Thamerus:

Different Nonlinear Regression Models with Incorrectly Observed Covariates

Sonderforschungsbereich 386, Paper 68 (1997)

Online unter: http://epub.ub.uni-muenchen.de/

Different Nonlinear Regression Models with Incorrectly Observed Covariates

Markus Thamerus, SFB 386
Institut für Statistik, Universität München

Abstract

We present quasi-likelihood models for different regression problems when one of the explanatory variables is measured with heteroscedastic error. In order to derive models for the observed data the conditional mean and variance functions of the regression models are only expressed through functions of the observable covariates. The latent covariable is treated as a random variable that follows a normal distribution. Furthermore it is assumed that enough additional information is provided to estimate the individual measurement error variances, e.g. through replicated measurements of the fallible predictor variable. The discussion includes the polynomial regression model as well as the probit and logit model for binary data, the Poisson model for count data and ordinal regression models.

Keywords: heteroscedastic measurement error, quasi-likelihood, polynomial regression, Poisson model, binary regression models, ordinal regression models

1 Introduction

It is a familiar situation for practical researchers that some of the predictors of a regression model cannot be observed correctly and instead are only measured with error. If this measurement error is not taken into account the estimators of the model parameters will be biased. This was shown by Stefanski (1985) in general for all regression models where the parameters of interest are estimated by an M-estimator, which is consistent in the absence of measurement error.

For all discussed models in this paper the response variables Y_i, $i = 1, \ldots, n$ are related to the explanatory variables $Z_i = (Z_{i1}, \ldots, Z_{ik})'$ and X_i by a nonlinear regression function. The continuous regressors X_i can only be observed by their incorrect measurements W_i. We assume that the true predictors X_i are related to the observed covariates W_i through $W_i = X_i + U_i$, where the measurement errors U_i, $i = 1, \ldots, n$ are independent stochastic variables with expectations zero and we do not restrict the error variances to be constant but allow for heteroscedasticity.

We will consider the structural case of errors-in-variables models and treat the latent regressors X_i as independent and identically distributed random variables.

The structural approach to regression models with covariate error consists of three main components:

a) the unobservable true regression model, that relates the response variables Y_i and the true regressors Z_i and X_i,
b) the error model, that characterizes the relationship between the latent regressors X_i and their measurements W_i and
c) the assumed marginal distribution of the X_i's.

For the main part of nonlinear regression models likelihood analysis depending on the associated distributions of all three parts a)–c) remains computationally difficult, since it requires numerical optimization routines to evaluate an integral in the likelihood function of the observed data. Carroll, Ruppert and Stefanski (1995) give an excellent overview of methods for treating measurement error in nonlinear regression models including likelihood models as well. For more details on the maximization of likelihood functions in errors-in-variables models see e.g. Crouch and Spiegelman (1990) or Liu and Pierce (1994). An indirect method to obtain maximum likelihood estimates of the regression parameters is to use an EM algorithm as it is proposed by Schafer (1993) for a probit regression model or by Schafer and Purdy (1996) for the linear regression model. Due to all the computational difficulties associated with likelihood analysis in the errors-in-variables problem we prefer an alternative method, that directly allows to model heteroscedastic measurement errors as well.

We will use quasi-likelihood models, see e.g. McCullagh (1991) for an introduction, that solely base on the first and second conditional moments of the response Y_i given the known explanatory variables Z_i and the observed measurements W_i. If the mean and variance function of the model in the observable variables can be specified, estimation is carried out by the usual iteratively reweighted least square algorithm for such models which is easier to implement than the numerical integration methods for the likelihood analysis. The unbiasedness of the quasi-score function guarantees the consistency and asymptotic normality of the parameter estimates. One of the first to use quasi-likelihood methods to analyze errors-in-variables models is Armstrong (1985). The subject is also considered by Liang and Liu (1991).

In this article we will discuss the quasi-likelihood approach for different nonlinear regression models with incorrectly observed covariates under the assumption of heteroscedastic measurement errors. We will state a structural approach for the polynomial regression model and the Poisson regression model. The use of quasi-likelihood

methods for binary regression models in the case of nonconstant measurement error variances is discussed and the idea is extended to the multicategorial case when ordinal response variables are observed. In the next section we discuss quasi-likelihood models for the errors-in-variables problem in general and state a heteroscedastic error model. Section three reviews the use of quasi-likelihood methods for the different nonlinear regression models with heteroscedastic measurement errors in one of the covariates and it is shown how a model in the observable variables is derived when the latent regressors follow a normal distribution.

