**International Journal of Scientific Research in Mathematical and Statistical Sciences** **Research Paper**
Volume-7, Issue-2, pp.17-33, April (2020) **E-ISSN:** 2348-4519

# Covariate Measurement Error in a Non-Linear Model for Longitudinal Data

**Rosa C. S. Oliveira**[1,2]

[1]Matemática, Faculdade de Ciências, Porto, Portugal
[2]CINTESIS, Faculdade de Medicina, Porto, Portugal

**Corresponding email: rosita21@gmail.com*
**Available online at: www.isroset.org**

Received: 06/Mar/2019, Accepted: 14/Apr/2020, Online: 30/Apr/2020

***Abstract*-**In this paper, we study nonlinear generalized regression models for the analysis of longitudinal data. We indicate a two-stage approach that, in the first stage models a linear model to the longitudinal data to estimate the random effect that characterizes the trajectory for each individual and in the second stage uses the estimated intercept and slope as predictors for the covariates used in a probit model. We show that the straight use of the estimates in the probit model produces biased estimates of their outcome. We show how to adapt a Regression Calibration and a Pseudo-Likelihood approach to this context and compare these approaches with a naive analysis where the estimation error is ignored. Regression Calibration and Pseudo-Likelihood methods seem to be the best choice as they perform well, even for small sample size, if data is not very noisy. It is fair to say, mainly, that the Regression Calibration and Pseudo-Likelihood method perform equally well, nonetheless, Regression Calibration seems to perform a bit better. Nevertheless, naive approach has smaller Mean Squared Error, has a great absolute bias. Regression Calibration seems to be the best. Our study indicates that correcting for measurement errors instead of falsely assume that errors are not present will produce less bias than ignoring exposure measurement error in the analysis.

***Keywords*-**measurement error, errors-in-variables, general nonlinear model, longitudinal data

## 1 INTRODUCTION

The study of risk factors of death is central to health metrics. Biomarkers are objective physical or biologic measures of health conditions. The availability of information about biomarkers has opened the possibility to include biomarkers in population health metrics and the study of biomarkers as risk factors of death has a long tradition in health sciences, especially of cardiovascular death or other severe outcomes such as heart attacks. Most evaluations of the diagnostic or prognostic ability of clinical biomarkers focus on a single measurement and its relation with an outcome of interest. Despite, there are settings where the longitudinal profile of a biomarker may be more informative than an isolated value of cross-sectional observations. One reason sticks from the chances provided by longitudinal data to improve prognostic models for succeeding outcomes for each individual. E.g., Póvoa[1] (2011) showed that the C-reactive protein (CRP) trend over several days for patients with sepsis - a whole-body inflammatory response to an infection that can cause organs to fail and in severe cases, death - was a better prognostic marker of sepsis resolution than the individual measurement of the CRP. However, the methodological approach used by the authors ignores intrinsic statistical problems regarding the joint analysis of the longitudinal measurements and their association with the binary outcome of interest, in that case, the patient's discharge status from the Intensive Care Unit (death or alive). The motivation problem of this article comes from the SACiUCI study[2], a prospective observational study aiming to identify demographic and clinical prognostic factors of patients admitted to intensive care with sepsis. In particular, we will revisit the evaluation of the longitudinal profile of CRP as a marker of sepsis resolution. Given that CRP levels increase dramatically in response to an inflammatory process, it is expected that the progressive reduction of these levels would indicate that the systemic inflammatory response that underlies sepsis is becoming under control and thus, associated with a good prognosis for the patient.

**Abbreviations:** CAS, Community-Acquired Sepsis; CRP, C-Reactive Protein; GLM, Generalized Linear Model; GLS, Generalized Least Squares; ICU, Intensive Care Unit; LL, Lower Limit; ME, Measurement Error; MSE, Mean Squared Error; OLS, Ordinary Least Squares; OR, Odds Ratio; PL, Pseudo-Likelihood; RC, Regression Calibration SAPS, Simplified Acute Physiology Score; SE, Standard Error; SOFA, Sequential Organ Failure Assessment; UL, Upper Limit; WBC, White Blood Cell count; 95% CI, 95% Confidence Interval

We focus on a two-stage approach where first we fit a linear model to the longitudinal data to estimate the random effect that characterises the trajectory for each individual and in a second stage, we consider the individual-specific intercept and slope of the linear regression as covariates in a model for the binary outcome. Prediction of the outcome variable, e.g. death, is based on the logistic or probit regression model (e.g.) in the second stage, and as we do not know the true values for the intercepts and slopes but only estimates of these quantities obtained from the linear model, simply substituting the true values by the estimates as covariates in the model for the binary outcomes. will bias the results of the second model. By recognizing that the use of the estimates as covariates instead of the true values is identical to a situation where the covariates are measured with error, we can use the same principles of Measurement Error (ME) theory to correct or at least compensate for the bias of the substitution. Tosteson[2] used a likelihood-based method to fit mixed models in which the covariates contain measurement error. Carroll[3](2006), on the other hand, used Regression Calibration (RC), a simple method applicable to any regression model and most valuable for estimating parameters in Generalized linear models (GLM) with covariate measurement error. However, there is an important difference from the classical error measurement problem. It is usually required that the true values are measured in a subset of the data or, at least, that we have available repeated measurements of the variable measured with error, for identifiability of the error components in the model. Nevertheless, it is well known that naive implementation such as conducting longitudinal data analysis covering measurement error conducts to biased inferences, and considering that, several methods have been proposed to reduce this bias. Even so, it has not been sufficiently investigated a correction of bias on GLM (a logistic regression model, more precisely) to correct for bias in a linear mixed model adjusted for other covariates. In our setting, we do not have the true intercepts or slopes for any of the participants nor multiple measurements of the intercepts and slopes. We will show how to adapt a Regression Calibration (RC) and a Pseudo-Likelihood (PL) approaches to this context and contrast these approaches with a naive analysis where the estimation error is ignored.

The paper is organized as follows. Section 1 provides a brief introduction to the problem. Section 2 presents an overview of the generalized linear models with *measurement errors*. Section 3 proposes a correction for the MLE estimator. For that, we present the response, covariate, and error models for the three estimators: Ordinary Least Squares (OLS), Regression Calibration and Pseudo-Likelihood methods. Section 4 includes Monte Carlo simulations that demonstrate the usefulness of our estimator and its goodness of fit on sample performance. The proposed approach is applied to the SACiUCI data in 5. Concluding remarks are given in Section 6. In **Supplementary Material** all technical details are given about measurement error and CRP.

## 2 | GENERALIZED LINEAR MODELS AND MEASUREMENT ERROR

### 2.1 | Measurement error models and methods - previous results

Various statistical procedures have been developed for statistical inference in measurement error models. Let $W$ be a $k \times 1$ vector of surrogate variable, measured with error, and $X$ a corresponding $k \times 1$ vector of a latent variable. $X$ is assumed to have a multivariate normal distribution.

Let $D$ be a dichotomous outcome variable and assume that a logistic regression model relates $D$ to $X$ through a $k \times 1$ vector of regression coefficients $\beta$.

In the classical error model framework, $W = X + U$, where $U$ has mean zero and variance equal to $\sigma_U^2$ and is independent of $X$. The classical measurement error model is an unbiased and additive error model, such that $E[W|X] = X$.

We will consider the probit measurement error model that can be written as

$$E(D|X) = \text{pr}(D = 1|X) = \Phi^{-1}(\alpha + \beta X),$$

where $\Phi(.)$ indicates the standard normal distribution.

