\boldsymbol{X}=x_j)$ when we select the same number of cases and controls in each matched stratum.

Next, we have

$$\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)} = \frac{P(Y=1,E=1|\boldsymbol{X}=x_j) + P(Y=1,E=0|\boldsymbol{X}=x_j)}{P(Y=0,E=1|\boldsymbol{X}=x_j) + P(Y=0,E=0|\boldsymbol{X}=x_j)}$$

$$= \frac{P(Y=1|E=1,\boldsymbol{X}=x_j)P(E=1|\boldsymbol{X}=x_j) + P(Y=1|E=0,\boldsymbol{X}=x_j)P(E=0|\boldsymbol{X}=x_j)}{P(Y=0|E=1,\boldsymbol{X}=x_j)P(E=1|\boldsymbol{X}=x_j) + P(Y=0|E=0,\boldsymbol{X}=x_j)P(E=0|\boldsymbol{X}=x_j)}$$

The explicit form of $\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}$ can be derived by substituting the conditional probability of having an event or exposure based on the outcome model (1) and the exposure model (2). If we have rare event, the logistic outcome model (1) can be approximated by a log linear model $e^{\beta_0+\beta_1 E+\beta_2' x_j}$ and then we have the following:

$$\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)} \approx \frac{e^{\beta_1}\frac{e^{\alpha_0+\alpha_1 x_j}}{1+e^{\alpha_0+\alpha_1 x_j}} + \frac{1}{1+e^{\alpha_0+\alpha_1 x_j}}}{\frac{e^{\alpha_0+\alpha_1 x_j}}{1+e^{\alpha_0+\alpha_1 x_j}} + \frac{1}{1+e^{\alpha_0+\alpha_1 x_j}}} e^{\beta_0+\beta_2' x_j}$$

Next, it follows

$$P(Y=1|\boldsymbol{X},E,S=1) = \frac{1}{1+\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}e^{-\beta_0-\beta_1 E-\beta_2' x_j}}$$

$$= \frac{1}{1+e^{c(x_j)-\beta_1 E}}$$

Where $c(x_j) = \log(\frac{e^{\beta_1}-1}{1+e^{-\alpha_0-\alpha_1' x_j}} + 1)$. The last equation shows that the logit model for the matched case control data does not have $\beta_2 x_j$ from the logit model for the source study population. Intuitively it is because of exact matching on X for cases and controls essentially removes the association between X and the outcome. It should be emphasized that even without assuming the disease is uncommon and the logit outcome model (1) cannot be approximated by a log linear model, $\beta_2 x_j$ will still be cancelled out. However, because of over-sampling cases and under-sampling controls (selection bias), we have this additional term $c(x_j)$ of X and $c(x_j)$ consists of conditional probabilities of being exposed and not being exposed. Essentially, if $c(x_j)$ is not approximately linear in , a logit model including E and a linear term of X will be a mis-specified model and fitting this mis-specified model could result in biased estimate of β_1. In this study, $\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}$ is derived based on known population outcome and exposure models.

We also designed a simulation study to validate the derived closed form expression of $c(x_j)$. The detailed simulation algorithm is outlined in the main text. We set $\alpha_0 = 0$, $\alpha_1 = 2$, $\beta_0 = -5$, and $\beta_1 = 1$. Thus, ~56% of subjects are exposed to the exposure and ~4% have disease. The matching variable X is a categorical variable with 4 levels: (-2, -1, 1, 2) that follows a uniform discrete distribution with probabilities of (0.25, 0.25, 0.25, 0.25). Therefore, we have four matching strata formed at each level of X and each stratum will have stratum specific intercept. We generated 10000 subjects and performed 1:1 case-control matching. For each simulated matched case control data set, we fit an unconditional logistic regression including exposure E and three dummy variables: X_1 (1 if $\mathrm{X} = -1, 0$ otherwises), X_2(1 if $\mathrm{X} = 1, 0$ otherwises), and X_3(1 if $\mathrm{X} = 2, 0$ otherwises). We repeated the process 10000 times and averaged the estimates

of intercepts. The results in the Web Table 1 show that derived expression $c(x_j) = \log(\frac{e^{\beta_1 - 1}}{1+e^{-\alpha_0 - \alpha_1' x_j}} + 1)$ give very similar results to the simulated results