2 Model of the Observed Data

The fundamental idea of using quasi-likelihood methods to analyse regression models with incorrectly observed covariates is to make a transition from the unobservable model, formulated in terms of the latent variables, to a model of the observable data. The unobservable model of interest is given by a nonlinear regression model, where the vector of the regression parameters β is estimated by solving an unbiased estimating equation. In the case of no measurement error estimation is based on the mean and variance function of the observed data, i.e., the conditional mean and variance of Y_i as a function of Z_i and X_i. These are given by

$$\mu\left(Z_i, X_i; \beta\right) = \mathrm{E}\left(Y_i \mid Z_i, X_i\right) \quad \text{for the mean and} \tag{1}$$

$$\sigma^2\left(Z_i, X_i; \beta, \nu\right) = \mathrm{V}(Y_i \mid Z_i, X_i) \quad \text{for the variance function,} \tag{2}$$

where ν denotes an additional (optional) variance parameter. This includes the class of generalized linear models where the conditional distribution of Y_i given Z_i and X_i belongs to the exponential family and the mean function (1) is given by

$$\mu(Z_i, X_i; \beta) = g(\beta_0 + Z_i'\beta_Z + \beta_X X_i) \quad \text{with} \quad \beta = (\beta_0, \beta_Z', \beta_X)$$

and where $g^{-1}(\cdot)$ is the link function of the model. To derive a model in the observable variables we denote the first two moments of the conditional distribution of Y_i given Z_i and W_i with

$$m_Q(Z_i, W_i; \beta) = \mathrm{E}\left(Y_i \mid Z_i, W_i\right) \quad \text{for the mean and}$$

$$v_Q(Z_i, W_i; \beta, \nu) = \mathrm{V}(Y_i \mid Z_i, W_i) \quad \text{for the variance function.}$$

If those two functions can be specified, the estimation of β can be carried out by the usual iteratively reweighted least square algorithm for quasi-likelihood models (see

e.g. McCullagh, 1991). The quasi-score function

$$\mathrm{s}\left(\beta\right)=\sum_{i=1}^{n}\frac{\partial\, m_Q\left(Z_i,W_i;\beta\right)}{\partial\,\beta}v_Q^{-1}(Z_i,W_i;\beta,\nu)(Y_i-m_Q(Z_i,W_i;\beta))=\sum_{i=1}^{n}\mathrm{s}_i(\beta)$$

provides an unbiased estimating equation for β and the quasi-likelihood estimator $\hat{\beta}_{q_l}$ is found as the root of the equation $s\left(\beta\right)=0$. This provides the asymptotic properties of $\hat{\beta}_{ql}$ and it holds that $\hat{\beta}_{ql}$ is a consistent estimator of β and that

$$\hat{\beta}_{ql}\overset{a}{\sim}N(\beta,n^{-1}F^{-1}(\beta)V(\beta)F^{-1}(\beta)).$$

The covariance matrix of $\hat{\beta}_{ql}$ is of the 'sandwich' form and its parts are the inverse of the expected quasi-information matrix and the estimated covariance matrix of the score function. It is estimated empirically by

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_{ql}\right)=n^{-1}\hat{F}^{-1}(\hat{\beta}_{ql})\hat{V}(\hat{\beta}_{ql})\hat{F}^{-1}(\hat{\beta}_{ql})$$

with its components given through

$$\hat{F}(\hat{\beta}_{ql})=\frac{1}{n}\left(\sum_{i=1}^{n}-\frac{\partial\ s_i(\beta)}{\partial\,\beta'}\bigg|_{\beta=\hat{\beta}_{ql}}\right)\quad\text{and}\quad\hat{V}(\hat{\beta}_{ql})=\frac{1}{n}\left(\sum_{i=1}^{n}s_i(\beta)\ \left(s_i(\beta)\right)'\big|_{\beta=\hat{\beta}_{ql}}\right).$$

It is supposed that the measurement error is nondifferential, which is defined as the conditional independence of Y_i and W_i given X_i and Z_i. For most measurement problems this assumption is reasonable since it implies that no additional information about the response Y_i is provided by the measurement W_i if the true explanatory variables Z_i and X_i are observed. In the case of nondifferential measurement error the fallible predictor W_i is called a surrogate. With this assumption the construction principle for the functions m_Q and v_Q can be demonstrated. For all models in the next section we have to compute the expectations

$$\begin{aligned}
m_Q\left(Z_i,W_i;\beta\right) &= \mathrm{E}\left(\mathrm{E}\left(Y_i\mid Z_i,X_i,W_i\right)\mid Z_i,W_i\right)=\mathrm{E}\left(\mathrm{E}\left(Y_i\mid Z_i,X_i\right)\mid Z_i,W_i\right)\\
&= \mathrm{E}\left(\mu\left(Z_i,X_i;\beta\right)\mid Z_i,W_i\right)\quad\text{and} \qquad (3)\\
v_Q\left(Z_i,W_i;\beta,\nu\right) &= \mathrm{V}(\mathrm{E}\left(Y_i\mid Z_i,X_i,W_i\right)\mid Z_i,W_i)+\mathrm{E}\left(\mathrm{V}(Y_i\mid Z_i,X_i,W_i)\mid Z_i,W_i\right)\\
&= \mathrm{V}(\mathrm{E}\left(Y_i\mid Z_i,X_i\right)\mid Z_i,W_i)+\mathrm{E}\left(\mathrm{V}(Y_i\mid Z_i,X_i)\mid Z_i,W_i\right)\\
&= \mathrm{V}(\mu\left(Z_i,X_i;\beta\right)\mid Z_i,W_i)+\mathrm{E}\left(\sigma^2\left(Z_i,X_i;\beta,\nu\right)\mid Z_i,W_i\right). \qquad (4)
\end{aligned}$$

