Lec 13: Kernel Regression

Weiping Zhang

2019.11.6

Univariate Kernel Regression

- Statistical Properties

Multivariate Kernel Regression

- Statistical Properties
- Bandwidth selection
- Kernel regression with mixed data

Univariate Kernel Regression

- The relationship between two variables,X and Y

$$Y = m(X)$$

  where $m(\cdot)$ is a function.

- Model:

$$Y_i = m(X_i) + \epsilon_i, i = 1, \ldots, n \tag{1}$$
$$E(Y|X = x) = m(x). \tag{2}$$

  (1): $Y = m(X)$ doesn't need to hold for the i'th observation. Error term ϵ.
  (2) The relationship holds on average, $m(x) = E(Y|X = x)$ is called the regression function (of Y on X).

- Goal: Estimate $m(\cdot)$ on basis of i.i.d observations $(X_i, Y_i), i = 1, \ldots, n$
- In a parametric approach

  $$m(x) = \alpha + \beta x$$

  and then estimate α and β.
- In the nonparametric approach: No prior restrictions on $m(\cdot)$.

Conditional Expectation (repetition)

- Let X and Y be two random variables with joint pdf $f(x, y)$. Then the conditional expectation of Y given that $X = x$

$$\begin{aligned} E(Y|X = x) &= \int y f(y|x) dy \\ &= \int y \frac{f(x, y)}{f_X(x)} dy \\ &= m(x) \end{aligned}$$

- Note: $m(x)$ might be quite nonlinear.

Example

- Consider the joint pdf

$$f(x,y) = x + y, \quad 0 \leq x, y \leq 1$$

then the marginal pdf is

$$f_X(x) = x + \frac{1}{2}, 0 \leq x \leq 1$$

- The conditional expectation is

$$\begin{aligned} E(Y|X=x) &= \int y \frac{f(x,y)}{f_X(x)} dy \\ &= \int_0^1 y \frac{x+y}{x+\frac{1}{2}} dy = \frac{x/2 + 1/3}{x + 1/2} = m(x) \end{aligned}$$

which is nonlinear.

Design

- Random design
  Observations $(X_i, Y_i), i = 1, \dots, n$ from bivariate distribution $f(x, y)$. The distribution of $f_X(x)$ is unknown.
- Fix design
  Control the predictor variable, X, then Y is the only random variable. The distribution of $f_X(x)$ is known.

Kernel regression–Random design

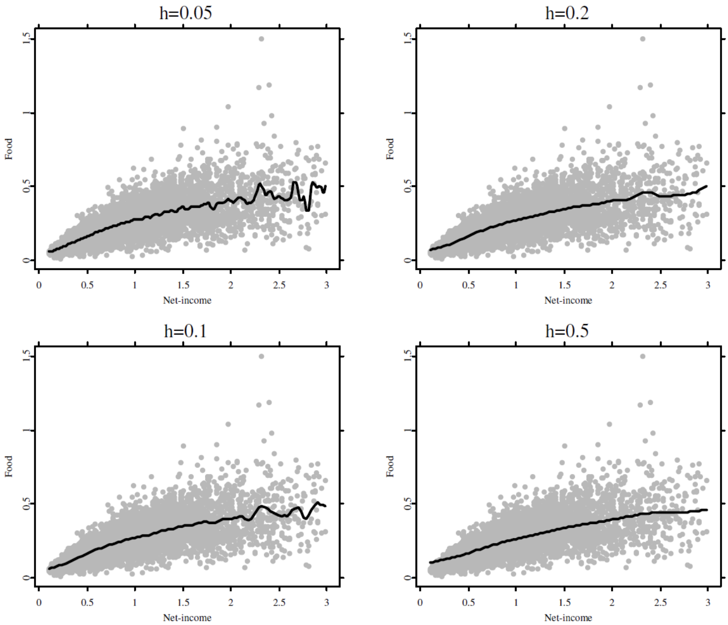
- The Nadaraya-Watson estimator

$$\hat{m}(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}$$

Rewrite the Nadaraya-Watson estimator

$$\begin{aligned}\hat{m}(x) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{K_h(x - X_i)}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}\Big)Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)Y_i\end{aligned}$$

- Weighted (local) average of Y_i (note: $\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x) = 1$)
- h determines the degree of smoothness.



