Chapter 2. Semi-parametric Models (I)
Part 4

February 21, 2007

1 Statistical inference of the Single-index model

There are two set of estimators in the model: parameter vector α_0 and nonparametric link function $\phi(.)$. Suppose the estimators are $\hat{\alpha}$ and $\hat{\phi}(.)$ respectively.

For the estimator of α_0,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, W^{-1}\sigma^2)$$

in distribution, where

$$W = E\{[\phi'(\alpha_0^\top X)]^2 (X - E(X|\alpha_0^\top X))(X - E(X|\alpha_0^\top X))^\top\}.$$

In practice, it can be replaced by

$$\{n^{-1}\sum_{i=1}^{n}[\hat{\phi}'(\hat{\alpha}^\top X_i)]^2 (X_i - \hat{\mu}(\hat{\alpha}^\top X_i))(X_i - \hat{\mu}(\hat{\alpha}^\top X_i))\}^{-1}$$

where

$$\begin{pmatrix} \hat{\phi}(\hat{\alpha}^\top X_i) \\ \hat{\phi}'(\hat{\alpha}^\top X_i) \end{pmatrix} = \Big\{ \sum_{j=1}^{n} K_h(d_{ij}) \begin{pmatrix} 1 \\ d_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ d_{ij} \end{pmatrix}^\top \Big\}^{-1} \sum_{j=1}^{n} K_h(d_{ij}) \begin{pmatrix} 1 \\ d_{ij} \end{pmatrix} Y_i$$

where $d_{ij} = \hat{\alpha}^\top (X_i - X_j)$ and

$$\hat{\mu}(\hat{\alpha}^\top X_i) = \sum_{j=1}^{n} K_h(d_{ij}) X_j / \sum_{j=1}^{n} K_h(d_{ij}).$$

and σ^2 can be replaced by

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

where $\hat{\varepsilon}_i = \hat{Y}_i - Y_i$ is the fitted residuals and $\hat{y}_i$ is the fitted values

$$\hat{Y}_i = \hat{\phi}(\hat{\alpha}^\top X_i)$$

[based on this, write down the 95% confidence interval for the parameters]

For the estimator of ϕ, if $nh^4 \to 0$, then

$$\sqrt{nh}(\hat{\phi}(v) - \phi(v)) \sim N(0, \frac{d_0\sigma^2}{nhf(v)})$$

where $f(v)$ is the density function of $\alpha_0^\top X$. [based on this, write down the 95% confidence band for the link function]

Prediction using single-index model: for a new point X_0, we can predict its response by $\hat{\phi}(\hat{\alpha}^\top X_0)$ and the 95% confidence interval is

$$\hat{\phi}(\hat{\alpha}^\top X_0) \pm 1.96\sqrt{\frac{d_0\hat{\sigma}_2}{nh\hat{f}(\hat{\alpha}^\top X_0)}}.$$

Example 1.1 (Swiss banknotes data) *Now, we try a single-index model*

$$Y = \phi({\alpha_0}^\top X) + \varepsilon$$

where $X = (\mathbf{x}_1, \cdots, \mathbf{x}_6)^\top$ *The confidence band for the link function is shown in figure 1*

The estimated α *is*

$$0.2420208, -0.7634609, 0.2461061, -0.3085073, -0.2624723, 0.3659464$$

with standard error

$$0.04341794, 0.02662028, 0.05757453, 0.03430258, 0.03561638, 0.03177162$$

Suppose we need to check the following two banknotes

$$[215131131910141], \quad [214, 130, 130, 10, 12, 140)]$$

Then the predicted values are 0.09460568 *and 0.999997 respectively. The first is genuine and the second one counterfeit.*

