UPPER EXPECTATION PARAMETRIC REGRESSION

Lu Lin[1], Ping Dong[1], Yunquan Song[2] and Lixing Zhu[3]

[1]*Shandong University,* [2]*China University of Petroleum*

and [3]*Hong Kong Baptist University*

Supplementary Material: Numerical Studies and Proof

S1 Numerical studies

S1.1 Simulation Studies

In this subsection we examine the finite sample behaviors of the newly proposed estimators by simulation studies. To obtain thorough comparisons, in addition to the new PMLS estimator, we comprehensively consider the OLS estimators that ignore the distribution randomness in the sense before. Mean squared error (MSE), prediction error (PE) and boxplots are used to evaluate the performances of the involved estimators and models. Also the simulation results for estimation bias are reported to emphasize the influence from the distribution randomness, especially from expectation uncertainty. In the following, we design 4 experiments. The first experiment is to compare the PMLS with the overall average $\overline{Y}$ for estimating the upper expectation of Y, the second and third experiments are designed for examining the performances of PMLS and OLS when estimating the parameter β and $\overline{\mu}$ in simple linear and multiple linear models. The fourth experiment is used to investigate the usefulness of PMLS for prediction.

Experiment 1. Consider the simplest case with $\beta = 0$:

$$Y_i = \varepsilon_i, \quad i = 1, \cdots, N,$$

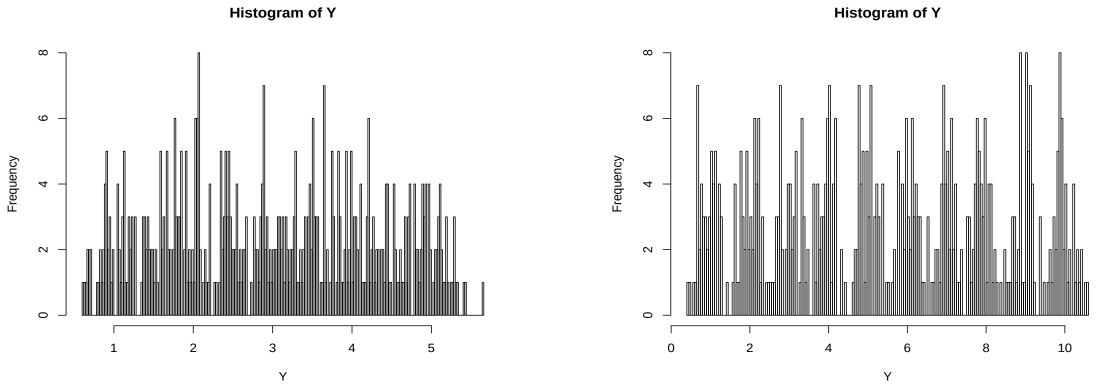
where $\varepsilon_i, i = 1, \cdots, N$, are independent and follow the distributions in the class $\mathscr{F} = \{N(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathcal{T}\}$. In the following, we respectively consider two cases of the distribution randomness:

Case 1. $\mathcal{T} = \{k/2 : k = 1, \cdots, 10\} \times \{0.20^2, 0.25^2\}$ and $T = (\mu, \sigma^2)$ is uniformly distributed on $\mathcal{T}$.

Case 2. $\mathcal{T} = \{k : k = 1, \cdots, 10\} \times \{0.25^2\}$ and $T = (\mu, \sigma^2)$ is uniformly distributed on $\mathcal{T}$.

For each k, the size of the sample from $N(\mu_k, \sigma_k^2)$ is designed as $[N/10]$.

Before performing the simulation, we first use the histograms of Y_i in the two cases to observe what pattern of the data appears to show the distribution randomness. It is very clear from Figure 1 that the distributions in the two cases look like multimodal although every distribution is unimodal. It shows that when we have a data set showing multimodal pattern, we may not simply believe the multimodality of an underlying distribution, the distribution randomness would also be a possibility. Under this situation, the classical statistical inferences such as the estimation of population expectation, have less accuracy. Instead, our goal is to consistently estimate the upper expectation $\overline{\mu} = \mathbb{E}[\varepsilon] = \mathbb{E}[Y]$.



(a) Case 1

(b) Case 2

Figure 1: Histograms for Experiment 1 with $N = 500$.

To examine the consequence of ignoring the distribution randomness in estimation, we compare the PMLS estimator with the OLS estimator that is the overall average $\overline{Y}$ of all of observations in this experiment. For the total sample sizes $N = 100, 500$ and 1000, the empirical bias and MSE, and the boxplots of the estimators with 500 replications are reported respectively in Table 1, and Figures 2 and 3. Note that for this very simple model, we cannot have a constant intercept term because of the distribution randomness. Therefore, theoretically, the intercept term for every observation is not identifiable, which is absorbed in the error term in the upper expectation of error term. The OLS estimator estimates nothing as its limit is in between, from the description in Section 1, the upper and lower expectation: $\underline{\mu} \leq \bar{Y} \leq \overline{\mu}$ with a probability going to one.

Table 1: Estimation bias and MSE for cases 1 and 2 in Experiment 1

	N	100		500		1000	
case	criterions	Bias	MSE	Bias	MSE	Bias	MSE
1	PMLS	-0.0842	0.0319	0.0640	0.0113	0.1210	0.0187
	OLS	-2.2517	5.0707	-2.2501	5.0631	-2.2493	5.0595
2	PMLS	-0.0408	0.0157	0.0562	0.0062	0.0920	0.0108
	OLS	-4.4980	20.2330	-4.4999	20.2497	-4.5000	20.2509

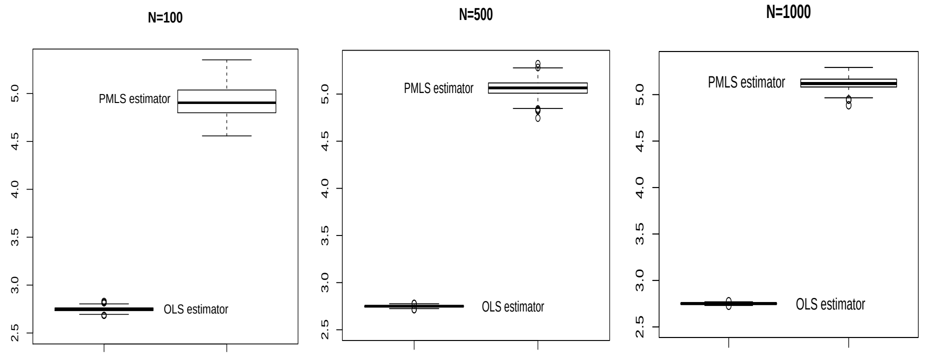


Figure 2: The boxplots of the PMLS estimator and the OLS estimator in case 1 with the true $\overline{\mu} = 5$ in Experiment 1.

