Working Paper Series, N. 8, June 2007

# A Flexible Approach to Measurement Error Correction in Case-Control Studies

**A. Guolo**

Department of Statistical Sciences

University of Padua

Italy

**Abstract:** We investigate the use of prospective likelihood methods to analyze retrospective case-control data where some of the covariates are measured with error. We show that prospective methods can be applied and the case-control sampling scheme can be ignored if one adequately models the distribution of the error-prone covariates in the case-control sampling scheme. Indeed, subject to this, the prospective likelihood methods result in consistent estimates and information standard errors are asymptotically correct. However, the distribution of such covariates is not the same in the population and under case-control sampling, dictating the need to model the distribution flexibly. In this paper, we illustrate the general principle by modeling the distribution of the error-prone covariates using the skewnormal distribution. The performance of the method is evaluated through simulation studies, which show satisfactory results in terms of bias and coverage. Finally, the method is applied to the analysis of two data sets which refer, respectively, to a cholesterol study and a study on breast cancer.

**Keywords:** Likelihood, Logistic regression, Measurement error, Regression calibration, Retrospective study, Skewnormal distribution.

# 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: +39 049 827 4192
guolo@stat.unipd.it
http://www.stat.unipd.it/~guolo

# A Flexible Approach to Measurement Error Correction in Case-Control Studies

**A. Guolo**

Department of Statistical Sciences

University of Padua

Italy

**Abstract:** We investigate the use of prospective likelihood methods to analyze retrospective case-control data where some of the covariates are measured with error. We show that prospective methods can be applied and the case-control sampling scheme can be ignored if one adequately models the distribution of the error-prone covariates in the case-control sampling scheme. Indeed, subject to this, the prospective likelihood methods result in consistent estimates and information standard errors are asymptotically correct. However, the distribution of such covariates is not the same in the population and under case-control sampling, dictating the need to model the distribution flexibly. In this paper, we illustrate the general principle by modeling the distribution of the error-prone covariates using the skewnormal distribution. The performance of the method is evaluated through simulation studies, which show satisfactory results in terms of bias and coverage. Finally, the method is applied to the analysis of two data sets which refer, respectively, to a cholesterol study and a study on breast cancer.

**Keywords:** Likelihood, Logistic regression, Measurement error, Regression calibration, Retrospective study, Skewnormal distribution.

# 1 Introduction

The problem of erroneously measuring variables is common in many scientific areas, as, for example, in biology, epidemiology, econometrics. It has long been recognized that ignoring the presence of measurement errors in statistical analyses can lead to bias of estimators, reduced power of tests and inaccurate coverage probabilities of confidence intervals (Armstrong, 2003). To alleviate these problems, many correction techniques have been proposed, see Carroll *et al.* (2006) for a detailed review.

In this paper, we take up the consideration of measurement error analysis for population-based case-control studies. Our goal is to develop a likelihood approach to this problem. In keeping with much of the literature in case-control studies, we will take a prospective approach, i.e., compute maximum likelihood estimators and make likelihood inferences ignoring the case-control study and treating the data as if it arises from a random sampling framework. This problem and approach, while seemingly simple, have not been considered in detail in the literature. Our main theoretical result is to show that, if one properly models the distribution of the error-prone covariates *in the case-control sampling scheme*, then prospective likelihood estimation and inferences are asymptotically correct.

Section 2 describes our notation, and in Section 2.2 we show a feature about prospective approaches to case-control data, namely that the distribution of the mismeasured covariates in the population differs from that in the case-control sampling scheme. This suggests the need for flexible families for this distribution in the case-control sampling scheme. While our point is quite general, we focus here on the skewnormal family of distributions (Azzalini, 1985). Section 3 reviews the various methods we will compare.

In Section 4, we show the new result, i.e. that if the distribution of the error-prone covariates is properly modeled in the case-control sampling scheme, then likelihood approaches are asymptotically valid. This point is related to a general discussion in Carroll *et al.* (1995), where the Authors do not actually consider the case of a prospective likelihood analysis of measurement error data, and especially they do not note the essential modeling requirement.

Section 5 gives the results of simulation studies that indicate the strength and applicability of prospective likelihood methods for measurement error models in case-control data. Finally, Section 6 illustrates the results of the application of our method to the analysis of two data examples. The first refers to a study on blood cholesterol level as risk factor for coronary heart disease, while the second one refers to a study about nutrition habits and occurrence of breast cancer.

# 2 Models

## 2.1 Notation

Suppose that case-control data are available. Let $D$ be the case ($D = 1$) or control ($D = 0$) status. Let $X$ be the set of covariates which are not directly observed. Instead of $X$, the mismeasured variables $W$ are observed. Other variables $Z$ can be observed with no measurement error. Suppose that $D$ is related to $X$ and $Z$ through the so-called *disease model*, whose density function is $f_{D|XZ}(d|x, z; \beta)$. In case-control studies, the logistic regression model is typically used, $f_{D|XZ}(d|x, z; \beta) = H(\beta_0 + x^T\beta_1)^d \left\{1 - H(\beta_0 + x^T\beta_1)\right\}^{1-d}$, where $H(\cdot)$ is the logistic distribution function, $H(v) = \{1 + \exp(-v)\}^{-1}$. Moreover, let $f_{W|XZ}(w|x, z; \gamma)$ be the density function of the model relating $W$ to $X$ and

$Z$ through the so-called nondifferential *measurement error model*. Likelihood methods also require an *exposure model*, that is, a model for the unobserved $X$ possibly depending on $Z$, whose density function is $f_{X|Z}(x|z;\delta)$. The inferential interest typically focuses on the vector of parameters $\beta_1$, which explains the influence of the unobserved $X$ on the disease status indicator. It is well known that if inferential analyses are performed by ignoring the presence of measurement error, that is a *naive* analysis in carried out, the results can be misleading, sometimes seriously.

## 2.2 Distribution of $X$ in the case-control sampling scheme

The usual approach to the analysis of case-control data is to ignore the case-control sampling scheme and to pretend that the data are collected according to a prospective sampling design. In the case of no measurement error, the equivalence between a prospective and a retrospective analysis of case-control data is proved by Prentice and Pyke (1979) for logistic regression. This result has been often invoked in literature to justify the use of the logistic regression model for case-control data analysis. Carroll *et al.* (1995) extend the results by Prentice and Pyke (1979) to models with missing data. They show that the prospective analysis of case-control data often works to estimate consistently all but the intercept parameter in the logistic models, and that in general it is expected to work. Further, they go on to show that if a prospective analysis yields consistent estimates, then standard errors from that analysis are at worst conservative, and often exact. However, Carroll *et al.* (1995) do not show that a prospective likelihood analysis in the type of measurement error analysis of interest to us actually results in consistent estimation of all but the intercept parameter.

