Statistica Sinica Preprint No: SS-2023-0249	
Title	Adjusting for Non-confounding Covariates in Case-control Association Studies
Manuscript ID	SS-2023-0249
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0249
Complete List of Authors	Siliang Zhang, Jinbo Chen, Zhiliang Ying and Hong Zhang
Corresponding Authors	Hong Zhang
E-mails	zhangh@ustc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

Adjusting for non-confounding covariates in case-control association studies

Siliang Zhang[1], Jinbo Chen[2], Zhiliang Ying[3], Hong Zhang[4]

[1]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, School of Statistics, East China Normal University, Shanghai 200062, China*

[2]*Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

[3]*Department of Statistics, Columbia University, New York, NY 10027, USA*

[4]*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China*

Abstract: There is a substantial literature in case-control logistic regression on whether or not non-confounding covariates should be adjusted for. However, only limited and ad hoc theoretical results are available on this important topic. A constrained maximum likelihood method was recently proposed, which appears to be generally more powerful than logistic regression methods with or without adjusting for non-confounding covariates. This note provides a theoretical clarification for the case-control logistic regression with and without covariate

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

adjustment and the constrained maximum likelihood method on their relative performances in terms of asymptotic relative efficiencies. We show that the benefit of covariate adjustment in the case-control logistic regression depends on the disease prevalence. We also show that the constrained maximum likelihood estimator gives an asymptotically uniformly most powerful test.

Key words and phrases: Asymptotic relative efficiency; Case-control design; Constrained maximum likelihood; Logistic regression; Non-confounding covariate.

1. Introduction

It is well known that adjusting for baseline covariates can lead to an improved statistical inference efficiency (Fisher, 1932). Through rigorous derivation, Robinson and Jewell (1991) showed that in logistic regression, adjusting for non-confounding covariates in randomized clinical trials always benefits the testing for treatment effect in terms of Pitman's asymptotic relative efficiency (ARE), albeit with some estimation precision loss. Neuhaus (1998) extended the results of Robinson and Jewell (1991) to generalized linear models.

It is more complicated, however, when data are collected retrospectively but analyzed using prospective logistic regression. In particular, Kuo and Feingold (2010) showed through simulations that adjusting for non-confounding covariates in case-control studies may decrease the statisti-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

cal inference efficiency for exposure-disease association; see Xing and Xing (2010) for additional comments. Pirinen et al. (2012) showed that adjusting for non-confounding covariates in case-control studies can decrease estimation precision, resulting in a loss of power provided that the estimation is approximately unbiased. In general, covariate adjustment could result in both bias and loss of efficiency in estimation, and it is not clear how they affect the hypothesis testing for exposure-disease association.

Zaitlen et al. (2012) proposed a liability threshold model that exploits covariate-specific prevalence information to improve the inference efficiency of exposure-disease association. Zhang et al. (2018) developed a constrained profile maximum likelihood method with known disease prevalence when the exposure and covariate are independent. Their simulation results indicate that the method outperforms the standard logistic regression with or without adjusting for covariates.

In this paper, we derive theoretical properties for the case-control logistic regression methods with and without covariate adjustment, and the constrained maximum likelihood method. Specifically, when both exposure and covariate are dichotomous, we derive the asymptotic biases, asymptotic variances and AREs of the three methods. Furthermore, we obtain the asymptotic distribution of the constrained maximum likelihood estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

tor with a possibly misspecified disease prevalence. Our theoretical findings are at least threefold. First, adjusting for non-confounding covariates can decrease power in case-control studies when the true disease prevalence is low, which extends the finding of Robinson and Jewell (1991) and Neuhaus (1998) to case-control studies. Second, the constrained maximum likelihood method has a uniform power advantage over the other two methods. Third, the constrained maximum likelihood method is robust against prevalence misspecification.

2. Models and methods

Consider a case-control study design involving a binary response variable D ($D = 1$: case; $D = 0$: control), a binary exposure variable of interest E ($E = 1$: exposed; $E = 0$: unexposed) and a binary covariate X ($X = 1$: high risk category; $X = 0$: low risk category). The exposure variable E could be a genetic mutation or an environmental exposure. The covariate X is assumed to be independent of E throughout this paper. Note that spurious association could be produced when X and E are dependent but X is not adjusted for. Let $f = \text{pr}(D = 1)$, $\theta = \text{pr}(X = 1)$ and $\pi = \text{pr}(E = 1)$ be the prevalences of D, X and E, respectively, in the population from which cases and controls are sampled. Throughout this paper, we assume

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

that X and E are non-degenerate so that $0 < \pi, \theta < 1$. Furthermore, we assume that the following logistic regression model holds true:

$$p_{ij}(\alpha, \beta, \gamma) = \text{pr}(D = 1 \mid X = i, E = j) = \frac{\exp(\alpha + \beta i + \gamma j)}{1 + \exp(\alpha + \beta i + \gamma j)}, \quad (2.1)$$

where α is the baseline log-relative risk, and β and γ are log odds ratios. Note that the above model does not involve E-X interaction term due to the assumption that the E-D odds ratio does not depend on X. We are interested in testing the null hypothesis of no association between D and E with the adjustment of X, i.e., $H_0 : \gamma = 0$. Under the case-control study design, data for (E, X) are randomly sampled from case population $(D = 1)$ and control population $(D = 0)$. Let n_{dij} denote the number of subjects with $D = d$, $X = i$ and $E = j$. The total numbers of cases and controls are denoted by n_{1++} and n_{0++}, respectively, and let $\nu = n_{1++}/n_{0++}$.

