The Behavior of Estimators in Misspecified Regression Models

Priscilla E. Greenwood[1], Karl Küsters[2], and Wolfgang Wefelmeyer[2]

[1] Department of Mathematics, The University of British Columbia, Vancouver, B.C., Canada V6T 1Z2
[2] Mathematical Institute, University of Cologne, 50931 Cologne, Germany (email: wefelm@math.uni-koeln.de)

Abstract. We consider nonlinear regression models and the usual estimators for the regression parameter and the response distribution. We assume that the model is partly misspecified in one of three ways: either the covariate is not independent of the error, or the regression function is mismodelled, or the error is not centered. We determine what the estimators then estimate, and calculate their influence function and asymptotic variance.
Keywords: Nonlinear regression, misspecified model, least squares estimator, von Mises statistic.

1 Introduction

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent and identically distributed observations of a nonlinear regression model

$$Y = r_\vartheta(X) + \varepsilon$$

in which the error ε is centered in the sense that $E\varepsilon = 0$, and also independent of the covariate X. There are well-studied estimators for the regression parameter ϑ and the response distribution. How do these estimators behave when certain features of the model are misspecified? Do they remain consistent? If so, how is their influence function and asymptotic variance affected? If they do not remain consistent, what do the estimators then estimate, and what is their influence function and asymptotic variance?

We answer these questions separately for three types of misspecification: when the error is not independent of the covariate, when the parametric model for the regression function is incorrect, and when the error is not centered. We also consider the special case of *linear* regression.

Of course, these misspecifications could also happen simultaneously. We will not consider this here.

We note that the behavior of the estimators depends only on features of the true underlying distribution, not on properties of some true model that contains this distribution.

To simplify the notation, we take the parameter ϑ *one-dimensional.* Generalization to higher dimensions is straightforward. Derivatives with respect to the parameter will be denoted by a dot over the function, for example $\dot r_\vartheta(x) = \partial_\vartheta r_\vartheta(x)$.

2 Regression parameter

The classical estimator for the regression parameter ϑ is the *least squares estimator*. It solves the estimating equation

$$\sum_{i=1}^n \dot r_\vartheta(X_i)(Y_i - r_\vartheta(X_i)) = 0. \tag{1}$$

Leaving questions of uniqueness aside, the least squares estimator converges to the solution of the equation

$$E[\dot r_\vartheta(X)(Y - r_\vartheta(X))] = 0. \tag{2}$$

Correct model. For comparison with the misspecified case, we recall results when the nonlinear regression model is correct. Then the true ϑ solves (2), and hence the least squares estimator is consistent for ϑ. Moreover, a Taylor expansion of (1) gives

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^n \dot r_{\hat\vartheta}(X_i)(Y_i - r_{\hat\vartheta}(X_i)) \\ &= \frac{1}{n}\sum_{i=1}^n \dot r_\vartheta(X_i)\varepsilon_i - (\hat\vartheta - \vartheta)\frac{1}{n}\sum_{i=1}^n \dot r^2_\vartheta(X_i) + o_p(n^{-1/2}). \end{aligned}$$

Hence

$$\hat\vartheta = \vartheta + (E\dot r^2_\vartheta(X))^{-1}\frac{1}{n}\sum_{i=1}^n \dot r_\vartheta(X_i)\varepsilon_i + o_p(n^{-1/2}). \tag{3}$$

This means that the least squares estimator $\hat\vartheta$ is *asymptotically linear* with *influence function* $(E\dot r^2_\vartheta(X))^{-1}\dot r_\vartheta(X)\varepsilon$. By the central limit theorem, it is therefore asymptotically normal with variance $(E\dot r^2_\vartheta(X))^{-1}E\varepsilon^2$. The least squares estimator is not efficient. See Schick [2] for the construction of efficient estimators.

For the *linear* regression model, with $r_\vartheta(X) = \vartheta X$, we have $\dot r_\vartheta(X) = X$. Then the influence function of $\hat\vartheta$ is $(EX^2)^{-1}X\varepsilon$, and the asymptotic variance is $(EX^2)^{-1}E\varepsilon^2$.

