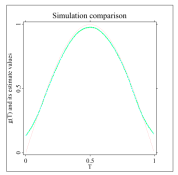
Partially Linear Models with Heteroskedastic Variance

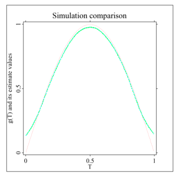
Wolfgang Härdle

Hua Liang



Outline

- Partially linear models (PLM)
- PLMHV
- Estimators
- Asymptotic Normality
- Remarks
- Simulation
- Example
- Conclusions

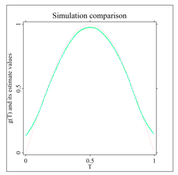


Partially Linear Models – PLM

- Engle, et al. (1986), JASA
- Many papers study **PLM** in theory and their applications in Economics, Biostatistics and so on.

In theory and applications

- Härdle, Mammen & Müller (1998). East–West German migration
- Liang, Härdle & Werwarz (1999). Income and age in west Germany.
- Schmalensee & Stoker (1999). Household gasoline consumption in the United States.

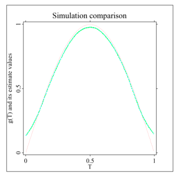


Partially Linear Models – PLM

In theory

- Speckman, P. (1988), JRSSB.
- Green & Silverman, (1991), Book
- Cuzick, J. (1992), JRSSB.
- Robinson, P. (1988), Econometrica.
- Liang, H. & Härdle, W. (2001).
- Hamilton & Troung (1997), *Journal of Multivariate Analysis* .

Härdle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*. *Springer Physica-Verlag*.

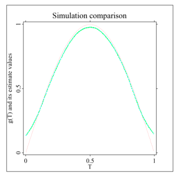


Nonparametric Smoothing

$$\begin{aligned} Y &= X^\top \beta + g(T) + \varepsilon \\ E(Y|T) &= \{E(X|T)\}^\top \beta + g(T) \\ Y - E(Y|T) &= \{X - E(X|T)\}^\top \beta + \varepsilon \end{aligned}$$

- $\widehat{E}(Y|T) = \sum_{i=1}^{n} \omega_{ni}(T) Y_i$
- $\widehat{E}(X|T) = \sum_{i=1}^{n} \omega_{ni}(T) X_i$
- "LS" estimator of β: regression of $Y - \widehat{E}(Y|T)$ on $X - \widehat{E}(X|T)$.

Of course, one may estimate nonparametric function by Smoothing Spline, Piecewise Polynomial, Local Linear, and even Wavelet.



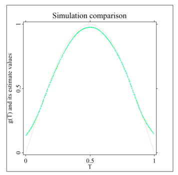
Nonparametric Smoothing

$$\beta_{LS} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{Y}}$$

$$\begin{aligned} \widetilde{\mathbf{X}}^\top &= (\widetilde{X}_1, \ldots, \widetilde{X}_n), \quad \widetilde{X}_i = X_i - \sum_{j=1}^{n} \omega_{nj}(T_i) X_j \\ \widetilde{\mathbf{Y}} &= (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^\top, \quad \widetilde{Y}_i = Y_i - \sum_{j=1}^{n} \omega_{nj}(T_i) Y_j \end{aligned}$$

$$n^{1/2}(\beta_{LS} - \beta) \xrightarrow{\mathcal{L}} N(0, B^{-1} C B^{-1})$$

$B = \text{cov}\{X - E(X|T)\}$ and $C = \text{cov}\big[\varepsilon * \{X - E(X|T)\}\big]$.

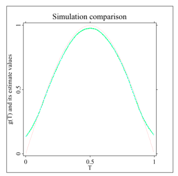


PLMHV

$$Y_i = X_i^\top \beta + g(T_i) + \sigma_i e_i, i = 1, \ldots, n,$$

with

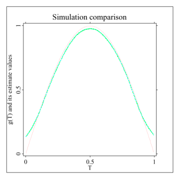
- **Case 1**. $\sigma_i^2 = H(W_i)$, where $\{W_i; i = 1, \ldots, n\}$ are also design points, which are assumed to be independent of e_i and (X_i, T_i) and defined on $[0, 1]$.
- **Case 2**. $\sigma_i^2 = H(T_i)$, i.e., the variance σ_i^2 is a function of the design points T_i.
- **Case 3**. $\sigma_i^2 = H\{X_i^\top \beta + g(T_i)\}$.