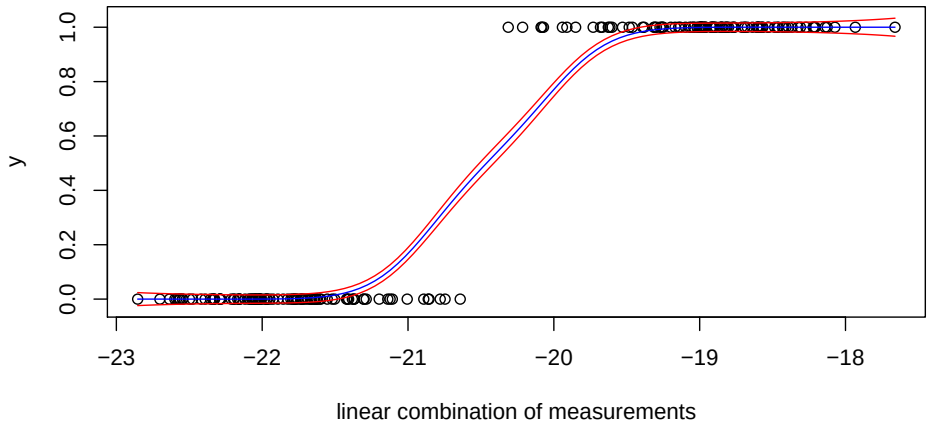

Figure 1: The estimated single-index model **(sim.R) (c2d0.R)**

2 Application of the single-index model

2.1 Model check for univariate linear regression model

Suppose we have data $(\mathbf{x}_i, Y_i), i = 1, ..., n$, we first try linear regression model

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_i + \varepsilon_i$$

The least squares estimator of β_0, β_1 are $\hat{\beta}_0, \hat{\beta}_1$. The fitted values are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_i.$$

Suppose the fitted residuals are

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

One way to check whether the linear regression model is adequate is by plotting the residuals against the regressor (independent variable). If any systematic departure from zero is found, then the model is not adequate.

Example 2.1 *For* **data 1** *and* **data 2**, *we consider linear regression model*

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_i + \varepsilon_i$$

the residuals are plotted in the two panels of figure 2. The linear regression for the first data is adequate but the second is not.

For the second data, we can consider higher order polynomial regression, say

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_i + \beta_2 \mathbf{x}_i^2 + \varepsilon_i$$

The residuals are plotted in the third panel. It can be regarded as adequate.

2.2 Model check for multivariate linear regression model

Suppose we have data $(X_i, Y_i), i = 1, ..., n$ where $X_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})^\top$. If we check whether a linear regression model

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + ... + \beta_p \mathbf{x}_{ip} + \varepsilon_i \tag{2.1}$$

is adequate or not, the above method usually fails.

Example 2.2 *Consider model*

$$Y = 2 + \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_1 \mathbf{x}_2 + \varepsilon$$

If $\mathbf{x}_1, \mathbf{x}_2, \varepsilon$ are independent and follow N(0, 1), you will find a linear regression model is adequate if you plot the residuals against $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively.

For **data 3** *(from the above model), we consider linear regression model*

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \varepsilon_i$$

the residuals are plotted in the two panels of figure 3 indicating that the model is adequate (Wrong!)

Instead of checking whether the residuals

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

has systematic departure from 0 for each regressor, we can consider checking whether there is a linear combination of the regressors $\alpha^\top X_i$, and plot the residuals against $\alpha^\top X_i$, such that the departure can be observed. In other words, we fit a single-index model