After we have set up the framework to estimate regression models with surrogate predictors we now regard the relationship between the latent regressors X_i and their

measurements W_i. For all discussed models in section three we assume an additive heteroscedastic measurement error model:

$$\begin{aligned} &\text{For } \; i = 1, \ldots, n \;\; \text{it holds that } \;\; W_i = X_i + U_i \text{ with } U_i \sim N(0, \sigma_i^2) \\ &\text{where } \;\; \mathrm{Cov}\,(U_i, U_j) = 0 \text{ for } i \neq j,\; j = 1, \ldots, n \;\; \text{and} \\ &\text{the errors } U_i \text{ are independent from the variables } Y_i,\; X_i \text{ and } Z_i. \end{aligned} \tag{5}$$

This implies that the measurement errors are nondifferential. For some applications the assumption of a heteroscedastic measurement error model is more reasonable than to assume constant error variances. Thamerus (1997) describes an example where the true regional concentration of radon X_i is approximated by the average of n_i single measurements W_{ij} of X_i within one region. For those individual measurements of X_i, $i = 1, \ldots, n$ the error model

$$\begin{aligned} W_{ij} = X_i + \varepsilon_{ij} \qquad &\text{with } \; \mathrm{E}\,(\varepsilon_{ij}) = 0 \;\; \text{and } \; \mathrm{Var}\,(\varepsilon_{ij}) = \sigma_{\varepsilon_i}^2 \;\; \text{for } j = 1, \ldots, n_i \\ &\text{and } \mathrm{Cov}\,(\varepsilon_{ij}, \varepsilon_{il}) = 0 \;\; \text{for } j \neq l,\; l = 1, \ldots, n_i \end{aligned}$$

was assumed. The observed averages $W_i = n_i^{-1} \sum_{j=1}^{n_i} W_{ij}$ therefore follow the additive heteroscedastic error model

$$W_i = X_i + U_i \;\; \text{ with } \;\; U_i \sim N(0, \sigma_i^2) \;\; \text{ for } \;\; i = 1, \ldots, n.$$

The heteroscedastic error variances $\sigma_i^2 = {n_i}^{-1}\sigma_{\epsilon_i}^2$ can be estimated with the help of the sample variances $s_{W_i}^2$ of the measurements within each region by $\hat{\sigma}_i^2 = {n_i}^{-1} s_{W_i}^2$. In general the analysis of errors-in-variables models is nonpractical without additional information on the measurement error process. For our discussion we will assume that the heteroscedastic error variances σ_i^2 are known or that enough information e.g. replicated measurements of the fallible predictor variable as in the example above is provided that the individual variances σ_i^2 at least can be estimated consistently.

In a structural errors-in-variables model the latent regressors X_i are treated as independently and identically distributed random variables and an assumption has to be made about the distribution of the true covariates X_i. This in fact is a crucial point of the analysis and requires careful examination of what is known about this distribution through the observed sample of the W_i's. The normal distribution is often used in the literature. If the distribution of the true covariates is skewed one choice is to assume that the true covariates X_i are lognormally distributed and often the transformation $\log(X_i)$ along with a multiplicative error model is used in the

analysis. For more details on transformations of the X-variables see the recent paper of Eckert, Carroll and Wang (1997). More complex situations lead to the assumption of a mixture of normal distributions, see e.g. Küchenhoff and Carroll (1997), which is computationally more demanding but yields similar convenient properties as the assumption of a normal distribution. Thamerus (1997) used a normal mixture distribution together with an heteroscedastic measurement error model in a Poisson regression model. For all discussed models in section three we assume that the true variables X_i are independently and identically distributed normal variables with expectation μ_X and variance σ_X^2. Under the assumption of an additive heteroscedastic measurement error model unbiased estimators of these parameters are given by $\hat{\mu}_X = \bar{w}$ and $\hat{\sigma}_X^2 = s_W^2 - \frac{n-1}{n^2}\sum_{i=1}^n \sigma_i^2$ where $\bar{w}$ is the sample mean and s_W^2 is the sample variance of the observed measurements W_i, $i = 1, \ldots, n$.

In order to state the mean and variance functions of the different quasi-likelihood models it is necessary to specify the conditional distributions of W_i given the true covariates Z_i and X_i. If we furthermore assume that X_i is independent of the other correctly observed covariates Z_i we find for the conditional distributions of X_i given the surrogate W_i for $i = 1, \ldots, n$:

$$\begin{aligned} X_i \mid Z_i, W_i \sim X_i \mid W_i &\sim N(\mu_i, \tau_i^2) \text{ with} \\ \mu_i &= \mu_X + \sigma_X^2(\sigma_X^2 + \sigma_i^2)^{-1}(W_i - \mu_X) \text{ and} \\ \tau_i^2 &= \sigma_X^2\left(1 - \sigma_X^2(\sigma_X^2 + \sigma_i^2)^{-1}\right). \end{aligned} \tag{6}$$

Note, that the variances τ_i^2 of this conditional distributions differ between individuals as a consequence of the heteroscedastic error variances σ_i^2. At the end of this section we summarize all the assumptions under which the functions m_Q and v_Q, given in (3) and (4), for the different models in the next section will be derived.

Assumptions

(A1) The variables X_i and W_i, $i = 1, \ldots, n$ are related by an additive heteroscedastic measurement error model as defined in (5).