This paper provides a theoretical clarification on the AREs for three methods. The first method, henceforth referred to as "ADJ", fits model (2.1) to the case-control data by adjusting for X as if the data were prospectively collected. The corresponding estimator of γ, $\hat{\gamma}_A$, is the maximizer of the prospective likelihood function, which is consistent, asymptotically normally distributed, and semiparametric efficient (Anderson, 1972; Prentice and Pyke, 1979; Breslow et al., 2000). Robinson and Jewell (1991) derived a closed-form estimator of the asymptotic variance of $\hat{\gamma}_A$. The null

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

hypothesis, $H_0 : \gamma = 0$, can be tested using a Wald statistic.

The second method, referred to as "MAR", tests the marginal association between D and E. MAR fits the following logistic regression without adjusting for X:

$$\text{pr}(D = 1 \mid E = j) = \frac{\exp(\alpha_0 + \gamma_0 j)}{1 + \exp(\alpha_0 + \gamma_0 j)},$$

where α_0 is the marginal baseline log-relative risk and γ_0 is the marginal E-D log odds ratio. Note that α_0 and γ_0 generally differ from α and γ in model (2.1) unless β equals zero. The corresponding maximum likelihood estimator of γ_0, denoted by $\hat{\gamma}_M$, takes the form $\hat{\gamma}_M = \log(n_{1+1}/n_{1+0}) - \log(n_{0+1}/n_{0+0})$. The null hypothesis of no association between D and E, formulated as $\gamma_0 = 0$, can again be tested using a Wald statistic.

The third method, referred to as "ADJCON", is based on a constrained maximum likelihood method (Zhang et al., 2018). ADJCON optimizes the same likelihood function adopted in ADJ, subject to the following additional condition and constraint:

(C1) the variables E and X are independent;

(C2) the disease prevalence is known to be f (i.e., $\text{pr}(D = 1) = f$).

The constraint (C2) can be expressed as

$$\theta = \{f - p_{01}\pi - p_{00}(1-\pi)\}/\{p_{11}\pi + p_{10}(1-\pi) - p_{01}\pi - p_{00}(1-\pi)\}, \quad (2.2)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

where $p_{ij} = p_{ij}(\alpha, \beta, \gamma)$ is defined in (2.1). The likelihood function under (C1) and (C2) can be written as:

$$\prod_{d=0}^{1}\prod_{i=0}^{1}\prod_{j=0}^{1}\{\text{pr}(D=d \mid X=i, E=j)\text{pr}(X=i)\text{pr}(E=j)\}^{n_{dij}}, \qquad (2.3)$$

where $\theta = \text{pr}(X=1) = 1 - \text{pr}(X=0)$ is replaced with the right-hand side of (2.2). The corresponding maximum likelihood estimator, denoted by $\hat{\gamma}_{AC}$, can then be numerically obtained using any non-linear optimization algorithm. The null hypothesis $H_0 : \gamma = 0$ can be tested using a Wald statistic based on $\hat{\gamma}_{AC}$. Intuitively, AdjCon should be more efficient than Adj since the former incorporates additional condition and constraint. Indeed, simulation results in Zhang et al. (2018) showed that AdjCon outperforms both Adj and Mar.

3. Main Results

In this section, we establish theoretical properties for the estimators $\hat{\gamma}_M$, $\hat{\gamma}_A$ and $\hat{\gamma}_{AC}$. Specifically, asymptotic biases are derived in Section 3.1; asymptotic distributions are presented in Section 3.2; AREs for testing the null hypothesis are given in Section 3.3; robustness of $\hat{\gamma}_{AC}$ when prevalence is misspecified is investigated in Section 3.4.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

3.1 Asymptotic bias without covariate adjustment

Here we derive an asymptotic expression and the corresponding asymptotic bias for $\hat{\gamma}_M$. Let $n = n_{1++} + n_{0++}$. To avoid singularity, assume that $\nu = n_{1++}/n_{0++}$ is bounded away from zero and infinity.

Lemma 1. *We have the following asymptotic expansion for the marginal maximum likelihood estimator* $\hat{\gamma}_M$*:*

$$\hat{\gamma}_M = \gamma + \delta + O_P(n^{-1/2}) \text{ as } n \to \infty,$$

where

$$\delta = \log\left\{1 + \frac{e^{\alpha}(b_1 - b_2)(1 - e^{\gamma})}{(1 + e^{\alpha+\gamma}b_1)(1 + e^{\alpha}b_2)}\right\}, \tag{3.1}$$

and $b_1 = 1 + (e^{\beta} - 1)(1 - \theta), b_2 = \{1 + (e^{-\beta} - 1)(1 - \theta)\}^{-1}$.