$$\hat{\varepsilon}_i = \phi(\alpha^\top X_i) + \xi_i \tag{2.2}$$

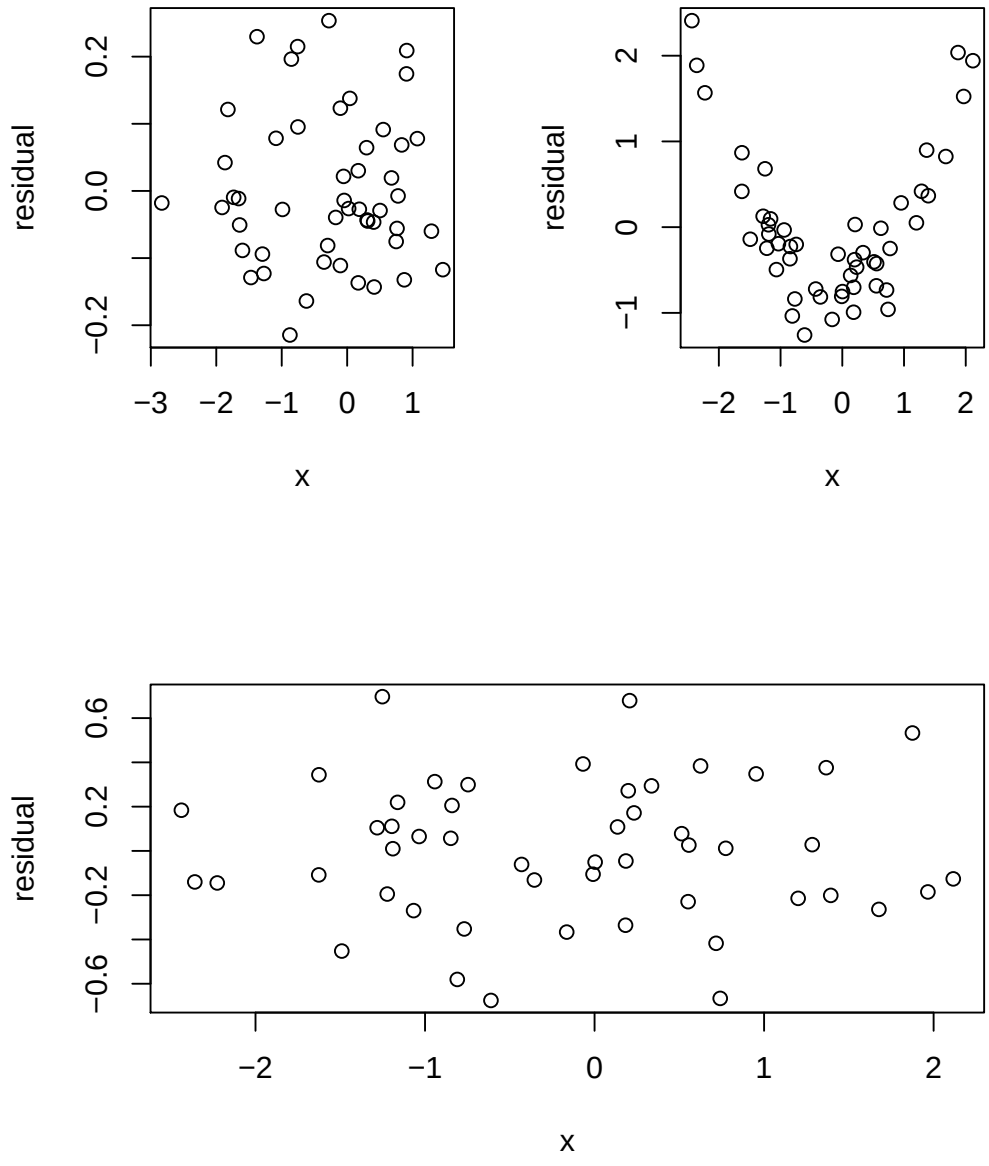

Figure 2: the plot of residuals against the covariate. **(c2d1.R) (c2d2.R) (c2d2s.R)**