In the measurement error context, since $X$ is a latent variable, care must be taken in formulating a likelihood analysis if one wishes to pursue methods that ignore the case-control sampling scheme. Indeed, the distribution of $X$ is the population does not equal that of $X$ in the case-control sample. To see this, consider Figure 1. We generated 10,000 observations from the scalar variable $X$ distributed according to a mixture of Lognormal$(-2.3, 0.9)$ and Lognormal$(-1.5, 0.9)$, with mixing weights 0.8 and 0.2. We performed case-control sampling, according to the probability of disease given by the logistic function, $\text{pr}(D = 1|X) = H(\beta_0 + \beta_1 X)$, by setting $\beta_0 = -1.5$ and $\beta_1$ assuming one of the values $(0.5; 1.2; 2.0)$. Figure 1 displays the density function of $X$ in the population (solid line) and in the case-control sample (dashed line). The discrepancies between the densities are evident and become greater as the value of $\beta_1$ increases. Similar results (not shown here) are obtained under different distributions for $X$.

# 3 Correction Techniques

## 3.1 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 idea underlying the method is to replace the unknown values of $X$ by an estimate of the expectation of $X$ given $(W, Z)$, that is $E[X|W, Z] = X^*$, by using additional information. Then, a standard inferential process on the observations from $(D, X^*)$ can be run.

## 3.2 Likelihood Methods

The likelihood approach for measurement error correction has received less attention in the measurement error literature with respect to alternatives. This is mainly due to its computational complexity and to the difficulties in checking the parametric assumptions it requires, especially that of the unobserved variables $X$. Nevertheless, some recent results have shown the advantages of the likelihood method, mainly based on the large sample optimality properties of the corresponding estimators (Schafer and Purdy, 1996; Küchenhoff and Carroll, 1997).

Here we pursue the idea of ignoring the case-control sampling scheme and applying standard likelihood methods. We will show in Section 4 that, as long as we adequately model the distribution of $X$ in the case-control sampling scheme, then prospective likelihood methods lead to consistent estimation and correct inference. From the discussion in Section 2.2, however, it is clear that a flexible family of distributions is required.

Here is how a consistent prospective likelihood approach can be implemented. Suppose that $n$ independent observations from $(D, W, Z)$ are available. Then, the likelihood is obtained by integrating out the product of the model densities with respect to the unknown quantity $X$

$$L(\beta_0, \beta_1, \gamma, \delta) = \prod_{i=1}^{n} \int f_{D|XZ}(D_i|x, Z_i; \beta_0, \beta_1) f_{W|XZ}(W_i|x, Z_i; \gamma) f_{XZ}(x|Z_i; \delta) dx. \tag{1}$$

The integral is replaced by a sum if $X$ is a discrete random variable.

The parameters in (1) cannot usually be estimated without additional information about the measurement error model. Suppose that extra information is available in terms of internal validation data. This means that, for a small group of $m$ subjects, $m << n$, observations from $(D, X, Z)$ are recorded. To

take account of this, the likelihood has the following expression

$$L(\beta_0, \beta_1, \gamma, \delta) = \prod_{i=1}^{n} \int f_{D|XZ}(D_i|x, Z_i; \beta, \beta_1) f_{W|XZ}(W_i|x, Z_i; \gamma) f_{X|Z}(x|Z_i; \delta) dx \times \prod_{j=1}^{m} f_{D|XZ}(D_j|X_j, Z_j; \beta_0, \beta_1) f_{W|XZ}(W_j|X_j, Z_j; \gamma) f_{X|Z}(X_j|Z_j; \delta). \quad (2)$$

Similar modifications of the likelihood are defined to take account of other types of additional data, for example, external validation data or replicates (Schafer, 2002).

Carroll *et al.* (1999a, 1999b) and later Richardson *et al.* (2002), for example, flexibly model the distribution of interest through a mixture of normal variables. Here, we suggest to flexibly modeling the distribution of $X$ through the skewnormal distribution (Azzalini, 1985), $X \sim \text{SN}(\mu, \sigma, \alpha)$, which has density function

$$f_X(x; \delta) = f_X(x; \mu, \sigma, \alpha) = (2/\sigma)\phi\left\{(x-\mu)/\sigma\right\}\Phi\left\{\alpha(x-\mu)/\sigma\right\},$$

where $\delta = (\mu, \sigma, \alpha)^{\text{T}}$, $\mu, \sigma, \alpha$ are, respectively, the location, the scale and the shape parameter and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions. The exposure model can be easily extended to include error-free variables $Z$. Note once again that we will use this parametric family as one possible model for the distribution of $X$ in the case-control sampling scheme.

As we show in Section 4, the use of the skewnormal distribution in place of the exact distribution of $X$ is justified as long as the skewnormal is a good approximation of the distribution of $X$ in the case-control sampling scheme.

# 4 Theoretical Context

Let the number of cases and controls be $n_1$ and $n_0 = n - n_1$, respectively. Moreover, suppose that $n_1/n$ remains fixed as $n \to \infty$. For simplicity, suppose that there are no error-free variables $Z$, although the results can be easily extended to include them. Let $f_{X,\mathrm{cc}}(x)$ be the actual density function for $X$ in the case-control sampling scheme, and let $f_X(x, \xi)$ be a parametric family of density functions. Then, in practice, as long as $f_X(x, \xi)$ is a good approximation of $f_{X,\mathrm{cc}}(x)$, a prospective likelihood analysis of the case-control data is legitimate. More precisely, in the Appendix we show the following result.

**Theorem 1** Let $\pi_d = \mathrm{pr}(D = d)$ in the population, and define $\beta_0^* = \beta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Suppose that $n_1/n$ remains fixed as $n \to \infty$ and that the distribution of $X$ in the case-control sampling scheme is $f_X(x, \xi)$ for some true parameter $\xi$. Let $\theta = (\beta_0^*, \beta_1^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$. Let $\widehat{\theta}$ be the prospective likelihood estimate of $\theta$. Then $\widehat{\theta}$ is consistent for $\theta$. Further, standard error estimates for $\beta_1$ derived from prospective likelihood information-based calculations are at worst asymptotically conservative.

# 5 Simulation Studies

We performed different simulation studies in order to evaluate the behaviour of the likelihood approach to correct for measurement error affecting $X$, when the distribution of $X$ is flexibly modeled by the skewnormal (SN). The results are compared to those derived from the likelihood approach where the distribution of $X$ is the actual one in the case-control sampling scheme (LIK), to those provided by the RC method (RC) as well as to the *naive* results (NAIVE), that is, the ones obtained by ignoring the presence of measurement error. Our

interest focuses on the parameter $\beta_1$.

