myjournal manuscript No.
(will be inserted by the editor)

Weighted least squares estimators in possibly misspecified nonlinear regression

Ursula U. Müller

Department of Statistics, Texas A&M University, College Station, TX 77843-3143
uschi@stat.tamu.edu

Received: date / Revised version: date

Abstract The behavior of estimators for misspecified parametric models has been well studied. We consider estimators for misspecified nonlinear regression models, with error and covariates possibly dependent. These models are described by specifying a parametric model for the conditional expectation of the response given the covariates. This is a parametric family of conditional constraints, which makes the model itself close to nonparametric. We study the behavior of weighted least squares estimators both when the regression function is correctly specified, and when it is misspecified and also involves possible additional covariates.

Keywords Conditional mean model · Efficient estimation · Estimating equation · Influence function · Misspecification

1 Introduction

Suppose we have specified a parametric regression function $E(Y|X) = r_\vartheta(X)$ with one-dimensional response Y, random covariate X with values in some arbitrary space, unknown p-dimensional parameter ϑ and no further assumptions on the structure of the law of (X, Y). This also covers linear regression models, with $r_\vartheta(X) = \vartheta^\top X$. We can write the model as $Y = r_\vartheta(X) + \varepsilon$, with $E(\varepsilon|X) = 0$ but the distribution of (X, ε) otherwise being unknown. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and want to estimate ϑ. The model $E(Y|X) = r_\vartheta(X)$ suggests estimators $\hat{\vartheta}$ that solve estimating equations

$$\sum_{i=1}^{n} w_t(X_i)(Y_i - r_t(X_i)) = 0 \tag{1.1}$$

with respect to t, where w_t is a p-dimensional vector of weight functions. Solutions $\hat{\vartheta}$ of estimating equations of the form (1.1) are called *weighted* least squares estimators. The ordinary least squares estimator solves an estimating equation (1.1) with weights $w_t(X) = \dot{r}_t(X)^\top$, where $\dot{r}_t$ is the vector of partial derivatives of r_t with respect to the parameter t. We show in Section 3 that the ordinary least squares estimator is not efficient in the model specified by $E(Y|X) = r_\vartheta(X)$ and prove that the random weights $w_t(X) = \dot{r}_t(X)^\top \hat{\sigma}_t^{-2}(X)$ lead to an efficient estimator; here $\hat{\sigma}_t^2(X)$ is an appropriate estimator of the conditional variance $E(\varepsilon^2|X)$ of the error given the covariate.

Suppose now that our model for the regression function is wrong and that the true regression function is $E(Y|X,Z) = r(X,Z)$, with Z being a possible additional covariate. What does $\hat{\vartheta}$ now estimate — and how well does it estimate what it estimates?

Again, the model can be written in the form $Y = r(X,Z) + \varepsilon$ with $E(\varepsilon|X,Z) = 0$. The reason for allowing additional covariates is that we wish to cover misspecifications in which the number of regression parameters is chosen incorrectly. We could, for example, have a linear model with additional covariates, $Y = \vartheta^\top X + \tau^\top Z + \varepsilon$, or with an additive nonparametric term, $Y = \vartheta^\top X + \beta(Z) + \varepsilon$. Note that formally these two regression models could be written as $Y = \vartheta^\top X + \eta$ with correct regression function $r_\vartheta(X) = \vartheta^\top X$ but with errors $\eta = \tau^\top Z + \varepsilon$ or $\eta = \beta(Z) + \varepsilon$ that are in general not conditionally centered any more, since $E(\eta|X)$ equals $\tau^\top E(Z|X)$ or $E(\beta(Z)|X)$.

To see how the weighted least squares estimator $\hat{\vartheta}$ behaves under misspecification, we view the estimating equation (1.1) as an empirical version of the equation

$$E(w_t(X)(Y - r_t(X))) = 0. \tag{1.2}$$

This is all the information that the weighted least squares estimator uses about the nonlinear model. Note that the constraint (1.2) is much weaker than the constraint $E((Y - r_t(X))|X) = 0$ that defines the regression model, the latter being equivalent to $E(w(X)(Y - r_t(X))) = 0$ for (essentially) *all* functions w and for some t. Let $t(P)$ denote the solution of equation (1.2), with P denoting the joint law of (X,Y). It follows that $\hat{\vartheta}$ estimates $t(P)$. Here we must assume that such a solution exists. Lindsay and Qu (2003) then call P *compatible*.

The estimator $\hat{\vartheta}$ solving (1.1) is an empirical estimator, in the sense that it solves an empirical version of the equation (1.2) that defines the functional $t(P)$. We therefore expect it to be not only consistent but also *efficient* for $t(P)$ in the nonparametric model, with nothing specified about the law of (X,Y,Z). The reason for this conjecture is that in a nonparametric model all regular and asymptotically linear estimators necessarily have the same influence function; see also Section 4. The simplest example is linear regression, $r_\vartheta(X) = \vartheta X$, and the unweighted estimating equation, i.e. (1.1) with $w_t(X) = 1$, which gives $\hat{\vartheta} = \sum Y_i / \sum X_i$. It

estimates $t_1(P) = Er(X,Z)/E(X)$. The ordinary least squares estimator $\hat{\vartheta} = \sum X_i Y_i / \sum X_i^2$, which solves the weighted estimating equation (1.1) with $w_t(X) = X$, estimates the functional $t_2(P) = E(Xr(X,Z))/EX^2$. Both estimators are consistent estimators of ϑ if the linear regression model is true, $r(X,Z) = \vartheta X$, but not efficient in this model since they do not use the specific structure of the model. They are, however, optimal for estimating the two functionals $t_1(P)$ and $t_2(P)$ in the nonparametric model.

In Section 4 we show that the weighted least squares estimator $\hat{\vartheta}$ is efficient for the solution $t(P)$ of (1.2) in the model $E(Y|X,Z) = r(X,Z)$. This implies asymptotic normality. We show that the asymptotic variance bound is not changed when Z is not observed. This result is not unexpected, because our efficient estimator does not use the additional covariates.

There is a large amount of literature on estimation under model misspecification in situations which differ from the one considered here. White (1982) showed that the quasi maximum likelihood estimator for the parameter of a probability model p_ϑ is strongly consistent for the parameter which minimizes the Kullback-Leibler distance between the unknown true density p and p_ϑ. Since then many authors have studied this topic in various contexts, some recently. Aguirre-Torres and Toribio (2004), for example, also assume a misspecified parametric model for the density (but with unobservable components) and apply the *efficient method of moments* to estimate the parameter. Articles on estimation in misspecified *regression* models are typically concerned with specific models and misspecifications. Gould and Lawless (1988) and Zhang and Liu (2003) study models with misspecified error distribution. Qu et al. (2000) address the problem of a misspecified correlation structure in longitudinal data analysis. Struthers and Kalbfleisch (1986) study the behavior of parameter estimates in proportional hazards regression where the misspecified model is an accelerated failure time model or where relevant covariates have been omitted. The latter problem of erroneously ignored covariates has also been examined by other authors, for example Sarkar (1989), and McKean et al. (1993), in the context of linear models. For more work on misspecifications in the linear model see, for example, Severini (1998) and Shi et al. (2003). Severini derives properties of the ordinary least squares estimator in the normal linear model with misspecified expectations. Shi et al. consider fixed design points x and derive minimax robust designs for misspecifications with the form of an additive function of x.

To our knowledge, the work most pertinent to our study is White (1981) who looks at the consequences of using approximations $Y = r_\vartheta(X)$ of an unknown true regression model $Y = r(X)$ (without error term ε). He shows that the least squares estimator is a consistent estimator of the parameter that minimizes the prediction mean squared error and also discusses consistency of certain weighted least squares estimators (for more details see Section 5). Our work covers and extends these results.

The rest of the paper is organized as follows. In Section 2 we derive a Taylor expansion of weighted least squares estimators. We will distinguish

two cases: where the nonlinear model is correctly specified and where it is misspecified. The expansion allows us to derive the asymptotic distribution of the estimators in both cases, in particular the covariance matrices, and optimal weights w_t^* for the case that the nonlinear model is the true one. In Section 3 we will focus on the nonlinear regression model. We show that the estimator using optimal weights w_t^* is efficient for ϑ if the model is correctly specified. The weights depend on the underlying distribution and must be estimated. This does not change the asymptotic variance. In Section 4 we assume that the model is misspecified and show that weighted least squares estimators are efficient for functionals $t(P)$ defined as solutions of (1.2). This section can be kept concise by using results from Müller and Wefelmeyer (2002a). In Section 5 we illustrate our results with some simulations and discuss open problems and examples, in particular linear regression and additional restrictions on the misspecification.

