Chapter 2. Semi-parametric Models (I) Part 2

February 12, 2007

1 The varying coefficient regression model

Recall in the linear regression model

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_q \mathbf{x}_q + \varepsilon.$$

where $\beta_0, \beta_1, ..., \beta_q$ are constant and does not change with any other factors.

Example 1.1 (An intuitive data) [1] *Relative spinal bone mineral density measurements on 261 North American adolescents. In the data, "idnum" identifies the child, and hence the repeat measurements; "age" is average age of child when measurements were taken; "gender" is male, denoted by 1 or female by 0; "spnbmd" is Relative Spinal bone mineral density measurement. Consider a model*

$$spnbmd = \beta_0 + \beta_1 * age + \varepsilon$$

for male and female separately, we have

$$\begin{array}{llcccc} Male: & spnbmd & = & 11.7367 & - & 0.4834 * age \\ & s.e. & & (1.2540) & & (0.0757) \\ Female: & spnbmd & = & 15.5963 & - & 0.7267 * age \\ & s.e. & & (0.9942) & & (0.0597) \end{array}$$

The estimations suggest that the gender changed the relation between the bone density and age (WHY). In other words, β_0 and β_1 should be a function of "gender"!

To explore how a factor Z affects the relation between Y and X, we consider the following **varying coefficient regression model** (or functional coefficient regression model) proposed by Hastie and Tibishirani (1992)

$$Y = a_0(Z) + a_1(Z)\mathbf{x}_1 + \cdots + a_q(Z)\mathbf{x}_q + \varepsilon. \tag{1.1}$$

[1] source: Bachrach LK, Hastie T, Wang M-C, Narasimhan B, Marcus R. Bone Mineral Acquisition in Healthy Asian, Hispanic, Black and Caucasian Youth. A Longitudinal Study. J Clin Endocrinol Metab (1999) 84, 4702-12.

where $a_0(z), a_1(z), ..., a_q(z)$ are unknown functions. We further assume that

$$E(\varepsilon|\mathbf{x}_1, \cdots, \mathbf{x}_p, Z) = 0$$

If Z is time, then model can describe how the model changes with time.

2 Estimation of the Varying coefficient regression model

Suppose a random sample $\{(Z_i, \mathbf{x}_{i1}, ..., \mathbf{x}_{iq}, Y_i), i = 1, ..., n\}$ is from model (1.1), i.e.

$$Y_i = a_0(Z_i) + a_1(Z_i)\mathbf{x}_{i1} + \cdots + a_q(Z_i)\mathbf{x}_{iq} + \varepsilon_i.$$

We need to estimate the coefficient functions $a_k(z), k = 0, 1, ..., q$.

For any two points z and z', we have the following approximation

$$a_j(z') \approx a_j(z) + b_j(z)(z' - z)$$

for any z' in a neighborhood of z. Apply this to each observation, we have

$$\begin{aligned} Y_i \approx \{a_0(z) + b_0(z)(Z_i - z)\} + \{a_1(z) + b_1(z)(Z_i - z)\}\mathbf{x}_{i1} \\ + \cdots + \{a_q(z) + b_q(z)(Z_i - z)\}\mathbf{x}_{iq} + \varepsilon_i. \end{aligned}$$

where $i = 1, ..., n$ or

$$\begin{aligned} Y_1 &\approx a_0(z) + a_1(z)\mathbf{x}_{11} + ... + a_q(z)\mathbf{x}_{1q} \\ &\quad + b_0(z)(Z_1 - z) + b_1(z)(Z_1 - z)\mathbf{x}_{11} + \cdots + b_q(z)(Z_1 - z)\mathbf{x}_{1q} + \varepsilon_1. \\ Y_2 &\approx a_0(z) + a_1(z)\mathbf{x}_{21} + ... + a_q(z)\mathbf{x}_{2q} \\ &\quad + b_0(z)(Z_2 - z) + b_1(z)(Z_2 - z)\mathbf{x}_{21} + \cdots + b_q(z)(Z_2 - z)\mathbf{x}_{2q} + \varepsilon_2, \\ &\quad ... \\ Y_n &\approx a_0(z) + a_1(z)\mathbf{x}_{n1} + ... + a_q(z)\mathbf{x}_{nq} \\ &\quad + b_0(z)(Z_n - z) + b_1(z)(Z_n - z)\mathbf{x}_{n1} + \cdots + b_q(z)(Z_n - z)\mathbf{x}_{nq} + \varepsilon_n. \end{aligned}$$

Note that with fixed z, this is a simple linear regression

Again, consider a weighted least squares estimation. We estimate the functions $a_j(z)$ be the minimizer of

$$\begin{aligned} \sum_{i=1}^{n} \{Y_i - [a_0 + a_n\mathbf{x}_{i1} + ... + a_q\mathbf{x}_{nq} + b_0(Z_n - z) \\ + b_n(Z_n - z)\mathbf{x}_{n1} + \cdots + b_q(Z_n - z)\mathbf{x}_{nq}]\}^2 K_h(Z_i - z) \end{aligned} \tag{2.2}$$

with respect to $a_0, ..., a_q, b_0, ...b_q$.

