A heteroscedastic polynomial regression with measurement error in both axes

Alexandre G. Patriota and Heleno Bolfarine
Institute of Mathematics and Statistics - USP, São Paulo, Brasil

Abstract

With epidemiological and astronomical data, it is common to observe variances that vary with the observations. Further, values for those variances typically are available from follow-up studies or replications. This paper deals with consistent estimation and hypothesis testing in a heteroscedastic polynomial model with measurement error in both axes and an equation error. For obtaining consistent estimators and consistently assessing their asymptotic variances, we embrace the corrected score approach. Furthermore, we applied the theoretical results in two real data sets: the WHO MONICA project data set on cardiovascular diseases and their risk factors and the *Chandra* observatory data set. We also simulate the rejection rates for the Wald statistic in order to study test size and power for small and moderate samples, indicating that the test behaves satisfactorily in those situations.

keywords Polynomial regression, measurement error, corrected score, asymptotic theory

1 Introduction

Recently, heteroscedastic linear errors-in-variables models have been proposed to fit epidemiological (Kulathinal et al., 2002; Cheng and Riu, 2006; de Castro et al., 2007) and astronomical (Akritas and Bershady, 1996; Kelly, 2007; Kelly et al., 2008) data sets. In Kulathinal et al. (2002) was proposed a simple EM (Expectation and Maximization) algorithm to find the maximum likelihood (ML) estimators of a linear heteroscedastic structural errors-in-variables model. The authors considered that the linear equation is subject to error and applied this model to a real dataset of the WHO MONICA project on cardiovascular disease and its risk factors. For this data set, it was found a significant variance for the equation error, which makes such more complex models useful in fitting real data sets. In the same way, de Castro et al. (2007) derived the Fisher information for the parameters which makes it possible to test conjointly the intercept and inclination parameters using Wald type statistics. They also proposed testing statistics based on

the likelihood ratio and score statistics. In a previous study, Akritas and Bershady (1996) has entertained a similar model (with known covariance between the errors) and applied it to an astronomical data set. The authors had proposed a method-of-moment to estimate the model parameters and gave an approximation for their asymptotic covariance matrix. Motivated by these applications, we assume a heteroscedastic polynomial model with error in both axes adding an equation error, which seems not available in literature. Then, using the data sets produced by the WHO MONICA project and by the *Chandra* X-ray observatory, we found evidence of a quadratic and cubic relationship, respectively, relating the response variable and the covariate (both inaccessible directly). It is also the case that the model presented in this paper extends the model considered in Zavala et al. (2007), where a heteroscedastic polynomial nonequation error model is considered. The approach is based on the corrected score methodology, which when feasible, yields consistent and asymptotically normal estimators for the model parameters. Moreover, consistent estimators for the asymptotic variances can also be obtained. In this paper, we also consider a "flexible" polynomial model where is possible to fit partial polynomial relationships between y_i (unobservable response variable) and x_i (unobservable covariate). That is, in a third degree polynomial, for example, coefficient for x^2 may be taken as zero.

Most of the literature deal with the homoscedastic case and error in just one of the axis. See, for example, Chan and Mak (1985), Fuller (1987), Cheng and Scheneeweis (1998) and Kukush (2005). An exception is Zavala et al. (2007) where a heteroscedastic polynomial errors-in-variables model without equation error is considered.

This article is organized as follows. Section 2 gives a way to fit partial polynomial models, specifically, in a heteroscedastic polynomial errors-in-variables model without equation error (the same model considered by Zavala et al., 2007). Section 3 considers a more general model which regards an equation error in the model presented in Section 2. Section 4 presents a simulation study where it is shown that the proposed approach yields Wald tests with empirical levels close to the nominal significance levels for small and moderate samples. Section 5 deals with applications to the WHO MONICA and *Chandra* data sets. Section 6 ends the paper with conclusions and remarks.

2 Partial polynomial errors-in-variables models without equation error

In this section we present a concise implementation of partial polynomial models with no equation error to the model considered in Zavala et al. (2007) where it is specified that:

$$\begin{array}{rcl} Y_i & = & y_i + e_i, \\ X_i & = & x_i + u_i, \end{array} \tag{1}$$

with $y_i = \beta_0 + \beta_1 x_i + \ldots, \beta_p x_i^p$ and

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{bmatrix} \lambda_i & 0 \\ 0 & \kappa_i \end{bmatrix} \right)$$

with λ_i and κ_i known for $i = 1, \ldots, n$. The authors provided consistent estimators of the model parameters and gave consistent estimates for their asymptotic covariance matrix. However, if some coefficients in the polynomial equation are equal to zero, then we have to derive the estimators and consistent estimators for the asymptotic covariance matrix for each case. As an alternative, we are going to present a general way to fit partial polynomial models which have some of the parameters equal to zero. For that, consider initially the equation $y_i = a_0\beta_0 + a_1\beta_1 x_i + \ldots, a_p\beta_p x_i^p$ which, in matrix notation, can be written as

$$y_i = \boldsymbol{\beta}_F^\top \boldsymbol{A} \ddot{\boldsymbol{x}}_i \tag{2}$$

where

$$\boldsymbol{\beta}_F = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{A} = \begin{bmatrix} a_0 & 0 & 0 & \ldots & 0 \\ 0 & a_1 & 0 & \ldots & 0 \\ 0 & 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & a_p \end{bmatrix} \qquad \text{and} \qquad \ddot{\boldsymbol{x}}_i = \begin{pmatrix} 1 \\ x_i \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix}.$$

The elements of the matrix $\boldsymbol{A}$ are known in such way that, $a_j = 0$ if $\beta_j = 0$ and 1, otherwise, for all $j = 0, 1, \ldots, p$. The model studied by Zavala et al. (2007) considers that $a_j = 1$ for all $j = 0, \ldots, p$.