A nonlinear function of a mismeasured covariate introduces an error term that is not additively detachable from the true value of the regressor, making standard linear instrumental variable estimators inconsistent. As a result, in general nonlinear specifications there exists, so far, no estimator that exhibits all the desirable properties of the estimators to correct for measurement error in linear specifications, such as asymptotic normality, absence of distributional assumptions and weak regularity conditions, among others.

Clear measurement modeling is crucial because when measurement error is neglected biased estimates are, frequently, achieved for the regression parameters. For instance, in the ordinary probit regression scenario, it can lead to biased estimates of

odds (Rosner, 1990)[4]. Carroll et al.[3] showed that joint modeling of the response and measurement process enables "disattenuated" the odds ratio for the true covariate estimation. Another well-known fact is that measurement modeling assists prediction of the true covariate, or exposure factor, for a subject.

Estimates of the measurement error bias in the restrictive case of nonlinear models where the measurement error variance is small were studied by Carroll et al.[3]. Though, to this point, there exists no overall framework to convert an estimator, that assumes that all regressors are accurately observed, into one, robust to the presence of *measurement errors* in the regressors. The easiest adjustment is given by a specified estimate of the measurement error variance, e.g. gated by replicate measures of the unknown covariate. Carroll et al.[3] points out that certain types of covariates can be subject to float such that later measurements from latent variables tend to reveal a modification in the mean across the observations, but there are no results on the general nonlinear model problem.

The problem is that without some information about the measurement error parameters one cannot estimate all the parameters in the model and the covariate coefficient - $\beta_1$, in particular, is typically non-identifiable. To correct for measurement error we need to assume something about the parameters or have additional data, which allows us to estimate them. There has been a lot of work on correcting for measurement error assuming certain variances or functions of them are known.

### 2.1.1 | Regression Calibration

One of the more popular methods for adjusting for measurement error in Generalized Linear Models (GLM) is Regression Calibration, proposed by Carroll et al.[3]. The justification of the RC is based on a first-order Taylor approximation of $E(D|W)$ and the assumption of non-differential error, i.e., $E(D|X,W) = E(D|X)$. The main idea of RC is to predict the true value of the covariate, $X$, from the observed data, $W$ and, then, use this predicted value in the regression analysis. Consider model (1), given that we do not have $X$ but $W$, we substitute $X$ in the model by $E(X|W)$.

$$\text{pr}(D_i = 1|\widehat{X}_i) = E(D_i|\widehat{X}_i) = E(E(D_i|\widehat{X}_i, X_i)|\widehat{X}_i) = E(E(D_i|X_i)|\widehat{X}_i),$$

This method assumes the use of the best approximation of the unobserved variable, given the known existing information. This approximation on which RC is based is exact for linear regression but is only "almost exact" for probit regression[3]. Nonetheless, RC algorithm still works well and gives good estimates, research was needed so better estimates were obtained.

In RC $E(X|W)$ is used as a covariate instead of $W$, where $E(X|W)$ is the predicted value of $X$ based on the surrogate measurement $W$. One notice that in classical measurement error framework $E(X|W) \neq W$.

Once RC requires that the true data is defined from replicated or validation data it is appropriate in the following two situations. First, it is appropriate when a gold standard is available in the main study/internal or external validation study and a linear measurement error model with constant variance applies. The main issue in the validation study is that this kind of study doesn't have data on the outcome variable for the primary regression models but focuses on the comparison of the surrogate measurement and the gold standard. However, gold standards, if exist, are usually only available for a small subset of the data, called a validation data set, so the vast majority of the data will not have a gold standard and when no gold standard is available, then replication data can be used to estimate the measurement error variance, sometimes only on a subset of the data. This requires the assumption that W = X+U. Once you can estimate that variance, you can do RC, SIMEX, etc. Second, it is appropriate when the main study has an internal validation study design with replicate measurements, and the value of the variance of measurement error can be assumed.

Suppose that $Y = \alpha + \beta X + \epsilon$, where $\epsilon$ is independent of $X$ and take expectations conditional on $W$. Then

$$E(Y|W) = \alpha + \beta E(X|W).$$

RC consists in three steps wherein we only need to develop and fit the calibration model for the regression of the unknown covariates $X$ on the observed covariate $W$. To attain a model for $X$ and $W$ is assumed using replicates or a validation sample and $\widehat{X}$ is estimated with the regression of $X$ on $W$, let be $\widehat{E}(X|W)$. A calibration function for estimating $X$ is then obtained. In a second stage, $X$ is replaced by its estimate, $\widehat{X}$, from the calibration model in a standard analysis as a covariate. Finally, standard errors and confidence intervals are adjusted to account for the estimation of the unknown parameter taking into consideration that $X$ is estimated in the calibration model and is not the true covariate value. This is accomplished, usually, performing bootstrapping or sandwich estimates.

The method generalises to the setting of multiple explanatory variables measured with error. Notwithstanding if there is more than one surrogate variable the situation is much more complex as the covariance matrix must be calculated and specified before.

Standard errors and confidence intervals should allow for the first stage of estimation, e.g. by using bootstrap methods. We should notice that RC also gives approximately unbiased estimates for probit regression, Cox regression, ... As replicated or validation data is generally inaccessible the difficulty remains, moreover, besides, the efficiency of the method depends on how well the calibration function is estimated, and it was proven to be inadequate in highly nonlinear models.

### 2.1.2 | Likelihood methods

The likelihood function assumes a demanding role in both Frequentist and Bayesian methods. Nonetheless in either framework, with the current blast in the span of information sets and the increment in multifaceted nature of the conditions among variables in numerous reasonable and real models, it is often unfeasible to build the full likelihood. In these circumstances, other frameworks can be considered, as PL or composite likelihood as a convenient framework.

The likelihood for an observed data point (Y, W) conditional on Z is

$f_{Y,W|Z} = \int f_{Y|Z,X,W} f_{W|Z,X} f_{X|Z} dx = \int f_{Y|Z,X} f_{W|Z,X} f_{X|Z} dx$ where the second equality follows from the assumption of non-differential measurement error.

The application of likelihood techniques requires the parametric specification of the distribution for the unobserved variable X, that is, the exposure model, together with the specification of the disease model and the measurement error model previously defined. As like in functional models, estimation of parameters in the disease model generally requires, for all intents and purposes, observations that allow estimation of parameters in the error model as replicate measurements, e.g.. likelihood methods require stronger distributional assumptions, but they can be applied to more general problems than the functional models, including those with discrete covariates subject to misclassification (Carroll et. al)[3]. Likelihood-based methods are used with structural models, i.e., assuming $X$ is random and requiring an exposure model for $X$, with the normal distribution as the default exposure model. On one hand, in nonlinear problems, likelihood-based confidence intervals are generally more reliable than the ones attained from normal approximations and they are also less computationally intensive than bootstrapping. On the other hand, likelihood methods are often computationally more demanding than the previous methods, whereas those require little more than the use of standard statistical packages. In the context of measurement error problems, the advantages of likelihood methods relative to functional methods have been studied in Shafer et al.[5] and Carroll et al.[6]. In this survey we won't go deep on this subject, redirecting interesting ones for the mentioned papers. Nonetheless, we should recall that the likelihood methods are advantageous only on correct specification of the likelihood, which is often a difficult task in measurement error, and so, finding the maximum likelihood is not always straightforward.

Concerning robustness and efficiency, likelihood methods are usually more robust then the simpler ones, however, they are also more difficult to understand. The same goes with efficiency. Nonetheless, a likelihood framework generally is more efficient than functional modeling, i.e produces smaller standard errors, it will also be computationally more difficult and sometimes the gains in the efficiency are very minor. As in measurement error context, the covariate is latent, to compute the likelihood function can be difficult and time-consuming, considering one has to integrate the possibly high dimensional latent variable and the scenario can become even worse when more variables are plugged in the model.