2 Influence function of weighted least squares estimators

In the following we keep the notation from the Introduction. We assume that Y is P-square integrable, and that ϑ is a p-dimensional parameter in some open parameter space Θ. The nonlinear model $E(Y|X) = r_\vartheta(X)$ can be regarded as a model defined by a *conditional* constraint, namely by $E(Y - r_\vartheta(X)|X) = 0$. The model might be misspecified. In this case we assume that there is no structural constraint which defines the model. In particular, the above conditional constraint may not hold. This is a nonparametric model. We allow for an additional vector of covariates Z and write $r(X, Z) = E(Y|X, Z)$ for the true (unknown) regression function.

The weighted least squares estimator $\hat{\vartheta}$ solving (1.1) estimates the solution $t(P)$ of equation (1.2), $E(w_t(X)(Y - r_t(X))) = 0$. By definition of r we have $E(w_t(X)(Y - r(X, Z))) = 0$ as well, for any weight function w_t and any t. Hence $t(P)$ solves

$$E(w_t(X)r_t(X)) = E(w_t(X)r(X, Z)). \tag{2.1}$$

This is useful for determining $t(P)$.

We assume that a solution $t(P)$ of the defining equation (2.1) exists. If the nonlinear model is correctly specified, the *unconditional* equation (2.1) obviously holds and, by definition, the *conditional* constraint $E(Y - r_\vartheta(X)|X) = 0$ is satisfied. The latter is not true under misspecification, which is the crucial difference between the correctly specified and the misspecified model: under misspecification the difference $Y - r_{t(P)}(X)$ is *not* conditionally centered. This will affect the asymptotic expansion of the weighted least squares estimator which we will carry out in Theorem 1, covering both misspecified and correctly specified nonlinear regression models.

We make the following assumptions.

Assumption 1 *The p-dimensional vector $w_\tau(X)$ and the regression function $r_\tau(X)$ are $L_2(P)$ differentiable at $\tau = t(P)$ with a $p \times p$ matrix of partial*

derivatives $\dot{w}_{t(P)}(X)$ *and a p-dimensional gradient* $\dot{r}_{t(P)}(X)$*, respectively,*

$$E(|w_\tau(X) - w_{t(P)}(X) - \dot{w}_{t(P)}(X)(\tau - t(P))|^2) = o(|\tau - t(P)|^2),$$
$$E(|r_\tau(X) - r_{t(P)}(X) - \dot{r}_{t(P)}(X)(\tau - t(P))|^2) = o(|\tau - t(P)|^2).$$

Assumption 1 guarantees that the expected value of $w_\tau(X)(Y - r_\tau(X))$ can be approximated as follows,

$$\begin{aligned} &E(w_\tau(X)(Y - r_\tau(X))) - E(w_{t(P)}(X)(Y - r_{t(P)}(X))) \\ &= -A(\tau - t(P)) + o(|\tau - t(P)|), \end{aligned} \tag{2.2}$$

where A is a $p \times p$ matrix of expectations, namely, with $\mu_{t(P)}(X) = E(Y - r_{t(P)}(X)|X)$,

$$A = E(w_{t(P)}(X)\dot{r}_{t(P)}(X)) - E(\dot{w}_{t(P)}(X)\mu_{t(P)}(X)). \tag{2.3}$$

We must assume that A is invertible:

Assumption 2 *The $p \times p$ matrix A from (2.3) is invertible.*

Assumptions 1 and 2 will be in force throughout this paper.

Remark. By Assumption 1, w_τ and r_τ are $L_2(P)$ differentiable. This implies that the empirical process

$$E_{n\tau} = n^{-1/2} \sum_{i=1}^{n} \big(w_\tau(X_i)(Y_i - r_\tau(X_i)) - E(w_\tau(X)(Y - r_\tau(X)))\big)$$

is *stochastically equicontinuous* at $\tau = t(P)$: for every $\varepsilon, \eta > 0$ there is a δ such that

$$\limsup_n P(\sup_{|\tau - t(P)| \le \delta} |E_{n\tau} - E_{nt(P)}| > \eta) \le \varepsilon. \tag{2.4}$$

See e.g. Andrews and Pollard (1994).

Theorem 1 *Any consistent solution $\hat{\vartheta}$ of (1.1) has the stochastic expansion*

$$n^{1/2}(\hat{\vartheta} - t(P)) = A^{-1} n^{-1/2} \sum_{i=1}^{n} w_{t(P)}(X_i)(Y_i - r_{t(P)}(X_i)) + o_p(1) \tag{2.5}$$

with $A = E(w_{t(P)}(X)\dot{r}_{t(P)}(X)) - E(\dot{w}_{t(P)}(X)\mu_{t(P)}(X))$. If the nonlinear regression model is correctly specified we have $\mu_{t(P)}(X) = \mu_\vartheta(X) = E(Y - r_\vartheta(X)|X) = 0$ and the stochastic expansion simplifies:

$$n^{1/2}(\hat{\vartheta} - \vartheta) = \big(E(w_\vartheta(X)\dot{r}_\vartheta(X))\big)^{-1} n^{-1/2} \sum_{i=1}^{n} w_\vartheta(X_i)(Y_i - r_\vartheta(X_i)) + o_p(1). \tag{2.6}$$

Proof. Consider the estimating equation (1.1) and the empirical process $E_{n\tau}$ from equation (2.4) in the above remark. We have

$$\begin{aligned} 0 &= n^{-1/2}\sum_{i=1}^{n} w_{\hat\vartheta}(X_i)(Y_i - r_{\hat\vartheta}(X_i)) \\ &= E_{n\hat\vartheta} + n^{-1/2}\sum_{i=1}^{n} E(w_{\hat\vartheta}(X)(Y - r_{\hat\vartheta}(X))) + E_{nt(P)} - E_{nt(P)} \end{aligned}$$

with $E_{n\hat\vartheta} - E_{nt(P)} = o_p(1)$ by (2.4). Hence

$$\begin{aligned} 0 &= E_{nt(P)} + n^{-1/2}\sum_{i=1}^{n} E(w_{\hat\vartheta}(X)(Y - r_{\hat\vartheta}(X))) + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n} \Big(w_{t(P)}(X_i)(Y_i - r_{t(P)}(X_i)) \\ &\quad - E(w_{t(P)}(X)(Y - r_{t(P)}(X))) + E(w_{\hat\vartheta}(X)(Y - r_{\hat\vartheta}(X)))\Big) + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n} w_{t(P)}(X_i)(Y_i - r_{t(P)}(X_i)) - An^{1/2}(\hat\vartheta - t(P)) \\ &\quad + n^{1/2} o(|\hat\vartheta - t(P)|) + o_p(1). \end{aligned}$$

In the last equation we used (2.2). Since the matrix A is invertible by Assumption 2 we have proved (2.5). □

Corollary 1 *Theorem 1 implies that $\hat\vartheta$ is asymptotically normally distributed with asymptotic mean $t(P)$. Under misspecification the approximate covariance matrix of $\hat\vartheta$ is $\Sigma_{t(P)}/n$ where $\Sigma_{t(P)}$ is the asymptotic covariance matrix,*

$$\Sigma_{t(P)} \;=\; A^{-1}E(\sigma^2_{t(P)}(X)w_{t(P)}(X)w_{t(P)}(X)^\top)(A^\top)^{-1},$$

with $\sigma^2_{t(P)}(X) = E((Y - r_{t(P)}(X))^2|X)$. This would be the conditional variance of Y given X if the nonlinear model were correctly specified. In that case the asymptotic mean of $\hat\vartheta$ is $t(P) = \vartheta$ and the asymptotic covariance matrix Σ_ϑ is

$$\big(E(w_\vartheta(X)\dot r_\vartheta(X))\big)^{-1}E\big(\sigma^2_\vartheta(X)w_\vartheta(X)w_\vartheta(X)^\top\big)\big(E(\dot r_\vartheta(X)^\top w_\vartheta(X)^\top)\big)^{-1}.$$

A degenerate case is *linear* regression and *ordinary* least squares. Then $r_\vartheta(X) = \vartheta^\top X$ and $w_\vartheta(X) = X$. Hence $\dot r_\vartheta(X) = X^\top$ and $\dot w_\vartheta(X) = 0$. Then the matrix A simplifies to $E(XX^\top)$.

