Thamerus:

Modelling Count Data with Heteroscedastic Measurement Error in the Covariates

Sonderforschungsbereich 386, Paper 58 (1997)

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

Modelling Count Data with Heteroscedastic Measurement Error in the Covariates

Markus Thamerus, SFB 386

Institut für Statistik, Universität München

Abstract

This paper is concerned with the estimation of the regression coefficients for a count data model when one of the explanatory variables is subject to heteroscedastic measurement error. The observed values W are related to the true regressor X by the additive error model W=X+U. The errors U are assumed to be normally distributed with zero mean but heteroscedastic variances, which are known or can be estimated from repeated measurements. Inference is done by using quasi likelihood methods, where a model of the observed data is specified only through a mean and a variance function for the response Y given W and other correctly observed covariates. Although this approach weakens the assumption of a parametric regression model, there is still the need to determine the marginal distribution of the unobserved variable X, which is treated as a random variable. Provided appropriate functions for the mean and variance are stated, the regression parameters can be estimated consistently. We illustrate our methods through an analysis of lung cancer rates in Switzerland. One of the covariates, the regional radon averages, cannot be measured exactly due to the strong dependency of radon on geological conditions and various other environmental sources of influence. The distribution of the unobserved true radon measure is modelled as a finite mixture of normals.

Keywords: measurement error; quasi likelihood; Poisson regression; radon data

1 Introduction

When ordinary regression techniques are applied to a model where one or several predictors are subject to measurement error, the regression parameter estimates are asymptotically biased. For nonlinear models the monograph of Carroll, Ruppert and Stefanski (1995) gives a fundamental introduction into the different methods to adjust for this effect. In this article we will focus on estimation and inference of a Poisson regression model with heteroscedastic measurement error in one of the predictors. Let the true model relate the response Y, given in counts, to the predictors (X, Z), where X denotes a continuous covariate that cannot be measured directly and is only observed through a proxy W, and Z is a set of covariates measured without error.

Throughout this paper we will focus on a structural model for the unobserved predictor X, which means that X is treated as a random variable and its distribution is parametrically modeled. Furthermore we make the assumption of nondifferential measurement error, which means that the conditional distribution of Y given X and Z is independent of W: $f_{Y|Z,X,W} = f_{Y|Z,X}$. The observed predictor W is then called a surrogate. This includes the frequently used additive measurement error model $W = X + U$, where the measurement error $U \sim (O, \sigma_u^2)$ is independent of (Y, X). Quasilikelihood methods for regression models with covariate measurement error require information on the posterior distribution of the true predictor X given the observed covariates (W, Z). If validation data for X are at hand and an assumption for the error distribution of U is made, one can proceed to estimate the distribution of $X \mid W, Z$. This is very often not the case and one has to make a strong assumption on the distribution of X and use the observations of W to estimate it. Therefore some knowledge about the error process U that generated the observations W is needed. In contrast to most applications which assume the error variances to be constant, we allow for heteroscedastic measurement errors, that is, $\mathrm{Var}\,(U_i) = \sigma_i^2, i = 1, \ldots, n$.

Our work was mainly motivated by a data set from a Swiss study (Minder and Völkle, 1995), where registered (mortal) lung cancer cases (Y) were related to regional average radon measurements (W) and other predictors (Z). The observed mean values W for regional radon exposure have to be regarded as proxy variables for an existing true mean X of each region. Since the number of individual radon measurements that were used to compute the average W for each region ranged from 16 to 511, the errors cannot be assumed to be homoscedastic.

The aim of this paper is to show how a quasilikelihood approach can be used for a count data model when one of the explanatory variables is subject to heteroscedastic measurement error. The assumption of a finite mixture of normal distributions as the marginal distribution for the latent variable X is very flexible and it is shown that the derivation of a regression model in the observable variables remains tractable.

In the following section we will introduce the quasilikelihood model for a Poisson regression and derive appropriate mean and variance functions when the latent variable

X follows a normal mixture distribution. In section three we will apply this approach to the Swiss data. The impact of measurement error on the estimation results and other related aspects will be discussed in the last section.

2 The Quasilikelihood Approach

The use of quasilikelihood techniques for regression models with covariate measurement error has been widely discussed in the literature. One of the first general approaches has been described by Armstrong (1985). Asymptotic results and a very detailed discussion of quasilikelihood methods for different observed data structures can be found in Carroll and Stefanski (1990).

