Chapter 1. Semiparametric models (I)
Part 1

February 9, 2007

1 Statistical inference for the partially linear regression model

The partially linear regression model is

$$Y = \beta_1 \mathbf{x}_1 + \cdots + \beta_q \mathbf{x}_q + g(Z) + \varepsilon.$$

where g is an unknown function, $\beta_1, ..., \beta_q$ are unknown parameters. We further assume that

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

Let $g_0(z) = E(Y|Z = z)$, $g_k(z) = E(\mathbf{x}_k|Z = z), k = 1, ..., q$. The model is equivalent to

$$Y - g_0(Z) = \beta_1\{\mathbf{x}_1 - g_1(Z)\} + \cdots + \beta_q\{\mathbf{x}_q - g_q(Z)\} + \varepsilon. \tag{1.1}$$

For each random sample, we have

$$\begin{aligned}
Y_1 - g_0(Z_1) &= \beta_1\{\mathbf{x}_{11} - g_1(Z_1)\} + \cdots + \beta_1\{\mathbf{x}_{1q} - g_{1q}(Z)\} + \varepsilon_1 \\
Y_2 - g_0(Z_2) &= \beta_1\{\mathbf{x}_{21} - g_1(Z_2)\} + \cdots + \beta_1\{\mathbf{x}_{2q} - g_{2q}(Z)\} + \varepsilon_2 \\
&\vdots \\
Y_n - g_0(Z_n) &= \beta_1\{\mathbf{x}_{n1} - g_1(Z_n)\} + \cdots + \beta_1\{\mathbf{x}_{nq} - g_q(Z_n)\} + \varepsilon_n.
\end{aligned}$$

Let

$$\tilde{\mathbf{Y}} = \begin{pmatrix} Y_1 - g_0(Z_1) \\ Y_2 - g_0(Z_2) \\ \vdots \\ Y_n - g_0(Z_n) \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_{11} - g_1(Z_1), & \cdots, & \mathbf{x}_{11} - g_q(Z_1) \\ \mathbf{x}_{21} - g_1(Z_2), & \cdots, & \mathbf{x}_{21} - g_q(Z_2) \\ \vdots & & \\ \mathbf{x}_{n1} - g_1(Z_n), & \cdots, & \mathbf{x}_{n1} - g_q(Z_n) \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

Similar to the linear regression model, the estimator of β

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \vdots, \\ \tilde{\beta}_q \end{pmatrix} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}} \tilde{\mathbf{Y}}$$

If $\varepsilon \sim N(0, \sigma^2)$, then

$$\tilde{\beta} - \beta \sim N\{0, \sigma^2 (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\}$$

In other words

$$\tilde{\beta}_k - \beta_k \sim N(0, \sigma^2 c_{kk})$$

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

In implementing the above procedure, we need to replace $g_k(z), k = 0, 1, \cdots, q$ by kernel smoothing (NW or local linear) estimator, say $\hat{g}_k(z)$ using NW,

$$\hat{g}_0(z) = \frac{\sum_{i=1}^n K_h(Z_i - z) Y_i}{\sum_{i=1}^n K_h(Z_i - z)}$$

and

$$\hat{g}_k(z) = \frac{\sum_{i=1}^n K_h(Z_i - z) \mathbf{x}_{ik}}{\sum_{i=1}^n K_h(Z_i - z)}$$

for $k = 1, \cdots, q$ respectively. Let

$$\hat{\mathbf{Y}} = \begin{pmatrix} Y_1 - \hat{g}_0(Z_1) \\ Y_2 - \hat{g}_0(Z_2) \\ \vdots \\ Y_n - \hat{g}_0(Z_n) \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_{11} - \hat{g}_1(Z_1), & \cdots, & \mathbf{x}_{11} - \hat{g}_q(Z_1) \\ \mathbf{x}_{21} - \hat{g}_1(Z_2), & \cdots, & \mathbf{x}_{21} - \hat{g}_q(Z_2) \\ & \vdots & \\ \mathbf{x}_{n1} - \hat{g}_1(Z_n), & \cdots, & \mathbf{x}_{n1} - \hat{g}_q(Z_n) \end{pmatrix}$$