Estimators

$$\beta_{nW} = \Big(\sum_{i=1}^{n} \widehat{\gamma}_i \widetilde{X}_i \widetilde{X}_i^{\top}\Big)^{-1} \Big(\sum_{i=1}^{k_n} \widehat{\gamma}_i^{(2)} \widetilde{X}_i \widetilde{Y}_i + \sum_{i=k_n+1}^{n} \widehat{\gamma}_i^{(1)} \widetilde{X}_i \widetilde{Y}_i\Big)$$

- k_n: the integer part of $n/2$
- $\widehat{\gamma}_i^{(1)}$: estimator of $1/\sigma_i^2$ based on $(X_1, T_1, Y_1), \ldots, (X_{k_n}, T_{k_n}, Y_{k_n})$
- $\widehat{\gamma}_i^{(2)}$: estimator of $1/\sigma_i^2$ based on $(X_{k_n+1}, T_{k_n+1}, Y_{k_n+1}), \ldots, (X_n, T_n, Y_n)$,



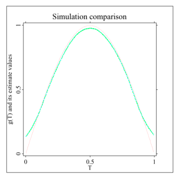
One of Main Results

- β_{nW} is an asymptotically normal estimator of β with asymptotic distribution

$$N\left\{0, \left(E\left[\frac{1}{\sigma_i^2}\text{cov}\{X - E(X|T)\}\right]\right)^{-1}\right\}$$

- Estimators of the nonparametric component $g(\cdot)$

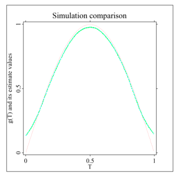
$$\widehat{g}_{nW}(t) = \sum_{i=1}^{n} \omega_{ni}^*(t)(\widetilde{Y}_i - \widetilde{X}_i^T \beta_{nW}),$$



One of Main Results

- A consistent estimator for asymptotic variance by a standard nonparametric regression as follows.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\widehat{\gamma}_i}\left\{X_i-\sum_{j=1}^{n}\omega_{nj}(T_i)X_j\right\}\left\{X_i-\sum_{j=1}^{n}\omega_{nj}(T_i)X_j\right\}^{\top}$$

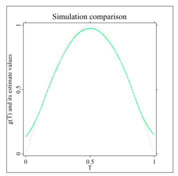


Remarks

- The efficiency bound (Chamberlain, 1992) for partially linear models:

$$E\left(\frac{1}{\sigma_i^2}X_iX_i^T\middle|T_i\right)-E\left\{E\left(\frac{1}{\sigma_i^2}X_i\middle|T_i\right)E\left(\frac{1}{\sigma_i^2}X_i\middle|T_i\right)^{\top}E^{-1}\left(\frac{1}{\sigma_i^2}\middle|T_i\right)\right\}$$

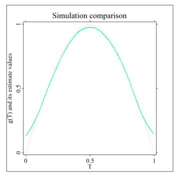
- The covariance of our estimators is identical to the bound of Chamberlain (1992) if σ_i^2 does not depend on T_i.
- For general structure, our estimators do not arrive this bound and new estimators are need



Simulation with XploRe

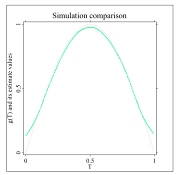
$$Y_i = X_i^\top \beta + g(T_i) + \sigma_i \varepsilon_i, \qquad i = 1, \ldots, n = 300$$

- $\{\varepsilon_i\}$: $N(0,1)$
- $\{X_i\}$ and $\{T_i\}$: $\sim$uniform$[0,1]$
- $\beta = (1, 0.75)^\top$
- $g(t) = \sin(t)$
- Run 500 situations
- Quartic kernel $(15/16)(1-u^2)^2 I(|u| \leq 1)$
- Cross-Validation criterion to select bandwidth



Simulation with XploRe

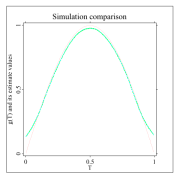
- Model 1: $\sigma_i^2 = T_i^2$
- Model 2: $\sigma_i^2 = W_i^3$ where W_i iid $\sim$ Uniform[0,1].
- Model 3: $\sigma_i^2 = a_1 \exp[a_2\{X_i^\top \beta + g(T_i)\}^2], (a_1, a_2) = (1/4, 1/3200)$.



Simulation with XploRe

Table 1: Simulation results ($\times 10^{-3}$)

	Model	$\beta_0 = 1$		$\beta_1 = 0.75$	
		Bias	MSE	Bias	MSE
β_{LS}	1	8.696	8.7291	23.401	9.1567
β_{nW}	1	4.230	2.2592	1.93	2.0011
β_{LS}	2	12.882	7.2312	5.595	8.4213
β_{nW}	2	5.676	1.9235	0.357	1.3241
β_{LS}	3	5.9	4.351	18.83	8.521
β_{nW}	3	1.87	1.762	3.94	2.642