It can be shown that $b_1 \geq b_2$, with the equality holding if and only if $\gamma = 0$ or $\beta = 0$. Furthermore, the signs of γ and δ are opposite and $|\delta| \leq |\gamma|$ - see (S1.6) and (S2.1) in the Supplementary Material for details. As a result, we have the following corollary:

Corollary 1. *The limiting value of* $\hat{\gamma}_M$*,* $\gamma + \delta$*, shrinks towards zero (i.e.,* $|\gamma + \delta| \leq |\gamma|$*). Furthermore, the asymptotic bias* δ *equals zero if and only if either* X *or* E *is not associated with* D*. Finally,* $|\delta|$ *is maximized at* $f = f^*$*,*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

where

$$f^* = \sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij}(\alpha^*, \beta, \gamma)\theta^i(1-\theta)^{1-i}\pi^j(1-\pi)^{1-j},$$

and

$$\alpha^* = -\frac{1}{2}\{\log(b_1 b_2) + \gamma\}.$$

Lemma 1 and Corollary 1 confirm the empirical observations that the maximum likelihood estimator of γ is conservative when ignoring non-confounding covariates (Stringer et al., 2011). The asymptotic unbiasedness conditions $\beta = 0$ and $\gamma = 0$ correspond to two non-confounding assumptions in prospective logistic regression (Robinson and Jewell, 1991). Gail et al. (1984) and Neuhaus and Jewell (1993) obtained similar results, but only for β around zero. In contrast, our results hold for general β. Moreover, $f \to 0$ implies $\delta \to 0$, i.e., adjusting for X results in a very small bias in the low disease prevalence situation, which is consistent with previous findings (Lee, 1982). Figure 1(A) displays the asymptotic bias in one parameter setting based on Lemma 1. The bias appears to increase with f for $f \in (0, f^*]$ and decrease with f for $f \in [f^*, 1)$. Unlike $\hat{\gamma}_M$, both $\hat{\gamma}_A$ and $\hat{\gamma}_{AC}$ are asymptotically unbiased under model (2.1). That is, $\hat{\gamma}_A = \gamma + O_P(n^{-1/2})$ (Anderson, 1972; Prentice and Pyke, 1979) and $\hat{\gamma}_{AC} = \gamma + O_P(n^{-1/2})$ (Zhang et al., 2018).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

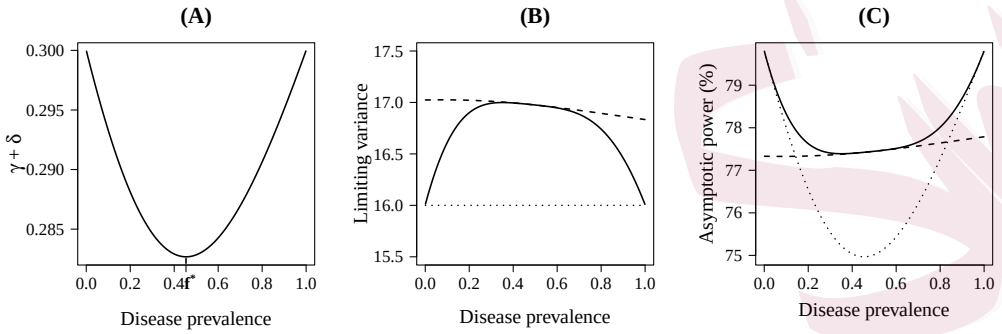

Figure 1: (A) The limiting value of $\hat{\gamma}_M$, $\gamma + \delta$, as a function of the disease prevalence f, with the underlying parameters being $\pi = 0.5$, $\theta = 0.4$, $\beta = 1$ and $\gamma = 0.3$; (B) The limiting variances of $n^{1/2}\hat{\gamma}_A$ (dashed line), $n^{1/2}\hat{\gamma}_{AC}$ (solid line) and $n^{1/2}\hat{\gamma}_M$ (dotted line), with the underlying parameters being $\pi = 0.5$, $\theta = 0.4$, $\beta = 1$, $\gamma = 0.05$ and $\nu = 1$; (C) The asymptotic powers of Mar (dotted line), Adj (dashed line) and AdjCon (solid line), with the underlying parameters being $n = 5 \times 10^4$, $q = 1$, $\theta = 0.4$, $\pi = 0.5$, $\beta = 1$ and $\gamma = 0.05$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

3.2 Asymptotic normality

In this subsection, we establish the asymptotic normality for $\hat{\gamma}_M$, $\hat{\gamma}_A$ and $\hat{\gamma}_{AC}$.