The final estimator of β

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots, \\ \hat{\beta}_q \end{pmatrix} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}} \hat{\mathbf{Y}}$$

If $\varepsilon \sim N(0, \sigma^2)$, then

$$\hat{\beta} - \beta \sim N\{0, \sigma^2 (\hat{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\}$$

In other words

$$\tilde{\beta}_k - \beta_k \sim N(0, \sigma^2 c_{kk})$$

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

The estimated model is

$$\hat{Y} = \hat{\beta}_1 \mathbf{x}_1 + ... + \hat{\beta}_q \mathbf{x}_q + \hat{g}(Z)$$

we can estimate σ^2 by

$$\hat{\sigma} = \sqrt{n^{-1} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}.$$

where

$$\hat{Y}_i = \hat{\beta}_1 \mathbf{x}_{i1} + ... + \hat{\beta}_q \mathbf{x}_{iq} + \hat{g}(Z_i).$$

You can also calculate the R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2},$$

where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

Example 1.1 (simulation) *$n = 100$ observations are sampled from a simulated model*

$$Y = \mathbf{x}_1 + 2\mathbf{x}_2 - \mathbf{x}_3 + \cos(2\pi Z) + 0.2\varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \varepsilon \sim N(0, 1), Z \sim Uniform(0, 1)$ are independent.

The estimated model is

$$\begin{array}{ccccccc} \hat{y} & = & 0.94\mathbf{x}_1 & + & 2.10\mathbf{x}_2 & - & 1.07\mathbf{x}_3 + \hat{g}(Z) \\ (s.e.) & & (0.02) & & (0.03) & & (0.03) \end{array}$$

the estimated function $\hat{g}(z)$ is shown in Figure 1

Example 1.2 (Ozone data) *We consider model*

$$Ozone = \beta_1 * Temperature + \beta_2 * Wind + g(Radiation) + \varepsilon$$

The estimated model is

$$\begin{array}{ccccc} \hat{Ozone} & = & 1.57 * Temperature & - & 3.23 * Wind \quad + \hat{g}(Z) \\ (s.e.) & & (0.26) & & (0.63) \end{array}$$

the estimated function $\hat{g}(z)$ is shown in Figure 2

The estimated model suggests that there is a threshold for the radiation, above which its effect on the level of ozone is very significant compare with that below the threshold.

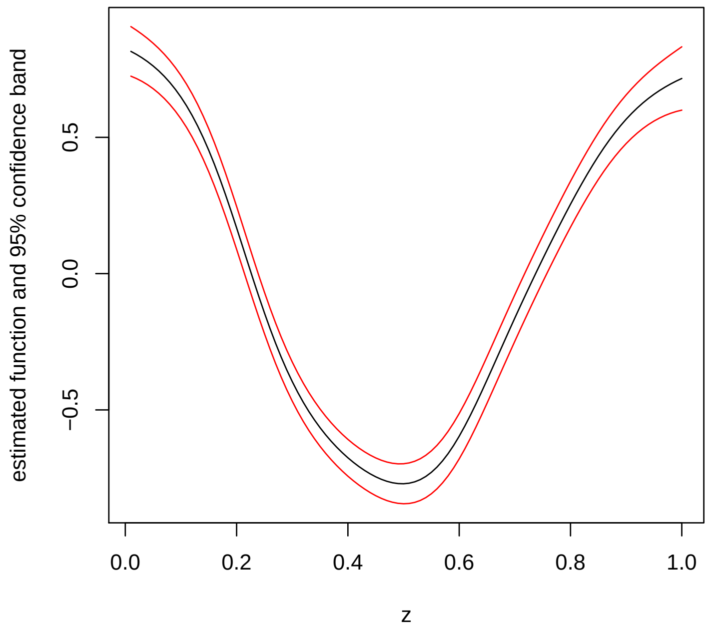


Figure 1: calculation reuslts for Example 1.1. The line in the central is the estimated function $\hat{g}(z)$, the upper and lower lines are the 95% point-wise confidence band. **(plr.R)** **(c1h1.R)**