Four kernel regression estimates for the 1973 U.K. Family Expenditure data with bandwidths $h = 0.05, h = 0.1, h = 0.2$, and $h = 0.5$

- What happens if the denominator of $W_{hi}(x)$ is equal to 0? Then the numerator is also equal to 0, and the estimator is not defined. This happened in regions of sparse data.
- **Local constant estimator(Nadaraya-Watson)**:

$$\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_h(x - X_i)$$

Then

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}$$

Kernel regression–Fix design

- $f_X(x)$ is known. Weights of the form

  $$W_{hi}^{FD}(x) = \frac{K_h(x - X_i)}{f_X(x)}$$

  Simpler structure and therefore easier to analyse.

- One particular fixed design kernel regression estimator: (Gasser-Müller)

  For the case of ordered design points $x_{(i)}, i = 1, \ldots, n$ from $[a, b]$

  $$W_{hi}^{GM}(x) = n \int_{s_{i-1}}^{s_i} K_h(x - u) du$$

  where $s_i = \frac{x_{(i)} + x_{(i+1)}}{2}$, so $s_0 = a, s_{n+1} = b$. Note that $W_{hi}^{GM}(x)$ sums to 1.

To show how the weights $W_{hi}^{GM}(x)$ are related to the intuitively appealing formula $W_{hi}^{FD}(x)$ note that by the mean value theorem

$$(s_i - s_{i-1})K_h(x - \xi) = \int_{s_{i-1}}^{s_i} K_h(x - u)du$$

for some ξ between s_i and s_{i-1}. Moreover

$$n(s_i - s_{i-1}) \approx \frac{1}{f_X(x)}$$

Therefore,

$$W_{hi}^{FD}(x) = \frac{K_h(x - X_i)}{f_X(x)} \approx n \int_{s_{i-1}}^{s_i} K_h(x - u)du = W_{hi}^{GM}(x)$$

Statistical Properties

Are the kernel regression estimator consistent?

Theorem (Consistence of Nadaraya-Watson)

Assume the univariate random design model and the regularity conditions: $\int |K(u)|du < \infty, uK(u) \to 0$ *for* $|u| \to \infty$, $EY^2 < \infty$. *Suppose also* $h \to 0$, $nh \to \infty$, *then*

$$\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)Y_i = \hat{m}(x) \xrightarrow{P} m(x)$$

where for x *holds* $f_X(x) > 0$ *and* x *is a point of continuity of* $m(x), f_X(x)$, *and* $\sigma^2(x) = Var(Y|X = x)$.

- Proof idea: Show that the numerator and the denominator of $\hat{m}_h(x)$ converge. Then $\hat{m}_h(x)$ converges (Slutsky's theorem).

What is the speed of the convergence in the random design?

Theorem (Speed of the convergence (Nadaraya-Watson))

Assume the univariate random design model and the regularity conditions: $\int |K(u)|du < \infty, uK(u) \to 0$ *for* $|u| \to \infty$, $EY^2 < \infty$. *Suppose also* $h \to 0$, $nh \to \infty$, *then*

$$MSE(\hat{m}(x)) \approx \frac{1}{nh}\frac{\sigma^2}{f_X(x)}\kappa_{02} + \frac{h^4}{4}\kappa_{21}^2\Big(m''(x) + 2\frac{m'(x)f_X'(x)}{f_X(x)}\Big)^2$$

where for x *holds* $f_X(x) > 0$ *and* x *is a point of continuity of* $m(x), f_X(x)$, *and* $\sigma^2(x) = Var(Y|X = x)$.

Proof idea: Rewrite the Nadaraya-Watson estimator:

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)}$$

Consider the numerator,

$$\begin{aligned} E\hat{r}_h(x) &= E\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)Y_i = EK_h(x-X_1)Y_1 \\ &= \iint yK_h(x-u)f(y|u)f_X(u)dydu \\ &= \int K_h(x-u)f_X(u)[\int yf(y|u)dy]du \\ &= \int K_h(x-u)f_X(u)g(u)du = \int K_h(x-u)r(u)du \end{aligned}$$

where we define $r(u) = f_X(u)g(u) = \int yf(y,u)dy$. Expanding, as in the kernel density case we have,