(A2) For the latent variables it holds: $X_i \overset{i.i.d}{\sim} N(\mu_X, \sigma_X^2)$, $i = 1, \ldots, n$.

(A3) The variables X_i and Z_i, $i = 1, \ldots, n$ are independent.

3 Application to Different Regression Models

Polynomial regression models

In a forthcoming paper Cheng and Schneeweiß (1998) develop a functional errors-in-variables model for the polynomial regression model by correcting the scorefunction of the model to adjust for measurement error in the observed covariates. Moon and Gunst (1995) give a summary of the work on polynomial regression models with covariate errors. In contrast to these two papers and the work cited therein we allow for heteroscedastic measurement error and show how under the assumptions (A1)–(A3) a structural model is accomplished.

In the notation of (1) and (2) the polynomial regression model without additional variables Z_i is given by

$$\mu\left(X_i;\beta\right) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k \quad \text{and} \quad \sigma^2\left(X_i;\beta,\nu\right) = \sigma_\varepsilon^2 \tag{7}$$

with $\beta = (\beta_0,\ldots,\beta_k)'$ and $\nu = \sigma_\varepsilon^2$. The model of the observable data depends on higher moments of the conditional distributions of W_i given X_i. The k-th moment of that distribution will be denoted by

$$\mu'_{k,i} = \mathrm{E}\left(X_i^k \mid W_i\right) \quad \text{for} \quad k = 1,2,\ldots \quad \text{with} \quad \mu'_{0,i} = 1.$$

If the two functions of the unobservable model (7) are inserted into the general formulas (3) and (4) for the mean and variance function m_Q and v_Q we find

$$\begin{aligned}
m_Q\left(W_i;\beta\right) &= \mathrm{E}\left(\mu\left(X_i;\beta\right) \mid W_i\right) \\
&= \beta_0 + \beta_1 \mathrm{E}\left(X_i \mid W_i\right) + \beta_2 \mathrm{E}\left(X_i^2 \mid W_i\right) + \ldots + \beta_k \mathrm{E}\left(X_i^k \mid W_i\right) \\
&= \beta_0 + \beta_1 \mu'_{1,i} + \beta_2 \mu'_{2,i} + \ldots + \beta_k \mu'_{k,i} \quad \text{and} \\
v_Q\left(W_i;\beta,\nu\right) &= \mathrm{E}\left(\sigma^2\left(X_i;\beta,\nu\right) \mid W_i\right) + \mathrm{V}(\mu\left(X_i;\beta\right) \mid W_i) \\
&= \sigma_\varepsilon^2 + \mathrm{V}(\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X^k \mid W_i) \\
&= \sigma_\varepsilon^2 + \sum_{j=1}^{k} \beta_j^2 \,\mathrm{V}(X_i^j \mid W_i) + 2\sum_{l=1}^{k-1}\sum_{m=l+1}^{k} \beta_l \beta_m \,\mathrm{Cov}\left(X_i^l, X_i^m \mid W_i\right) \\
&= \sigma_\varepsilon^2 + \sum_{j=1}^{k} \beta_j^2 \left(\mathrm{E}\left(X_i^{2j} \mid W_i\right) - \left(\mathrm{E}\left(X_i^j \mid W_i\right)\right)^2\right) \\
&\quad + 2\sum_{l=1}^{k-1}\sum_{m=l+1}^{k} \beta_l \beta_m \left(\mathrm{E}\left(X_i^{l+m} \mid W_i\right) - \mathrm{E}\left(X_i^l \mid W_i\right)\mathrm{E}\left(X_i^m \mid W_i\right)\right) \\
&= \sigma_\varepsilon^2 + \sum_{j=1}^{k} \beta_j^2 \left(\mu'_{2j,i} - \left(\mu'_{j,i}\right)^2\right) + 2\sum_{l=1}^{k-1}\sum_{m=l+1}^{k} \beta_l \beta_m \left(\mu'_{l+m,i} - \mu'_{l,i}\mu'_{m,i}\right).
\end{aligned}$$

The mean function m_Q of the quasi-likelihood model is a linear function of the first k moments of the conditional distributions of W_i given X_i. The variance function m_Q uses moments up to the order $2k$. All those moments can be computed under the normal assumption (A2) for the latent variables X_i since the conditional distributions for X_i given W_i, defined in (6), are normal as well with parameters μ_i and τ_i^2. The k-th central moments of that distributions will be denoted by

$$\mu_{k,i} = \mathrm{E}\left((X_i - \mu'_{1,i})^k \mid W_i\right), \quad \text{for} \quad k = 1, 2, \ldots \quad \text{with} \quad \mu_{0,i} = 1.$$

Normality yields the calculation of the moments $\mu_{k,i}$ in dependence of the variances τ_i^2. For $r = 0, 1, 2, \ldots$ it holds that

$$\mu_{k,i} = \begin{cases} 0 & \text{for} \quad k = 2r+1, \\ \tau_i^k (k-1)(k-3) \cdot \ldots \cdot 3 \cdot 1 & \text{for} \quad k = 2r. \end{cases} \tag{8}$$

The different moments are connected by the Binomial theorem. In general we have

$$\mu'_{k,i} = \sum_{j=0}^{k} \binom{k}{j} \mu_{k-j,i} (\mu'_{1,i})^j \quad \text{for} \quad k = 1, 2, \ldots \,. \tag{9}$$

