The Lineartiy Condition and Adaptive Estimation in Single-index Regressions

Yongwu Shao*† Dennis Cook†
Sanford Weisberg
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

Sep 15, 2006

Abstract

We show that under a linearity condition on the distribution of the predictors, the coefficient vector in a single-index regression can be estimated with the same efficiency as in the case when the link function is known. Thus, the linearity condition seems to substitute for knowing the exact conditional distribution of the response given the linear combinations of the predictors.

1 Introduction

1.1 Single-index Regressions

Consider a continuous univariate response Y and a vector of continuous predictors $X \in \mathbb{R}^p$. The most general goal of a regression is to infer about the conditional distribution of $Y|X$. In this paper we consider single-index regressions, in which $Y|X$ depends on X through at most one linear combination $\beta_0^{\mathrm{T}} X$ of the predictors.

Focusing on the mean function $\mathrm{E}(Y|X)$, Härdle and Stoker (1989) developed a nonparametric method called average derivative estimation for estimating β_0 in the single-index conditional mean $\mathrm{E}(Y|X) = g(\beta_0^{\mathrm{T}} X)$, where the mean function

*Corresponding author. E-mail: ywshao@stat.umn.edu

†R. D. Cook and Y. Shao were suppoted by Grant DMS-0405360 from the National Science Foundation.

g is unknown. Weisberg and Welsh (1994) considered the case in which $Y|X$ follows a generalized linear model, where the linear coefficient β_0 and the link function are unknown. Both pairs of authors gave estimates for β_0 that are $\sqrt{n}$-consistent.

Yin and Cook (2005) proposed the problem of single-index regressions, in which the conditional distribution of $Y|X$ is completely characterized by a linear combination $\beta_0^{\mathrm{T}}X$, so there is no loss of information about Y if we replace X with $\beta_0^{\mathrm{T}}X$. More specifically, we assume that

$$Y \perp\!\!\!\perp X|\beta_0^{\mathrm{T}}X \tag{1}$$

where for identifiability purposes, we require that $||\beta_0|| = 1$. (1) is equivalent to the statement that $Y|X$ has a conditional density $\eta_0(y|\beta_0^{\mathrm{T}}x)$, where η_0 is unknown. Single-index regression is a special case of sufficient dimension reduction when the dimension of the central subspace (Cook, 1996) is one. It does not require a pre-specified single-index model.

We show that under the linearity condition (Li and Duan, 1989), for single-index regressions there exists an adpative estimate for β_0 that can be estimated with the same efficiency as the maximum likelihood estimate when the conditional density η_0 is completely specified. For example, if the true model is $Y = g(\beta_0^{\mathrm{T}}X) + \epsilon$, where the link function g and the density of the error ϵ are unknown, then β_0 can be estimated with the same efficiency as in the case when g and the error distribution are known.

1.2 Linearity Condition

Many sufficient dimension reduction methods require the linearity condition: $E(X|\beta_0^{\mathrm{T}}X)$ is a linear function of $\beta_0^{\mathrm{T}}X$ (Li and Duan, 1989). It is used in popular methods like sliced inverse regression (Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991) and principal Hessian directions (Li, 1992, Cook, 1998).

The linearity condition holds if the predictor has an elliptically distribution (Eaton, 1986), so it holds when X has a multivariate normal distribution. Moreover, Diaconis and Freedman (1984) showed that most low-dimension projections of a high-dimension data cloud are close to being normal. Hall and Li (1993) argued that the linearity condition holds approximately when p is large. The linearity condition applies only to the marginal distribution of the predictors and not to the conditional distribution of $Y|X$ as is common in regression modeling. Consequently at the stage of data collection, we might design the experiment so that

the distribution of X will not blatantly violate elliptic symmetry. We can also transform the predictors to normality, or we can re-weight the data (Cook and Nachtscheim, 1994) to approximate an ellliptical distribution.