By writing

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_{11} & ... & \mathbf{x}_{1q} & (Z_1 - z) & (Z_1 - z)\mathbf{x}_{11} & ... & (Z_i - z)\mathbf{x}_{1q} \\ 1 & \mathbf{x}_{21} & ... & \mathbf{x}_{2q} & (Z_2 - z) & (Z_2 - z)\mathbf{x}_{21} & ... & (Z_2 - z)\mathbf{x}_{2q} \\ ... \\ 1 & \mathbf{x}_{n1} & ... & \mathbf{x}_{nq} & (Z_n - z) & (Z_n - z)\mathbf{x}_{n1} & ... & (Z_n - z)\mathbf{x}_{nq} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}.$$

and

$$\mathbf{W} = \begin{pmatrix} K_h(Z_1 - z) & 0 & ... & 0 \\ 0 & K_h(Z_2 - z) & ... & 0 \\ ... \\ 0 & 0 & ... & K_h(Z_n - z) \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix}.$$

Then the minimizer to (2.2), i.e. the estimators, are

$$\begin{pmatrix} \hat{a}_0(z) \\ \hat{a}_1(z) \\ ... \\ \hat{a}_q(z) \\ \hat{b}_0(z) \\ \hat{b}_1(z) \\ ... \\ \hat{b}_q(z) \end{pmatrix} = \{\mathbf{X}^\top \mathbf{W} \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y}. \tag{2.3}$$

3 Statistical inference of the Varying coefficient regression model

If the design of $(Z_i, \mathbf{x}_{i1}, ..., \mathbf{x}_{iq}), i = 1, ..., n$ is not random, then by (2.3) we have

$$\begin{aligned} \begin{pmatrix} \hat{a}_0(z) \\ \hat{a}_1(z) \\ ... \\ \hat{a}_q(z) \\ \hat{b}_0(z) \\ \hat{b}_1(z) \\ ... \\ \hat{b}_q(z) \end{pmatrix} &\approx \{\mathbf{X}^\top \mathbf{W} \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{W} \{\mathbf{X} \begin{pmatrix} a_0(z) \\ a_1(z) \\ ... \\ a_q(z) \\ b_0(z) \\ b_1(z) \\ ... \\ b_q(z) \end{pmatrix} + \mathcal{E}\} \\ &= \begin{pmatrix} a_0(z) \\ a_1(z) \\ ... \\ a_q(z) \\ b_0(z) \\ b_1(z) \\ ... \\ b_q(z) \end{pmatrix} + \{\mathbf{X}^\top \mathbf{W} \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{W} \mathcal{E} \end{aligned}$$

If

$$\mathcal{E} \sim N(0, \sigma^2 I),$$

then

$$\{\mathbf{X}^\top \mathbf{W}\mathbf{X}\}^{-1}\mathbf{X}^\top \mathbf{W}\mathcal{E} \sim N(0, \{\mathbf{X}^\top \mathbf{W}\mathbf{X}\}^{-1}\mathbf{X}^\top \mathbf{W}^2\mathbf{X}\{\mathbf{X}^\top \mathbf{W}\mathbf{X}\}^{-1}\sigma^2)$$

The 95% confidence band for $a_k(z)$ is approximately

$$\hat{a}_k(z) \pm 1.96\sigma\sqrt{c_{kk}},$$

where c_{kk} is the (k,k)th entry of $\{\mathbf{X}^\top \mathbf{W}\mathbf{X}\}^{-1}\mathbf{X}^\top \mathbf{W}^2\mathbf{X}\{\mathbf{X}^\top \mathbf{W}\mathbf{X}\}^{-1}$.