With the help of the results (8) and (9) all necessary moments to compute the functions m_Q and v_Q can be determined under the knowledge of the means $\mu'_{1,i} = \mu_i$ and variances $\mu_{2,i} = \tau_i^2$. We will demonstrate this for the quadratic regression model ($k = 2$) that is given by

$$\mu\left(X_i; \beta\right) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \quad \text{and} \quad \sigma^2\left(X_i; \beta, \nu\right) = \sigma_\varepsilon^2.$$

Mean and variance function of the model of the observed data are found as

$$\begin{aligned} m_B\left(W_i; \beta\right) &= \beta_0 + \beta_1 \mu'_{1,i} + \beta_2 \mu'_{2,i} = \beta_0 + \beta_1 \mu_i + \beta_2 (\tau_i^2 + \mu_i^2) \text{ and} \qquad (10) \\ v_B\left(W_i; \beta, \nu\right) &= \sigma_\varepsilon^2 + \beta_1^2 \left(\mu'_{2,i} - (\mu'_{1,i})^2\right) + \beta_2^2 \left(\mu'_{4,i} - (\mu'_{2,i})^2\right) \\ &\quad + 2\,\beta_1 \beta_2 \left(\mu'_{3,i} - \mu'_{1,i} \mu'_{2,i}\right) \\ &= \sigma_\varepsilon^2 + \beta_1^2 \left(\tau_i^2 + \mu_i^2 - \mu_i^2\right) \\ &\quad + \beta_2^2 \left(3\tau_i^4 + 6\tau_i^2 \mu_i^2 + \mu_i^4 - \tau_i^4 - 2\tau_i^2 \mu_i^2 - \mu_i^4\right) \\ &\quad + 2\,\beta_1 \beta_2 \left(3\tau_i^2 \mu_i + \mu_i^3 - \mu_i \tau_i^2 - \mu_i^3\right) \\ &= \sigma_\varepsilon^2 + \beta_1^2 \tau_i^2 + 2\beta_2^2 \left(\tau_i^4 + 2\tau_i^2 \mu_i^2\right) + 4\beta_1 \beta_2 \tau_i^2 \mu_i. \qquad (11) \end{aligned}$$

A particular property of the polynomial regression model with incorrectly observed covariates is that the measurement error itself is raised to the power of k. The

regressors of the naive approach $W_i^j = (X_i + U_i)^j$, $j = 1, \ldots, k$ are replaced in the quasi-likelihood model by the moments $\mu'_{j,i}$, $j = 1, \ldots, k$ and additionally the different variance structure of the model that is caused by the measurement errors is taken into account. This can already be seen in the example of the quadratic regression model. The variance σ_ε^2 that occurs in the variance function v_Q has to be updated in every iteration step of the estimation procedure. This can be done by the residual variance which in the $r+1$-th step can be computed with the help of the estimator $\hat{\beta}_1^{(r)}$ from the previous step as

$$(\hat{\sigma}_\varepsilon^2)^{(r+1)} = (n-k-1)^{-1} \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} \mu'_{1,i} + \hat{\beta}_2^{(r)} \mu'_{2,i} + \ldots + \hat{\beta}_k^{(r)} \mu'_{k,i}) \right)^2 .$$

Poisson regression model

Quasi-likelihood methods in Poisson regression models were mostly used to account for overdispersion, see e.g. Breslow (1990). Additional variation in the data can also be caused by measurement error in the covariates. Armstrong (1985) derives a model in the observable variables for the Poisson regression model when the erroneous predictor variables follow a normal distribution. This approach was adopted by Thamerus (1997) and modified for a mixture of normal distributions combined with a heteroscedastic measurement error model. We will demonstrate how a model under the assumptions (A1)–(A3) can be derived.

The mean and variance function (1) and (2) of the unobservable Poisson regression model are identical and given by

$$\mu\left(Z_i, X_i; \beta\right) = \sigma^2\left(Z_i, X_i; \beta\right) = \exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X X_i\right). \tag{12}$$

To derive the model of the observable data we insert the mean and variance function given in (12) into the equations (3) and (4). Under the assumptions (A1) and (A3) we at first obtain for the mean and variance function m_Q and v_Q the expressions

$$\begin{aligned} m_Q\left(Z_i, W_i; \beta\right) &= \mathrm{E}\left(\exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X X_i\right) \mid Z_i, W_i\right) \\ &= \exp\left(\beta_0 + Z_i^{'}\beta_Z\right) \mathrm{E}\left(\exp\left(\beta_X X_i\right) \mid W_i\right) \quad \text{and} \qquad (13) \\ v_Q\left(Z_i, W_i; \beta\right) &= \mathrm{V}(\exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X X_i\right) \mid Z_i, W_i) \\ &\quad + \mathrm{E}\left(\exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X X_i\right) \mid Z_i, W_i\right) \\ &= \exp\left(2\beta_0 + 2Z_i^{'}\beta_Z\right) \mathrm{E}\left(\exp\left(2\beta_X X_i\right) \mid W_i\right) \\ &\quad - \left(\exp\left(\beta_0 + Z_i^{'}\beta_Z\right) \mathrm{E}\left(\exp\left(\beta_X X_i\right) \mid W_i\right)\right)^2 \\ &\quad + \exp\left(\beta_0 + Z_i^{'}\beta_Z\right) \mathrm{E}\left(\exp\left(\beta_X X_i\right) \mid W_i\right). \qquad (14) \end{aligned}$$