In Section 4 we will show that the weighted least squares estimator $\hat\vartheta$ is efficient for $t(P)$ in the nonparametric model where $t(P)$ is defined by $E(w_t(X)r_t(X)) = E(w_t(X)r(X,Z))$. Since the meaning of $t(P)$ depends on the weights, and since the estimator is efficient, it is clear that changing the weights would not produce an improved estimator but, rather, an estimator

for a different functional. The situation is different in the nonlinear model. Here, any weighted estimator $\hat{\vartheta}$ estimates ϑ. Hence it is reasonable to choose weights such that the resulting estimator is efficient for ϑ. In order to determine such optimal weights we consider the asymptotic covariance matrix Σ_ϑ of $\hat{\vartheta}$ given in Corollary 1, $\Sigma_\vartheta = E(k(X,Y)k(X,Y)^\top)$, where k is the influence function given in (2.6), $k(x,y) = \big(E(w_\vartheta(X)\dot{r}_\vartheta(X))\big)^{-1} w_\vartheta(x)(y - r_\vartheta(x))$. Then, by arguments given below, the weights that minimize the asymptotic covariance are

$$w^*_\vartheta(x) = \dot{r}_\vartheta(x)^\top \sigma^{-2}_\vartheta(x), \tag{2.7}$$

where $\sigma^2_\vartheta(x)$ is the conditional variance of Y given $X = x$ as defined above. This can be seen as follows. Write k^* for the influence function using weights w^*_ϑ, i.e. $k^*(x,y) = \big(E(w^*_\vartheta(X)\dot{r}_\vartheta(X))\big)^{-1} w^*_\vartheta(x)(y - r_\vartheta(x))$. For ϑ one-dimensional we have, by the Cauchy-Schwarz inequality, $(E(k^*k))^2 \le E(k^{*2})E(k^2)$ where $E(k^{*2})$ is the asymptotic variance of the estimator $\hat{\vartheta}$ using weights w^*_ϑ. This is the desired $E(k^{*2}) \le E(k^2)$ if $E(k^*k) = E(k^{*2})$, which is equivalent to $E((k - k^*)k^*) = 0$, i.e., to k^* and $k - k^*$ orthogonal. It is now easy to check that this holds for our specific choice w^*_ϑ from (2.7). For higher dimensional ϑ a related argument applies. The weights above are chosen such that the spaces spanned by the influence function k^* and by $k - k^*$ are orthogonal, $E((k - k^*)k^{*\top}) = 0$ (see also the discussion of efficiency in Section 3). Writing $k = k^* + (k - k^*)$ we obtain $\Sigma_\vartheta = E(kk^\top) = E(k^*k^{*\top}) + E((k - k^*)(k - k^*)^\top)$, i.e. the weights w^*_ϑ minimize the asymptotic covariance.

Note that the optimal weights $w^*_\vartheta(x)$ involve the conditional variance $\sigma^2_\vartheta(x)$, which depends on the underlying distribution. Since we are considering nonlinear regression, where we do not have a parametric model for $\sigma^2_\vartheta(x)$, it must be estimated using nonparametric methods. In Section 3 we show that the weights w^*_ϑ do indeed yield an efficient estimator in the nonlinear model: we will see that an estimator is efficient if it is asymptotically linear as stated in (2.6) with $w_\vartheta = w^*_\vartheta$.

Remark. There is literature on the behavior of the *maximum likelihood estimator* in misspecified *parametric* models (see, for example, White, 1982, and Greenwood and Wefelmeyer, 1997). Consider independent observations $V_1, \dots, V_n$ with a (misspecified) parametric density $p_\vartheta, \vartheta \in \Theta$. Let p denote the true density of the distribution P. By the *Kullback–Leibler information* for the family $p_\vartheta, \vartheta \in \Theta$, we mean the expectation $P \log p_\vartheta = \int \log p_\vartheta \, dP = \int \log p_\vartheta(v) p(v) \, dv$. Write $t(P)$ for the parameter ϑ that maximizes $P \log p_\vartheta$. Then the parameter $\hat{\vartheta}$ that maximizes the empirical version of $P \log p_\vartheta$, namely $1/n \sum_{i=1}^n \log p_\vartheta(V_i)$, is an estimator for the maximum Kullback–Leibler information functional $t(P)$. In particular, $\hat{\vartheta}$ is the maximum likelihood estimator.

The fact that the maximum likelihood estimator estimates the maximum Kullback-Leibler information functional relates to the regression setting as follows. Write $Y = r_\vartheta(X) + \varepsilon$ and suppose that X and ε are independent.

Further assume that ε is normally distributed with known variance and that the covariate distribution is known, say with density m. Then the observations $V_i = (X_i, Y_i)$ are independent with density

$$p_\vartheta(x,y) = m(x)\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y-r_\vartheta(x))^2}{2\sigma}\Big).$$

Hence the maximum likelihood estimator has a score function proportional to $\dot r_\vartheta(x)^\top(y-r_\vartheta(x))$. Since the least squares estimator solves the estimating equation (1.1) with weights $w_\vartheta(x) = \dot r_\vartheta(x)^\top$, it coincides with the maximum likelihood estimator for the family of densities p_ϑ, $\vartheta \in \Theta$. Now assume that the regression model is misspecified and the regression function is r. (It makes no difference whether an additional covariate Z is involved or not.) Then the density of the observations does not depend on ϑ and can be called p. The above argument shows that the functional $t(P)$ from (2.1) is the maximum Kullback–Leibler information functional for the family p_ϑ.

The score function for ϑ is not changed if we assume a model for the distribution of the covariate X, as long as it does not involve ϑ. The maximum likelihood estimator then becomes a *conditional* maximum likelihood estimator. In particular, we may take m to be completely unknown. Hence the least squares estimator estimates the maximum Kullback-Leibler information functional $t(P)$ for this larger family of (semiparametric) densities $p_\vartheta, \vartheta \in \Theta$.

3 Efficiency in constrained models

In this section we assume that the nonlinear regression model, $E(Y|X) = r_\vartheta(X)$, is correctly specified. We will show that the least squares estimator that uses the weights given in (2.7) is efficient for $t(P) = \vartheta$ in the sense of Hájek and Le Cam. We keep the discussion brief and refer to the exposition of Bickel et al. (1998), which treats this efficiency concept in detail.

The model is described by a constraint on the conditional distribution Q of Y given X. It is therefore convenient to factor P,

$$P(dx,dy) \;=\; M(dx)Q(x,dy).$$

The marginal distribution M of X is arbitrary. For reasons of clarity we will use operator notation in this section. For instance, the constraint $E(Y - r_\vartheta(X)|X = x) = 0$ is written

$$Q[y - r_\vartheta(x)] = \int (y - r_\vartheta(x))Q(x,dy) = 0$$

and, similarly, the unconditional expectation $E(Y-r_\vartheta(X))$ is $P[y-r_\vartheta(x)] = \int (y - r_\vartheta(x))P(dx,dy)$.

In order to be able to characterize efficient estimators in an adequate way, we first need to determine the tangent space of the model which is

the set of possible perturbations of P within the model. An estimator of a particular functional is, roughly speaking, efficient if its influence function equals the *canonical gradient* of the functional, which is an element of the tangent space.