Error depends on covariate. Suppose that X and ε are dependent, but that the model for the regression function is still correct. This means that $E(Y|X) = r_\vartheta(X)$, so $E(Y - r_\vartheta(X)|X) = 0$. It follows that the true ϑ still solves (2), and hence the least squares estimator $\hat\vartheta$ remains consistent for

ϑ. Then its stochastic expansion (3) also remains valid, but the asymptotic variance does not simplify as in our initial model with X and ε independent. It now is

$$(E\dot r_\vartheta^2(X))^{-2}E[\dot r_\vartheta^2(X)E(\varepsilon^2|X)].$$

For the *linear* regression model $r_\vartheta(X) = \vartheta X$, the influence function of $\hat\vartheta$ still is $(EX^2)^{-1}X\varepsilon$, now with asymptotic variance is $(EX^2)^{-2}E[X^2E(\varepsilon^2|X)]$.

Misspecified regression function. For this subsection, compare also White [4] and Stute *et al.* [3]. Suppose the true model is $Y = r(X) + \varepsilon$ with X and ε independent and $E\varepsilon = 0$, but the regression function is not of the form r_ϑ for some ϑ. For a given distribution P of (X, Y), let $\vartheta(P)$ denote the solution of (2). The least squares estimator estimates $\vartheta(P)$. Write $\varepsilon_P = Y - r_{\vartheta(P)}(X)$. This decomposes as

$$\varepsilon_P = \varepsilon + r(X) - r_{\vartheta(P)}(X).$$

In particular, ε_P is neither independent of X nor conditionally centered,

$$E(\varepsilon_P|X) = r(X) - r_{\vartheta(P)}(X).$$

We have $E[\dot r_\vartheta(X)\varepsilon] = 0$ for all ϑ, so equation (2) defining $\vartheta(P)$ can be rewritten as

$$E[\dot r_\vartheta(X)(r(X) - r_\vartheta(X))] = 0.$$

Hence $\vartheta = \vartheta(P)$ minimizes the mean squared distance $E[(r(X) - r_\vartheta(X))^2]$. The stochastic expansion of the estimating equation (1) now contains an additional term that involves the *second* derivative of the regression function,

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^n \dot r_{\hat\vartheta}(X_i)(Y_i - r_{\hat\vartheta}(X_i)) \;&=\; \frac{1}{n}\sum_{i=1}^n \dot r_{\vartheta(P)}(X_i)\varepsilon_{Pi} \\ -(\hat\vartheta - \vartheta(P))\Big(\frac{1}{n}\sum_{i=1}^n \dot r_{\vartheta(P)}^2(X_i) &- \frac{1}{n}\sum_{i=1}^n \ddot r_{\vartheta(P)}(X_i)\varepsilon_{Pi}\Big) + o_p(n^{-1/2}), \end{aligned}$$

hence

$$\hat\vartheta = \vartheta(P) + c^{-1}(P)\frac{1}{n}\sum_{i=1}^n \dot r_{\vartheta(P)}(X_i)\varepsilon_{Pi} + o_p(n^{-1/2})$$

with

$$c(P) = E\dot r_{\vartheta(P)}^2(X) - E[\ddot r_{\vartheta(P)}(X)\varepsilon_P].$$

Hence $\hat\vartheta$ is asymptotically linear with influence function $c^{-1}(P)\dot r_{\vartheta(P)}(X)\varepsilon_P$. With $E(\varepsilon_P^2|X)] = E\varepsilon^2 + (r(X) - r_{\vartheta(P)}(X))^2$, the asymptotic variance of $\hat\vartheta$ can be written as

$$c^{-2}(P)\big(E\dot r_{\vartheta(P)}^2(X)E\varepsilon^2 + E[\dot r_{\vartheta(P)}^2(X)(r(X) - r_{\vartheta(P)}(X))^2]\big).$$