For simplicity, suppose that the unobservable and mismeasured $X$ is scalar. Extensions to the multi-dimensional case are straightforward from a theoretical point of view, while they can lead to an increased computational complexity because of the difficulties in solving the integrals in the likelihood expression (1) or (2). Moreover, suppose that there are no error-free variables $Z$.

## 5.1 Details

In our simulation studies we generate sets of case-control data of size $n = 600$. The generation is repeated 500 times. Different distributions for $X$ are taken into account, which are common in practice: a mixture of lognormal distributions, $\text{Ln}(\mu_1, \sigma^2_{ln})$ and $\text{Ln}(\mu_2, \sigma^2_{ln})$, with mixing weights 0.8 and 0.2, respectively, a $\chi^2_1$ distribution and a Weibull distribution, $\text{Weibull}(\mu_{wei}, \sigma_{wei})$. The binary outcome variable $D$ is generated under the model

$$\text{logit}\{\text{pr}(D = 1)|X\} = \beta_0 + \beta_1 X.$$

We fix $(\beta_0, \beta_1)^\top = (-1.5, 0.8)^\top$ and $(\mu_1, \mu_2, \sigma_{ln}, \mu_{wei}, \sigma_{wei})^\top = (-2.3, -1.5, 0.9, 1.4, 0.6)^\top$. The measurement error is assumed to be multiplicative, $W = X \exp\{U\}$, with $U \sim \text{Normal}(0, \sigma^2_U)$, as it is reasonable in many applications. Different amounts of measurement error are considered, $\sigma_U = \{0.3, 0.5, 0.75\}$. Simulations were performed using the `R` programming language (R Development Core Team, 2005), Version 2.2.1.

Both the RC and the likelihood analysis are performed by considering that additional information is available, together with the primary data set of observations from $(Y, W)$. In particular, we suppose that internal validation data have been collected, in form of observations from $(Y, X, W)$ for a small subset

of the primary data. The subsample is randomly selected as the 10% of the primary data.

The maximization of the likelihood function is performed by using the `R` routine `optim`, which is based on the optimization algorithm by Nelder and Mead (1965). An alternative algorithm based on the Newton-Raphson algorithm gives similar results. The likelihood maximization requires numerical evaluation of integrals. As suggested by Higdon and Schafer (2001) and by Schafer (2002), integrals are evaluated through Gauss-Hermite quadrature. A study (not shown here) that we performed to evaluate the accuracy of the integral evaluation with different number of Gauss-Hermite nodes indicated that 12 nodes is a satisfactory choice. Moreover, 12 nodes are sufficient to cover the range of values of $X$ which are simulated under the different distributions previously summarized.

The optimization algorithm requires finding reasonable initial estimates of parameters. We considered the *naive* estimators as initial estimate of the disease model parameters $(\beta_0, \beta_1)$. With respect to the exposure model we calculated moment-based estimators of the parameters using the additional data.

As explained in Section 4, the use of a flexible distribution as an alternative to the real distribution of $X$ in the case-control sample is justified as long as the approximation is good in the case-control sampling scheme. In the simulation studies we performed, an empirical evaluation shows that the skewnormal distribution is a satisfactory solution in that it is close to the distribution of $X$ in the case-control sampling scheme. This behaviour is shown in Figure 2. We plotted the density function of $X$ in the case-control sample (solid line), derived from 10000 observations generated from a mix-

ture of Lognormal$(-2.3, 0.9)$ and Lognormal$(-1.5, 0.9)$ with mixing weights 0.8 and 0.2, a $\chi^2_1$ and a Weibull$(1.4, 0.6)$. The probability of disease is given by $\text{logit}\{\text{pr}(D = 1|X)\} = \beta_0 + \beta_1 X$, with $(\beta_0, \beta_1)^{\text{T}} = (-1.5, 0.8)^{\text{T}}$. Then, we estimated the skewnormal parameters on the simulated data and added the density function of the skewnormal to the plot (dashed line). The approximation of the skewnormal density to the actual density of $X$ in the case-control sample is satisfactory, both for the values of $\beta_1$ given here and for larger values (results not shown).

## 5.2 Results

In Tables 1–3 we report the results of the simulation studies performed to test the behaviour of the correction techniques, when $X$ follows one of the abovementioned distributions, mixture of lognormals, $\chi^2_1$ and Weibull, and with three increasing amounts of measurement error.

The measurement error correction techniques we focus on are compared with respect to bias (Bias) and standard error (s.e.) of the corresponding estimators of $\beta_1$. Moreover, the associated standard deviations of these quantities are reported in parentheses. Finally, we compute the empirical coverage of confidence intervals. The $(1-\alpha)\%$ confidence interval is computed as $\widehat{\beta}_1 \pm z_{\alpha/2}$ times the estimated standard error of $\widehat{\beta}_1$, where $\widehat{\beta}_1$ is the estimate provided by the adopted correction technique or the *naive* analysis and $z_\alpha$ is the $\alpha^{\text{th}}$ quantile of the standard normal distribution. We focus on $\alpha = 0.05$ and $\alpha = 0.1$. The estimated standard error of the RC estimator is computed using the bootstrap, with 1000 bootstrap samples. For the maximum likelihood estimator, instead, both when the actual distribution of $X$ or the skewnormal distribution are used to model the exposure, and for the *naive* estimator we refer to the

estimated standard error provided by the Hessian matrix.

First of all, the simulation results highlight the need for correction techniques in order to improve the *naive* analysis. In fact, the *naive* approach experiences considerable bias of the estimator of $\beta_1$ and nominal levels of confidence intervals that overestimate the empirical coverages, sometimes seriously. This situation is emphasized under a $\chi^2_1$ or a Weibull distribution for $X$ and it gets worse as the measurement error variance increases. In this cases, RC provides less biased estimators, although they are affected by a larger variance. This variance obviously increases for larger measurement error variance. However, outside the first scenario where $X$ is assumed to follow a mixture of lognormal distributions, also RC experiences poor coverages of confidence intervals.

When applying the likelihood approach to error correction, instead, the advantages are seen in all the examined situations. Consider first the likelihood approach based on the actual distribution for $X$ in the case-control sample. As expected from our theory, simulations results indicate that the method globally performs very satisfactorily, mainly if we consider the coverage of confidence intervals as evaluation criterion. Moreover, these results do not seem to be affected by the increasing variance of meausurement error. If we focus now on the flexible approach we suggest to correct for measurement error, simulation results are encouraging. The use of the skewnormal as a flexible tool to describe the distribution of $X$ is satisfactory in order to correct for measurement error in all the examined situations. The method provides an estimator of $\beta_1$ which has bias and standard error comparable to those from the likelihood approach based on the actual distribution of $X$. This behaviour is maintained also under increasing values of the measurement error variance. If we consider

the empirical coverages of confidence intervals, the method provides values that are close to the nominal ones. They are almost always close to those provided by the likelihood approach based on the actual distribution of $X$. Small discrepancies are related to the specification of a Weibull distribution for $X$, under large meausurement error variance.