For random design, we assume that ε_i is independent of $(Z_i, \mathbf{x}_{i1}, ..., \mathbf{x}_{iq}), i = 1, ..., n$ and $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

Let

$$r_{ij}(z) = E(\mathbf{x}_i\mathbf{x}_j | Z = z), \quad i, j = 1, ..., q$$

$$\alpha_{k,j} = (r_{1j}(z), ...r_{k-1,j}, r_{k+1,j}, ..., r_{pj})^\top$$

and

$$\Omega_k = E\{(\mathbf{x}_1, ..., \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, ..., \mathbf{x}_q)^\top(\mathbf{x}_1, ..., \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, ..., \mathbf{x}_p) | Z = z)$$

If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then under some assumptions, we have

$$Bias(\hat{a}_k(z)) = E(\hat{a}_k(z)) - a_k(z) \approx -\frac{h^2 c_2}{2r_{kk}} \sum_{\substack{j=1 \\ j\neq k}}^{q} r_{kj} a_j''(u)$$

and

$$var(\hat{a}_k(z)) \approx \frac{\sigma^2(\lambda_2 r_{kk} + \lambda_3 \alpha_k^\top \Omega_k^{-1}\alpha_k}{nhf(z)\lambda_1 r_{kk}(r_k k - \alpha_k^\top \Omega_k^{-1}\alpha_k)}$$

where $\lambda_1 = (c_4 - c_2^2)^2, \lambda_2 = d_0c_4^2 - 2d_2^2c_2c_4 + c_2d_4$ and $\lambda_3 = 2c - 2d_2c_4 - 2d_0c_2^2c_4 - c_2^2d_4 + d_0c_2^4$, with

$$c_k = \int v^k K(v)dv, \quad d_k = \int v^k K^2(v)dv.$$

Here $f(z)$ is the density function of Z.

If further $nh^5 \to 0$, then

$$\hat{a}_k(z) - a_k(z) \to N\{0, \frac{\sigma_{kk}^2(z)}{nhf(z)}\}$$

where

$$\sigma_{kk}^2(z) = \frac{\sigma^2(\lambda_2 r_{kk} + \lambda_3 \alpha_k^\top \Omega_k^{-1}\alpha_k)}{nhf(z)\lambda_1 r_{kk}(r_{kk} - \alpha_k^\top \Omega_k^{-1}\alpha_k)}$$

The 95% confidence band is

$$[L_n(z), U_n(z)]$$

where

$$L_n(z) = \hat{a}_k(z) - 1.96\Big(\frac{\sigma^2_{kk}(z)}{nhf(z)}\Big)^{1/2}$$

and

$$U_n(z) = \hat{a}_k(z) + 1.96\Big(\frac{\sigma^2_{kk}(z)}{nhf(z)}\Big)^{1/2}$$

4 Bandwidth selection

The cross-validation method and other methods can also be used.

5 Simulations and Examples for real data analysis

Example 5.1 (simulation) *We consider the following model*

$$Y = \exp(-40(Z-0.5)^2) + \cos(2\pi Z)\mathbf{x}_1 + \sin(2\pi Z)\mathbf{x}_2 + 0.5\varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2, \varepsilon \sim N(0,1)$ *and* $Z \sim uniform(0,1)$ *are IID. In the model,* $a_0(z) = \exp(-40(z-0.5)^2), a_1(v) = \cos(2\pi v)$ *and* $a_2(z) = \sin(2\pi z)$.

100 samples are drawn from the model. The estimated functions of $a_0(z) = \exp(-40(z-0.5)^2), a_1(v) = \cos(2\pi v)$ *and* $a_2(z) = \sin(2\pi z)$ *are shown in Figure 1.*

Example 5.2 (Pollution and health data) *The pollutants (NO_2, SO_2, O_3, Particulate matters (PM)) were observed daily in Hong Kong from 1994-1997. The daily hospital admission of patients suffering circulatory diseases and respiratory diseases are also recorded. We consider the following model*

$$Y_t = a_0(t)+a_1(t)*NO2_t+a_2(t)*SO2_t+a_3(t)*O3_t+a_4(t)*PM_t+a_5(t)*Temperature+a_6(t)*Humdity$$

Y_t *is the number of hospital admission suffering respiratory diseases.*

The estimated coefficient functions are shown in Figure 2.