To consistently estimate the model (1) parameters considering (2) we consider the corrected score approach (for details, see Nakamura, 1990) which depends on a pseudo log-likelihood function $\ell^*(\boldsymbol{\theta}, \boldsymbol{X}) = \sum_{i=1}^n \ell_i^*(\boldsymbol{\theta}, \boldsymbol{X})$ satisfying

$$E(\ell^*(\boldsymbol{\theta}, \boldsymbol{X}) | \boldsymbol{Y}, \boldsymbol{x}) = \ell(\boldsymbol{\theta}, \boldsymbol{x}),$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ is the (unobserved) log-likelihood function of $(\boldsymbol{Y}, \boldsymbol{x})$ and $\ell^*(\boldsymbol{\theta}, \boldsymbol{X})$ is called the corrected log-likelihood function which depends only on the observable data $(\boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$. Note that we omit the response variable $\boldsymbol{Y}$ in the expressions $\ell(.)$ and $\ell^*(.)$ to simplify notation. For the nonequation error model, the unknown parameter $\boldsymbol{\theta}$ is the unknown $\boldsymbol{\beta}$. We can then define the following quantities

$$\boldsymbol{U}^*(\boldsymbol{\theta}, \boldsymbol{X}) = \sum_{i=1}^{n} \frac{\partial \ell_i^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \boldsymbol{I}^*(\boldsymbol{\theta}, \boldsymbol{X}) = -\sum_{i=1}^{n} \frac{\partial^2 \ell_i^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

and suppose that $\widehat{\boldsymbol{\theta}}_n$ is such that $\boldsymbol{U}^*(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{X}) = \boldsymbol{0}$, which is the corrected score estimator of $\boldsymbol{\theta}$. Under the regularity conditions stated in Gimenez and Bolfarine (1997) the corrected score estimator, $\hat{\boldsymbol{\theta}}_n$, is consistent and asymptotically normal that is, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_n)$, where $\boldsymbol{\Omega}_n$ is a sandwich type matrix which can be consistently estimated by

$$\widehat{\boldsymbol{\Omega}}_n = \frac{1}{n} \boldsymbol{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}}_n) \boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) \boldsymbol{\Lambda}_n^{-1}(\widehat{\boldsymbol{\theta}}_n),$$

where

$$\boldsymbol{\Lambda}_n(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{I}_i^*(\boldsymbol{\theta}, \boldsymbol{X}) \quad \text{and} \quad \boldsymbol{\Gamma}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i^*(\boldsymbol{\theta}, \boldsymbol{X}) \boldsymbol{U}_i^*(\boldsymbol{\theta}, \boldsymbol{X})^\top,$$

with $\boldsymbol{U}_i^*(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{\partial \ell_i^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{I}_i^*(\boldsymbol{\theta}, \boldsymbol{X}) = \frac{\partial^2 \ell_i^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$.

In order to apply the Nakamura's approach, we start by writing the (unobserved) log-likelihood function for model (1) which is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}_F, \boldsymbol{x}) &\propto -\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\lambda_i} (Y_i - y_i)^2 \\ &\propto \boldsymbol{\beta}_F^\top \boldsymbol{A} \sum_{i=1}^{n} \frac{Y_i \ddot{\boldsymbol{x}}_i}{\lambda_i} - \frac{1}{2} \boldsymbol{\beta}_F^\top \boldsymbol{A} \sum_{i=1}^{n} \left(\frac{\ddot{\boldsymbol{x}}_i \ddot{\boldsymbol{x}}_i^\top}{\lambda_i} \right) \boldsymbol{A} \boldsymbol{\beta}_F. \end{aligned}$$

Hence, for the purpose of implementing the corrected score approach, we have to find the quantities $t_{i,k}$ such that $E(t_{i,k}|x_i) = x_i^k$, $k = 1, \ldots, 2p$ in which, under normality, we have that (see Zavala et al., 2007)

$$t_{i,0} = 1, \qquad t_{i,1} = X_i \qquad \text{and} \qquad t_{i,(j+1)} = X_i t_{i,j} - j\kappa_i t_{i,(j-1)},$$

