Local Polynomial Variance-Function Estimation

Mushfiqur Rahman

- Motivation
- Methodology
- Data and variables
- Results
- Other works

References:

- David Ruppert et. al., "Local polynomial Variance -Function Estimation", Technometrics, 1997, Vol. 39, No. 3
- J.S. Simonoff, "Smoothing Methods in Statistics",1996, Springer.
- M.J.Buckley et al.,"The estimation of residual variance in nonparametric regression", Biometika,1988, Vol. 75, 189-199.

Motivation

- In regression analysis, it is often the case that the homoscedasticity assumption is violated.
- To see heteroscedasticity in linear regression we fit

$$Y = X\beta + \epsilon$$

  where $var(\epsilon_i) = \sigma_i^2, \Rightarrow \sigma_i^2 = \sigma^2\omega_i$, i=1,...n
- In the presence of heteroscedasticity the estimates remain unbiased but they loose efficiency.

- In case of nonparametric regression with severe heteroscedasticity, the variance function must be estimated to obtain a satisfactory bandwidth for the derivative and to estimate the variance of the response.
- In nonparametric regression we fit

$$Y = m(X) + \epsilon$$

where $var(\epsilon_i) = \sigma_i^2, \Rightarrow \sigma_i^2 = \sigma^2 \omega_i$, i=1,...n

Methodology

- Model and Assumptions: Let (X_1, Y_1),. . . ,(X_n, Y_n) be a sample of random pairs that are assumed to satisfy the heteroscedastic nonparametric regression model

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, ..., n \tag{1}$$

  where, the errors ϵ_1, ... ,ϵ_n are independent zero mean random variables satisfying

$$i) var(\epsilon_i) = v(X_i), \qquad ii) E(\epsilon_i)^4 < \infty$$

- Here m is mean function and v is variance function. Both $m(X_i)$ and $v(X_i)$ are column vectors.
- Suppose that $\hat{m} = SY$ where S is $n \times n$ smoother matrix. Let S_1 = smoother matrix corresponding to an initial smooth of the data and $r = (I - S_1)Y$ is the vector of residuals. Smoothing the squard residuals we obtain $S_2 r^2$, where S_2 is another smoother matrix.
- Under homoscedasticity:

$$E[S_2 r^2 | X_1, ..., X_n] = S_2[\{E(S_1 Y | X_1, ..., X_n) - m\}^2 + \sigma^2(1 + \Delta)]$$

where Δ = column vector of $diag(S_1 S_1' - 2S_1)$
- If $S_1 Y$ is conditionally unbiased we can write

$$E[S_2 r^2 | X_1, ..., X_n] = S_2[\sigma^2(1 + \Delta)] = \sigma^2(1 + S_2 \Delta)$$

- The estimate of the variance function:

$$\hat{v} = (S_2 r^2)/(1 + S_2 \Delta) \tag{2}$$

Examples

- Parametric Linear model: $Y_i = (X\beta)_i + \epsilon_i$, where $var(\epsilon_i) = v(X_i)$. One should replace S_1 by the hat matrix $R = X(X'X)^{-1}X'$. Using symmetry and idempotency of R we obtain the variance function estimator

$$\hat{v} = S_2(R - I)Y^2/[1 - S_2 diag(R)]$$

- For homoscedastic parametric regression the estimator of the variance function reduces to the usual

$$\hat{\sigma}^2 = Y'(I - R)Y/(n - p)$$

- If the homoscedastic nonparametric regression is assumed. Taking $S_2 = n^{-1}11'$ the varience is estimated by

$$\hat{\sigma}^2 = [Y'(S_1 - I)'(S_1 - I)Y]/[n + tr(S_1 S_1' - 2S_1)]$$

Local polynomial variance function

- Local polynomial regression estimator is the minimizer of

$$\sum_{i=1}^{n}[Y_i - \beta_0 - ... - \beta_p(x - x_i)^p]^2 K\left(\frac{x - x_i}{h}\right) \tag{3}$$

- Let X_x be the design matrix

$$\begin{pmatrix} 1 & x - x_1 & \dots & (x - x_1)^p \\ \vdots & \vdots & \dots & \vdots \\ 1 & x - x_n & \vdots & (x - x_n)^p \end{pmatrix}$$

- Let

$$W_x = h^{-1} diag \left[K\left(\frac{x-x_1}{h}\right), \ldots, K\left(\frac{x-x_n}{h}\right)\right]$$

be the weight matrix.

- The estimator $\hat{m}_p(x)$ is then the intercept $\hat{\beta}_0$

$$\hat{m}_p(x) = e_1'(X_x'W_xX_x)^{-1}X_x'W_xY = S_xY,$$

where e_r is the $(p+1)\times$ 1 vector having the value 1 in the rth entry and zero elsewhere and

$$S_x = e_1'(X_x'W_xX_x)^{-1}X_x'W_x.$$

- The (i, j)th element of the pth-degree local polynomial smoother matrix, S is

$$(S_x)_{ij} = e_1'(X_x'W_xX_x)^{-1}X_x'W_xe_j.$$

- Using this notation, one can define the local polynomial estimate of $v(x)$ to be

$$\begin{aligned}\hat{v}(x) = \hat{v}(x; p_1, h_1, p_2, h_2) \\ &= \frac{e_1'(X_{xp_2}' W_{xh_2} X_{xp_2})^{-1} X_{xp_2}' W_{xh_2} r^2}{1 + e_1'(X_{xp_2}' W_{xh_2} X_{xp_2})^{-1} X_{xp_2}' W_{xh_2} \Delta}.\end{aligned}$$

where $r = (I - S_{p_1,h_1})Y$ and $\Delta = diag(S_{p_1,h_1} S_{p_1,h_1}' - 2S_{p_1,h_1})$

Data and Variables

- I used the data from /home/nancy/data/simonoff/whale.dat.
- I fit both Hudson and Casey nursing time data using Splus/R function ”gam”, ”ksmooth”, ”locpoly” where time is taken as smooth.
- Residual plots indicates the heteroscedasticity.
- Using the second degree polynomial I got the variance esimates.
- Also I have taken a simulation study.

Other work on nonparametric variance-function estimation

Work on nonparametric variance estimation can ne categorized according to the following criteria:

- The mean function may be parametrically or nonparametrically modeled.
- The variance function may be considered constant(homoscedastic) or nonconstant(heteroscedastic).
- In a first stage of estimation, the variance at $X_i, v(X_i)$, may be estimated by a residual from a preliminary fit. Alternatively, one may use a squared "pseudo-residual", which is a weighted average of a fixed (independent of n) number of the $Y_i's$.
- In a second estimation stage, the squared residuals or pseudo-residuals may be kernel-smoothed of smoothed by local polynomial regression.

- The bandwidth for smoothing the residuals or pseudo-residuals may be chosen subjectively or by a data-based method.

Bias and covariance of the estimate

- Let $V = diag(v)$ then the bias of $\hat{v}$ can be written as

$$E(\hat{v}-v|X) = \frac{(S_2 - I)v + S_2[b_1^2 + diag(S_1 V S_1' - 2S_1 V)] - (S_2\Delta)v}{1 + S_2\Delta}$$

and if we assume that the errors are normally distributed then covariance of the estimate can be written as

$$cov(\hat{v}|X) = \frac{2S_2[\{(S_1-I)V(S_1-I)'\}\{(S_1-I)V(S_1-I)'+2b_1b_1'\}]S_2'}{(1+S_2\Delta)(1+S_2\Delta)'}$$

where $b_1 = (S_1 - I)m$ denote the basis vector of the smooth S_1.