Web Table 1: Validating the derived closed form expression of $c(x_j)$

Intercept parameter for matching stratum	True intercept: $-c(x_j)$	Simulated estimates
$X = -2$	-0.6201145	-0.62064563
$X = -1$	-0.7227731	-0.7105566
$X = 1$	-0.5751923	-0.5716599
$X = 2$	-0.7914948	-0.7956336

1.2) <u>When the disease outcome is not rare.</u>

When the outcome is not rare, in the stratum formed by x_j where the number of cases n_{1j} is less than the number of controls n_{0j}, we will select all cases (selection probability=1) and select the equal number of controls. Thus, we have

$$P(S = 1 | Y = 1, E, \boldsymbol{X} = x_j) = 1$$

and

$$P(S = 1 | Y = 0, E, \boldsymbol{X} = x_j)$$

$$= \frac{P(Y = 1 | \boldsymbol{X} = x_j)}{P(Y = 0 | \boldsymbol{X} = x_j)}$$

In the stratum formed by x_j where the number of cases n_{1j} is larger than the number of controls n_{0j}, we will select all controls (selection probability=1) and select the equal number of cases. Thus, we have

$$P\left(S=1\middle|Y=0,E,\boldsymbol{X}=x_j\right)=1$$

and

$$P\left(S=1\middle|Y=1,E,\boldsymbol{X}=x_j\right)=\frac{P(Y=0|\boldsymbol{X}=x_j)}{P(Y=1|\boldsymbol{X}=x_j)}$$

Regardless of the relative size of n_{1j} verse n_{0j}, we always have

$$\frac{P\left(S=1|Y=0,E,\boldsymbol{X}=x_j\right)}{P(S=1|Y=1,E,\boldsymbol{X}=x_j)}=\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}$$

Since the outcome is not rare, the outcome model is logit form $\frac{e^{\beta_0+\beta_1E+\beta_2'x_j}}{1+e^{\beta_0+\beta_1E+\beta_2'x_j}}$, not log-linear form. $e^{\beta_0+\beta_1E+\beta_2x_j}$:

$$\frac{P(Y=1|\boldsymbol{X}=x_j)}{P(Y=0|\boldsymbol{X}=x_j)}=\frac{\frac{e^{\beta_1}}{1+e^{\beta_0+\beta_1+\beta_2'x_j}}\frac{e^{\alpha_0+\alpha_1'x_j}}{1+e^{\alpha_0+\alpha_1'x_j}}+\frac{1}{1+e^{\beta_0+\beta_2'x_j}}\frac{1}{1+e^{\alpha_0+\alpha_1'x_j}}}{\frac{1}{1+e^{\beta_0+\beta_1+\beta_2'x_j}}\frac{e^{\alpha_0+\alpha_1'x_j}}{1+e^{\alpha_0+\alpha_1'x_j}}+\frac{1}{1+e^{\beta_0+\beta_2'x_j}}\frac{1}{1+e^{\alpha_0+\alpha_1'x_j}}}e^{\beta_0+\beta_2'x_j}$$

$$=\left(\frac{\frac{e^{\beta_1}}{1+e^{\beta_0+\beta_1+\beta_2'x_j}}\frac{e^{\alpha_0+\alpha_1'x_j}}{1+e^{\alpha_0+\alpha_1x_j}}-\frac{1}{1+e^{\beta_0+\beta_1+\beta_2'x_j}}\frac{e^{\alpha_0+\alpha_1'x_j}}{1+e^{\alpha_0+\alpha_1'x_j}}}{\frac{1}{1+e^{\beta_0+\beta_1+\beta_2'x_j}}\frac{e^{\alpha_0+\alpha_1'x_j}}{1+e^{\alpha_0+\alpha_1'x_j}}+\frac{1}{1+e^{\beta_0+\beta_2'x_j}}\frac{1}{1+e^{\alpha_0+\alpha_1'x_j}}}+1\right)e^{\beta_0+\beta_2'x_j}$$

$$=\left(\frac{e^{\beta_1}-1}{1+\frac{1+e^{\beta_0+\beta_1+\beta_2'x_j}}{1+e^{\beta_0+\beta_2'x_j}}\frac{1}{1+e^{-\alpha_0-\alpha_1'x_j}}}+1\right)e^{\beta_0+\beta_2'x_j}$$

$$= \left(\frac{e^{\beta_1}-1}{1+(1+\frac{e^{\beta_1}-1}{1+e^{-\beta_0-\beta_2'x_j}})e^{-\alpha_0-\alpha_1'x_j}}+1\right)e^{\beta_0+\beta_2'x_j}$$