# 6 Examples

In this section, we report the results of the application of our flexible approach to measurement error correction in situations referred to two different data sets. The first example refers to a cholesterol study, while the second to a study on breast cancer.

## 6.1 A Cholesterol Study

The first data set refers to a study on the risk of coronary heart disease (CHD) as a function of blood cholesterol level. These data are extracted from the Lipids Research Clinics study, which was previously discussed by Satten and Kupper (1993). Later, a portion of these data, involving men aged 60-70 who do not smoke, for a total of 256 records, have been analyzed by Roeder *et al.* (1996). The case status ($D = 1$) occurs if a subject has had a heart attack, an abnormal exercise electrocardiogram, history of angina pectoris, and so forth. Covariates are low-density lipoprotein (LDL) cholesterol level and total cholesterol level (TCL). TCL may be considered as a surrogate of LDL, whose direct measure is expensive and time consuming. The error in measuring TCL relies in that it provides a measure of LDL plus unknown quantities of other components as triglycerides and high density lipoprotein. According to the

notation used above, $CHD = D$, $X = LDL/100$, $W = TCL/100$. Using $X$ as the predictor, we obtain an estimate of $\beta_1$ equal to 0.656, with a standard error of 0.336. The *naive* analysis, instead, provides an estimate of $\beta_1$ equal to 0.549, with a standard error of 0.313.

In examining this problem, Roeder *et al.* (1996) suggest that a nondifferential lognormal measurement error, that translates into a multiplicative error in the natural scale, adequately fits the data. They correct for measurement error by a semiparametric approach, in which the marginal distribution of $X$ is modeled through a nonparametric mixture distribution. In order to illustrate the information contained in the data, they analyze a sample of data with complete and reduced observations. The complete data are randomly selected as the 28% of the data set. Using only the complete data, the estimate of $\beta_1$ is equal to 0.943, with a standard error of 0.620. Instead, using complete and reduced data provide an estimate equal to 0.765.

We analyzed the same data, by using the likelihood approach with the distribution of $X$ flexibly modeled through the skewnormal. We randomly selected the complete data from $(Y, X, W)$ as the 10%, 15%, 20% and 25% of the data set, in order to evaluate the change of the results with respect to the amount of additional information provided. The fitted skewnormal results in a good approximation for the distribution of $X$, as can be seen in Figure 3 with reference to complete data equal to 15% and 25% of the data set. Our approach results in an estimate of $\beta_1$ equal to 0.629 (s.e.= 0.388), 0.634 (s.e.= 0.336), 0.624 (s.e.= 0.378), 0.676 (s.e.= 0.346), for the four different amounts of complete data, respectively. The estimate is close to the one based on the LDL measures, that is 0.656. Moreover, it can be noted that the estimates only change slightly as the amount of the complete data decreases, thus adding to

the results a measure of robustness with respect to the amount of additional information.

## 6.2 A Breast Cancer Study

The second data set we focus on refers to the NHANES-I Epidemiologic Study Cohort (Jones *et al.*, 1987), originally consisting of 8596 women who were interviewed about their nutrition habits and later examined for evidence of cancer. Carroll *et al.* (2006, Section 4.3) examined a portion of the data, involving 3145 women aged 20-50, who have no missing data on the variables of interest. The case status ($D = 1$) occurs in the presence of breast cancer. There are 59 cases of breast cancer in the data. The interest focuses on the long-term saturated fat intake, which in our notation plays the role of $X$. Instead of $X$, a measure $W$ is collected, which is a 24-hour recall done by the partecipants. Together with $W$ other variables are observed, age, poverty index ratio, body mass index, assumption of alcohol, family history of breast cancer, age at menarche and menopausal status. They are supposed to be correctly measured. In our notation these variables are indicated by $Z$.

The saturated fat is measured with considerable error (Beaton *et al.*, 1979; Wu *et al.*, 1986). This led to considerable controversy as regards its use to assess breast cancer risk (Prentice *et al.*, 1989; Willett *et al.*, 1987). Moreover, in the NHANES data there is no information enough to adequately describe the measurement error structure. By using external validation data, where the measurement error is assumed additive on a logarithmic scale, Carroll *et al.* (2006, Section 4.3) suggest that the measurement error variance can be set equal to 0.171. Furthermore, they estimate that over 75% of the variance of the 24-hour recall is made up by measurement error. A *naive* analysis performed

by Carroll *et al.* (2006, Section 4.3) turns out in a negative logistic regression coefficient of saturated fat, equal to -0.97 (s.e.= 0.29). This results agrees with the nonparametric density estimate of the risk factor computed for cases and controls, which indicates a protective effect of higher levels of saturated fat in the diet (Figure 4). However it is in opposition with one popular hypothesis on the influence of saturated fat in diet. With respect to variables $Z$, only the age and menopausal status are significant predictors of risk. Carroll *et al.* (2006, Section 4.3) correct for measurement error in $X$ through regression calibration. Their estimate of the influence of saturated fat in the diet is equal to -4.67 (s.e.= 2.26), with a 95% confidence interval ranging from -10.37 to -1.38.

We analyzed the same data, by using the likelihood approach with the distribution of $X$ flexibly modeled through the skewnormal. We assume that the measurement error variance is known and equal to 0.171. The analysis provides an estimate of the influence of saturated fat on risk cancer equal to -3.26 (s.e.= 1.59). This estimate indicates a negative effect of saturated fat intake on the risk of breast cancer, smaller and less variable than the one provided by RC. The associated 95% confidence interval ranges from -6.39 to -0.14. With respect to variables $Z$, our analysis indicates that the age, the poverty index ratio, the body mass index and the menopausal status of the subjects are significant predictors of risk.

# 7 Conclusions

We have investigated the use of prospective likelihood methods in the analysis of case-control data with measurement error affecting a covariate $X$. We showed that properly modeling the distribution of the mismeasured covariates

*in the case-control sampling scheme* results in asymptotically valid inferences. Because of the fact that the distribution of $X$ in the population differs from that in the case-control sampling scheme, we proposed to use flexible families of distributions for $X$, illustrating with the skewnormal family of distributions.

Simulations indicate that the prospective likelihood approach can work very well in terms of corrections for bias and achieving nominal confidence levels. The method not surprisingly works better than such standard devices as regression calibration. Thus, properly constructed, likelihood analysis of case-control studies subject to covariate measurement error is a viable option.

## Acknowledgements

This research was partially supported by the Italian Ministry for Education, University and Research.

## References

[1] Armstrong, B. (2003). Exposure measurement error: consequences and design issues. In *Exposure Assessment in Occupational and Environmental Epidemiology* (M. J. Nieuwenhuijsen, Ed.), Oxford University Press, Oxford.