The estimated function changes with time indicating that the effect of pollutants on the respiratory diseases changes with time. The reason need to be further investigated

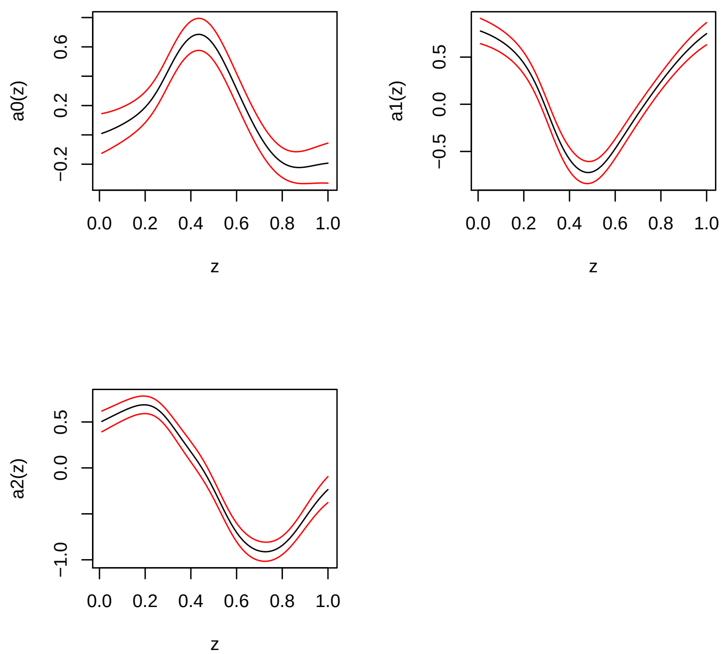

Figure 1: In each panel, the curve in the middle is the estimated functions of $a_0(z), a_1(v)$ and $a_2(z)$, the upper and lower lines are the 95% point-wise confidence bands respectively. **(vcm.R) (c2b1.R)**

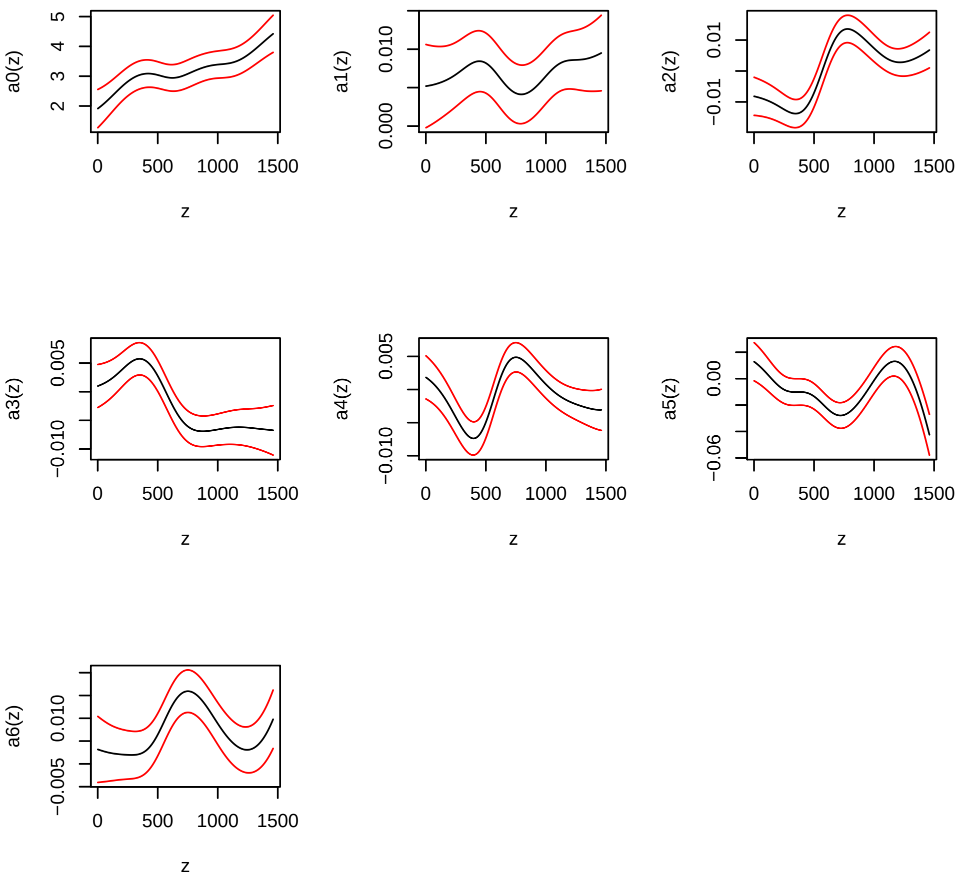

Figure 2: In each panel, the curve in the central is the estimated functions of $a_0(z), a_1(z), ..., a_6(z)$, the upper and lower lines are the 95% point-wise confidence bands respectively. **(vcm.R) (c2b2.R)**

References

Chen, R. and Tsay, S.(1993) Functional-coefficient autoregressive models. *J. Amer. Statist. Ass.*, 88, 298-308.

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc. B.* **55**, 757-796.