A critical decision in a likelihood analysis is the choice of the error model. After the likelihood model is selected (Classical, Berkson, ...) the likelihood function should be computed and maximized.

## 3 | MODEL SETUP AND METHODS

This section starts with the formulation of the error measurement. Then we show how to turn our problem into a measurement error problem. Subsequently, we present the covariate model encompassing distributional information of the latent variable and the error model relating the latent variable to the observable measurement (X). Then the response model linking the true to the dependent variable death/survival state (D) in a longitudinal setting is presented. Three methods of estimating the linear measurement model coefficients are discussed: OLS, Pseudo-Likelihood (PL) and RC.

### 3.1 | Data and Basic Model

In the simplest case of across-sectional bivariate regression where the independent variable measured is reported with additive error, the proportional attenuation bias is given asymptotically by $1 - \lambda$, where $\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$.

The quantity $\lambda$ is often referred to as the reliability of the data. The reliability of a variable in the general case is the slope coefficient from an OLS regression of the correctly measured variable on the mismeasured variable, i.e, the proportion of a change in the observed variable that translates into a change in the true, latent, variable. In this survey, we propose a reliability factor when errors are made on the measure/report of a variable in a dichotomous regression scenario.

Estimates which do not account for measurement error are typically biased. Correcting for this bias entails what is often referred to as a bias versus variance tradeoff. What this means is that in most problems, the very nature of correcting for bias is that the resulting corrected estimator will be more variable than the biased estimator. Of course, when an estimator is more variable, the confidence intervals associated with it are longer.

Because $\lambda < 1$ it is clear that while the correction-for-attenuation in $\beta_x$ reduces it's biased to zero there is an increase in variability relative to the variance of the biased estimator $\beta_x$. The variability is inflated even further if an estimate $\hat{\lambda}$ is used in place of $\lambda$. The price for reduced bias is increased variance, just like in Pseudo-likelihood in which the decrease of bias comes with a price on the variance.

Remember that Mean Squared Error (MSE) is the sum of the variance plus the square of the bias. This is an interesting criterion to use, because uncorrected estimators have more bias but smaller variance than corrected estimators, and the bias versus variance tradeoff is transparent.

We present our model within the framework of the counting process. Denote by $i = 1, ..., n$ individuals, each of whom has a binary disease status $D_i$. Also, we have further variables $Y_{ij}$ measured at times $(t_{i1}, ..., t_{ij})$, with $t_{i1} = 0$ denoting baseline.

We suppose that, unconditionally,

$$\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i})^{\mathrm{T}} = \text{Normal}\{\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}, \Sigma_\beta\}.$$

We link the binary outcome to the continuous measurements via a probit regression, so that

$$\text{pr}(D_i = 1|\boldsymbol{\beta}_i) = \Phi(\alpha_0 + \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\alpha}_1) = \Phi(\alpha_0 + \alpha_1\beta_{0i} + \alpha_2\beta_{1i}). \tag{1}$$

In other words, the binary outcome depends on the true baseline value $\beta_{0i}$ and the individual slope $\beta_{1i}$. While the response variable $D_i$ is observed, the covariates $\boldsymbol{\beta}_i$ are latent variables, in the sense that we do not observe the true $\boldsymbol{\beta}_i$, instead, we observe an estimates version of this $\boldsymbol{\beta}_i$. There are different estimates we can use and will use OLS.

We do not observe $\boldsymbol{\beta}_i$, but when $J > 2$, we can estimate it by $\widehat{\boldsymbol{\beta}}_i$, the ordinary least squares estimate of $\boldsymbol{\beta}_i$. Write $\mathcal{Y}_i = (Y_{i1}, ..., Y_{ij})^{\mathrm{T}}$, and

$$\mathrm{V}_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{ij} \end{bmatrix}.$$

Notice that $\mathcal{Y}_i = V_i\boldsymbol{\beta}_i + (\epsilon_{i1}, ..., \epsilon_{ij})^{\mathrm{T}}$. Hence, conditional upon the person,

$$\begin{aligned} \text{cov}(\mathcal{Y}_i|\text{ person } i) &= \sigma_\epsilon^2 I_J; \\ \widehat{\boldsymbol{\beta}}_i &= \boldsymbol{\beta}_i + (V_i^{\mathrm{T}}V_i)^{-1}V_i^{\mathrm{T}}(\epsilon_{i1}, ..., \epsilon_{ij})^{\mathrm{T}}. \end{aligned} \tag{2}$$

To understand its properties, recall that the ordinary least squares estimate is

$$\widehat{\boldsymbol{\beta}}_i = (V_i^{\mathrm{T}}V_i)^{-1}V_i^{\mathrm{T}}\mathcal{Y}_i.$$

Further, we see that from (2), unconditionally,

$$\begin{aligned} \text{cov}(\boldsymbol{\beta}_i) &= \Sigma_\beta; \\ \text{cov}(\widehat{\boldsymbol{\beta}}_i) &= \Sigma_\beta + \sigma_\epsilon^2(V_i^{\mathrm{T}}V_i)^{-1}; \\ \text{cov}(\widehat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_i) &= \Sigma_\beta. \end{aligned}$$

## 3.2 | Measurement Error Formulation

We now show how to turn our problem into a measurement error problem. We have the model $\text{pr}(D_i = 1|\boldsymbol{\beta}_i) = \Phi(\alpha_0 + \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\alpha})$ ; where $\widehat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + U_i$ given $\boldsymbol{\beta}_i = \text{Normal}\{\boldsymbol{\beta}, \Sigma_\beta\}$ and $U_i = \text{Normal}\{0, \Sigma_{ui} = \sigma_\epsilon^2(V_i^{\mathrm{T}}V_i)^{-1}\}$.

We assume that the measurement error is nondifferential. Besides, unconditionally,

$$E(\boldsymbol{\beta}_i|\widehat{\boldsymbol{\beta}}_i) = \mathcal{G}(\widehat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui}) = \boldsymbol{\beta} + \Sigma_\beta(\Sigma_\beta + \Sigma_{ui})^{-1}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}); \tag{3}$$

$$\text{cov}(\boldsymbol{\beta}_i|\widehat{\boldsymbol{\beta}}_i) = \mathcal{V}(\Sigma_\beta, \Sigma_{ui}) = \Sigma_\beta\{I - (\Sigma_\beta + \Sigma_{ui})^{-1}\Sigma_\beta\}. \tag{4}$$

Finally, since $\int \Phi(a + bx)\phi(x)dx = \Phi\{a/(1 + b^2)^{1/2}\}$, it is easily seen that

$$\text{pr}(D_i|W_i) = \Phi\left[\frac{\alpha_0 + \mathcal{G}^{\text{T}}(\widehat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui})\boldsymbol{\alpha}}{\{1 + \boldsymbol{\alpha}^{\text{T}}\mathcal{V}(\boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui})\boldsymbol{\alpha}\}^{1/2}}\right]. \tag{5}$$

### 3.3 | Estimation of $(\sigma_\epsilon^2, \boldsymbol{\beta}, \Sigma_\beta)$

To implement a measurement error analysis, we need to estimate $\sigma_\epsilon^2$, $\boldsymbol{\beta}$ and $\Sigma_\beta$.

