Generated Covariates in Nonparametric Estimation: A Short Review.

Enno Mammen, Christoph Rothe, and Melanie Schienle

Abstract In many applications, covariates are not observed but have to be estimated from data. We outline some regression-type models where such a situation occurs and discuss estimation of the regression function in this context. We review theoretical results on how asymptotic properties of nonparametric estimators differ in the presence of generated covariates from the standard case where all covariates are observed. These results also extend to settings where the focus of interest is on average functionals of the regression function.

1 Introduction

Consider a nonparametric regression model of the form

$$Y = m_0(R) + \varepsilon,$$
$$E[\varepsilon|R] = 0$$

where Y is a one-dimensional response variable and R is a q-dimensional covariate vector. The statistical goal is to nonparametrically estimate the regression function $m_0 : \mathbf{R}^q \to \mathbf{R}$ or a functional of the regression function, e.g. a weighted average $T(m_0) = \int m_0(x)w(x)dx$. We consider the case where the covariate R is unobserved

Enno Mammen
Department of Economics, University of Mannheim, D-68131 Mannheim, Germany, e-mail: emammen@rumms.uni-mannheim.de

Christoph Rothe
Toulouse School of Economics, 21 Alle de Brienne, F-31000 Toulouse, France e-mail: rothe@cict.fr

Melanie Schienle
School of Business and Economics, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany e-mail: melanie.schienle@wiwi.hu-berlin.de

but an estimator $\widehat{R}$ of R is available. In this note, we provide some examples where such a situation occurs. Furthermore, appropriate forms of nonparametric estimators of m_0 are discussed and results on their asymptotic distribution are reviewed. In particular, we analyse how the real feasible estimator of m_0 obtained via regression on $\widehat{R}$ differs from the infeasible one obtained by regressing on R. With stochastic expansions for the difference of these two estimators, the asymptotic distribution of the real estimator of m_0 can be accurately described.

The note is organized as follows. In the next section, some examples illustrate how and where generated covariates typically appear in practice. Section 3 provides an overview of the asymptotic theory when m_0 is estimated by local linear estimation. In particular, the theory can also be applied to cases where the main interest is in averages of the regression function m_0, which is also important for some of the stated examples.

2 Examples

2.1 Simultaneous Nonparametric Equation Models without Additivity (Imbens and Newey, 2009)

In economic models, there are often unobserved covariates which affect both response and observed covariates. Generally, such covariates which are correlated with the disturbance are called endogenous. Imbens and Newey (2009) propose a regression model with endogenous covariates where the error variable does not enter additively into the model. This allows for general forms of unobserved heterogeneity which has led to recent popularity of such nonseparable models among economists.

They consider a general regression relation of the form

$$Y = \mu(X_1, Z_1, e)$$

where X_1 and Z_1 are observed covariates and Y is a one-dimensional response. While Z_1 is independent of the error variable e, no assumptions are made on the dependence between X_1 and e at this stage. For identification, however, assume that the endogenous variable X_1 is generated as

$$X_1 = h(Z_1, Z_2, V),$$

where Z_2 is an observed so-called instrumental variable not contained in the original equation, and (Z_1, Z_2) is independent of the joint vector of errors (e, V).

If the function h is strictly monotone in V, one can set without loss of generality that the conditional distribution of V given (Z_1, Z_2) is the uniform law on [0,1]. This can be achieved by putting

$$V = F_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)$$

and choosing h as the inverse of $F_{X_1|Z_1,Z_2}$. Then by definition, the conditional distribution of V given (Z_1, Z_2) does not depend on (Z_1, Z_2). Thus, V is independent of (Z_1, Z_2). Note that the above independence assumption is slightly more restrictive, because it does not only require that (Z_1, Z_2) is independent of each e and V separately, but also of (e, V) jointly.

