Bias of the Quasi Score Estimator of a Measurement Error Model Under Misspecification of the Regressor Distribution.

Hans Schneeweiss, Department of Statistics, University of Munich, Akademiestr. 1, 80799 Munich, Germany, email: schneew@stat.uni-muenchen.de
Chi-Lun Cheng, Institute of Statistics, Academia Sinica, Taipei 11529, Taiwan R.O.C., email: clcheng@stat.sinica.edu.tw

Abstract: In a structural error model the structural quasi score (SQS) estimator is based on the distribution of the latent regressor variable. If this distribution is misspecified the SQS estimator is (asymptotically) biased. Two types of misspecification are considered. Both assume that the statistician erroneously adopts a normal distribution as his model for the regressor distribution. In the first type of misspecification the true model consists of a mixture of normal distributions which cluster round a single normal distribution, in the second type the true distribution is a normal distribution admixed with a second normal distribution of low weight. In both cases of misspecification the bias, of course, tends to zero when the size of misspecification tends to zero. However, in the first case the bias goes to zero in a flat way so that small deviations from the true model lead to a negligible bias, whereas in the second case the bias is noticeable even for small deviations from the true model.

1 Introduction

A measurement error (or errors-in-variables) model is essently a regression model with measurement errors in the regressor variables, Schneeweiss and Mittag (1986), Fuller (1987), Carroll et al (1995), Cheng and Van Ness (1999), Wansbeek and Meijer (2000). It is well-known that disregarding the measurement error when estimating the regression parameters, i.e., using an estimation procedure as if there were no measurement errors, results in inconsistent estimates. In some cases, the (asymptotic) bias can be devastating, rendering the results of a regression analysis completely useless.

On the assumption that the measurement process as such is known, in particular, that the measurement error variance is known, it is often possible to correct the estimation procedure for the measurement error so as to recover consistency.

Structural Quasi Score (SQS) (or Quasi Likelihood) is a method that achieves this objective, Armstrong (1985), Carroll et al (1995), Kukush et al (2002). It takes advantage of the distribution of the regressor variables, which is supposed to be known, or known at least up to a finite number of parameters (structural model).

There are other, so-called functional, methods that do not depend on the knowledge of the regressor distribution, Carroll et al (1995). But these methods typically yield estimates with larger standard errors than SQS, just because they do not use the information inherent in the regressor distribution, for an example see Shklyar and Schneeweiss (2002).

On the other hand, a method like SQS that so much depends on knowledge of the regressor distribution may react adversely to any form of misspecification of this distribution, in other words, there is the risk of reintroducing a bias in the estimation of the regression parameters if the method is based on the wrong model for the regressor distribution.

This bias has been found in various simulation studies, Schneeweiss and Nittner (2001), and has been systematically studied for small misspecification deviations in a polynomial regression model by Schneeweiss et al (2003).

In the latter paper a normal distribution for the scalar regressor variable was assumed as the misspecified model, whereas the true model consisted of a mixture of two normals with equal variances but different means. The difference of means was taken as a measure of deviation of the misspecified model from the true model. When that measure tended to zero the misspecification bias of the estimated polynomial regression coefficients also tended to zero - as was to be expected -, but the bias was of the order of the squared deviation rather than just of the order of the deviation, so that the bias tended to zero in a flat way.

This result can be interpreted to mean that the bias of SQS under small deviations from the assumed model is comparatively very small and may often be neglected. SQS reacts rather insensitively to a misspecification of the regressor distribution as long as the misspecification is of this particular form and is not too strong.

This result was derived for a very special case: a special, viz., polynomial, regression model was assumed, a normal distribution was taken for the regressor, a very limited form of deviation from this assumption was considered, finally only the scalar case was investigated. In this paper we generalize this result to an arbitrary nonlinear regression model, a more general distribution for the regressor variables, a wider form of deviation from this distribution, and we consider the multivariate case. The true regressor distribution is taken to be a finite mixture of normals which cluster round a single normal, whereas the erroneously assumed distribution is just a normal distribution. This type of misspecification, which generalizes the situation studied in Schneeweiss et al (2003) and which can still be generalized further, is called the clustering type of misspecification.

We also study a completely different form of deviation from the model distribution, which consists of mixing the model distribution with another normal distribution with a mean far away from the mean of the model distribution, but with a small weight, which eventually tends to zero. This is the typical situation studied in the theory of robust statistics. Let us call it the admixture type of misspecification. It turns out that in this case the bias of the regression estimator, although going to zero when the weight of the additional mixture component tends to zero, does not do so in a flat way. Thus deviations of this kind have a stronger impact on the bias than deviations of the kind considered before. On the other hand, admixture type deviations are easier to detect and then can be taken care of in the estimation of the regression parameter, see Section 9.

In Section 2 we introduce the model in its simplest form, where the regressor distribution of the assumed model is just a normal distribution (i.e., it is a mixture of normals, with just one component), but the true model is a mixture of normals clustering round a single normal. Section 3 prepares for the SQS estimation method to be introduced in Sections 4 and for the computation of the SQS bias in Section 5. The behavior of this bias for small deviations from the true model and also for small measurement errors is studied in Section 6 and 7, respectively. When both, measurement error and deviation, are small, the bias is particularly small as found in Section 8. Section 9 generalizes these results to the case where the latent distribution of the assumed model consists of a mixture of normals. Section 10 studies the bias resulting from the admixture type of misspecification. Section 11 contains some simulation results and Section 12 some concluding remarks.

2 The model

A structural measurement error model consists of three parts: a regression model linking a scalar response variable y to a vector ξ of latent regressors, a measurement model linking the latent ξ to an observable vector x, and a model describing the distribution of ξ.

Here we assume that the regression model is given by conditional mean and variance functions:

$$\begin{aligned} \mathrm{E}(y|\xi) &= m^*(\xi, \beta) && (1) \\ \mathrm{V}(y|\xi) &= v^*(\xi, \beta, \gamma), && (2) \end{aligned}$$

where m^* and v^* are known functions of ξ and a parameter vector β to be estimated and v^* is additionally dependent an a variance parameter γ, which in some cases may be missing. These functions are supposed to be differentiable with respect to β and γ.

Examples are the polynomial model, see Cheng and Schneeweiss (2002), where

$$m^*(\xi, \beta) = \sum_{j=0}^{m} \beta_j \xi^j, \quad v^*(\xi, \beta, \gamma) = \gamma,$$

and the Poisson model, see Kukush et al (2001), Shklyar and Schneeweiss (2002), where

$$m^*(\xi, \beta) = v^*(\xi, \beta) = \exp(\beta_0 + \sum_{j=1}^{m} \beta_j \xi_j).$$

In a Poisson model with overdispersion we may have a variance function given by

$$v^*(\xi, \beta, \gamma) = \gamma \exp(\beta_0 + \sum_{j=1}^{m} \beta_j \xi_j).$$

We consider a "classical" measurement model, where

$$x = \xi + \delta \tag{3}$$

with $\delta \sim N(0, \Sigma)$ and δ being independent of ξ and y. The covariance matrix Σ of the measurement error δ is supposed to be known.

In Section 7 we want to study the properties of the model when Σ tends to zero. A convenient way to achieve this goal is to extract a common factor σ^2, writing

$$\Sigma = \sigma^2 \Omega,$$

and to let σ^2 tend to zero keeping Ω fixed.

Finally, ξ is taken to be a random vector with a parametric distribution. We have to distinguish between the (misspecified) model adopted by the statistician and the true model (the "data generating process"). We assume that the statistician erroneously adopts a normal distribution for ξ:

$$\xi \sim N(\mu_\xi, \Sigma_\xi), \tag{4}$$

which, because of its simplicity, turns out to be particularly suited for the ensuing analysis of the regression model. By contrast, the true model is a mixture of normals:

$$\xi \sim \sum_{k=1}^{K} p_k N(\mu_k, \Sigma_k), \tag{5}$$

with nonnegative weights p_k summing to 1 and p.d. covariance matrices Σ_k. Such a model is very flexible and can approximate many distributions that may turn up in practice.

Here, however, we assume that the mixture model comes close to a reference normal model $N(\mu_0, \Sigma_0)$, Σ_0 p.d., in the sense that

$$\begin{aligned} \mu_k &= \mu_0 + \kappa_k \vartheta + O(\vartheta^2) && (6) \\ \Sigma_k &= \Sigma_0 + \Lambda_k \vartheta + O(\vartheta^2) && (7) \end{aligned}$$

with some fixed vectors κ_k and symmetric matrices Λ_k and a scalar ϑ, which should be small and which later on we will let tend to zero. The O-functions are assumed to be differentiable. Thus the adopted model and the true model do not deviate too much from each other, and both come close to the reference model. Indeed, we have for the true model

$$\begin{aligned} \mu_\xi &= \sum_1^K p_k \mu_k = \mu_0 + \bar{\kappa}\vartheta + O(\vartheta^2) && (8) \\ \Sigma_\xi &= \sum_1^K p_k \Sigma_k + \sum_1^K p_k (\mu_k - \mu_\xi)(\mu_k - \mu_\xi)' \\ &= \Sigma_0 + \bar{\Lambda}\vartheta + O(\vartheta^2), && (9) \end{aligned}$$

where $\bar{\kappa} = \sum_1^K p_k \kappa_k$ and $\bar{\Lambda} = \sum_1^K p_k \Lambda_k$.

The adopted model, although differing in its distribution law from the true model, has the same parameters μ_ξ and Σ_ξ as the true model. These (nuisance) parameters can be consistently estimated directly from the data $x_i, i = 1, \cdots, n$, without recourse to the regression model:

$$\hat{\mu}_\xi = \bar{x}, \;\; \hat{\Sigma}_\xi = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})(x_i - \bar{x})' - \Sigma.$$

For simplicity we here assume that μ_ξ and Σ_ξ are known to the statistician. This assumption does not invalidate the ensuing results as the consistency properties of the SQS estimator do not change when the nuisance parameters are replaced with consistent estimates, see Kukush et al (2002).