Define

$$\widehat{\sigma}_{\epsilon i}^2 = (J - 2)^{-1}\sum_{j=1}^{J}(Y_{ij} - V_i\widehat{\boldsymbol{\beta}}_i)^2,$$

the Mean Squared Error of the regression for person $i$. Each $\widehat{\sigma}_{\epsilon i}^2$ is unbiased for $\sigma_\epsilon^2$, so an unbiased estimate of $\sigma_\epsilon^2$ is just

$$\widehat{\sigma}_\epsilon^2 = n^{-1}\sum_{i=1}^{n}\widehat{\sigma}_{\epsilon i}^2.$$

The unconditional covariance matrix of $\widehat{\boldsymbol{\beta}}_i$ is $\Sigma_\beta + \sigma_\epsilon^2(V_i^{\text{T}}V_i)^{-1}$. Let $\widehat{\Omega}$ be the sample covariance of $(\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_n)$. Then since $\widehat{\Omega}$ estimates $\Sigma_\beta + \sigma_\epsilon^2 n^{-1}\sum_{i=1}^{n}(V_i^{\text{T}}V_i)^{-1}$, a consistent estimate of $\Sigma_\beta$ is

$\widehat{\Sigma}_\beta = \widehat{\Omega} - \widehat{\sigma}_\epsilon^2 n^{-1}\sum_{i=1}^{n}(V_i^{\text{T}}V_i)^{-1}$.

Finally, of course, an unbiased estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_i.$$

### 3.4 | Pseudo-Likelihood Estimation of $(\alpha_0, \boldsymbol{\alpha})$

For longitudinal binary data with non-repeated and non-ignorable missing outcomes over time, a full likelihood approach is very complicated algebraically and MLE can be computationally infeasible, so, as in this research problem we consider that the outcome is not, always, observed at all times we use PLs to estimate the covariate effects on the marginal probabilities of the outcomes. This method in the probit case is difficult to compute, perform poorly in moderate sample sizes and is non-robust to misspecification on the conditional density $f_{X|W}$. The PL requires specification of the distribution for the data at all pairs of times on the same subject, but makes no assumptions about the joint distribution of the data at three or more times on the same subject, so the method can be considered semi-parametric. To formulate a full likelihood for non-ignorable non-monotone missing outcomes over time, one must specify a model for the repeated binary outcomes of interest, and a model for the missingness mechanism. To estimate the parameters, a full likelihood approach has many nuisance parameters, is complicated algebraically, and maximum likelihood estimation can be computationally prohibitive, especially when the number of times is large. We propose a 'Pseudo-Likelihood' to estimate the covariate effects on the marginal probabilities of the outcomes, besides to the association parameters and missingness parameters. The PL requires only partial specification of the distribution of observations and missingness indicators at pairs of times, and can be much less computationally prohibitive than maximum likelihood.

#### 3.4.1 | General Case

Define

$$\mathcal{K}(\alpha_0, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui}) = \frac{\alpha_0 + \mathcal{G}^T\left(\widehat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui}\right)\boldsymbol{\alpha}}{\{1 + \boldsymbol{\alpha}^T\mathcal{V}\left(\boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui}\right)\boldsymbol{\alpha}\}^{1/2}} = \mathcal{K}_i(\cdot)$$

Given the estimates formed in Section 3.3, we will estimate $(\alpha_0, \boldsymbol{\alpha})$ by maximizing the pseudo-loglikelihood

$$\sum_{i=1}^{n}Y_i\text{log}[\Phi\{\mathcal{K}(\alpha_0, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\beta}}, \widehat{\Sigma}_\beta, \widehat{\Sigma}_{u_i})\}] + \sum_{i=1}^{n}(1 - Y_i)\text{log}[1 - \Phi\{\mathcal{K}(\alpha_0, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\beta}}, \widehat{\Sigma}_\beta, \widehat{\Sigma}_{u_i})\})\}]$$

in $(\alpha_0, \boldsymbol{\alpha})$. This can be performed either by an optimizer or by Fisher's method of scoring, illustrated as follows.

The Fisher Scoring algorithm is an iterative maximum likelihood procedure and can be implemented fitting iteratively reweighted least square. Is a similar method to Newton-Raphson but modified to overcome the convergence problem of the Newton-Raphson method. Both algorithms give the same parameter estimates; however, the estimated covariance matrix of the parameter estimators can differ slightly. Because Fisher scoring is based on the expected information matrix while the Newton-Raphson method is based on the observed information matrix. In the case of a binary logit/probit model, the observed and expected information matrices are identical, resulting in identical estimated covariance matrices for both algorithms.

Define $Q_i$

$$\begin{pmatrix} \partial \mathcal{K}_i(\cdot)/\partial \alpha_0 \\ \partial \mathcal{K}_i(\cdot)/\partial \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} \{1 + \boldsymbol{\alpha}^T \mathcal{V}\left(\boldsymbol{\beta}, \Sigma_\beta, \Sigma_{ui}\right) \boldsymbol{\alpha}\}^{-1/2} \\ \mathcal{G}_i(\cdot)/\{1 + \boldsymbol{\alpha}^T \mathcal{V}_i \boldsymbol{\alpha}\}^{1/2} - \{\alpha_0 + \mathcal{G}_i^{\mathrm{T}}(\cdot)\boldsymbol{\alpha}\}\{\boldsymbol{\alpha}\mathcal{V}_i\}^T/\{1 + \boldsymbol{\alpha}^T \mathcal{V}_i \boldsymbol{\alpha}\}^{3/2} \end{pmatrix}$$

It is readily seen that the Fisher Score is

$$\mathcal{S}_i(\cdot) = \sum_{i=1}^{n} \frac{\varphi\{\mathcal{K}_i(\cdot)\}Q_i}{\Phi\{\mathcal{K}_i(\cdot)\}\left[1 - \Phi\{\mathcal{K}_i(\cdot)\}\right]} \times \left[D_i - \Phi\{\mathcal{K}_i(\cdot)\}\right],$$

while the Fisher Hessian will be:

$$\mathcal{H}_i(\cdot) = -\sum_{i=1}^{n} \frac{\varphi^2\{\mathcal{K}_i(\cdot)\}Q_i Q_i^T}{\Phi\{\mathcal{K}_i(\cdot)\}\left[1 - \Phi\{\mathcal{K}_i(\cdot)\}\right]}.$$

The Fisher scoring update is

$$(\alpha_{0\mathrm{new}}, \boldsymbol{\alpha}_{\mathrm{new}}^{\mathrm{T}})^{\mathrm{T}} = (\alpha_{0\mathrm{old}}, \boldsymbol{\alpha}_{\mathrm{old}}^{\mathrm{T}})^{\mathrm{T}} - \{\mathcal{H}(\cdot)\}^{-1}\mathcal{S}(\cdot),$$

where $\mathcal{H}(\cdot)$ and $\mathcal{S}(\cdot)$ are evaluated at $(\alpha_{0\mathrm{old}}, \boldsymbol{\alpha}_{\mathrm{old}}^{\mathrm{T}})^{\mathrm{T}}$. The covariance matrix of $(\widehat{\alpha}_0, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}})^{\mathrm{T}}$ can be estimated by the bootstrap, thus taking into account the estimation of the parameters as described in Section 3.3. It is possible to use estimating equation theory as well to derived estimates of the asymptotic covariance matrix.

Finally, the measurement error model has to be specified. It relates the CRP measurements to their evolution and baseline value, i.e., the surrogate covariate $W_{it}$ to the latent variable $X_{it}$ . Lastly, we have to stipulate the model that relates this CRP evolution to the prediction of disease.

## 4 | SIMULATION STUDY

In this section, we evaluate the finite sample behavior of the proposed estimators and compare them with the naive maximum likelihood estimates that ignore prediction error.