$$E\hat{r}_h(x) = r(x) + \frac{h^2}{2}r''(x)\kappa_{21} + o(h^2)$$

and where the linear term in h vanishes due to the mean zero assumption on K just as in density estimation,

$$\begin{aligned} Var(\hat{r}_h(x)) &= \frac{1}{n} Var(K_h(x - X_1)Y_1) \\ &= \frac{1}{n}\Big[\int K_h^2(x-u)\tau^2(u)f_X(u)du - (\int K_h(x-u)r(u)du)^2\Big] \\ &= \frac{1}{nh}\int K^2(v)\tau^2(x+vh)f_X(x+vh)dv + o(\frac{1}{nh}) \\ &= \frac{1}{nh} f_X(x)\tau^2(x)\kappa_{02} + o(\frac{1}{nh}) \\ &= \frac{1}{nh} f_X(x)\sigma^2(x)\kappa_{02} + o(\frac{1}{nh}) \end{aligned}$$

where $\sigma^2(x) = Var(Y|X = x)$, $\tau^2(x) = E[Y^2|X = x]$. So

$$MSE(\hat{r}_h(x)) = \frac{1}{nh} f_X(x)\sigma^2(x)\kappa_{02} + \frac{h^4}{4}(r''(x))^2\kappa_{21}^2 + o(h^4) + o(\frac{1}{nh})$$

Since

$$\begin{aligned}\hat{m}_h(x) - m(x) &= \Big(\frac{\hat{r}_h(x)}{\hat{f}_h(x)} - m(x)\Big)\Big(\frac{\hat{f}_h(x)}{f_X(x)} + \Big(1 - \frac{\hat{f}_h(x)}{f_X(x)}\Big)\Big) \\ &= \frac{\hat{r}_h - m\hat{f}_h}{f_X} + (\hat{m} - m)\Big(\frac{f_X - \hat{f}_h}{f_X}\Big) \\ &= O_p(n^{-2/5}) + o_p(1)O_p(n^{-2/5}) = O_p(n^{-2/5}).\end{aligned}$$

where we have let $h = O(n^{-1/5})$ to compute the orders in the last line. So we can focus on the first term

$$\begin{aligned}&\frac{1}{f_X^2}E(\hat{r}_h - m\hat{f})^2 = \frac{1}{(nf_X)^2}E\Big[\sum_{i=1}^{n} K_h(x - X_i)(Y_i - m(x))\Big]^2 \\ &\quad = \frac{1}{nf_X^2}Var\Big[K_h(x - X_1)(Y_1 - m(x))\Big] \\ &\quad + \frac{1}{f_X^2}E^2\Big[K_h(x - X_1)(Y_1 - m(x))\Big]\end{aligned}$$

and this yields,

$$MSE(\hat{m}(x)) = \frac{1}{nh}\frac{\sigma^2}{f_X(x)}\kappa_{02} + \frac{h^4}{4}\kappa_{21}^2\Big(m''(x) + 2\frac{m'(x)f_X'(x)}{f_X(x)}\Big)^2 \\ + o(\frac{1}{nh}) + o(h^4)$$

So

$$h = O(n^{-1/5}) \Rightarrow MSE(\hat{m}_h(x)) = O(n^{-4/5}).$$

- Asymptotic MSE:

$$AMSE(\hat{m}(x)) = \frac{1}{nh}C_1 + h^4 C_2$$

Minimizing wrt. h gives the optimal bandwidth

$$h_{opt} \sim n^{-1/5}$$

- Rate of convergence af AMSE is of order $O(n^{-4/5})$.
- Slower than the rate obtained by LS estimation in linear regression, but the same as in nonparametric density estimation.

Multivariate Kernel Regression

- we will mostly be interested in specifying how the response variable Y depends on a vector of exogenous variables, denoted by $\mathbf{X}$. This means we aim to estimate the conditional expectation

  $$E(Y|\mathbf{X}) = E(Y|X_1, \ldots, X_d) = m(\mathbf{X}),$$

  where $\mathbf{X} = (X_1, \ldots, X_d)'$
- Goal: estimate $m(\mathbf{x})$ based on i.i.d observations $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$

- Consider

$$E(Y|\mathbf{X} = \mathbf{x}) = \int y f(y|\mathbf{x}) dy = \frac{\int y f(y, \mathbf{x}) dy}{f_X(\mathbf{x})}$$

