

Estimation in Covariate-Adjusted Regression

ESRA KURUM

Collaborative Work with Damla Senturk
Department of Statistics, Penn State University

ABSTRACT

We propose a new estimation procedure for covariate adjusted nonlinear regression models for situations where both the predictors and response in a nonlinear regression model are not directly observed, however distorted versions of the predictors and response are observed. The distorted versions are assumed to be contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate. We demonstrate how the regression coefficients can be estimated by establishing a connection to nonlinear varying-coefficient models. Simulation studies are used to illustrate the efficacy of the proposed estimation algorithm.

Nonlinear CAR

Cui et al. (2008) have extended CAR to nonlinear regression models. Consider the following nonlinear regression model,

$$Y_i = f(X_i, \beta) + \epsilon_i, \quad (4)$$

where $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. The distorted data have the multiplicative distortion given as follows.

$$\begin{aligned} \tilde{X}_{ri} &= \phi_r(U_i)X_{ri} \\ \tilde{Y}_i &= \psi(U_i)Y_i. \end{aligned} \quad (5)$$

Here, $\phi_r(U_i)$ and $\psi(U_i)$ are unknown smooth functions of U , $r = 1, \dots, p$ and $i = 1, \dots, n$.

The estimation procedure for Nonlinear CAR proposed by Cui et al. (2008) starts with regressing the observed response and predictors on U , respectively, to obtain estimators of the smooth functions, $\hat{\psi}(U)$ and $\hat{\phi}(U)$. Then, these estimators are used to estimate the underlying unobserved predictors and response,

$$\hat{X}_{ri} = \frac{\tilde{X}_{ri}}{\hat{\phi}_r(u_i)} \text{ and } \hat{Y}_i = \frac{\tilde{Y}_i}{\hat{\psi}(u_i)}, \quad r = 1, \dots, p. \quad (6)$$

Then, the estimators in (6) are used to minimize a nonlinear least squares criterion to estimate the coefficients in model (4).

Proposed Estimation Procedure

The model for proposed estimation procedure is same as the one proposed by Cui et al. (2008) and we also make use of the identifiability constraints imposed by Senturk and Muller (2005) on the smoothing functions,

$$E\{\phi(U_i)\} = 1 \text{ and } E\{\psi(U_i)\} = 1,$$

i.e the mean of the distorted variables are same with the adjusted variable.

In the Covariate-adjusted nonlinear regression with multiplicative distortion the following model holds,

$$E(\tilde{Y}_i | \tilde{X}_i, U_i) = f\{\tilde{X}_i, \alpha(U_i)\},$$

where $\tilde{X}_i = (\tilde{X}_{1i}, \dots, \tilde{X}_{pi})^T$, $\alpha(U_i) = \{\alpha_1(U_i), \dots, \alpha_p(U_i)\}^T$, $\alpha_1(U_i) = \psi(U_i)\beta_0$ and $\alpha_r(U_i) = \beta_r \frac{\psi(U_i)}{\phi_r(U_i)}$, $r = 1, \dots, p$. Therefore,

$$\tilde{Y}_i = f\{\tilde{X}_i, \alpha(U_i)\} + \epsilon_i,$$

which is a nonlinear varying-coefficient model (Wang, 2007).

Nonlinear Varying-Coefficient Models

The estimation procedure proposed by Wang (2007) for Nonlinear VCM starts with approximating the regression coefficients for any u in the neighborhood of u_0 by Taylor's expansion,

$$\alpha_r(u) \approx \alpha_r(u_0) + \alpha_r(u - u_0) \equiv a_r + b_r(u - u_0), \quad r = 1, \dots, p.$$