For $i = 1, \ldots, n$ let Y_i and Z_i be the response and a vector of covariates measured without error, X_i denotes the unobservable regressor variable and W_i the measured surrogate. We assume an additive heteroscedastic error model,

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

where the U_i's are mutually independent, U_i and (Y_i, X_i) are independent and the error variances σ_i^2 are known or can be estimated from independent replications of W_i. The quasilikelihood approach only requires the specification of a mean and variance function for the regression model, which will be denoted by f_m and f_v. The first step to obtain a quasilikelihood model in the observable variables is to set up the 'unobservable' mean and variance function as it is implied by the distribution of Y_i given Z_i and X_i. We will write those first two conditional moments as

$$\mathrm{E}\,(Y_i \mid Z_i, X_i) = \mu\,(Z_i, X_i, \beta) \quad \text{for the mean and} \tag{2}$$

$$\mathrm{V}\,(Y_i \mid Z_i, X_i) = \sigma^2\,(Z_i, X_i, \beta) \quad \text{for the variance} \tag{3}$$

function, where β is the vector of the regression parameters. In a more general formulation the variance function depends on additional variance parameters ν or/and is expressed as a function of μ, but as we will concentrate on a Poisson regression merely, there is no need for a more general notation. To proceed to the mean

and variance functions for the observed data, $f_m(Z_i, W_i, \beta) = \mathrm{E}(Y \mid Z_i, W_i)$ and $f_v(Z_i, W_i, \beta) = \mathrm{V}(Y_i \mid Z_i, W_i)$, one iterates expectations and uses the nondifferential error property. A quasilikelihood model in the observable variables can therefore be stated as

$$\begin{aligned} f_m(Z_i, W_i, \beta) &= \mathrm{E}(\mathrm{E}(Y_i \mid Z_i, X_i, W_i) \mid Z_i, W_i) \\ &= \mathrm{E}(\mu(Z_i, X_i, \beta) \mid Z_i, W_i) \quad \text{and} \qquad (4) \\ f_v(Z_i, W_i, \beta) &= \mathrm{V}(\mathrm{E}(Y_i \mid Z_i, X_i, W_i) \mid Z_i, W_i) + \mathrm{E}(\mathrm{V}(Y_i \mid Z_i, X_i, W_i) \mid Z_i, W_i) \\ &= \mathrm{V}(\mu(Z_i, X_i, \beta) \mid Z_i, W_i) + \mathrm{E}(\sigma^2(Z_i, X_i, \beta) \mid Z_i, W_i). \qquad (5) \end{aligned}$$

An unbiased estimating equation for $\beta = (\beta_0, \beta_Z', \beta_X)'$ is given by the 'quasi' score-function

$$\mathrm{s}^{(n)}(\beta) = \sum_{i=1}^{n} \frac{\partial f_m(Z_i, W_i, \beta)}{\partial \beta} \frac{Y_i - f_m(Z_i, W_i, \beta)}{f_v(Z_i, W_i, \beta)} = \sum_{i=1}^{n} \mathrm{s}_i(\beta)$$

and the consistent quasilikelihood estimator $\hat{\beta}_{ql}$ is found as the root of the equation $\mathrm{s}(\beta) = 0$. Its asymptotic normality is also established via the theory of unbiased estimating equations and it holds that

$$\hat{\beta}_{ql} \overset{a}{\sim} \mathrm{N}\left(\beta, n^{-1} F^{-1}(\beta) V(\beta) F^{-1}(\beta)\right). \qquad (6)$$

The parts of the asymptotic covariance matrix of $\hat{\beta}_{ql}$ are given by

$$\begin{aligned} F(\beta) &= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}\left(-\frac{\partial \mathrm{s}^{(n)}(\beta)}{\partial \beta'}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(-\frac{\partial \mathrm{s}_i(\beta)}{\partial \beta'}\right) \quad \text{and} \qquad (7) \\ V(\beta) &= \lim_{n\to\infty} \frac{1}{n} \mathrm{cov}\left(\mathrm{s}^{(n)}(\beta)\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left(\mathrm{s}_i(\beta) \left(\mathrm{s}_i(\beta)\right)'\right). \qquad (8) \end{aligned}$$

It is estimated by

$$\widehat{\mathrm{cov}}(\hat{\beta}_{ql}) = n^{-1} \hat{F}^{-1}(\hat{\beta}_{ql}) \hat{V}(\hat{\beta}_{ql}) \hat{F}^{-1}(\hat{\beta}_{ql}) \quad \text{with} \qquad (9)$$

$$\hat{F}(\hat{\beta}_{ql}) = \frac{1}{n}\left(\sum_{i=1}^{n} -\frac{\partial \mathrm{s}_i(\beta)}{\partial \beta'}\Big|_{\beta=\hat{\beta}_{ql}}\right) \quad \text{and} \quad \hat{V}(\hat{\beta}_{ql}) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{s}_i(\beta) \left(\mathrm{s}_i(\beta)\right)'\Big|_{\beta=\hat{\beta}_{ql}}\right). \qquad (10)$$

Both, mean (4) and variance function (5) make use of the conditional distribution of X given Z and W. If this distribution can be specified parametrically, it is in

principle possible to calculate them directly. In our case, we will state f_m and f_v under the assumption that the heteroscedastic error variances σ_i^2 are given and that the parameters of the marginal distribution of X can be estimated.