$j = 1, \ldots, 2p$. Moreover, defining

$$\boldsymbol{H}_i = \begin{bmatrix} 1 & t_{i,1} & t_{i,2} & \ldots & t_{i,p} \\ t_{i,1} & t_{i,2} & t_{i,3} & \ldots & t_{i,(p+1)} \\ t_{i,2} & t_{i,3} & t_{i,4} & \ldots & t_{i,(p+2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{i,p} & t_{i,(p+1)} & t_{i,(p+2)} & \ldots & t_{i,2p} \end{bmatrix},$$

$$\boldsymbol{h}_i = Y_i \begin{pmatrix} 1 & t_{i,1} & t_{i,2} & \ldots & t_{i,p} \end{pmatrix}^\top, \quad \boldsymbol{T}_n = \sum_{i=1}^{n} \boldsymbol{H}_i/\lambda_i \quad \text{and} \quad \boldsymbol{F}_n = \sum_{i=1}^{n} \boldsymbol{h}_i/\lambda_i,$$

it can be verified, by showing that $E[\ell^*(\boldsymbol{\beta}_F, \boldsymbol{X})|\boldsymbol{x}] = \ell(\boldsymbol{\beta}_F, \boldsymbol{x})$ holds, that the corrected log-likelihood function for the observed data $(\boldsymbol{Y}, \boldsymbol{X})$ is given by

$$\ell^*(\boldsymbol{\beta}_F, \boldsymbol{X}) \propto \boldsymbol{\beta}_F^\top \boldsymbol{A}\boldsymbol{F}_n - \frac{1}{2}\boldsymbol{\beta}_F^\top \boldsymbol{A}\boldsymbol{T}_n\boldsymbol{A}\boldsymbol{\beta}_F.$$

Notice that we can use the decomposition $\boldsymbol{A} = \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$, in such a way that $\boldsymbol{\beta} = \boldsymbol{\Delta}^\top\boldsymbol{\beta}_F$ is the vector which has all components different from zero, so that the corrected log-likelihood function for $\boldsymbol{\beta}$ can be written as

$$\ell^*(\boldsymbol{\beta}, \boldsymbol{X}) \propto \boldsymbol{\beta}^\top \dot{\boldsymbol{F}}_n - \frac{1}{2}\boldsymbol{\beta}^\top \dot{\boldsymbol{T}}_n\boldsymbol{\beta}, \tag{3}$$

where $\dot{\boldsymbol{F}}_n = \boldsymbol{\Delta}\boldsymbol{F}_n$ and $\dot{\boldsymbol{T}}_n = \boldsymbol{\Delta}^\top\boldsymbol{T}_n\boldsymbol{\Delta}$. Therefore, differentiating the corrected log-likelihood function (3) we have the corrected score function and it is given by

$$\boldsymbol{U}^*(\boldsymbol{\beta}, \boldsymbol{X}) = \frac{\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{X})}{\partial \boldsymbol{\beta}} = \dot{\boldsymbol{F}}_n - \dot{\boldsymbol{T}}_n\boldsymbol{\beta} \tag{4}$$

Equating (4) to zero, we assess consistent estimators for $\boldsymbol{\beta}$ by solving the equation $\dot{\boldsymbol{T}}_n\widehat{\boldsymbol{\beta}}_n = \dot{\boldsymbol{F}}_n$, leading to

$$\widehat{\boldsymbol{\beta}}_n = \dot{\boldsymbol{T}}_n^{-1}\dot{\boldsymbol{F}}_n \tag{5}$$

We can estimate the asymptotic covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}_n$ using the asymptotic distribution for the corrected score estimator (see Gimenez and Bolfarine, 1997) given by

$$\dot{\boldsymbol{T}}_n^{-1}\boldsymbol{\Lambda}_n\dot{\boldsymbol{T}}_n^{-1}, \tag{6}$$

where $\boldsymbol{\Lambda}_n = \sum_{i=1}^n \boldsymbol{U}_i^*(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{X})\boldsymbol{U}_i^*(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{X})^\top$ and $\boldsymbol{U}_i^*(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{X}) = \frac{1}{\lambda_i}\left(\ddot{\boldsymbol{h}}_i - \ddot{\boldsymbol{H}}_i\widehat{\boldsymbol{\beta}}_n\right)$ with $\ddot{\boldsymbol{H}}_i = \boldsymbol{\Delta}^\top \boldsymbol{H}_i \boldsymbol{\Delta}$ and $\ddot{\boldsymbol{h}}_i = \boldsymbol{\Delta}^\top \boldsymbol{h}_i$.

In addition, for testing $H_0 : \boldsymbol{G\beta} = \boldsymbol{d}$ we may use the Wald statistic given by

$$\xi_n = (\boldsymbol{G}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{d})^\top \left(\boldsymbol{G}\dot{\boldsymbol{T}}_n^{-1}\boldsymbol{\Lambda}_n\dot{\boldsymbol{T}}_n^{-1}\boldsymbol{G}^\top\right)^{-1}(\boldsymbol{G}\widehat{\boldsymbol{\beta}}_n - \boldsymbol{d}).$$

Under the assumptions stated in Zavala et al. (2007), that is, there exists a $\gamma > 0$ such that

$$\lim \frac{1}{n^{1+\gamma/2}} \sum_{i=1}^n |x_i^p|^{(2+\gamma)} = 0 \qquad \text{and} \tag{7}$$

$$0 < \liminf \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^{2(2p-1)} \leq \limsup \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^{2(2p-1)} < \infty,$$

then $\xi_n \xrightarrow{D} \chi^2(g)$, where $g = \text{rank}(G)$ and "$\xrightarrow{D}$" means convergence in distribution.

