Working Paper Series, N. 15, November 2009

Measurement Error Correction in Exploiting Gene-Environment Independence in Family-Based Case-Control Studies

Annamaria Guolo
Department of Economics, Society and Institutions, University of Verona
Via dell'Artigliere, 8, I-37129, Verona, Italy

Abstract: Family-based case-control designs are commonly used in epidemiological studies for evaluating the role of genetic susceptibility and environmental exposure to risk factors in the etiology of rare diseases. Within this framework, it is often reasonable to assume genetic susceptibility and environmental exposure being conditionally independent of each other within families in the source population. We focus on this setting to consider the common situation of measurement error affecting the assessment of the environmental exposure. We propose to correct for measurement error through a likelihood-based method, by exploiting the conditional likelihood of Chatterjee, Kalaylioglu and Carroll (2005) to relate the probability of disease to the genetic and the mismeasured environmental risk factors. Simulation studies show that this approach provides less biased and more efficient results than that based on traditional logistic regression. The likelihood approach for measurement error correction is also compared to regression calibration, the last resulting in severely biased estimators of the parameters of interest.

Keywords: conditional likelihood, conditional logistic regression, family-based case-control study, gene-environment independence, measurement error, regression calibration.

Department of Statistical Sciences
University of Padua
Italy

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Measurement Error Correction in Exploiting Gene-Environment Independence in Family-Based Case-Control Studies

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Annamaria Guolo
tel: +39 045 802 8716
annamaria.guolo@univr.it

Measurement Error Correction in Exploiting Gene-Environment Independence in Family-Based Case-Control Studies

Annamaria Guolo
Department of Economics, Society and Institutions, University of Verona
Via dell'Artigliere, 8, I-37129, Verona, Italy

Abstract: Family-based case-control designs are commonly used in epidemiological studies for evaluating the role of genetic susceptibility and environmental exposure to risk factors in the etiology of rare diseases. Within this framework, it is often reasonable to assume genetic susceptibility and environmental exposure being conditionally independent of each other within families in the source population. We focus on this setting to consider the common situation of measurement error affecting the assessment of the environmental exposure. We propose to correct for measurement error through a likelihood-based method, by exploiting the conditional likelihood of Chatterjee, Kalaylioglu and Carroll (2005) to relate the probability of disease to the genetic and the mismeasured environmental risk factors. Simulation studies show that this approach provides less biased and more efficient results than that based on traditional logistic regression. The likelihood approach for measurement error correction is also compared to regression calibration, the last resulting in severely biased estimators of the parameters of interest.

Keywords: conditional likelihood, conditional logistic regression, family-based case-control study, gene-environment independence, measurement error, regression calibration.

1 Introduction

Evaluating the influence of genetic susceptibility (G) and environmental exposure (X) as well as gene-environment (G-X) interaction on disease risks is a topic of increasing interest in epidemiologic studies. To this aim, a powerful tool is represented by family-based case-control designs, where controls are selected from families the cases belong to (Witte *et al.*, 1999; Gauderman, 2002).

In this paper, we focus on a feature of particular interest in gene-environment interaction problems, that is G-X independence within families in the source population. Assuming that the subject's genetic susceptibility G, which is determined from birth, does not affect the subject's environmental exposure X, is often reasonable especially in case of external environmental exposures, as, for example, to pollution or radioactive substances.

Within this framework, we face the problem of error affecting the measure of the environmental exposure for cases and controls. We propose a prospective likelihood-based approach (Carroll *et al.*, 2006, Chapter 8) to correct for the presence of measurement error. In particular, we take advantage of the conditional likelihood of

Chatterjee *et al.* (2005), which is successful in exploiting the within family G-X independence restriction. Conversely, we show that measurement error analysis based upon conditional logistic regression (Breslow and Day, 1980, p. 247–249) can lead to inefficient inferential results.

Moreover, we compare our approach to the method proposed by McShane *et al.* (2001), a modified version of the commonly used regression calibration approach. We show that the method, while being satisfactory in the case of no interaction models, experiences substantial bias in the estimators of all the parameters of interest when G-X interaction is taken into account.

The paper is organized as follows. Section 2 focuses on the problem of gene-environment interaction in family-based case-control studies and fixes the notation we will use throughout the paper. In addition, conditional logistic regression and the conditional likelihood of Chatterjee *et al.* (2005) are introduced. The problem of measurement error affecting the environmental exposure is presented in Section 3. Section 4 is devoted to the specification of regression calibration and of the likelihood approach to correct for measurement error in covariates. Section 5 shows the results of several simulation studies performed in order to compare the correction techniques. The paper ends with the discussion in Section 6.

2 Models and Notation

Let D be the binary indicator of case, $D = 1$, or control status, $D = 0$. Let G be the subject's genetic factors and X be the subject's exposure to environmental risk factors. For simplicity of exposition, here we focus on G and X being scalar variables, although the results can be readily extended to the multidimensional case. Suppose that a fixed number of cases and associated controls are sampled within a given family F in the population. Without loss of generality, we focus on the 1:1 matched study, e.g. sibling cases and controls. We assume that, within a family F, the prospective risk model for the disease for the j^{th} $(j = 1, 2)$ relative is given by the logistic regression model

$$\mathrm{pr}(D_j = 1|G_j, X_j, F) = H\left\{\alpha_F + m(G_j, X_j; \beta)\right\}, \tag{1}$$

where $H(v) = \{1 + \exp(-v)\}^{-1}$ is the logistic distribution function, $m(\cdot)$ is a known but arbitrary function and α_F are family-based intercepts. Often, the multiplicative interaction model

$$m(G_j, X_j; \beta) = \beta_0 + \beta_G G_j + \beta_X X_j + \beta_{GX} G_j X_j$$

is of interest, where $\beta = (\beta_0, \beta_G, \beta_X, \beta_{GX})^\top$ and the exponents of the parameters in β have the usual interpretation in terms of odds ratios.

Likelihood analysis based on (1) suffers for the Neymann-Scott problem, as the number of nuisance parameters α_F increases with the number of families. This leads to inconsistent estimators of the parameter of interest β. The problem is generally solved through conditioning on the number of cases within each family F, in order to eliminate the nuisance parameters α_F. This gives rise to conditional logistic regression (Breslow and Day, 1980, p. 247–249), which has the following expression.