Define the vectors $a = (a_1, \dots, a_p)^T$, $b = (b_1, \dots, b_p)^T$, $Y = (Y_1, \dots, Y_n)$ and $\alpha(U_i) = \{\alpha_1(U_i), \dots, \alpha_p(U_i)\}^T$. Local linear regression estimator of $(a^T, b^T)^T$ is obtained by minimizing,

$$l(a, b) = \frac{1}{2} \sum_{i=1}^n [Y_i - f\{X_i, a + b(U_i - u_0)\}]^2 K_h(U_i - u_0)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is the kernel function and h is the bandwidth. An iterated least squares algorithm is used to compute the value that minimizes the likelihood. At the end we obtain values of the estimators of $\alpha(u_0)$ such that $\hat{\alpha}(u_0) = \hat{a}$ and $\hat{\alpha}'(u_0) = \hat{b}$.

Proposed Estimation Procedure (Con't)

To obtain the estimator of β in the nonlinear regression model, we first need to estimate $\alpha(U_i)$ in the nonlinear VCM. We adapted the iterative procedure of Wang (2007) to obtain the estimators of $\alpha(U_i)$. Then, the underlying regression parameters of interest are targeted via nonlinear VCM as follows,

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^n \hat{\alpha}_1(U_i),$$

$$\hat{\beta}_r = \frac{1}{n \hat{\mu}_{\tilde{X}_r}} \sum_{i=1}^n \hat{\alpha}_r(U_i) \tilde{X}_{ri},$$

$r = 1, \dots, p$.

These estimators are motivated by the equalities $E\{\alpha_1(U)\} = \beta_1$ and $E\{\alpha_r(U)\tilde{X}_r\} = \beta_r \tilde{X}_r$.

Simulation Study

Consider the following nonlinear regression model,

$$Y_i = \beta_1 \{1 - \exp(-\beta_2 X_i)\} + \epsilon_i, \quad (7)$$

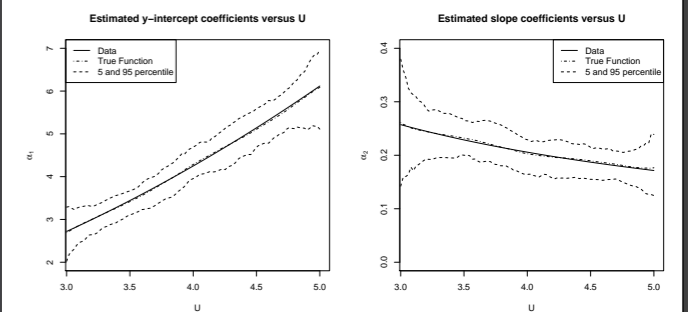
where $\beta_1 = 4.31$, $\beta_2 = 0.21$, $\epsilon \sim N(0, 0.02)$ and $X \sim N(5, 3)$. We also assume that the covariate, $U \sim Uniform(3, 5)$. The distortion functions are $\psi(U_i) = (U_i + 1)^2/25.3$ and $\phi(U_i) = (U_i + 1)/5$, which satisfy the identifiability conditions, $E(\psi(U)) = 1$ and $E(\phi(U)) = 1$. Therefore, the distorted versions of the predictor and response are defined as in (5).

We can write the observed nonlinear VCM as,

$$\tilde{Y}_i = \alpha_1(U_i) \left[1 - \exp\{-\alpha_2(U_i)\tilde{X}_i\} \right] + \epsilon, \quad (8)$$

where $\alpha_1(U_i) = \psi(U_i)\beta_1$ and $\alpha_2(U_i) = \beta_2/\{\phi(U_i)\}$.