The simulation results can verify that our PMLS estimator is clearly superior to the OLS estimator. More precisely, we have the following findings:

(1) From Table 1, the distribution randomness mainly results in the estimation bias of the OLS estimator, and the estimation bias almost obliterates the effect of variance in the MSE of the estimator. However, this distribution randomness has no significant impact for the PMLS estimator for the upper expectation

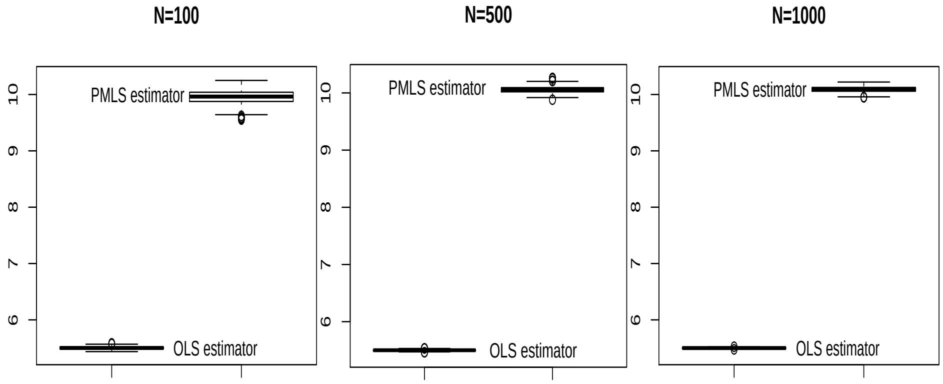


Figure 3: The boxplots of the PMLS estimator and the OLS estimator in case 2 with the true $\overline{\mu} = 10$ in Experiment 1.

$\overline{\mu}$. The estimation bias of the PMLS estimator are very obviously smaller than those of the OLS estimator in both cases. The centerlines of the boxplots of the PMLS estimator in Figures 2 and 3 are just located respectively at the true values 5 and 10 of the upper expectations. But the centerlines of the boxplots of the OLS estimator are far below the true values.

(2) From Figures 2 and 3, we can see that although the boxplots of the PMLS estimator are nearly centralized around the centerlines, the values have more dispersion than those of the OLS estimator, implying the new estimator has larger variance and a slow convergence rate. It is because the new method only uses a part of the data. However, this enlarged variance is negligible compared with the significant estimation bias from which the OLS estimator suffers.

Experiment 2. Consider the following univariate linear regression:

$$Y_i = \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, N,$$

where $X_i, i = 1, \cdots, N$, are independent and identically distributed as $N(1,1)$. Suppose that $\varepsilon_i, i = 1, \cdots, N$, are independent and follow the distributions in the class $\mathscr{F} = \left\{N(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathcal{T}\right\}$ with $\mathcal{T} = \{k : k = 1, \cdots, 10\} \times \{(0.05k)^2 : k = 1, \cdots, 10\}$. As we commented in Section 2, the model cannot contain a nonzero constant intercept term because even an intercept term, β_0, is imposed, it is impossible to be identified and consistently estimated. In fact, the intercept is absorbed in $\overline{\mu}$. Hence, we do not report the simulation result for β_0. The histograms of Y_i and the residuals $\hat{\varepsilon}_i$ derived from the OLS are multimodal to present the distribution randomness (the histograms are not reported herewith for saving space).

In the simulation, we set $\beta_1 = 2$, and let (μ, σ^2) be uniformly distributed on $\mathcal{T}$. For each k, the size of the sample from $N(\mu_k, \sigma_k^2)$ is designed as $[N/10]$. For total sample sizes $N = 100, 500$ and 1000, the empirical bias and MSE, and the boxplots of the estimators over 500 replications are reported respectively in Table 2, and Figures 4 and 5. Although the model used here is totally different from that in Experiment 1, a conclusion from the simulation results is similar to the finding (1) obtained in Experiment 1. That is to say, PMLS can accurately estimate the regression coefficient and the upper expectation of the error, while OLS gets the estimators that are far away from the true values. Unlike that in the finding (2) in Experiment 1, the variance of the OLS estimator is larger than that of the PMLS estimator in this experiment. The PMLS estimator of β_1 performs better than the PMLS estimator of $\overline{\mu}$ with smaller bias and MSE particularly when the sample size is large. Perhaps it is because the two-step estimation procedure for $\overline{\mu}$ introduces more estimation error.

Table 2: Estimation bias and MSE in Experiment 2

parameters	N	100		500		1000	
	criterions	Bias	MSE	Bias	MSE	Bias	MSE
β_1	PMLS	-0.07497	0.03101	-0.01854	0.00341	-0.00449	0.00063
	OLS	2.75808	7.66282	2.74548	7.55041	2.74865	7.56090
$\overline{\mu}$	PMLS	0.00636	0.13711	0.16631	0.05203	0.26512	0.08386

We note that the OLS estimator of β_1 has a significant bias. One may expect to centralize data to reduce the bias. As we explained before, for every observation, the center $E_{f_{t_i}}(\varepsilon_i)$ is a conditional expectation when $T = t_i$ is given, and such a center is actually a random variable because of the distribution randomness defined in Section 2. Thus, in theory, using the overall average of Y_i's as the center of every Y_i is not meaningful and it is also not estimable. On the other hand, in practice, when we use it as if the distribution randomness did not exist, its practical performance can be promoted because ε_i is not centered. We now pretend that the observations do not have the distribution randomness. If only Y_i's in the above regression are centered, it can be easily verified

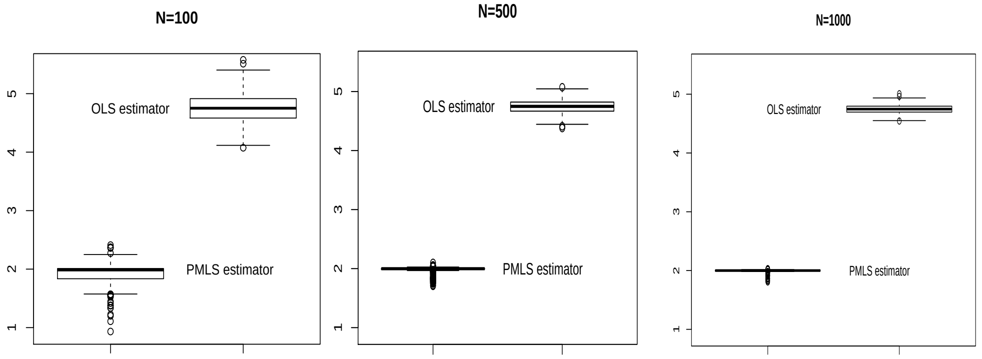


Figure 4: The boxplots of the PMLS estimators and the OLS estimators for β_1 in Experiment 2 with the true $\beta_1 = 2$.