1.3 Adaptive Estimation

The problem of adaptive estimation was introduced by Stein (1956). One wishes to estimate a Euclidean parameter θ in the presence of an infinite-dimensional shape parameter G, usually the density. An adaptive estimate performs asymptotically as well with G unknown as the maximum likelihood estimate does when G is known (Bickel, 1982). A general method of constructing adaptive estimates was constructed by Bickel (1982). Schick (1986, 1993) generalized and improved Bickel's method.

It has been shown that adaptive estimation is possible in the symmetric location problem, in which we need to estimate the center of symmetry of an unknown distribution (Stone, 1975). It is also possible in linear regressions where the error density is symmetric and unknown and we need to estimate the linear coefficient (Bickel, 1982). When the observations are not independent, Koul and Pflug (1990), Schick (1993), Koul and Schick (1996) showed adaptive estimation is possible in certain autoregressive models. Early literature on adaptive estimation generally focused on these models and their generalizations. In this paper we show that under the linearity condition, adaptive estimation is also possible for single-index regressions.

2 Main Results

Without loss of generality we assume that X has mean zero and covariance I_p. We also assume that $\beta_0 \in \Theta$, where

$$\Theta = \{\beta \in \mathbb{R}^p : ||\beta|| = 1\}.$$

Let $l(t, y) = (\partial/\partial t)\eta_0(y|t)/\eta_0(y|t)$ be the derivative of the log density or equivalently the log likelihood. By using a Lagrange multiplier, the score equation for β_0 is

$$Q_{\beta_0}\mathrm{E}[Xl(\beta_0^{\mathrm{T}}X, Y)] = 0, \tag{2}$$

where $Q_\zeta = I_p - P_\zeta$ and P_ζ is the orthogonal projection onto the subspace spanned by the columns of the matrix ζ.

It can be shown that (2) holds not only for l, but for any $f(\cdot, \cdot) \in \mathbb{R}$.

LEMMA 1. *Assume that the linearity condition holds. Assume $f(\cdot,\cdot) \in \mathbb{R}$. Then β_0 is a solution of the equation*

$$Q_\beta \mathrm{E}[X f(\beta^{\mathrm{T}} X, Y)] = 0. \tag{3}$$

Proof. Since X has covariance matrix I_p, according to Cook (1998, pp. 57), we have $\mathrm{E}[Q_{\beta_0} X | \beta_0^{\mathrm{T}} X] = 0$. Therefore

$$\begin{aligned} Q_{\beta_0}\mathrm{E}[X f(\beta_0^{\mathrm{T}} X, Y)] &= \mathrm{E}[Q_{\beta_0} X f(\beta_0^{\mathrm{T}} X, Y)] \\ &= \mathrm{E}\{\mathrm{E}[Q_{\beta_0} X f(\beta_0^{\mathrm{T}} X, Y)] | \beta_0^{\mathrm{T}} X\} \\ &= \mathrm{E}\{\mathrm{E}[Q_{\beta_0} X | \beta_0^{\mathrm{T}} X]\mathrm{E}[f(\beta_0^{\mathrm{T}} X, Y) | \beta_0^{\mathrm{T}} X]\} \\ &= 0 \end{aligned}$$

□

The above lemma shows that a misspecified l still produces a Fisher consistent estimate of β_0. According to van der Vaart (1998, Theorem 25.27), Lemma 1 together with some regularity conditions would enable us to construct an adaptive estimate for β_0. The regularity conditions are typically satisfied in practice. A proof of the following theorem is given in the appendix.

THEOREM 1. *Assume that the Fisher information $\mathcal{I}(\beta) = \mathrm{E}_\beta[X X^{\mathrm{T}} l^2(\beta^{\mathrm{T}} X, Y)]$ is finite, nonsingular and differentiable with respect to β in a neighborhood of β_0. Let $\hat{l}_n(t, y)$ be an estimate of $l(t, y)$ that satisfies*

$$\mathrm{E}_{\beta_0}[||X||^2(\hat{l}(\beta_0^{\mathrm{T}} X, Y) - l(\beta_0^{\mathrm{T}} X, Y))^2] = o_p(1). \tag{4}$$

Then under the linearity condition we can construct an adaptive estimate of β_0 in (1) based on $\hat{l}_n(t, y)$.