Simulation Results



Simulation Results

Estimated bias, variance and mean squared error of the proposed intercept coefficient

Sample Size		Proposed	Cui
n=50	bias	0.036	0.088
	var	0.070	0.071
	mse	0.071	0.078
n=100	bias	0.011	0.034
	var	0.026	0.026
	mse	0.026	0.027
n=200	bias	0.008	0.028
	var	0.015	0.015
	mse	0.015	0.016

Simulation Results

Estimated bias, variance and mean squared error of the proposed slope coefficient

Sample Size		Proposed	Cui
n=50	bias	0.0010	0.0064
	var	0.0003	0.0003
	mse	0.0003	0.0003
n=100	bias	0.0006	0.0029
	var	0.0001	0.00013
	mse	0.0001	0.00013
n=200	bias	0.0002	0.0019
	var	0.00006	0.00007
	mse	0.00006	0.00007

References

- Cui, X., Guo, W., Lin, L. and Zhu, L. (2008) Covariate adjusted nonlinear regression. *Annals of statistics* (to appear)
- Hastie, T.J. and Tibshirani, R.J. (1993), Varying-coefficient models (with discussion), *J. Royal Statist. Soc., Ser. B*, **55**, 757-796
- Senturk, D. and Muller, H.G. (2005a). Covariate adjusted regression. *Biometrika* **92**, 59-74.
- Senturk, D. and Muller, H.G. (2006). Inference for covariate adjusted regression via varying-coefficient models. *Annals of Statistics* **34**, 654-679.
- Wang, Y. (2007) Varying-coefficient models: New models, inference procedures and applications, unpublished thesis (PhD), Penn State University.

Literature Review: CAR

Consider the multiple regression model,

$$Y_i = \beta_0 + \sum_{r=1}^p \beta_r X_{ri} + \epsilon_i, \quad (1)$$

where X_{ri} 's are the predictors, Y_i is the response, ϵ_i is the error and $i = 1, \dots, n$.

Senturk and Muller (2005, 2006) proposed CAR for cases where the predictors and response are not directly observed in (1). Instead, their distorted versions \tilde{X} and \tilde{Y} , along with a univariate covariate U are observed,

$$\begin{aligned} \tilde{X}_{ri} &= \phi_r(U_i)X_{ri}, \\ \tilde{Y}_i &= \psi(U_i)Y_i. \end{aligned} \quad (2)$$

Here, $\phi_r(U)$ and $\psi(U)$ are unknown smooth functions of U . The goal is to estimate $(\beta_0, \beta_1, \dots, \beta_p)$ from distorted \tilde{X} , \tilde{Y} , U .

Senturk and Muller (2005) utilized the connection between the above distortion setup and varying-coefficient models (VCM) that holds between the observed variables. Let $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p)^T$, then we can write,

$$\begin{aligned} E(\tilde{Y}_i | \tilde{X}_i, U_i) &= \psi(U_i)\beta_0 + \psi(U_i) \sum_{r=1}^p \beta_r \frac{\phi_r(U_i)X_{ri}}{\phi_r(U_i)} \\ &= \alpha_0(U_i) + \sum_{r=1}^p \alpha_r(U_i)\tilde{X}_{ri}, \end{aligned}$$

where $\alpha_0(U_i) = \psi(U_i)\beta_0$ and $\alpha_r(U_i) = \beta_r \frac{\psi(U_i)}{\phi_r(U_i)}$.

As a result,

$$\tilde{Y}_i = \alpha_0(U_i) + \sum_{r=1}^p \alpha_r(U_i)\tilde{X}_{ri} + \psi(U_i)\epsilon_i. \quad (3)$$

Model (3) is a multiple VCM, where the regression coefficients are allowed to vary smoothly with the value of the covariate U (Hastie and Tibshirani, 1993; Cleveland, Grosse and Shyu, 1991).

Motivation : MDRD Dataset

Glomerular filtration rate (GFR) is used as an indicator of kidney health and the stage of kidney disease. Because of the difficulty of the other techniques to measure GFR, serum creatinine (SCR) has become a traditional way to estimate GFR. It is known that there is a nonlinear relationship between GFR and SCR. In addition, GFR and SCR are both affected by Body Surface Area (BSA). Therefore, Cui et al. (2008) has suggested covariate-adjusted nonlinear models to adjust for the non-parametric effects of BSA on GFR and SCR in the regression relationship of GFR and SCR in Modification of Diet in Renal Disease (MDRD) study.