For the partial cubic model $y_i = \beta_0 + \beta_1 x_i + \beta_3 x_i^3$ we have that $a_0 = 1$, $a_1 = 1$, $a_2 = 0$ and $a_3 = 1$. The matrix $\boldsymbol{\Delta}$ is as follows

$$\boldsymbol{\Delta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and a consistent estimator for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_3)^\top$ is computed by (5) and its asymptotic covariance by (6) where

$$\dot{\boldsymbol{T}}_n = \sum_{i=1}^n \frac{1}{\lambda_i}\begin{bmatrix} 1 & t_{i,1} & t_{i,3} \\ t_{i,1} & t_{i,2} & t_{i,4} \\ t_{i,3} & t_{i,4} & t_{i,6} \end{bmatrix} \qquad \text{and} \qquad \dot{\boldsymbol{F}}_n = \sum_{i=1}^n \frac{1}{\lambda_i}\begin{pmatrix} Y_i \\ Y_i t_{i,1} \\ Y_i t_{i,3} \end{pmatrix}.$$

3 Partial polynomial errors-in-variables models with equation error

In this paper, we also consider that

$$y_i | x_i \sim \mathcal{N}(\boldsymbol{\beta}_F \boldsymbol{A}\ddot{\boldsymbol{x}}_i; \sigma^2), \tag{8}$$

i.e., the equation is subject to error. This means that the true variables y_i and x_i are not perfectly related (Cheng and Riu, 2006). Therefore, the log-likelihood function considering (8) is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}_F, \boldsymbol{x}) &\propto -\frac{1}{2}\sum_{i=1}^{n}\log\tau_i - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\tau_i}\left(Y_i - y_i\right)^2 \\ &\propto -\frac{1}{2}\sum_{i=1}^{n}\log\tau_i - \frac{1}{2}\sum_{i=1}^{n}\frac{Y_i^2}{\tau_i} + \\ &\quad +\boldsymbol{\beta}_F^\top \boldsymbol{A}\sum_{i=1}^{n}\frac{Y_i\ddot{\boldsymbol{x}}_i}{\tau_i} - \frac{1}{2}\boldsymbol{\beta}_F^\top \boldsymbol{A}\sum_{i=1}^{n}\left(\frac{\ddot{\boldsymbol{x}}_i\ddot{\boldsymbol{x}}_i^\top}{\tau_i}\right)\boldsymbol{A}\boldsymbol{\beta}_F,\end{aligned}$$

where $\tau_i = \lambda_i + \sigma^2$ and $\boldsymbol{\theta}_F = (\boldsymbol{\beta}_F^\top, \sigma^2)^\top$. Define

$$\boldsymbol{T}_n(\sigma^2) = \sum_{i=1}^{n}\boldsymbol{H}_i/\tau_i, \qquad \boldsymbol{F}_n(\sigma^2) = \sum_{i=1}^{n}\boldsymbol{h}_i/\tau_i, \qquad \ddot{\boldsymbol{T}}_n(\sigma^2) = \sum_{i=1}^{n}\dot{\boldsymbol{H}}_i/\tau_i$$

$$\text{and} \qquad \ddot{\boldsymbol{F}}_n(\sigma^2) = \sum_{i=1}^{n}\dot{\boldsymbol{h}}_i/\tau_i,$$

where $\boldsymbol{H}_i$, $\boldsymbol{h}_i$, $\dot{\boldsymbol{H}}_i$ and $\dot{\boldsymbol{h}}_i$ are the same quantities defined in Section 2. Then, the corrected log-likelihood function for the observed data $(\boldsymbol{Y}, \boldsymbol{X})$ is given by

$$\ell^*(\boldsymbol{\theta}, \boldsymbol{X}) \propto -\frac{1}{2}\sum_{i=1}^{n}\log\tau_i - \frac{1}{2}\sum_{i=1}^{n}\frac{Y_i^2}{\tau_i} + \boldsymbol{\beta}^\top\sum_{i=1}^{n}\frac{\ddot{\boldsymbol{h}}_i}{\tau_i} - \frac{1}{2}\boldsymbol{\beta}^\top\sum_{i=1}^{n}\frac{\ddot{\boldsymbol{H}}_i}{\tau_i}\boldsymbol{\beta},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. Differentiating the corrected log-likelihood function we have the corrected score functions which are given by

$$\boldsymbol{U}^*_{\sigma^2}(\boldsymbol{\beta}, \boldsymbol{X}) = \frac{\partial\ell^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\left\{\frac{\ddot{\boldsymbol{h}}_i - \ddot{\boldsymbol{H}}_i\boldsymbol{\beta}}{\tau_i}\right\} \tag{9}$$

and

$$U^*(\sigma^2, \boldsymbol{X}) = \frac{\partial\ell^*(\boldsymbol{\theta}, \boldsymbol{X})}{\partial\sigma^2} = \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{c_i}{\tau_i^2} - \frac{1}{\tau_i}\right\}, \tag{10}$$

