Non and Semiparametric Methods of Regression Analysis

May 30, 2014

Regression Analysis

In statistics, regression analysis is a statistical process for estimating the relationships among variables.

A regression model relates Y to a function of X and β

$$Y \approx f(X, \beta),$$

where the approximation is formalized as $E[Y|X] = f(X, \beta)$.

Linear Regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Consumer demand for electricity.

Nonparametric Model

Allowing for nonlinearity:

$$Y_i = g(V_i) + \epsilon_i, \tag{2}$$

where $(Y, V) \doteq \{(Y_i, V_i)\}_{i=1}^n \subset \mathbb{R} \times \mathbb{R}^d$ are assumed to be $i.i.d.$, and $E(\epsilon|v) = 0$.

When $d > 3$, (2) suffers from curse-of-dimensionality:

$$E||\hat{g}(v) - g(v)||^2 \leq O(n^{-2/(2+d)}).$$

Semiparametric models as alternatives:

Consumer demand for electricity.

Partially Linear (PL) Semiparametric Time Series Model

Semi-Linear Model:

$$Y_t = X_t'\beta + g(V_t) + \epsilon_t, \tag{3}$$

where $g(\cdot) : \mathbb{R}^q \to \mathbb{R}$, $(Y, X, V) \doteq \{(Y_t, X_t, V_t)\}_{t=1}^n \doteq \mathcal{W}_t \subset \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q$,

$E(\epsilon_t | X_t, V_t, \mathcal{M}_{-\infty}^{t-1}) = 0$ and $\mathcal{M}_s^t$ is $\sigma(\mathcal{W}_s, \ldots \mathcal{W}_t)$ for $s \leq t$.

A number of studies in the literature show $\sqrt{n}$ consistency of $\hat{\beta}$:

1. Robinson (1988) and Speckman (1988) for *i.i.d.* case.
2. Fan and Li (1999) for stationary time series with an absolutely regular process-i.e., β-mixing condition.
3. Xia, Tong and Li (1999) for stationary time series with a strongly regular process-.i.e, α-mixing condition.

Partially Linear (PL) Semiparametric Time Series Model

- Estimation procedure:
  1. Partially-out $g(v)$ function from (3) using the conditional expectation relation:

$$\begin{aligned} E(y|v) &= E(x|v)'\beta + g(v) \\ Y_t - E(Y_t|V_t) &= \{X_t - E(X_t|V_t)\}'\beta + \epsilon_t \\ W_t &= U_t'\beta + \epsilon_t. \end{aligned}$$

  2. Estimate β using the least square estimation method:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{t=1}^{n}\hat{U}_t\hat{U}_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\hat{U}_t\hat{W}_t\right),$$

  where $\hat{U}_t = X_t - \hat{E}(X_t|V_t)$ and $\hat{W}_t = Y_t - \hat{E}(Y_t|V_t)$.

  3. Given $\hat{\beta}$, identify $\hat{g}(v)$:

$$\hat{g}(v) = \hat{E}(y|v) - \hat{E}(x|v)'\hat{\beta}.$$

Partially Linear (PL) Semiparametric Time Series Model

Semi-Linear Model:

$$Y_t = \mu + X_t'\beta + g(V_t) + \epsilon_t \tag{4}$$

Identification condition for (4) is $E(g(v)) = 0$; see Gao(2007).

Additive Semiparametric Time Series Model

$$Y_t = X_t'\beta + \sum_{l=1}^{q} g_l(V_{lt}) + \epsilon_t, \tag{5}$$

where $g_l(\cdot) : \mathbb{R} \to \mathbb{R}$ and $E(g_l(v_l)) = 0$ for all $l = 1, \ldots, q$.

See Gao, Lu and Tjøstheim (2006) for details.

Estimation procedure:

1. Repeat the estimation procedure Steps 1. - 3., given that $g(v) = \sum_{l=1}^{q} g_l(v_l)$ and $E(g_l(v_l)) = 0$.
2. Perform the marginal integration technique of Linton and Nielson (1995) to identify individual $g_l(\cdot)$-function.

Additive Semiparametric Time Series Model

Marginal integration technique:

For a fixed value for v_k, $g_k(v_k) = E[g(V_{t1}, \ldots, v_k, \ldots, V_{tq})]$,

$$g_k(v_k) = \int g(V_{t1}, \ldots, v_k, \ldots, V_{tq}) dQ(v_1), dQ(v_2), \ldots, dQ(v_q),$$

where Q is a deterministic weighting function with $\int dQ(v_i) = 1$ for $i = 1, \ldots k$.