As can be seen from (13) and (14) the essential task is to compute expectations of the form $\mathrm{E}\left(\exp\left(c\,X_i\right) \mid W_i\right)$, where the constant factor c has to be replaced by β_X or $2\beta_X$. As a result of the further assumption (A2) the conditional distributions of X_i given W_i are normal with the associated parameters μ_i and τ_i^2, see (6), and hence all expectations can be expressed as

$$\mathrm{E}\left(\exp\left(c\,X_i\right) \mid W_i\right) = \exp\left(c\,\mu_i + \frac{c^2\,\tau_i^2}{2}\right).$$

With this result all expectations in (13) and (14) can be calculated and we finally get the mean and variance function of the observable model as

$$\begin{aligned}
m_Q\left(Z_i, W_i; \beta\right) &= \exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X\mu_i + \frac{\beta_X^2\tau_i^2}{2}\right) \quad \text{and} \\
v_Q\left(Z_i, W_i; \beta\right) &= \exp\left(2\beta_0 + 2Z_i^{'}\beta_Z + 2\beta_X\mu_i + 2\beta_X^2\tau_i^2\right) \\
&\quad - \left(\exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X\mu_i + \frac{\beta_X^2\tau_i^2}{2}\right)\right)^2 \\
&\quad + \exp\left(\beta_0 + Z_i^{'}\beta_Z + \beta_X\mu_i + \frac{\beta_X^2\tau_i^2}{2}\right) \\
&= \exp\left(2\beta_0 + 2Z_i^{'}\beta_Z + 2\beta_X\mu_i + 2\beta_X^2\tau_i^2\right) \\
&\quad - \left(m_B\left(Z_i, W_i; \beta\right)\right)^2 + m_B\left(Z_i, W_i; \beta\right).
\end{aligned}$$

Binary regression models

The most popular method to treat covariate measurement error in a logistic regression model is the regression calibration approach. It was initiated by Rosner, Willett an Spiegelman (1989) and generalized for any regression model by Carroll and Stefanski (1990). Carroll et al. (1995) give a detailed description of the different approaches, structural and functional, to model measurement errors in binary regression models. Quasi-likelihood methods for such models have also been studied by Liang and Liu (1991) assuming homoscedastic measurement errors.

We will extend this idea to the heteroscedastic case and consider the probit and logit model for binary responses. For the conditional distribution of the response variables given the true covariates it holds that $Y_i \mid Z_i, X_i \sim B\left(1, \pi_i(\beta)\right)$.

Probit regression: Mean and variance function (1) and (2) of the unobserved probit regression model are given by

$$\mu\left(Z_i, X_i; \beta\right) = \Phi(\beta_0 + Z_i'\beta_Z + \beta_X X_i) = \pi_i(\beta) \quad \text{and} \tag{15}$$

$$\sigma^2\left(Z_i, X_i; \beta\right) \;=\; \pi_i(\beta)(1-\pi_i(\beta)). \tag{16}$$

With $\phi(\cdot)$ and $\Phi(\cdot)$ we denote the density and distribution function of the standard normal distribution. For establishing the quasilikelihood model under the assumptions (A1)–(A3) we make use of a probit integral argument, see e.g. Tosteson, Schafer and Stefanski (1989): For the fixed quantities $s > 0$ and m it holds in general that

$$\int_{-\infty}^{+\infty} s^{-1}\Phi(a+bx)\phi\left(\frac{x-m}{s}\right) dx = \Phi\left(\frac{a+b\,m}{(1+b^2s^2)^{\frac{1}{2}}}\right). \tag{17}$$

If we insert the functions (15) and (16) into the general equations (3) and (4) of the quasi-likelihood model, the relation (17) enables us to determine the functions m_Q and v_Q. For the mean function m_Q of the observable data we find

$$\begin{aligned} m_Q\left(Z_i, W_i; \beta\right) &= \mathrm{E}\left(\Phi(\beta_0 + Z_i'\beta_Z + \beta_X X_i) \mid Z_i, W_i\right) \\ &= \int_{-\infty}^{+\infty} \Phi(\beta_0 + Z_i'\beta_Z + \beta_X x_i)\, f_{X_i|W_i}(x_i)\, dx_i \\ &= \int_{-\infty}^{+\infty} \frac{1}{\tau_i}\Phi(\beta_0 + Z_i'\beta_Z + \beta_X x_i)\, \phi\left(\frac{x_i-\mu_i}{\tau_i}\right) dx_i \\ &= \Phi\left(\frac{\beta_0 + Z_i'\beta_Z + \beta_X\mu_i}{(1+\beta_X^2\tau_i^2)^{\frac{1}{2}}}\right) = \pi_i^*(\beta). \end{aligned}$$

The variance function of that model simply results in $v_B\left(Z_i, W_i; \beta\right) = \pi_i^*(\beta)(1-\pi_i^*(\beta))$. Hence we find that under the assumptions (A1)–(A3) the model in the observable data again is a probit regression model. For the conditional distribution of Y_i given the observable regressors Z_i and W_i we find that $Y_i \mid Z_i, W_i \sim B\left(1, \pi_i^*(\beta)\right)$ with the probabilities $\pi_i^*(\beta)$ given in (18). The functions m_Q and v_Q therefore provide a likelihood model for the probit regression with heteroscedastic measurement error.