Following van der Vaart (1998, pp. 393), an adaptive estimate can be constructed in the following way. Suppose β_n is a $\sqrt{n}$-consistent estimate of β_0. For instance, under the linearity condition β_n can be chosen as the ordinary least squares estimator (Li and Duan, 1989). Let Γ_n be a $p \times (p-1)$ matrix such that (Γ_n, β_n) is an orthogonal matrix. Let

$$\tilde{\mathcal{I}}_n = \sum_{i=1}^{n} [X_i X_i^{\mathrm{T}} \hat{l}_n^2(\beta_n^{\mathrm{T}} X_i, Y_i)]$$

be an estimator of the information matrix for β. Let $\hat{\beta}_n$ be a one-step iteration of the Newton-Raphson algorithm for solving the equation

$$Q_\beta \sum_{i=1}^{n} [X_i \hat{l}_n(\beta^{\mathrm{T}} X_i, Y_i)] = 0$$

with respect to β on the manifold Θ, starting at the initial guess β_n. We can write $\hat{\beta}_n$ as

$$\hat{\beta}_n = \beta_n + \frac{1}{n} \Gamma_n [\Gamma_n^{\mathrm{T}} \tilde{\mathcal{I}}_n \Gamma_n]^{-1} \Gamma_n^{\mathrm{T}} \sum_{i=1}^{n} [X_i \hat{l}_n(\beta_n^{\mathrm{T}} X_i, Y_i)] \tag{5}$$

Van der Vaart (1998, Theorem 25.27) showed that, by using discretization and sample-splitting devices, $\hat{\beta}_n$ is an adaptive estimate of β_0 if $\hat{l}$ satisifies (4). One such $\hat{l}$ based on the kernel density estimation in Härdle and Stoker (1989) is constructed in the Appendix.

Since $\hat{\beta}_n$ is an adaptive estimator, it has the same asymptotic distribution as the maximum likelihood estimator. Next we will derive the asymptotic distribution of the maximum likelihood estimator. Let $\hat{\beta}_{\mathrm{mle}}$ be the maximum likelihood estimator of β_0. It is shown in the appendix that under mild regularity conditions $\hat{\beta}_{\mathrm{mle}}$ has the following asymptotic distribution.

THEOREM 2. *Assume that the regularity conditions for the asymptotic normality of the maximum likelihood estimate hold. Then*

$$\hat{\beta}_{\mathrm{mle}} = \beta_0 + \frac{1}{n} \Gamma_0 [\Gamma_0^{\mathrm{T}} \mathcal{I}(\beta_0) \Gamma_0]^{-1} \Gamma_0^{\mathrm{T}} \sum_{i=1}^{n} [X_i l(\beta_0^{\mathrm{T}} X_i, Y_i)] + o_p(n^{-1/2}) \tag{6}$$

where Γ_0 is a $p \times (p-1)$ matrix such that (Γ_0, β_0) is an orthogonal matrix.

Since $\hat{\beta}_n$ is an adaptive estimator, it has the same asymptotic distribution as $\hat{\beta}_{\mathrm{mle}}$, we conclude that $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges to a normal distribution with zero mean and covariance matrix equal to the covariance matrix of $\Gamma_0 [\Gamma_0^{\mathrm{T}} \mathcal{I}(\beta_0) \Gamma_0]^{-1} \Gamma_0^{\mathrm{T}} X l(\beta_0^{\mathrm{T}} X, Y)]$.

3 Discussion

In this article we showed that under the linearity condition, there exists an adaptive estimate of the coefficient vector in a single-index regression. From this result we can see the important role of the lineartiy condition in single-index regression,

and more generally, in sufficient dimension reduction. The linearity condition is unusual, as it does not occur commonly outside of sufficient dimension reduction. We have shown that the linearity condition asymptotically takes the place of a known density. We conjecture that if the linearity condition fails, then an adaptive estimate does not exist. As a consequence, the coefficient vector cannot be estimated as well as it can be with the maximum likelihood estimator.