For the *linear* regression model $r_\vartheta(X) = \vartheta X$ we obtain $\vartheta(P)$ as solution of $E[X(r(X) - \vartheta X)] = 0$, so

$$\vartheta(P) = (EX^2)^{-1} E[Xr(X)].$$

We have $\ddot{r}_\vartheta(X) = 0$ and $c(P) = EX^2$, so the influence function of $\hat{\vartheta}$ is $(EX^2)^{-1} X(Y - \vartheta(P)X)$, and its asymptotic variance is

$$(EX^2)^{-1} E\varepsilon^2 + (EX^2)^{-2} E[X^2(r(X) - \vartheta(P)X)^2].$$

Error is not centered. Suppose $E\varepsilon$ is not zero. Then the true model can be written as $Y = r_\tau(X) + \mu + \varepsilon$ with $E\varepsilon = 0$. This is a special case of a misspecified regression function $r(X) = r_\tau(X) + \mu$. The previous subsection remains unchanged otherwise.

For the *linear* regression model $r_\vartheta(X) = \vartheta X$ we obtain $\vartheta(P)$ as solution of $E[X(\tau X + \mu - \vartheta X)] = 0$, so

$$\vartheta(P) = \tau + \mu EX/EX^2.$$

We have $\ddot{r}_\vartheta(X) = 0$ and $c(P) = EX^2$, so the influence function of $\hat{\vartheta}$ is $(EX^2)^{-1} X(Y - \vartheta(P)X)$, and its asymptotic variance is

$$(EX^2)^{-1} E\varepsilon^2 + (EX^2)^{-2} E[X^2(\tau X + \mu - \vartheta(P)X)^2].$$

3 Response distribution

The usual estimator for an expectation $M = Eh(Y)$ is the empirical estimator $(1/n) \sum_{i=1}^n h(Y_i)$. It does not use the regression model and is therefore robust under any misspecification of the model. We can however estimate M using the regression model as follows. Since X and ε are independent, $Y = r_\vartheta(X) + \varepsilon$ is a convolution. Estimate ϑ by the *least squares estimator*. Estimate the error ε_i by the *residual* $\hat{\varepsilon}_i = Y_i - r_{\hat{\vartheta}}(X_i)$. An estimator for $M = Eh(r_\vartheta(X) + \varepsilon)$ is now given by the *von Mises statistic*

$$\hat{M} = \frac{1}{n^2} \sum_{i,j=1}^n h(r_{\hat{\vartheta}}(X_i) + \hat{\varepsilon}_j) = \frac{1}{n^2} \sum_{i,j=1}^n h(r_{\hat{\vartheta}}(X_i) - r_{\hat{\vartheta}}(X_j) + Y_j).$$

Correct model. For comparison with the misspecified case, we recall results when the nonlinear regression model is correct. We refer to Müller [1] who considers, more generally, the case when responses are missing at random. Then $\hat{\vartheta}$ converges to ϑ, so $\hat{M}$ converges to $Eh(r_\vartheta(X) + \varepsilon) = Eh(Y) = M$.

Hence $\hat{M}$ is consistent for M. In order to determine the influence function of $\hat{M}$, we first consider $\hat{M}$ as a function of $\hat{\vartheta}$ and expand it around ϑ,

$$\begin{aligned}
\hat{M} &= \frac{1}{n^2} \sum_{i,j=1}^{n} h(r_\vartheta(X_i) + \varepsilon_j) \\
&\qquad + (\hat{\vartheta} - \vartheta) \frac{1}{n^2} \sum_{i,j=1}^{n} (\dot{r}_\vartheta(X_i) - \dot{r}_\vartheta(X_j)) h'(r_\vartheta(X_i) + \varepsilon_j) + o_p(n^{-1/2}) \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} h(r_\vartheta(X_i) + \varepsilon_j) + (\hat{\vartheta} - \vartheta) H + o_p(n^{-1/2}) \qquad (4)
\end{aligned}$$

with

$$\begin{aligned}
H &= E[\dot{r}_\vartheta(X) h'(r_\vartheta(X) + \varepsilon)] - E\dot{r}_\vartheta(X) E h'(r_\vartheta(X) + \varepsilon) \\
&= E[\dot{r}_\vartheta(X) h'(Y)] - E\dot{r}_\vartheta(X) E h'(Y).
\end{aligned}$$