where $c_i = Y_i^2 - 2\boldsymbol{\beta}^\top\ddot{\boldsymbol{h}}_i + \boldsymbol{\beta}^\top\ddot{\boldsymbol{H}}_i\boldsymbol{\beta}$. Equating (9) and (10) to zero we assess consistent estimators for $\boldsymbol{\beta}$ and σ^2 that are obtained solving the following equations

$$\ddot{\boldsymbol{T}}_n(\widehat{\sigma}^2)\widehat{\boldsymbol{\beta}}_n = \ddot{\boldsymbol{F}}_n(\widehat{\sigma}^2) \quad \text{and} \quad \sum_{i=1}^{n} \frac{1}{\widehat{\tau}_i} = \sum_{i=1}^{n} \frac{\widehat{c}_i}{\widehat{\tau}_i^2}, \tag{11}$$

where $\widehat{\tau}_i = \lambda_i + \widehat{\sigma}^2$ and $\widehat{c}_i = Y_i^2 - 2\widehat{\boldsymbol{\beta}}_n^\top \ddot{\boldsymbol{h}}_i + \widehat{\boldsymbol{\beta}}_n^\top \ddot{\boldsymbol{H}}_i \widehat{\boldsymbol{\beta}}_n$. Equations in (11) do not have analytical solutions, though we can find the estimates using the following numerical procedure:

1. Start the procedure by setting $v = 0$ and find the initial estimates $\widehat{\boldsymbol{\theta}}^{(v)} = (\widehat{\boldsymbol{\beta}}_n^{(v)^\top}, \widehat{\sigma}^{2(v)})^\top$, where $\widehat{\boldsymbol{\beta}}_n^{(0)}$ and $\widehat{\sigma}^{2(0)}$ are the initial estimates for $\boldsymbol{\beta}$ and σ^2;
2. Compute

$$\widehat{\boldsymbol{\theta}}^{(v+1)} = \widehat{\boldsymbol{\theta}}^{(v)} + k \left[\boldsymbol{V}^* \left(\widehat{\boldsymbol{\theta}}^{(v)}, \boldsymbol{X} \right) \right]^{-1} \boldsymbol{U}^* \left(\widehat{\boldsymbol{\theta}}^{(v)}, \boldsymbol{X} \right),$$

where $k \in (0, 1]$ is a constant to avoid non-convergence (usually $k = 1$),

$$\boldsymbol{V}^* \left(\widehat{\boldsymbol{\theta}}, \boldsymbol{X} \right) = \sum_{i=1}^{n} E \left(-\frac{\widehat{\partial U_i^*(\boldsymbol{\theta})}}{\partial \boldsymbol{\theta}} \right) = \begin{bmatrix} \ddot{\boldsymbol{T}}_n(\widehat{\sigma}^2) & \mathbf{0} \\ \mathbf{0}^\top & L(\widehat{\sigma}^2) \end{bmatrix}$$

with

$$L(\sigma^2) = E \left(-\frac{\partial U^*(\sigma^2, \boldsymbol{X})}{\partial \sigma^2} \right) = \frac{1}{2} \sum_{i=1}^{n} 1/\tau_i^2;$$

and
3. Increment v by one and repeat the step 2. until convergence.

It is allowed to assess the asymptotic covariance of the estimators produced equating (9) and (10) to zero by using the sandwich estimator given by $\frac{1}{n}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\Lambda}_n\boldsymbol{\Gamma}_n^{-1}$, where

$$\boldsymbol{\Gamma}_n = \frac{1}{n}\boldsymbol{V}^* \left(\widehat{\boldsymbol{\theta}}, \boldsymbol{X} \right) \qquad \text{and} \qquad \boldsymbol{\Lambda}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i^*(\widehat{\boldsymbol{\theta}}, \boldsymbol{X}) \boldsymbol{U}_i^*(\widehat{\boldsymbol{\theta}}, \boldsymbol{X})^\top$$

with

$$\boldsymbol{U}_i^*(\widehat{\boldsymbol{\theta}}, \boldsymbol{X}) = \begin{pmatrix} \frac{1}{\widehat{\tau}_i}(\ddot{\boldsymbol{h}}_i - \ddot{\boldsymbol{H}}_i \widehat{\boldsymbol{\beta}}_p) \\ \frac{1}{2}\frac{\widehat{c}_i}{\widehat{\tau}_i^2} - \frac{1}{2\widehat{\tau}_i} \end{pmatrix}.$$

Hence, when an equation error is added to the model, the corrected score approach requires numerical (or iterative) procedures, which is not the case with the model considered in Zavala et al. (2007). The linear case, where

$$y_i | x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$

has to be treated as a subcase of the general polynomial model. That is, there is no analytical solution as in the case of the moments estimators (Kulathinal et al., 2002; Cheng and Riu, 2006) and the algorithm described above for the polynomial case may also be used for the linear situation. Estimates for the asymptotic covariance matrix can also be obtained from the general expression described above for the polynomial situation. To the best of our knowledge this approach is not in the literature.