Logistic regression: In order to state the logistic regression model we will denote the logistic distribution function with $H(t) = (1+\exp(-t))^{-1}$ and therefore can write the mean function (1) of the true model in this case as

$$\mu\left(Z_i, X_i; \beta\right) = H(\beta_0 + Z_i'\beta_Z + \beta_X X_i) = \pi_i(\beta). \tag{18}$$

To derive the mean function m_Q under the assumptions (A1)–(A3) we insert (18) into (3) and with the conditional distribution given in (6) we end up with the integral

$$\begin{aligned} m_Q\left(Z_i, W_i; \beta\right) &= \mathrm{E}\left(H(\beta_0 + Z_i'\beta_Z + \beta_X X_i) \mid Z_i, W_i\right) \\ &= \int_{-\infty}^{+\infty} \frac{1}{\tau_i} H(\beta_0 + Z_i'\beta_Z + \beta_X x_i)\phi\left(\frac{x_i-\mu_i}{\tau_i}\right) dx_i\,, \end{aligned}$$

which can not be expressed in closed form. Alternatively to numerical integration methods the function m_Q can be sufficiently approximated by exploiting the relation $H(t) \approx \Phi(\frac{t}{c})$ with $c = \frac{15\pi}{16\sqrt{3}}$, see Johnson and Kotz (1970, Ch. 22). With the help of the relation (17) this leads to an approximation of the mean function m_Q for the logistic model by a 'scaled' probit model

$$\begin{aligned} m_Q\left(Z_i, W_i; \beta\right) &\approx \int_{-\infty}^{+\infty} \frac{1}{\tau_i} \Phi\left(\frac{1}{c}(\beta_0 + Z_i'\beta_Z + \beta_X x_i)\right) \phi\left(\frac{x_i - \mu_i}{\tau_i}\right) dx_i \\ &= \Phi\left(\frac{1}{c}\left(\frac{\beta_0 + Z_i'\beta_Z + \beta_X \mu_i}{(1 + c^{-2}\beta_X^2 \tau_i^2)^{\frac{1}{2}}}\right)\right) =: \pi_i^a. \end{aligned}$$

The variance function v_Q of the observed model is thus approximated by $v_Q\left(Z_i, W_i; \beta\right) \approx \pi_i^a(\beta)(1 - \pi_i^a(\beta))$. Just like in the probit case but merely approximately we find for the conditional distribution $Y_i \mid Z_i, W_i \overset{a}{\sim} B\left(1, \pi_i^a(\beta)\right)$.

Ordinal regression models

Ordinal regression models with measurement error in the covariates were examined by Tosteson et al. (1989) who supposed to estimate the model parameters by adjusting the estimator that is obtained if the regression calibration method is used. We will restrict our discussion on the cumulative probit model for ordinal data and show that the results of the probit model for binary data under the assumptions (A1)–(A3) can easily be transferred to the multicategorial case.

With $Y_i^* \in \{1, \ldots, p\}$ we will denote the observed ordinal response variables that are modeled by the vector of dummy variables $Y_i = (Y_{i1}, \ldots, Y_{iq})'$ and it holds that $Y_{ir} = 1 \Leftrightarrow Y_i^* = r$ for $r = 1, \ldots, q = p - 1$. The observed variables Y_i^* are categorized versions of the latent response variables ξ_i where

$$Y_i^* = r \Leftrightarrow \vartheta_{r-1} < \xi_i \leq \vartheta_r, \; r = 1, \ldots, p \;\text{ with }\; -\infty = \vartheta_0 < \vartheta_1 < \ldots < \vartheta_p = \infty.$$

The latent response variables ξ_i are related to the true covariates Z_i and X_i by

$$\xi_i = -\left(Z_i'\beta_Z + \beta_X X_i\right) + \varepsilon_i \;\text{ with }\; \varepsilon_i \sim N(0, 1). \tag{19}$$

The vectors Y_i given the true covariates Z_i and X_i are multinominal distributed with $Y_i \mid Z_i, X_i \sim M(1, \Pi_i(\vartheta, \beta))$. The q-dimensional mean function (1) of the cumulative probit model is given by

$$\begin{aligned} &\mu\left(Z_i, X_i; \vartheta, \beta\right) = \Pi_i(\vartheta, \beta) = \left(\pi_{i1}(\vartheta_1, \vartheta_0, \beta), \ldots, \pi_{iq}(\vartheta_q, \vartheta_{q-1}, \beta)\right)' \qquad (20) \\ &\text{with the elements } \pi_{ir}(\vartheta_r, \vartheta_{r-1}, \beta) \text{ for } r = 1, \ldots, q \text{ given by} \\ &\pi_{ir}(\vartheta_r, \vartheta_{r-1}, \beta) = \Phi(\vartheta_r + Z_i'\beta_Z + \beta_X X_i) - \Phi(\vartheta_{r-1} + Z_i'\beta_Z + \beta_X X_i). \end{aligned}$$