Lemma 2. *As n goes to infinity, $n^{1/2}(\hat{\gamma}_M - \gamma - \delta)$, $n^{1/2}(\hat{\gamma}_A - \gamma)$ and $n^{1/2}(\hat{\gamma}_{AC} - \gamma)$ converge in distribution to normal with mean zero and variances σ_M^2, σ_A^2 and σ_{AC}^2, whose explicit expressions are given in (S3.1), (S3.3), and (S3.7) in the Supplementary Material.*

We can analytically compare σ_M^2, σ_A^2 and σ_{AC}^2 using their explicit expressions, especially when $\gamma \to 0$ (and $f \to 0$), as detailed in the following corollary.

Corollary 2. *For σ_M^2, σ_{AC}^2 and σ_A^2, we have the following:*

1. $\sigma_M^2 \leq \sigma_A^2$, with equality holding if and only if $\beta = 0$.

2. If $\gamma \to 0$, then $\sigma_M^2 \to \sigma_0^2$ and $\sigma_A^2 \to \lambda\sigma_0^2$, where $\sigma_0^2 = (2 + \nu + 1/\nu)/\{\pi(1-\pi)\}$ and

$$\lambda = 1 + \frac{\nu\theta(1-\theta)}{(1+\nu)} \frac{(1-e^{\beta})^2}{\left\{(1-\theta)\phi + e^{\beta}\theta\phi^{-1}\right\}^2 + \nu e^{\beta}\left\{(1-\theta)\phi + \theta\phi^{-1}\right\}^2} \tag{3.2}$$

with $\phi = \sqrt{\frac{1+e^{\alpha+\beta}}{1+e^{\alpha}}}$. Furthermore, $\lambda \geq 1$ and the equality holds if and only if $\beta = 0$.

3. If $\gamma \to 0$ and $f \to 0$, then we have $\lambda \to \lambda_0$, $\sigma_{AC}^2 \to \sigma_0^2$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

$\sigma_A^2 \to \lambda_0 \sigma_0^2$, *where*

$$\lambda_0 = 1 + \frac{\nu\theta(1-\theta)}{(1+\nu)} \frac{(1-e^{\beta})^2}{(1-\theta+e^{\beta}\theta)^2 + \nu e^{\beta}} \geq 1, \tag{3.3}$$

and $\lambda_0 = 1$ *if and only if* $\beta = 0$.

Unlike linear models, adjusting for non-confounding covariates in case-control logistic regression always leads to an increase in the variance of γ estimator, i.e., $\sigma_M^2 \leq \sigma_A^2$, as claimed in Cololary 2. This result agrees with the finding of Robinson and Jewell (1991) for prospective studies. In the rare disease situation, Corollary 2 states that $\hat{\gamma}_M$ and $\hat{\gamma}_{AC}$ have the same asymptotic variance, and that $\hat{\gamma}_A$ has a larger asymptotic variance unless the covariate X is independent of the disease D. This result complements that of Pirinen et al. (2012), which only derives the approximated ratio of the variances for $\hat{\gamma}_M$ and $\hat{\gamma}_A$. Figure 1(B) displays the asymptotic variances as functions of f in one parameter setting. The asymptotic variance of $\hat{\gamma}_M$ appears to be the smallest in general. The variance of $\hat{\gamma}_A$ is the largest among the three estimators. On the other hand, the variance of $\hat{\gamma}_{AC}$ falls in between the other two, and it first increases then decreases with f.

We now compare the performances of the Wald test for the three methods (Mar, Adj and AdjCon) under the contiguous alternative hypothesis $H_1 : \gamma = c_1 n^{-1/2}$, where c_1 is a fixed non-zero constant. With the asymptotic mean and variance for each method in Lemmas 1 and 2, we can derive

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

the limiting power for the corresponding Wald test under H_1. As shown in Figure 1(C), AdjCon appears to be more powerful than Adj, which is due to the asymptotic unbiasedness of $\hat{\gamma}_A$ and $\hat{\gamma}_{AC}$ and the smaller asymptotic variance of AdjCon, especially when f is close to 0 or 1. When f approaches 0.5, the power gain of AdjCon over Adj becomes negligible, as their asymptotic variances converge. When f is close to 0 or 1, Mar appears to be more powerful than Adj, as they have similar means but Adj gives a larger variance. However, Mar becomes less powerful than the other two methods as f takes value around 0.5. This stems from the fact that $\hat{\gamma}_M$ is considerably biased toward zero and its variance advantage cannot be compensated for the bias disadvantage. AdjCon appears to be slightly more powerful than Mar when f is close to 0 or 1 (Figure 1(C)). The next subsection gives theoretical results related to Figure 1(C).

3.3 Asymptotic relative efficiencies

Asymptotic power comparison of various methods is carried out analytically through Pitman's ARE (Pitman, 1979; Serfling, 2009). For test statistics T_1 and T_2, the T_1 vs. T_2 Pitman ARE is equal to $e_P(T_1, T_2) = \lim(m_2/m_1)$ (refer to the Supplementary Material for details), where m_1 and m_2 are the sample sizes of T_1 and T_2 for achieving the same asymptotic power

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

under the contiguous alternative hypotheses $\gamma = cm_2^{-1/2}$ and $\gamma = cm_1^{-1/2}$, respectively. Therefore, $e_P(T_1, T_2) > 1$ indicates that T_1 is asymptotically more powerful than T_2, and vice versa. Denote by T_M, T_A and T_{AC} the Wald test statistics corresponding to $\hat{\gamma}_M$, $\hat{\gamma}_A$ and $\hat{\gamma}_{AC}$, respectively. Let $\rho = e^{\alpha}$, which is related to the disease prevalence. We evaluate $e_P(T_M, T_A)$ and $e_P(T_M, T_{AC})$ in Theorem 1 and Theorem 2, respectively.