Appendix

Proof of Theorem 1. Since Lemma 1 holds, according to van der Vaart (1998, Theorem 25.27), we only need to prove the following two statements.

1. The conditional density $\eta_0(y|\beta_0^{\mathrm{T}}x)$ is differentiable in quadratic mean with respect to β_0.

2. Let $h(\beta^{\mathrm{T}}x, y)$ be the joint density of $\beta^{\mathrm{T}}X$ and Y, then

$$\int ||x||^2 \left[l(\beta_n^{\mathrm{T}}x, y)\sqrt{h(\beta_n^{\mathrm{T}}x, y)} - l(\beta_0^{\mathrm{T}}x, y)\sqrt{h(\beta_0^{\mathrm{T}}x, y)} \right]^2 dxdy \to 0.$$

The first statement is true by van der Vaart (1998, Theorem 7.2). So we only need to prove the second statement.

Since

$$\mathcal{I}(\beta_0) = \int xx^{\mathrm{T}} \left[l(\beta_0^{\mathrm{T}}x, y)\sqrt{h(\beta_0^{\mathrm{T}}x, y)} \right]^2 dxdy$$

and

$$\mathcal{I}(\beta_n) = \int xx^{\mathrm{T}} \left[l(\beta_n^{\mathrm{T}}x, y)\sqrt{h(\beta_0^{\mathrm{T}}x, y)} \right]^2 dxdy$$

By the assumptions, $\mathcal{I}(\beta)$ is continuous on a neighborhood of β_0 and $\mathcal{I}(\beta_0)$ is finite, we conclude that $\mathcal{I}(\beta_n)$ is also finite, hence $\mathrm{tr}[\mathcal{I}(\beta_0) + \mathcal{I}(\beta_n)] < \infty$. By the triangular inequality,

$$\begin{aligned} & \int ||x||^2 \left[|l(\beta_n^{\mathrm{T}}x, y)|\sqrt{h(\beta_n^{\mathrm{T}}x, y)} + |l(\beta_0^{\mathrm{T}}x, y)|\sqrt{h(\beta_0^{\mathrm{T}}x, y)} \right]^2 dxdy \\ \leq \;& \mathrm{tr}[\mathcal{I}(\beta_0) + \mathcal{I}(\beta_n)] < \infty \end{aligned}$$

Then by the dominate convergence theorem,

$$\int ||x||^2 \left[l(\beta_n^{\mathrm{T}}x, y)\sqrt{h(\beta_n^{\mathrm{T}}x, y)} - l(\beta_0^{\mathrm{T}}x, y)\sqrt{h(\beta_0^{\mathrm{T}}x, y)} \right]^2 dxdy \to 0$$

Therefore the second statement is also true.

□

Proof. of Theorem 2. We first transform the manifold Θ to $\mathbb{R}^{p-1}$ by using the following linear transformation. For any $\beta \in \Theta$, let $\alpha = \varphi(\beta) = \Gamma_0\beta$. Then $\beta = \varphi^{-1}(\alpha) = \Gamma_0\alpha + (1 - ||\alpha||^2)\beta_0$, and $\eta_0(y|\beta^{\mathrm{T}}x) = \eta_0(y|\varphi^{-1}(\alpha)^{\mathrm{T}}x)$. By taking the derivative of $\eta_0(y|\varphi^{-1}(\alpha)^{\mathrm{T}}x)$ with respect to α, we can derive the asymptotic distribution of the maximum likelihood estimate for α as following,

$$\hat{\alpha}_{\mathrm{mle}} = \frac{1}{n}[\Gamma_0^{\mathrm{T}}\mathcal{I}(\beta_0)]^{-1}\Gamma_0^{\mathrm{T}}\sum_{i=1}^{n}[X_i l(\beta_0^{\mathrm{T}}X_i, Y_i)] + o_p(n^{-1/2})$$