In order to obtain the influence function of $\hat{M}$, we use the Hoeffding decomposition

$$\frac{1}{n^2} \sum_{i,j=1}^{n} h(r_\vartheta(X_i) + \varepsilon_j) = M + \frac{1}{n} \sum_{i=1}^{n} \big(h_{X_i} - M + h_{\varepsilon_i} - M \big) + o_p(n^{-1/2}) \qquad (5)$$

with

$$\begin{aligned}
h_X &= E(h(r_\vartheta(X) + \varepsilon)|X) = E(h(Y)|X), \\
h_\varepsilon &= E(h(r_\vartheta(X) + \varepsilon)|\varepsilon) = E(h(Y)|\varepsilon).
\end{aligned}$$

The influence function of $\hat{\vartheta}$ was obtained as $(E\dot{r}_\vartheta^2(X))^{-1} \dot{r}_\vartheta(X)\varepsilon$ in Section 2. Hence $\hat{M}$ is asymptotically linear with influence function

$$h_X - M + h_\varepsilon - M + H(E\dot{r}_\vartheta^2(X))^{-1} \dot{r}_\vartheta(X)\varepsilon.$$

Since X and ε are independent, the conditional expectations can be written

$$h_X = \int h(r_\vartheta(X) + y) P^\varepsilon(dy), \qquad h_\varepsilon = \int h(r_\vartheta(x) + \varepsilon) dP^X(dx). \qquad (6)$$

For the *linear* regression model $r_\vartheta(X) = \vartheta X$, the influence function of $\hat{\vartheta}$ was obtained as $(EX^2)^{-1} X\varepsilon$. The influence function of $\hat{M}$ is therefore

$$h_X - M + h_\varepsilon - M + \big(E[Xh'(Y)] - EXEh'(Y) \big) (EX^2)^{-1} X\varepsilon$$

with $h_X = \int h(\vartheta X + y) dP^\varepsilon(dy)$ and $h_\varepsilon = \int h(\vartheta x + \varepsilon) dP^X(dx)$.

Error depends on covariate. If X and ε are dependent, but the model for the regression function is still correct, then the least squares estimator

$\hat\vartheta$ remains consistent, and its stochastic expansion (3) also remains valid. The von Mises statistic $\hat M$ is dominated by the terms with $i \neq j$. Hence it converges to

$$M_{12} = Eh(r_\vartheta(X_1) + \varepsilon_2) = Eh(Y_2 + r_\vartheta(X_1) - r_\vartheta(X_2)).$$

Here the time indices 1 and 2 indicate that $r_\vartheta(X_1)$ and ε_2 are *independent*. But X and ε are not independent any more, so in general M_{12} is now different from $M = Eh(Y) = Eh(r_\vartheta(X) + \varepsilon)$, and $\hat M$ does not remain consistent for M.

The stochastic expansion (4) of $\hat M$ remains unchanged, but we must now write

$$H = E[\dot r(X_1)h'(r_\vartheta(X_1) + \varepsilon_2)] - E[\dot r(X_2)h'(r_\vartheta(X_1) + \varepsilon_2)].$$

The Hoeffding decomposition (5) is valid with M_{12} in place of M. We must now write

$$h_{X_1} = E(h(r_\vartheta(X_1) + \varepsilon_2)|X_1), \qquad h_{\varepsilon_2} = E(h(r_\vartheta(X_1) + \varepsilon_2)|\varepsilon_2),$$

which is still the same as in (6).