Theorem 1. *The T_M vs. T_A Pitman ARE has the following asymptotic representation:*

$$e_P(T_M, T_A) = \left\{ \frac{b_1 b_2 \rho^2 + 2b_2\rho + 1}{b_1 b_2 \rho^2 + (b_1 + b_2)\rho + 1} \right\}^2 \lambda,$$

where $\lambda \geq 1$ is defined in (3.2) *and b_1 and b_2 are defined below* (3.1).

Theorem 1 gives an analytical form for $e_P(T_M, T_A)$, which allows us to evaluate their AREs under different prevalence levels. In particular, we have the following result in the rare disease situation.

Corollary 3. *For rare disease (i.e., $\rho \to 0$), the T_M vs. T_A Pitman ARE has the following asymptotic expansion:*

$$e_P(T_M, T_A) = \lambda_0 + O(\rho),$$

where $\lambda_0 \geq 1$ with λ_0 being defined in (3.3)*, and the equality holds if and only X and D are independent (or equivalently $\beta = 0$).*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

It is not surprising that AdjCon is generally more powerful than Adj since the two methods are based on the same model but the former incorporates additional condition and constraint. Furthermore, as indicated in Figure 1(C), AdjCon also appears to be more powerful than Mar. The following theorem gives a theoretical justification.

Theorem 2. *For rare disease (i.e., $\rho \to 0$), the T_M vs. T_{AC} Pitman ARE has the following asymptotic representation:*

$$e_P(T_M, T_{AC}) = 1 + \tau\rho^2 + o(\rho^2),$$

where

$$\tau = -\frac{(1-\theta)\theta(e^\beta - 1)^2\left[(1+1/\nu)\{(\theta(e^\beta-1)+1)^2 + e^\beta\nu\} + 2(1+(e^{2\beta}-1)\theta)\right]}{\{\theta\,(e^\beta - 1) + 1\}^2}.$$

Remark: Clearly, $\tau \leq 0$. Furthermore, $\tau = 0$ holds if and only if $\beta = 0$, which is equivalent to X and D being independent.

3.4 Constrained maximum method under prevalence misspecification

This section studies robustness of AdjCon against misspecification of disease prevalence. Numerical studies of Zhang et al. (2018) suggested that AdjCon is not very sensitive to the misspecification. Before stating our theoretical result, we introduce some more notations and assumptions. Let

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

$\boldsymbol{s} = (\beta, \gamma, \theta, \pi)^{\mathrm{T}}$ denote unknown model parameters. Note that the intercept parameter α is determined by f and $\boldsymbol{s}$ according to (2.2). Denote by $l_f(\cdot)$ the log-likelihood function with given prevalence f. Let the true prevalence be f_0. Let $\boldsymbol{s}_f^*$ denote the maximizer of $E_{f_0}\{l_f(\boldsymbol{s})\}$. Let $\hat{\boldsymbol{s}}_f$ denote the maximum likelihood estimator of $\boldsymbol{s}_f$ with the disease prevalence being specified to be f. Throughout this section, we assume that f is bounded away from 1 (i.e., $f \in (0, 1-\epsilon]$ for some $\epsilon > 0$) and for all $(\beta^*, \gamma^*, \theta^*, \pi^*) \in \{\boldsymbol{s}_f^* : f \in (0, 1-\epsilon]\}$, β^* and γ^* are bounded away from infinity and θ^* and π^* are bounded away from zero and one.

Theorem 3. *For any specified prevalence $f \in (0, 1-\epsilon]$, we have the following asymptotic properties:*

$$\sqrt{n}(\hat{\boldsymbol{s}}_f - \boldsymbol{s}_f^*) \to N(0, \Sigma_f(\boldsymbol{s}_f^*)) \text{ in distribution,}$$

$$\|\boldsymbol{s}_f^* - \boldsymbol{s}_{f_0}^*\| \leq C_1|f - f_0|,$$

and

$$\|\Sigma_f(\boldsymbol{s}_f^*) - \Sigma_{f_0}(\boldsymbol{s}_{f_0}^*)\| \leq C_2|f - f_0|,$$

where $\|\cdot\|$ is the Euclidean norm, $\Sigma_f(\boldsymbol{s}_f^*)$ is the asymptotic covariance matrix of $\hat{\boldsymbol{s}}_f$ evaluated at $\boldsymbol{s}_f^*$, and C_1 and C_2 are constants independent of f.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

Theorem 3 establishes the asymptotic normality of the maximum likelihood estimator with a possibly misspecified disease prevalence. Furthermore, the limiting value $\boldsymbol{s}_f^*$ of the maximum likelihood estimator $\hat{\boldsymbol{s}}_f$ and the corresponding asymptotic covariance matrix $\Sigma_f(\boldsymbol{s}_f^*)$ are Lipschitz continuous with respect to f, indicating that the statistical inference is not very sensitive to disease prevalence misspecification.

4. Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of Mar, Adj, and AdjCon. Our focus is to comparatively evaluate the performance of these methods in hypothesis testing. Additional results that further demonstrate the robustness of AdjCon in the context of prevalence misspecification are presented in the Supplementary Material.

We evaluate the hypothesis test performance of the three considered methods using simulation data. First, the covariate X and the exposure E are generated from Bernoulli distributions with success probabilities $\theta = 0.5$ and $\pi = 0.5$, respectively. Next, the disease status D is generated from the logistic regression model (2.1) with $\beta = 1.0$, $\gamma = 0$ or 0.075, and $f = 0.01, 0.05, 0.10, \ldots, 0.30$. A large population of size 10^7 are generated for each disease prevalence, and $n_{1++} = 10,000$ cases and $n_{0++} = 10,000$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

controls are randomly sampled from diseased individuals and non-diseased individuals, respectively. Wald test statistics for AdjCon, Mar, and Adj are calculated for each generated dataset, and type-I error rates ($\gamma = 0$) and powers ($\gamma = 0.075$) under nomininal level 0.05 are obtained based on 100,000 simulation replicates.

As expected, all methods have well controlled type-I error rates (Figure 2(A)). As shown in Figure 2(B), the three methods have power trends the same as those in Figure 1(C). That is, AdjCon is uniformly more powerful than Mar and Adj, while Mar is more powerful than Adj for small f and vice versa for large f.

We also conduct a sensitivity analysis with misspecified disease prevalence in AdjCon. It turns out that AdjCon is quite robust in the sense that the type-I error rates can be well controlled and the powers are still satisfying. We refer to Section S8 of the Supplementary Material for detailed simulation description and results (Table S1 and Figure S1).

5. Real Data Analysis

In this section, we analyze a case-control dataset on high-density lipoprotein cholesterol (HDL-C) (Edmondson et al., 2011). Individuals with HDL-C level above the 90th percentile are considered as cases and those below

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

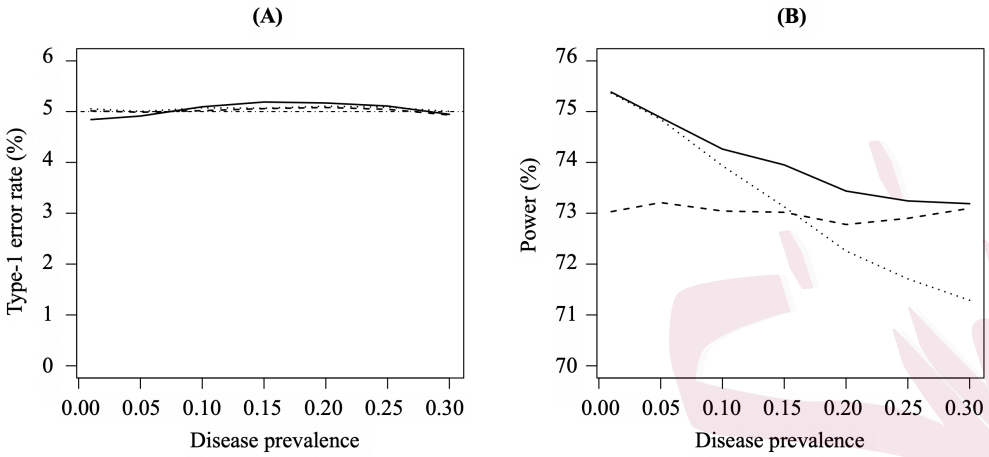

Figure 2: (A) Type-I error rates of MAR (dotted line), ADJ (dashed line), and ADJCON (solid line) for testing exposure-disease association ($H_0 : \gamma = 0$) with $\gamma = 0$, $\beta = 1$, $\theta = \pi = 0.5$, $n_0 = n_1 = 10000$; (B) Powers of MAR (dotted line), ADJ (dashed line), and ADJCON (solid line) for testing exposure-disease association ($H_0 : \gamma = 0$) with $\gamma = 0.075$, $\beta = 1$, $\theta = \pi = 0.5$, $n_0 = n_1 = 10000$.

the 30th percentile as controls. Of the 1231 subjects recruited from the University of Pennsylvania Hospital, 625 are identified as cases and 606 as controls. We consider a single covariate by dichotomizing body mass index (BMI) (1 = "BMI $>$ 26"; 0 = "BMI $\leq$ 26). This study involves 64 single nucleotide polymorphisms (SNPs) across 13 candidate genetic regions (PCSK5, NR1H3, FADS1-2-3, MVK/MMAB, LCAT, APOE, PLTP,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

GALNT2, LPL, ABCA1, LIPC, CETP, and LIPG), which have been previously reported to be associated with HDL-C levels (Edmondson et al., 2011). We aim to examine the association between the HDL-C level and these SNPs while adjusting for BMI.