In a first step we consider the nonparametric model, without constraint $Q[y - r_\vartheta(x)] = 0$, and (Hellinger differentiable) perturbations of M and Q,

$$M_{nv}(dx) \doteq M(dx)(1+n^{-1/2}v(x)), \quad Q_{nh}(x, dy) \doteq Q(x, dy)(1+n^{-1/2}h(x, y))$$

with derivatives v in $L_{2,0}(M)$ and h in H where

$$\begin{aligned} L_{2,0}(M) &= \{v \in L_2(M) : Mv(x) = 0\}, \\ H &= \{h(x, y) \in L_2(P) : Qh(x, y) = 0\}. \end{aligned}$$

Note that we require $Mv(x) = 0$ and $Qh(x, y) = 0$ in order to guarantee that M_{nv} and Q_{nh} are probability distributions. The perturbed joint distribution is

$$P_{nvh}(dx, dy) = M_{nv}(dx)Q_{nh}(x, dy) \doteq P(dx, dy)(1 + n^{-1/2}(v(x) + h(x, y)))$$

with derivative $v(x) + h(x, y)$. Since the functions v in $L_{2,0}(M)$ and h in H are orthogonal, $P[v(x)h(x, y)] = M[v(x)Qh(x, y)] = 0$, the tangent space of the nonparametric model at P is the orthogonal sum

$$L_{2,0}(M) \oplus H = \{v(x) + h(x, y) : v \in L_{2,0}(M), h \in H\}.$$

We now consider the nonlinear regression model, i.e. we additionally take the constraint $Q[y - r_\vartheta(x)] = 0$ into account in order to determine the tangent space of this constrained model. The perturbed distribution $P_{nvh}(dx, dy)$ must now fulfill a perturbed constraint, $Q_{nh}[y - r_{\vartheta_{nu}}(x)] = 0$ for some ϑ_{nu} close to ϑ, say $\vartheta_{nu} = \vartheta + n^{-1/2}u$ with u in $\mathbf{R}^p$. Using $Q[y - r_\vartheta(x)] = 0$ and $Q[h(x, y)] = 0$ we obtain

$$\begin{aligned} 0 = Q_{nh}[y - r_{\vartheta_{nu}}(x)] &\doteq Q(1 + n^{-1/2}h(x, y))[y - r_{\vartheta_{nu}}(x)] \\ &\doteq Q(1 + n^{-1/2}h(x, y))[y - r_\vartheta(x) - n^{-1/2}\dot{r}_\vartheta(x)u] \\ &= n^{-1/2}(Q[h(x, y)y] - \dot{r}_\vartheta(x)u), \end{aligned}$$

which leads to a constraint $Q[h(x, y)y] = \dot{r}_\vartheta(x)u$ on h in H. For fixed $u \in \mathbf{R}^p$ we write H_u for the solution space of this equation,

$$H_u = \{h \in H : Q[h(x, y)y] = \dot{r}_\vartheta(x)u\},$$

and H_* for the union of all affine spaces H_u, $u \in \mathbf{R}^p$. With this notation the tangent space of the constrained model is

$$\mathcal{H} = L_{2,0}(M) \oplus H_*.$$

Note that if the marginal distribution M is known then we do not have to perturb M and the tangent space reduces to the solution space H_*.

If ϑ is known, H_* reduces to the space of solutions of the corresponding homogeneous equation,

$$H_0 = \{h \in H : Q[h(x,y)y] = 0\},$$

and the tangent space reduces to $L_{2,0}(M) \oplus H_0$.

The tangent space of the constrained model is now specified, namely as the orthogonal sum of $L_{2,0}(M)$ and the solution space H_*, but we find it convenient to go further and decompose H_* into the homogeneous solution space H_0 and its orthogonal complement in H_*. We again write $\sigma^2_\vartheta(x)$ for the conditional variance given x, $\sigma^2_\vartheta(x) = Q[y - r_\vartheta(x)]^2$, and introduce the p-dimensional vector

$$\ell(x,y) = \dot{r}_\vartheta(x)^\top \sigma_\vartheta^{-2}(x)(y - r_\vartheta(x)). \tag{3.1}$$

Let ℓ_j denote the j-th component of ℓ and e_j the p-dimensional standard basis vector, i.e. the j-th component is one and the other components are zero. In order to describe the solution space of the inhomogeneous equation $Q[h(x,y)y] = \dot{r}_\vartheta(x)u$, $u \in \mathbf{R}^p$, we solve the equation for the standard basis vectors $u = e_j$, $j = 1, \ldots, p$. It is easily seen that $h = \ell_j$ is the unique solution of $Q[h(x,y)y] = \dot{r}_\vartheta(x)e_j$ that is *orthogonal* to H_0. Write $[\ell]$ for the linear span of $\ell_1, \ldots, \ell_p$, $[\ell] = \{u^\top \ell : u \in \mathbf{R}^p\}$. Then H_* has the orthogonal decomposition into the homogeneous solution space H_0 and the inhomogeneous solution space $[\ell]$, $H_* = H_0 \oplus [\ell]$, and the tangent space of the constrained model is

$$\mathcal{H} = L_{2,0}(M) \oplus H_0 \oplus [\ell].$$

This decomposition is useful for estimating arbitrary differentiable functionals $t(P)$.

We restrict our attention to estimating ϑ and consider it as a p–dimensional functional of P by setting $t(P) = \vartheta$ if $Q[y - r_\vartheta(x)] = 0$. The vector ℓ in (3.1) will play the role of *score function* for ϑ. The *Fisher information* is

$$I = P[\ell\ell^\top] = M[\dot{r}_\vartheta(x)^\top \dot{r}_\vartheta(x)\sigma_\vartheta^{-2}(x)]. \tag{3.2}$$

At this point we need to recall some results on the characterization of efficient estimators. For parametric models the results are due to Le Cam (1960) and Hájek (1970). For semiparametric models, as considered here, we again refer to Bickel et al. (1998).

A p-dimensional functional $t(P)$ is called *differentiable at* P with *gradient* $\tilde{g}$ if $\tilde{g} = \tilde{v} + \tilde{h}$ with $\tilde{v}$ in $L_{2,0}(M)^p$ and $\tilde{h}$ in H^p and

$$n^{1/2}(t(P_{nvh}) - t(P)) \to P[(\tilde{v} + \tilde{h})(v + h)] \quad \text{for } v \in L_{2,0}(M), h \in H_*.$$

The *canonical* gradient g is the component-wise projection of $\tilde{g}$ onto the tangent space $\mathcal{H} = L_{2,0}(M) \oplus H_*$, i.e. $g = \tilde{v} + \tilde{h}_*$ with $\tilde{h}_*$ the component-wise projection of $\tilde{h}$ onto H_*.

An estimator T_n for $t(P)$ is called *regular* at P with *limit* L if

$$n^{1/2}(T_n - t(P_{nvh}) \Rightarrow L \quad \text{under } P_{nvh} \text{ for } v \in L_{2,0}(M), h \in H_*.$$

The convolution theorem says that

$$L = (P[gg^\top])^{1/2} N_p + R \quad \text{in distribution,}$$

where N_p is a p-dimensional standard normal vector and R is independent of N_p. This justifies calling an estimator T_n *efficient* for $t(P)$ if

$$n^{1/2}(T_n - t(P) \Rightarrow (P[gg^\top])^{1/2} N_p \quad \text{under } P_n.$$

An estimator T_n for $t(P)$ is called *asymptotically linear* at P with *influence function* k if $k \in L_{2,0}(P)^p$ and

$$n^{1/2}(T_n - t(P)) = n^{-1/2} \sum_{i=1}^{n} k(X_i, Y_i) + o_p(1).$$

We have the following characterizations:

1. An asymptotically linear estimator is regular if and only if its influence function is a gradient.

2. An estimator is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient.

We will utilize the second characterization as follows: in the next lemma we determine the canonical gradient of ϑ and then go on to show that the weighted least squares estimator $\hat{\vartheta}$ with suitably estimated weights w_ϑ^* as given in (2.7) has the expansion (2.6) (with $w_\vartheta = w_\vartheta^*$), and that the influence function given there equals the canonical gradient. This gives efficiency.