For the *linear* regression model $r_\vartheta(X) = \vartheta X$ we have $\dot r_\vartheta(X) = X$, so $H = E[Xh'(Y)]$ and

$$h_X = E(h(Y)|X) = \int h(\vartheta X + y)P^\varepsilon(dy),$$
$$h_\varepsilon = E(h(Y)|\varepsilon) = \int h(\vartheta x + \varepsilon)P^X(dx).$$

The influence function of $\hat\vartheta$ is $(EX^2)^{-1}X\varepsilon$, so the influence function of $\hat M$ is

$$h_X - M + h_\varepsilon - M + H(EX^2)^{-1}X\varepsilon.$$

Misspecified regression function. Suppose the true model is $Y = r(X)+\varepsilon$ with X and ε independent and $E\varepsilon = 0$, but the regression function is not of the form r_ϑ for some ϑ. For a given distribution P of (X, Y), let $\vartheta(P)$ denote the solution of (2). As seen in Section 2, the least squares estimator $\hat\vartheta$ estimates $\vartheta(P)$. Set $\varepsilon_P = Y - r_{\vartheta(P)}(X)$. Then $\hat\varepsilon_j = Y_j - r_{\hat\vartheta}(X_j)$ approximates $\varepsilon_{Pj} = Y_j - r_{\vartheta(P)}(X_j)$. Since the von Mises statistic $\hat M$ is dominated by the terms with $i \neq j$, it converges to

$$M(P) = Eh(r_{\vartheta(P)}(X_1) + \varepsilon_{P2}) = Eh(Y_2 + r_{\vartheta(P)}(X_1) - r_{\vartheta(P)}(X_2)).$$

Here the different time indices 1 and 2 again indicate that the two random variables are independent. But $r_{\vartheta(P)}(X)$ and $\varepsilon_P = \varepsilon + r(X) - r_{\vartheta(P)}(X)$

are now in general *dependent*. Hence $M(P)$ differs from $M = Eh(Y) = Eh(r_{\vartheta(P)}(X) + \varepsilon_P)$, and $\hat{M}$ does not remain consistent for M.

In order to determine the influence function of $\hat{M}$, we proceed as in the correct model. We first consider $\hat{M}$ as a function of $\hat{\vartheta}$ and expand it around $\vartheta(P)$, now neglecting the terms with $i = j$,

$$\begin{aligned}\hat{M} = & \frac{1}{n^2}\sum_{i,j=1}^{n} h(r_{\vartheta(P)}(X_i) + \varepsilon_{Pj}) \\ & +(\hat{\vartheta} - \vartheta(P))\frac{1}{n^2}\sum_{i,j=1}^{n}(\dot{r}_{\vartheta(P)}(X_i) - \dot{r}_{\vartheta(P)}(X_j))h'(r_{\vartheta(P)}(X_i) + \varepsilon_{Pj}) \\ & +o_p(n^{-1/2}) \\ = & \frac{1}{n^2}\sum_{i,j=1}^{n} h(r_{\vartheta(P)}(X_i) + \varepsilon_{Pj}) + (\hat{\vartheta} - \vartheta(P))H(P) + o_p(n^{-1/2})\end{aligned}$$

with

$$\begin{aligned}H(P) = & E[r_{\vartheta(P)}(X_1)h'(r_{\vartheta(P)}(X_1) + \varepsilon_{P2})] \\ & -E[r_{\vartheta(P)}(X_2)h'(r_{\vartheta(P)}(X_1) + \varepsilon_{P2})].\end{aligned}$$