For fixed values of z_1, z_2 and v and for $x_1 = h(z_1, z_2, v)$ it is straightforward to show

$$\begin{aligned}
&E[\mu(x_1, z_1, e)|V = v] \\
&= E[\mu(X_1, Z_1, e)|Z_1 = z_1, Z_2 = z_2, V = v] \\
&= E[\mu(X_1, Z_1, e)|X_1 = x_1, Z_1 = z_1, V = v] \\
&= E[Y|Z_1 = z_1, Z_2 = z_2, V = v].
\end{aligned}$$

Thus we can write

$$Y = m_0(R) + \varepsilon$$

where

$$\begin{aligned}
&S = (X_1, Z_1, Z_2), \\
&R = r_0(S) = (X_1, Z_1, F_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)) = (X_1, Z_1, V), \\
&m_0(x_1, z_1, v) = E[\mu(x_1, z_1, e)|V = v], \\
&\varepsilon = Y - E[Y|S].
\end{aligned}$$

In this model, the covariate V is unobserved, but an estimate $\hat{V} = \hat{F}_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)$ of $\hat{F}_{X_1|Z_1,Z_2}$ is available. Thus, instead of R also use the feasible $\hat{R} = (X_1, Z_1, \hat{V})$. Then the function m_0 can be estimated by regressing Y onto $\hat{R}$. Let us denote this estimator as real, feasible estimator $\hat{m}$. One may compare this estimator to the theoretical, infeasible estimator $\tilde{m}$ obtained from regressing Y onto R. If the asymptotics of the theoretical estimator $\tilde{m}$ are well-understood, an asymptotic understanding of $\hat{m}$ can be based on a stochastic expansion of the difference of $\hat{m} - \tilde{m}$.

The function m_0 is not of direct interest because it contains the nuisance covariate V. In general, the focus is on the so-called average structural function $E[\mu(x_1, z_1, e)]$, the expected response if one exogenously fixes X_1 at x_1 and Z_1 at z_1. This function can be estimated by

$$\int_0^1 \widehat{m}(x_1, z_1, v)dv.$$

Other functionals of interest are averages of the derivative $\partial\mu(x_1, z_1, e)/\partial(x_1, z_1)$.

2.2 Simultaneous Nonparametric Equation Models with Additivity (Newey, Powell, Vella 1999)

In Newey, Powell, Vella (1999) a submodel of the regression model of the last subsection is considered. The setup differs from the last subsection by assuming that the error enters additively into the regression function, i.e.

$$Y = \mu(X_1, Z_1) + e.$$

For the control equation also an additive specification is used:

$$X_1 = h(Z_1, Z_2) + V,$$

but one could also proceed with the control equation of the last section.

With (Z_1, Z_2) independent of (e, V) as before, it is

$$E[Y|X_1, Z_1, Z_2] = \mu(X_1, Z_1) + \lambda(V) = E[Y|X_1, Z_1, V]$$

with $\lambda(V) = E[e|V]$. Thus we get an additive model where the regressor in the second additive component is not observed. This additive model can also be obtained under slightly weaker conditions, namely that $E[e|Z_1, Z_2, V] = E[e|V]$ and $E[V|Z_1, Z_2] = 0$.

There are two major approaches to fit an additive nonparametric model: marginal integration and backfitting. In Marginal Integration (Newey (1994), Tjøstheim and Auestad (1994), Linton and Nielsen (1995)), first a full dimensional regression function $E[Y|X_1 = x_1, Z_1 = z_1, V = v]$ is estimated. And then in a second step, v is integrated out to obtain an estimate of $\mu(x_1, z_1)$. The first step of this procedure can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where

$$\begin{aligned}
&S = (X_1, Z_1, Z_2),\\
&R = r_0(S) = (X_1, Z_1, X_1 - h(Z_1, Z_2)) = (X_1, Z_1, V),\\
&m_0(r) = E[Y|R = r],\\
&\varepsilon = Y - E[Y|R].
\end{aligned}$$

A fit of the unobserved R is given by $\hat{R} = (X_1, Z_1, \hat{V})$ with $\hat{V} = X_1 - \hat{h}(Z_1, Z_2)$ where $\hat{h}$ is a (nonparametric) estimator of the control function h.