Next, it follows

$$P(Y=1|\boldsymbol{X},E,S=1) = \frac{1}{1+\frac{P(Y=1|X=x_j)}{P(Y=0|X=x_j)}e^{-\beta_0-\beta_1E-\beta_2'x_j}}$$

$$= \frac{1}{1+e^{c(x_j)-\beta_1E}}$$

Where the nuisance term $c(x_j) = \left(\frac{e^{\beta_1-1}}{1+(1+\frac{e^{\beta_1-1}}{1+e^{-\beta_0-\beta_2'x_j}})e^{-\alpha_0-\alpha_1'x_j}}+1\right)$.

Thus, whether the outcome is rare or not, the main effect confounding term $\beta_2'x_j$ will be removed from the logit outcome model for the matched sample. When the outcome is not rare, the nuisance term $c(x_j)$ becomes more complex and involves the confounding term from both the population outcome and exposure models. If $f(\boldsymbol{X})$ and $g(\boldsymbol{X})$ take more complex forms (e.g., non-linear terms and interaction terms, the nuisance term $c(x_j)$ will become even more complex.

In this study we revisit the case-control matching design from a perspective of a stratified sampling design and derive the corresponding correct ***ULR*** from the pre-specified population outcome and exposure models. Thus, the complexity of fitting a correct ***ULR*** for matched data can be clearly revealed**.** This same technique has been used by Qian *et al.* [1] to develop a two-stage variable selection procedure and prediction rule for a nested, matched case-control study.

Web Appendix 2: The potential bias of ULR+L in matched case-control data.

We assume $(X) = \rho X + \omega$, where ω is some random error and when we fit a ***ULR*** with linear term of matching variable in the matched case control data, the model (3) becomes

$$P(Y = 1|X, E, S = 1) = \frac{1}{1 + e^{\rho X + \omega - \beta_1 E}}, (4)$$

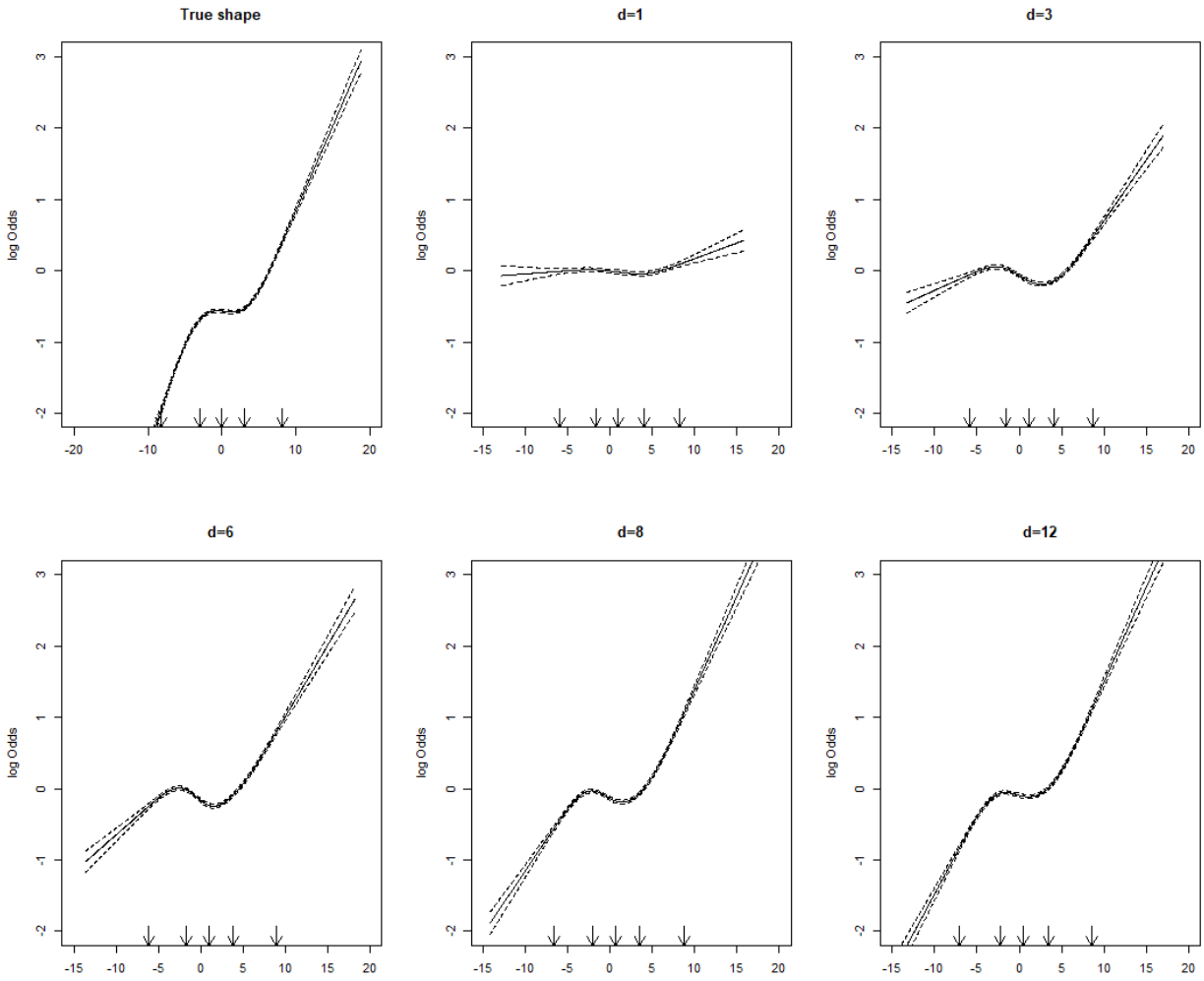
Since we cannot include ω in the model (4), omitting ω will change the coefficients of remaining variables including β_1 because the logit model is not collapsible. Large variance of ω- $var(\omega)$, and large bias we have for estimating β_1 using adjusted ***ULR***. There are two factors could impact the size of $var(\omega)$: variance and non-linearity of $c(X)$. Larger variance of $c(X)$ and more severe deviation from linearity, $var(\omega)$ becomes larger.