Hence,

$$\hat{g}_k(v_k) = \frac{1}{n} \sum_{t=1}^{n} \hat{g}(V_{t1}, \ldots, v_k, \ldots, V_{tq}).$$

Single-Index (SI) Semiparametric Time Series Model

$$Y_t = g(V_t'\alpha_0) + \epsilon_t, \tag{6}$$

where $g(\cdot) : \mathbb{R} \to \mathbb{R}$, $(Y, V) \doteq \{(Y_t, V_t)\}_{t=1}^n \doteq \mathcal{W}_t \subset \mathbb{R} \times \mathbb{R}^q$ and

$E(\epsilon_t | V_t, \mathcal{M}_{-\infty}^{t-1}) = 0$ and $\mathcal{M}_s^t$ is $\sigma(\mathcal{W}_s, \ldots \mathcal{W}_t)$ for $s \leq t$.

A number of studies in the literature show $\sqrt{n}$ consistency of $\hat{\alpha}$:

1. Ichimura (1993) under *i.i.d.* case.
2. Härdle, Hall and Ichimura (1993) under *i.i.d.* case for choosing the optimal smoothing parameter.
3. Xia, Tong and Li (1999) under strongly regular condition.
4. Blundell and Power (2004) under *i.i.d.* case for the discrete dependent variable case.

Single-Index (SI) Semiparametric Time Series Model

Estimation procedure (see Horowitz (2003) for details):

1. Given α, estimate $\hat{g}(v'\alpha)$ using the nonparametric technique.
2. Estimate $\hat{\alpha}$ by minimising the objective function such as:

$$\min_{\alpha} S(\alpha) = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{g}(V_t'\alpha)\right)^2 .$$

3. Given $\hat{\alpha}$, estimate $\hat{g}(v'\hat{\alpha})$.

Note that Step 2 is using the similar argument as in the parametric nonlinear regression case (see Amemiya (1985), for details).

Single-Index (SI) Semiparametric Time Series Model

A few extensions:

1. Generalised partially linear single-index (GPLSI) model of Carroll, Fan, Gijbels and Wands (1997):

$$Y_t = X_t'\beta + g(V_t'\alpha_0) + \epsilon_t.$$

2. Extended generalised partially linear single-index (EGPLSI) model of Xia, Tong and Li (1999):

$$Y_t = X_t'\beta + g(X_t'\alpha_0) + \epsilon_t,$$

where $\beta \perp \alpha_0$ and $||\alpha_0|| = 1$.

Specification Testing for Nonparametric Models

Hypotheses:

$$H_{01} : m(x) = m_{\theta_0}(x) \ versus \ H_{11} : m(x) = m_{\theta_1}(x) + C_n \Delta_n(x) \quad (7)$$

Härdle and Mammen (1993):

$$M_{1n}(h) = nh^{d/2} \int \{\hat{m}(x) - \tilde{m}_{\hat{\theta}}(x)\}^2 w(x) dx.$$

Horowitz and Spokoiny (2001) (discrete approximation):

$$M_{2n}(h) = \sum_{t=1}^{n} \left(\hat{m}(X_t) - \tilde{m}_{\hat{\theta}}(X_t)\right)^2.$$

Gao and Gijbels (2008):

$$M_{3n}(x) = \frac{h^{d/2}}{n} \sum_{s=1}^{n} \sum_{t=1}^{n} \hat{\epsilon}_s K_h(X_t - X_s) \hat{\epsilon}_t,$$

where $\hat{\epsilon}_t = y_t - \tilde{m}_{\hat{\theta}}(X_t)$.

Specification Testing for Partially Linear Models

Hypotheses:

$$H_{02} : m(x, v) = x'\beta + g(v) \tag{8}$$

$$H_{12} : m(x, v) = x'\beta + g(v) + C_n\Delta_n(x, v)$$

The test statistic is

$$L_{2n}(x) = \sum_{s=1}^{n}\sum_{t=1}^{n} \hat{\epsilon}_s K_h(X_t - X_s)\hat{\epsilon}_t,$$

where $\hat{\epsilon}_t = y_t - X_t'\hat{\beta} - \hat{g}(V_t)$.

- Fan and Li (1996) consider the PL test for *i.i.d.* case.
- Possible extension of Fan and Li (1996) to the time series case using the CLT established in Fan and Li (1999a) under absolutely regulare condition (β-mixing).

Specification Testing for Partially Linear Models

Hypotheses:

$$H_{03} : m(v) = g(v'\alpha_0) \text{ vs } H_{13} : m(x,v) = g(v'\alpha_0) + C_n\Delta_n(x,v). \quad (9)$$

The test statistic can be derived from Xia, Tong and Li (1999) results:

$$L_{3n}(x) = \sum_{s=1}^{n}\sum_{t=1}^{n} \tilde{\epsilon}_s K_h\left((X_t - X_s)'\hat{\alpha}\right)\tilde{\epsilon}_t,$$

where $\tilde{\epsilon}_t = y_t - \hat{g}(V_t'\hat{\alpha})$.