In the Smooth Backfitting approach (Mammen, Linton, Nielsen, 1999) for an additive model, estimates are obtained by iteration. As ingredients for the iteration algorithm, one needs estimators of the marginal expectations $E[Y|X_1, Z_1]$, $E[Y|V]$, and of the joint density of (X_1, Z_1, V). Here estimation of $E[Y|V]$ can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where now

$$\begin{aligned}
&S = (X_1, Z_1, Z_2),\\
&R = r_0(S) = X_1 - h(Z_1, Z_2) = V,
\end{aligned}$$

$$m_0(v) = E[Y|V = v],$$
$$\varepsilon = Y - E[Y|V].$$

2.3 Marginal Treatment Effects (Heckman, Vytlacil, 2005, 2009)

In Heckman, Vytlacil (2005, 2009) the following model for treatment effects is discussed: we observe D, Y_D, X, Z in

$$Y_d = \rho(X, U_d, \theta_d) \qquad \text{for } d = 0, 1$$
$$D = 1, \text{ if } V \leq \mu(Z), \text{ and } D = 0, \text{ otherwise.}$$

Here θ_0 and θ_1 are unknown parameters that are finite or infinite-dimensional. Furthermore, ρ is a known function. An example for a specification would be $\rho(X, U_d, \theta_d) = m_d(X) + U_d$ with a "nonparametric parameter" $\theta_d = m_d$. The variable D is a dummy variable that indicates if a person is treated or not. The model contains counterfactual outcomes. If a person is treated ($D = 1$) the outcome Y_1 is observed, assuming that there also exists an unobserved outcome Y_0 that would have been observed if the person had not been treated. The participation of the person in the treatment is driven by an unobserved variable V. Without loss of generality, set V as uniform distribution on $[0, 1]$. For identification of the model the following condition is required:

(U_0, V) and (U_1, V) are conditionally independent of Z given X.

Note that the norming of V implies that $P(D = 1|Z) = \mu(Z)$.

Here, a function of interest is the Marginal Treatment Effect $MTE(x, v) = E[Y_1 - Y_0|X = x, V = v]$, the expected treatment effect for an individual with covariate $X = x$ that lies on the v-quantile of the unobserved propensity to participate in the treatment. It holds that

$$\begin{aligned} MTE(x, v) &= E[Y_1 - Y_0|X = x, V = v] \\ &= -\frac{\partial}{\partial v} E[Y_D|X = x, \mu(Z) = v]. \end{aligned}$$

This follows because for $\delta > 0$ small:

$$\begin{aligned} &MTE(x, v) = E[Y_1 - Y_0|X = x, V = v] \\ &\approx \delta^{-1} \left(-E[Y_1 I[V \geq v + \delta]|X = x] - E[Y_0 I[V < v + \delta]|X = x]\right. \\ &\qquad \left. + E[Y_1 I[V \geq v]|X = x] + E[Y_0 I[V < v]|X = x]\right) \\ &= \delta^{-1} \left(-E[Y_1 I[V \geq v + \delta]|X = x, \mu(Z) = v + \delta]\right. \\ &\qquad -E[Y_0 I[V < v + \delta]|X = x, \mu(Z) = v + \delta] + E[Y_1 I[V \geq v]|X = x, \mu(Z) = v] \\ &\qquad \left. + E[Y_0 I[V < v]|X = x, \mu(Z) = v]\right) \\ &= \delta^{-1} \left(-E[Y_D I[V \geq v + \delta]|X = x, \mu(Z) = v + \delta]\right. \end{aligned}$$

$$
\begin{aligned}
&-E[Y_D I[V < v+\delta]|X = x, \mu(Z) = v+\delta] + E[Y_D I[V \geq v]|X = x, \mu(Z) = v] \\
&+E[Y_D I[V < v]|X = x, \mu(Z) = v]) \\
&= \delta^{-1} \left(-E[Y_D|X = x, \mu(Z) = v+\delta] + E[Y_D|X = x, \mu(Z) = v]\right) \\
&\approx -\frac{\partial}{\partial v} E[Y_D|X = x, \mu(Z) = v].
\end{aligned}
$$

Here estimation of (the partial derivative of) $E[Y_D|X = x, \mu(Z) = v]$ can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where now

$$
\begin{aligned}
&Y = Y_D, \\
&S = (X, Z), \\
&R = r_0(S) = (X, \mu(Z)), \\
&m_0(r) = E[Y|(X, \mu(Z)) = r], \\
&\varepsilon = Y_D - E[Y_D|(X, \mu(Z))].
\end{aligned}
$$

Many treatment effects parameters and other parameters can be written as weighted averages of $MTE(x, v)$. Estimation of the MTE function is again based on a regression problem with an unobserved covariate $\mu(Z)$. Here interest is in a partial derivative of the regression function.