Parameters similar to our real data example were used to simulate disease status for the same number of units and repeated measurements of exposure as in the study data. We carried out Monte Carlo replications in each simulation study and report the empirical bias, the mean square error (MSE) and the coverage, both for model framework and bootstrap. Coverage is computed as the relative frequency of datasets of which the 95% confidence interval includes the true value of $\boldsymbol{\alpha}$ and is computed as $\widehat{\boldsymbol{\beta}} \pm 1.96$ times the estimated standard error of $\widehat{\boldsymbol{\beta}}$. Here the estimated standard error of $\widehat{\boldsymbol{\beta}}$ is computed by bootstrapping (200 samples) for regression calibration and is taken from the estimated Hessian matrix the naive estimator. Due to the computational intensity of the parametric bootstrap, we restrict to 500 Monte Carlo runs for each of these results. We will concentrate on the estimation of $\boldsymbol{\alpha}$.

The MSE is calculated relative to the mean square error of the estimate, which is the estimate subsequent from a logistic regression in the theoretical situation where the true value X is known. Taking into consideration the unstable estimation in the 250 observation the smaller 2.5% and the larger 2.5% are trimmed, in that case. We base our estimates of MSE, bias, and coverage on the remaining data in these situations.

In each simulated dataset, we generated disease, $D_{ij}$, within each subject by the hazard where the $X_{ij}$ was generated from a standard normal distribution. Data were generated from a normal distribution $Y_i j = \beta_{0_i} + \beta_{1_i} \times t_{ij} + \epsilon_{ij}$. So that for each subject $i$, $Y$ denotes the continuous outcome (CRP), $\beta_0$ denote baseline CRP value, $\beta_1$ each subject CRP slope rate taking into account the five CRP values at measurements time $j = 0, 1, 2, 3, 4$. The error term was independently and identically distributed. We assume that the distribution of surrogate exposure given true exposure is the same for both diseased and nondiseased subjects, i.e., Pr(W|X,D = 1) = Pr(W|X,D = 0).

We consider the sample sizes $N = 250$, and $N = 1000$ and attenuation factor $\lambda = \{0.7; 0.75; 0.8; 0.85; 0.9; 0.95; 0.98\}$. We took $\boldsymbol{\alpha}_1 = -2$ and $\boldsymbol{\alpha}_2 = -0.5$. The choice of $\boldsymbol{\alpha}$ represents cases where the expected $D = 1$ is 78.24%.

The method proposed here started with a multivariate linear model with varying intercept and slope groups effect obtaining estimates of the parameters in linear mixed-effect models and then a probit regression model was fitted to find covariate (W) values related to a binary response (D), obtaining estimates of the uncorrected probit regression coefficients. The competing coefficients estimates were assessed and our correction parameter was applied using Fisher Scoring method. Our simulations were based on the idea of calibrating our method in situations ideal for parametric methods, i.e linear additive normal measurement error. Finally, the coefficients were compared.

All computations are done in R 3.2.0 GUI 1.33 Mavericks.

## 4.1 | Simulation Results

Figures 1 and 2,show the Bias, Mean Squared Error and Coverage Average 95% Confidence Interval estimates for the intercept and slope regression parameters using simulated data with three distinct methods for two separate sample sizes (n=250 and n=1000) and seven $\lambda$ levels. The empirical standard errors were gotten by computing the standard deviation of the MLEs for the regression parameters for the set of replications (data not shown). These values were identical to the average of the standard errors (model- based standard errors) obtained in each replication.

Our results confirm that coefficients estimates were attenuated in the presence of confounders (measurement error), i.e., when confounders are present in the model, the exposure to disease associations will not be properly adjusted and, in that sense, variables can have less influence when models trying to predict the outcome are taken.

It should be noticed that concerning naive estimator, OLS overestimate $\alpha_1$ and underestimate $\alpha_2$. Pseudo-Likelihood underestimates $\alpha_1$ and behaves just like the naive estimator for $\alpha_1$ except for big sample sizes for higher $\lambda$ values. Regression Calibration overestimates both the coefficients.

The $\alpha_1$ estimates are severely biased, compared with the true values especially for lower $\lambda$ values. The $\alpha_2$ estimates are moderately biased.

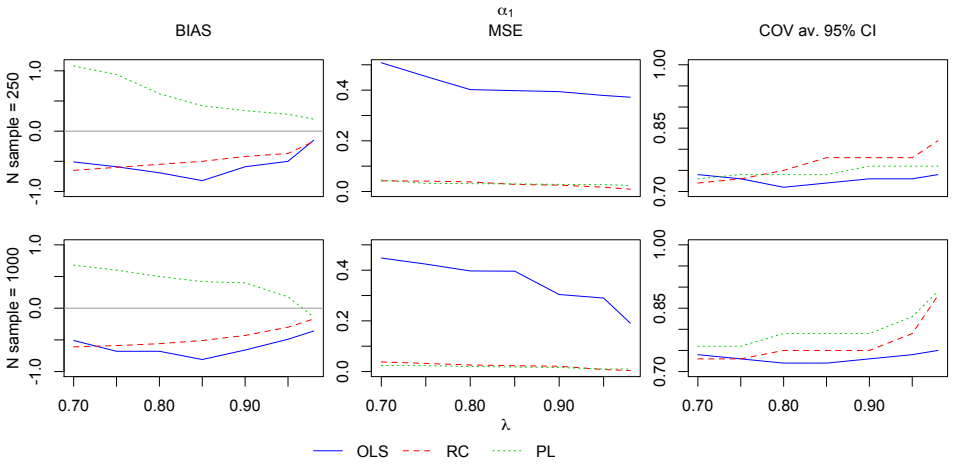


**FIGURE 1** Simulated Bias, Mean Squared Error (MSE) and Coverage Average 95% Confidence Interval (COV av. 95% CI) for $\alpha_1$ estimate with $\alpha_1 = -2$ and $\alpha_2 = -0.5$. Based on 10000 samples of size 250 and 1000. OLS is the Ordinary Least Square regression estimate, RC is the Regression Calibration estimate and PL is the Pseudo-Likelihood regression estimate.

Concerning MSE, the Naive approach is, without doubt, the worse estimator, concerning $\alpha_1$ estimate. We notice that naive estimator outperform RC and PL concerning $\alpha_2$ estimate for lower $\lambda$ values, especially for small sample sizes. Pseudo-Likelihood behaves better than Regression Calibration in small samples, nonetheless when the sample size is big Regression Calibration behaves better then Pseudo-Likelihood. The naive gain in MSE is only due to the variance.

The confidence limits from all three correction methods showed adequate coverage for all scenarios above 70%, nevertheless, Regression Calibration perform better than both the Pseudo-Likelihood and the OLS estimators for all sample s and $\lambda$.

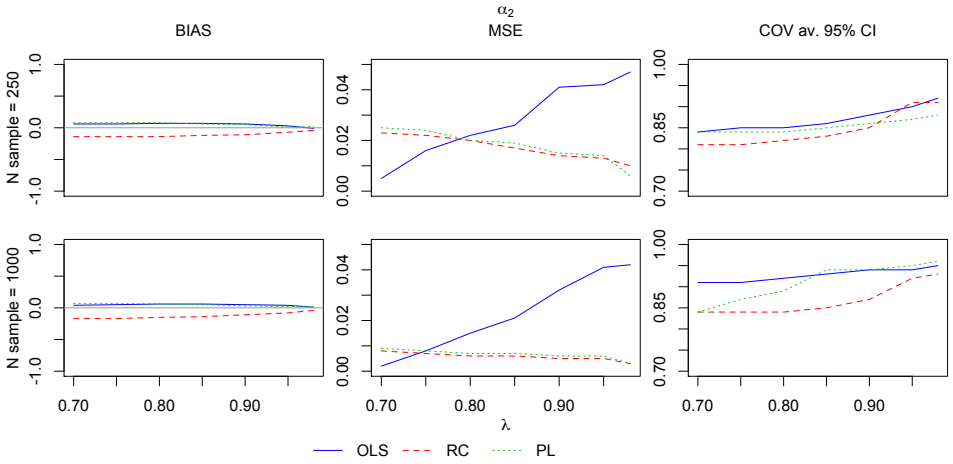


**FIGURE 2** Simulated Bias, Mean Squared Error (MSE) and Coverage Average 95% Confidence Interval (COV av. 95% CI) for $\alpha_2$ estimate with $\alpha_1 = -2$ and $\alpha_2 = -0.5$. Based on 10000 samples of size 250 and 1000. OLS is the Ordinary Least Square regression estimate, RC is the Regression Calibration estimate and PL is the Pseudo-Likelihood regression estimate.