Remark 1: ϑ is a dimensionless parameter, while κ_k and Λ_k take care of the dimensions in the components of x. It might seem inadequate and inconsistent to measure the deviations in (7) via Λ_k with squared and cross-product dimensions, whereas in (6) they are measured via κ_k in the same dimensions as the components of x. However an alternative specification of variance deviation like, e.g., $\Sigma_k^{\frac{1}{2}} = \Sigma_0^{\frac{1}{2}} + M_k \vartheta + O(\vartheta^2)$ will again lead to equation (7).

Remark 2: When ϑ is small, the components of the true model cluster round the reference distribution, which consists of only one component, just as the misspecified model. We therefore call this type of misspecification the clustering type of misspecification to distinguish it from another type to be considered in Section 10.

3 Conditioning

In order to explain the SQS procedure, but also to derive the bias of SQS, we need to condition the mean and variance of y on the observable regressor x.

Let I be an indicator variable taking the values $1, \cdots, K$ such that $I = k$ means that x comes from the k'th component of the mixture model. Then

$$p_k = P(I = k),$$

and

$$\xi | I = k \sim N(\mu_k, \Sigma_k).$$

We find from (1)

$$\begin{aligned} \mathrm{E}(y|x, I = k) &= \mathrm{E}[\mathrm{E}(y|x, \xi, I = k)|x, I = k] \\ &= \mathrm{E}[\mathrm{E}(y|\xi)|x, I = k] \\ &= \mathrm{E}[m^*(\xi, \beta)|x, I = k] \\ &=: m_k(x, \beta) =: m_k. \end{aligned} \tag{10}$$

Here we assumed implicitly that, given ξ, the additional knowledge of I has no influence on y, more precisely: $y|\xi, I \sim y|\xi$, and also that I is independent of δ. This, however, is not a model assumption. One can always construct an indicator variable I that has these properties. The variable I is a completely artificial variable, designed so as to describe the mixed distribution of ξ, without reflecting any real structure of the underlying population of ξ, see A1.

In order to be able to evaluate the conditional mean m_k of (10), we need to find the conditional distribution of ξ given x and given $I = k$. As (ξ', x') given $I = k$ is jointly normally distributed, the conditional distribution is normal, too, i.e.,

$$(\xi | x, I = k) \sim N(\mu_k(x), T_k)$$

with (see Shklyar and Schneeweiss (2002) and A2)

$$\mu_k(x) = x - \Sigma(\Sigma_k + \Sigma)^{-1}(x - \mu_k) \tag{11}$$

$$\begin{aligned} T_k &= \Sigma_k - \Sigma_k(\Sigma_k + \Sigma)^{-1}\Sigma_k \\ &= \Sigma - \Sigma(\Sigma_k + \Sigma)^{-1}\Sigma. \end{aligned} \tag{12}$$

It follows that the expression m_k of (10) is a function of $\mu_k(x)$ and T_k, which is the same for all k and will be designated by $\tilde{m}$:

$$m_k(x, \beta) = \tilde{m}(\mu_k(x), T_k, \beta). \tag{13}$$

The function $\tilde{m}$ is differentiable in all three arguments. In principle $m_k(x, \beta)$ is found by integrating $m^*(\xi, \beta)$ with respect to the c.d.f. $N(\mu_k(x), T_k)$. In several cases this integration can be easily carried out explicitly, leading to closed expressions for $m_k(x, \beta)$. The polynomial and the Poisson model of Section 2 are cases in point. Here, however, we do not need explicit formulas for m_k or for $\tilde{m}$.

The conditional distribution of ξ given x is a mixture of the conditional distribution of ξ given x and given $I = k$, $k = 1, \cdots, K$:

$$\xi | x \sim \sum_{k=1}^{K} \pi_k(x) N(\mu_k(x), T_k). \tag{14}$$

Here the weights $\pi_k(x)$ are the conditional (posterior) probabilities that $I = k$ pertains, given the observation x:

$$\pi_k(x) = P(I = k|x) = \frac{p_k \varphi_k(x)}{\sum_{j=1}^{K} p_j \varphi_j(x)}, \tag{15}$$

where $\varphi_j(x)$ is the density of $N(\mu_j, \Sigma_j + \Sigma)$.

It follows that

$$\mathrm{E}(y|x) = \sum_{1}^{K} \pi_k(x) m_k(x, \beta) =: \sum \pi_k m_k. \tag{16}$$

Let us now turn to the misspecified model (4). If (4) were true, the conditional distribution of ξ given x would be $N(\mu(x), T)$ with

$$\begin{aligned} \mu(x) &= x - \Sigma(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi) && (17) \\ T &= \Sigma_\xi - \Sigma_\xi(\Sigma_\xi + \Sigma)^{-1}\Sigma_\xi \\ &= \Sigma - \Sigma(\Sigma_\xi + \Sigma)^{-1}\Sigma, && (18) \end{aligned}$$

similar to (11) and (12), where μ_ξ and Σ_ξ come from (8) and (9), respectively. The conditional expectation of y given x, then, would be

$$\mathrm{E}_o(y|x) = \mathrm{E}_o[m^*(\xi,\beta)|x] =: m(x,\beta) = \tilde{m}(\mu(x), T, \beta), \tag{19}$$

with the same function $\tilde{m}$ as in (13), but with $\mu(x)$ and T in place of $\mu_k(x)$ and T_k, respectively. Similarly, the conditional variance of y given x under (4) is denoted by

$$\mathrm{V}_o(y|x) =: v(x,\beta,\gamma). \tag{20}$$

It is computed as

$$v(x,\beta,\gamma) = \mathrm{E}_o[v^*(\xi,\beta,\gamma)|x] + \mathrm{V}_o[m^*(\xi,\beta)|x].$$

The subscript o for E and V was chosen to designate the dependence on the misspecified distribution (4). This is the model, which is taken as the basis for constructing the SQS estimator.

4 The SQS estimator

Starting from the mean - variance model (19), (20) under the misspecified distribution (4), the SQS estimator of β and γ is defined as the solution to the system of estimating equations

$$\sum_{i=1}^{n} \left[\frac{y_i - m(x_i,\hat\beta)}{v(x_i,\hat\beta,\hat\gamma)} \; \frac{\partial m(x_i,\hat\beta)}{\partial\hat\beta} \right] = 0 \tag{21}$$

$$\sum_{i=1}^{n} \{ v(x_i,\hat\beta,\hat\gamma) - [y_i - m(x_i,\hat\beta)]^2 \} = 0, \tag{22}$$

where (x_i, y_i), $i = 1, \cdots, n$, is a sample of i.i.d. observations x_i, y_i having the distribution of the true model (1), (2), (3), and (5). It should be noted that other estimating equations instead of (22) are also possible, see Carroll et al (1995), but these would not change the subsequent results.

If (4) were true, the SQS estimators $\hat{\beta}$, $\hat{\gamma}$ would be consistent under rather general regularity assumptions, Kukush et al (2002). This follows essentially from the fact that after dividing (21) and (22) by n and going to the limit $n \to \infty$ the resulting equations

$$\begin{aligned} \mathrm{E}\left[\frac{y - m(x,\beta)}{v(x,\beta,\gamma)}\ \frac{\partial m(x,\beta)}{\partial \beta}\right] &= 0 && (23) \\ \mathrm{E}[v(x,\beta,\gamma)] - E[y - m(x,\beta)]^2\} &= 0, && (24) \end{aligned}$$

would be satisfied for the true parameter values β and γ. However, under the assumption that (4) is misspecified, as it is in our case, the solutions to (23) and (24) will not be the true β and γ any more and the estimators (i.e., the solutions to (21) and (22)) will be asymptotically biased. We want to investigate this bias.

Since we are here only interested in the regression parameter β, we can reduce the system (21), (22) to just one (vector-valued) equation for $\hat{\beta}$ only by solving (22) for $\hat{\gamma}$ given $\hat{\beta}$ and substituting the resulting solution $\hat{\gamma}(\hat{\beta})$ for $\hat{\gamma}$ in (21). If, for simplicity, we then change the notation by writing $v(x,\hat{\beta})$ for $v(x,\hat{\beta},\hat{\gamma}(\hat{\beta}))$, the estimating equation for $\hat{\beta}$ becomes

$$\sum_{i=1}^{n}\left[\frac{y_i - m(x_i,\hat{\beta})}{v(x_i,\hat{\beta})}\ \frac{\partial m(x_i,\hat{\beta})}{\partial \hat{\beta}}\right] = 0. \tag{25}$$

Its solution is the SQS estimator of β. Due to the misspecification of the underlying model, $\hat{\beta}$ will be (asymptotically) biased.

5 Bias

If we denote the probability limit plim$\hat{\beta}$ by b, the (asymptotic) bias of $\hat{\beta}$ is given by

$$B = b - \beta.$$

We find b by solving the limit ($n \to \infty$) of the estimating equation (25) divided by n; i.e., b is the solution to

$$\mathrm{E}\left[\frac{y - m(x,b)}{v(x,b)}\ \frac{\partial m(x,b)}{\partial b}\right] = 0, \tag{26}$$

which for short we may also write as

$$\mathrm{E}[(y-m)v^{-1}m_b] = 0,$$

where we introduced the abbreviations

$$m := m(x,b), \ \ v := v(x,b), \ \ m_b := \frac{\partial}{\partial b} m(x,b).$$

The expectation is to be taken with respect to the true model. Thus

$$\mathrm{E}[(y-m)v^{-1}m_b] = \mathrm{E}\{\mathrm{E}[(y-m)v^{-1}m_b|x]\}$$

and, by (16),

$$\begin{aligned} \mathrm{E}[(y-m)v^{-1}m_b] &= \mathrm{E}[(\sum_{k=1}^{K} \pi_k m_k - m)v^{-1}m_b] \\ &= \sum_{k=1}^{K} \mathrm{E}[\pi_k (m_k - m)v^{-1}m_b] \end{aligned} \tag{27}$$

because $\sum \pi_k = 1$.

Clearly, the right hand side of (27) is a function of ϑ, σ, and b, as its constituents are functions of these parameters. (We keep all the remaining parameters, like β, γ, μ_0, Σ_0, Ω, p_k κ_k, Λ_k fixed). Thus

$$\mathrm{E}[(y-m)v^{-1}m_b] = f(\vartheta, \sigma, b), \tag{28}$$

and f is differentiable. Equation (26) now becomes

$$f(\vartheta, \sigma, b) = 0$$

with a solution b which is a (differentiable) function of ϑ and σ:

$$b = b(\vartheta, \sigma).$$

The same is true for the Bias: $B = B(\vartheta, \sigma)$.