If we replace the multivariate density $f(y, \mathbf{x})$ by its kernel density estimate

$$\hat{f}_{h,H}(y, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(y - Y_i) \mathcal{K}_H(\mathbf{x} - \mathbf{X}_i)$$

and $f_X(\mathbf{x})$ by

$$\hat{f}_H(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_H(\mathbf{x} - \mathbf{X}_i)$$

we arrive at the multivariate generalization of the Nadaraya-Watson estimator:

$$\begin{aligned}\hat{m}_H(\mathbf{x}) &= \frac{\sum_{i=1}^{n} \mathcal{K}_H(\mathbf{x} - \mathbf{X}_i) Y_i}{\sum_{i=1}^{n} \mathcal{K}_H(\mathbf{x} - \mathbf{X}_i)} \\ &= \sum_{i=1}^{n} W_{Hi}(\mathbf{x}) Y_i\end{aligned}$$

- Hence, the multivariate kernel regression estimator is again a weighted sum of the observed responses Y_i.
- Note also, that the multivariate Nadaraya-Watson estimator is a local constant estimator:

$$\hat{m}_H(\mathbf{x}) = \arg\min_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 \mathcal{K}_H(\mathbf{x} - \mathbf{X}_i)$$

Statistical Properties

Theorem (Asymptotic Normality)

Under some regularity conditions, and $n \to \infty$, $nh_1 \cdots h_d \to \infty$, $nh_1 \cdots h_d \sum_{s=1}^{d} h_s^4 \to 0$ *and* $h_s \to 0, s = 1, \ldots, d$, *we have*

$$\sqrt{nh_1 \cdots h_d}[\hat{m}(\mathbf{x}) - m(\mathbf{x}) - \frac{\kappa_{21}}{2} \sum_{s=1}^{d} B_s(\mathbf{x}) h_s^2] \rightsquigarrow N(0, \kappa_{02}^d \sigma^2(\mathbf{x})/f(\mathbf{x}))$$

where $B_s(\mathbf{x}) = 2f_s(\mathbf{x})m_s(\mathbf{x})/f(\mathbf{x}) + m_{ss}(\mathbf{x})$ *with* $f_s(\mathbf{x}), m_s(\mathbf{x})$ *being the partial derivatives of* $f(\mathbf{x})$ *and* $m(\mathbf{x})$ *w.r.t the* s*th coordinate* x_s*, respectively.* $m_{ss}(\mathbf{x})$ *is the second order partial derivative of* $m(\mathbf{x})$ *w.r.t* x_s*.*

Statistical Properties

Theorem

The conditional asymptotic bias and variance of the multivariate Nadaraya-Watson kernel regression estimator are

$$\begin{aligned} MSE(\hat{m}(\mathbf{x})) &= E[\hat{m}(\mathbf{x}) - m(\mathbf{x})]^2 \\ &= \Big[\frac{\kappa_{21}}{2}\sum_{s=1}^{d} B_s(\mathbf{x}) h_s^2\Big]^2 + \frac{1}{nh_1\cdots h_d f_X(\mathbf{x})}\sigma^2(\mathbf{x})\kappa_{02}^d \\ &+ O\Big(\sum_{s=1}^{d} h_s^4\Big) + O\Big(\Big[\frac{1}{nh_1\cdots h_d}\sum_{s=1}^{d} h_s^4\Big]^{1/2}\Big). \end{aligned}$$

Theorem

Under certain regularity conditions, we have

$$(1) \sup_{x \in \mathcal{S}} |\hat{m}(\mathbf{x}) - m(\mathbf{x})| = o\Big(\frac{1}{\sqrt{nh_1 \cdots h_d}} \sqrt{lnn}\Big) + o\Big(\sum_{s=1}^{d} h_s^2\Big) \quad a.s.$$

$$(2) \sup_{x \in \mathcal{S}} E|\hat{m}(\mathbf{x}) - m(\mathbf{x})|^2 = o\Big(\frac{1}{nh_1 \cdots h_d}\Big) + o\Big(\sum_{s=1}^{d} h_s^4\Big) \quad a.s.$$

where $\mathcal{S}$ is a compact set on $\mathcal{R}^d$ contained in the support of f_X.

See Masry (1996a,b) for the proof of the above theorem.