2.4 Further Examples.

Further examples of regression problems with unobserved covariates are sample selection models, censored regression models, generalized Roy models, stochastic volatility models and semiparametric GARCH-in-Mean models. For a discussion and/or references of these models we refer to Mammen, Rothe and Schienle (2011).

3 Nonparametric Regression with Nonparametrically Generated Covariates.

In all examples of the last section, the fit $\hat{R}$ of the unobserved covariate is of the form $\hat{R} = \hat{r}(S)$, where $\hat{r}$ is an estimator of a function r_0 that fulfills $R = r_0(S)$ for an observed covariate S. Thus we have the following nonparametric regression model

$$
\begin{aligned}
&Y = m_0(r_0(S)) + \varepsilon, \\
&E[\varepsilon|S] = 0.
\end{aligned}
$$

In this section, we give a brief description of the asymptotics of a nonparametric estimator $\hat{m}$ that is based on regressing Y onto the fitted covariate $\hat{R} = \hat{r}(S)$. For illustration, we restrict the discussion to the special case where $\hat{m} = \hat{m}_{LL}$ is a local

linear estimator for an i.i.d. sample (S_i, Y_i), i.e. $\widehat{m}_{LL}(x) = \widehat{\alpha}$, where $(\widehat{\alpha}, \widehat{\beta})$ minimizes

$$\sum_{i=1}^{n} [Y_i - \alpha - \beta^T (\widehat{R}_i - x)]^2 K_h(\widehat{R}_i - x).$$

Here is $K_h(u)$ a product kernel:

$$K_h(u) = (h_1 \cdot \ldots \cdot h_q)^{-1} K_1(u_1) \cdot \ldots \cdot K_q(u_q)$$

for kernel functions $K_1, \ldots, K_q$ and a bandwidth vector $h = (h_1, \ldots, h_q)$. We call this estimator also the real estimator, in contrast to the theoretical estimator $\widetilde{m}_{LL}$ which is defined as $\widetilde{m}_{LL}(x) = \widetilde{\alpha}$ where $(\widetilde{\alpha}, \widetilde{\beta})$ minimizes

$$\sum_{i=1}^{n} [Y_i - \alpha - \beta^T (R_i - x)]^2 K_h(R_i - x).$$

Since the R_i's are unobserved, this theoretical estimator is infeasible. It is, however, introduced here because its asymptotic behaviour is well understood. Thus, for the asymptotic properties of the real estimator we only need a stochastic expansion of $\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x)$. Such an expansion was derived in Mammen, Rothe and Schienle (2011) (MRS in the following).

For the comparison of $\widehat{m}_{LL}$ and $\widetilde{m}_{LL}$, MRS use three types of assumptions: besides standard smoothing assumptions, these are conditions on accuracy (A) and complexity (C) of the estimator $\widehat{r}$ of r_0. The assumption (A) requires that $\hat{r}$ converges to r_0 with a rate that is fast enough. The assumption (C) states that there exist sequences of sets $\mathscr{M}_n$ with two properties: (i) $\widehat{r} \in \mathscr{M}_n$ with probability tending to one. (ii) The sets $\mathscr{M}_n$ are not too large, where size is measured by entropy. The main result in MRS is the following expansion

$$\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x) \approx -m'(x) \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)}.$$

This result can be interpreted as follows: The real estimator $\widehat{m}_{LL}(x)$ and the oracle estimator $\widetilde{m}_{LL}(x)$ differ by a local weighted average of the estimator of r_0:

$$-m'(x) \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)}.$$

This local average is of the order of the bias of $\widehat{r}$ but it may have a faster rate as the variance part of $\widehat{r}$. Thus we can conclude that for achieving a certain rate of convergence for estimating m_0, it is not necessary that an estimator of r_0 has the same or a faster rate. A similar result can be obtained for derivatives of the regression function.