Let (D_{k1}, G_{k1}, X_{k1}) indicate the observations for the case and (D_{k2}, G_{k2}, X_{k2}) those for the control in the k^{th} matched set, $k = 1, \ldots, K$. Thus, the likelihood obtained from the density function $f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta)$, $k = 1, \ldots, K$, is given by

$$L_{CLR}(\beta) = \prod_{k=1}^{K} f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta) \quad (2)$$

where

$$\begin{aligned} & f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta) = \\ & \quad \mathrm{pr}(D_{k1} = 1, D_{k2} = 0|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta) = \\ & \quad \frac{\exp(\beta_G G_{k1} + \beta_X X_{k1} + \beta_{GX} G_{k1} X_{k1})}{\exp(\beta_G G_{k1} + \beta_X X_{k1} + \beta_{GX} G_{k1} X_{k1}) + \exp(\beta_G G_{k2} + \beta_X X_{k2} + \beta_{GX} G_{k2} X_{k2})}. \end{aligned} \quad (3)$$

Expression (3) is obtained by conditioning on the number of cases within each stratum k, that is on $D_{k1} + D_{k2} = 1$. The method requires no assumptions on the distribution of the risk factors in the underlying population.

Chatterjee *et al.* (2005) focus on the situation of G and X being independent within families in the source population. The G-X independence within families in the source population, while being likely to occur as, for example, for exposure to external factors, is a weak assumption. The reason is that the assumed family-level independence is less likely to be affected by spurious association between G and X in the population. Indeed, the G-X independence at the population-level may be violated as a consequence of the population stratification on variables such as age group, ethnic background and family history (Mukherjee *et al.*, 2007). The within-family independence assumption, instead, is much less likely to fail because of these factors. Chatterjee *et al.* (2005) point out that the assumption is the weakest in the sibling-case-control design, in which no ethnic or family substructure may be supposed to induce differences among siblings.

Under the family-level G-X independence assumption, Chatterjee *et al.* (2005) show that conditional logistic regression (2) is inefficient. They suggest a novel conditional likelihood that is highly efficient in exploiting the within family G-X independence assumption. Let $\mathcal{G}_k$ be the unordered set of genotypes observed in the k^{th} matched set. By conditioning on $\mathcal{G}_k$ and under a rare disease assumption, Chatterjee *et al.* (2005) obtain their conditional likelihood from the density function $f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, \mathcal{G}_k, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta)$ in the k^{th} matched set,

$$L_{CC}(\beta) = \prod_{k=1}^{K} f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, \mathcal{G}_k, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta), \quad (4)$$

where

$$
\begin{aligned}
&f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, \mathcal{G}_k, X_{k1}, X_{k2}, D_{k1}+D_{k2}=1;\beta) = \\
&\quad \mathrm{pr}(D_{k1}=1, D_{k2}=0|G_{k1}, G_{k2}, \mathcal{G}_k, X_{k1}, X_{k2}, D_{k1}+D_{k2}=1) = \\
&\frac{\exp(\beta_G G_{k1} + \beta_X X_{k1} + \beta_{GX} G_{k1} X_{k1})}{\sum_{j=1}^{2}\{\exp(\beta_G G_{kj} + \beta_X X_{k1} + \beta_{GX} G_{kj} X_{k1}) + \exp(\beta_G G_{kj} + \beta_X X_{k2} + \beta_{GX} G_{kj} X_{k2})\}}.
\end{aligned}
\tag{5}
$$

For matched pairs where different genotypes are observed, $\mathcal{G}_k$ contains the information about the two observed genotypes, but no specification of the individual genotype of the case G_{k1} and the control G_{k2}. In fact, all the possible genotype-exposure configurations in the k^{th} matched set, that is $\{(G_{k1}, X_{k1}), (G_{k1}, X_{k2}), (G_{k2}, X_{k1}), (G_{k2}, X_{k2})\}$, are taken into account in the denominator of (5). The numerator, instead, is equal to that in (3).

Chatterjee *et al.* (2005) show analytically that the proposed method is asymptotically at least as efficient as standard conditional logistic regression and confirm these results by simulation studies.

In this paper, we investigate the problem of measurement error affecting the subject's exposure to an environmental risk factor X. We consider a likelihood analysis of family-based case-control data in order to estimate the G-X interaction, under the G-X independence assumption within families in the source population. Within this framework, we take advantage of the method suggested by Chatterjee *et al.* (2005) and compare the results to those from a measurement error version of standard conditional logistic regression and the regression calibration method of McShane *et al.* (2001).

3 Measurement Error

The problem of measurement error affecting covariates arises in many scientific areas. It is well known that uncorrect measures of covariates can yield misleading inferential results, the most relevant being the bias induced on the estimators. Considerable attention has been paid on this problem in literature. Carroll *et al.* (2006) provide a review of the techniques proposed to correct for measurement error.

In this paper, the focus is on measurement error affecting continuous covariates. In particular, we suppose that the subject's exposure to an environmental risk factor X is not directly observed. Instead of X, we observe a mismeasured quantity W, which is related to X according to the so-called *measurement error model.* We mainly focus on a classical measurement error model, $W = X + U$, where $U \sim \mathrm{Normal}(0, \sigma_u^2)$, which can be often thought as a good approximation of more complex measurement error structures. However, the possibility of alternative measurement error distributions is investigated. In particular, we take account of an asymmetric and a multiplicative measurement error structure. We assume that the measurement error is nondifferential, i.e. independent of the disease status D, as reasonably holds when the exposure is assessed by an instrument.

We evaluate two approaches to correct for measurement error affecting the observations of X, regression calibration and a likelihood-based approach. These are briefly reviewed in the following section.

4 Correction Techniques

4.1 Regression Calibration

Regression calibration (RC, for short) is one of the most commonly used methods to correct for measurements errors (Carroll *et al.*, 2006, Chapter 4). This is mainly due to its simple applicability with existing packages. The method involves two steps. In the first one (calibration step) the unknown values of X are estimated by the conditional expectation of X given (W, G), that is $E(X|W, G) = X^*$. In the second step, conditional logistic regression is performed with X replaced by X^*.

The RC approach we consider in this paper is the one suggested by McShane *et al.* (2001). While in ordinary case-control studies only controls are used within the calibration step, McShane *et al.* (2001), in analyzing matched data, show that using all the observations from cases and controls produces nearly unbiased results. They show that their proposal generally works well, except when the covariate distribution is highly skewed.