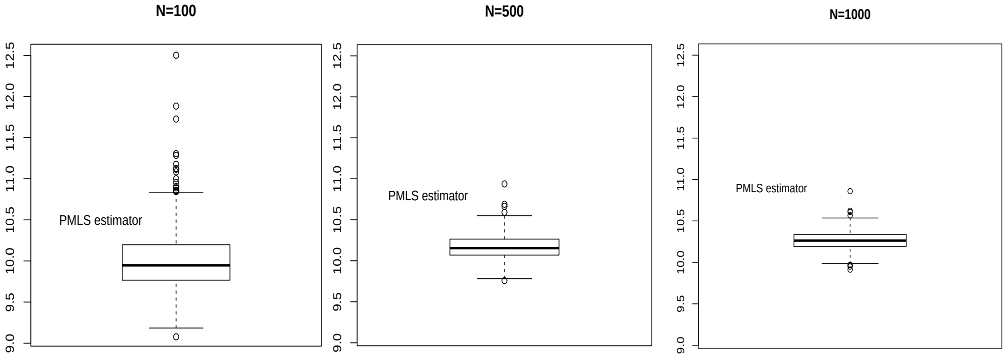


Figure 5: The boxplots of the PMLS estimators for $\overline{\mu}$ in Experiment 2 with the true $\overline{\mu} = 10$.

that, in this example, $\widehat{\beta}_{1LS} = \frac{1}{2}\beta_1 + O_p\left(\frac{1}{\sqrt{N}}\right)$, which is also biased. If both X_i' and Y_i' are centered, it can be seen that $\widehat{\beta}_{1LS} = \beta_1 + O_p\left(\frac{1}{\sqrt{N}}\right)$. The simulation result in Table 3 shows that when we blindly use OLS with centered data, the estimation efficiency does be promoted. However, even for the latter, the estimation bias is slightly larger than that of the PMLS estimator, and the MSE of the centered LS estimator is about 5 times of that of the the PMLS estimator although in the case without the distribution randomness, the bias-reduced LS estimator should have a variance achieving Fisher information bound.

Experiment 3. Consider the multiple linear regression:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1, \cdots, N,$$

Table 3: Estimation bias and MSE for bias-reduced LS estimator in Experiment 2

$N = 500$	only Y_i's centralized		both X_i' and Y_i's centralized	
	Bias	MSE	Bias	MSE
$\widehat{\beta}_{1LS}$	-1.00689	1.02000	-0.01953	0.01569

where $\beta_1 = 3$, $\beta_2 = 2$, $X_{1i} \sim N(1, 1)$ and $X_{2i} \sim N(2, 1)$, the other settings are designed as those in Experiment 2. The simulation results are reported in Table 4 and Figures 6-8 and further indicate that the PMLS estimator is consistently superior to the OLS estimator in estimation bias, MSE and variance. Again we can see that OLS overestimates regression coefficients because the distribution randomness makes it impossible to remove the bias which is absorbed in the error terms. The effect of the overestimated regression coefficients by OLS is to compensate the loss of ignoring the positive error. Then, a new problem emerges naturally: *Is OLS able to give a proper prediction?* We discuss this issue in the following experiment.

Table 4: Estimation bias and MSE in Experiment 3

parameters	N	100		500		1000	
	criterion	Bias	MSE	Bias	MSE	Bias	MSE
β_1	PMLS	0.05841	0.03646	0.01916	0.00362	0.01247	0.00169
	OLS	0.90110	0.92334	0.91106	0.85260	0.91335	0.84461
β_2	PMLS	0.12348	0.05200	0.03135	0.00495	0.02113	0.00282
	OLS	1.85004	3.45390	1.83649	3.38013	1.83498	3.37050
$\overline{\mu}$	PMLS	-0.33971	0.34478	0.05134	0.04601	0.18741	0.06053

Experiment 4. The model and experiment conditions are completely identical to those in Experiment 3, but the purpose is to examine the prediction behavior. Before comparing the predictions derived by OLS and PMLS, we give a definition of prediction under the situation with the distribution randomness. Because the classical methods ignore the distribution randomness, a natural prediction of Y based on OLS that is *blindly* used is given as

$$\widehat{Y}_{LS} = \widehat{\beta}_{LS}^{\top} X_0$$

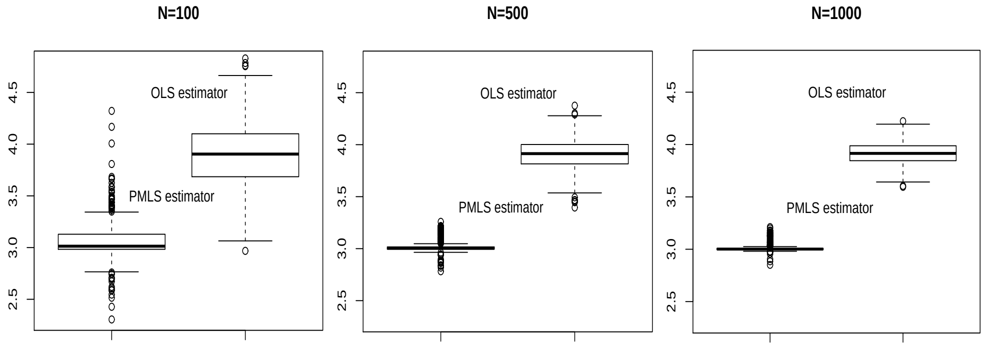


Figure 6: The boxplots of the PMLS estimators for β_1 in Experiment 3 with the true $\beta_1 = 3$.