Bandwidth selection

- The bandwidth is important:

$$\hat{m}(X_i) \to Y_i, \quad h \to 0$$

$$\hat{m}(X_i) \to \bar{Y}, \quad h \to \infty$$

- **Rule-of-thumb** When using a second order kernel, it can be shown that the optimal bandwidth is of order $O\left(n^{-1/(4+d)}\right)$. A popular rule-of-thumb procedure is to choose

$$h_s = c_s \hat{\sigma}_s n^{-1/(4+d)}$$

  where $\hat{\sigma}_s$ is the sample standard deviation of $\{X_{is}\}_{i=1}^{n}$ and c_s is a constant depending on the kernel in use. For example, if one uses the Gaussian kernel, $\hat{c}_s$ is often chosen to be 1.06 or 1 in practice, if one uses the Epanechnikov kernel, $\hat{c}_s$ is often chosen to be 2.34.

- The argument is same to the rule-of-thumb reference rule for density estimation. This method is easy to use but its disadvantage is obvious. It lacks flexibility because different bandwidth sequences may be called for depending on the tail density behavior of the different components of X_i.
- **Plug-in Methods** An alternative method is to use the "plug-in" method which is based upon minimizing a weighted integrated mean square error (WMISE) of the form

$$WMISE(h_1, \ldots, h_d) = \int E[\hat{m}(\mathbf{x}) - m(\mathbf{x})]^2 w(\mathbf{x}) d\mathbf{x}$$

where the expectation is taken with respect to the random sample $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ and $w(\mathbf{x})$ is a non-negative weight function which truncates some bad estimates of $\hat{m}(\mathbf{x})$, say, when $\mathbf{x}$ is close to the boundary of its support. Simple calculations show

$$WMISE(h_1, \ldots, h_d) = \int \Big\{ \Big[\frac{\kappa_{21}}{2} \sum_{s=1}^{d} B_s(\mathbf{x}) h_s^2 \Big]^2 + \frac{\sigma^2(\mathbf{x}) \kappa_{02}^d}{n h_1 \cdots h_d f_X(\mathbf{x})} \Big\} w(\mathbf{x}) d\mathbf{x}$$

Let h_{s0} be the smoothing parameter that minimizes $WMISE$. One can show that

$$h_{s0} = a_s n^{-1/(4+d)}$$

for $s = 1, \ldots, d.$, where a_s depends on the unknown functions m, f and their derivatives. It is fairly straightforward to obtain explicit expression for a_s for $d \leq 2$ but not necessarily for large d. When a closed form expression for a_s exists, one can obtain a consistent estimator $\hat{a}_s$ for a_s usually by plugging the pilot estimates for m, f, and their derivatives.

- **Least Squares Cross Validation** Now we introduce a completely data-driven method for choosing the bandwidth parameters known as the "least-squares cross-validation" method. We choose $h = (h_1, \ldots, h_d)$ to minimize the following least squares cross-validation criterion function

$$CV_{lc}(h) = \frac{1}{n} \sum_{i=1}^{m} [Y_i - \hat{m}_{-i}(\mathbf{X}_i)]^2 w(\mathbf{X}_i),$$

where

$$\hat{m}_{-i}(\mathbf{X}_i) = \frac{\sum_{j=1, j \neq i}^{n} K_h(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{j=1, j \neq i}^{n} K_h(\mathbf{X}_i - \mathbf{X}_j)}$$

is the leave-one-out kernel estimator of $m(\mathbf{X}_i)$, and $w(\cdot)$ is a nonnegative weight function which truncates some bad estimates of $\hat{m}_{-i}(\mathbf{X}_i)$ caused by the so-called boundary effect or random denominator issue.

Let $\hat{h} = (\hat{h}_1, \ldots, \hat{h}_d)$ denote the solution to the above cross-validation problem. Then one can show

$$\frac{\hat{h}_s}{h_{s0}} \to 1$$

in probability for $s = 1, \ldots, d$.