In the calibration step, we consider W data from both cases and controls in order to obtain the conditional expectation of X given (W, G)(McShane *et al.*, 2001). In particular, let Σ_{ab} denote the covariance matrix between two random variables A and B and μ_a denote the mean of a random variable A. In order to guarantee the parameter identifiability, we suppose that the measurement error variance matrix Σ_{uu} is known. Let n be the number of observations subdivided in K matched pairs, so that $n = 2K$. Let (X_i, W_i, G_i) the i-th observation of (X, W, G), for $i = 1, \ldots, n$. The best linear approximation to $E(X|W, G)$ is

$$E(X|G, W) \approx \widehat{\mu}_w + \begin{pmatrix} \widehat{\Sigma}_{xx} \\ \widehat{\Sigma}_{xg} \end{pmatrix}^{\top} \begin{bmatrix} \widehat{\Sigma}_{xx} + \Sigma_{uu} & \widehat{\Sigma}_{xg} \\ \widehat{\Sigma}_{xg}^{\top} & \widehat{\Sigma}_{gg} \end{bmatrix} \begin{pmatrix} W - \widehat{\mu}_w \\ G - \widehat{\mu}_g \end{pmatrix},$$

where

$$\begin{aligned} \widehat{\mu}_x &= \widehat{\mu}_w = \overline{W} = \sum_{i=1}^{n} W_i/n; \quad \widehat{\mu}_g = \overline{G} = \sum_{i=1}^{n} G_i/n; \\ \widehat{\Sigma}_{gg} &= (n-1)^{-1} \sum_{i=1}^{n} (G_i - \widehat{\mu}_g)(G_i - \widehat{\mu}_g)^{\top}; \\ \widehat{\Sigma}_{xg} &= (n-1)^{-1} \sum_{i=1}^{n} (W_i - \widehat{\mu}_w)(G_i - \widehat{\mu}_g)^{\top}; \\ \widehat{\Sigma}_{xx} &= \left[\left\{ \sum_{i=1}^{n} (\overline{W} - \widehat{\mu}_w)(\overline{W} - \widehat{\mu}_w)^{\top} \right\} - (n-1)\Sigma_{uu} \right] / (n-1). \end{aligned}$$

Within the calibration step we ignore the matching nature of the data. This means that we substitute the resulting predictions of X to the corresponding unknown values and perform a conditional logistic regression on the matched data. Standard errors for RC estimates are readily obtained by bootstrap.

4.2 Likelihood Methods

Within the context of family-based case-control data, let W_{k1} and W_{k2} be the observations from W for the case and the associated control in the k^{th} matched set, respectively. Then, the likelihood function is obtained starting from the joint density function of the observed quantities (D, G, W, X) over the K pairs, and then integrating out the unobserved X (Carroll *et al.*, 2006, Chapter 8),

$$L(\theta) = \prod_{k=1}^{K} \int_{\mathbb{R}^2} f_{DGWX}(D_{k1}, D_{k2}, G_{k1}, G_{k2}, W_{k1}, W_{k2}, x_{k1}, x_{k2}; \theta) dx_{k1} dx_{k2}. \qquad (6)$$

The integral is replaced by a sum in case of discrete X.

Suppose that the relationship between D and (G, X, W) in the k^{th} matched set is specified through the density function $f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \beta)$ according to the standard conditional logistic regression, as in (3), and under the nondifferential measurement error assumption. Then, likelihood (6) can be rewritten as

$$\begin{aligned} L(\theta) \;\; = \;\; & \prod_{k=1}^{K} \int_{\mathbb{R}^2} f_{D|GX}(D_{k1}, D_{k2}|G_{k1}, G_{k2}, x_{k1}, x_{k2}, D_{k1} + D_{k2} = 1; \beta) \\ & f_{W|GX}(W_{k1}, W_{k2}|G_{k1}, G_{k2}, x_{k1}, x_{k2}, D_{k1} + D_{k2} = 1; \gamma) \\ & f_{X|G}(x_{k1}, x_{k2}|G_{k1}, G_{k2}, D_{k1} + D_{k2} = 1; \delta) \\ & f_G(G_{k1}, G_{k2}|D_{k1} + D_{k2} = 1; \lambda) dx_{k1} dx_{k2}, \end{aligned}$$

where $\theta = (\beta^\top, \gamma^\top, \delta^\top, \lambda^\top)^\top$, $f_{W|GX}(\cdot; \gamma)$ is the density function for the measurement error model relating W to (X, G), depending on γ, $f_{X|G}(\cdot; \delta)$ is the density function of the model for the mismeasured variable X given G, depending of δ and $f_G(\cdot; \lambda)$ is the density function of G, depending of λ. The expression of the likelihood can be simplified as follows. Since it is reasonable to assume that the error in measuring X is independent of G and that it is also independent when evaluated for the case or for the control, then

$$\begin{aligned} f_{W|GX}(W_{k1}, W_{k2}|G_{k1}, G_{k2}, X_{k1}, X_{k2}, D_{k1} + D_{k2} = 1; \gamma) \;\; = & \\ f_{W|X}(W_{k1}|X_{k1}, D_{k1} + D_{k2} = 1; \gamma) f_{W|X}(W_{k2}|X_{k2}, D_{k1} + D_{k2} = 1; \gamma). & \end{aligned}$$

Moreover, given the G-X independence assumption within families, it follows that

$$f_{X|G}(X_{k1}, X_{k2}|G_{k1}, G_{k2}, D_{k1} + D_{k2} = 1; \delta) = f_X(X_{k1}, X_{k2}|D_{k1} + D_{k2} = 1; \delta).$$

Finally, the marginal density of G carries no information about the parameter of interest β and it does not depend on X, thus it can be neglected in the likelihood computation.

A similar expression for the likelihood can be obtained also when the conditional likelihood of Chatterjee *et al.* (2005), as in (5), is considered in place of conditional logistic regression.

The likelihood computation can be simplified by considering the environmental exposure X for cases and controls independent within each matched family. In this case,

$$f_X(X_{k1}, X_{k2}|D_{k1} + D_{k2} = 1; \delta) = f_X(X_{k1}|D_{k1} + D_{k2} = 1; \delta) f_X(X_{k2}|D_{k1} + D_{k2} = 1; \delta).$$

An extensive simulation study shows that the simplification is reasonable. We generated 200 matched case-control pairs from the logistic regression model, under different values for the influence of G, X and G-X interaction on the probability of disease. More details about the simulation design can be found in Section 5.1. For each combination of the values for $\beta_G, \beta_X, \beta_{GX}$, we tested the independence of the exposure between the cases and the controls. The independence test is based on the empirical copula process as proposed by Genest and Rémillard (2004). The results of the simulation study are reported in Table 5 in Appendix A, suggesting that modeling the covariate distributions of the cases and the controls as independent is a reasonable approach.

When performing the likelihood analysis, some additional information, e.g. the measurement error variance σ_U^2 known, is needed to guarantee the identifiability of the parameters (Carroll *et al.*, 2006, Section 8.1.2).