Therefore, for testing $H_0 : \boldsymbol{G\theta} = \boldsymbol{d}$ it allows to use the following Wald statistic

$$\xi_n = n(\boldsymbol{G}\widehat{\boldsymbol{\theta}} - \boldsymbol{d})^\top \left(\boldsymbol{G}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\Lambda}_n\boldsymbol{\Gamma}_n^{-1}\boldsymbol{G}^\top\right)^{-1} (\boldsymbol{G}\widehat{\boldsymbol{\theta}} - \boldsymbol{d}), \tag{12}$$

which, under (7), has asymptotic chi-square distribution with rank-of-$\boldsymbol{G}$ degrees of freedom. Naturally, this convergence is only valid for testing values when $\boldsymbol{d} \in \mathcal{R}^p \times \mathcal{R}_+$ where $\mathcal{R}_+$ is the positive real set (excluding zero).

4 Simulation

This section presents the results of simulation studies in order to guide us regarding the behavior of the statistic (12) for small and moderate sample sizes. The asymptotic distribution of (12) can be used, however, as an approximation for testing when the sample size is small or moderate. To further study this issue, we designed a Monte Carlo study by generating 10 000 samples which were used to compute the empirical level and power of the statistics at the 5% nominal level. The simulation setting was taken to represent the real data set of the WHO MONICA project presented in the next section.

We carry out two types of simulation, namely: linear and quadratic relationship between the unobservable variables y_i and x_i. In both cases, we consider that $X_i|x_i \sim \mathcal{N}(x_i, \kappa_i)$ where $\sqrt{\kappa_i} \sim U(0.5, 1.5)$ and $Y_i|x_i \sim \mathcal{N}(y_i, \lambda_i)$ where $\sqrt{\lambda_i} \sim U(0.5, 4)$. The (unknown) values of x_i was generated from the normal distribution with mean $\mu_x = -2$ and variance $\sigma_x^2 = 4$. The approach presented in this paper is distribution free concerning x_i and the results, for the corrected score approach, are similar whatever be the values of x_i. As for the linear relationship, we consider (β_0, β_1) on a neighborhood of $(0, 1)$. For the quadratic relationship, we consider $\beta_0 = 0$ and (β_1, β_2) on a neighborhood of $(1, 0)$.

Table 1 presents the results for a linear relationship, which considers the following model: $y_i|x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, 10)$ and also depicts the results of a

Table 1: Rejection rates for the linear and quadratic models (at a 5% nominal level) using the Wald statistics (12) and for $n = 40$, $n = 80$ and $n = 100$.

	Linear relationship				*Quadratic relationship*		
β_0	β_1			β_2	β_1		
$n = 40$	0.6	1	1.4	$n = 40$	0.6	1	1.4
-2	0.5831	0.8261	0.9798	-0.15	0.5568	0.5402	0.9990
0	0.3469	**0.0993**	0.3171	0.00	0.8897	**0.1543**	0.9455
2	0.9862	0.8394	0.5486	0.15	0.9918	0.4964	0.6014
$n = 80$	0.6	1	1.4	$n = 80$	0.6	1	1.4
-2	0.8764	0.9866	0.9999	-0.15	0.2917	0.5591	0.9989
0	0.4386	**0.0721**	0.3986	0.00	0.8121	**0.0904**	0.8880
2	1.0000	0.9861	0.8443	0.15	0.9936	0.4999	0.3557
$n = 100$	0.6	1	1.4	$n = 100$	0.6	1	1.4
-2	0.8779	0.9940	0.9998	-0.15	0.3945	0.7667	0.9999
0	0.6361	**0.0667**	0.5870	0.00	0.9731	**0.0825**	0.9917
2	1.0000	0.9950	0.8508	0.15	1.0000	0.7300	0.4964

quadratic relationship that considers the following model: $y_i|x_i \sim \mathcal{N}(\beta_1 x_i + \beta_2 x_i^2, 10)$. It can be seen from Table 1 that the empirical nominal levels (middle cells in bold) get closer to the nominal level (5%) as n increases and for $n = 100$ the results are quite satisfactory.

5 Applications

5.1 Data set of the WHO MONICA project

The WHO MONICA project is a monitoring study of cardiovascular diseases, for more information go to http://www.ktl.fi/monica which provides a full description of the project. The data analyzed in this section are trends of the risk scores for women ($n = 36$) and for men ($n = 38$) in each population. According to Kulathinal et al. (2002), the risk score was defined as a linear combination of smoking status, systolic blood pressure, body mass index and total cholesterol level. Furthermore, a proportional hazards models was taken in order to derive its coefficients and the sampling errors of the trend estimates were considered as measurement errors. Therefore, it is possible to assess the variances in each experimental unit. Additional

information about data sets can be found in Kuulasmaa et al. (2000). The data set has been previously analyzed in the literature (Kulathinal et al., 2002; de Castro et al., 2007; Kuulasmaa et al., 2000), where a linear model has been considered. We consider now the possibility of fitting a quadratic model to this data set, that is,

$$y_i|x_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i + \beta_2 x_i^2, \sigma^2). \tag{13}$$

Table 2 shows the estimates (and the standard-errors) of the model parameters in (13). Figure 1 presents the scatter plot with a estimated quadratic and linear regressions for both men and women data sets.