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

[3] Beaton, G.H., Milner, J. and Little, J.A. (1979). Sources of variation in 24-hour dietary recall data: implications for nutrition study design and interpretation. *American Journal of Clinical Nutrition*, **32**, 2546–2559.

[4] Carroll, R.J., Maca, J.D. and Ruppert, D. (1999a). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541–554.

[5] Carroll, R.J., Roeder, K. and Wasserman, L. (1999b). Flexible parametric measurement error models. *Biometrics*, **55**, 44–54.

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

[7] Carroll, R.J., Wang, S. and Wang, C.Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, **90**, 157–169.

[8] Higdon, R. and Schafer, D.W. (2001). Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis*, **35**, 283–299.

[9] Jones, D.Y., Schatzkin, A., Green, S.B., Block, G., Brinton, L.A., Ziegler, R.G., Hoover, R. and Taylor, P.R. (1987). Dietary fat and breast cancer in the National Health and Nutrition Survey I: Epidemiologic follow-up study. *Journal of the National Cancer Institute*, 79, 465–471.

[10] Küchenhoff, H. and Carroll, R.J. (1997). Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine*, **16**, 169–188.

[11] Nelder, J.A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, **7**, 308-313.

[12] Prentice, R.L., Pepe, M. and Self, S.G. (1989). Dietary fat and breast cancer: a review of the literature and a discussion of methodologic issues. *Cancer Research*, **49**, 3147–3156.

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

[14] R Development Core Team (2005). *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`.

[15] Richardson, S., Leblond, L., Jaussent, I. and Green, P.J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A*, **165**, 549–566.

[16] Roeder, K., Carroll, R.J. and Lindsay, B.G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, **91**, 722–732.

[17] Satten, G.A. and Kupper, L.L. (1993). Inferences about exposure-disease association using probability of exposure information. *Journal of the American Statistical Association*, **88**, 200–208.

[18] Schafer, D. (2002). Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computation and Simulation*, **72**, 33–45.

[19] Schafer, D.W. and Purdy, K.G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika*, **83**, 813–824.

[20] Willett, W.C., Meir, J.S., Colditz, G.A., Rosner, B.A., Hennekens, C.H. and Speizer, F.E. (1987). Dietary fat and the risk of breast cancer. *New England Journal of Medicine*, **316**, 22–25.

[21] Wu, M.L., Whittemore, A.S. and Jung, D.L. (1986). Errors in reported dietary intakes. *American Journal of Epidemiology*, **124**, 826–835.

# A : Proof of Theorem 1

## A.1 Prospective Formulation

To prove the asymptotic validity of the method, make the following definitions. Let $f_{X|D}(\cdot)$ be the density of $X$ given $D$. Let $p_d = n_d/n$, $\pi_d = \mathrm{pr}(D = d)$, $\beta_0^* = \beta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Let $f_{X,\mathrm{cc}}(x) = \sum_{d=0}^{1} p_d f_{X|D}(x|D = d)$ be the density function of $X$ in the case-control study.

According to the methodology, we parameterize the density of $X$ in the case control study as $f_{X,\mathrm{cc}}(x) = f_X(x, \xi)$.

The prospective loglikelihood function, in which we explicitly indicate the dependence on the data for clarity, is

$$L(\beta_0^*, \beta_1, \xi; d, w) = \log\Big(\int H\{\beta_0^* + m(x, \beta_1)\}^d [1 - H\{\beta_0^* + m(x, \beta_1)\}]^{1-d} \times f_{W|X}(w|x) f_X(x, \xi) dx\Big),$$

where $m(\cdot)$ is a known and arbitrary function of $\beta_1$ and $x$. The most common choice is $m(x, \beta_1) = x^T \beta_1$. Let $\theta = (\beta_0^*, \beta_1^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$. The score function is

$$U(\theta) = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} L(\beta_0^*, \beta_1, \xi; D_i, W_i).$$

A major condition for the prospective method to consistently estimate $\theta$ is $E_{\mathrm{cc}}\{U(\theta)\} = 0$, where $E_{\mathrm{cc}}(\cdot)$ is expectation in the case-control sampling

scheme. Specifically, we must show

$$\begin{aligned} 0 &= \sum_{d=0}^{1} p_d E\left\{\frac{\partial}{\partial\theta} L(\beta_0^*, \beta_1, \xi; d, W)|D=d\right\} \\ &= \int \sum_{d=0}^{1} p_d f_{W|D}(w|D=d)\frac{\partial}{\partial\theta} L(\beta_0^*, \beta_1, \xi; d, w)dw. \end{aligned} \tag{3}$$

## A.2 Proof of (3)

Showing (3) directly algebraically as in Carroll *et al.* (1995) is complex. Instead, our approach is to define a new "pretend" study that is an actual prospective study, and show that the expectation of the derivative of the log-likelihood in this alternative sampling framework is exactly the right hand side of (3), which hence equals zero by properties of likelihood functions.

Consider a random sample from the population, but let $\delta = 1$ mean that $(D, W)$ are observed while $\delta = 0$ means that they are not. Suppose that a correctly specified parameterized model $f_{X|\delta=1}(x, \xi, |\delta = 1)$ is available for the density function of $X$ given $\delta = 1$. The probability of observing $(D, W)$ is

$$\text{pr}(\delta = 1|D, W, X) = \text{pr}(\delta = 1|D) = \frac{p_d/\pi_d}{p_0/\pi_0 + p_1/\pi_1} \propto p_d/\pi_d.$$

Marginally, in this sampling scheme,

$$\text{pr}(\delta = 1) = \sum_{d=0}^{1} \pi_d \text{pr}(\delta = 1|D = d)\{p_0/\pi_0 + p_1/\pi_1\}^{-1}.$$

Because

$$H\{\beta_0 + m(x, \beta_1)\}^d = \exp[d\{\beta_0 + m(x, \beta_1)\}][1 - H\{\beta_0 + m(x, \beta_1)\}],$$

an easy calculation shows that the observed data satisfy

$$\text{pr}(D = 1|X, W, \delta = 1) = H\{\beta_0^* + m(X, \beta_1)\}.$$

Further, with a slight abuse of notation, since $W$ is independent of $D$ given $X$, and since $\delta$ depends only on $D$, we have that

$$\begin{aligned} f_{DXW|\delta=1}(d,x,w|\delta=1) &= \text{pr}(D=d|W,X,\delta=1)f_{W|X,\delta=1}(w|x,\delta=1)f_{X|\delta=1}(x|\delta=1) \\ &= \text{pr}(D=d|X,\delta=1)f_{W|X}(w|x)f_{X|\delta=1}(x|\delta=1) \\ &= H\{\beta_0^*+m(x,\beta_1)\}^d[1-H\{\beta_0^*+m(x,\beta_1)\}]^{1-d}f_{W|X}(w|x) \\ &\quad f_{X|\delta=1}(x|\delta=1). \end{aligned}$$