Remark 1

Within the likelihood approach, Guolo (2008) studied the connection between the case-control sampling scheme and the possibility of model misspecification, with the consequent risk of unreliable inferential results. The problem arises from the difficulty in specifying a model for the unobserved Xs and it is also exacerbated when handling case-control data. In fact, the distribution of the covariates in the population can notably differ from that under the case-control sampling scheme. The problem can be addressed by specifying the model for X in the case-control sample through a flexible distribution. By this way, Guolo (2008) shows that likelihood estimation and inferences are asymptotically correct. Following Guolo (2008), we flexibly model the distribution of X in the case-control sample through the skewnormal distribution (Azzalini, 1985), $X \sim \mathrm{SN}(\mu_X, \sigma_X, \alpha_X)$, with density function

$$f_X(x; \delta) = f_X(x; \mu_X, \sigma_X, \alpha_X) = (2/\sigma_X)\phi\left\{(x - \mu_X)/\sigma_X\right\}\Phi\left\{\alpha_X(x - \mu_X)/\sigma_X\right\},$$

where $\delta = (\mu_X, \sigma_X, \alpha_X)^\top$, $\mu_X, \sigma_X, \alpha_X$ are, respectively, the location, the scale and the shape parameter and $\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard normal density and distribution functions.

Remark 2

Together with the possibility of model misspecification for X described below, the case-control sampling scheme has some effects also on the G-X relationship. In fact, despite the G-X independence assumed within families in the population, a G-X dependence may hold in the case-control data as a consequence of the sampling scheme. Consider, for example, Figure 1. In the figure, the nonparametric estimate of the density of X conditionally on $G = 0$ and $G = 1$ in a population (left panel) of

$50,000$ families and in the extracted case-control sample (right panel) is reported. While the estimated density of X is the same for $G = 0$ and $G = 1$ in the population, its different behavior in the case-control sample is an evidence of the lack of G-X independence. Actually, the case-control sampling scheme can be shown to induce the G-X density function being different from that in the population and, in particular, dependent on the disease status D. For details, see Appendix C. When performing the likelihood analysis, we aim to correct for the spurious G-X association in the case-control sample and recover the within families independence holding in the population. To this aim, in spite of conditional logistic regression, we take advantage of the conditional likelihood by Chatterjee *et al.* (2005) since it equally includes all the combinations of gene susceptibility and environmental exposure for the subjects independently of their disease status, as it would be expected under the G-X independence assumption.

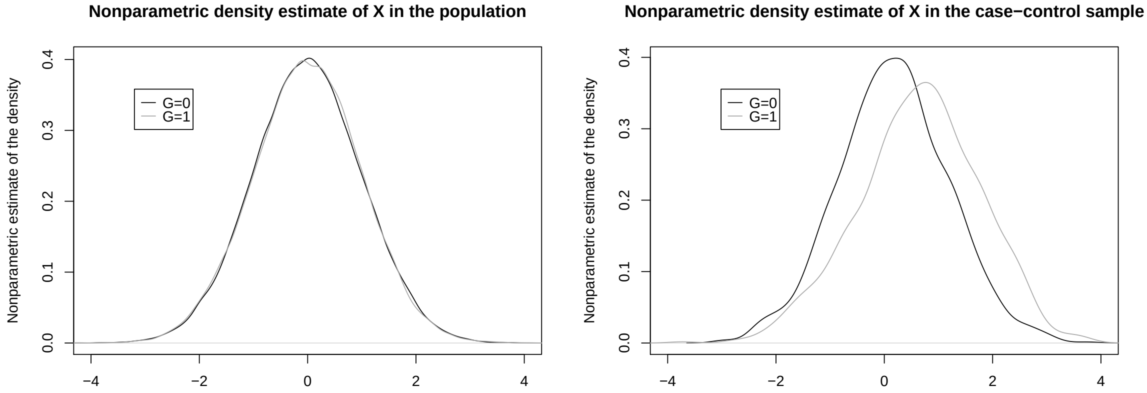

Figure 1: Nonparametric estimate of the density of the environmental exposure X in the population (left panel) of 50,000 families and in the case-control sample (right panel) conditionally on the values of the genetic susceptibility G, $G = 0$ (black line) and $G = 1$ (grey line). While G and X are independent in the population, they are not independent in the case-control sampling scheme.

5 Simulation Studies

We performed extensive simulation studies in order to evaluate the behaviour of the likelihood-based approach to correct for measurement error affecting X under the assumption of G-X independence within families in the population. We refer both to the measurement error generalization of conditional logistic regression (CLR) and of the conditional likelihood (CC) of Chatterjee *et al.* (2005) to relate D to (G, X). Both the methods are compared to regression calibration (RC) in the version by McShane *et al.* (2001) and to the *naive* analysis (NAIVE), which ignores the presence of measurement error. Several simulation scenarios are examined, allowing for different distributions of X in the population, as well as different sample sizes and measurement error structures.

5.1 Simulation Design

In the simulation studies we took a sample of K case-control pairs. In Section 5.3 we will report the results referred to the simulation with $K = 1,000$. Also the moderate sample performance of the correction techniques has been examined, with $K = 500$. The corresponding results are reported in Appendix B.

We simulate data for families composed of two siblings and their parents. We suppose the gene variant of interest is a bi-allelic locus, with susceptibility allele a and normal allele A. Let p denote the prevalence of the risk allele. By assuming Hardy-Weinberg equilibrium holding in the population, the distribution of the genotypes $\mathcal{G}$ in the population is given by $\mathrm{pr}(\mathcal{G}|p) = p^2, 2p(1-p), (1-p)^2$ for $\mathcal{G} = aa, Aa/aA, AA$, respectively. We focus on two different settings of interest, the dominant (Aa/aA or aa) and the recessive model (aa) for the effect of the gene-variant.

Following the same design as Chatterjee *et al.* (2005), for each family F we generate a family-specific allele frequency parameter θ_F in order to allow for between family variability. We first generate a normal random variable u_F with mean μ and variance σ^2, and then we obtain θ_F belonging to the $0-1$ scale as $\theta_F = e^{u_F}/(1+e^{u_F})$. Choosing $\sigma^2 = 0.5$ allows the $\pm 2\sigma$ limit of the distribution of u_F corresponding to approximately 15-fold variation in allele frequency across the families. The value of μ is chosen in order to guarantee that the marginal probability of the genotype variant of interest for the dominant and for the recessive model in the population is fixed to 0.2. Given the allele frequency parameter θ_F, we simulate the parental genotypes according to Hardy-Weinberg equilibrium, under the assumption of independence between the two parents. Then, conditionally on the parental genotypes, the genotype of the siblings is simulated according to the Mendelian mode of inheritance. The genotype information G for the siblings is mapped into a genetic covariate, which is taken to be binary, thus indicating the presence or absence of the genetic mutation. The siblings environmental exposure X is assumed to be independently distributed of the genetic susceptibility within families in the source population. Two distributions for X are taken into account: a $\log \chi^2_2$ distribution and a mixture of normal distributions distributions $\mathrm{Normal}(0.5, 1)$ and $\mathrm{Normal}(-1.5, 1)$, with mixing weights 0.6 and 0.4. Within a given family F, let X_j be the environmental exposure for the case ($j = 1$) or the control ($j = 2$). Having generated the values of G and X, the binary disease outcome D for the siblings in each family F is drawn from the logistic regression model

$$\mathrm{pr}(D_j = 1|G_j, X_j, F) = H\left(\beta_0 + \alpha_F + \beta_G G_j + \beta_X X_j + \beta_{GX} G_j X_j\right),$$

where $j = 1, 2$, the parameters $(\beta_G, \beta_X, \beta_{GX})^\top$ are set equal to $(\log(1.3), \log(1.5), 1)^\top$ and the parameter β_0 is chosen such that $\mathrm{pr}(D = 1) = 0.01$. The family-specific intercepts α_F allow for heterogeneity in the probability of disease among families. Values of α_F are simulated from a standard normal distribution.