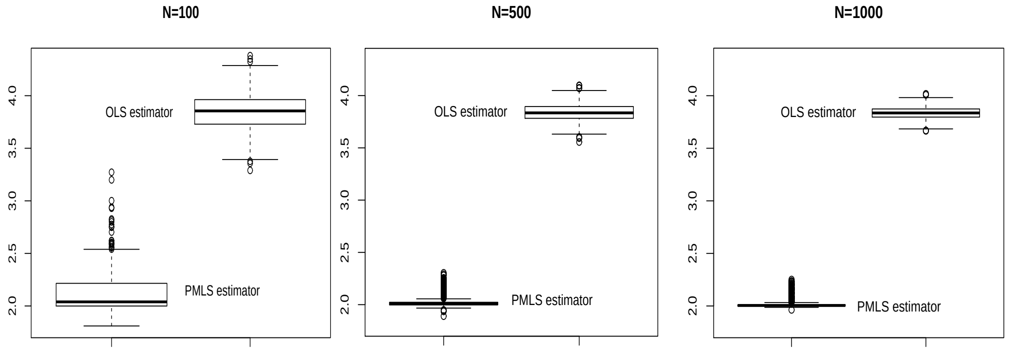


Figure 7: The boxplots of the PMLS estimators for β_2 in Experiment 3 with the true $\beta_2 = 2$.

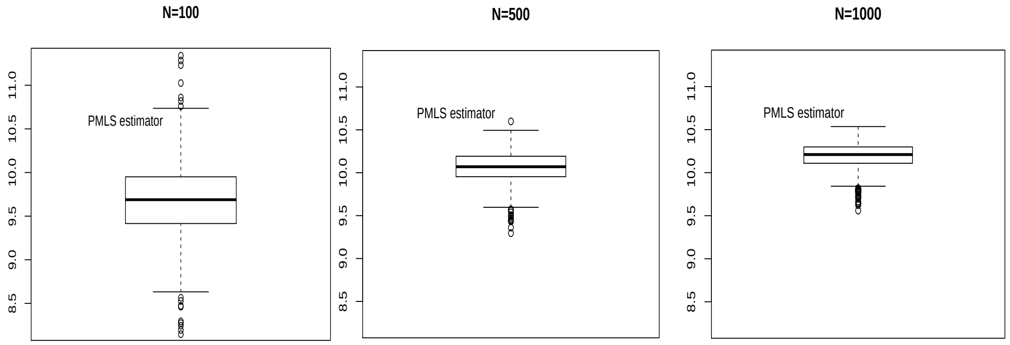


Figure 8: The boxplots of the PMLS estimators for $\overline{\mu}$ in Experiment 3 with the true $\overline{\mu} = 10$.

for a given predictor X_0. In contrast, the main goal of the upper expectation regression is to predict maximum values of Y conditional on predictor X_0. Thus, under the framework of upper expectation regression, the prediction is defined by

$$\widehat{Y}_M = \widehat{\beta}^\top X_0 + \widehat{\overline{\mu}},$$

where both $\widehat{\beta}$ and $\widehat{\overline{\mu}}$ are the MPLS estimators proposed in the previous sections. On the other hand, if our goal is to predict all the values Y_i, not merely the maximum values, based on PMLS, a reasonable prediction is defined as

$$\widehat{Y}_{PMLS} = \widehat{\beta}^\top X_0 + \mu_M,$$

where μ_M is a suitable value in the interval $[\widehat{\underline{\mu}}, \widehat{\overline{\mu}}]$. Here the estimator $\widehat{\underline{\mu}}$ of the lower expectation of ε can be obtained by the similar argument proposed in the previous sections. If without additional information about expectation uncertainty of ε, we simply choose the middle point $\mu_M = (\widehat{\underline{\mu}} + \widehat{\overline{\mu}})/2$. It is worth pointing out that although the overall average of $\overline{Y} - \widehat{\beta}^\top \overline{X}$ can also be between $\widehat{\underline{\mu}}$ and $\widehat{\overline{\mu}}$, it does not converge to a fixed value under the distribution randomness and thus, its use makes no theoretical ground.

We first consider the performances of the predictions for some larger values of Y. For n values $Y_1, \cdots, Y_n$, we rearrange them in descending order as $Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(n)}$. In this case, average prediction error (APE_m) of predicting the first m largest values of Y is defined by

$$APE_m = \frac{1}{m}\sum_{i=1}^{m}(Y_{(i)} - \widehat{Y}_{(i)})^2,$$

where $\widehat{Y}_{(i)}$ is a prediction value of $Y_{(i)}$. When $m = n$, APE_m is actually the standard average prediction error; when m is small, APE_m mainly evaluates the behaviors of prediction for the large values of Y. The simulation results are presented in Figure 9, in which the curves are the medians of APE_ms of 500 replications. It clearly shows that the upper expectation regression can relatively accurately predict the larger values of Y. More precisely, for 100 values of Y, the upper expectation regression gives relatively successful prediction for the first 34 largest values of Y. However, if ignoring the distribution randomness, the OLS-based prediction behaves poorly for predicting the larger values of Y.

Finally, we investigate the behaviors of the predictions $\widehat{Y}_{LS}$ and $\widehat{Y}_{PMLS}$ for all the values of Y. The APE_ms of the two predictions are reported in Table 5. It is clear that the APE_m of $\widehat{Y}_{PMLS}$ is significantly smaller than that of $\widehat{Y}_{LS}$. Because of the distribution randomness, however, both the two predictions have relatively large APE_m even for large sample size. It shows that it is impossible to improve the predictions if without further information about the distribution of ε, in other words, we can not completely characterize the regression under the situation with the distribution randomness.

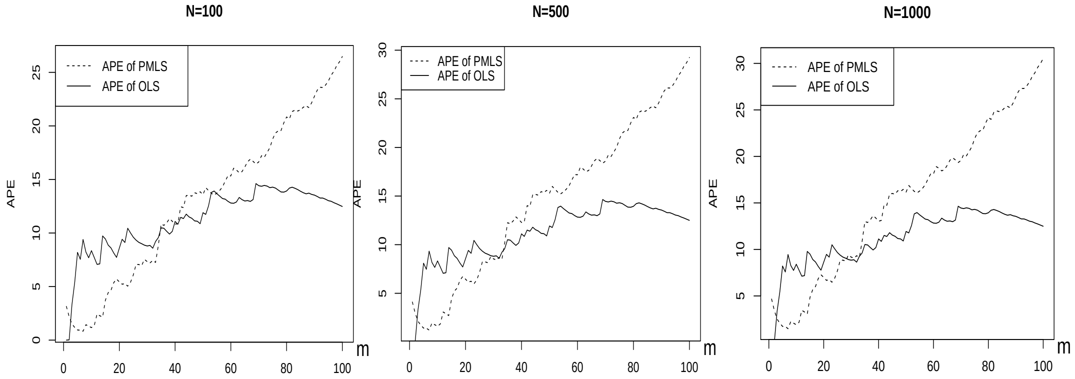


Figure 9: The medians of APE_ms for the first m largest values of Y in Experiment 4.