Lemma 1 *The functional $t(P)$ defined by $t(P) = \vartheta$ if $Q[y - r_\vartheta(x)] = 0$ is differentiable at P with canonical gradient*

$$I^{-1}\ell(x,y) = (M[\dot{r}_\vartheta(x)^\top \dot{r}_\vartheta(x)\sigma_\vartheta^{-2}(x)])^{-1}\dot{r}_\vartheta(x)^\top \sigma_\vartheta^{-2}(x)(y - r_\vartheta(x)).$$

Proof. Let $v \in L_{2,0}(M)$ and $h \in H_*$, i.e. $v + h \in \mathcal{H}$. By definition, $t(P)$ is differentiable at P with gradient g if

$$n^{1/2}(t(P_{nvh}) - t(P)) \to P[g(v+h)] \quad \text{with } g \in L_{2,0}(M)^p \oplus H^p.$$

Since $n^{1/2}(t(P_{nvh}) - t(P)) = n^{1/2}(\vartheta_{nu} - \vartheta) = u$, we therefore need to determine g such that $P[g(v+h)] = u$. In particular, we want g to be the canonical gradient, i.e. with components in the tangent space $\mathcal{H}$. Due to the orthogonal decomposition $\mathcal{H} = L_{2,0}(M) \oplus H_0 \oplus [\ell]$, where $[\ell]$ comes from ϑ being unknown, we can assume that the components of g are in the linear span $[\ell]$, i.e. g is of the form $J\ell$ where J is some $p \times p$ matrix. Indeed, using $P[\ell v] = 0$, $Q[h(x,y)y] = \dot{r}_\vartheta(x)u$ and formulas (3.1) and (3.2) for the score function ℓ and the Fisher information I, we obtain for $g = J\ell$

$$\begin{aligned} &P[J\ell(v+h)] = JP[\ell h] = JP[\dot{r}_\vartheta(x)^\top \sigma_\vartheta^{-2}(x)(y - r_\vartheta(x))h(x,y)] \\ &= JM[\dot{r}_\vartheta(x)^\top \sigma_\vartheta^{-2}(x) Q[h(x,y)y]] = JM[\dot{r}_\vartheta(x)^\top \sigma_\vartheta^{-2}(x)\dot{r}_\vartheta(x)u] = JIu. \end{aligned}$$

The choice $J = I^{-1}$ thus gives the desired $P[g(v+h)] = P[I^{-1}\ell(v+h)] = u$. Since the components of $g = I^{-1}\ell$ are in the tangent space by construction, this shows that $g = I^{-1}\ell$ is the canonical gradient of $t(P)$. □

We will now show that any consistent solution $\hat{\vartheta}$ of the weighted estimating equation (1.1) with optimal weights $w^*_\vartheta(x) = \dot{r}_\vartheta(x)^\top \sigma^2_\vartheta(x)^{-1}$ is efficient for ϑ. As mentioned earlier, the equation is *undetermined* since the weights involve the *unknown* conditional variance $\sigma^2_\vartheta(x)$. Hence it cannot be used as it stands. We can, however, replace $\sigma^2_\vartheta(x)$ with an appropriate consistent estimator $\hat{\sigma}^2_\vartheta(x)$. This does not change the asymptotic variance since the stochastic approximation (2.6) remains valid: the term $n^{-1/2}\sum_{i=1}^n \dot{r}_\vartheta(X_i)^\top(\hat{\sigma}^2_\vartheta(X_i)^{-1} - \sigma^2_\vartheta(X_i)^{-1})(Y_i - r_\vartheta(X_i)))$ is (approximately) conditionally centered and hence negligible. Two simple estimators come immediately to mind. Since $\sigma^2_\vartheta(x) = E(Y^2|X = x) - (E(Y|X = x))^2 = E(Y^2|X) - r_\vartheta(x)^2$, we can use a kernel estimator such as the Nadaraya–Watson estimator for the first term of the sum. For the second term we can use the regression function directly or a second kernel estimator.

Note that if we had a parametric model for the conditional variance, i.e. if $\sigma^2_\vartheta(x)$ were a known function (up to the parameter ϑ), we could use the weights directly but would lose efficiency. The reason is as follows. A parametric model for the conditional variance constitutes a second constraint on the model, namely $E((Y - r_\vartheta(X))^2 - \sigma^2_\vartheta(X)|X) = 0$. This would yield a two-dimensional constraint which would have to be incorporated into the estimating equation. In particular it would lead to new different optimal weights involving higher moments (see Müller and Wefelmeyer, 2002b, for autoregressive models with multidimensional constraints).

In the following let $\hat{\sigma}^2_\vartheta(x)$ be some consistent estimator for $\sigma^2_\vartheta(x)$ (which possibly depends on ϑ) and consider a consistent solution $\hat{\vartheta}$ of the (determined) estimating equation with estimated optimal weights,

$$\sum_{i=1}^n \dot{r}_t(X_i)^\top \hat{\sigma}^2_t(X_i)^{-1}(Y_i - r_t(X_i)) = 0. \tag{3.3}$$

As explained in the above remark, the approximation (2.6) remains valid:

$$\begin{aligned} n^{1/2}(\hat{\vartheta} - \vartheta) &= (Ew^*_\vartheta(X)\dot{r}_\vartheta(X))^{-1} n^{-1/2} \sum_{i=1}^n w^*_\vartheta(X_i)(Y_i - r_\vartheta(X_i)) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^n \big(E[\dot{r}_\vartheta(X)^\top \dot{r}_\vartheta(X)\sigma_\vartheta^{-2}(X)]\big)^{-1} \\ &\qquad \dot{r}_\vartheta(X_i)^\top \sigma_\vartheta^{-2}(X_i)(Y_i - r_\vartheta(X_i)) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^n k^*(X_i, Y_i) + o_p(1) \end{aligned}$$

with influence function $k^*(x,y) = I^{-1}\ell(x,y)$ as given in Lemma 1. Hence $\hat{\vartheta}$ is asymptotically linear with influence function equal to the canonical gradient, i.e. it is efficient. We formulate the result as a corollary.

Corollary 2 *In the nonlinear regression model $E(Y|X) = r_\vartheta(X)$, any consistent solution $\hat{\vartheta}$ of the estimating equation (3.3) is efficient for ϑ.*

Note that $\hat{\vartheta}$ is asymptotically normally distributed with asymptotic covariance matrix I^{-1} where I is the Fisher information given in (3.2). An illustration of the estimation procedure is in Section 5 where we compare the performance of the efficient estimator and the ordinary least squares estimator by means of a simple simulation study.

4 Efficiency under misspecification

We now consider the case where the parametric model is misspecified and assume that the true regression model is nonparametric, i.e. $E(Y|X,Z) = r(X,Z)$. A weighted least squares estimator $\hat{\vartheta}$ solving equation (1.1) is an empirical estimator in the nonparametric model. This indicates that it is efficient for $t(P)$ given by (2.1), $E(w_{t(P)}(X)r_{t(P)}(X)) = E(w_{t(P)}(X)r(X,Z))$.

For a rigorous proof we refer to Müller and Wefelmeyer (2002a), who construct efficient estimators for ϑ in a more general context. Let us briefly sketch the connection. Müller and Wefelmeyer consider parametric models defined by an unconditional constraint, $Ea_\vartheta(V) = 0$, where a_ϑ is a k-dimensional vector of functions and ϑ a p-dimensional parameter vector with $k \geq p$. Note that only the case $k > p$ defines a proper constraint. If $k = p$ then the model is nonparametric, which is the situation considered here. Using $E(Y|X,Z) = r(X,Z)$, we can rewrite (2.1) as an (improper) unconditional constraint,

$$E(w_{t(P)}(X)(Y - r_{t(P)}(X))) = 0. \tag{4.1}$$

Since the dimension of $w_{t(P)}$ is p, our misspecified model is a special case of $Ea_{t(P)}(V) = 0$ with $V = (X,Y)$ and $k = p$. We can now apply Lemma 2 in Müller and Wefelmeyer (2002a), which gives the efficient influence function of $\hat{\vartheta}$ in the general constrained model. Simple calculations show that it equals the influence function of $\hat{\vartheta}$ given in our expansion (2.5) in Theorem 1. This shows that $\hat{\vartheta}$ is efficient for $t(P)$ defined by (4.1) and therefore, if we write $E(Y|X,Z) = r(X,Z)$, is efficient for $t(P)$ defined by (2.1). We formulate the result as a corollary.

Corollary 3 *In the nonparametric regression model $E(Y|X,Z) = r(X,Z)$, any consistent solution $\hat{\vartheta}$ of (1.1) is efficient for the functional $t(P)$ defined by (2.1).*

Note that for the above result the additional covariate Z need not be an additional observation but could also stand for an unobservable variable such as an error. In either case $\hat{\vartheta}$ is asymptotically normally distributed with asymptotic mean $t(P)$ and asymptotic covariance matrix $\Sigma_{t(P)}$ as in Corollary 1.

The efficiency proof in Müller and Wefelmeyer (2002a) is similar to the proof carried out in the last section, yet simpler. The reason is the following. The nonlinear regression model is defined by a *conditional* constraint,

$$E(w_\vartheta(X)(Y - r_\vartheta(X))|X) = 0. \tag{4.2}$$

Since (4.2) implies (4.1), model (4.2) is a submodel of (4.1) with more structure. In the last section, in order to determine a local model we had to factor the joint distribution $P(dx, dy)$ into the marginal distribution $M(dx)$ and the conditional distribution $Q(x, dy)$ (for which the conditional constraint is defined) and perturbed both distributions. This is not necessary in the unconditionally constrained model (4.1), where it suffices to perturb P. It should, however, be noted that the efficiency proof from Section 3 adapts to this simpler situation in a straightforward way.