However,

$$\begin{aligned} f_{X|\delta=1}(x,\xi,|\delta=1) &= \sum_{d=1}^{1} f_{DX,\delta=1}(d,x,\delta=1)/\text{pr}(\delta=1) \\ &= \sum_{d=1}^{1} \text{pr}(\delta=1|D=d,X)f_{X|D}(x|d)\text{pr}(D=d)/\text{pr}(\delta=1) \\ &= \sum_{d=1}^{1}(p_d\pi_d)f_{X|D}(x|d)\pi_d \\ &= \sum_{d=1}^{1} p_d f_{X|D}(x|d) = f_X(x,\xi). \end{aligned}$$

In other words, $f_X(x,\xi)$ is the properly parameterized version of $f_{X|\delta=1}(x|\delta=1)$ in this alternative sampling scheme. Collecting terms, we see that

$$\begin{aligned} &f_{D,X,W|\delta=1}(d,x,w|\delta=1) \\ &\quad = H\{\beta_0^*+m(x,\beta_1)\}^d[1-H\{\beta_0^*+m(x,\beta_1)\}]^{1-d}f_{W|X}(w|x)f_X(x,\xi), \end{aligned}$$

and thus

$$\begin{aligned} f_{D,W|\delta=1}(d,w|\delta=1) &= \int H\{\beta_0^*+m(x,\beta_1)\}^d[1-H\{\beta_0^*+m(x,\beta_1)\}]^{1-d} \\ &\qquad \times f_{W|X}(w|x)f_X(x,\xi)dx \\ &= \exp\{L(\beta_0^*,\beta_1,d,w,\xi)\}. \end{aligned} \tag{4}$$

Because this is a proper likelihood function of observed data in the parameters $\theta = (\beta_0^*, \beta_1^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$, it follows that

$$0 = E\left\{\frac{\partial}{\partial\theta}L(\beta_0^*, \beta_1, \xi; D, W)|\delta = 1\right\}. \tag{5}$$

However, because $\delta$ is a function of $D$ alone and is independent of $W$ given $D$,

$$\begin{aligned} f_{D,X,W|\delta=1}(d, x, w|\delta = 1) &= f_{W|D,\delta=1}(w|d, \delta = 1)\mathrm{pr}(D = d|\delta = 1) \\ &= f_{W|D}(w|d)\mathrm{pr}(D = d|\delta = 1) \\ &= p_d f_{W|D}(w|d), \end{aligned} \tag{6}$$

the last step following from some detailed algebra. Thus, (5) means that

$$0 = \int \sum_{d=0}^{1} f_{W|D}(w|d)\frac{\partial}{\partial\theta}L(\beta_0^*, \beta_1, \xi; d, w)dw, \tag{7}$$

which is the same as (5), and hence (7) shows (3).

## A.3 Inference

For inference, two steps are required. The first is the basic information equality. That is, we need to show that

$$\begin{aligned} -I_{\mathrm{F}} &= \sum_{d=0}^{1} p_d E\left\{\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}L(\beta_0^*, \beta_1, \xi; d, W)|d\right\} \\ &= -\sum_{d=0}^{1} p_d E\left\{\frac{\partial}{\partial\theta}L(\beta_0^*, \beta_1, \xi; d, W)\frac{\partial}{\partial\theta^{\mathrm{T}}}L(\beta_0^*, \beta_1, \xi; d, W)|d\right\} \end{aligned} \tag{8}$$

where $I_{\mathrm{F}}$ is the Fisher information matrix. However, in the alternative sampling framework, because (4) is a proper likelihood function, we must have that

$$\begin{aligned} &E\left\{\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}L(\beta_0^*, \beta_1, \xi; D, W)|\delta = 1\right\} \\ &\quad -E\left\{\frac{\partial}{\partial\theta}L(\beta_0^*, \beta_1, \xi; D, W)\frac{\partial}{\partial\theta^{\mathrm{T}}}L(\beta_0^*, \beta_1, \xi; D, W)|\delta = 1\right\}. \end{aligned}$$

Now apply (6) to show that equation (8) holds in the actual case-control sample.

In practice, asymptotic inference would be based on assuming the asympotic distribution of $n^{1/2}(\widehat{\theta} - \theta)$ given by

$$n^{1/2}(\widehat{\theta} - \theta) \overset{d}{\sim} \text{Normal}(0, I_{\text{F}}^{-1}). \tag{9}$$

Because of (8), Carroll *et al.* (1995, Section 4.1) show that for a positive semidefinite matrix $\Lambda$, the asymptotic covariance matrix of $\widehat{\theta}$ is $I_{\text{F}}^{-1}(I_{\text{F}} - \Lambda)I_{\text{F}}^{-1}$, and thus at worst prospectively derived information-based standard errors are asymptotically conservative.

**Table 1:** Bias, standard error and coverages of confidence intervals for $\beta_1$ based on 500 replications, for $\sigma_U = 0.3$. The true value for $\beta_1$ is 0.8.

| Mix-Lognormals | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.142 (0.395) | 0.012 (0.493) | 0.020 (0.450) | 0.010 (0.437) |
| s.e. | 0.369 (0.056) | 0.499 (0.085) | 0.438 (0.062) | 0.436 (0.060) |
| 90% | 0.854 | 0.928 | 0.914 | 0.910 |
| 95% | 0.938 | 0.960 | 0.964 | 0.960 |

| $\chi^2_1$ | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.086 (0.097) | -0.058 (0.115) | 0.005 (0.102) | 0.011 (0.103) |
| s.e. | 0.087 (0.007) | 0.097 (0.013) | 0.102 (0.008) | 0.103 (0.008) |
| 90% | 0.684 | 0.770 | 0.906 | 0.906 |
| 95% | 0.752 | 0.866 | 0.956 | 0.954 |

| Weibull | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.209 (0.197) | -0.108 (0.232) | 0.013 (0.244) | 0.011 (0.321) |
| s.e. | 0.192 (0.011) | 0.233 (0.023) | 0.253 (0.014) | 0.251 (0.015) |
| 90% | 0.682 | 0.864 | 0.907 | 0.904 |
| 95% | 0.790 | 0.916 | 0.958 | 0.947 |

**Table 2:** Bias, standard error and coverages of confidence intervals for $\beta_1$ based on 500 replications, for $\sigma_U = 0.5$. The true value for $\beta_1$ is 0.8.