Table 5: The APE_m of predicting all the values of Y in Experiment 4

N	100	500	1000
PMLS	8.49825	8.48973	8.49585
OLS	12.47704	12.49915	12.49017

S1.2 Real data analysis

In this subsection we use a real data example to show how the upper expectation regression works under the setting with the distribution randomness. Consider the data set of the Fifth National Bank of Springfield based on data from 1995 (see examples 11.3 and 11.4 in Albright et al. (1999)). This data set has been analyzed by such as Fan and Peng (2004). The bank, whose name has since changed, was charged in court with paying its female employees substantially lower salaries than its male employees. For each of its 208 employees, the data set includes the following variables:

- EduLev: education level, a categorical variable with categories 1 (finished high school), 2 (finished some college courses), 3 (obtained a bachelors degree), 4 (took some graduate courses), 5 (obtained a graduate degree).

- JobGrade: a categorical variable indicating the current job level, the possible levels being 1-6 (6 highest).
- YrHired: year that an employee was hired.
- YrBorn: year that an employee was born.
- Gender: a categorical variable with values “Female” and “Male”.
- YrsPrior: number of years of work experience at another bank prior to working at the Fifth National Bank.

- PCJob: a dummy variable with value 1 if the empolyee’s current job is computer related and value 0 otherwise.
- Salary: current (1995) annual salary in thousands of dollars.

Fan and Peng (2004) fitted the data by a linear model as

$$\text{Salary} = \beta_0 + \beta_1\text{Gender} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i + \sum_{i=1}^{5}\beta_{6+i}\text{JobGrd}_i + \varepsilon. \tag{S1.1}$$

Figure 1 in Section 1 presents the histogram, scatter plot and a nonparametric fit of the residuals via the above model, showing that the errors may not be identically distributed. It would be because some other factors, such as the years of working experience and the age of an employee, may affect the salary. Therefore, we regard these potential factors as unobserved factors in an upper expectation regression with the distribution randomness. The model still has a linear regression function as follows:

$$\mathbb{E}(\text{Salary}) = \beta_1\text{Gender} + \beta_2\text{PCJob} + \sum_{i=1}^{4}\beta_{2+i}\text{Edu}_i + \sum_{i+1}^{5}\beta_{6+i}\text{JobGrd}_i + \overline{\mu}. \tag{S1.2}$$

It is worth pointing out that the differences from the model (S1.1) are that the error ε in (S1.2) is supposed to be of the distribution randomness, and the model (S1.2) does not have the intercept term, which is absorbed into the upper expectation of ε. Moreover, if the model actually has distribution randomness, the upper expectation $\overline{\mu}$ is not the same as the intercept β_0 in the OLS fitting of (S1.1). In fact, the so-called “intercept”, say β_0, is not identifiable, and the OLS fitting, it is determined by minimizing least squares. Thus, the “intercept” has on a unique true value and is different from $\overline{\mu}$ because $\overline{\mu}$ is the largest one among all the different “intercepts”.

We use 170 data to estimate the model parameters and then use the obtained models to fit the rest of the data (to predict 38 values of "Salary"). The predictions and prediction errors are defined in Experiment 4. The results of parameter estimation and the APE_m are listed in Table 6. Compared with the OLS regression (S1.1), the upper expectation regression (S1.2) has the following interesting features:

(1) The absolute values of the estimators of the coefficients of JobGrd$_i$ are significantly reduced, but the others, especially the coefficients of Gender and Edu$_2$, are largened. The phenomenon could be explained as follows. As JobGrd$_i$ may be related to the years of working experience and the age of the employee, when these factors are not included in model (S1.1), the model requires larger coefficients of JobGrd$_i$ to draw the information of these factors. On the other hand, the effect of JobGrd$_i$ is absorbed into the error term of model (S1.2).

(2) The difference of the APE_ms between the two models is not significant.

Table 6: Parameter estimation and APE_m for real data

parameters	OLS	PMLS
β_1 (Gender)	-1.314	-3.115
β_2(PCJob)	4.532	5.082
β_3(Edu$_1$)	1.523	1.969
β_4(Edu$_2$)	0.086	0.474
β_5(Edu$_3$)	-0.335	0.387
β_6(Edu$_4$)	-1.439	-1.692
β_7(JobGrd$_1$)	-36.191	-34.143
β_8(JobGrd$_2$)	-34.304	-31.574
β_9(JobGrd$_3$)	-29.392	-27.046
β_{10}(JobGrd$_4$)	-24.341	-22.201
β_{10}(JobGrd$_5$)	-17.579	-17.210
$\overline{\mu}$	–	68.832
$\underline{\mu}$	–	63.628
β_0 (Intercept)	68.814	–
APE_m	32.615	32.544

On the other hand, as shown in Experiment 4, the upper expectation regression is more concerned about the maximum information. Figure 10(a) presents the medians of the APE_ms for the m largest values of “Salary” via 100 replications. From this figure, we can get the following finding:

(3) The upper expectation regression can relatively accurately predict the larger values of “Salary”. For example, for the first 15 largest values of “Salary”, the APE_m of the upper expectation regression is 10.6892; and for the first 24 largest values of of “Salary”, the APE_m of the upper expectation regression is 6.5338. Both are smaller than the corresponding values that are based on the OLS regression.