Table 2: Estimates (standard-error) of the model parameters (13) using Monica data for men and women

	Men		Women	
	Linear	Quadratic	Linear	Quadratic
β_0	-2.0888 (0.4352)	-2.7183 (0.5095)	-0.0705 (0.8602)	-0.5811 (0.9123)
β_1	0.4705 (0.2381)	0.4857 (0.1901)	0.6434 (0.3376)	1.2133 (0.4050)
β_2	-	0.1278 (0.0477)	-	0.2047 (0.0975)
σ^2	4.8746 (1.4308)	4.4000 (1.6195)	11.1092 (5.0150)	10.0241 (5.0665)

5.2 Data set of the *Chandra* X-ray Observatory Center

The *Chandra* X-ray observatory is the NASA's flagship mission for X-ray astronomy. One of the most studied astronomical problem is to investigate how the "X-ray photon index" emission depends on the Eddington ratio of quasars (see Kelly et al., 2008, for details). There are many problems regarding the data collection such as sample selection and censoring, as discussed in Kelly (2007) and Akritas and Bershady (1996). The data set analyzed in this paper has no censored observations, however, it is subject to sample selection as reported in Kelly (2007). We modeled this data set disregarding the bias produced by the data collection just to show the applicability of our approach. We are engaged in future researches to take into account these sample peculiarities in a polynomial relationship relating the response variable (X-ray photon index) and the covariate (base-10 logarithm of the Eddington ratio of quasars).

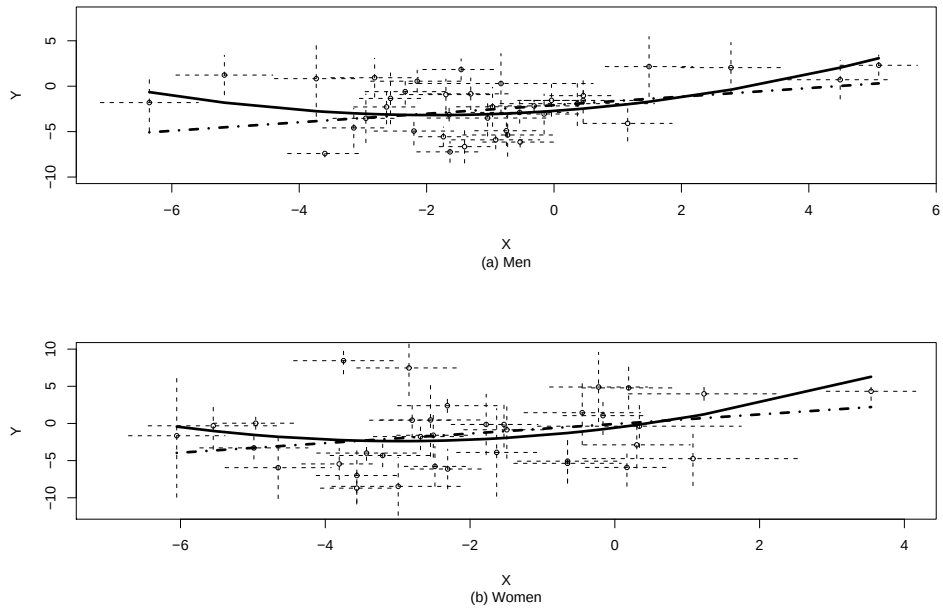

Figure 1: Quadratic (full line) and Linear (dash-dot line) regressions using MONICA data for men (a) and women (b)

The Eddington ratio of quasars is a function of the black hole mass, which is necessary to estimate. Therefore, the covariate is subject to error. In addition, it is allowed to assess the precision related to this measure in each experimental unit (defining heteroscedastic errors). The "X-ray photon index" and its precision was obtained from *Chandra* observatory measurements. The equation error (or intrinsic scatter in the astronomy jargon) is expected for this problem. For a full information, see Kelly et al. (2008).

Kelly et al. (2008) found that the relationship between the response variable, y_i, and the covariate, x_i, is not linear. Figure 2 shows the scatter plot (with the error bars) of the observed X-ray photon index and the observed base-10 logarithm of the Eddington ratio of quasars ($n = 153$), which suggests a quadratic relationship. However, the algorithm for a quadratic model diverges from this data set which indicates false relationship or larger measurement error in the surrogate variable (our simulation study shows the

larger the measurement error the greater the chance for non-convergence. Besides, we found that when a polynomial model of order p is true and a polynomial model of order $q < p$ is fitted, non-convergence might happen). Also, non-convergence occurs with the linear regression. Then, we consider a cubic model. The only configuration that presents statistic significance is considering that $\beta_0 = \beta_2 = 0$. That is, the model formulated for this data set is given by

$$y_i|x_i \sim \mathcal{N}(\beta_1 x_i + \beta_3 x_i^3, \sigma^2), \tag{14}$$

Table 3: Estimates (standard-error) of the model parameters (14) using *Chandra* data

	Estimates
β_1	-2.1202 (0.3944)
β_3	0.3415 (0.2156)
σ^2	0.1156 (0.0293)

Table 3 gives the estimates and their standard-error (in parenthesis) for the parameters of the model (14). Figure 2 presents the scatter plot with a estimated cubic regression for the *Chandra* data set.