The model parameters are denoted by $\vartheta = (\vartheta_1, \ldots, \vartheta_q)'$ and $\beta = (\beta_Z', \beta_X)'$. For the variance function (2) of the unobserved regression model we have

$$\sigma^2 (Z_i, X_i; \vartheta, \beta) = \operatorname{diag} \Pi_i(\vartheta, \beta) - \Pi_i(\vartheta, \beta)(\Pi_i(\vartheta, \beta))'.$$

To derive a model in the observable variables under the assumptions (A1)–(A3) we have to compute the conditional expectation (3) of the mean function (20), that is

$$m_Q (Z_i, W_i; \vartheta, \beta) = \mathrm{E} \left(\Pi_i(\vartheta, \beta) \mid Z_i, W_i \right).$$

The same arguments that were used to find the mean function (18) in the univariate case can be applied to every element of the vector m_Q. This leads to the result

$$\begin{aligned} & m_B (Z_i, W_i; \vartheta, \beta) = \Pi_i^*(\vartheta, \beta) = \left(\pi_{i1}^*(\vartheta_1, \vartheta_0, \beta), \ldots, \pi_{iq}^*(\vartheta_q, \vartheta_{q-1}, \beta) \right)' \qquad (21) \\ & \text{with the elements } \pi_{ir}^*(\vartheta_r, \vartheta_{r-1}, \beta) \text{ for } r = 1, \ldots, q \text{ given by} \\ & \pi_{ir}^*(\vartheta_r, \vartheta_{r-1}, \beta) = \Phi \left(\frac{\vartheta_r + Z_i' \beta_Z + \beta_X \mu_i}{(1 + \beta_X^2 \tau_i^2)^{\frac{1}{2}}} \right) - \Phi \left(\frac{\vartheta_{r-1} + Z_i' \beta_Z + \beta_X \mu_i}{(1 + \beta_X^2 \tau_i^2)^{\frac{1}{2}}} \right). \end{aligned}$$

The variance function (4) of that model is just given by

$$v_B (Z_i, W_i; \vartheta, \beta) \;\; = \;\; \operatorname{diag} \Pi_i^*(\vartheta, \beta) - \Pi_i^*(\vartheta, \beta)(\Pi_i^*(\vartheta, \beta))'.$$

The vectors Y_i given Z_i and W_i follow a multinomial distribution and it holds that $Y_i \mid Z_i, W_i \sim M(1, \Pi_i^*(\vartheta, \beta))$ with the probability vector $\Pi_i^*(\vartheta, \beta)$ given in (21). The model of the observed data again is a cumulative probit model.

4 Conclusion

Quasi-likelihood models offer an useful method to analyze nonlinear regression models with measurement error in the covariates. Provided additional information on the error process is given an heteroscedastic variance structure of the measurement errors can be embedded into the models as well. In this approach the latent regressor variables are treated as stochastic variables following a normal distribution. Even if mandatory this assumption may not be fulfilled for some applications and one way to consider this would be a straightforward extension to the case of a mixture of normal distributions.

References

[1] Armstrong, B. (1985). Measurement error in the generalized linear model. *Communications in Statistics.* **14**, 529-544.

[2] Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association.* **85**, 565-571.

[3] Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall, London.

[4] Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association.* **85**, 652-663.

[5] Cheng, C.L. and Schneeweiß, H. (1998). Polynomial regression with errors in the variables. To appear in *Journal of the Royal Statistical Society B.*

[6] Crouch, E.A.C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association.* **85**, 464-469.

[7] Eckert, R.S., Carroll, R.J. and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics.* **53**, 262-272.

[8] Johnson, N.L. and Kotz, S. (1970). *Continuous univariate distributions - 2.* John Wiley, New York.

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

[10] Liang, K. Y. and Liu, X. (1991). Estimating equations in generalized linear models with measurement error. In *Estimating Functions.* V.P. Godambe (ed.). Oxford University Press, New York.

[11] Liu, Q. and Pierce D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika.* **81**, 624-629.

[12] McCullagh, P. (1991). Quasi-likelihood and estimating functions. in *Statistical Theory and Modelling*, D.V. Hinkley, N. Reid and E.J. Snell (eds.). Chapman and Hall, London.

[13] Moon, M.S. and Gunst, R.F. (1995) Polynomial measurement error modelling. *Computational Statistics & Data Analysis.* **19**, 1-21.

[14] Rosner, B., Willett, W.C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine.* **8**, 1051-1069.

[15] Schafer, D.W. (1993). Likelihood analysis for probit regression with measurement error. *Biometrika.* **80**, 899-904.

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

[17] Stefanski, L.A. (1985). The effect of measurement error on parameter estimation. *Biometrika.* **72**, 583-592.

[18] Thamerus, M. (1997). Modelling Count Data with Heteroscedastic Measurement Error in the Covariates. *Discussion Paper 58, SFB 386.* Ludwig-Maximilians-Universität München.

[19] Tosteson, T.D., Stefanski, L.A. and Schafer, D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine.* **8**, 1139-1147.