Clearly, the bias vanishes when $\vartheta = 0$, because then we have no misspecification and SQS becomes a consistent estimation procedure. Thus

$$b(0, \sigma) = \beta.$$

Similarly, $\sigma = 0$ implies a vanishing bias, because in this case the model has no measurement errors and SQS reduces to an ordinary quasi-score method to be applied to the model (1), (2), which then is a consistent procedure. Thus

$$b(\vartheta, 0) = \beta.$$

6 Small deviations

We want to study the bias $B(\vartheta, \sigma)$ near $\vartheta = 0$. We know that $B(0, \sigma) = 0$. To find out about the behavior of $B(\vartheta, \sigma)$ in a neighborhood of $\vartheta = 0$ we shall evaluate the derivative $\partial B/\partial\vartheta$ or, equivalently, $\partial b/\partial\vartheta$ at $\vartheta = 0$. By the implicit function theorem

$$\frac{\partial b}{\partial \vartheta} = -\left(\frac{\partial f}{\partial b'}\right)^{-1} \frac{\partial f}{\partial \vartheta}. \tag{29}$$

At $\vartheta = 0$, we have $b = \beta$, $\mu_\xi = \mu_k = \mu_0$, $\Sigma_\xi = \Sigma_k = \Sigma_0$, and consequently $\mu(x) = \mu_k(x) =: \mu_0(x)$, and $T = T_k =: T_0$, see (6) to (9), (11), (12), and (17), (18). It follows that

$$m(x, b) = m_k(x, \beta) = \tilde{m}(\mu_0(x), T_0, \beta) \tag{30}$$

at $\vartheta = 0$, see (13), (19). In addition

$$\pi_k(x) = p_k \tag{31}$$

at $\vartheta = 0$, see (15).

We want to evaluate (29) at $\vartheta = 0$. In the sequel all functions and their derivatives are taken at $\vartheta = 0$ and $b = \beta$. Now, by (28) and (27), noting that π_k and m_k are not functions of b,

$$\begin{aligned} \frac{\partial f}{\partial b'} &= -\mathrm{E}(v^{-1} m_b m_b') + \sum_{k=1}^{K} \mathrm{E}[\pi_k (m_k - m) \frac{\partial}{\partial b'}(v^{-1} m_b)] \\ &= -\mathrm{E}(v^{-1} m_b m_b') \end{aligned} \tag{32}$$

owing to (30). This matrix is negative definite.

The other factor in (29) is

$$\begin{aligned}\frac{\partial f}{\partial \vartheta} &= \mathrm{E}[(\sum_{k=1}^{K} \pi_k \frac{\partial m_k}{\partial \vartheta} - \frac{\partial m}{\partial \vartheta}) v^{-1} m_b] \\ &+ \sum_{k=1}^{K} \mathrm{E}[(m_k - m) \frac{\partial}{\partial \vartheta} (\pi_k v^{-1} m_b)].\end{aligned}$$

The last term, again, is zero owing to (30), and, due to (31), we get

$$\frac{\partial f}{\partial \vartheta} = \mathrm{E}[(\sum_{k=1}^{K} p_k \frac{\partial m_k}{\partial \vartheta} - \frac{\partial m}{\partial \vartheta}) v^{-1} m_b]. \tag{33}$$

Now by (13),

$$\frac{\partial m_k}{\partial \vartheta} = \frac{\partial \tilde{m}}{\partial \mu_k'(x)} \frac{\partial \mu_k(x)}{\partial \vartheta} + tr(\frac{\partial \tilde{m}}{\partial T_k} \frac{\partial T_k}{\partial \vartheta}).$$

According to (30) we may also write, at $\vartheta = 0$:

$$\frac{\partial m_k}{\partial \vartheta} = \frac{\partial \tilde{m}}{\partial \mu_0'(x)} \frac{\partial \mu_k(x)}{\partial \vartheta} + tr(\frac{\partial \tilde{m}}{\partial T_0} \frac{\partial T_k}{\partial \vartheta}). \tag{34}$$

Similarly, by (19), and (30),

$$\frac{\partial m}{\partial \vartheta} = \frac{\partial \tilde{m}}{\partial \mu_0'(x)} \frac{\partial \mu(x)}{\partial \vartheta} + tr(\frac{\partial \tilde{m}}{\partial T_0} \frac{\partial T}{\partial \vartheta}). \tag{35}$$

Furthermore, by (11), see also A2,

$$\begin{aligned}\frac{\partial \mu_k(x)}{\partial \vartheta} &= \Sigma(\Sigma_k + \Sigma)^{-1} \frac{\partial \Sigma_k}{\partial \vartheta} (\Sigma_k + \Sigma)^{-1} (x - \mu_k) \\ &+ \Sigma(\Sigma_k + \Sigma)^{-1} \frac{\partial \mu_k}{\partial \vartheta},\end{aligned}$$

see Dhrymes (1984), especially Corollary 41, p.125, for the rules of differentiating matrices. To continue: by (6) and (7), at $\vartheta = 0$,

$$\frac{\partial \mu_k(x)}{\partial \vartheta} = \Sigma(\Sigma_0 + \Sigma)^{-1} \Lambda_k (\Sigma_0 + \Sigma)^{-1} (x - \mu_0) + \Sigma(\Sigma_0 + \Sigma)^{-1} \kappa_k. \tag{36}$$

Similarly, by (17), (8), and (9),

$$\frac{\partial \mu(x)}{\partial \vartheta} = \Sigma(\Sigma_0 + \Sigma)^{-1} \bar{\Lambda} (\Sigma_0 + \Sigma)^{-1} (x - \mu_0) + \Sigma(\Sigma_0 + \Sigma)^{-1} \bar{\kappa}. \tag{37}$$

Finally, by (12) and (7), see also A2,

$$\frac{\partial T_k}{\partial \vartheta} = \Sigma(\Sigma_0 + \Sigma)^{-1}\Lambda_k(\Sigma_0 + \Sigma)^{-1}\Sigma, \tag{38}$$

and, by (18) and (9),

$$\frac{\partial T}{\partial \vartheta} = \Sigma(\Sigma_0 + \Sigma)^{-1}\bar{\Lambda}(\Sigma_0 + \Sigma)^{-1}\Sigma. \tag{39}$$

Collecting terms and using the definitions of $\bar{\Lambda}$ and $\bar{\kappa}$ just after (9), we see that (33) to (39) imply

$$\frac{\partial f}{\partial \vartheta} = 0 \tag{40}$$

at $\vartheta = 0$. Finally (29), (32), and (40) imply

$$\frac{\partial b}{\partial \vartheta} = 0$$

at $\vartheta = 0$ and consequently also $\frac{\partial B}{\partial \vartheta} = 0$. We thus have proved the following theorem:

Theorem 1

The bias B of the SQS estimator $\hat{\beta}$ in the model (1), (2), (3), (5), computed under the misspecified model (4), tends to zero, when the deviation of the misspecified model from the true model as measured by the parameter ϑ tends to zero, in such a way that

$$\frac{\partial B}{\partial \vartheta} = 0$$

at $\vartheta = 0$.

7 Small measurement errors

We have a similar theorem when $\sigma \to 0$, when again the bias goes to zero in a flat way. But the arguments leading to this result are somewhat different.

Theorem 2

Under the conditions of theorem 1 and for any fixed misspecification, the bias *of $\hat{\beta}$ tends to zero, when the measurement error variances tend to zero, in such a way that*

$$\frac{\partial B}{\partial \sigma} = 0$$

at $\sigma = 0$.

Proof: Similar to (29) we have

$$\frac{\partial b}{\partial \sigma^2} = -(\frac{\partial f}{\partial b'})^{-1} \frac{\partial f}{\partial \sigma^2}. \tag{41}$$

For $\sigma^2 = 0$ and therefore $\Sigma = 0$ we have $b = \beta$, as noted at the end of Section (5); furthermore $\mu(x) = \mu_k(x) = x = \xi$ and $T = T_k = 0$ for $\sigma^2 = 0$, see (11), (12), and (17), (18). It follows that

$$m(x, b) = m_k(x, \beta) = \tilde{m}(x, 0, \beta) \tag{42}$$

at $\sigma^2 = 0$, see (13), (19). This is similar to (30); however (31) is not valid any more if $\sigma^2 = 0$, but $\vartheta \neq 0$. In the sequel functions and their derivatives are taken at $\sigma^2 = 0$.

Just as in (33) we have, owing to (42),

$$\frac{\partial f}{\partial \sigma^2} = \mathrm{E}[(\sum_{k=1}^{K} \pi_k \frac{\partial m_k}{\partial \sigma^2} - \frac{\partial m}{\partial \sigma^2}) v^{-1} m_b]$$

and, similar to (34) and (35),

$$\begin{aligned} \frac{\partial m_k}{\partial \sigma^2} &= \frac{\partial \tilde{m}}{\partial \mu'(x)} \frac{\partial \mu_k(x)}{\partial \sigma^2} + tr(\frac{\partial \tilde{m}}{\partial T} \frac{\partial T_k}{\partial \sigma^2}) \\ \frac{\partial m}{\partial \sigma^2} &= \frac{\partial \tilde{m}}{\partial \mu'(x)} \frac{\partial \mu(x)}{\partial \sigma^2} + tr(\frac{\partial \tilde{m}}{\partial T} \frac{\partial T}{\partial \sigma^2}). \end{aligned}$$

But now by (11), (12) and (17), (18), see A2,

$$\begin{aligned} \frac{\partial \mu_k(x)}{\partial \sigma^2} &= -\Omega \Sigma_k^{-1} (x - \mu_k) \\ \frac{\partial \mu(x)}{\partial \sigma^2} &= -\Omega \Sigma_\xi^{-1} (x - \mu_\xi) \\ \frac{\partial T_k}{\partial \sigma^2} &= \frac{\partial T}{\partial \sigma^2} = \Omega. \end{aligned}$$

Collecting terms, we get

$$\frac{\partial f}{\partial \sigma^2} = \mathrm{E}\{\frac{\partial \tilde{m}}{\partial \mu'(x)}\Omega[\Sigma_\xi^{-1}(x-\mu_\xi) - \sum_{k=1}^{K} \pi_k \Sigma_k^{-1}(x-\mu_k)]v^{-1}m_b\}$$

at $\sigma^2 = 0$. We cannot say (as we could with $\partial f/\partial \vartheta$) that this term is zero. However, we may state that $|\partial f/\partial \sigma^2| < \infty$ at $\sigma^2 = 0$, and so, by (41),

$$|\frac{\partial b}{\partial \sigma^2}| < \infty.$$

It then follows that

$$\frac{\partial b}{\partial \sigma} = \frac{\partial b}{\partial \sigma^2}\frac{d\sigma^2}{d\sigma} = \frac{\partial b}{\partial \sigma^2} \cdot 2\sigma = 0$$

at $\sigma = 0$, which proves Theorem 2.