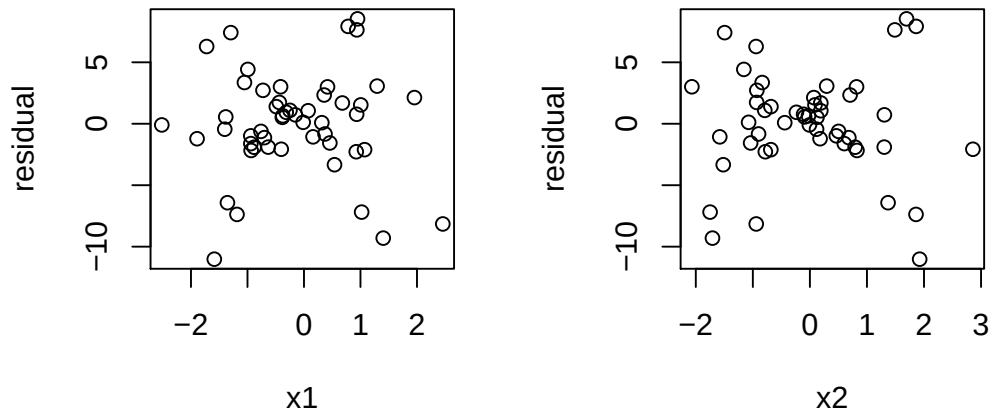

Figure 3: the plot of residuals against the covariate. **(c2d3.R)**

Suppose the estimated α in model (2.2) is $\hat{\alpha}$. Then we plot $\hat{\varepsilon}_i$ against $\hat{\alpha}^\top X_i$. If the departure is obvious, then the linear regression model (2.1) is not adequate, otherwise adequate.

Example 2.3 *For* **data 3** *(from the above model), we consider linear regression model*

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \varepsilon_i$$

for the residuals, we fit model

$$\hat{\varepsilon}_i = \phi(\alpha_1 \mathbf{x}_{i1} + \alpha_2 \mathbf{x}_{i2}) + \xi_i$$

The estimated single-index parameters are $0.7339013, -0.6792561$. *Now we plot the residuals against* $0.7339013\mathbf{x}_{i1} - 0.6792561\mathbf{x}_{i2}$ *in figure 4 indicating that the model is not adequate (Correct)*

3 Extended partially linear single-index models

If we write the above procedure as a model, we can consider

$$Y_i = \beta_1 \mathbf{x}_{i1} + ... + \beta_p \mathbf{x}_{ip} + \phi(\alpha_0^\top X_i) + \varepsilon_i.$$

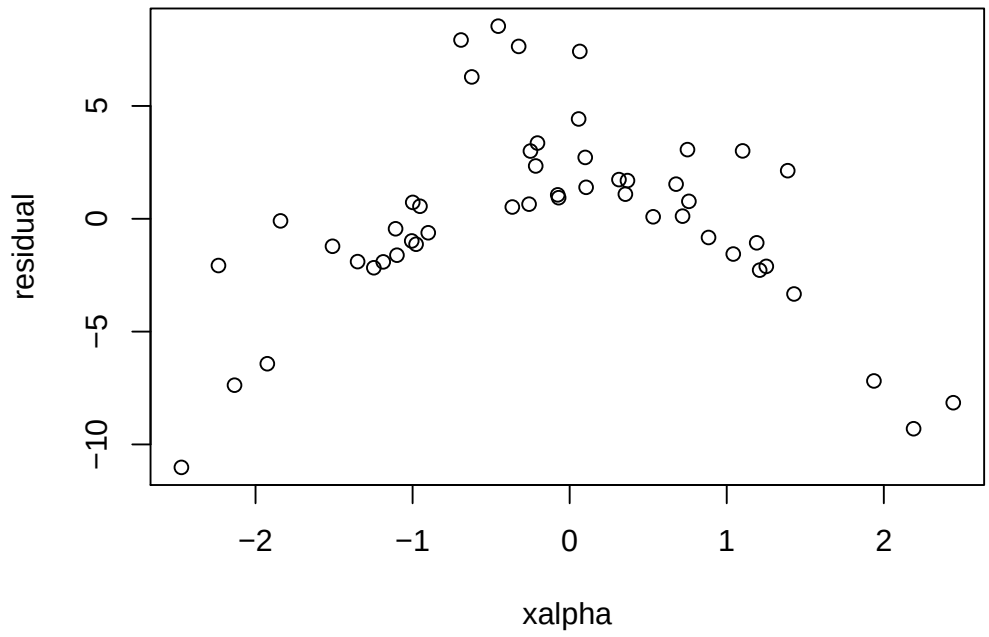

Figure 4: the plot of residuals against the covariate. **(c2d4.R)**

Here, we introduce a simple estimation method. we first estimate the linear part use least squares estimation and calculate the residuals. Then use single-index model to fit the residuals

Example 3.1 *For* **data** *about the baseball's players and their performance, we consider linear regression model*

$$Y_i = \beta_1 \mathbf{x}_{i,1} + ... + \beta_p \mathbf{x}_{i,16} + \phi(\alpha_0^\top X_i) + \varepsilon_i.$$