Among the 64 SNPs, 41 are not significantly associated with BMI at level 0.05 by Pearson χ^2 tests (Table S2 in the Supplementary Material). BMI is strongly associated with the HDL-C level (Polychoric correlation coefficient $= 0.65$, p-value $< 2 \times 10^{-16}$), so that BMI can be regarded as a non-confounding covariate for the association analysis between each of the 41 SNPs and the HDL-C level. We apply AdjCon, Mar, and Adj to test the association between each SNP and the HDL-C level by adjusting for BMI. We fix the prevalence rate at 25% in the AdjCon method as in Zhang et al. (2018). The resulting p-values are presented in Table S3 in the Supplementary Material. An association is considered to be significant if the corresponding Bonferroni corrected p-value is smaller than 0.05.

Table 1 shows all significant associations among the 41 SNPs. Evidently, the p-values of AdjCon are uniformly smaller than those of Mar and Adj except for one SNP, and AdjCon uniquely identify two significant SNPs. These empirical results align with the theoretical results in Section 3.3.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

Table 1: Bonferroni adjusted p-values for SNP vs. HDL-C association tests

SNP	Mar	Adj	AdjCon	SNP	Mar	Adj	AdjCon
rs3779788	**0.039**	**0.020**	**0.015**	rs256	0.111	**0.041**	**0.031**
rs263	**0.008**	**0.002**	**0.001**	rs264	0.105	**0.040**	**0.034**
rs328	0.056	0.057	**0.032**	rs12679834	0.087	0.074	**0.045**
rs3208305	**0.017**	**0.025**	**0.016**	rs13702	**0.018**	**0.022**	**0.015**
rs11076174	**1.8E-3**	**2.8E-3**	**1.9E-3**	rs11076176	**5.6E-7**	**6.1E-8**	**5.2E-8**
rs289714	**5.3E-7**	**3.2E-8**	**3.0E-8**				

This table includes those SNPs significantly associated with the HDL-C level by at least one method at level 0.05 after Bonferroni adjustment.

6. Discussion

Adjusting for independent risk factors in randomized clinical trials can help improve estimation efficiency and test power in linear regression analyses (Fisher, 1932; Kahan et al., 2014). In case-control studies, there is still debate on whether independent covariates should be adjusted for in logistic regression analyses. We theoretically explored three methods's estimation efficiency and power when both the covariate and exposure of interest are binary. Our results can be summarized as follows. First, the estimated odds ratio of the exposure effect with the independent covariate ignored

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

($\hat{\gamma}_M$) is smaller than that of the covariate-adjusted estimate ($\hat{\gamma}_A$). This provided theoretical justification for the empirical observations in the literature (Stringer et al., 2011). Second, the variance of $\hat{\gamma}_M$ is smaller than that of $\hat{\gamma}_A$. This extended results in Pirinen et al. (2012) for rare outcome. Third, the variance of the estimated odds ratio for the covariate-adjusted exposure effect lies between those of Mar and Adj if the covariate-exposure independence is explicitly accommodated in the maximum likelihood estimation (AdjCon). AdjCon is always more powerful than both Mar and Adj, Mar is more powerful than Adj at low outcome prevalence, and Adj is more powerful than Mar when the outcome prevalence is close to 0.5. Last, we show the statistical inference for the AdjCon method is not sensitive to the outcome prevalence misspecification. These results theoretically confirm the empirical findings in Zhang et al. (2018).

The main results provide us with useful guidance for choosing appropriate approaches in case-control studies. In particular, we suggest using the constrained maximum likelihood method if the computational burden is not an issue. The marginal approach is preferred if the outcome prevalence is small, especially when one is interested in screening variables among a large number of potential risk factors (e.g., in genomewide association analysis studies).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

Our theoretical results are developed in a simple situation where the exposures of interest and the covariate are both binary. Further work is warranted to extend the current results to more general situations. For example, the exposure and covariate can be categorical or even continuous, there could be multiple independent covariates, and the sampling of cases and controls could be stratified. The three methods considered here can be extended to allow for link functions other than the logit link function. In Section S10 of the Supplementary Material, we conduct a simulation study for the probit link function and find that the corresponding results are similar to those under the logit link function.