Kernel regression with mixed data

■ Non-continuous predictors can be also taken into account in nonparametric regression. The key for doing so is an adequate definition of a suitable kernel function for any random variable X, not just continuous. Therefore, we need to find *a positive function that is a pdf on the support of X and that allows to assign more weight to observations of the random variable that are close to a given point.*

■ We analyse next the two main possibilities for non-continuous variables:

- *Categorical or unordered discrete variables.* Categorical variables are specified in base R by factor(). Due to the lack of ordering, the basic mathematical operation behind a kernel, a distance computation, is senseless. This motivates the Aitchison and Aitken (1976) kernel:

- Assume that the categorical random variable X_d has c_d different levels. Then, it can be represented as $X_d \in C_d := \{0, 1, \ldots, c_d - 1\}$. For $x_d, X_d \in C_d$, the Aitchison and Aitken (1976) [1] unordered discrete kernel is

$$l_u\left(x_d, X_d; \lambda\right) := \begin{cases} 1 - \lambda, & \text{if } x_d = X_d \\ \frac{\lambda}{c_d - 1}, & \text{if } x_d \neq X_d \end{cases}$$

  where $\lambda \in \left[0, \left(c_d - 1\right)/c_d\right]$ is the bandwidth. Observe that this kernel is constant if $x_d \neq X_d$ since the levels of the variable are unordered, there is no sense of proximity between them.

- *Ordinal or ordered discrete variables.* These variables are specified by *ordered* (an ordered factor in base R). Despite the existence of an ordering, the possible distances between the observations of these variables are discrete.

[1] https://doi.org/10.1093/biomet/63.3.413.

- If the ordered discrete random variable X_d can take c_d different ordered values, then it can be represented as $X_d \in C_d := \{0, 1, \ldots, c_d - 1\}$. For $x_d, X_d \in C_d$, a possible (Li and Racine 2007[2]) ordered discrete kernel is

  $$l_o\left(x_d, X_d; \eta\right) := \eta^{|x_d - X_d|}$$

  where $\eta \in [0, 1]$ is the bandwidth.
- Once we have defined the suitably kernels for ordered and unordered discrete variables, we can define the Nadaraya-Watson for mixed multivariate data. Assume that, among the $p_c + p_u + p_o$ predictors, the first p_c are continuous, the next p_u are discrete unordered (or categorical), and the last p_o are discrete ordered (or ordinal).

[2] Li, Qi, and Jeffrey Scott Racine. 2007. Nonparametric Econometrics. Princeton, NJ: Princeton University Press.

$$\hat{m}\left(\mathbf{x};0,\left(\mathbf{h}_c,\boldsymbol{\lambda}_u,\boldsymbol{\eta}_o\right)\right):=\sum_{i=1}^{n}W_i^0(\mathbf{x})Y_i$$

where $\mathbf{h}_c=\left(h_1,\ldots,h_{p_c}\right),\boldsymbol{\lambda}_u=\left(\lambda_1,\ldots,\lambda_{p_u}\right),\boldsymbol{\eta}_o=\left(\eta_1,\ldots,\eta_{p_o}\right),$ and

$$W_i^0(\mathbf{x})=\frac{L_\Pi\left(\mathbf{x}-\mathbf{x}_i\right)}{\sum_{j=1}^{n}L_\Pi\left(\mathbf{x}-\mathbf{x}_i\right)}$$

$$L_\Pi\left(\mathbf{x}-\mathbf{x}_i\right):=\prod_{j=1}^{p_c}K_{h_j}\left(x_j-X_{ij}\right)\prod_{k=1}^{p_u}l_u\left(x_k,X_{ik};\lambda_j\right)\prod_{\ell=1}^{p_o}l_o\left(x_\ell,X_{it};\eta_j\right)$$

- The np package employs a variation of the previous kernels and implements the local constant and linear estimators for mixed multivariate data.

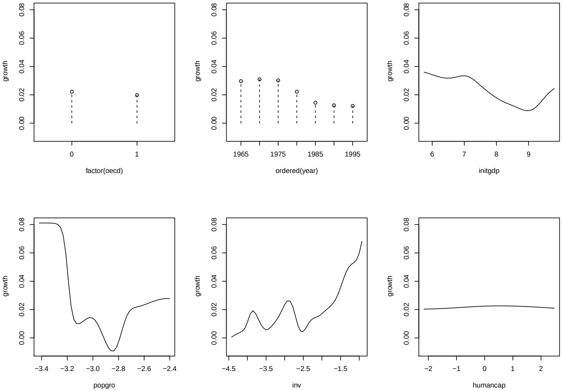


Figure 1: Marginal effects of each predictor on the response for the oecdpanel data in np package