Model for the Poisson Regression

For $i = 1, \ldots, n$ let $Y_i \sim Po(\lambda_i)$ with $\lambda_i = \exp(\beta_0 + Z_i'\beta_Z + \beta_X X_i)$. The underlying 'unobservable' regression model is given through

$$\mu(Z_i, X_i, \beta) = \sigma^2(Z_i, X_i, \beta) = \exp(\beta_0 + Z_i'\beta_Z + \beta_X X_i) \tag{11}$$

and by using the formulas as given in (4) and (5), it is easily seen that the mean and variance functions of the 'observable' model is of the form

$$\begin{aligned} \mathrm{E}(Y_i \mid Z_i, W_i) &= \exp(\beta_0 + Z_i'\beta_Z) \cdot \mathrm{E}(\exp(\beta_X X_i) \mid Z_i, W_i) \quad \text{and} & (12) \\ \mathrm{V}(Y_i \mid Z_i, W_i) &= \exp(\beta_0 + Z_i'\beta_Z) \cdot \mathrm{E}(\exp(\beta_X X_i) \mid Z_i, W_i) \\ &\quad + \exp(2\beta_0 + 2Z_i'\beta_Z) \cdot \mathrm{E}(\exp(2\beta_X X_i) \mid Z_i, W_i) \\ &\quad - [\exp(\beta_0 + Z_i'\beta_Z) \cdot E(\exp(\beta_X X_i) \mid Z_i, W_i)]^2. & (13) \end{aligned}$$

Now expectations of the form $\mathrm{E}(\exp(cX_i) \mid Z_i, W_i)$ have to be computed. To derive closed form expressions for f_m and f_v we will proceed in the following way: Under the assumption of a structural model we state a parametric distribution for the latent variables X_i. From these we find the conditional distribution of X_i given W_i and, as only normal distributions are involved, it is then possible to compute the conditional expectation $\mathrm{E}(\exp(cX_i) \mid Z_i, W_i)$.

We will model the distribution of the i.i.d variables $X_i, i = 1, \ldots, n$, parametrically as a mixture of normal distributions and write its density as

$$f_{X_i}(x_i) = \sum_{k=1}^{m} \alpha_k \, \varphi(x_i \mid \mu_k, \tau_k^2),$$

where $\varphi(\cdot \mid \mu_k, \tau_k^2)$ denotes the normal density function with parameters μ_k and τ_k^2. Finite mixture distributions provide a flexible class of distributions and often represent a more realistic choice in practice as they do not demand that the observed

variables come from one homogeneous population. As we will see later in the example, the assumption of a mixture distribution was indicated by the observed data. Additionally it is assumed that the latent variables X_i are independent of the other covariates Z_i. The random variable of the k-th component of the mixture distribution of X_i will be denoted by X_{ki} with density $\varphi(x_i \mid \mu_k, \tau_k^2)$. Since we have an additive error model $W_i = X_i + U_i$, $U_i \sim N(0, \sigma_i^2)$, it is easily seen that the distribution of W_i is a mixture of normal distributions as well. Indeed we find that on each component variance of that mixture distribution an heteroscedastic variance part induced by the measurement error is added. Therefore the density of the k-th component W_{ki} of W_i is given by $\varphi(w_i \mid \mu_k, \tau_k^2 + \sigma_i^2)$. In order to find the conditional distribution of $X_i \mid W_i$, we simplify the notation and write the densities of X_i, W_i and U_i as $f_{X_i}(x_i) = \sum_{k=1}^m \alpha_k \, f_{X_{ki}}(x_i)$, $f_{W_i}(w_i) = \sum_{k=1}^m \alpha_k \, f_{W_{ki}}(w_i)$ and $f_{U_i}(u_i)$, respectively. By applying a linear transformation to the joint density of X_i and W_i we find

$$\begin{aligned} f_{X_i|W_i}(x_i) &= \frac{f_{X_i,W_i}(x_i, w_i)}{f_{W_i}(w_i)} = \frac{f_{X_i}(x_i) f_{U_i}(w_i - x_i)}{f_{W_i}(w_i)} = \frac{\sum_{k=1}^m \alpha_k f_{X_{ki}}(x_i) f_{U_i}(w_i - x_i)}{\sum_{k=1}^m \alpha_k f_{W_{ki}}(w_i)} \\ &= \sum_{k=1}^m \frac{\alpha_k f_{W_{ki}}(w_i)}{\sum_{j=1}^m \alpha_j f_{W_{ji}}(w_i)} \frac{f_{X_{ki}}(x_i) f_{U_i}(w_i - x_i)}{f_{W_{ki}}(w_i)} = \sum_{k=1}^m \lambda_{ki} f_{C_{ki}}(x_i). \qquad (14) \end{aligned}$$