Simulation with XploRe

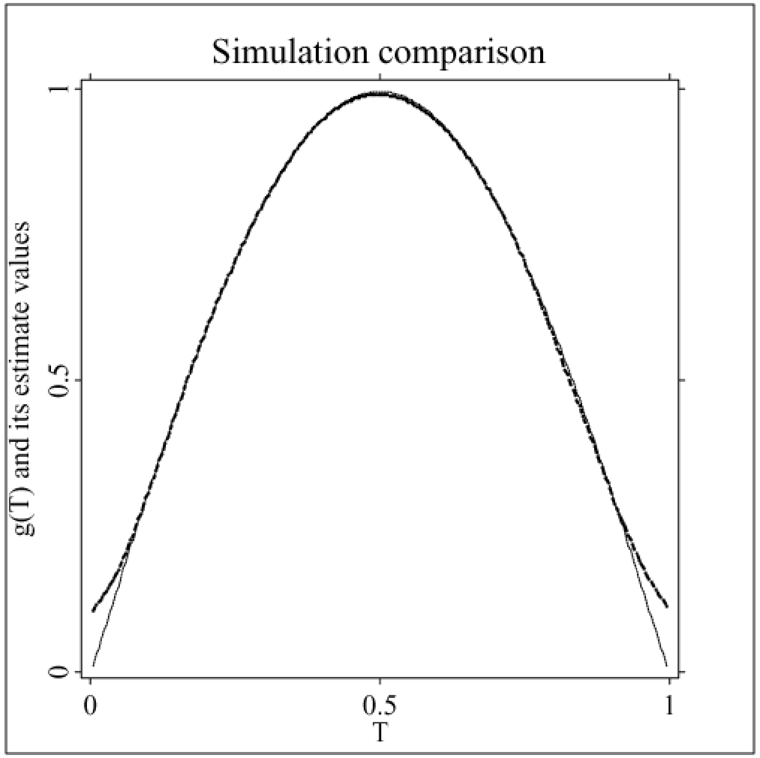
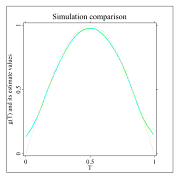


Figure 1: *Estimates of the function $g(T)$ for the first model. Solid-lines stand for true values and dished-lines for our estimate values.*



Simulation with XploRe

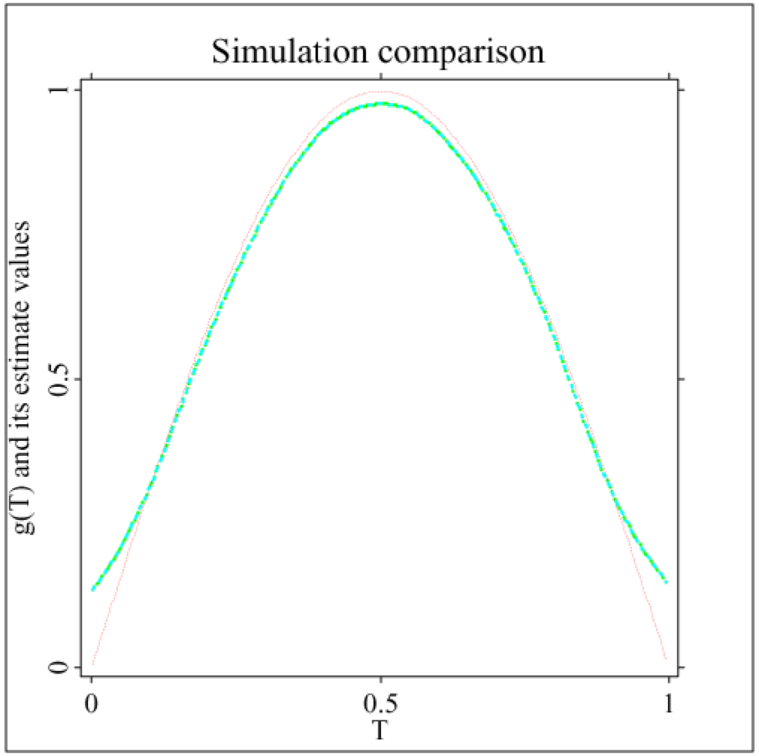
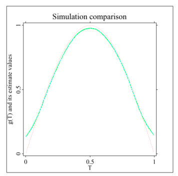