By the delta method, we have

$$\hat{\beta}_{\mathrm{mle}} = \beta_0 + \frac{1}{n}\Gamma_0[\Gamma_0^{\mathrm{T}}\mathcal{I}(\beta_0)\Gamma_0]^{-1}\Gamma_0^{\mathrm{T}}\sum_{i=1}^{n}[X_i l(\beta_0^{\mathrm{T}}X_i, Y_i)] + o_p(n^{-1/2})$$

□

Construction of $\hat{l}$ that satisfies (4). Let $h(\beta_0^{\mathrm{T}}x, y)$ be the joint density of $(\beta_0^{\mathrm{T}}X, Y)$, and $g(\beta_0^{\mathrm{T}}x)$ be the density of $\beta_0^{\mathrm{T}}X$, then

$$\eta_0(y|\beta_0^{\mathrm{T}}x) = h(\beta_0^{\mathrm{T}}x, y)/g(\beta_0^{\mathrm{T}}x)$$

and

$$l = h'/h - g'/g,$$

where h', g' are the derivative of h, g w.r.t. the first argument. To estimate l, we only need to estimate h'/h and g'/g.

We only consider the estimation of g'/g in detail here, because h'/h can be estimated in the same way, except that the dimension of the density estimation is different. Let d be the dimension of the density estimation, $d = 1$ for g and $d = 2$ for h.

Let $T_i = \beta_0^{\mathrm{T}}X_i$. For a fixed twice continuously differentiable probability density w with compact support, a bandwidth parameter σ, and a cut-off tuning parameter δ, set

$$\begin{aligned} \hat{g}_n(s) &= \sigma_n^{-d}\sum_{i=1}^{n} w(\frac{s - T_i}{\sigma_n}) \\ \hat{\xi}_n(s) &= \frac{\hat{g}_n'}{\hat{g}_n}(s)1_{\hat{g}_n(s)>\delta} \end{aligned} \tag{7}$$

where $\hat{\xi}_n(s)$ is our estimator of $g'(s)/g(s)$. Then $\mathrm{E}[(\hat{\xi}_n(X) - g'(X)/g(X))^2||X||^2]$ converges to zero in probability provided $\delta \uparrow \infty$ and $\sigma \downarrow 0$ at appropriate speeds.

Hardle and Stoker, 1991, page 992) showed that under some regularity conditions we have for any $\epsilon > 0$,

$$\sup[|\hat{g}(s) - g(s)|1_{g(s)>(\delta/2)}] = O_p[(n^{1-(\epsilon/2)}\sigma^d)^{-1/2}]$$

and

$$\sup[|\hat{g}'(s) - g'(s)|1_{g(s)>(\delta/2)}] = O_p[(n^{1-(\epsilon/2)}\sigma^{d+2})^{-1/2}]$$

Therefore

$$\sup[|(\hat{g}'/\hat{g}) - (g/g)|1_{g>(\delta/2)}] = O_p[\delta^{-2}(n^{1-(\epsilon/2)}\sigma^{d+2})^{-1/2}]$$

Hence for large n we have

$$\begin{aligned} & \mathrm{E}[(\hat{\xi}_n - g')^2||X||^2] \\ = \; & \mathrm{E}[(g'/g)^2||X||^2 1_{\hat{g}<\delta}] + \mathrm{E}[((\hat{g}'/\hat{g}) - (g'/g))^2||X||^2 1_{\hat{g}>\delta}] \\ \leq \; & \mathrm{E}[(g'/g)^2||X||^2 1_{g<2\delta}] + \mathrm{E}[((\hat{g}'/\hat{g}) - (g'/g))^2||X||^2 1_{g>(\delta)/2}] \\ \leq \; & \mathrm{E}[(g'/g)^2||X||^2 1_{g<2\delta}] + O_p[\delta^{-2}(n^{1-(\epsilon/2)}\sigma^{d+2})^{-1/2}] \cdot \mathrm{E}[||X||^2] \end{aligned}$$