| Mix-Lognormals | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.277 (0.327) | 0.153 (0.592) | 0.003 (0.490) | 0.012 (0.539) |
| s.e. | 0.312 (0.048) | 0.592 (0.120) | 0.485 (0.072) | 0.495 (0.087) |
| 90% | 0.718 | 0.917 | 0.918 | 0.894 |
| 95% | 0.826 | 0.949 | 0.964 | 0.950 |

| $\chi^2_1$ | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.244 (0.089) | -0.149 (0.118) | 0.008 (0.114) | 0.016 (0.117) |
| s.e. | 0.075 (0.007) | 0.100 (0.015) | 0.111 (0.010) | 0.115 (0.011) |
| 90% | 0.118 | 0.522 | 0.902 | 0.894 |
| 95% | 0.162 | 0.612 | 0.938 | 0.942 |

| Weibull | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.415 (0.148) | -0.166 (0.260) | 0.035 (0.306) | 0.008 (0.316) |
| s.e. | 0.149 (0.012) | 0.243 (0.036) | 0.288 (0.019) | 0.276 (0.024) |
| 90% | 0.124 | 0.778 | 0.884 | 0.865 |
| 95% | 0.218 | 0.882 | 0.944 | 0.931 |

**Table 3:** Bias, standard error and coverages of confidence intervals for $\beta_1$ based on 500 replications, for $\sigma_U = 0.75$. The true value for $\beta_1$ is 0.8.

| Mix-Lognormals | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.518 (0.234) | 0.200 (0.793) | -0.012 (0.590) | 0.099 (0.721) |
| s.e. | 0.224 (0.048) | 0.748 (0.176) | 0.563 (0.105) | 0.610 (0.169) |
| 90% | 0.254 | 0.890 | 0.928 | 0.896 |
| 95% | 0.346 | 0.946 | 0.972 | 0.937 |

| $\chi^2_1$ | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.440 (0.076) | -0.277 (0.131) | 0.013 (0.123) | 0.015 (0.137) |
| s.e. | 0.057 (0.007) | 0.101 (0.021) | 0.123 (0.012) | 0.129 (0.017) |
| 90% | 0.002 | 0.232 | 0.898 | 0.884 |
| 95% | 0.002 | 0.282 | 0.932 | 0.927 |

| Weibull | NAIVE | RC | LIK | SN |
|---|---|---|---|---|
| Bias | -0.606 (0.105) | -0.193 (0.308) | -0.017 (0.316) | -0.072 (0.328) |
| s.e. | 0.097 (0.012) | 0.277 (0.057) | 0.331 (0.028) | 0.308 (0.034) |
| 90% | 0.002 | 0.748 | 0.912 | 0.870 |
| 95% | 0.002 | 0.826 | 0.970 | 0.920 |

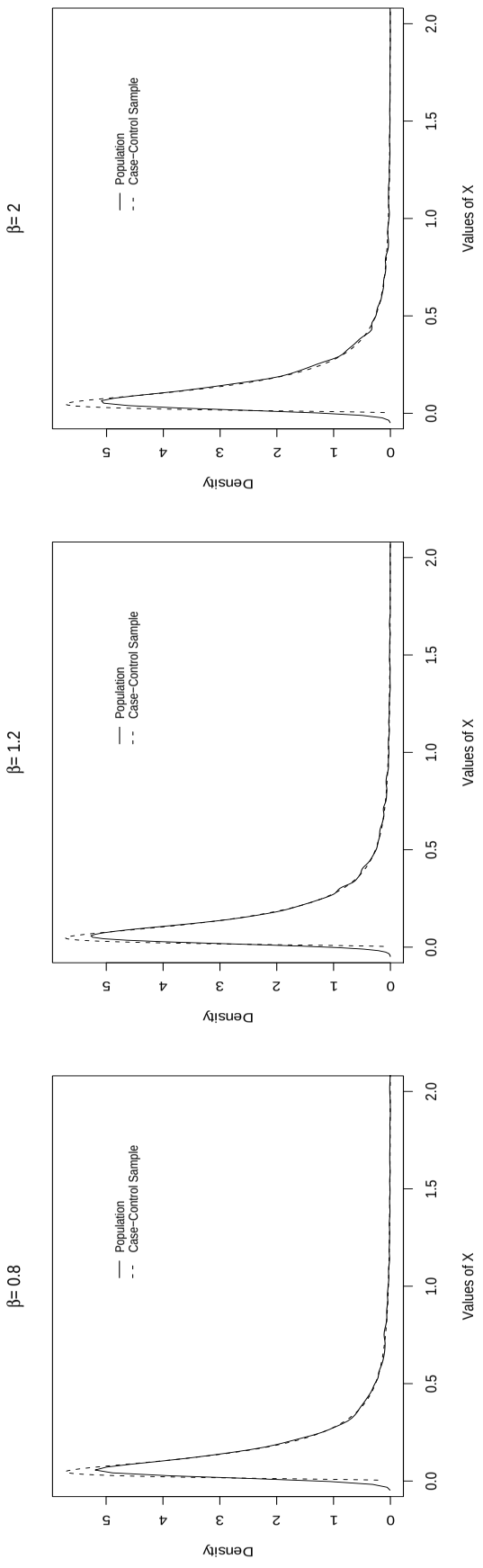


**Figure 1:** Density function of $X$ obtained by simulating from a mixture of Lognormal$(-2.3, 0.9)$ and Lognormal$(-1.5, 0.9)$, with mixing weights 0.8 and 0.2 (solid line). Density function of $X$ in the case-control sample extracted