8 Small deviations and small measurement errors

In this section we want to study the behavior of the bias function $B(\vartheta, \sigma)$ when both ϑ and σ tend to zero. Let us denote the first and second derivatives of B with respect to ϑ and σ by using ϑ and σ as subscripts. The results of Theorem 1 and 2 can then be summarized as

$$\begin{aligned} B_\vartheta(0,\sigma) &= 0 \\ B_\sigma(\vartheta,0) &= 0, \end{aligned}$$

from which follows that also

$$\begin{aligned} B_{\vartheta\sigma}(0,\sigma) &= B_{\vartheta\sigma\sigma}(0,\sigma) = 0 \\ B_{\sigma\vartheta}(\vartheta,0) &= B_{\sigma\vartheta\vartheta}(\vartheta,0) = 0. \end{aligned}$$

On the other hand, $B(0,\sigma) = 0$, which implies

$$B_\sigma(0,\sigma) = B_{\sigma\sigma}(0,\sigma) = B_{\sigma\sigma\sigma}(0,\sigma) = 0.$$

Similarly, $B(\vartheta, 0) = 0$, and thus

$$B_\vartheta(\vartheta, 0) = B_{\vartheta\vartheta}(\vartheta, 0) = B_{\vartheta\vartheta\vartheta}(\vartheta, 0) = 0.$$

A Taylor expansion of $B(\vartheta, \sigma)$ at $(0, 0)$ results in

$$B(\vartheta, \sigma) = B_{\sigma\sigma\vartheta\vartheta}(0, 0)\vartheta^2\sigma^2 + \text{terms of higher order}.$$

Suppose ϑ and σ tend to zero simultaneously in such a way that $\vartheta = a\lambda$, $\sigma = b\lambda$ and $\lambda \to 0$. Then

$$\frac{\partial B}{\partial \lambda} = \frac{\partial^2 B}{\partial \lambda^2} = \frac{\partial^3 B}{\partial \lambda^3} = 0$$

at $\lambda = 0$. In other words, $B(\lambda) = O(\lambda^4)$. This means that the bias tends to zero in an extremely flat way if both ϑ and σ become small.

9 Mixture of normals for the ξ-distribution

Up to now we assumed that the true distribution of ξ was a mixture of normals, the components of which, however, come close to one fundamental normal distribution, and this distribution is taken by the statistician as the model basis for estimating the parameter vector β. Indeed, when the components of the mixture cluster round one main component, it will typically be difficult for the statistician to even recognize that a mixture is present let alone to distinguish between its various components. In this situation, a misspecification as considered in this paper can easily come up, leading to a bias in the estimation of β. But, as we have seen, this bias will most often be negligible.

However, if the true distribution of ξ is a mixture with clearly distinguishable components, i.e., with means and/or variances for apart, the bias resulting from the use of a misspecified model with only one component can be very large, and our Theorem 1 does not apply. On the other hand, in such a situation the statistician will easily recognize the presence of a mixture distribution, even though the distribution of ξ is latent. The statistician can "observe" the distribution of x and this will be a similar mixture of normals as the distribution of ξ except that the component covariance matrices will be enlarged by the addition of the measurement error covariance matrix Σ.

Let us therefore now assume that the statistician recognizes a mixture of normals for the distribution of ξ so that his (more flexible) model, instead of (4), becomes

$$\xi \sim \sum_{k=1}^{K} p_k N(\mu_{k\xi}, \Sigma_{k\xi}).$$

But again this model is misspecified. The true model is a more complicated (and even more flexible) mixture of normals

$$\xi \sim \sum_{k=1}^{K} \sum_{j=1}^{J_k} p_{kj} N(\mu_{kj}, \Sigma_{kj}),$$

where the components cluster round the K components of a "reference" mixture distribution $\sum_{k=1}^{K} p_k N(\mu_k, \Sigma_k)$ with $p_k = \sum_{j=1}^{J_k} p_{kj}$, i.e.,

$$\begin{aligned} \mu_{kj} &= \mu_k + \kappa_{kj}\vartheta + O(\vartheta^2) \\ \Sigma_{kj} &= \Sigma_k + \Lambda_{kj}\vartheta + O(\vartheta^2) \end{aligned}$$

and ϑ is small.

The SQS estimator of β can now be constructed in a similar way as before. Conditioning ξ on x results in the mixture model

$$\xi | x \sim \sum_{k=1}^{K} \sum_{j=1}^{J_k} \pi_{kj}(x) N(\mu_{kj}(x), T_{kj})$$

with

$$\begin{aligned} \mu_{kj}(x) &= x - \Sigma(\Sigma_{kj} + \Sigma)^{-1}(x - \mu_{kj}) \\ T_{kj} &= \Sigma - \Sigma(\Sigma_{kj} + \Sigma)^{-1}\Sigma \\ \pi_{kj}(x) &= \frac{p_{kj}\varphi_{kj}(x)}{\sum_{l=1}^{K} \sum_{h=1}^{J_k} p_{lh}\varphi_{lh}(x)}, \end{aligned}$$

$\varphi_{lh}(x)$ being the p.d.f. of $N(\mu_{lh}, \Sigma_{lh} + \Sigma)$, and consequently

$$\mathrm{E}(y|x) = \sum_k \sum_j \pi_{kj}(x) m_{kj}(x, \beta),$$

where

$$\begin{aligned} m_{kj}(x, \beta) &= \mathrm{E}[m^*(\xi, \beta)|x, I = (k, j)] \\ &= \tilde{m}(\mu_{kj}(x), T_{kj}, \beta) \end{aligned}$$

with the same function $\tilde{m}$ as in (13).

Starting from the misspecified model, however, the statistician will work with the conditional expectation

$$\mathrm{E}_0(y|x) = \sum_{k=1}^{K} \pi_k(x) m_k(x, \beta),$$

where

$$\pi_k(x) = \frac{p_k \varphi_{k\xi}(x)}{\sum_{h=1}^{K} p_h \varphi_{h\xi}(x)},$$

$\varphi_{h\xi}(x)$ being the p.d.f. of $N(\mu_{h\xi}, \Sigma_{h\xi} + \Sigma)$, and

$$m_k(x, \beta) = \tilde{m}(\mu_{k\xi}(x), T_{k\xi}, \beta)$$

with

$$\begin{aligned} \mu_{k\xi}(x) &= x - \Sigma(\Sigma_{k\xi} + \Sigma)^{-1}(x - \mu_{k\xi}) \\ T_{k\xi} &= \Sigma - \Sigma(\Sigma_{k\xi} + \Sigma)^{-1}\Sigma \end{aligned}$$

and

$$\begin{aligned} \mu_{k\xi} &= \sum_{j=1}^{J_k} p_{kj} \mu_{kj} / p_k \\ \Sigma_{k\xi} &= \sum_{j=1}^{J_k} p_{kj} \Sigma_{kj} / p_k + \sum_{j=1}^{J_k} p_{kj} (\mu_{kj} - \mu_{k\xi})(\mu_{kj} - \mu_{k\xi})' / p_k. \end{aligned}$$

The conditional variance of y given x computed under the misspecified model is denoted by $\bar{v}(x, \beta, \gamma)$. It is given by

$$\bar{v}(x, \beta, \gamma) = \sum_{k=1}^{\mathrm{K}} \pi_k(x) v_k(x, \beta, \gamma) + \sum_{k=1}^{\mathrm{K}} \pi_k(x) m_k^2(x, \beta) - [\sum_{k=1}^{\mathrm{K}} \pi_k(x) m_k(x, \beta)]^2$$

with $v_k(x, \beta, \gamma)$ computed in a similar way as $v(x, \beta, \gamma)$ in Section 3 - just replace $\mu(x)$ and T of (17) and (18) with $\mu_{k\xi}(x)$ and $T_{k\xi}$, respectively. The SQS estimator is now found as the solution to the equations

$$\begin{aligned} \sum_{i=1}^{n} \left[\frac{y_i - \sum_{k=1}^{K} \pi_k(x_i) m_k(x_i, \hat{\beta})}{\bar{v}(x_i, \hat{\beta}, \hat{\gamma})} \frac{\partial \sum_{k=1}^{K} \pi_k(x_i) m_k(x_i, \hat{\beta})}{\partial \hat{\beta}} \right] &= 0 \\ \sum_{i=1}^{n} \left[\bar{v}(x_i, \hat{\beta}, \hat{\gamma}) - \{y_i - \sum_{k=1}^{K} \pi_k(x_i) m_k(x_i, \hat{\beta})\}^2 \right] &= 0. \end{aligned}$$

The probability limit of $\hat{\beta}$, b =plim$\hat{\beta}$, is found as the solution to

$$\mathrm{E} \left[\frac{y - \sum_{k=1}^{K} \pi_k(x) m_k(x, b)}{\bar{v}(x, b, \gamma(b))} \frac{\partial \sum_{k=1}^{K} \pi_k(x) m_k(x, b)}{\partial b} \right] = 0$$

where the expectation is taken with respect to the true model. The solution b is a function of ϑ and σ.

By arguments very similar to those in Sections 6 and 7 one finds that Theorems 1 and 2 hold true even in this more general setting. The same applies to the results of Section 8.