Simulation results in Section 5.3 refer to the case of a classical, i.e. linear and additive, measurement error affecting the covariate X. We assume that $W = X + U$, where the random error U follows a normal distribution, $U \sim \text{Normal}(0, \sigma_U^2)$. Different amounts of measurement error are considered, $\sigma_U^2 \in \{0.7, 1.0\}$. The performance of the correction techniques was also examined under nonclassical measurement error structures, on a subset of the data, as follows. First, when X is generated from a $\log \chi_2^2$ distribution, we examined the performance of the correction methods under an asymmetric measurement error, $W = X + U$, with U following a skewnormal distribution with location, scale and shape parameters equal, respectively, to 0, 0.8 and 1.0. Secondly, when X is generated from a mixture of normals, we allowed W following a multiplicative structure, $W = XU$, with U distributed according to a normal variable with mean 0.8 and variance 0.25. This includes heteroschedasticity in the measure of X.

5.2 Details

We considered 500 replicates of the simulation scheme described in Section 5.1.

Regression calibration estimates are obtained according to the algorithm described in Section 4.1. Standard errors are obtained by the bootstrap on the matched pairs of subjects, with 100 boostrap samples.

The measurement error analysis in (6) is based upon conditional logistic regression (2) or the conditional likelihood (4) of Chatterjee *et al.* (2005). The resulting likelihood function is maximized by using the optimization procedures provided by the R programming language (R Development Core Team, 2009). Integrals involved in the likelihood maximization are numerically evaluated, through Gauss-Hermite multidimensional quadrature with 14 nodes. The optimization procedure requires initial estimates of the parameters. We chose the estimates provided by regression calibration for $(\beta_G, \beta_X, \beta_{GX})^\top$, while we used the moment-based estimates on the observations from W for the parameters involved in the distribution of X.

The variance estimates for the likelihood estimators are obtained using the sandwich method. Let $\ell(\theta)$ be the log-likelihood for θ obtained from (6) and let $\hat{\theta}$ be the maximizer of $\ell(\theta)$. Then, the sandwhich estimator of the covariance matrix for $\hat{\theta}$ is

$$\text{cov}(\hat{\theta}) = K^{-1}\, J_k^{-1}(\theta) I_k(\theta) J_k^{-1}(\theta)\big|_{\theta=\hat{\theta}},$$

where

$$J_k(\theta) = K^{-1} \sum_{k=1}^{K} \frac{\partial^2}{\partial\theta\partial\theta^{\mathcal{T}}} \ell_k(\theta)$$

and

$$I_k(\theta) = K^{-1} \sum_{k=1}^{K} \frac{\partial}{\partial\theta} \ell_k(\theta) \left(\frac{\partial}{\partial\theta} \ell_k(\theta) \right)^{\mathcal{T}}.$$

See, for example, Carroll *et al.* (2006, Section A.6.1).

5.3 Results

The simulation results, performed under both the specifications of X, are summarized in Tables 1-2, under a classical measurement error structure, for both the recessive and the dominant genetic model. They refer, respectively, to X distributed as a $\log \chi_2^2$ distribution and as a mixture of normals. The simulation results under nonclassical errors are reported in Tables 3-4, under the recessive genetic model. In each table, the estimates of $\beta_G, \beta_X, \beta_{GX}$, the estimated standard errors of the parameter estimators and the empirical coverages of confidence intervals at nominal level equal to 0.95 are reported.

First, the simulation results highlight that the *naive* approach provides estimators for all the parameters of interest which are notably more biased than alternatives, under all the examined measurement error structures. Moreover, the empirical coverage of their resulting confidence intervals is very low and far from the nominal level. In case of classical measurement error, the results become worse as the amount of measurement error increases, under both the recessive and the dominant genetic variant. The use of RC only slightly outperforms the *naive* analysis, again retaining high levels of bias of the estimators and poor empirical coverage of confidence intervals. In particular, this situation is quite dramatic for the estimation of the interaction parameter β_{GX}, see, for example, Table 1. This result is in marked contrast to the results of McShane *et al.* (2001), who pointed out the good performance of the method, both in terms of bias and mean squared error of the estimators. This difference may stem from the fact that McShane *et al.* (2001) do not consider interaction models. Their method is successful when interaction is not present, otherwise leading to biased results. This conjecture is confirmed by a simulation study (100 replicates) of $K = 500$ matched case-control pairs, performed under the assumption of no G-X interaction, that is, under $\beta_{GX} = 0$ (see Table 6 in Appendix A). In this case, in fact, RC works quite well. For simplicity, only the results for a recessive genetic variant under a classical measurement error model with $\sigma_U^2 = 1$ are reported, although simulation conclusions hold more generally.

Correcting for the presence of measurement error through a likelihood-based approach within the conditional likelihood framework provides some advantages with respect to the *naive* analysis or to regression calibration. For example, consider the bias reduction for the estimators of β_{GX} and the related improvement of the empirical coverages of confidence intervals under different measurement error structures. However, results are still far from being satisfactory. In particular, while correcting for the measurement error affecting the estimate of β_X, the approach experiences severe bias of the estimator of β_G. Similarly, the empirical coverage of confidence intervals is far from the nominal level. According to this performance, the presence of the measurement error affecting X has repercussions on the estimators of parameters related to other variables, mainly to G. This can be thought of as a consequence of the spurious G-X relationship introduced by the case-control sampling and that the likelihood approach based on CLR cannot correct for, see Remark 2 in Section 4.2. Conditional logistic regression cannot recover the within families independence holding in the population while estimating the parameters, because it takes into account only the observed gene susceptibility and environmental exposure for each subject of the case-control sample.