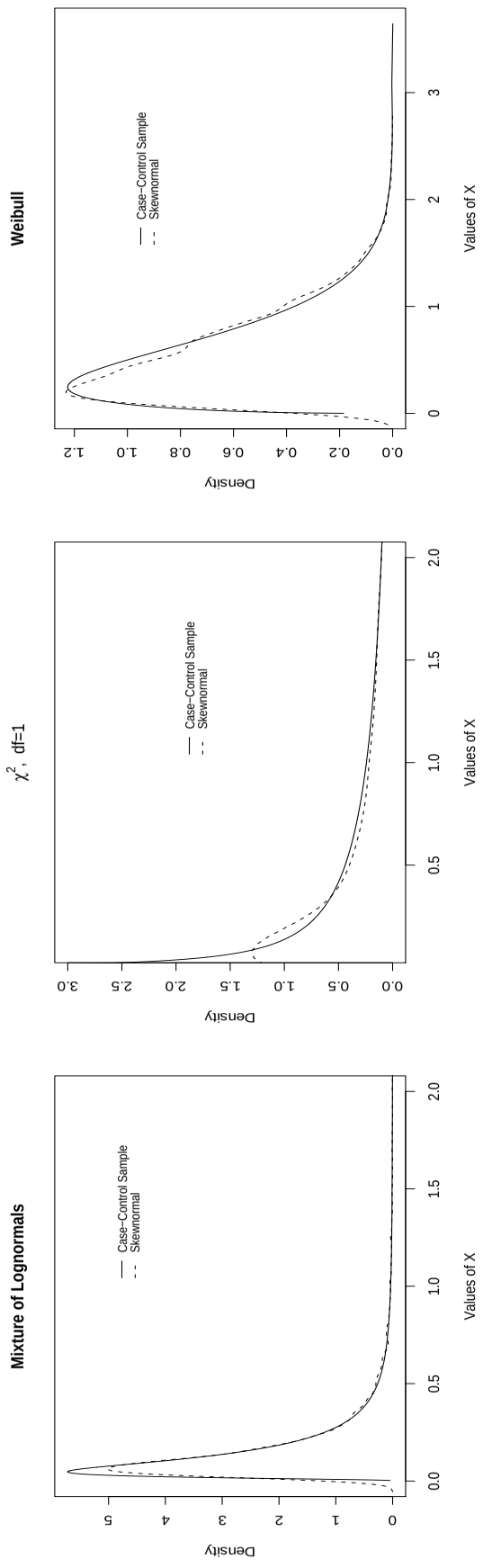


**Figure 2:** Density function of $X$ in the case-control sample (solid line) based on 10000 simulated values and density function of the best skewnormal fitted on the observations from $X$ (dashed line).

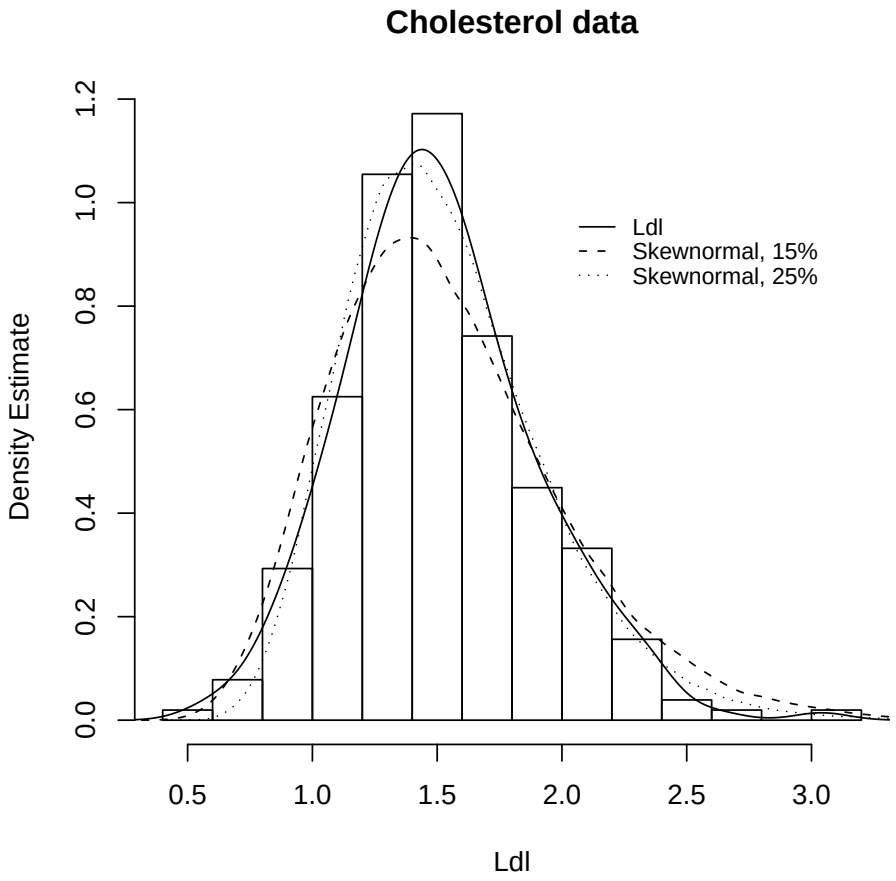


**Figure 3:** Density function of Ldl (solid line) and of the skewnormal fitted using a 15% or a 25% validation data sample (dashed lines): Cholesterol data.

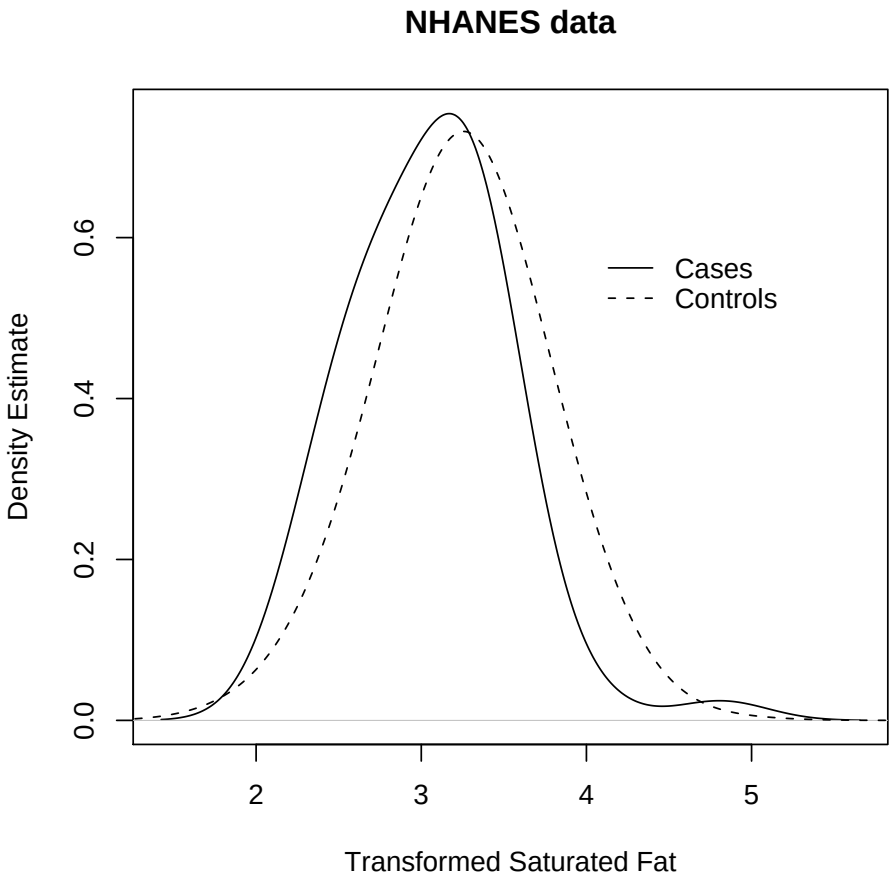


**Figure 4:** Density estimates of logarithm of the saturated fat for cases (solid line) and controls (dashed line): NHANES data.

## Acknowledgements

This research was partially supported by the Italian Ministry for Education, University and Research.

## Working Paper Series

## Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

**Department of Statistical Sciences**
*University of Padua*
*Italy*