Example 1.3 (Baseball Hitter's salary in America data) *Let* $Y = \log(salary)$. *We consider model:*

$$Y = \beta_1 * x_1 + ... + \beta_{15} * x_{15} + g(age) + \varepsilon$$

The estimated model is

$$\begin{array}{cccccccc} \hat{Y} & = & -0.001403159 x_1 & + & 0.004141342 x_2 & +...+ & 0.0001828098 x_{15} & +\hat{g}(Z) \\ (s.e.) & & (0.0009708911) & & (0.0037462473) & ... & (0.0002597766) & \end{array}$$

the estimated function $\hat{g}(z)$ *is shown in Figure 4*

The estimated model suggests that there is an "aging effect": too "old" is an adverse factor for a player's salary.

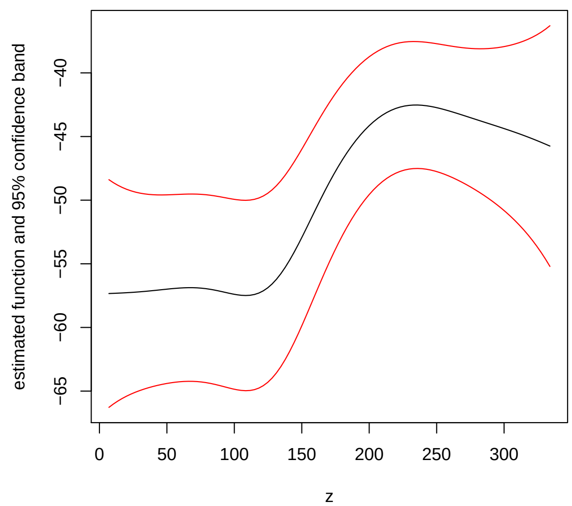
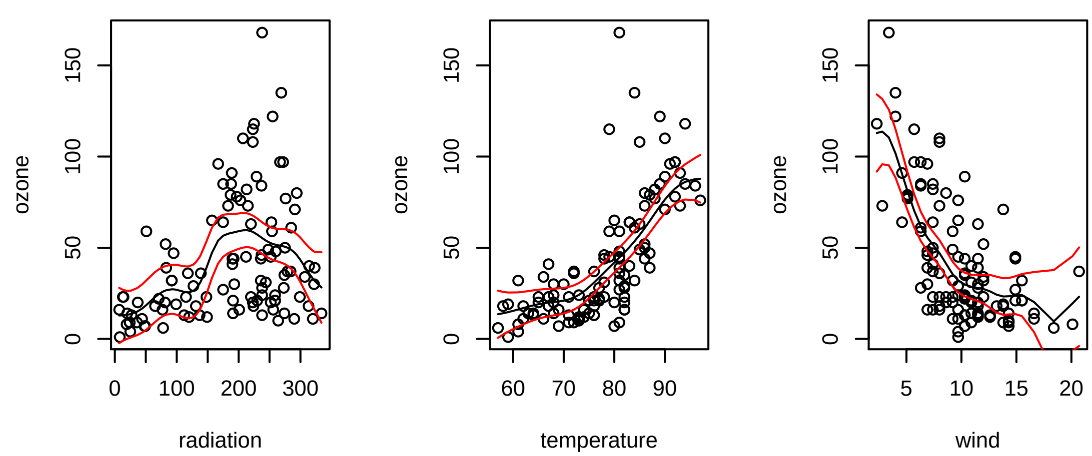


Figure 2: calculation results for Example 1.2. The line in the central is the estimated function $\hat{g}(z)$, the upper and lower lines are the 95% point-wise confidence band. **(plr.R) (c1h2.R)**