Conversely, when adopting the measurement error generalization of CC, all the combinations of gene susceptibility and environmental exposure of the pairs of relatives are taken into account independently of the subjects' disease status, as it would be expected under the G-X independence assumption. Thus, the likelihood analysis provides more satisfactory results than the alternative based on CLR. According to the examined simulation studies, the measurement error generalization of CC properly corrects for errors affecting X, while at the meantime accounting for repercussions on the other covariates. The most remarkable result is the dramatic bias reduction of the estimators of the parameters β_G and β_{GX} which can be obtained by relying on CC in place of CLR. See, for example, Table 1 and Table 2. This performance, together with a sensible reduction of the standard error of the estimates, turns out in empirical coverages of confidence intervals which are very close to the 0.95 nominal level. The satisfactory behavior of the likelihood approach based on CC is maintained under different measurement error structures and magnitude, as well as in case of rare or common genetic mutation.

The moderate sample performance of the methods in correcting for measurement errors, for $K = 500$ matched case-control pairs, is summarized in Tables 7-8 in Appendix B. The relative performance of the methods substantially recovers that with $K = 1,000$ matched case-control pairs, dictating the likelihood approach within CC framework as the preferable solution. The main difference is an expected increased value of the estimated standard error of the parameter estimators. Within a likelihood approach, a slightly increased bias of the estimators is a consequence of the finite sample distribution of the maximum likelihood estimators.

6 Conclusions

In this paper, we have investigated the measurement error problem in the context of family-based case-control studies about the role of genetic susceptibility and environmental exposure on the risk of rare disease. We have focused on the weak assumption of gene-environment independence within families in the source population. This is a much weaker assumption than the independence at the population level and, at the meantime, very likely to be satisfied, for example in case of external environmental risk factors.

Within this framework, we suggest to correct for measurement error affecting the exposure to environmental risk factors through a likelihood-based approach. Simulation studies show that this solution is much more preferable to regression calibration. Regression calibration, in fact, is successful only in case of no G-X interaction models, otherwise yielding very biased results. Likelihood analysis is shown to provide notable efficiency advantages in terms of inferential results if the assumption of G-X independence within families is properly exploited. This goal is achieved by basing the analysis on the conditional likelihood of Chatterjee *et al.* (2005) in place of the standard conditional logistic regression. In fact, while the former results in almost no bias of the estimators, the latter fails to exploit the G-X independence assumption and suffers from notable effects of the measurement error on all the parameters of interest. This behavior occurs under different frameworks, namely different measurement error structure, recessive or genetic variants, moderate or large sample sizes. The reason of the success of basing measurement error analysis on the conditional likelihood by Chatterjee *et al.* (2005) is related to the treatment of the G-X relationship in the case-control sample. While the analysis based on conditional logistic regression does not exploit the G-X independence within families, the approach based on the conditional likelihood of Chatterjee *et al.* (2005), instead, attacks this problem directly by treating cases and controls in a symmetric way and allowing for all the possible gene susceptibilities for the subjects.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171-178.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Volume 1 – The Analysis of Case-Control Studies.* Oxford: Oxford University Press.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Boca Raton: Chapman & Hall, CRC Press.

Chatterjee, N., Kalaylioglu, Z. and Carroll, R.J. (2005). Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genetic Epidemiology* **28**, 138–156.

Gauderman, W.J. (2002). Sample size requirements for matched case-control studies of gene-environment interaction. *Statistics in Medicine* **21**, 35–50.

Genest, C. and Rémillard, B. (2004). Tests of independence and randomness based on the empirical copula process. *Test* **13**, 335–369.

Guolo, A. (2008). A flexible approach to measurement error correction in case-control studies. *Biometrics* **64**, 1207–1214.

McShane, L.M., Midthune, D.N., Dorgan, J.F., Freedman, L.S. and Carroll, R.J. (2001). Covariate measurement error adjustment for matched case-control studies. *Biometrics* **57**, 62–73.

Mukherjee, B., Zhang, L., Ghogh, M. and Sinha, S. (2007). Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. *Biometrics* **63**, 834–844.

R Development Core Team (2009). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Witte, J.S., Gauderman, W.J. and Thomas, D.C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *American Journal of Epidemiology* **149**, 693–705.

A Simulation studies: preliminary results

Table 5 contains the results of a simulation study performed in order to evaluate the p-value of the empirical copula test of independence (Genest and Rémillard, 2004) of the environmental exposure between cases and controls, for $\beta_G \in \{\log(1.3), \log(1.8), \log(2.3)\}$, $\beta_E \in \{\log(1.5), \log(2.0), 1.0\}$ and $\beta_{GX} \in \{1.0, 1.5, 2.0\}$. Values from X are simulated from a mixture of normal distributions, as in Section 5.1. The large values of the p-values provide an empirical justification for modeling the covariate distributions for cases and controls as independent, see Section 4.2.

Table 6 summarizes the results of the simulation study performed in order to evaluate the behavior of the correction methods under no gene-environment interaction in the population. In this case, both RC and the likelihood approach within CLR framework seem to perform well, while they generally do not under nonzero gene-environment interaction.

B Simulation studies: moderate sample performance

Tables 7-8 report the results of simulation studies performed in order to examine the moderate sample performance of the NAIVE approach, regression calibration and likelihood analysis, within both conditional logistic regression or the conditional likelihood by Chatterjee *et al.* (2005). We consider 500 replicates of $K = 500$ matched case-control pairs, under both the recessive and the dominant genetic model, according to X following a $\log \chi^2_2$ distribution or a mixture of normals. The focus is on the classical measurement error model, with $\sigma^2_U \in \{0.7, 1.0\}$. Results globally recover the main findings derived from the simulation study of 500 replicates of $K = 1,000$ matched case-control pairs, see Section 5.3. As expected, as the sample size reduces, the estimated standard errors of the parameter estimators increase and the bias of the likelihood estimators slightly increases.

C Gene-environment relationship in the case-control sample

The G-X independence assumed within families in the source population may not hold in the case control data because of the sampling scheme. To show this, make the following definitions. Let $p_d = n_d/n$ be the proportion of subjects with $D = d$ in the sample and let τ be the indicator of whether or not a subject has been selected in the case-control sample, so that $\tau = 1$ if (D, W) are observed and $\tau = 0$ otherwise.