Even though $r_{\vartheta(P)}(X)$ and $\varepsilon_P = \varepsilon_j + r(X) - r_{\vartheta(P)}(X)$ are now dependent, $r_{\vartheta(P)}(X_i)$ and $\varepsilon_{Pj} = \varepsilon_j + r(X_j) - r_{\vartheta(P)}(X_j)$ are still independent for $i \neq j$. Hence we still have a Hoeffding decomposition

$$\begin{aligned}& \frac{1}{n^2}\sum_{i,j=1}^{n} h(r_{\vartheta(P)}(X_i) + \varepsilon_{Pj}) \\ & = M(P) + \frac{1}{n}\sum_{i=1}^{n}\big(h_{PX_i} - M(P) + h_{P\varepsilon_{Pi}} - M(P)\big) + o_p(n^{-1/2})\end{aligned}$$

with

$$\begin{aligned}h_{PX_1} &= E(h(r_{\vartheta(P)}(X_1) + \varepsilon_{P2})|X_1), \\ h_{P\varepsilon_{P2}} &= E(h(r_{\vartheta(P)}(X_1) + \varepsilon_{P2})|\varepsilon_{P2}).\end{aligned}$$

The influence function of $\hat{\vartheta}$ was obtained in Section 2 as $c^{-1}(P)\dot{r}_{\vartheta(P)}(X)\varepsilon_P$ with

$$c(P) = E\dot{r}^2_{\vartheta(P)}(X) - E[\ddot{r}_{\vartheta(P)}(X)\varepsilon_P].$$

Taken together, $\hat{M}$ estimates $M(P)$ and is asymptotically linear with influence function

$$h_{PX} - M(P) + h_{P\varepsilon_P} - M(P) + H(P)c^{-1}(P)\dot{r}_{\vartheta(P)}(X)\varepsilon_P.$$

As seen in Section 2, for the *linear* regression model $r_\vartheta(X) = \vartheta X$ we have $\vartheta(P) = (EX^2)^{-1}E[Xr(X)]$. So

$$\begin{aligned} M(P) &= Eh(\vartheta(P)X_1 + \varepsilon_{P2}) \\ &= Eh(Y_2 + (EX^2)^{-1}E[Xr(X)](X_1 - X_2)). \end{aligned}$$

Furthermore, $\hat{\vartheta}$ estimates $\vartheta(P)$ and has influence function $(EX^2)^{-1}X(Y - \vartheta(P)X)$. With these notations, $\hat{M}$ estimates $M(P)$ and is asymptotically linear with influence function

$$h_{PX} - M(P) + h_{P\varepsilon_P} - M(P) + H(P)(EX^2)^{-1}X(Y - \vartheta(P)X)$$

with

$$h_{PX_1} = E(h(\vartheta(P)X_1 + \varepsilon_{P2})|X_1), \quad h_{P\varepsilon_{P2}} = E(h(\vartheta(P)X_1 + \varepsilon_{P2})|\varepsilon_{P2})$$

and

$$H(P) = E[\vartheta(P)X_1 h'(\vartheta(P)X_1 + \varepsilon_{P2})] - E[\vartheta(P)X_2 h'(\vartheta(P)X_1 + \varepsilon_{P2})].$$

Error is not centered. If $E\varepsilon$ is not zero, the true model can be written $Y = r_\tau(X) + \mu + \varepsilon$ with $E\varepsilon = 0$. This is a special case of a misspecified regression function $r(X) = r_\tau(X) + \mu$. The previous subsection remains unchanged otherwise.

For the *linear* regression model $r_\vartheta(X) = \vartheta X$ we have obtained $\vartheta(P) = \tau + \mu EX/EX^2$. This means that $\hat{M}$ now estimates

$$\begin{aligned} M(P) &= Eh\Big(\Big(\tau + \mu\frac{EX}{EX^2}\Big)X_1 + \varepsilon_{P2}\Big) \\ &= Eh\Big(Y_2 + \Big(\tau + \mu\frac{EX}{EX^2}\Big)(X_1 - X_2)\Big). \end{aligned}$$

The results of the previous subsection concerning *linear* regression remain unchanged otherwise.

References

1. Müller, Ursula U. Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, 37, 2245–2277, 2009.
2. Schick, A. On efficient estimation in regression models. *Ann. Statist.*, 21, 1486–1521, 1993. Correction 23, 1862–1863, 1995.
3. Stute, W., Xu, W. L. and Zhu, L. X. Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, 95, 451–467, 2008.
4. White, H. Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.*, 76, 419–33, 1981.