It should be noted that for small samples, the techniques occasionally do not converge. For all three estimators, there was a large reduction in the standard error of the regression parameter estimates when the sample size was increased. Generally, all the correction procedures are best suited when the sample size is large.

## 5 | ANALYSIS OF SACIUCI DATA

Póvoa et al.[1] described a prospective, cohort, multi-centered study, conducted over one year (1 December, 2004 to 30 November, 2005) of biomarkers as prognostic factors in SEPSIS. Data of community-acquired sepsis study was collected in seventeen intensive care units (ICU) in Portugal. All adult patients (age $>$ 18 years) consecutively admitted in the participating ICUs were enrolled and screened for CAS. Patients were then followed up until death or hospital discharge. Sepsis and sepsis-related conditions were defined according to the criteria proposed by the American College of Chest Physicians/Society of Critical Care Medicine[7] .

Using data from the Portuguese Sepsis Adquirida na Comunidade e internada em Unidade de Cuidados Intensivos (SACiUCI) study[1] we investigated the relationship between serial CRP measurements after prescription of antibiotics and the clinical course of Community-Acquired Sepsis (CAS) admitted in Intensive Care Units (ICU). To build the relation, we consider the five measurements of CRP within the GLM model framework and the mortality (D). Only patients with available measurements at Day 1 till Day 5 were considered for this analysis. Patients were followed-up during the first five ICU days, day of ICU discharge or death and hospital outcome. CRP-ratio was calculated concerning Day 1 CRP concentration. Patients were classified according to the pattern of CRP-ratio response to antibiotics: fast response if Day 5 CRP-ratio was $< 0.4$, slow response if Day 5 CRP-ratio was between 0.4 and 0.8, and no response if Day 5 CRP-ratio was $> 0.8$. Data collection included demographic data and comorbid diseases. Clinical and laboratory data at the time of hospital admission was recorded. The Simplified Acute Physiology Score (SAPS) II was calculated. Microbiological and clinical infectious data were reported. C-reactive protein, body temperature, White Blood Cell count (WBC) and the Sequential Organ Failure Assessment (SOFA) score were evaluated during the first five days of ICU stay. Blood samples were obtained from an arterial line at ICU admission and subsequently every morning. For purposes of the time dependent analysis Day 1 (D1) was defined as the day of ICU admission. The following days were successively named as D2 up to D5. The withdrawal of the inflammatory stimulus results in a sharp decrease of the CRP serum concentration in a way similar to first-order elimination kinetics and, in that sense, relative CRP variations are more

**TABLE 1** Regression Estimates from SACiUCI data.

| | | | | $OR_{95\%}CI$ | |
|---|---|---|---|---|---|
| **Estimator** | **Estimate** | **SE** | **OR** | **LL** | **UL** |
| **CRP init** ($\alpha_1$) | | | | | |
| $Naive$ | .099 | .253 | 1.576 | ( .960, | 2.588) |
| $Regression\ Calibration$ | .105 | .270 | 1.554 | ( .916, | 2.636) |
| $Pseudo-Likelihood$ | .209 | .074 | 1.13 | (1.04 , | 1.23 ) |
| **CRP change** ($\alpha_2$) | | | | | |
| $Naive$ | .292 | .747 | 13.115 | (3.031, | 56.755) |
| $Regression\ Calibration$ | .330 | .845 | 18.187 | (3.474, | 95.197) |
| $Pseudo-Likelihood$ | 1.230 | .23 | 2.06 | (1.580, | 2.69 ) |

informative than absolute changes. As a result, we fitted a linear model for the CRP for each patient, i, and assessed the CRP baseline value (intercept) and change described by the slope in the model. Subsequently, we considered a prognostic model using the change of CRP, $P(D_i = 1) = g(\beta_0 + \widehat{\boldsymbol{\beta}}X)$ using $probit^{-1}$ as the link function.

The uncorrected analysis of these data was published by Póvoa et al.[8]. Table 1 compares standard methods to enhancements to RC and Pseudo-Likelihood discussed in this paper. The results from the uncorrected regression of Sepsis on X, the CRP baseline value and change, taken to be perfectly measured with a binary indicator for disease use the study data in an ordinary logistic regression model. The point estimate $\widehat{\boldsymbol{\beta}}$, its estimated standard error ($SE(\widehat{\boldsymbol{\beta}})$), the odds ratio and its 95% confidence interval are given. The bias-corrected point estimates were all three-fold smaller than their uncorrected counterpart. Note that the corrected CRP change estimate is approximately 10% smaller than the usual estimate. We are convinced that this is optimistic because the corrected SE estimator is smaller, approximately 10%, than the standard counterpart.

In this example, the augmented coefficient for CRP change index after correction can be interpreted as meaning that the confounding effects of CRP change in Sepsis prediction are only partially accounted for if one does not correct for measurement error.

## 6 | CONCLUDING REMARKS

The present article presents an efficient method for repeated estimation in a nonlinear model scenario. In contrast to previous works, we studied a robust method for consistent estimation in logistic/probit regression models when *measurement errors* exist in the predictors and the response is binary. Contrary to previous studies the *measurement error* distribution does not need to be known neither replicate data need to be available. The methods discussed in this paper assume non-differential *measurement error*, required for valid application of Regression Calibration method. Though, this assumption is usually fulfilled in prospective studies.

Our simulations confirm that even when the original model could be nicely estimated, the presence of the *measurement error* can severely bias the estimates if the surrogate is used in the estimation. It was shown that the performance of the various methods follows the patterns suggested by both the analytic results and the simulation studies presented in Sections 4 and 5, respectively.

Our simulations show that compared to OLS, RC and PL offers little or no advantage when sample sizes are small, but perform best when samples are reasonably large and especially when the measurement error or the effects are not small.

The OLS method performed poorly, with OLS overestimating the true values consistently except concerning the covariate measured without error, in which case, underestimate is performed.

Compared to the naive OLS approach, Regression Calibration grossly attenuates the estimated error in prediction, i.e, moreover, the magnitudes of these estimates are relatively close to the true coefficients, especially respecting the intercept. Concerning greater $\lambda$ values, i.e., $> 0.90$ estimates are very good. However, one would not have known this before performing this measurement error correction.

RC outperform the PL estimator, due in part to the large number of nuisance parameters, which must be managed for the PL estimator. Both the RC and the PL estimators may have a non-negligible bias for large sample sizes.

Even though PL approach is more sophisticated it seems to perform worse, especially for small samples, i.e., under rare disease conditions, as it varies profusely and, nevertheless bias "absolute value" decreases, it is unstable for small sample size, probably because when calculating PL we take into account many estimates witch makes the coefficients estimates unstable.

The naive OLS has a larger bias but smaller SE than Regression Calibration and Pseudo-Likelihood.

Generally, it is fair to say that the Regression Calibration and Pseudo-Likelihood method perform equally well, nonetheless, Regression Calibration seems to perform a bit better. In fact, predictably requires less than 10 Fisher-Scoring iterations to maximize the likelihood. Nevertheless, in the case of one variable, curiously, it is more efficient to use the OLS. We conclude that correcting for measurement errors instead of falsely assume that errors are not present will produce less bias than ignoring exposure measurement error in the analysis.