We now shortly outline the main ideas of how the expansion of $\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x)$ was obtained in MRS. We want to compare:

real estimator $\widehat{m}_{LL}$ = SMOOTH of $\widehat{r}(S)$ versus $m_0(r_0(S))+\varepsilon$,
oracle estimator $\widetilde{m}_{LL}$ = SMOOTH of $r_0(S)$ versus $m_0(r_0(S))+\varepsilon$.

Now, because of additivity of the operator SMOOTH, it is

$$\begin{aligned}\widehat{m}_{LL} = \ & \text{SMOOTH of } \widehat{r}(S) \text{ versus } m_0(\widehat{r}(S))+\varepsilon \\ & + \text{SMOOTH of } \widehat{r}(S) \text{ versus } m_0(r_0(S)) - m_0(\widehat{r}(S)).\end{aligned}$$

If $\widehat{r}$ was non-random we get, because $|\widehat{r}(S)-r_0(S)|$ is small by assumption (A),

$$\begin{aligned}\widehat{m}_{LL} \approx \ & \text{SMOOTH of } r_0(S) \text{ versus } m_0(r_0(S))+\varepsilon \\ & + \text{SMOOTH of } \widehat{r}(S) \text{ versus } m_0(r_0(S)) - m_0(\widehat{r}(S)) \\ \approx \ & \widetilde{m}_{LL} \\ & + \text{SMOOTH of } r_0(S) \text{ versus } m_0'(r_0(S))(r_0(S)-\widehat{r}(S)).\end{aligned}$$

This is (nearly) the formula of the desired expansion.

It remains to take care of the fact that $\widehat{r}$ is random and not purely deterministic. In order to do so, the argument must be uniform over the set of possible realizations of $\widehat{r}$. This can be achieved by an empirical process worst case analysis. We must show that

$$\begin{aligned}&|\widehat{m}_{LL,r}(x) - \widetilde{m}_{LL}(x) \\ &\qquad + m'(x)\frac{\frac{1}{n}\sum_{i=1}^n K_h(r_0(S_i)-x)(r(S_i)-r_0(S_i))}{\frac{1}{n}\sum_{i=1}^n K_h(r_0(S_i)-x)}|\end{aligned}$$

is small uniformly for r in $\mathscr{M}_n$. Here $\widehat{m}_{LL,r}$ is the local linear estimator based on regressing Y onto $r(S)$. At this stage of the proof one makes use of Assumption (C).

References

1. Heckman, J.J. and Vytlacil, E. J.: Structural equations, treatment effects, and econometric policy evaluation. Econometrica **73**, 669–738 (2005)
2. Heckman, J.J. and Vytlacil, E. J.: Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments. In: Heckman, J.J. and Leamer, E.E. (eds.) Handbook of Econometrics 6, chapter 71, Elsevier (2007)
3. Imbens, G.W. and Newey, W.K.: Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. Econometrica **77**, 1481–1512 (2009)
4. Linton, O. and Nielsen, J.P.: Kernel estimation of partial means and a general variance estimator. Biometrika **82**, 93-100 (1995)
5. Mammen, E. , Linton, O. and Nielsen, J.P.: The existence and asymptotic properties of a backfitting algorithm under weak conditions. Annals of Statistics **27**, 1443–1490 (1999)
6. Mammen, E. , Rothe, C. and Schienle, M. : Nonparametric regression with nonparametrically generated covariates. Preprint (2011)
7. Newey, W.K.: A kernel method of estimating structured nonparametric regression based on marginal integration. Econometric Theory **10**, 233-253 (1994)

8. Newey, W.K. and Powell, J.L. and Vella, F.: Nonparametric estimation of triangular simultaneous equations models. Econometrica **67**, 565–603 (1999)
9. Tjøstheim, D. and Auestad, B.H.: Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags. Journal of the American Statistical Association, **89**, (1994)