[as comparison, if we do the pairwise analysis, we have the following figures]



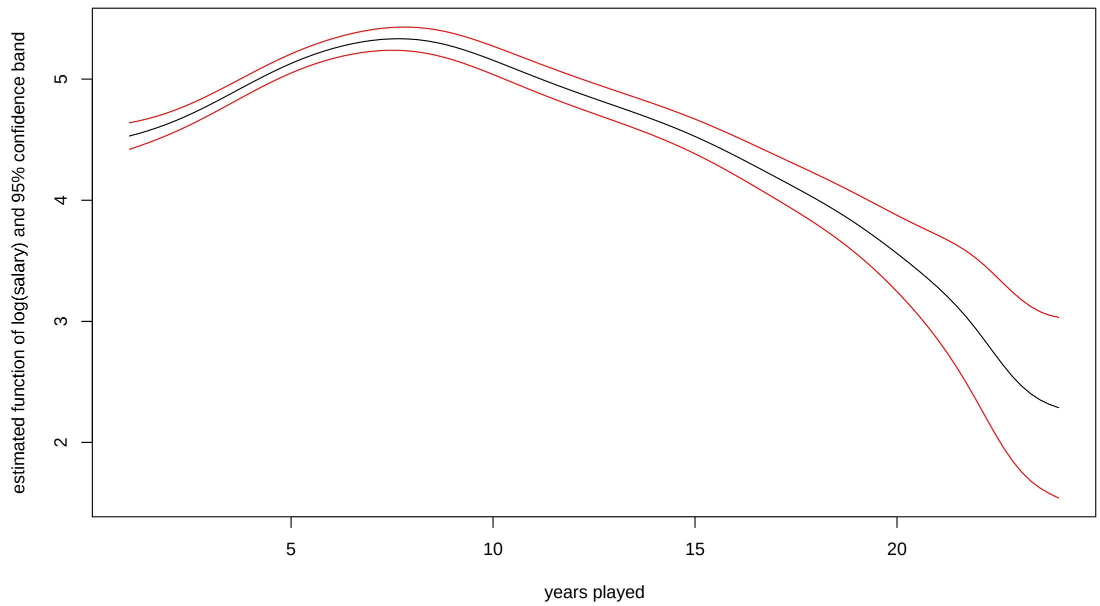


Figure 4: calculation results for Example 1.3. The line in the central is the estimated function $\hat{g}(z)$, the upper and lower lines are the 95% point-wise confidence band. **(plr.R)** **(c1h3.R)**

2 The distribution of $\hat{g}(z)$

Theorem 2.1 *Under some conditions, we have*

$$\sqrt{nh}\{\hat{g}(z) - g(z) - Bias\} \to N(0, \frac{\sigma^2 d_0}{f(x)})$$

where $\sigma^2 = E\varepsilon_i^2$ *and* $d_0 = \int K^2$. *If* $nh^5 \to 0$, *then we have the following point-wise 95% confidence band for* $m(x)$

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

where

$$L(x) = \hat{g}(x) - 1.96\sqrt{\frac{\hat{\sigma}^2 d_0}{nh\hat{f}(x)}},$$

$$m(x) = \hat{g}(x) + 1.96\sqrt{\frac{\hat{\sigma}^2 d_0}{nh\hat{f}(x)}},$$

and

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{\beta}_1 Z_{i1} - ... - \hat{\beta}_q Z_{iq} - \hat{g}(Z_i)\}^2, \quad \hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(X_i - x).$$

The bias is $\frac{1}{2} g''(x) h^2$ *if local linear kernel estimator is used or* $\frac{1}{2} c_2 g''(x) h^2 + c_2 f^{-1}(x) f'(x) g'(x) h^2$ *if NW estimator is used.*

References

R. Engle, C. Granger, J. Rice & A. Weiss (1986). Semiparametric estimation of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81, 310-320.