5 Applications

There are many examples of nonlinear regression models that are covered by our parametric model $E(Y|X) = r_\vartheta(X)$, for example cases where r_ϑ is a power function, $r_\vartheta(x) = \vartheta_1 x^{\vartheta_2}$, the exponential decay model $r_\vartheta(x) = \vartheta_1 \exp(-\vartheta_2 x)$ and polynomial and polynomial-trigonometric regression. A popular model for enzymatic reactions and for other applications is the Michaelis–Menten model (1913), $r_\vartheta(x) = \vartheta_1 x/(\vartheta_2 + x)$. It features no typical error structure (see e.g. Ruppert et al. 1989) and thus fits well into our model class where only the conditional expectation $E(Y|X)$ is specified.

As seen in the previous discussion, when the nonlinear model is correctly specified the optimally weighted least squares estimator $\hat{\vartheta}$ is efficient for ϑ. It is straightforward to write down the estimating equation (3.3) for specific examples such as the regression models from above. For an illustration, and in order to compare the ordinary least squares estimator (OLS) and the efficient estimator, we performed a simple simulation study considering the regression model $Y = \vartheta_1 \cos(\vartheta_2 X) + \varepsilon$ with amplitude parameter $\vartheta_1 = 1$ and frequency parameter $\vartheta_2 = 2$. In applications, data often suggest that the conditional variance of Y given $X = x$, $\sigma^2(x)$, is increasing with x or that it is larger in the boundary regions. In order to study these two situations we generated the error variables ε as transformations of the covariates X and auxiliary standard normally distributed variables N, namely $\varepsilon = \sigma(X)N$ with σ (a) an increasing line and (b) a parabola. For the sake of simplicity we sampled the covariates X from a uniform distribution with support $[-1, 1]$. The regression function and two typical samples with size $n = 200$ are shown in Figure 1.

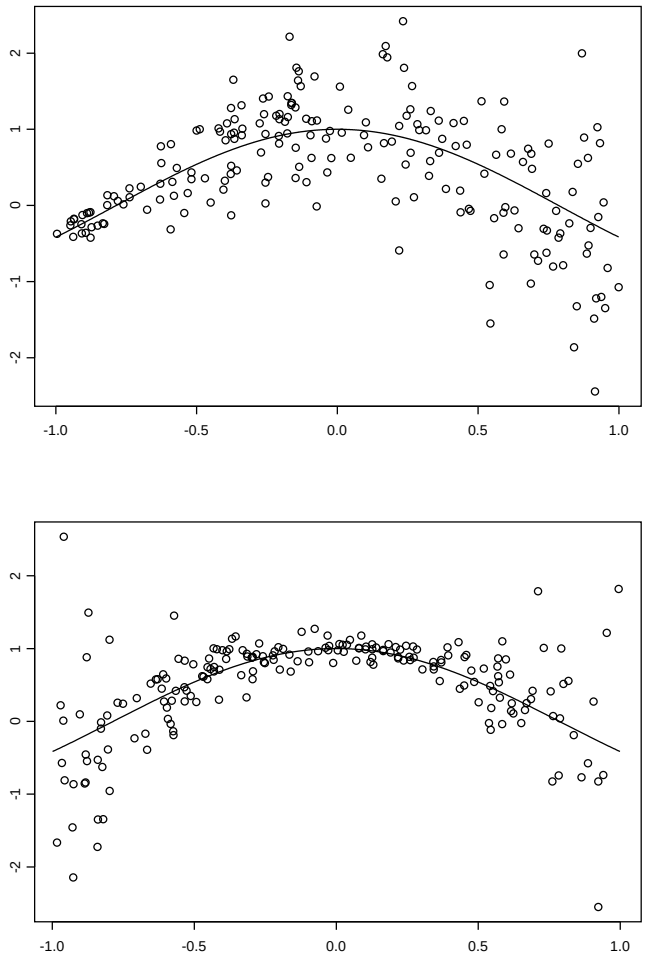

Fig. 1 Graph of the regression function $r_\vartheta(x) = \vartheta_1 \cos(\vartheta_2 x)$ with $\vartheta_1 = 1$, $\vartheta_2 = 2$, and typical sample data (x_i, y_i), $i = 1, \ldots, 200$, with (a) σ an increasing line, $\sigma(x) = 0.5x + 0.6$ (upper panel) and (b) σ a parabola, $\sigma(x) = x^2 + 0.1$ (lower panel).

Table 1 shows the simulated mean squared errors of the OLS and the efficient estimator for the regression parameters ϑ_1 and ϑ_2 for samples of size $n = 50$, $n = 100$ and $n = 200$. The simulations are based on 10,000 repetitions. The simulated mean squared error is computed as the average of the 10,000 sample mean squared errors $(\hat{\vartheta}_1 - 1)^2$ and $(\hat{\vartheta}_2 - 2)^2$. The efficient estimator requires an estimator of the conditional variance

$\sigma^2(x) = E(Y^2|X = x) - (E(Y|X = x))^2$. We estimated the two terms of $\sigma^2(x)$ by box kernels with bandwidths $b = 0.3, 0.6, 0.9$ and 1.2. (For $n = 50$ we skipped the case $b = 0.3$, for which the simulations were unstable; the same applies for $n = 100$ and $n = 200$ and bandwidths $b < 0.3$.) For comparison, we also computed the simulated mean squared error of the efficient estimator with the true $\sigma(x)$ in place of the estimated $\hat{\sigma}(x)$. As expected, it outperforms both the efficient estimator (with estimated $\sigma(x)$) and the ordinary least squares estimator. Nevertheless, the efficient estimator is noticeably better than the OLS for all choices of b (and in a few cases even very close to optimal). We expect that the performance of the efficient estimator can even be improved further if a more elaborate estimator of $\sigma(x)$ is chosen. A deliberate estimation approach will in particular be necessary if the covariate data contain outliers (e.g. if their distribution is close to normal) which can lead to very biased estimators of $\sigma(x)$, especially if the conditional variance at those data points is large.

Table 1 The table entries are the simulated mean squared errors of estimators $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$ of $\vartheta_1 = 1$ and $\vartheta_2 = 2$ in the regression model $r_\vartheta(x) = \vartheta_1 \cos(\vartheta_2 x)$ with (a) $\sigma(x) = 0.5x + 0.6$ and (b) $\sigma(x) = x^2 + 0.1$. We considered the ordinary least squares estimator (OLS), the efficient estimator with kernel estimators for $\sigma(x)$ and bandwidths $b = 0.3, 0.6, 0.9, 1.2$, and, for comparison, the efficient estimator (CP) using the true $\sigma(x)$.

Amplitude ϑ_1

σ	n	OLS	0.3	0.6	0.9	1.2	CP
(a)	50	0.020674	–	0.015551	0.015873	0.016186	0.012804
	100	0.009914	0.007113	0.006972	0.007376	0.007702	0.006053
	200	0.005157	0.003360	0.003559	0.003831	0.003995	0.003089
(b)	50	0.006247	–	0.001506	0.001733	0.002338	0.001201
	100	0.003031	0.000656	0.000653	0.000748	0.001014	0.000580
	200	0.001473	0.000301	0.000309	0.000348	0.000467	0.000286

Frequency ϑ_2

σ	n	OLS	0.3	0.6	0.9	1.2	CP
(a)	50	0.041415	–	0.008080	0.009865	0.012396	0.005311
	100	0.020047	0.003260	0.003117	0.004058	0.005354	0.002310
	200	0.009375	0.001205	0.001382	0.001815	0.002409	0.001104
(b)	50	0.049602	–	0.038141	0.043202	0.048663	0.026118
	100	0.023624	0.016170	0.016419	0.020000	0.022275	0.012550
	200	0.011537	0.006930	0.007928	0.009722	0.010721	0.006355

If the regression model is correctly specified, the weights of the asymptotically optimal estimator are uniquely determined (see estimating equation (3.3)). In the following we will therefore focus on the more interesting situation where the model is misspecified, i.e. when arbitrary weighted least squares estimators $\hat{\vartheta}$ are used which estimate, depending on the choice of weights, different functionals $t(P)$ of the distribution. It is clear that in most applications one can not find an explicit solution $t = t(P)$ of the defining equation (2.1), $E(w_t(X)r_t(X)) = E(w_t(X)r(X, Z))$, even if one chooses specific weights such as those used by the OLS, $w_t(X) = \dot{r}_t(X)^\top$.