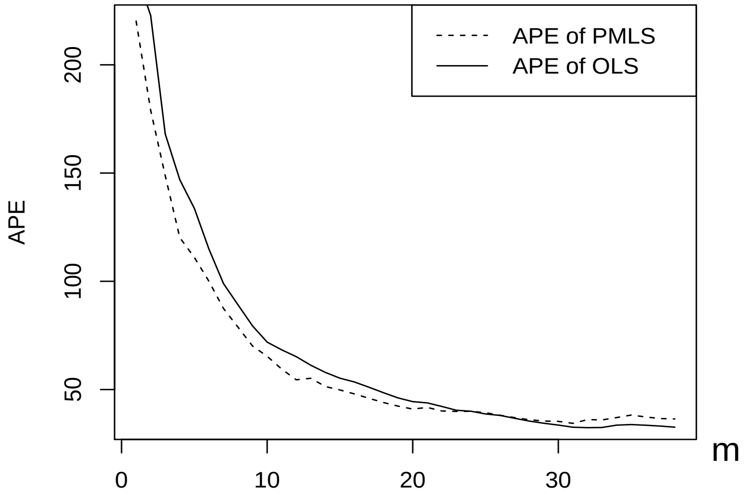
Finally, we examine the R^2 values of the two models, which is defined by

$$R_m^2 = 1 - \frac{\sum_{j=1}^m (Y_{(j)} - \widehat{Y}_{(j)})^2}{\sum_{j=1}^m (Y_{(j)} - \overline{Y}_m)^2} \quad \text{for } m \leq N,$$

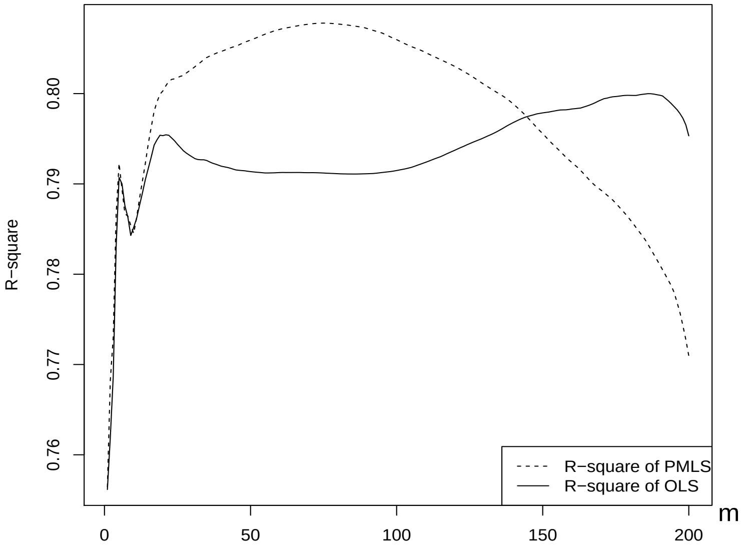
where $\overline{Y}_m = \sum_{j=1}^m Y_{(j)}/m$ with $Y_{(j)}$ given in Experiment 4. When $m = N$, the value of R_m^2 defined here is equivalent to the standard R^2 value. As stated above, for the case of distribution randomness, we mainly focus the relatively large values of Y_i, implying that the relatively small values of m are used to evaluate the behaviors of the fitting. That is to say that we use the values of R_m^2 of the first m largest values of “Salary” to check if the upper expectation regression can capture the information on relatively large values. The result is reported in Figure10(b). It suggests the following conclusion:

(4) For the two models, most values of R^2 are larger than 0.79, while the upper expectation regression has a relatively high R^2 for the larger values of “Salary”.

All the numerical results aforementioned are coincident with the theoretical conclusions.



(a) Medians of APE_ms.



(b) Values of R^2.

Figure 10: Figures in real data analysis.

S2 Proof of Lemma 1

It can be easily proved that

$$\widehat{\beta}_{LS} - \beta = E^{-1}[XX^T]\frac{1}{n}\sum_{j=1}^{n} E[X_j\mu_j] + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\widehat{\beta}_{LS}$ is the least squares estimator of β. The result leads to

$$\frac{1}{n}\sum_{j=1}^{n} E[X_j\mu_j] = E[XX^T](\widehat{\beta}_{LS} - \beta) + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Note that

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{\mu})^2 &= \frac{1}{n}\sum_{j=1}^{n}(Y_j - X_j^T\beta - \overline{\mu} + X_j^T\beta)^2 \\ &= C + \frac{1}{n}\sum_{j=1}^{n} E[G_j(\beta, \overline{\mu})] + \frac{2}{n}\sum_{j=1}^{n} E[\mu_j X_j^T]\beta + O_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where $C = \beta^T E[XX^T]\beta - 2\overline{\mu}E[X^T]\beta$. Then, we have

$$\frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{\mu})^2 = C + \frac{1}{n}\sum_{j=1}^{n} E[G_j(\beta, \overline{\mu})] + 2(\widehat{\beta}_{LS} - \beta)^T E[XX^T]\beta + O_p\left(\frac{1}{\sqrt{n}}\right).$$

The relation above leads to the conclusion of the lemma. □

S3 Proof of Theorem 1

It can be see from Lemma 1 and (S3.4) given below that for the asymptotic property of parameter estimation that for the asymptotic property of parameter estimation, the objective functions in (3.3) and (3.4) respectively have the following equivalent forms:

$$\frac{1}{2}\gamma' V\gamma + U_{n_\tau}\gamma \ \text{ and } \ \frac{1}{2}\gamma' V\gamma + U_{n_\lambda}\gamma + r_{n_\lambda}(\gamma),$$

where γ is a parameter vector, V is a positive definite matrix, U_n is stochastically bounded and $r_n(\gamma)$ goes to zero in probability for each γ. Thus, by the basic corollary of Hjørt and Pollard (unpublished report, 1993), we have that the objective functions in (3.3) and (3.4) are equivalent for parameter estimation with respect to asymptotic property. We thus only investigate the asymptotic properties of the estimator defined by (3.3).