As can be seen from (14) the conditional distribution of X_i given W_i again is a mixture of normally distributed random variables $C_{ki}, k = 1, \ldots, m$ with its associated densities found by conditioning X_{ki} on W_{ki} for each k. The proportions λ_{ki}, given by

$$\lambda_{ki} = \frac{\alpha_k \cdot \varphi(w_i \mid \mu_k, \tau_k^2 + \sigma_i^2)}{\sum_{j=1}^m \alpha_j \cdot \varphi(w_i \mid \mu_j, \tau_j^2 + \sigma_i^2)},$$

are the posteriori probabilities that the unobserved variable X_i belongs to component k when W_i was observed. Furthermore it holds that $C_{ki} \sim N(\mu_{ki}, \sigma_{ki}^2)$ with its parameters defined as

$$\begin{aligned} \mu_{ki} &= \mu_k + \frac{\tau_k^2}{\tau_k^2 + \sigma_i^2}(W_i - \mu_k) \quad \text{and} \\ \sigma_{ki}^2 &= \tau_k^2 (1 - \frac{\tau_k^2}{\tau_k^2 + \sigma_i^2}). \end{aligned}$$

Küchenhoff and Carroll (1997) used a similar argumentation for a homoscedastic measurement error model and a marginal mixture distribution with two components.

Now since X_i given W_i is a mixture distribution, we can rewrite the conditional expectations required for the definition of the mean and variance function of the quasilikelihood model and it holds that

$$\mathrm{E}\left(\exp\left(c\,X_i\right) \mid W_i, Z_i\right) = \sum_{k=1}^{m} \lambda_{ki} \mathrm{E}\left(\exp\left(c\,X_{ki}\right) \mid W_{ki}, Z_i\right)). \tag{15}$$

The properties of the moment generating function for normal distributions enables us to express these expectations as

$$\mathrm{E}\left(\exp\left(c\,X_{ki}\right) \mid W_{ki}, Z_i\right) = \mathrm{E}\left(\exp\left(c\,C_{ki}\right)\right) = \exp\left(c\,\mu_{ki} + c^2\sigma_{ki}^2 \cdot 0.5\right). \tag{16}$$

With this result and (15) plugged into (12) and (13) the derived model in the observable variables is given by

$$f_m(Z_i, W_i, \beta) = \exp\left(\beta_0 + Z_i'\beta_Z\right)\left[\sum_{k=1}^{m} \lambda_{ki} \exp\left(\beta_X \mu_{ki} + \beta_X^2 \cdot \sigma_{ki}^2 \cdot 0.5\right)\right], \tag{17}$$

$$\begin{aligned} f_v(Z_i, W_i, \beta) &= f_m(Z_i, W_i, \beta) - \left[f_m(Z_i, W_i, \beta)\right]^2 \\ &\quad + \exp\left(2\beta_0 + 2Z_i'\beta_Z\right)\left[\sum_{k=1}^{m} \lambda_{ki} \exp\left(2\beta_X \mu_{ki} + 2\beta_X^2 \cdot \sigma_{ki}^2\right)\right]. \end{aligned} \tag{18}$$

This model is clearly different from the unobservable Poisson regression model as stated in (11). Estimation is carried out by the usual iteratively reweighted least square algorithm for mean and variance models and requires to differentiate $f_m(Z_i, W_i, \beta)$ with respect to β. For details on fitting methods for such models see Carroll and Ruppert (1988).

3 Lung Cancer Data

In a recent study (Minder and Völkle, 1995) the objective was to find out if there exists a positive association between regional average radon measurements and registered, mortal lung cancer cases. The study was carried out in Switzerland, which was divided into 46 different regions. In each region the numbers of registered lung cancer cases were given for each of sixteen age groups. Regional average values of radon were obtained by repeated indoor measurements from different sites across each

region. Besides location the sites differed from each other by the type of building and the chosen floor level. As the latent covariate X_i we define the true average radon concentration for region i. For each of the 46 regions a mean value W_i was obtained through n_i single observations W_{ir}, $r = 1, \ldots, n_i$. The sample variances S_i^2 from these repeated measurements were given as well. The concentration of the radon gas strongly depends on local geological and atmospherical conditions. Furthermore the physical property of radon to decompose into other substances makes it difficult to obtain exact values. The location of the measuring devices and the instruments themselves are thus possible sources of measurement error. We will state the following additive model for the measurement error process: each observation W_{ir} is a proxy variable for the true regional average X_i and therefore we define for all i

$$W_{ir} = X_i + \epsilon_{ir} \text{ with } \mathrm{E}(\epsilon_{ir}) = 0 \text{ and } \mathrm{Var}(\epsilon_{ir}) = \sigma_{\epsilon_i}^2 \text{ for } r = 1, \ldots, n_i.$$