The G-X restriction within each family F in the population implies that the joint density of (G, X) can be factorized as follows

$$f_{GX|F}(G_i, X_j) = f_{G|F}(G_i|F) f_{X|F}(X_j|F),$$

where indices i and j refer to the i-th and the j-th subject of the family, i allowed to be equal to j, independently of the family components' disease status. In the case-control sample, instead, the density function of (G, X) can be shown (see Guolo, 2008, Appendix A) to be equal to

$$f_{GX|\tau=1,F}(G_i, X_j|\tau = 1, F) = \sum_{d=0}^{1} p_d f_{GX|F,D}(G_i, X_j|F, D).$$

The dependence of each term of the previous sum on D does not allow the joint density of (G, X) being factorized in a component dependent on G, say $f_{G|F,D}(G_i|F, D)$, and a component dependent of X, say $f_{X|F,D}(X_j|F, D)$. The factorization is achievable, instead, at the population level. The reason is that the population substructure induced by the disease status D is not included in the family substructure. In this situation, in fact, the G-X independence would be maintained. Conversely, the two partitions induced by the disease status and by the family groups intersect. This implies that the G-X independence within families is lost in the case-control sample.

	β_G			β_X			β_{GX}		
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
Recessive model									
$\sigma_U^2 = 0.7$									
NAIVE	0.497	0.183	0.236	-0.147	0.047	0.116	-0.504	0.090	0.000
RC	0.383	0.269	0.774	-0.081	0.086	0.932	-0.381	0.168	0.320
CLR	0.254	0.249	0.792	0.002	0.079	0.961	0.129	0.254	0.972
CC	-0.029	0.260	0.967	0.013	0.079	0.969	0.067	0.211	0.967
$\sigma_U^2 = 1.0$									
NAIVE	0.576	0.178	0.091	-0.181	0.043	0.014	-0.592	0.080	0.000
RC	0.442	0.263	0.646	-0.099	0.083	0.836	-0.447	0.156	0.120
CLR	0.324	0.259	0.706	0.004	0.085	0.965	0.137	0.288	0.981
CC	-0.029	0.275	0.975	0.017	0.086	0.975	0.077	0.236	0.977
Dominant model									
$\sigma_U^2 = 0.7$									
NAIVE	0.506	0.178	0.189	-0.146	0.054	0.223	-0.509	0.086	0.000
RC	0.384	0.261	0.742	-0.081	0.099	0.938	-0.384	0.161	0.248
CLR	0.246	0.244	0.811	0.007	0.091	0.944	0.118	0.231	0.970
CC	-0.039	0.249	0.950	0.023	0.092	0.954	0.062	0.196	0.968
$\sigma_U^2 = 1.0$									
NAIVE	0.590	0.174	0.055	-0.180	0.049	0.057	-0.596	0.077	0.000
RC	0.444	0.255	0.628	-0.099	0.095	0.876	-0.449	0.150	0.098
CLR	0.319	0.252	0.735	0.011	0.099	0.951	0.130	0.263	0.976
CC	-0.040	0.264	0.941	0.028	0.100	0.947	0.074	0.221	0.967

Table 1: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 1,000$ matched case-control pairs, under recessive or dominant genetic model. Classical measurement error with variance $\sigma_U^2 \in \{0.7, 1.0\}$. X distributed as a $\log \chi_2^2$ in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.

	β_G			β_X			β_{GX}		
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
				Recessive model					
				$\sigma_U^2 = 0.7$					
NAIVE	0.367	0.195	0.536	-0.105	0.043	0.314	-0.361	0.095	0.058
RC	0.247	0.283	0.932	-0.058	0.074	0.944	-0.247	0.169	0.692
CLR	0.277	0.258	0.814	0.012	0.066	0.950	0.298	0.266	0.908
CC	-0.039	0.232	0.952	0.028	0.067	0.948	0.103	0.180	0.970
				$\sigma_U^2 = 1.0$					
NAIVE	0.448	0.195	0.336	-0.136	0.042	0.086	-0.448	0.085	0.002
RC	0.301	0.274	0.886	-0.076	0.072	0.874	-0.310	0.154	0.450
CLR	0.369	0.285	0.732	0.016	0.074	0.950	0.402	0.367	0.948
CC	-0.054	0.254	0.956	0.032	0.074	0.944	0.158	0.205	0.960
				Dominant model					
				$\sigma_U^2 = 0.7$					
NAIVE	0.370	0.192	0.510	-0.110	0.049	0.382	-0.354	0.090	0.038
RC	0.262	0.276	0.928	-0.062	0.085	0.954	-0.247	0.156	0.666
CLR	0.245	0.255	0.844	0.004	0.075	0.954	0.306	0.244	0.862
CC	-0.084	0.230	0.940	0.026	0.077	0.956	0.135	0.175	0.920
				$\sigma_U^2 = 1.0$					
NAIVE	0.453	0.186	0.314	-0.142	0.046	0.156	-0.441	0.081	0.002
RC	0.319	0.268	0.866	-0.079	0.082	0.900	-0.310	0.144	0.374
CLR	0.328	0.271	0.782	0.007	0.082	0.952	0.415	0.308	0.878
CC	-0.110	0.247	0.948	0.031	0.085	0.958	0.196	0.207	0.912

Table 2: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 1,000$ matched case-control pairs, under recessive or dominant genetic model. Classical measurement error with variance $\sigma_U^2 \in \{0.7, 1.0\}$. X distributed as a mixture of normals in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.

		β_G			β_X			β_{GX}	
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
				Asymmetric measurement error structure					
NAIVE	0.113	0.208	0.895	-0.104	0.051	0.461	-0.388	0.103	0.048
RC	-0.026	0.470	0.998	-0.056	0.130	0.988	-0.285	0.272	0.888
CLR	0.169	0.244	0.867	0.004	0.073	0.956	0.119	0.221	0.972
CC	-0.030	0.240	0.940	0.014	0.073	0.952	0.053	0.181	0.966

Table 3: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 1,000$ matched case-control pairs, under recessive genetic model and asymmetric measurement error structure. X distributed as a $\log \chi^2_2$ in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.

		β_G			β_X			β_{GX}	
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
				Multiplicative measurement error structure					
NAIVE	0.389	0.192	0.494	-0.029	0.056	0.888	-0.147	0.129	0.712
RC	-0.609	0.611	0.932	0.005	0.138	0.994	-0.064	0.365	0.990
CLR	0.285	0.252	0.794	0.007	0.064	0.966	0.131	0.203	0.954
CC	-0.008	0.230	0.948	0.021	0.064	0.944	0.035	0.161	0.956

Table 4: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 1,000$ matched case-control pairs, under recessive genetic model and multiplicative measurement error structure. X distributed as a mixture of normals in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.

	$\beta_{GX} = 1.0$			$\beta_{GX} = 1.5$			$\beta_{GX} = 2.0$		
	β_X			β_X			β_X		
β_G	log(1.5)	log(2.0)	1.0	log(1.5)	log(2.0)	1.0	log(1.5)	log(2.0)	1.0
log(1.3)	0.612	0.499	0.355	0.530	0.546	0.484	0.365	0.430	0.308
log(1.8)	0.349	0.241	0.729	0.331	0.529	0.589	0.516	0.436	0.446
log(2.3)	0.571	0.416	0.558	0.522	0.461	0.516	0.559	0.568	0.378

Table 5: P-values of the test of independence of environmental exposure X between cases and controls, obtained from $K = 200$ simulated matched case-control pairs, for $\beta_G \in \{\log(1.3), \log(1.8), \log(2.3)\}$, $\beta_E \in \{\log(1.5), \log(2.0), 1.0\}$ and $\beta_{GX} \in \{1.0, 1.5, 2.0\}$, under recessive genetic model.