Our work was novel therein to the current time methods concerning measurement error invoked assumptions difficult to verify. Our method only rests on the assumption of non-differential measurement error and normality of the distribution. Regression Calibration method has gained some popularity because it's simplicity, and nevertheless, is an approximate we showed it perform well when compared to likelihood methods, an interesting alternative that many researchers put apart due to the difficulty of likelihood calculation in some designs. There continue to exist a huge number of unanswered questions about correction for covariate measurement error in nonlinear models when applied to longitudinal data or repeated measures. For instance, when there is more than one variable or when the number of covariates is greater than the number of measurements and for so, deserve further study. The first question is nowadays one of our research aims.

## 7 | AUTHORS PROFILE

MSc Rosa Santos Oliveira had his B.Sc. and M.Sc. Degrees in Mathematics from University of Porto in Portugal. She is currently working as a teacher of Mathematics at Instituto Politécnico do Cávado e do Ave and at Instituto Politécnico de Bragança. She is also a researcher at CINTESIS - Faculdade de Medicina do Porto. She has 17 years of teaching experience and 7 years of research experience.

## SOFTWARE

All programs were written in R, and are available from the first author upon request.To illustrate our programs, we have simulated a data set taking into care the real data set concerning the relevant covariate values. That simulated data set is also available. We will use lme4 package in R to fit linear and nonlinear mixed effect models, and use rstan to fit fully Bayesian multilevel models.

## ACKNOWLEDGMENTS

For helpful comments, discussions, and data, we thanks Profesor Raymond Carroll and MD Pedro Póvoa, respectively.

The work has been mostly supported by an unrestricted grant from ASSUCIP (Associação dos Amigos da Unidade de Cuidados Intensivos Polivalente, Hospital Geral de Santo António, Porto, Portugal) as well as by grants from GIS (Grupo de Infecção e Sepsis, Hospital de São João, Porto, Portugal), Merck, Sharp & Dohme and Eli Lilly. AMT-P is a researcher of the project PIC/IC/83312/2007, funded by Fundação para a Ciência e Tecnologia. for making the data available to us.

## APPENDIX

## 8 | THEORETICAL PROOFS

This section reports exact calculations designed to investigate the accuracy of the approximations proposed. These expressions depend on features of the joint distribution of the covariates $X$ and of the measurement error $U$ , as well as on assumptions on the conditional expectation of $Y$ given $D$ and $X$. In what follows we present results from numerical calculations obtained for

the simple setup maintained in the leading example allowing for departures from normality for both the distribution of $X$ and the distribution of $U$.

Advantage of the Bayesian approach vs Frequencist approach: - we can write the model using a more complex approach in the sense we do not have to write the model stepwise

- In bayesian approach, we can have a simultaneous model: with longitudinal and the probit regression. Then, after, we do a posterior for all the model at the same time.
- In frequencist approach, writing this model all at once is a very complicated likelihood.

Having the data:

$\int \Phi(a+bx)\phi(x)dx = \Phi\{a/(1+b^2)^{1/2}\}\ \mathrm{pr}(D=1|W)\ \sigma_\epsilon^2$

$$\begin{aligned} L(pcr_1,...,pcr_5,death) &= \prod[f(pcr_1,pcr_5,death)] \\ &= \prod[f(death|pcr_1,...,pcr_5)\times f(pcr_1,...,pcr_5)] \\ &= \prod[\int[f(death|pcr_1,...,pcr_5)\times f(pcr_1,...,pcr_5,u)],du] \\ &= \prod\int[f(death|pcr_1,...,pcr_5,u)\times f(pcr_1|u)...f(pcr_5|u)f(u)du. \end{aligned}$$

My calculus to the correcting factor:

$$\begin{aligned} E(y) &= 0\times Pr(y=0)+1\times Pr(y=1) \\ &= Pr(y=1) \\ &= \int yf(y)dy \\ &= \Phi(x^{\mathrm{T}}\beta) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x^{\mathrm{T}}\beta}{\sqrt{2}}} exp(-\tfrac{1}{2}t^2)dt \end{aligned}$$

assuming b is normal distributed

$$\begin{aligned} E(y) &= \int yf(y)dy \\ &= \int\int yf(y,b)dydb \\ &= \int\int yf(y|b)f(b)dydb \\ &= \int \Phi(\alpha+\beta x+b)\Phi(b)db \end{aligned}$$

$$F(\alpha+\beta x+b,\mu,\sigma^2) = \Phi\left(\frac{(\alpha+\beta x+b)-\mu}{\sigma^2}\right)$$

Assuming $\mu=0$ and $\sigma=1$ comes:

$$F(\alpha+\beta x+b) = \Phi(\alpha+\beta x+b)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x+b} exp\left(-\frac{1}{2}y^2\right) dy$$

And, as $\Phi(b) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)$, follows:

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x+b} exp\left(-\frac{1}{2}y^2\right) dy \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right) db$$

as we assumed $b \sim \mathcal{N}(0,\sigma)$ follows

$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x+b} exp\left(-\frac{1}{2}y^2\right) dy \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{b^2}{2\sigma^2}\right) db$$

$$\int_{-\infty}^{+\infty} \Phi(\alpha+\beta x+b)\Phi(b) db = \int_{-\infty}^{+\infty}\int_{-\infty}^{\alpha+\beta x+b} \Phi(\alpha+\beta x+b) dy \Phi(b) db$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{\alpha+\beta x+b} \Phi(\alpha+\beta x+b)\Phi(b) db dy$$
$$= \int_{-\infty}^{+\infty}\int_{0}^{\alpha+\beta x+b} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}y^2\right) dy \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{b^2}{2\sigma^2}\right) db$$

Let $y = y^* + (\alpha+\beta x+b)$, i.e, $y^* = y - (\alpha+\beta x+b)$, then

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(y_{1i}^* - \alpha_i - \beta_1 x_i - b_i)^2}{2}\right) dy^* \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{b_i^2}{2\sigma^2}\right) db$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(y_{1i}^* - \alpha_i - \beta_1 x_i - b_i)^2}{2}\right) \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{b_i^2}{2\sigma^2}\right) dy^* db$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{0} \frac{1}{\sigma 2\pi} exp\left(-\frac{(y_{1i}^* - \alpha_i - \beta_1 x_i - b_i)^2}{2}\right) exp\left(-\frac{b_i^2}{2\sigma^2}\right) dy^* db$$
$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp\left(-\frac{(y_{1i}^* - \alpha_i - \beta_1 x_i - b_i)^2}{2}\right) exp\left(-\frac{b_i^2}{2\sigma^2}\right) db dy^*$$
$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp\left(-\frac{\sigma^2(y_{1i}^* - \alpha_i - \beta_1 x_i - b_i)^2 + b_i^2}{2\sigma^2}\right) db dy^*$$