So we do not assume a particular distribution for the sampling errors ϵ_{ir}, we only require that they have expectation zero and equal variances. For the observed values $W_i = n_i^{-1} \sum_{r=1}^{n_i} W_{ir} = X_i + U_i$ with $U_i = n_i^{-1} \sum_{r=1}^{n_i} \epsilon_{ir}$ the central limit theorem permits us to assume a normal error distribution and we write

$$W_i = X_i + U_i \text{ with } U_i \sim N(0, \sigma_i^2) \text{ for } i = 1, \ldots, 46.$$

Its easily seen that the error variances $\sigma_i^2 = \sigma_{\epsilon_i}^2 / n_i$ are different for each region, even when $\sigma_{\epsilon_i}^2 = \sigma_\epsilon^2$ for $i = 1, \ldots, 46$, since they depend on the number of measurements n_i as well. This number n_i varies regionally from 16 to 511 observations and for the estimated error variances $\hat{\sigma}_i^2 = S_i^2 / n_i$ we find $\hat{\sigma}_i^2 \in [0.769, 426.983]$. The estimates $\hat{\sigma}_i^2$ will be treated as the variances σ_i^2, which we formerly assumed to be known. Figure 1 shows a scatterplot of the regional radon averages versus the estimated standard deviations $\hat{\sigma}_i$ of their error distributions. Marked by triangles and squares are averages computed from less, respectively, equal or more than one hundred single measurements. The plot clearly shows the heteroscedastic pattern of the error variances and although it is obvious that $\hat{\sigma}_i^2$ will tend to zero if n_i increases, this data show enough variability within each region to produce nonignorable measurement error. In the original study a number of Poisson regression models for different subgroups of the Swiss

population were estimated. We will restrict our analysis on that model that includes all Swiss women only. The response variable is the number of registered mortal lung cancer cases in region i and age interval j and will be denoted by Y_{ij}. As described above the predictor of main interest, the regional average radon concentration X_i, could only be observed through the surrogate W_i with known error variances σ_i^2.

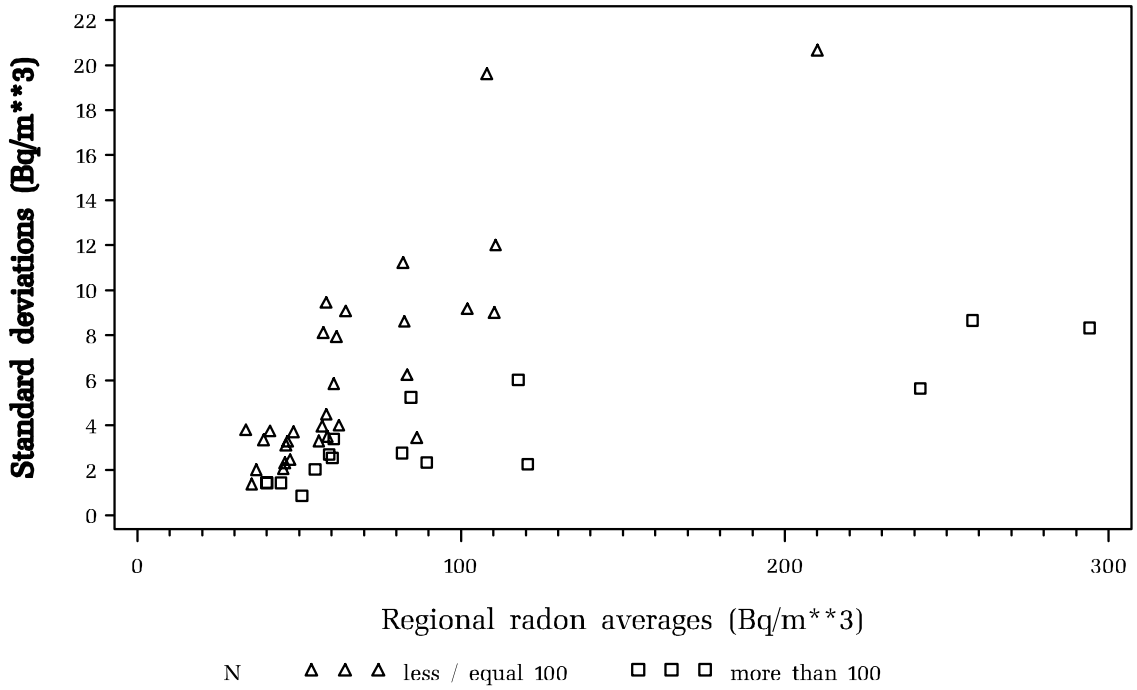

Figure 1. *Scatterplot of the regional radon averages against their error standard deviations. The triangles and squares indicate if* $n_i \leq 100$ *or if* $n_i > 100$.