10 Admixtures

In this section, we study another kind of misspecification of the ξ-distribution. We return to the situation of Section 2, where the assumed model consists of just one (normal) component, see (4), whereas now the true model is a mixture of two normals, one being the main component with high weight $1-p$ and the other one being an admixture to the main component with low weight p. We keep the two components fixed and let p tend to zero. In doing so, we study the bias of the SQS estimator as a function of p.

The model now is given by (5) with $K = 2$ components, where $p_1 = 1-p$ and $p_2 = p$. The assumed (misspecified) model is again given by (4).

The definitions of $\mu(x)$, T, $\mu_k(x)$, T_k, $k = 1, 2$ are the same as before, see (17), (18), (11), (12), except that now μ_ξ and Σ_ξ are functions of p.

$$\begin{aligned} \mu_\xi &= (1-p)\mu_1 + p\mu_2 & (43) \\ \Sigma_\xi &= (1-p)\Sigma_1 + p\Sigma_2 + (1-p)p(\mu_1 - \mu_2)(\mu_1 - \mu_2)'. & (44) \end{aligned}$$

Similarly π_1 and π_2 are functions of p, in particular, see (15),

$$\pi_2(x) = \frac{p\varphi_2(x)}{(1-p)\varphi_1(x) + p\varphi_2(x)}. \tag{45}$$

The bias is found as before by solving (26), and (26) can again be transformed into (27), which here becomes

$$g(p, b) = \mathrm{E}[(\pi_1 m_1 + \pi_2 m_2 - m)v^{-1}m_b]. \tag{46}$$

Note that, because the right hand side of (27) is now to be considered as a function of p, we used the notation $g(p, b)$ instead of $f(\vartheta, \sigma, b)$. The symbols m, m_b, v, and m_k, $k = 1, 2$, have the same meaning as before, in particular we will use again (13) and (19), the latter with $\beta = b$, as before.

Now, since b is the solution to $g(p, b) = 0$, b is a function of p and its derivative with respect to p is given by

$$\frac{db}{dp} = -\left(\frac{\partial g}{\partial b'}\right)^{-1} \frac{\partial g}{\partial p}. \tag{47}$$

We want to evaluate db/dp at $p = 0$. In the sequel all functions and their derivatives are to be taken at $p = 0$ and $b = \beta$.

We note that as $p \to 0$:

$$\begin{aligned} \mu_\xi \to \mu_1 \quad &, \quad \Sigma_\xi \to \Sigma_1, \\ \mu(x) \to \mu_1(x) \quad &, \quad T \to T_1, \\ m \to m_1 \quad &, \quad \pi_2 \to 0. \end{aligned} \tag{48}$$

Let us first find $\partial g/\partial b'$ from (44):

$$\begin{aligned}\frac{\partial g}{\partial b'} &= \mathrm{E}(v^{-1}m_b m_b') \\ &+ \mathrm{E}[\pi_1(m_1-m)\frac{\partial(v^{-1}m_b)}{\partial b'}] + \mathrm{E}[\pi_2(m_2-m)\frac{\partial(v^{-1}m_b)}{\partial b'}],\end{aligned}$$

which by (48) simplifies to

$$\frac{\partial g}{\partial b'} = -\mathrm{E}(v^{-1}m_b m_b'), \tag{49}$$

and this is negative definite, see also (32).

Now as to the derivative with respect to p, we have

$$\frac{\partial g}{\partial p} = \mathrm{E}\left[\left(\frac{\partial \pi_1}{\partial p}m_1 + \frac{\partial \pi_2}{\partial p}m_2 - \frac{\partial m}{\partial p}\right)v^{-1}m_b\right].$$

plus two more terms, which however turn out to be zero at $p=0$, just as in the derivation of (49). Because $\pi_1 = 1-\pi_2$, this simplifies to

$$\frac{\partial g}{\partial p} = \mathrm{E}\left[\left(\frac{\partial \pi_2}{\partial p}(m_2-m_1) - \frac{\partial m}{\partial p}\right)v^{-1}m_b\right]. \tag{50}$$

Let us evaluate the various derivatives one by one. From (45) we see that

$$\frac{\partial \pi_2}{\partial p} = \frac{\varphi_2}{\varphi_1} \tag{51}$$

at $p=0$. Furthermore, by (19),

$$\frac{\partial m}{\partial p} = \frac{\partial \tilde{m}}{\mu'(x)}\frac{\mu(x)}{\partial p} + tr\left(\frac{\partial \tilde{m}}{\partial T}\frac{\partial T}{\partial p}\right), \tag{52}$$

where, by (17), see also A2,

$$\begin{aligned}\frac{\partial \mu(x)}{\partial p} &= \Sigma(\Sigma_\xi+\Sigma)^{-1}\frac{\partial \Sigma_\xi}{\partial p}(\Sigma_\xi+\Sigma)^{-1}(x-\mu_\xi) \\ &+ \Sigma(\Sigma_\xi+\Sigma)^{-1}\frac{\partial \mu_\xi}{\partial p}\end{aligned}$$

and, by (18) and (48), see also A2,

$$\frac{\partial T}{\partial p} = \Sigma(\Sigma_1+\Sigma)^{-1}\frac{\partial \Sigma_\xi}{\partial p}(\Sigma_1+\Sigma)^{-1}\Sigma.$$

from (43) and (44) we find, at $p = 0$,

$$\begin{aligned}
\frac{\partial \mu_\xi}{\partial p} &= \mu_2 - \mu_1 \\
\frac{\partial \Sigma_\xi}{\partial p} &= \Sigma_2 - \Sigma_1 + (\mu_1 - \mu_2)(\mu_1 - \mu_2)'.
\end{aligned}$$

Substituting these last expressions in $\partial\mu(x)/\partial p$ and $\partial T/\partial p$ and using (48), we get

$$\frac{\mu(x)}{\partial p} = \Sigma\Delta(x - \mu_\xi) + \Sigma(\Sigma_1 + \Sigma)^{-1}(\mu_2 - \mu_1) \tag{53}$$

$$\frac{\partial T}{\partial p} = \Sigma\Delta\Sigma, \tag{54}$$

where

$$\Delta := (\Sigma_1 + \Sigma)^{-1}[(\Sigma_2 - \Sigma_1) + (\mu_1 - \mu_2)(\mu_1 - \mu_2)'](\Sigma_1 + \Sigma)^{-1}$$

is a matrix which, in a sense, measures the distance of the two components of the model. Substituting (53) and (54) in (52) and (52) and (51) in (50), we get

$$\begin{aligned}
\frac{\partial g}{\partial p} &= \mathrm{E}[\frac{\varphi_2}{\varphi_1}(m_2 - m_1)v^{-1}m_b] \\
&- \mathrm{E}[\frac{\partial \tilde{m}}{\partial \mu'(x)}\{\Sigma\Delta(x - \mu_1) - \Sigma(\Sigma_1 + \Sigma)^{-1}(\mu_2 - \mu_1)\}v^{-1}m_b] \\
&- \mathrm{E}[tr\left(\frac{\partial \tilde{m}}{\partial T}\Sigma\Delta\Sigma\right)v^{-1}m_b].
\end{aligned}$$

The expectations are to be taken with respect to $N(\mu_1, \Sigma_1 + \Sigma)$, as this is the distribution of x at $p = 0$. The first term on the right hand side can also be written as

$$\mathrm{E}_2[(m_2 - m_1)v^{-1}m_b],$$

where E_2 is the expectation with respect to $N(\mu_2, \Sigma_2 + \Sigma))$. Clearly $\partial g/\partial p$ does generally not vanish, and so in general

$$\frac{db}{dp} \neq 0$$

at $p = 0$.

Remark 3: It should be noted that an exception is the linear model $y = \beta_0 + \beta_1\xi + u$. For the linear model, the bias of SQS is always zero no matter whether the correct distribution of ξ or a misspecified one is used. Therefore in this case, $db/dp = 0$. In all other cases db/dp is a complicated function of $(\mu_2 - \mu_1)$ and $\Sigma_2 - \Sigma_1$.

Remark 4: Letting $\Sigma_2 = 0$, the resulting expression db/dp as a function of μ_2 is nothing but the influence function of the bias at the underlying normal distribution $N(\mu_1, \Sigma_1)$, see Hampel et al (1986).

Remark 5: One might object against the use of p as a measure of deviation from the true model. Why not use $\sqrt{p}$, for instance ? It would seem plausible to use a measure with the same dimension as x, e.g., $e := \mu_\xi - \mu_1$. But this measure is proportional to p and would lead to the same result: $db/de \neq 0$ at $e = 0$. This justifies the use of p as a measure of deviation in the case of the admixture type of misspecification.

11 Simulation

Simulation results showing the effect of the clustering type of misspecification can be found in Schneeweiss et al (2003). We here repeat only one result using however a somewhat different model. We also study two examples of the admixture type of misspecification.

The regression model of our simulation study is a polynomial of degree 2:

$$y = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \epsilon$$

with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = -0.5$, and $\sigma_\epsilon^2 = 1$. The measurement error variance is $\sigma_\delta^2 = 0.2$. The misspecified model for the distribution of ξ is $N(\mu_\xi, \sigma_\xi^2)$.

For the clustering type of misspecification the true distribution (model 1) is the mixture

$$\xi \sim 0.5 \cdot N(0,1) + 0.5 \cdot N(\vartheta, 1-\vartheta)$$

and $N(0,1)$ is the reference distribution. We let ϑ go to zero at steps of 0.05 starting from $\vartheta = 0.5$

For the admixture type of misspecification we tried two distributions for the true model. In the first (model 2)

$$\xi \sim (1-p)N(0,1) + pN(3,1)$$

and in the second (model 3)

$$\xi \sim (1-p)N(0,1) + pN(0,3).$$

In both cases, $N(0,1)$ is again the reference distribution ($p = 0$). We let p go to zero at steps of 0.01 starting from $p = 0.1$.

We chose a sample size $n = 500$, large enough so that asymptotic results apply. The number of replications was $N = 1000$.

Table 1 shows the bias of the SQS estimations of β_0, β_1, and β_2 in model 1 for various values of ϑ. In the corresponding Figure 1 one can see that the bias, in all three cases tends to 0 in a flat way when $\vartheta \to 0$.