There are some works related to ours in the literature. Methods have been developed to exploit gene-environment independence and prevalence information in the analysis of case-control data (Piegorsch et al., 1994; Chatterjee and Carroll, 2005; Mukherjee and Chatterjee, 2008; Chen and Chen, 2011; Clayton, 2012; Qin et al., 2014), to improve estimation efficiency and test power. Piegorsch et al. (1994) observed an improved efficiency for estimating gene-environment interaction effects using case-control data when the gene and environmental risk factors were independent in the population and the outcome was rare. Chatterjee and Carroll (2005) extended this method to incorporate covariates and allowed for a stratified sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

in the context of logistic regression models. Mukherjee and Chatterjee (2008) developed an empirical Bayes shrinkage method to relax the gene-environment independence assumption required in Chatterjee and Carroll (2005). Chen and Chen (2011) observed that no power improvement can be achieved by incorporating gene-environment independence if both gene and environmental factors are dichotomous. Qin et al. (2014) developed a rigorous statistical procedure to utilize covariate-specific outcome prevalence in the context of an exponential tilt model. The improvement in statistical efficiency of the method AdjCon is similar in spirit to these methods.

Supplementary Material

Detailed proofs of all theorems and lemmas, as well as definitions not included in the main paper due to space limitations, are available in the Supplementary Material accessible online at Statistica Sinica.

Acknowledgements

The research is supported in part by the Natural Science Foundation of China (12171451, 72091212), the US National Institutes of Health (R01-CA236468, R01-HL138306), the US National Science Foundation (DMS-2015417), the National Natural Science Foundation of China (12301373)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

and the Shanghai Science and Technology Committee Rising-Star Program (22YF1411100).

References

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika 59(1), 19–35.

Breslow, N., J. Robins, and J. Wellner (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. Bernoulli 6, 447–455.

Chatterjee, N. and R. J. Carroll (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92(2), 399–418.

Chen, H. Y. and J. Chen (2011). On information coded in gene-environment independence in case-control studies. American Journal of Epidemiology 174(6), 736–743.

Clayton, D. (2012). Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. Genetic Epidemiology 36(4), 409—-418.

Edmondson, A. C., P. S. Braund, I. M. Stylianou, A. V. Khera, C. P. Nelson, M. L. Wolfe, S. L. DerOhannessian, B. J. Keating, L. Qu, J. He, et al. (2011). Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol. Circulation: Cardiovascular Genetics 4(2), 145–155.

Fisher, R. A. (1932). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd (13th ed., 1958).

Gail, M. H., S. Wieand, and S. Piantadosi (1984). Biased estimates of treatment effect in ran-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

domized experiments with nonlinear regressions and omitted covariates. Biometrika 71(3), 431–444.

Kahan, B. C., V. Jairath, C. J. Doré, and T. P. Morris (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 15, 139.

Kuo, C.-L. and E. Feingold (2010). What's the best statistic for a simple test of genetic association in a case-control study? Genetic Epidemiology 34(3), 246–253.

Lee, L.-F. (1982). Specification error in multinomial logit models: Analysis of the omitted variable bias. Journal of Econometrics 20(2), 197–209.

Mukherjee, B. and N. Chatterjee (2008). Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64(3), 685–694.

Neuhaus, J. (1998). Estimation efficiency with omitted covariates in generalized linear models. Journal of the American Statistical Association 93(443), 1124–1129.

Neuhaus, J. and N. Jewell (1993). A geometric approach to assess bias due to omitted covariates in generalized linear-models. Biometrika 80(4), 807–815.

Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine 13(2), 153–162.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

REFERENCES

Pirinen, M., P. Donnelly, and C. C. A. Spencer (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. Nature Genetics 44(8), 848–851.

Pitman, E. J. G. (1979). Some Basic Theory of Statistical Inference. London: Chapman & Hall.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case-control studies. Biometrika 66(3), 403–411.

Qin, J., H. Zhang, P. Li, D. Albanes, and K. Yu (2014). Using covariate-specific disease prevalence information to increase the power of case-control studies. Biometrika 102(1), 169–180.

Robinson, L. D. and N. P. Jewell (1991). Some surprising results about covariate adjustment in logistic regression models. International Statistical Review/Revue Internationale de Statistique 59(2), 227–240.

Serfling, R. J. (2009). Approximation Theorems of Mathematical Statistics, Volume 162. John Wiley & Sons.

Stringer, S., N. R. Wray, R. S. Kahn, and E. M. Derks (2011). Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS ONE 6(11), e27964.

Xing, G. and C. Xing (2010). Adjusting for covariates in logistic regression models. Genetic Epidemiology 34(7), 769.

Zaitlen, N., B. Paşaniuc, N. Patterson, S. Pollack, B. Voight, L. Groop, D. Altshuler, B. E.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0249

REFERENCES

Henderson, L. N. Kolonel, L. Le Marchand, K. Waters, C. A. Haiman, B. E. Stranger, E. T. Dermitzakis, P. Kraft, and A. L. Price (2012). Analysis of case-control association studies with known risk variants. Bioinformatics 28(13), 1729–1737.

Zhang, H., N. Chatterjee, D. Rader, and J. Chen (2018). Adjustment of non-confounding covariates in case-control genetic association studies. Annals of Applied Statistics 12, 200–221.

East China Normal University

E-mail: slzhang@fem.ecnu.edu.cn

University of Pennsylvania

E-mail: jinboche@mail.med.upenn.edu

Columbia University

E-mail: zying@stat.columbia.edu

University of Science and Technology of China

E-mail: zhangh@ustc.edu.cn