For simplicity, here we only consider the case when n is a even number: $n = 2m$. As shown in Section 3, $\frac{1}{n}\sum_{j=1}^{n} G_{(j)}(\beta, \overline{\mu})$ is a decreasing function of n. On the other hand, the first condition in *C4* implies $\lambda\Delta_n = o(1)$

if $n \le n_0$, and the condition in *C0* shows Δ_n will be increasing when n exceeds n_0. These lead to that the selected n should satisfy $n \ge n_0$. Suppose without loss of generality that, for $n \ge n_0$, only the last d_n elements $G_{(n-d_n+1)}(\beta, \overline{\mu}), \cdots, G_{(n)}(\beta, \overline{\mu})$ in the set $\mathscr{G}_n$ do not come from f_*, with $n-d_n$ being an even number: $n-d_n = 2l$. Then, using the original indices, we have

$$\begin{aligned}\Delta_n &= \frac{1}{n}\left(\sum_{j=1}^{l} E[G_{k_j}(\beta,\overline{\mu})] - \sum_{j=1}^{l} E[G_{k_j}(\beta,\overline{\mu})]\right) \\ &\quad + \frac{2}{n}\sum_{j=1}^{m-l} E[G_{k_{l+j}}(\beta,\overline{\mu})] - \frac{1}{n}\sum_{j=1}^{d_n} E[G_{k_{2l+j}}(\beta,\overline{\mu})] \\ &=: \frac{1}{n}I_1 + \frac{1}{n}I_2 - \frac{1}{n}I_3.\end{aligned}$$

It can be seen that

$$I_1 = 0, \ I_2 = d_n E_{f_*}(Y - \beta^T X - \overline{\mu})^2.$$

By the treatments above, we have

$$\lambda\Delta_n = \frac{1}{n^{1-\epsilon}}I_2 - \frac{1}{n^{1-\epsilon}}I_3.$$

By the above results and the conditions in *C4*, it is clear that if $n = O(n_0)$, then $\lambda\Delta_n = o(1)$. If n/n_0 is diverging, then $(n_* + d_n)/n_0 \to \infty$, implying $d_n/n_0 \to \infty$. Note that $n^{1-\epsilon}/n_0$ is supposed to be bounded and

$$\frac{d_n}{n_0} = \frac{d_n}{n^{1-\epsilon}}\frac{n^{1-\epsilon}}{n_0}.$$

It implies $\frac{d_n}{n^{1-\epsilon}} \to \infty$. Moreover,

$$\frac{1}{n^{1-\epsilon}}I_2 - \frac{1}{n^{1-\epsilon}}I_3 \ge \frac{d_n}{n^{1-\epsilon}}\left[E_{f_*}(Y - \beta^T X - \overline{\mu})^2 - \max_{j=1,\cdots,d_n} E[G_{k_{l+j}}(\beta,\overline{\mu})]\right],$$

in which

$$E_{f_*}(Y - \beta^T X - \overline{\mu})^2 > \max_{j=1,\cdots,d_n} E[G_{k_{l+j}}(\beta,\overline{\mu})].$$

Thus

$$\frac{1}{n^{1-\epsilon}}I_2 - \frac{1}{n^{1-\epsilon}}I_3 \to \infty.$$

In this case, $\lambda\Delta_n$ is diverging as well and, consequently, the minimum value of the objective function $\frac{1}{n}\sum\limits_{j=1}^{n} G_{(j)}(\beta,\overline{\mu}) + \lambda\Delta_n$ does not exist. We then need only to consider the objective function $\frac{1}{n}\sum_{j=1}^{n} G_{(j)}(\beta,\overline{\mu}) + \lambda|\Delta_n|$ with $n = O(n_0)$ for the asymptotic properties of the estimation.

Furthermore, because $n = O(n_0)$, the objective function can be further expressed as

$$\frac{1}{n}\sum_{j=1}^{n-d_n} G_{(j)}(\beta,\overline{\mu}) + o_p(1) = \frac{1}{n}\sum_{j=1}^{n_*} G_{(j)}(\beta,\overline{\mu}) + o_p(1). \tag{S3.1}$$

Denoted by β^0 and μ_*^0 the true values of β and μ_*, respectively, and let β and $\overline{\mu}$ satisfy $\|\beta-\beta^0\| = O(1/\sqrt{n})$ and $|\overline{\mu}-\mu_*^0| = O(1/\sqrt{n})$. Because $n = O(n_0)$, we can assume $n_*/n \to 1$, without loss of generality. Then, the objective function (S3.1) can be replaced by

$$\begin{aligned}
&\frac{1}{n_*}\sum_{j=1}^{n_*} G_{(j)}(\beta,\overline{\mu}) + o_p(1)\\
&= \frac{1}{n_*}\sum_{j=1}^{n_*}\left(\varepsilon_{k_j} + \beta^{0\prime}X_{k_j} - \beta' X_{k_j} - \overline{\mu}\right)^2 + o_p(1)\\
&= \frac{1}{n_*}\sum_{j=1}^{n_*}\big\{(\varepsilon_{k_j}-\mu_*^0)^2 - 2[(\beta-\beta^0)'X_{k_j} + (\overline{\mu}-\mu_*^0)](\varepsilon_{k_j}-\mu_*^0)\\
&\qquad\qquad + [(\beta-\beta^0)'X_{k_j} + (\overline{\mu}-\mu_*^0)]^2\big\} + o_p(1)
\end{aligned} \tag{S3.2}$$

Because $\sum_{j=1}^{n_*}(\varepsilon_{k_j}-\mu_*^0)^2$ is free of β and $\overline{\mu}$, the objective function in (S3.2) is equivalent to

$$\begin{aligned}
&\frac{1}{n_*}\sum_{j=1}^{n_*}\big\{-2[(\beta-\beta^0)'X_{k_j} + (\overline{\mu}-\mu_*^0)](\varepsilon_{k_j}-\mu_*^0)\\
&\qquad\qquad + [(\beta-\beta^0)'X_{k_j} + (\overline{\mu}-\mu_*^0)]^2\big\} + o_p(1).
\end{aligned} \tag{S3.3}$$

By the basic corollary of Hjørt and Pollard (unpublished report, 1993), the term of order $o_p(1)$ can be ignored for the asymptotic property of the estimation. We then rewrite the above objective function as

$$Z_n(\gamma) = \sum_{j=1}^{n_*}\left\{\frac{-2}{\sqrt{n_*}}[\varepsilon_{k_j}-\mu_*^0][X_{k_j}', 1]\gamma + \frac{1}{n_*}\gamma'\Phi(X_{k_j})\gamma\right\}. \tag{S3.4}$$

The objective function $Z_n(\gamma)$ is obviously convex and is minimized at

$$\Gamma_n = \sqrt{n_*}[(\widehat{\beta}-\beta^0)', \widehat{\overline{\mu}}-\mu_*^0]'.$$