Tables 2 and 3 show the bias for models 2 and 3, respectively, as a function of p. Figures 2 and 3 show that the bias tends to 0 when $p \to 0$, but not in a flat way.

These results are in full accordance with the theoretical findings of Sections 6 and 10.

12 Conclusion

The structural quasi-score estimation method in a measurement error model starts from the conditional mean and variance functions of the response variable given the observable regressor. In order to be able to obtain these functions a distribution of the regressor must be specified. Clearly, when the statistician assumes a misspecified distribution, the resulting SQS estimator will be bias and the bias will depend on the degree of misspecification. Two types of misspecification have been considered.

In the first type the statistician assumes erroneously a single normal distribution, whereas the true distribution is a (finite) mixture of normals clustering round this single normal. (More generally the assumed distribution can be a mixture of normals and the true distribution a mixture of submixtures each of which cluster round one of the components of the assumed mixture.)

If the cluster comes very close to the assumed distribution, the bias will hardly be perceptible. More precisely, if the bias is measured as a function of a suitable measure of deviation ϑ between cluster and single component, the derivative of the bias with respect to ϑ is zero at the point $\vartheta = 0$.

This result is very handy for practical purposes. It says that in situations where the statistician has difficulties in making out the true distribution because it consists of several components clustering closely round one single component, the assumption of a simplified model by the statistician is of no great harm as the resulting estimation bias will be negligible.

There is, however, another type of misspecification, which occurs when the statistician again assumes a normal distribution but the true distribution consists of the assumed distribution admixed with another component (or, more generally, several components) far away from the assumed one, but with low weight. One can hope that due to the low weight the estimation bias will be small, but it turns out that it will rise steeply with growing weight p of the admixture so that the bias even for small p , though small, will most often not be negligible. Technically speaking: The derivative of the bias with respect to p is generally not zero at $p = 0$.

Fortunately, this more critical situation, where the bias reacts sensitively on misspecification, can be much better mastered by the statistician than the former case of the cluster type of misspecification. Admixtures with components markedly different from the assumed normal distribution can often be easily detected even if the weights of those admixture components should be small.

The essential implication of these results is that SQS seems to work nicely as far as a possible bias of SQS is concerned, as long as the statistician takes care in modelling the distribution of the latent regressor as accurately as possible. Slight misspecifications that might be difficult to avoid are of no great danger.

On the other hand, as far as efficiency is concerned, SLS is probably superior to other methods that do not take the regressor distribution into account. One such method is corrected score (CS); for CS applied to polynomials see Cheng and Schneeweiss (1998) and Cheng et al (2000). Although one can show that CS and SQS are practically of equal efficiency when the measurement error variance is small, c.f. Kukush et al (2002), for larger measurement errors SQS is likely to be more efficient than CS. An example is the Poisson regression model, c.f. Shklyar and Schneeweiss (2002).

This is another arguments which points to SQS as the preferred estimation method. However, SQS sometimes may be difficult to apply. An approximate, easy to apply, method, which is also based on knowledge of the regressor distribution, is Regression Calibration (RC), c.f. Carroll et al (1995).

In Table 4 and Figure 4, the bias from a simulation study of five estimators of β_1 from model 1 is shown as a function of ϑ: naive, CS, SQS, RC, and MSQS. The latter is SQS applied to a mixture of normals, as described in Section 9. The naive method is biased even for $\vartheta = 0$. CS and NSQS have hardly any bias (but are less efficient than SQS). SQS and RC behave similarly in this case (but not in others).

Acknowledgement: The authors want to thank Jia-Ren Tsai and Roland Wolf for their technical help in setting up the simulation study.

Appendix

A1. The indicator variable I

Let Ξ be the sample space of the random variable ξ with p.d.f. $f(\xi) = \sum_{k=1}^{K} f_k(\xi)p_k$, and let $\mathcal{K} = \{1, \cdots, k\}$. Let

$$p_k(\xi) = \frac{f_k(\xi)p_k}{f(\xi)}.$$

$p_k(\xi)$ is a probability function on $\mathcal{K}$ depending on ξ. Define a probability measure on $\Xi \times \mathcal{K}$ and implicitly a random variable I with values in $\mathcal{K}$ by letting

$$P(\xi \leq x, I = k) = F(\xi)p_k(\xi),$$

where $F(\xi)$ is the c.d.f. of ξ. Then the marginal distribution of I is given by

$$P(I = k) = \int f(\xi)p_k(\xi)d\xi = p_k.$$

and the conditional distribution of ξ given I by

$$f(\xi|I = k) = \frac{f(\xi)p_k(\xi)}{p_k} = f_k(\xi).$$

Thus I is an indicator variable for the components of the distribution of ξ. As by construction I depends only on ξ and ξ is independent of δ, so also I is independent of δ. Again because I depends only on ξ, we have $y|\xi, I \sim y|\xi$.

A2. The conditional distribution of ξ given x

Let $x = \xi + \delta$ as in (3) with $(\xi', \delta') \sim N(\mu_*, \Sigma_*)$, where $\mu_* = (\mu_\xi', 0')$ and $\Sigma_* =$ block diagonal (Σ_ξ, Σ). Then $\xi|x \sim N(\mu(x), T)$, where $\mu(x)$ and T can be found from the linear regression

$$\xi - \mu_\xi = B(x - \mu_x) + \epsilon,$$

where ϵ is independent of x. Obviously

$$B = \Sigma_{\xi x}\Sigma_x^{-1}.$$

Because of (3),

$$B = \Sigma_\xi(\Sigma_\xi + \Sigma)^{-1}.$$

Thus, because $\mu_x = \mu_\xi$,

$$\begin{aligned}\mu(x) &= \mu_\xi + \Sigma_\xi(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi) \\ &= x - \Sigma(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi),\end{aligned}$$

which proves (17) and by analogy (11).

From the first equation

$$\begin{aligned}\frac{\partial\mu(x)}{\partial\sigma^2} &= -\Sigma_\xi(\Sigma_\xi + \Sigma)^{-1}\Omega(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi) \\ &= -\Omega\Sigma_\xi^{-1}(x - \mu_\xi)\end{aligned}$$

at $\sigma^2 = 0$. Similarly from the last equation

$$\begin{aligned}\frac{\partial\mu(x)}{\partial\vartheta} &= \Sigma(\Sigma_\xi + \Sigma)^{-1}\frac{\partial\Sigma_\xi}{\partial\vartheta}(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi) \\ &+ \Sigma(\Sigma_\xi + \Sigma)^{-1}\frac{\partial\mu_\xi}{\partial\vartheta}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\mu(x)}{\partial p} &= \Sigma(\Sigma_\xi + \Sigma)^{-1}\frac{\partial\Sigma_\xi}{\partial p}(\Sigma_\xi + \Sigma)^{-1}(x - \mu_\xi) \\ &+ \Sigma(\Sigma_\xi + \Sigma)^{-1}\frac{\partial\mu_\xi}{\partial p}.\end{aligned}$$

Finally,

$$\begin{aligned}T &= \mathrm{E}\epsilon\epsilon' = \Sigma_\xi - B\Sigma_x B' \\ &= \Sigma_\xi - \Sigma_\xi(\Sigma_\xi + \Sigma)^{-1}\Sigma_\xi \\ &= \Sigma - \Sigma(\Sigma_\xi + \Sigma)^{-1}\Sigma,\end{aligned}$$

which proves (18) and by analogy (12).

Clearly from the next to last equation

$$\begin{aligned}\frac{\partial T}{\partial \sigma^2} &= \Sigma_\xi(\Sigma_\xi+\Sigma)^{-1}\Omega(\Sigma_\xi+\Sigma)^{-1}\Sigma_\xi \\ &= \Omega\end{aligned}$$

at $\sigma^2 = 0$. Similarly from the last equation

$$\begin{aligned}\frac{\partial T}{\partial \vartheta} &= \Sigma(\Sigma_\xi+\Sigma)^{-1}\frac{\partial \Sigma_\xi}{\partial \vartheta}(\Sigma_\xi+\Sigma)^{-1}\Sigma \\ \frac{\partial T}{\partial p} &= \Sigma(\Sigma_\xi+\Sigma)^{-1}\frac{\partial \Sigma_\xi}{\partial p}\Sigma_\xi+\Sigma)^{-1}\Sigma.\end{aligned}$$

We used the differentiation rule for a nonsingular matrix $A(x)$ being a differentiable function of x:

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1},$$

see Dhrymes (1984), especially Corollary 41.

References

- Armstrong, B. (1985), Measurement error in the generalized linear model, *Comm. in Stat. - Simul. and Comp. 14, 529-544.*

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

- Cheng, C.-L. and Schneeweiss, H. (1998), Polynomial regression with errors in the variables, *J. Roy. Stat. Soc.* B 60, 189-199.

- Cheng, C.-L., Schneeweiss, H., and Thamerus, M. (2000), A small sample estimator for a polynomial regression with errors in the variables, *J. Roy. Stat. Soc.* B 62, 699-709.

- Cheng, C.-L. and Schneeweiss, H. (2002), On the Polynomial Measurement Error Model. In S. van Huffel and P. Lemmerling (Eds.): *Total Least Squares and Errors-in-Variables Modeling.* Kluwer, Dordrecht, 131-143.

- Cheng, C.-L. and Van Ness, J.W. (1999), *Statistical Regression with Measurement Error.* Arnold, London.

- Dhrymes, P.J. (1984), *Mathematics for Econometrics.* (2nd ed.) Springer, New York.

- Fuller, W. A., (1987), *Measurement Error Models.* Wiley, New York.

- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics. The Approach Based on Inference Functions.* Wiley, New York.

- Kukush, A., Schneeweiß, H., and Wolf, R. (2001), Three estimators for the Poisson regression model with measurement errors, *Discussion*

Paper 243, Sonderforschungsbereich 386, University of Munich. To appear in *Statistical Papers*.

- Kukush, A., Schneeweiss, H. and Wolf, R. (2002), Comparing Different Estimators in a Nonlinear Measurement Error Model, *Discussion Paper 244*, Sonderforschungsbereich 386, University of Munich.