The correctly observed covariables are the population under risk N_{ij} and age A_j measured as the transformed midpoint of the j-th age interval, which equals zero for women between 15 and 19 years, equals one for age between 20 and 24 years, a.s.o.. Additionally an indicator variable C_i of the regional structure (1 for urban, 0 for rural) is given. The observed data structure for the regression model is summarized in Table 1. The 'unobservable' loglinear offset regression model relates the proportions Y_{ij}/N_{ij} to the covariates $Z_{ij} = (1, A_j, A_j^2, C_i)'$ and X_i, so that the logarithm of the first conditional moment of Y_{ij} given N_{ij}, Z_{ij} and X_i can be written as

$$\ln\left(\mu\left(N_{ij}, Z_i, X_i, \beta\right)\right) = \ln\left(N_{ij}\right) + \beta_0 + \beta_{A1} A_j + \beta_{A2} A_j^2 + \beta_C C_i + \beta_X X_i$$

and (11) holds.

	j-th age group (j=1,...,16): A_j
i-th region (i=1,...,46):	Y_{ij} ... registered lung cancer cases
W_i with σ_i^2, C_i	N_{ij} ... population under risk

Table 1. *Data structure of Swiss Study: observed variables.*

Figure 2 shows the regional radon averages plotted against $\ln(Y_{ij}/N_{ij})$. The regions considered as urban are marked by triangles. The plot itself gives no clear hint for the presence of an effect of radon on the occurrence of lung cancer. Markedly visible is the characteristic of the radon averages to appear in three distinct clusters. The main part of the data clusters around 50 Bq/m^3, the second group scatters around 100 Bq/m^3 and on the right hand side of the plot are four regions with averages above 200 Bq/m^3.

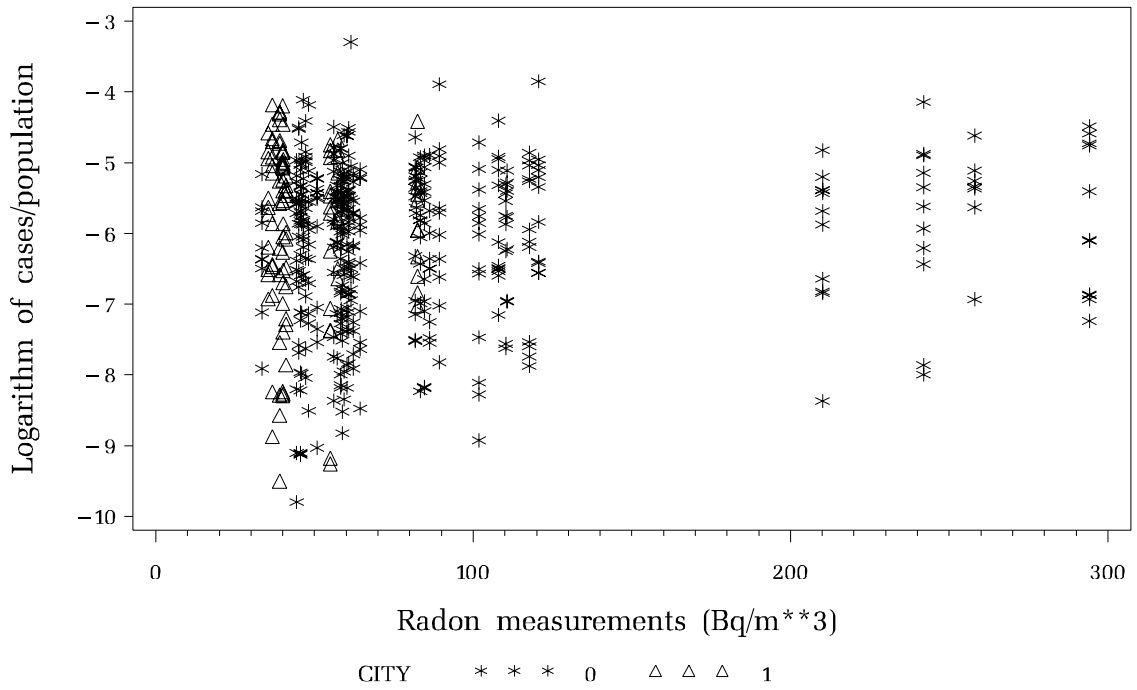

Figure 2. *Scatterplot of* $\ln(Y_{ij}/N_{ij})$ *against the observed radon averages* W_i. *Regions considered as urban/rural* ($C_i = 1/0$) *are marked by triangles/stars.*

As the marginal distribution of the true radon averages X_i we assume a normal mixture distribution with three components. Maximum likelihood estimates of the

parameters were obtained by applying an EM algorithm to the observed radon means W_i The results are shown in Table 2, for more details see Thamerus (1996).