	β_G			β_X			β_{GX}		
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
				X distributed as a $\log\chi^2_2$					
NAIVE	0.010	0.133	0.900	-0.168	0.052	0.120	0.002	0.065	0.950
RC	-0.008	0.193	0.940	-0.095	0.101	0.910	0.003	0.126	0.955
CLR	0.025	0.150	0.920	0.034	0.105	0.920	-0.002	0.129	0.950
CC	0.011	0.149	0.940	0.038	0.099	0.930	-0.008	0.114	0.950
				X distributed as a mixture of normals					
NAIVE	0.001	0.134	0.960	-0.137	0.047	0.220	-0.008	0.059	0.970
RC	-0.004	0.136	0.980	-0.077	0.062	0.760	-0.003	0.076	0.950
CLR	0.016	0.142	0.960	0.024	0.083	0.950	-0.015	0.101	0.970
CC	0.005	0.136	0.950	0.022	0.077	0.950	-0.013	0.088	0.960

Table 6: Bias, estimated standard error (S.e.) and empirical coverage of confidence intervals of nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 100 replicates of $K = 500$ matched case-control pairs, under recessive genetic model and no *G*-*X* interaction. Classical measurement error with variance $\sigma^2_U = 1$. X distributed as a $\log\chi^2_2$ or as a mixture of normals in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 0.000$.

	β_G			β_X			β_{GX}		
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
					Recessive model				
					$\sigma_U^2 = 0.7$				
NAIVE	0.506	0.260	0.512	-0.140	0.066	0.425	-0.509	0.128	0.057
RC	0.386	0.407	0.952	-0.074	0.125	0.956	-0.380	0.248	0.678
CLR	0.260	0.364	0.867	0.017	0.113	0.967	0.144	0.374	0.985
CC	-0.034	0.369	0.952	0.027	0.113	0.969	0.070	0.300	0.972
					$\sigma_U^2 = 1.0$				
NAIVE	0.583	0.253	0.366	-0.175	0.061	0.170	-0.597	0.113	0.002
RC	0.445	0.397	0.916	-0.092	0.121	0.934	-0.446	0.229	0.468
CLR	0.329	0.379	0.811	0.021	0.133	0.968	0.158	0.467	0.975
CC	-0.036	0.394	0.963	0.033	0.124	0.972	0.079	0.341	0.972
					Dominant model				
					$\sigma_U^2 = 0.7$				
NAIVE	0.516	0.254	0.473	-0.145	0.076	0.519	-0.507	0.123	0.042
RC	0.387	0.387	0.922	-0.078	0.145	0.972	-0.386	0.235	0.656
CLR	0.244	0.356	0.870	0.013	0.130	0.956	0.167	0.355	0.975
CC	-0.041	0.356	0.950	0.028	0.132	0.962	0.072	0.282	0.966
					$\sigma_U^2 = 1.0$				
NAIVE	0.598	0.248	0.311	-0.180	0.070	0.292	-0.593	0.110	0.002
RC	0.447	0.379	0.888	-0.096	0.140	0.948	-0.452	0.219	0.420
CLR	0.312	0.373	0.825	0.015	0.143	0.961	0.197	0.417	0.976
CC	-0.049	0.385	0.952	0.034	0.146	0.963	0.089	0.327	0.972

Table 7: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 500$ matched case-control pairs, under recessive or dominant genetic model. Classical measurement error with variance $\sigma_U^2 \in \{0.7, 1.0\}$. X distributed as a $\log \chi_2^2$ in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.

		β_G			β_X			β_{GX}	
	Bias	S.e.	Coverage	Bias	S.e.	Coverage	Bias	S.e.	Coverage
					Recessive model				
					$\sigma_U^2 = 0.7$				
NAIVE	0.358	0.279	0.772	-0.106	0.062	0.574	-0.337	0.137	0.340
RC	0.242	0.429	0.986	-0.054	0.107	0.970	-0.233	0.253	0.896
CLR	0.258	0.381	0.906	0.012	0.094	0.962	0.414	0.431	0.988
CC	-0.072	0.335	0.962	0.028	0.096	0.958	0.141	0.265	0.978
					$\sigma_U^2 = 1.0$				
NAIVE	0.440	0.271	0.636	-0.138	0.058	0.318	-0.427	0.122	0.090
RC	0.298	0.416	0.978	-0.072	0.104	0.958	-0.298	0.232	0.776
CLR	0.374	0.419	0.864	0.015	0.102	0.968	0.553	0.572	0.986
CC	-0.119	0.417	0.970	0.033	0.104	0.960	0.204	0.316	0.980
					Dominant model				
					$\sigma_U^2 = 0.7$				
NAIVE	0.373	0.273	0.740	-0.111	0.070	0.618	-0.347	0.128	0.250
RC	0.260	0.409	0.984	-0.062	0.124	0.968	-0.232	0.233	0.896
CLR	0.237	0.373	0.890	0.005	0.107	0.952	0.377	0.380	0.966
CC	-0.110	0.333	0.952	0.027	0.110	0.962	0.167	0.258	0.952
					$\sigma_U^2 = 1.0$				
NAIVE	0.456	0.266	0.598	-0.142	0.066	0.402	-0.436	0.115	0.058
RC	0.316	0.397	0.968	-0.080	0.120	0.944	-0.295	0.216	0.784
CLR	0.353	0.404	0.860	0.009	0.116	0.962	0.493	0.475	0.988
CC	-0.127	0.363	0.954	0.033	0.120	0.970	0.233	0.308	0.968

Table 8: Bias, estimated standard error (S.e.) and empirical coverage of confidence interval at nominal level 0.95 for the estimators of β_G, β_X and β_{GX}, obtained from the *naive* analysis (NAIVE), regression calibration (RC), conditional logistic regression (CLR) and conditional likelihood of Chatterjee *et al.* (2005) (CC). Results based on 500 replicates of $K = 500$ matched case-control pairs, under recessive or dominant genetic model. Classical measurement error with variance $\sigma_U^2 \in \{0.7, 1.0\}$. X distributed as a mixture of normals in the source population. True values of the parameters: $\beta_G = \log(1.3) = 0.262$, $\beta_X = \log(1.5) = 0.405$, $\beta_{GX} = 1.000$.