We first call $lm(y \sim x)$ *amd* $summary()$ *to estiamte the linear regression part. The estiamted model is*

$$\begin{aligned}\hat{Y}_i \quad = \quad & 4.618 - 0.003\mathbf{x}_{i,1} + 0.0139\mathbf{x}_{i,2} + 0.0087\mathbf{x}_{i,3} - 0.0014\mathbf{x}_{i,4} - 0.0003\mathbf{x}_{i,5} \\ & +0.011\mathbf{x}_{i,6} + 0.0558\mathbf{x}_{i,7} + 0.0001\mathbf{x}_{i,8} - 0.0006\mathbf{x}_{i,9} - 0.0003\mathbf{x}_{i,10} + 0.0017\mathbf{x}_{i,11} \\ & +0.0002\mathbf{x}_{i,12} - 0.0014\mathbf{x}_{i,13} + 0.0004\mathbf{x}_{i,14} + 0.0006\mathbf{x}_{i,15} - 0.0100\mathbf{x}_{i,16}\end{aligned}$$

Calculate residuals

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

We now fit the single-index model $\hat{\varepsilon}_i = \phi(\alpha^\top X_i) + \xi_i$. *The estimator of* α *are (0.005300291, 0.02133750, 0.1187392, -0.03996845, -0.03573508, -0.005695383, 0.9866392, -0.005173214, 0.01985905, 0.07435363, -0.007662053 -0.02958123, 0.007101752, 0.00, 0.01958818 -0.04166778) and their standard errors are (0.0033766704, 0.0119333402, 0.0348575584, 0.0153552103, 0.0135591564, 0.0100362666, 0.0048650937, 0.0006381937, 0.0033524727, 0.0100123527, 0.00385537 39, 0.0041333497, 0.0017020911, 0.0004796418, 0.0014283351, 0.0210736614). The estimated link function is shown in right panel of Figure 5*

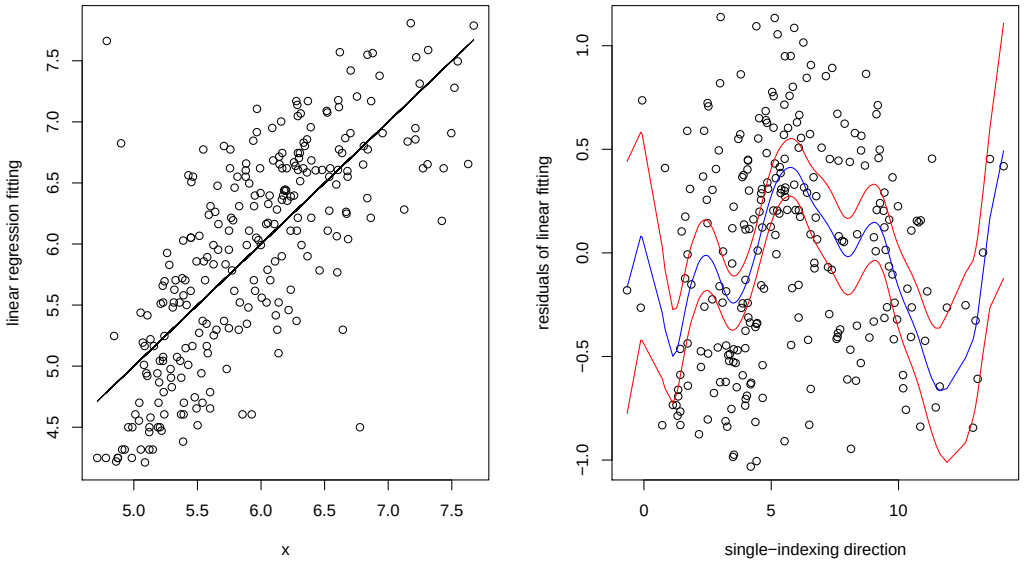

Figure 5: the plot of residuals against the covariate. **(sim.R) (c2d5.R)**