- Schneeweiss, H. and Mittag, H.J. (1986), *Lineare Modelle mit fehlerbehafteten Daten*, Physica-Verlag, Heidelberg.

- Schneeweiss, H. and Nittner, T. (2001), Estimating a polynomial regression with measurement errors in the structural and in the functional case - a comparison, in M. Sadeh (Ed.): *Data Analysis from Statistical Foundations, A Festschrift in Honour of the 75th Birthday of D.A.S. Fraser*. Nova Science, New York, 195-205.

- Schneeweiss, H., Cheng, C.-L., and Wolf, R. (2003), On the bias of structural estimation methods in a polynomial regression with measurement error when the distribution of the latent covariate is misspecified, in Klein, I. and Mittnik, S. (Eds.): *Contributions in Modern Economics, From Data Analysis to Economic Policy, In Honour of Gerd Hansen*. Kluwer, Boston, 209-222.

- Shklyar, S. and Schneeweiss, H. (2002). A comparison of asymptotic covariance matrices of three consistent estimators in the Poisson regression model with measurement errors, *Discussion Paper 283*, Sonderforschungsbereich 386, University of Munich.

- Wansbeek, T. and Meijer, E. (2000), *Measurement Error and Latent Variables in Econometrics*. Elsevier, Amsterdam.

Tables and Figures

Table 1: Bias of SQS-Estimators as a function of ϑ

Model 1: $\xi \sim 0.5N(0,1) + 0.5N(\vartheta, 1-\vartheta)$

ϑ	β_0	β_1	β_2
0.00	-0.001	0.003	-0.001
0.05	-0.001	0.003	-0.001
0.10	-0.001	0.004	-0.002
0.15	0.000	0.005	-0.003
0.20	0.001	0.008	-0.004
0.25	0.003	0.011	-0.006
0.30	0.004	0.015	-0.009
0.35	0.006	0.020	-0.013
0.40	0.008	0.026	-0.017
0.45	0.010	0.034	-0.022
0.50	0.011	0.044	-0.027

Table 2: Bias of SQS-Estimators as a function of p

Model 2: $\xi \sim (1-p)N(0,1) + pN(3,1)$

p	β_0	β_1	β_2
0.00	-0.001	0.000	0.000
0.01	0.014	-0.013	-0.014
0.02	0.022	-0.024	-0.018
0.03	0.026	-0.032	-0.019
0.04	0.029	-0.040	-0.018
0.05	0.029	-0.046	-0.016
0.06	0.030	-0.051	-0.014
0.07	0.031	-0.055	-0.012
0.08	0.031	-0.060	-0.010
0.09	0.030	-0.064	-0.007
0.10	0.029	-0.067	-0.005

Table 3: Bias of SQS-Estimators as a function of p

Model 3: $\xi \sim (1-p)N(0,1) + pN(0,3)$

p	β_0	β_1	β_2
0.00	-0.001	0.000	0.000
0.01	0.002	0.000	-0.003
0.02	0.005	0.000	-0.006
0.03	0.008	0.000	-0.009
0.04	0.011	0.000	-0.011
0.05	0.013	0.000	-0.013
0.06	0.014	0.000	-0.014
0.07	0.016	-0.001	-0.015
0.08	0.018	-0.001	-0.016
0.09	0.019	-0.001	-0.017
0.10	0.020	-0.001	-0.018

Table 4: Bias of 5 Estimators as a function of ϑ

Model 1: $\xi \sim 0.5N(0,1) + 0.5N(\vartheta, 1-\vartheta)$

ϑ	naive	CS	SQS	RC	MSQS
0.00	-0.164	0.002	0.003	0.004	0.005
0.05	-0.171	0.002	0.003	0.004	0.005
0.10	-0.177	0.002	0.004	0.005	0.005
0.15	-0.183	0.002	0.005	0.007	0.004
0.20	-0.187	0.003	0.008	0.010	0.004
0.25	-0.191	0.003	0.011	0.015	0.004
0.30	-0.194	0.003	0.015	0.020	0.005
0.35	-0.196	0.003	0.020	0.027	0.005
0.40	-0.197	0.003	0.026	0.036	0.006
0.45	-0.196	0.004	0.034	0.046	0.006
0.50	-0.195	0.004	0.044	0.058	0.007

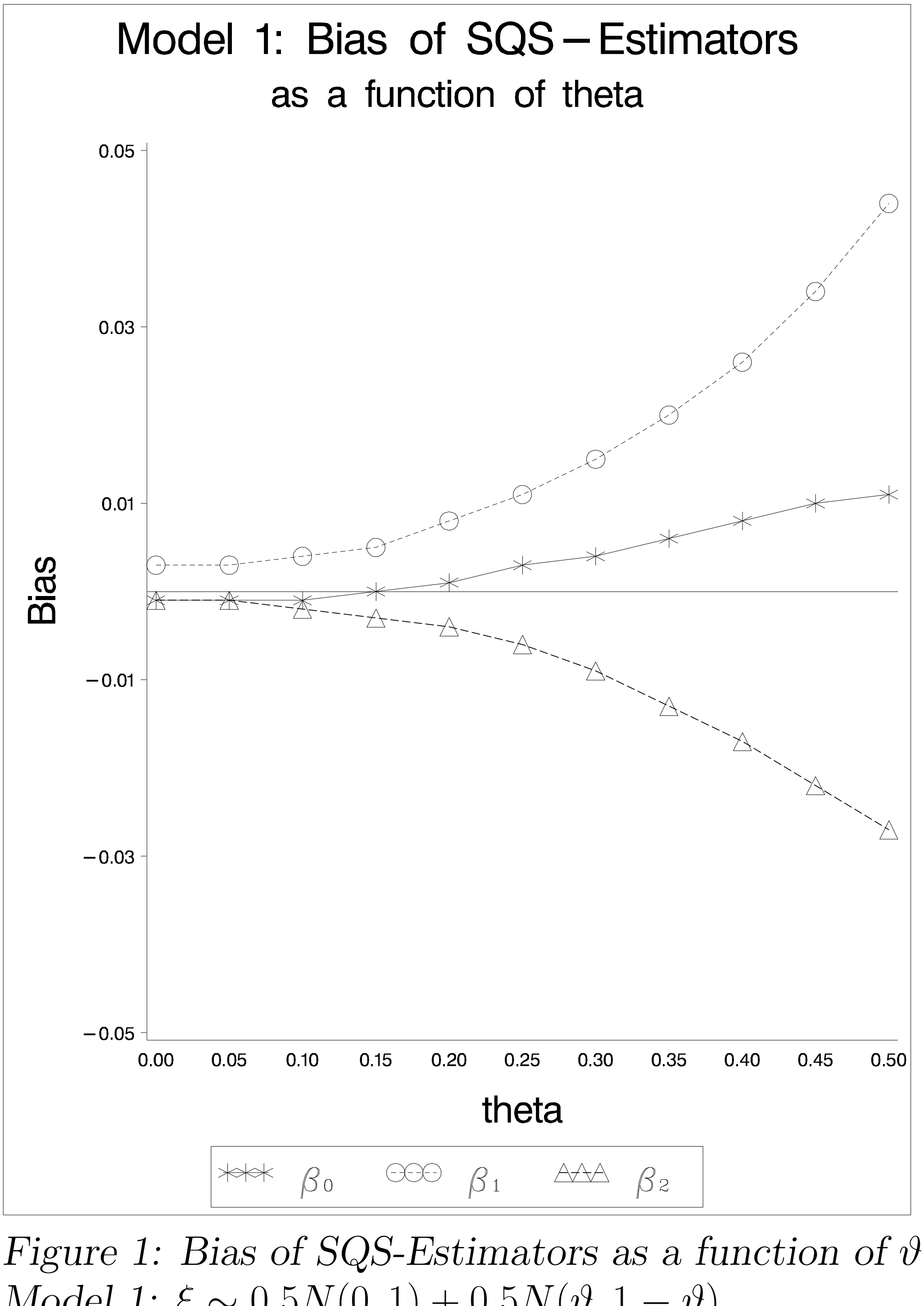

Figure 1: Bias of SQS-Estimators as a function of ϑ
Model 1: $\xi \sim 0.5N(0,1) + 0.5N(\vartheta, 1-\vartheta)$