Assume that $\mathrm{E}[||X||^2] < \infty$. Since $(g'/g)^2||X||^2 1_{g<2\delta}$ is dominated by $(g'/g)^2||X||^2$, and $\mathrm{E}[(g'/g)^2||X||^2 1_{g<2\delta}]$ is finite by assumptions, therefore $\mathrm{E}[(g'/g)^2||X||^2 1_{g<2\delta}]$ converges to zero when δ goes to zero. By the assumptions, $\delta^{-2}(n^{1-(\epsilon/2)}\sigma^{d+2})^{-1/2}$ also converges to zero, therefore $\mathrm{E}[(\hat{\xi}_n - (g'/g))^2||X||^2] = o_p(1)$.

In the same fashion we can construct $\hat{\zeta}_n(s, y)$ to estimate h'/h, except that we use (T_i, Y_i) as observations. Then an estimator for l can be defined as

$$\hat{l}_n = \hat{\zeta}_n - \hat{\xi}_n. \tag{8}$$

Since $\mathrm{E}[(\hat{\xi}_n - g'/g)^2||X||^2]$ and $\mathrm{E}[(\hat{\zeta}_n - h'/h)^2||X||^2]$ converges to zero in probability, we have $\mathrm{E}[(\hat{l}_n(\beta_0^{\mathrm{T}}X, Y) - f(\beta_0^{\mathrm{T}}X, Y))^2||X||^2]$ converges to zero in probability, and (4) is satisfied. □

References

Bickel, P. J. (1982) On adaptive estimation. Ann. of Statist., 10, 647-671.

Cook, R. D. (1996) Graphics for regressions with a binary response. J. Amer. Statist. Assoc., 91, 983-992.

Cook, R. D. (1998) Principle Hessian directions revisited (with discussion). J. Amer. Statist. Assoc., 93, 84-100.

Cook, R. D. and Nachtsheim, C. J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc., 89, 592-600.

Cook, R. D. and Weisberg, S. (1991) Discussion of "sliced inverse regression for dimension reduction". J. Amer. Statist. Assoc., 86, 328-332.

Diaconis, P. and Freedman, D. (1984) Asymptotics of graphical projection pursuit. Ann. of Statist., 12, 793-815.

Eaton, M. L. (1986) A characterization of spherical distributions. J. Mult. Anal., 20, 272-6.

Hall, P. and Li, K. C. (1993) On almost linearity of low dimensional projections from high dimensional data. Ann. of Statist., 21, 867-889.

Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by the method of average derivatives. J. Amer. Statist. Assoc., 84, 986-995.

Koul, H. L. and Pflug, G. (1990) Weakly adatptive estimators in explosive regression. Ann. of Statist., 18, 939-960.

Koul, H. L. and Schick, A. (1996) Adaptive estimation in a random coefficient autoregressive model. Ann. of Statist., 24, 1025-1052.

Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc., 86, 316-342.

Li, K. C. (1992) On Principle Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. J. Amer. Statist. Assoc., 87, 1025-1039.

Li, K. C. and Duan, N. (1989) Regression analysis under link violation. Ann. of Statist., 17, 1009-1052.

Schick, A. (1986) On asymptotically efficient estimation in semi-parametric models. Ann. of Statist., 14, 1139-1151.

Schick, A. (1993) On efficient estimation in regression models. Ann. of Statist., 21, 1481-1521.

Stein, C. (1956) Efficient nonparametric testing and estimation. Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 187-196. Unversity of California Press.

Stone, C. J. (1975) Adaptive maximum likelihood estimators of a location parameter. Ann. of Statist., 3, 276-284.

van der Vaart, A. W. (1998) Asmyptotic Statistics. Cambridge.

Weisberg, S. and Welsh, A. H. (1994) Adapting for the missing link. Ann. of Statist., 22, 1674-1700.

Yin, X. and Cook, R. D. (2005) Direction estimation in single-index regressions. Biometrika, 92, 371-384.