Note that $\varepsilon_{k_j}, j = 1,\cdots,n_*$, are identically distributed with the common mean μ_*^0 by the condition *C2*. It follows from the Lindeberg-Feller central limit theorem that

$$Z_n(\gamma) \xrightarrow{d} Z_0(\gamma) = -2W'\gamma + \gamma' E[\Phi(X)]\gamma,$$

where $W \sim N(0, u_*^2 E[\Phi(X)])$. The convexity of the limiting objective function, $Z_0(\gamma)$, assures the uniqueness of the minimizer and consequently, that

$$\sqrt{n_*}\left[(\widehat{\beta}-\beta^0)', \widehat{\overline{\mu}}-\mu_*^0\right]' = \hat{\gamma}_n = \arg\min \widetilde{Z}_n(\gamma) \xrightarrow{d} \hat{\gamma}_0 = \arg\min Z_0(\gamma).$$

(See, e.g., Pollard (1991), Hjørt and Pollard (unpublished report, 1993), Knight (1989)). Finally, we see $\hat{\gamma}_0 = E^{-1}[\Phi(X)]W$ and $(n_*)/n \to 1$. Then the result follows. □

S4 Proof of Theorem 2

By the same argument as used in the proof of Theorem 1, $n_{\widetilde{\tau}}$ can be replaced by the sample size $\widetilde{n}$. Let $\{H^0_{(j)} = Y_{s_j} - \beta^T X_{s_j} : j = 1, \cdots, N\}$ be the order statistic of $\{H^0_j = Y_j - \beta^T X_j : j = 1, \cdots, N\}$, satisfying $H^0_{(1)} \geq H^0_{(2)} \geq \cdots \geq H^0_{(n)}$. Write the corresponding index decomposition as $C^0_n = U^0_n \cup L^0_n$ and let

$$\Gamma^0_{n_{\widetilde{\lambda}}} = \frac{1}{[n_{\widetilde{\lambda}}/2]} \sum_{j \in U^0_n} H^0_j - \frac{1}{\widetilde{n} - [n_{\widetilde{\lambda}}/2]} \sum_{j \in L^0_n} H^0_j.$$

It follows from Theorem 1 that $\Gamma_{n_{\widetilde{\lambda}}} = \Gamma^0_{n_{\widetilde{\lambda}}} + O_p(1/\sqrt{n_{\widetilde{\lambda}}})$. Denoted by β^0 and $\overline{\mu}^0$ the true values of β and $\overline{\mu}$, respectively, and let β and $\overline{\mu}$ satisfy $\|\beta - \beta^0\| = O(1/\sqrt{n_{\widetilde{\lambda}}})$ and $|\overline{\mu} - \overline{\mu}^0| = O(1/\sqrt{n_{\widetilde{\lambda}}})$. Thus, the objective function in (3.6) can be expressed as

$$\frac{1}{n_{\widetilde{\lambda}}} \sum_{j=1}^{n_{\widetilde{\lambda}}} \Big([\varepsilon_{s_j} - \overline{\mu}^0] - (\beta^0 - \widehat{\beta})^T X_{s_j} - [\overline{\mu} - \overline{\mu}^0] \Big)^2 - \widetilde{\lambda} \Gamma_{n_{\widetilde{\lambda}}} + O_p(1/\sqrt{n_{\widetilde{\lambda}}}).$$

Note that $\{\varepsilon_{k_j}, j = 1, \cdots, n_*\}$ and $\{\varepsilon_{s_j}, j = 1, \cdots, n_{\widetilde{\lambda}}\}$ are independent, and $\widehat{\beta}$ depends only on $\{\varepsilon_{k_j}, j = 1, \cdots, n_*\}$. By the conclusion of Theorem 1 and the same argument as used in the proof of Theorem 1, we can prove the theorem. □

S5 Proof of Theorems 3-5

The proofs are similar to that of Theorem 1. □

Bibliography

Albright, S. C., Winston, W. L. and Zappe, C. J. (1999). *Data Analysis and Decision Making with Microsoft Excel*. Duxbury, Pacific Grove, CA.

Chen, Z. and Epstein, L. (2002). Ambiguity, Risk and Asset Returns in Continuous Time. *Econometrica*, **70**(4), 1403-1443.

Chung, K. L. (2006). *Elementary Probability Theory.* Springer; 4th ed. edition.

Clyde, M. and George, E. I. (2004). Model Uncertainty. *Statistical Science*, **19** (1), 81-94.

Fan, J. and Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *Ann. Statist.* **32**, 928-961.

Gilboa, I. and Schmeidler, D. (1989). Maxmin Expected Utility with Non-unique Prior. *Journal of Mathematical Economics*, **18**(2), 141-153.

Härdle, W., Liang, H. and Gao, J. (2000). *Partial linear Models.* Springer-Verlag, New York.

Hjørt, N. and D. Pollard (1993). Asymptotics for minimizers of convex processes. *unpublished Statistical Research Report.*

Huber, P. J. (1981). *Robust Statistics*, John Wiley and Sons.

Knight, P. H. (1921). *Risk, Uncertainty and Profit.* Sentry Press, Kelly, Bookseller.

Knight, K. (1989). Limit Theory for Autoregressive-Parameter Estimates in an Infinite-Variance Random Walk. *Canadian Journal of Statistics*, **17**, 261-278.

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

Lam, C. and Fan, J. (2008). Profile-kernel Likelihood Inference with Diverging Number of Parameters. *Ann. Statist.* **36**, 2232-2260.

Lin, L., Zhu, Lixing and Gai, Y. (2016). Inference for Biased Models: a Quasi-instrumental Variable approach. *Journal of Multivariate Analysis*, **145**, 22-36.

Peng, S. (1997). Backward SDE and Related g-expectations. *Pitman Research Notes in Mathematics Series*, **364**, 141-159.

Peng, S. (2007). G-Expectation, G-Brownian Motion and Related Stochastic Calculus of Itôs type. *Stochastic Analysis and Applications*, 541-567.

Peng, S. (2008). Multi-dimensional G-Brownian Motion and Related Stochastic Calculus Under G-expectation. *Stochastic Processes and Their Applications*, **118**(12), 2223-2253.

Peng, S. (2009). Survey on Normal Distributions, Central Limit Theorem, Brownian Motion and the Related Stochastic Calculus under Sublinear Expectations. *Science in China Series*, **52**, 7, 1391-1411.

Pollard, D. (1991). Asymptotics for least absolute deviation regression Estimators. *Econometric Theory*, **7**, 186-199.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* **6**, 461-464.

Zhang, C. M. (2008). Prediction Error Estimation under Bregman Divergence for Non-parametric Regression and Classification. *Scand. J. Statist.* **35**, 496-523.