	Component 1		Component 2		Component 3	
proportions α_k	0.6225	(0.0815)	0.2905	(0.0776)	0.0870	(0.0416)
means μ_k	49.7573	(1.8943)	94.2082	(6.4696)	254.5958	(13.9715)
stand. dev. τ_k	8.4732	(1.4605)	15.5282	(5.3460)	25.3031	(11.4102)

Table 2. *Estimation results for a three component normal mixture distribution of the true radon means. Standard errors are given in brackets.*

A naive estimator for $\beta = (\beta_0, \beta_{A1}, \beta_{A2}, \beta_C, \beta_X)'$ was originally obtained by replacing X_i with the observed averages W_i. It is well known, that this method yields inconsistent estimates. The estimated regression coefficients of the quasilikelihood model are found by applying an IRLS algorithm to the model given through the mean and variance functions (17) and (18). These estimates are presented in Table 3 together with those of the naive approach .

	naive model			quasilikelihood model		
variable	$\hat{\beta}$	se	p	$\hat{\beta}$	se	p
	-11.78876	0.15535	0.00000	-11.79294	0.15522	0.00000
age	0.98625	0.03276	0.00000	0.98624	0.03276	0.00000
age**2	-0.03688	0.00172	0.00000	-0.03688	0.00172	0.00000
urban	0.42568	0.03223	0.00000	0.42693	0.03199	0.00000
radon	0.00072	0.00038	0.05652	0.00078	0.00037	0.03387

Table 3. *Estimation results for the regression model of the Swiss lung cancer data. Given are the estimated regression coefficients, their standard errors and associated p values.*

For the naive procedure we found that the null hypothesis for the presence of a radon effect could not be rejected on a 5 % significance level. Note that the statistical inference is different for the quasilikelihood model that considers the individual measurement errors. Relative to their standard errors, both models produce similar results for the correctly observed covariates age, age squared and the urbanization indicator. The estimated radon effect of the quasilikelihood model however is greater than the one obtained from the naive model and its accompanying p value confirms a significant effect for the radon variable at the 5 % level. This difference in the p values of the two models is explained by the almost identical values of their standard errors. As a result we may state, that for this particular model, the naive estimation method finds a non-significant radon effect and that in comparison, the quasilikelihood approach leads to a different result.

4 Discussion

Most epidemiologists will confirm that age and smoking status have the strongest effects on the occurrence of lung cancer and that in this data set the absence of an appropriate smoking variable produces misleading results. This issue is also discussed in the original paper of Minder and Völkle (1994). They compared their estimation results of separate models for distinct age groups under the alternative assumptions whether the overall smoking behavior of the population remained constant or was dynamic. Since there is no information that smoking will be a confounding factor for radon we cannot contribute anything new to this discussion.

We will rather concentrate on two other topics. The first one is about the asymptotic covariance matrix of $\hat{\beta}_{ql}$. In our model the parameters of the distribution of X_i, for simplicity denoted by δ, are treated as known. The 'sandwich' estimator (9) that was used to estimate the covariance of $\hat{\beta}_{ql}$ does not consider the estimation of δ. According to Liang and Liu (1991) an estimator of similar form as (9) for the covariance can be constructed if $\hat{V}(\hat{\beta}_{ql})$ is replaced by a term that contains one part for the estimation of β and an additional part for the estimation of δ. It remains open whether the

estimated standard errors for $\hat{\beta}_{ql}$ would increase significantly if the additional part was used.

A very common method to describe the degree of attenuation of the estimated regression coefficients in the presence of measurement error is the definition of a ratio that relates the error variance to the variance of the latent variable X. If the error variance is homoscedastic, the so called *noise-to-signal ratio* of X, that is $\gamma = \mathrm{Var}\,(U_i)/\mathrm{Var}\,(X_i)$, is often used (see e.g. Fuller, 1987). As our error model $W_i = X_i + U_i$, $U_i \sim N(0, \sigma_i^2)$ is heteroscedastic, we will use this idea to define a *mean noise-to-signal ratio* of X as $\gamma_m = n^{-1} \sum_{i=1}^{n} \sigma_i^2/\mathrm{Var}\,(X_i)$ and estimate it by replacing $\mathrm{Var}\,(X_i) = \sigma_X^2$ with an estimate and make use of the known error variances σ_i^2.