let $y_{1i}^* - \alpha_i - \beta_1 x_i = t_i$, to simplify notation

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp\left(-\frac{\sigma^2(t_i-b_i)^2+b_i^2}{2\sigma^2}\right) dbdy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp\left(-\frac{\sigma^2(t_i^2-2t_ib_i+b_i^2)+b_i^2}{2\sigma^2}\right) dbdy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp - \left[tb_i^2 + \frac{1+\sigma^2}{2\sigma^2}b_i^2\right] dbexp - \frac{t_i^2}{2} dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp - \left[\left(\frac{1+\sigma^2}{2\sigma^2}b_i^2 + \frac{\sigma^2 t_i^2}{2(1+\sigma^2)} - tb_i\right)\right] exp\left(\frac{\sigma^2 t_i^2}{2(1+\sigma^2)}\right) dbexp\left(-\frac{t_i^2}{2}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp - \left[\sqrt{\left(\frac{1+\sigma^2}{2\sigma^2}\right)}b_i - \frac{\sigma t_i}{\sqrt{2(1+\sigma^2)}}\right]^2 dbexp\left(\frac{-t_i^2}{2}\right) exp\left(\frac{\sigma^2 t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sigma 2\pi} exp - \left[\sqrt{\left(\frac{1+\sigma^2}{2\sigma^2}\right)}b_i - \frac{\sigma t_i}{\sqrt{2(1+\sigma^2)}}\right]^2 dbexp\left(\frac{t_i^2 - sigma^2 t_i^2 + sigma^2 t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \sqrt{\left(\frac{\pi}{\frac{\sigma^2+1}{2\sigma^2}}\right)} \frac{1}{\sigma 2\pi} exp\left(\frac{-t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi}\sqrt{\frac{1+\sigma^2}{2}}} exp\left(\frac{-t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi(1+\sigma^2)}} exp\left(\frac{-t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{(1+\sigma^2)}} exp\left(\frac{-t_i^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{(1+\sigma^2)}} exp\left(\frac{-(y_{1i}^* - \alpha_i - \beta_1 x_i)^2}{2(1+\sigma^2)}\right) dy^*$$

$$= \int_{-\infty}^{0}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{(1+\sigma^2)}} exp\left[-\frac{\left[\frac{(y_{1i}^* - (\alpha_i+\beta_1 x_i))^2}{\sqrt{1+\sigma^2}}\right]}{2}\right] dy^*$$

let $y_{1i}^{**} = \frac{y_{1i}^* - (\alpha_i+\beta_1 x_i)}{\sqrt{1+\sigma^2}}$, then $dy^* = \sqrt{1+\sigma^2}$, so

$$
\begin{aligned}
&= \int_{-\infty}^{-\frac{\alpha_i+\beta_1 x_i}{\sqrt{1+\sigma^2}}} \frac{1}{\sqrt{2\pi}\sqrt{(1+\sigma^2)}} exp\left(-\frac{\left(y_{1i}^{**}\right)^2}{2}\right)\sqrt{1+\sigma^2}dy^{**} \\
&= \int_{-\infty}^{-\frac{\alpha_i+\beta_1 x_i}{\sqrt{1+\sigma^2}}} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{\left(y_{1i}^{**}\right)^2}{2}\right) dy^{**} \\
&= \int_{-\infty}^{-\frac{\alpha_i+\beta_1 x_i}{\sqrt{1+\sigma^2}}} \phi\left(\alpha^* + \beta^* x + b\right) dy^* \\
&= \Phi\left[\frac{\alpha_i + \beta_1 x_i}{\sqrt{1+\sigma^2}}\right] \\
&= \Phi\left(\alpha_i^* + \beta^* x\right)
\end{aligned}
$$

$$
\lambda = \left[\frac{1}{\sqrt{1+\sigma^2}}\right]
$$

$$
\lambda \times \sqrt{1+\sigma^2} = \left[\frac{1}{\sqrt{1+\sigma^2}}\right]
$$

## References


1. Póvoa P. C-reactive protein: a valuable marker of sepsis.. *Intensive Care Medicine* 2002; 28 (3): 235–243.
2. Tosteson JP, Demidenko E. Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicina* 1998; 17: 1959–1971.
3. Carroll R, Ruppert D, Stefanski L, Crainiceanu C. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall . 2006.
4. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* 1990; 132: 734–745.
5. W. SD, . PK. Likelihood analysis for errors-in-variables regression with replicate measurements.. *Biometrika* 1996; 83: 813–824.
6. Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association,* 1996; 91: 242–250.
7. R.C. B, R.A. B, F.B. C, R.P. D, A.M. F, W.J. KWSRS. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis.. In: . 101. ; 1992: 1644-1655.
8. Póvoa P, M. TPA, Carneiro AH, SACiUCI. PCASSG. C-reactive protein, an early marker of community-acquired sepsis resolution: a multi-center prospective observational study. *Critical Care* 2011; 15 (4).
9. Lawrence Gould CMIJQGMSBF. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group.. *Stat Med.* 2015; 34(14): 2181-95.

10. McCulloch CE, Searle SR. *Generalized, Linear and Mixed Models*. Wiley InterScience . 2001.

11. Akaike H. Information Theory and an Extension of The Maximum Likelihood Principle. *International Symposium on Information Theory* 1973; 2: 267–281.

12. Erling BA. Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society, Series B* 1971; 32: 283–301.

13. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* 1997; 16: 103–116.

14. Carroll RJ, Spiegelman CH, Gordon Lan KK, Bailey KT, Abbott RD. On errors-in-variables for binary regression models. *Biometrika* 1984; 71: 19–25.

15. Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association,* 1996; 91: 242–250.

16. Chen B, Yi GY, Cook RJ. Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data That Are Missing at Random. *Journal of the American Statistical Association* 2010; 105: 336–353.

17. J. G. C-reactive protein: risk factor, biomarker and/or therapeutic target–. *Canadian Journal of Cardiology* 2010; 26, Suppl A: 41A–44A.

18. Huang YJ, Wang CY. Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association* 2001; 96: 1469–1482.

19. Ma Y, Tsiatis AA. Closed form semiparametric estimators for measurement error models. *Stat. Sin.* 2006; 16: 183–193.

20. Chen MH, Ibrahim JG, Shao QM. Propriety of the Posterior Distribution and Existence of the MLE for Regression Models with Covariates Missing at Random. *Journal of the American Statistical Association* 2004; 99: 421–438.

21. Nugent W, Graycheck L, Basham R. A Devil Hidden in the Details: The Effects of Measurement Error in Regression Analysis. *Journal of Social Service Research* 2000; 27: 53–75.

22. Cox C. Nonlinear quasi-likelihood models: applications to continuous proportions. *Computational Statistics and Data Analysis* 1996; 21: 449–461.

23. Cox DR. Partial likelihood. *Biometrika* 1975; 62: 269–276.

24. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data, 2nd edition*. Oxford, U.K.: Oxford University Press. . 2002.

25. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. Wiley-Interscience . 2004.

26. Fitzmaurice GM, Laird NM, Zahner GEP. Multivariate Logistic Models for Incomplete Binary Responses. *Journal of the American Statistical Association* 1996; 91: 99–108.

27. Fitzmaurice GM, Lipsitz SR, Molenberghs G, Ibrahim JG. A Protective Estimator for Longitudinal Binary Data Subject to Non-Ignorable Non-Monotone Missingness. *Journal of the Royal Statistical Society. Series A* 2005; 168: 723–735.

28. Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression Models for Longitudinal Binary Responses with Informative Drop-Outs. *Journal of the Royal Statistical Society. Series B* 1995; 57: 691–704.

29. Fuller W. *Measurement Error Models*. John Wiley and Sons . 1987.

30. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; 38: 963–974.

31. Li Y, Lin X. Covariate Measurement Errors in Frailty Models for Clustered Survival Data. *Biometrika* 2000; 87 (4): 849–866.

32. Chaganty NR, Joe H. Efficiency of Generalized Estimating Equations for Binary Responses. *Journal of the Royal Statistical Society, Series B* 2004; 66: 851–860.

33. Cook JR, Stefanski L. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association,* 1994; 89: 1314–1328.

34. W. SD. Semiparametric maximum likelihood for measurement error regression.. *Biometrika* 2001; 57: 53–61.

35. Stefanski LA, Buzas JS. Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association* 1995; 90: 541–550.

36. Stefanski LA, Carroll R. Covariate measurement error in logistic regression. *Annals of Statistics* 1985; 13: 1335–1351.

37. Stefanski LA, Carroll R. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* 1987; 74: 703–716.