6 Conclusions and remarks

We studied a heteroscedastic polynomial with measurement error model in both axes, allowing to model partial polynomial regressions. Furthermore, it is possible to test general linear hypothesis using a Wald statistic with an asymptotic (central) chisquare distribution. We also modeled a quadratic regression with measurement error in both axes to the real dataset of the WHO MONICA project and a cubic partial regression model to the *Chandra* observations showing the usefulness of our approach. We remark that the regressions fitted in this paper are valid only for the observed range of the covariate, extrapolations of it might not be reliable. Moreover, the model studied here can be used as an approximation for complex functions in order to fitting data sets more accurately than the linear regression.

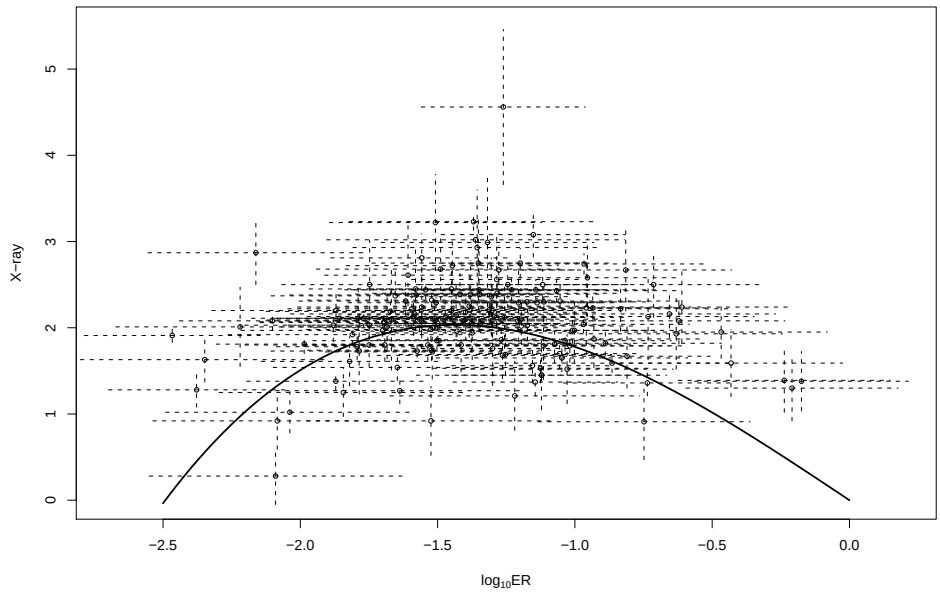

Figure 2: Cubic regression using *Chandra* observations.

acknowledgements

The authors thank Dr. Kari Kuulasmaa (National Public Health Institute, Finland) for gently supplying the data of our first example. The authors also thank Brandon C. Kelly for supplying the data of our second example and for his kindly disposition to explain it by email. This work was partially supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References

Akritas MG, Bershady MA. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal*. **470**:706–714.

Chan LK and Mak TK. (1985). On the polynomial function relationship. *Journal of the Royal Statistical Society Series B* **47**:510–518.

Cheng CL, Riu J. (2006). On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics*. **48**(4):511–519.

Cheng CL and Scheneeweis H. (1998). Polynomial regression with errors in variables. *Journal of the Royal Statistical Society Series B* **60**:189–199.

de Castro M, Galea M and Bolfarine H. (2007). Hypothesis testing in an errors-in-variables model with heteroscedastic measurement errors. (Submitted) *Statistics in Medicine*.

Fuller W. (1987). *Measurement Error Models*. Wiley: Chichester.

Gimenez P and Bolfarine H. (1997). Corrected score functions in classical error-in-variables and incidental parameter models. *Australian Journal of Statistics* **39**(3): 325–344.

Kelly BC. (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal*. **665**:1489–1506.

Kelly BC, Bechtold J, Trump JR, Vertergaard M. (2008). Observational constraints on the dependence of ratio-quiet quasar X-ray emission on black hole mass and accretion disk. (To appear) *The Astrophysical Journal*.

Kukush A., Schneeweiss H and Wolf R. (2005). Relative efficiency of three estimators in a polynomial regression with measurement errors *Journal of Statistical Planning and Inference* **127**:179–203.

Kulathinal SB, Kuulasmaa K and Gasbarra D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement error vary between the observations. *Statistics in Medicine*. **21**:1089–1101.

Kuulasmaa K, Tunstall-Pedoe H, Dobson A, Fortmann S, Tolonen H, Evans A, Ferrario M, Tuomilehto J for the WHO MONICA project. Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. Lancet 2000; 355:675 687.

Nakamura T. (1990). Corrected score functions for errors-in-variables models: methodology and applications to generalized linear models. *Biometrika* **77**:127–137.

Zavala AAZ, Bolfarine H and de Castro M. (2007). Consistent estimation and testing in heteroscedastic polynomial errors-in-variables models. *AISM* **59**:515–530.

Alexandre Galvão Patriota and Heleno Bolfarine
Institute of Mathematics and Statistics
University of São Paulo, Brazil
Rua do Matão, 1010 - Cidade Universitária
E-mail: patriota@ime.usp.br
hbolfar@ime.usp.br