The situation is different if the model r_ϑ has a linear structure, for example if we have a linear or polynomial regression model, which we will discuss in the following.

Linear regression and ordinary least squares. Consider the linear regression model $r_\vartheta(X) = \vartheta^\top X$ and the least squares estimator, i.e. the estimating equation (1.1) with weight vector $\dot{r}_t(X)^\top = X$. The defining equation (2.1) for $t(P)$ is $E(Xt^\top X) = E(XX^\top)t = E(Xr(X,Z)$. Hence, assuming $EXX^\top$ is invertible, the least squares estimator $\hat{\vartheta}$ estimates

$$t(P) = (EXX^\top)^{-1}E(Xr(X,Z)).$$

We now consider some special cases of misspecification.

1. Suppose the true model is an *additive* model $r(X,Z) = \alpha(X) + \beta(Z)$ involving a second covariate vector Z, where α and β are known or unknown functions. In this case $\hat{\vartheta}$ estimates

$$t(P) = (EXX^\top)^{-1}[E(X\alpha(X)) + E(X\beta(Z))].$$

2. Another candidate for a misspecified model is a special additive model, the *partially linear* model, $r(X,Z) = \vartheta^\top X + \beta(Z)$, where β is an unknown function of the second covariate vector Z. If the true model is partially linear, $\hat{\vartheta}$ estimates

$$t(P) = \vartheta + (EXX^\top)^{-1}E(X\beta(Z)). \tag{5.1}$$

Typically one can neither assume independence of the two covariate vectors X and Z nor make any further assumptions, such as $EX = 0$, which taken together would guarantee that the bias term was zero. However, if one knows about the type of the misspecification, and if the function β (and thus r) satisfies certain smoothness conditions, one can use this information to estimate ϑ consistently with a bias-corrected estimator $\tilde{\vartheta}$,

$$\tilde{\vartheta} = \hat{\vartheta} - \Big(\sum_{i=1}^n X_i X_i^\top\Big)^{-1} \sum_{i=1}^n X_i \hat{\beta}(Z_i), \tag{5.2}$$

where $\hat{\beta}$ is some nonparametric function estimator (see below for an illustration). Note that the model $r(X,Z) = \vartheta^\top X + \beta(Z)$ is also appropriate if the error in the linear regression model is misspecified. Then $\beta(Z)$ represents an additional additive error and Z an unobservable random variable. In this case it is plausible to assume $E\beta(Z) = 0$ and independence of X and Z. Then $\hat{\vartheta}$ is a consistent estimator of ϑ.

3. Another special case is given if the true model is, in fact, linear but involves *additional parameters* τ and covariates Z, i.e., $r(X,Z) = \vartheta^\top X + \tau^\top Z$. Then $\hat{\vartheta}$ estimates $t(P) = \vartheta + E(XX^\top)^{-1}E(XZ^\top)\tau$ and is consistent under obvious assumptions, e.g. if X and Z are independent with $EX = 0$ or $EZ = 0$. If not, the information can, in a similar way to the above, be used to construct a bias-corrected estimator $\tilde{\vartheta}$.

4. Now assume that the true model is *polynomial* of order q (without additional covariate Z). For simplicity let X be one-dimensional, so that the misspecified model is linear, $r_\vartheta(X) = \vartheta X$, and the true model is $r(X,Z) = r(X) = \vartheta X + \tau_1 X^2 + \ldots + \tau_{q-1} X^q$. In this case $\hat{\vartheta}$ estimates

$$t(P) \;=\; \vartheta + \frac{\sum_{i=2}^{q} \tau_{i-1} E(X^{i+1})}{EX^2}.$$

Hence $\hat{\vartheta}$ is, for example, consistent if the true model is quadratic and if the distribution of X is symmetric around zero.

For an illustration of the above we performed a simple simulation study for one-dimensional ϑ with $Y = X + \beta(Z) + \varepsilon$ as the true model, i.e. $r(X,Z) = \vartheta X + \beta(Z)$ with $\vartheta = 1$. We considered two cases, namely (a) $\beta(Z) = Z$ and (b) $\beta(Z) = Z^3$. By (5.1) the OLS $\hat{\vartheta}$ in the model $Y = \vartheta X + \varepsilon$ estimates (a) $t(P) = 1 + E(XZ)/EX^2$ and (b) $t(P) = 1 + E(XZ^3)/EX^2$. In order to consider the more interesting situation with a non-zero bias, we generated *dependent* covariates X and Z, namely X from a uniform distribution on $[0,2]$ and Z from a uniform distribution on $[X-1, X+1]$. The errors ε are standard normally distributed and independent of (X,Z). It is easy to see that in this example $\hat{\vartheta}$ is biased: it estimates (a) $t(P) = 2$ and (b) $t(P) = 1 + 17/5$. We therefore expect the simulated mean squared error of $\hat{\vartheta}$ to be dominated by the squared bias $(t(P) - \vartheta)^2 = (t(P) - 1)^2$ which is 1 in (a) and $(17/5)^2 = 11.56$ in (b). This is indeed supported by the data in the second column of Table 2, "OLS 1", which contains our simulation results for samples of size $n = 50, 100$ and 200.

Table 2 The table entries are the simulated mean squared errors of estimators of $\vartheta = 1$ in the regression model $Y = \vartheta X + \beta(Z) + \varepsilon$ with (a) $\beta(Z) = Z$ and (b) $\beta(Z) = Z^3$. We considered the OLS of ϑ in the model $Y = \vartheta X + \varepsilon$ (OLS 1) and in the model $Y = \vartheta X + \tau Z + \varepsilon$ (OLS 2), the bias-corrected estimator (5.2) using the true $\beta(Z)$ (CP) and estimator (5.2) using a box kernel $\hat{\beta}(Z)$ and different bandwidths $b = 0.1, \ldots, 1.5$. Note that $\hat{\beta}(Z)$ involves a pilot estimator (P) of ϑ which equals (5.2) if $b \approx 0$. The simulations are based on 20,000 repetitions.

(a)

n	OLS 1	OLS 2	CP	$b \approx 0$ (P)	$b = 0.5$	$b = 1.0$	$b = 1.5$
50	1.018891	0.001547	0.015312	0.154243	0.150550	0.146785	0.144919
100	1.007946	0.000384	0.007435	0.067519	0.066632	0.066154	0.067287
200	1.005097	0.000093	0.003722	0.031184	0.031034	0.031573	0.033589

(b)

n	OLS 1	OLS 2	CP	$b \approx 0$ (P)	$b = 0.1$	$b = 0.3$	$b = 0.5$
50	11.84135	0.035280	0.015333	0.203137	0.197967	0.197271	0.199133
100	11.70474	0.015979	0.007611	0.085443	0.082668	0.084776	0.093209
200	11.62956	0.007583	0.003721	0.042339	0.040499	0.044307	0.055980

We also calculated a simple bias-corrected estimator $\tilde{\vartheta}$ as suggested in (5.2). We estimated $\beta(Z)$ with a box kernel $\hat{\beta}(Z)$ based on "pseudo-observations"

$Y_i - \hat{\vartheta}_p X_i$ where $\hat{\vartheta}_p$ denotes a nonparametric pilot estimator for ϑ. (For the construction of $\hat{\vartheta}_p$ one can, for example, choose a variable W such that $E(W|Z) = 0$. Then $E(WY) = E(WX)\vartheta$ and $\vartheta = E(WY)/E(WX)$ can be estimated empirically. We took $W = X - E(X|Z)$ with $E(X|Z)$ replaced by a kernel estimator.) Note that here the OLS is simply $\hat{\vartheta} = \sum X_i Y_i / \sum X_i^2$. By (5.2), the bias-corrected estimator $\tilde{\vartheta}$ therefore equals the pilot estimator $\hat{\vartheta}_p$ if the kernel estimator $\hat{\beta}$ interpolates the data, i.e. if $\hat{\beta}(Z_i) = Y_i - \hat{\vartheta}_p X_i$, which is the case if the bandwidth b of $\hat{\beta}$ is very small, $b \approx 0$. The last four columns of Table 2 show the simulated mean squared errors of the bias-corrected estimator $\tilde{\vartheta}$ from (5.2) for different bandwidths b of $\hat{\beta}(Z)$ including $b \approx 0$ where $\tilde{\vartheta} = \hat{\vartheta}_p$ ("P" in Table 2). For a comparison we also simulated the mean squared errors of $\tilde{\vartheta}$ with the true $\beta(Z)$ inserted (see column "CP"). These are obviously significantly smaller than those of $\tilde{\vartheta}$ using kernel estimators $\hat{\beta}(Z)$, which are of similar order for different bandwidths b, indicating that the pilot estimator does not perform very well.