The latent variables $X_i, i = 1, \ldots, n$ were assumed to be i.i.d. variables of a mixture of normal distributions, so we write $X_i \sim MixNV(\alpha_1, \ldots, \alpha_m, \mu_1, \ldots, \mu_m, \tau_1^2, \ldots, \tau_m^2)$ and suggest two ways of estimating σ_X^2. The first method uses the estimated parameters of the mixture distribution from the EM algorithm. Let H_i be a classification variable that defines to which component of the mixture X_i belongs. Then the variance of X_i can be found by

$$\mathrm{Var}\,(X_i) = \mathrm{E}\,(\mathrm{Var}\,(X_i \mid H_i)) + \mathrm{Var}\,(\mathrm{E}\,(X_i \mid H_i)) = \sum_{k=1}^{m} \alpha_k \, \tau_k^2 + \sum_{k=1}^{m} \alpha_k (\mu_k - \mu)^2$$

where $\mu = \mathrm{E}\,(X_i) = \sum_{k=1}^{m} \alpha_k \mu_k$. The ML estimate $\hat{\sigma}_X^2$ is simply obtained by replacing the distribution parameters with its estimates. The method of moments uses the sample variance S_W^2 of the observed variables W_i and an estimator for σ_X^2 is found by

$$s_X^2 = S_W^2 - \frac{n-1}{n^2} \sum_{i=1}^{n} \sigma_i^2.$$

It is easily seen that s_X^2 is unbiased. This estimator is of great practical use since it can be computed without any knowledge of the distribution of the latent variable X. Most variation in X is caused by the four radon means that constitute the third component of the mixture distribution. To get an idea of the measurement error effect on the estimation results we performed an experiment and removed the four regions with radon averages above 200 Bq/m^3 from the data and fitted a normal mixture distribution with two components to the remaining averages. Table 4 gives

the estimated variances and mean signal-to-noise ratios for the original Swiss radon data (three components) and the reduced data (two components).

	three components: $n = 46$		two components: $n = 42$	
	variance	ratio	variance	ratio
$\hat{\sigma}_X^2$	3448.6645	0.0137	554.2709	0.0675
s_X^2	3383.6864	0.0139	554.5751	0.0675

Table 4. *Estimated variances and mean noise-to-signal ratios of X for the three components mixture model (full data) and the two components mixture modell (four data points omitted).*

The ratios for the full data model are rather small, a fact which is mainly caused by the different locations of the three components of the mixture. Therefore the impact of measurement error on the estimated radon effect is small and the naive estimator is only little biased. That the error variances influence the estimation results can be seen from the model of the observed data, given in (17) and (18). Both functions depend nonlinearily on the ratios σ_i^2/τ_k^2 and the location parameters μ_k through the conditional moments μ_{ki} and σ_{ki}.

The mean noise-to-signal ratios for the reduced data (two components) are approximately five times bigger than those for the original data and the biasing effect of measurement error on the naive estimates should be seen more clearly. Indeed, we computed the regression coefficients for the naive and the quasilikelihood regression for those data and got estimated radon effects of $\hat{\beta}_{X,naive} = -0.00086\,(0.00087)$ for the naive and $\hat{\beta}_{X,ql} = -0.00048\,(0.00091)$ for the quasilikelihood model (with standard errors given in brackets). Relative to their standard errors those two estimates differ from each other by a factor around two. Not surprisingly this example also reveals that the positive effect of radon as it was found by the full data model, disappears once the four highest radon exposed regions are not considered.

Quasilikelihood models are useful tools to analyze regression models when some of the covariates are subject to measurement error. Collecting repeated measurements of the erroneous regressor variable provides additional information on the measurement error process and is recommended to the researchers. If the marginal distribution of the latent variable is normal or a mixture of normal distributions, even a heteroscedastic error structure can be embedded into a quasilikelihood model for count data. Especially weak effects like the discussed effect of radon exposure on lung cancer can be detected by a model that considers the individual measurement error.

Acknowledgement

I want to thank W.A. Fuller, H. Schneeweiss and H. Küchenhoff for their useful comments and Ch. E. Minder for providing the problem.

5 References

Armstrong, B. (1985). Measurement Error in the Generalized Linear Model. *Communications in Statistics.* Series B, 14, 529-544.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, London.

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

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.

Fuller, W.A. (1980). *Measurement Error Models.* John Wiley, New York.

Küchenhoff, H. and Carroll, R. J.(1997). Segmented Regression with Errors in Predictors: Semi-Parametric and Parametric Methods. *Statistics in Medicine.* Vol. 16, 169-188.

Liang, K. Y. and Liu, X. (1991). Estimating Equations in Generalized Linear Models with Measurement Error. In *Estimating Functions,* V.P. Godambe (ed.). New York. Oxford University Press.

Minder, Ch. E. and Völkle, H. (1995). Radon und Lungenkrebssterblichkeit in der Schweiz. *324. Bericht der mathematisch-statistischen Sektion der Forschungsgesellschaft Johanneum*, 115-124.

Thamerus, M. (1996). Fitting a Finite Mixture Distribution to a Variable Subject to Heteroscedastic Measurement Error. *University of Munich, Discussion Paper List of the SFB 386*, Nr. 48.