Web Appendix 3: The bias of CLR vs. adjusted ULR when matching is not exact

In practice, we often select controls within a certain range of the matching variable values of the cases. Failing to control for this matching variable in ***CLR*** as a covariate in non-exact matched case-control data or failing to control for this matching variable correctly in adjusted ***ULR*** could potentially result in biased estimates of β_1, particularly if the matching variable is a strong confounder. Similar to described above for adjusted ***ULR*** in the exact matched setting, this bias can be interpreted as the omitted variable bias problem in the non-collapsible logit model[14,15]. To show this heuristically, we let P_i denote the ith matched pair indicator ($P_i = 1$ if the participant is in the ith matched pair; 0 otherwise), and let X_{i0} and X_{i1}denote the matching variable for the case ($j=1$) and control (j=0), where $|X_{i1} - X_{i0}| < d$ and d is the caliper size for matching. Thus, we assume that $X_{ij} = kP_i + \varepsilon_{ij}$, where ε_{ij} measures the deviation from the pair mean for the case and control in the ith matched pair and k is a constant. The variability of ε_{ij} is also expected to increase

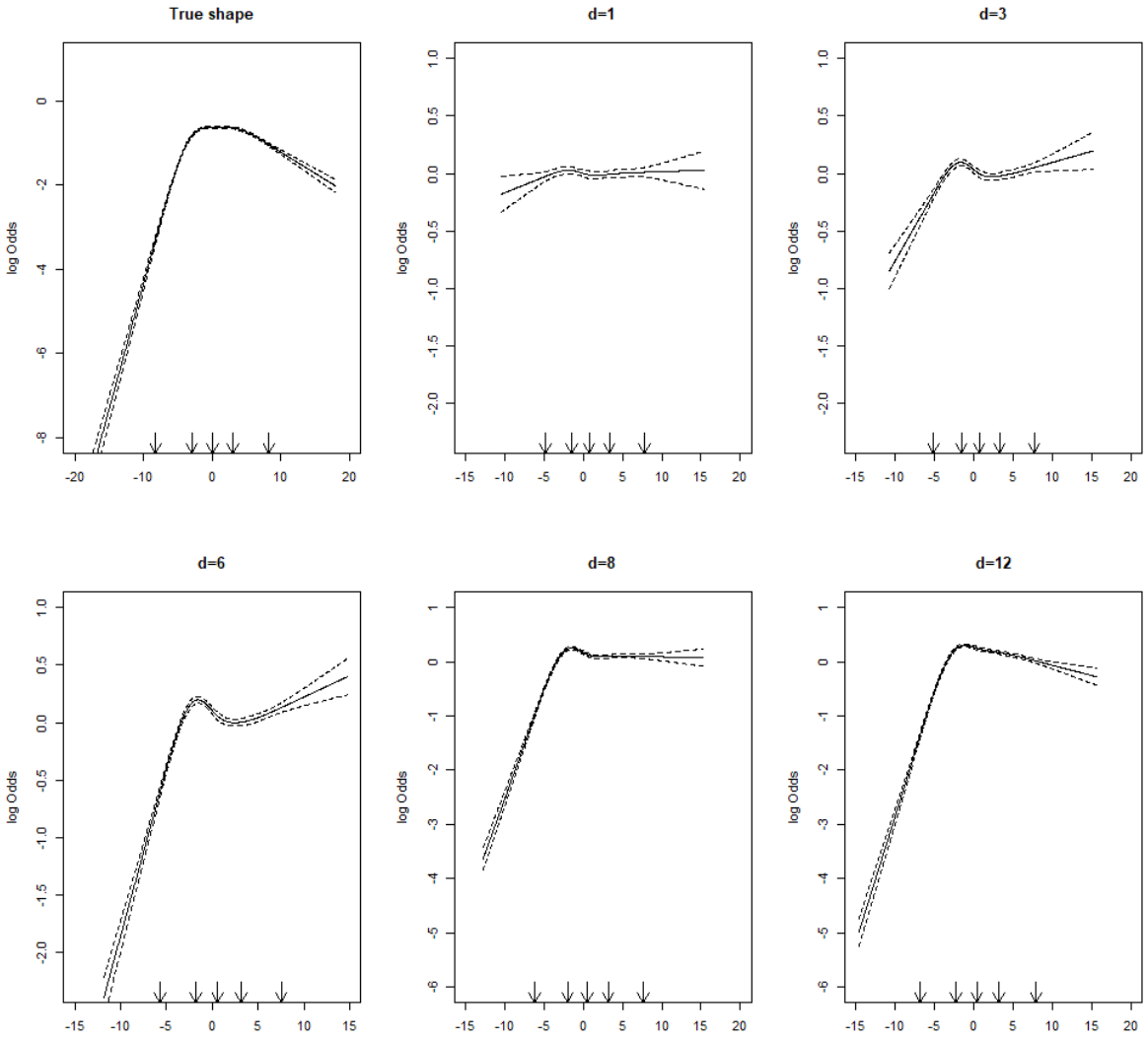
with increasing caliper size. By assuming that cases and controls in the same matched set have the same P_i and by omitting ε_{ij} in the likelihood function, ***CLR*** fails to account for the fact that cases and controls have different matching values and thus different risks of developing the outcome. When matching is not exact, ***CLR*** may also result in a biased estimate because of the non-collapsibility of the odds ratio and the residual confounding arising from not accounting for ε_{ij} in the logit model. The resulting bias is expected to increase with increasing caliper size.

Web Figure 1- The functional form of X_1 in the matched data with changing caliper size in scenario (3)



Web Figure 1: The first figure presents the functional form of the association between X_1 and the outcome in the simulated unmatched source population from the scenario (3) where X_1 in the outcome model has a linear term. The other figures presents the functional forms of the association between X_1 and the outcome in the 1:1 matched samples given different matching calipers. As caliper size increases, the functional form of X_1 in the matched sample is getting closer and closer to the true shape.

Web Figure 2-The functional form of X_1 in the matched data with changing caliper size in scenario (6)



Web Figure 2: The first figure presents the functional form of the association between X_1 and the outcome in the simulated unmatched source population from the scenario (6) where X_1 in the outcome model has a quadratic term. The other figures presents the functional forms of the association between X_1 and the outcome in the 1:1 matched samples given different matching

calipers. As caliper size increases, the functional form of X_1 in the matched sample is getting closer and closer to the true shape.

Reference:

1. Qian J, Payabvash S, Kemmling A, Lev MH, Schwamm LH, Betensky RA. Variable selection and prediction using a nested, matched case-control study: Application to hospital acquired pneumonia in stroke patients. *Biometrics*. 2014;70(1):153-163.