Finally, column "OLS 2" of Table 2 contains the simulated mean squared errors of the first component of the OLS $(\hat{\vartheta}, \hat{\tau})$ in the linear regression model $Y = \vartheta X + \tau Z + \varepsilon$ with $(\vartheta, \tau) = (1, 1)$ (which is the true regression model in case (a)). In (a), due to its optimality property as best linear unbiased estimator, this estimator, as expected, clearly outperforms the other estimators. This does not apply in (b) where the true model involves the cubic term $\beta(Z) = Z^3$. Here the values in column CP are smaller. We expect that the bias-corrected estimator with a better estimator of β would, besides being consistent, perform reasonably well.

Polynomial regression and ordinary least squares. In this paragraph we consider a polynomial model r_ϑ for the regression function, $r_\vartheta(X) = \vartheta_0 + \vartheta_1 X + \ldots + \vartheta_{p-1} X^{p-1}$, and assume we have no information about a possible misspecification, i.e. the true model is the nonparametric model $r(X, Z)$. The least squares estimator $\hat{\vartheta}$ uses the weight vector $w(X) = \dot{r}_\vartheta(X)^\top = (1, X, X^2, \ldots, X^{p-1})^\top$. Hence we can write $r_\vartheta(X) = \vartheta^\top w(X)$. The defining equation (2.1) for $t(P)$ is $E(w(X)w(X)^\top)t = E(w(X)r(X, Z))$. Hence the weighted least squares estimator estimates

$$t(P) \;=\; (E(w(X)w(X)^\top))^{-1} E(w(X)r(X, Z))$$

with $w(X) = (1, X, X^2, \ldots, X^{p-1})^\top$. Special misspecifications can now be studied analogously to the above on linear regression.

Linear regression and weighted least squares. We again consider linear regression, $r_\vartheta(X) = \vartheta^\top X$. The optimal estimator uses weights $w_t^*(X) = \dot{r}_t(X)^\top \hat{\sigma}(X)^{-2} = X\hat{\sigma}(X)^{-2}$. (Since estimators of σ need not depend on ϑ we have dropped the subscript t for this illustration.) If the true model is nonparametric, $E(Y|X, Z) = r(X, Z)$, $\hat{\vartheta}$ estimates

$$t(P) \;=\; (E(XX^\top \sigma(X)^{-2}))^{-1} E(Xr(X, Z)\sigma(X)^{-2}).$$

The situation is more complicated if we consider arbitrary weights, in particular if the weights depend on ϑ. The defining equation is then $t = (Ew_t(X)X^\top)^{-1}E(w_t(X)r(X,Z))$ and must be solved with respect to t. It is possible that $t(P)$ is not identifiable and $\hat{\vartheta}$ not a point but a set-valued estimator. A solution exists, for example, if the weights factorize, $w_t(X) = g(t)h(X)$. Then the weighted least squares estimator estimates $(Eh(X)X^\top)^{-1}E(h(X)r(X,Z))$.

In both situations, i.e. when the model is correctly specified and when it is misspecified, the choice of weights is of particular importance. If the parametric model holds, weighted least squares estimators $\hat{\vartheta}$ are always consistent, but only the estimator using optimal weights w_t^* is efficient for ϑ. If the model is misspecified, then the weights determine *what* $\hat{\vartheta}$ estimates: in order to estimate a particular functional $t(P)$, weights should be chosen carefully so that $t(P)$ is identifiable. If one has certain knowledge about a possible misspecification it may sometimes be advisable, in order to achieve consistency, to choose simple weights and/or use the additional information to construct a bias-corrected estimator, as suggested in (5.2) for linear regression.

Another question of interest in this context is: is it possible to choose weights such that $\hat{\vartheta}$ is robust against specific misspecifications, i.e. such that $\hat{\vartheta}$ is consistent for ϑ? There is no satisfactory answer yet, in particular no general answer. The above examples for linear regression have revealed that the ordinary least squares estimator is consistent under certain special assumptions on the misspecification and especially if the true model equals the parametric model except for an additional independent error. If the misspecification is described by such an additive structure, $r(X,Z) = r_\vartheta(X) + \beta(Z)$ with X and Z uncorrelated and $E\beta(Z) = 0$, then weighted least squares estimators are clearly consistent: $t = t(P) = \vartheta$ solves the defining equation (2.1), in this case $E(w_t(X)r_t(X)) = E(w_t(X)r_\vartheta(X))$. A similar conclusion was drawn by White (1981, Corollary 2.3), who also considered weighted least squares estimators (with weights not depending on the parameter). His parametric model is $Y = r_\vartheta(X)$ and his misspecified model is $Y = r_\vartheta(X) + \varepsilon$. The latter can be rewritten as our parametric model $E(Y|X) = r_\vartheta(X)$. Hence our parametric model is already flexible enough to cover a misspecification in form of an additive error. Consistency in a misspecified model of this type is thus not surprising, neither here nor in the above examples.

Acknowledgments. I thank a referee for suggestions that improved the exposition considerably.

References

Aguirre–Torres V and Domínguez MA (2004) Efficient method of moments in misspecified i.i.d. models. Econometric Theory 20: 513-534.

Andrews DWK and Pollard D (1994) An introduction to functional central limit theorems for dependent stochastic processes. Internat. Statist. Rev. 62: 119-132.

Bickel PJ, Klaassen CAJ, Ritov Y and Wellner JA (1998) Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.

Gould A and Lawless JF (1988) Consistency and efficiency of regression coefficient estimates in location-scale models. Biometrika 75: 535-540.

Greenwood PE and Wefelmeyer W (1997) Maximum likelihood estimator and Kullback-Leibler information in misspecified Markov chain models. Teor. Veroyatnost. i Primenen. 42: 169-178.

Hájek J (1970) A characterization of limiting distributions of regular estimates. Z. Wahrsch. Verw. Gebiete 14: 323-330.

Le Cam L (1960) Locally asymptotically normal families of distributions. Univ. California Publ. Statist. 3: 37-98.

Lindsay B and Qu A (2003) Inference functions and quadratic score tests. Statist. Sci. 18: 394-410.

McKean JW, Sheather SJ and Hettmansperger TP (1993) The use and interpretation of residuals based on robust estimation. J. Amer. Statist. Assoc. 88: 1254-1263.

Michaelis L and Menten ML (1913) Kinetik der Invertinwirkung. Biochemische Zeitschrift 49: 333-369.

Müller UU and Wefelmeyer W (2002a) Estimators for models with constraints involving unknown parameters. Math. Methods Statist. 11: 221-235.

Müller UU and Wefelmeyer W (2002b) Autoregression, estimating functions, and optimality criteria. In: Gulati C, Lin YX, Rayner J and Mishra S (eds.) Advances in Statistics, Combinatorics and Related Areas. World Scientific Publishing, Singapore, pp. 180-195.

Qu A, Lindsay BG and Li B (2000) Improving generalised estimating equations using quadratic inference functions. Biometrika 87: 823-836.

Ruppert D, Cressie N and Carroll RJ (1989) A transformation/weighting model for estimating Michaelis-Menten parameters. Biometrics 45: 637-656.

Sarkar N (1989) Comparisons among some estimators in misspecified linear models with multicollinearity. Ann. Inst. Statist. Math. 41: 717-724.

Severini TA (1998) Some properties of inferences in misspecified linear models. Statist. Probab. Lett. 40: 149-153.

Shi P, Ye JJ and Zhou J (2003) Minimax robust designs for misspecified regression models. Canad. J. Statist. 31: 397-414.

Struthers CA and Kalbfleisch JD (1986) Misspecified proportional hazard models. Biometrika 73: 363-369.

White H (1981) Consequences and detection of misspecified nonlinear regression models. J. Am. Stat. Assoc. 76: 419-433.

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1-26.

Zhang J and Liu A (2003) Local polynomial fitting based on empirical likelihood. Bernoulli 9: 579-605.