Endogeneity in Non and Semiparametric Models

Endogeneity in Partially Linear (PL) type of semiparametrics:

$$Y_t = X_t'\beta + g(V_t) + \epsilon_t,$$

where $E(\epsilon|x, v) \neq 0$.

Endogeneity in Single-index (SI) type of semiparametrics:

$$Y_t = g(V_t'\alpha_0) + \epsilon_t,$$

where $E(\epsilon|v) \neq 0$.

The endogeneity issues:

- Identification problem.
- Alternatives to address endogeneity.

Semiparametric Models with Nonstationary data

Gao and Hawthorne (2006) consider the trend stationary case:

$$Y_t = X_t'\beta + g\left(\frac{t}{n}\right) + \epsilon_t \tag{10}$$

Gao and Hawthorne (2006) also consider the nonstationary case,

where ϵ_t is $I(1)$:

$$\delta_{y,t} = \delta_{x,t}'\beta + m\left(\frac{t}{n}\right) + \delta_t,$$

where $\delta_{y,t} = Y_t - Y_{t-1}$, $\delta_{x,t} = X_t - X_{t-1}$, $m(t/n) = g(t/n) - g((t-1)/n)$ and $\delta_t = \epsilon_t - \epsilon_{t-1}$.

Semiparametric Models with Nonstationary data

Chen et al. (2013):

$$Y_t = \beta(U_t, \theta_1) + g(U_t) + \varepsilon_t, \tag{11}$$

where

$$U_t = H\left(\frac{t}{n}\right) + u_t, \tag{12}$$

where $H(t)$ is unknown functions defined on $\mathbb{R}^d$ and $\{u_t\}$ is a sequence of i.i.d. random errors.

Endogeneity and Nonstationary

Gao and Phillips (2013):

$$\begin{aligned} Y_t &= AX_t + g(V_t) + e_t; && (13)\\ X_t &= H(V_t) + U_t \quad t = 1, 2, \ldots, n; \\ E[e_t|V_t] &= E[e_t] = 0; \text{ and} && (14)\\ E[U_t|V_t] &= 0 && (15) \end{aligned}$$

Some Selected Papers

Kim and Saart (2013):

Endogeneity in both parametric and nonparametric

Saart, Kim and Reale (2013):

Generated Regressors + Asymp Optimality + Stationary Time Series

Saart and Gao (2012):

Generated Regressors + Estimation Algorithm + Application to High Frequency Finance

Saart and Gao (2013):

Nonparametric Hypothesis Testing + Generated Variable + Bandwidth Selection

Some Selected Papers

Kim, Saart and Gao (2013):

EGSI Model + Endogeneity Problem + Shape Invariant Analysis +

Application to Demand Analysis

Jiang, Xia and Saart (201x):

Common Factor of Functional Data

Kim, Pohlmeier and Saart (201x):

New Issues on Regression Models with Weak Instruments

Saart (201x):

Analysis of Financial Events and their Interaction

Summary

- The most well-known semiparametric models are:
  - Partially Linear (PL) model:

$$Y_t = X_t'\beta + g(V_t) + \epsilon_t.$$

  - Additive semiparametric model:

$$Y_t = X_t'\beta + \sum_{l=1}^{q} g_l(V_{lt}) + \epsilon_t.$$

  - Single-index (SI) model:

$$Y_t = g(V_t'\alpha_0) + \epsilon_t.$$

- Semiparametric specification tests are also available.
- Endogeneity in semiparametric models.
- Nonstationarity in semiparametric models.

Semiparametric time series models

- Estimation procedures of GPLSI and EGPLSI models:
  1. Given α, partially-out $g(v'\alpha)$ from the structural relations.
  2. Perform least square estimation on the reduced form to estimate β.
  3. Given $\hat{\beta}$, obtain the minising objective function to estimate α_0.
  4. Given $\hat{\beta}$ and $\hat{\alpha}$, recover the nonparametric function $\hat{g}(v'\hat{\alpha})$.

Mixing Conditions

- An absolutely regular mixing condition (β-mixing):

$$\beta(t) = \sup E[\sup\{|P(B|A) - P(B)| : A \in \Omega_1^s, B \in \Omega_{s+t}^\infty\}] \leq C_\beta \beta',$$

for all $s, t \geq 1$, where $0 < C_\beta < \infty$ and $0 < \beta' < 1$ are constants and Ω_i^j denotes the σ-field generated by $\{X_k : i \leq k \leq j\}$.

- A strong mixing regular condition (α-mixing):

$$\alpha(t) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \Omega_1^s, B \in \Omega_{s+t}^\infty\} \leq C_\alpha \alpha',$$

for all $s, t \geq 1$, where $0 < C_\alpha < \infty$ and $0 < \alpha' < 1$ are constants and Ω_i^j denotes the σ-field generated by $\{X_k : i \leq k \leq j\}$.