Figure 2: *Estimates of the function $g(T)$ for the second model*



Simulation with XploRe

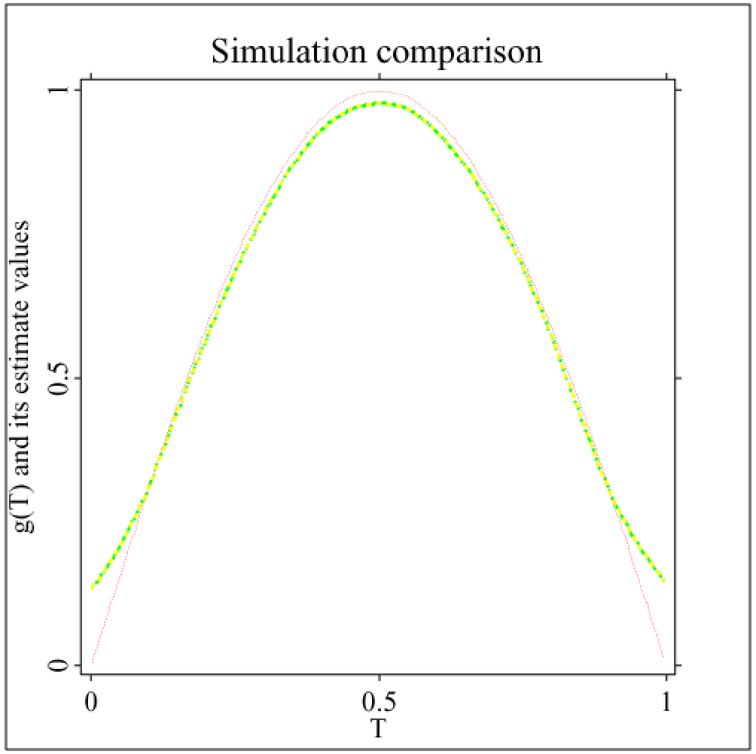
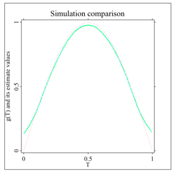
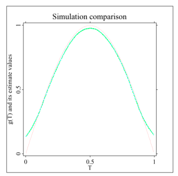


Figure 3: *Estimates of the function $g(T)$ for the third model*



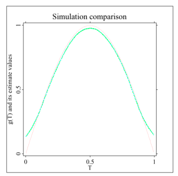
National Health and Nutrition Survey I: epidemiologic follow-up study in USA (NHANES)

- Data: $3,145$ women aged 25-50 and interviewed about their nutrition habits and when later examined for evidence of cancer.
- Y: saturated fat
- T: age
- X: body mass index (BMI), protein and vitamin A and B intakes



National Health and Nutrition Survey I: epidemiologic follow-up study in USA (NHANES)

- Y depends nonlinearly on age but linear upon other dummy variables.
- σ_i^2 is a function of age (case 2)
- XploRe was used
- $\beta_{nW} = (-0.162, 0.317, -0.00002, -0.0047)^\top$
- The pattern reaches to the summit at about age 37.



National Health and Nutrition Survey I: epidemiologic follow-up study in USA (NHANES)

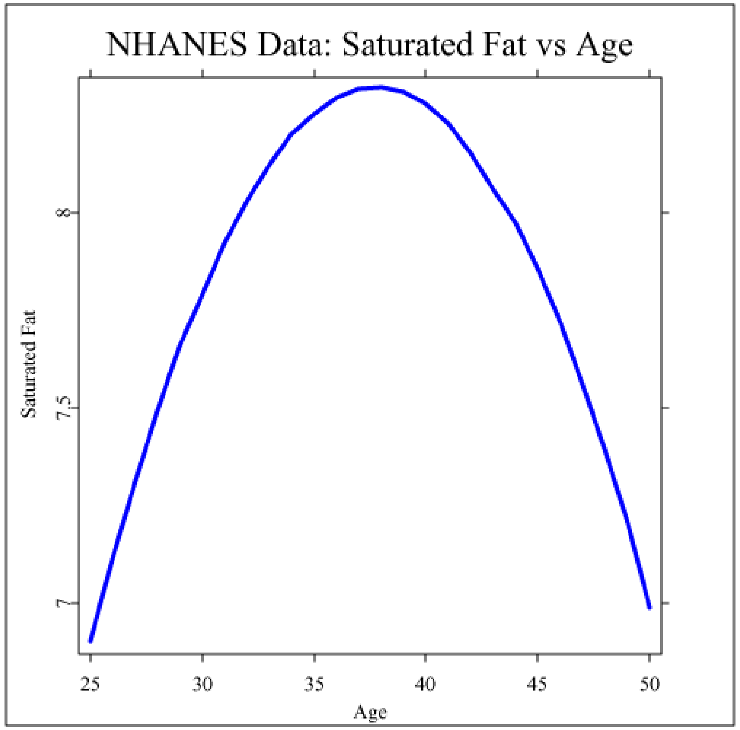
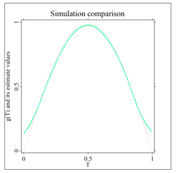


Figure 4: NHANES regression of saturated fat on age



Conclusion

- Partially linear models with heteroskedastic variances has been considered;
- Three classes of variance functions and corresponding estimators have been proposed;
- More efficient estimator β_{nW} has been constructed;
- Several simulations have been carried to illustrate our estimators;
- A real data set has been studied;
- Future work: More general variance function???