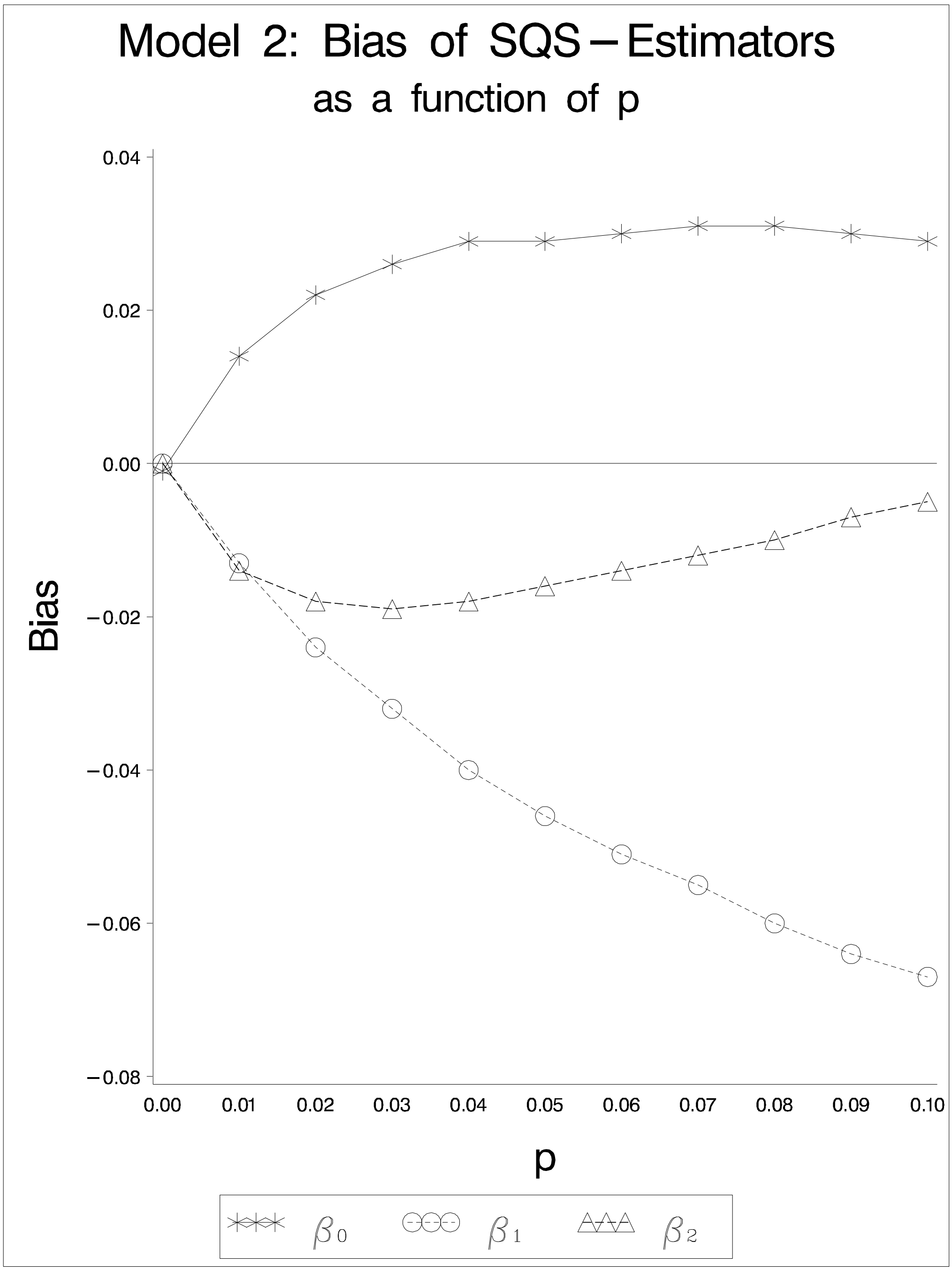

Figure 2: Bias of SQS-Estimators as a function of p
Model 2: $\xi \sim (1-p)N(0,1) + pN(3,1)$

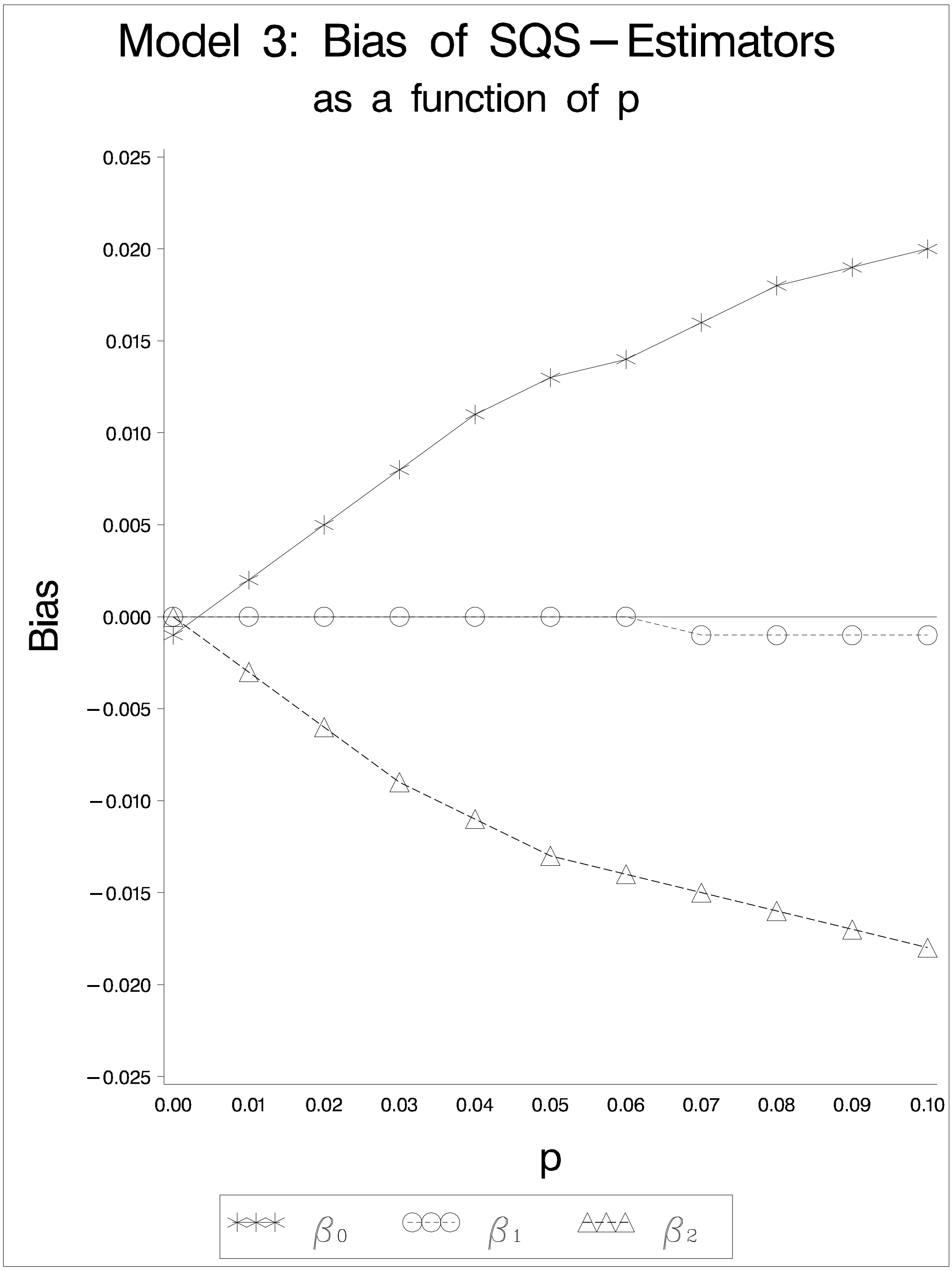

Figure 3: Bias of SQS-Estimators as a function of p
Model 3: $\xi \sim (1-p)N(0,1) + pN(0,3)$

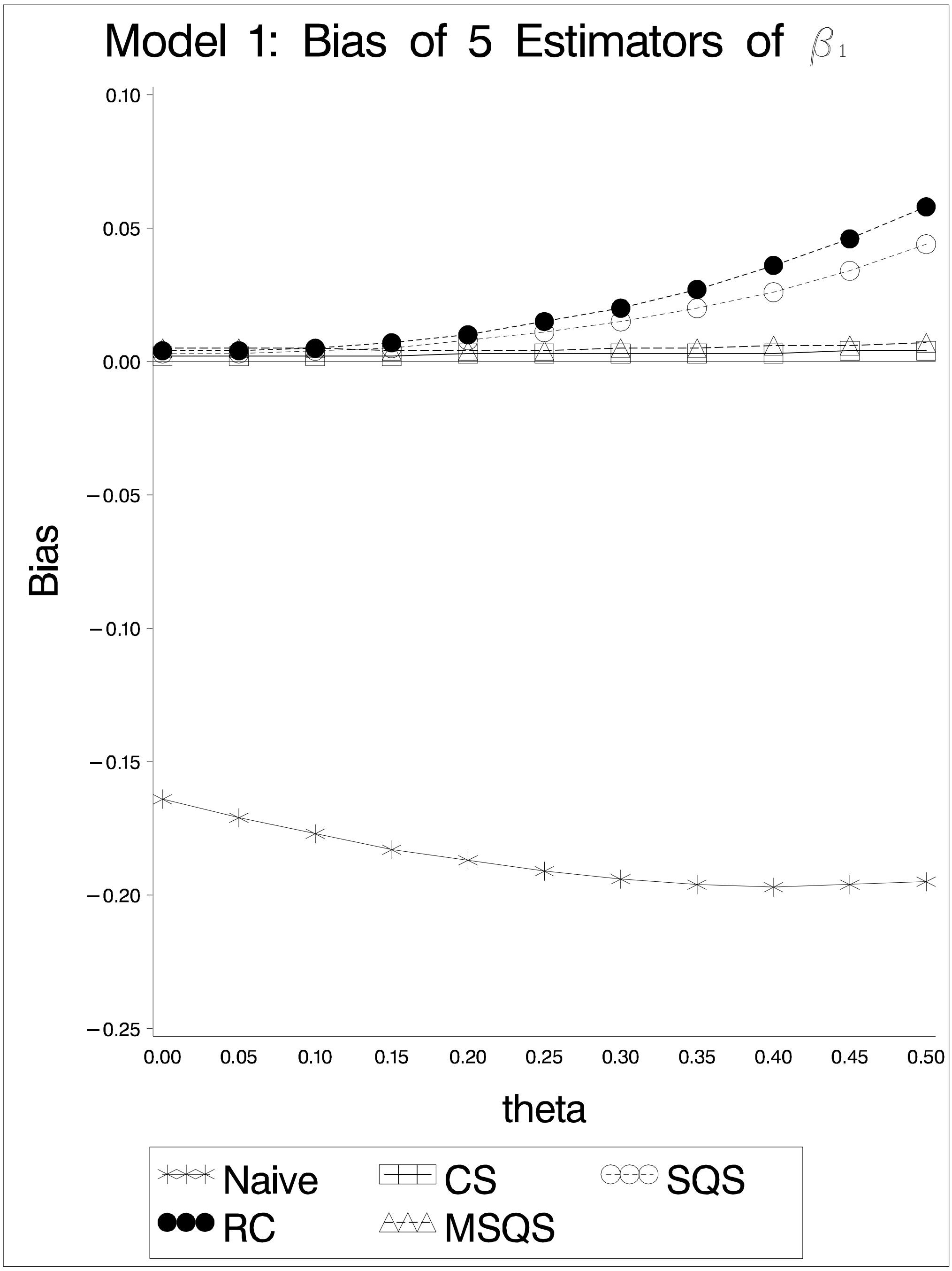

Figure 4: Bias of 5 Estimators as a function of ϑ
Model 1: $\xi \sim 0.5N(0,1) + 0.5N(\vartheta, 1-\vartheta)$