Statistica Sinica **27** (2017), 1265-1280
doi:https://doi.org/10.5705/ss.202016.0243

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*

Abstract: In regression analysis, some predictors might be unobservable, not observed, or ignored. These factors actually affect the response randomly. The observed data thus follows a conditional distribution when these factors are given. This phenomenon is called the distribution randomness. For such a working model, we propose an upper expectation regression and a two-step penalized maximum least squares procedure to estimate parameters in the mean function and the upper expectation of the error. The resulting estimators are consistent and asymptotically normal under certain conditions. Simulation studies and a data example are used to show that the classical least squares estimation fails but the proposed estimation performs well.

Key words and phrases: Distribution randomness, penalized least squares, upper expectation.

1. Introduction

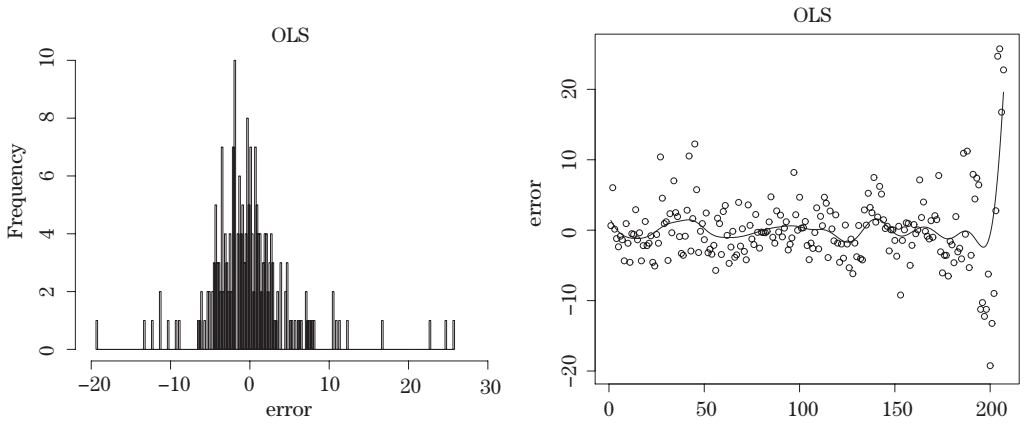
In classical regression modelling, collected data are often assumed to contain a response and all relevant predictors. Here we consider that some predictors are unobservable, unobserved, or ignored so that the working model can be sufficiently parsimonious. In high-dimensional paradigms, this is typically the case because a selected working model is parsimonious and unlikely to (or simply cannot) include all the predictors.

Suppose then that a random sample $\{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ is available in a regression setting, but there is a relevant predictor T_i that is unobservable, unobserved, or ignored. We call it the unobserved factor. Let

$$Y_i = g(\beta, X_i) + \varepsilon_i \quad \text{for } i = 1, \cdots, N, \tag{1.1}$$

where $g(\cdot, \cdot)$ is a given function, $X_i = (X_i^{(1)}, \cdots, X_i^{(p)})^T$ are the associated p-dimensional predictors with a probability density $f_X(\cdot)$. The parameters of interest are β and those associated with the distribution of ε, here $\varepsilon_i(T_i)$.

Since the *unobserved factor* T_i can affect the response randomly, ε_i has a conditional on T_i. We call this distribution randomness. A relevant example is in



(a) Histogram of residuals. (b) Model fitting of residuals.

Figure 1. Figure for OLS fitting.

Huber (1981) where the model is $Y_i = \varepsilon_i$, and the data may contain gross errors. Often one only uses a fraction of the predictors to build a parsimonious working model, ignoring the impact of unobserved or ignored factors on the response. As an example we consider, Fan and Peng (2004) where a linear regression model was used, together with OLS, to fit a data set of the Fifth National Bank of Springfield (see also examples 11.3 and 11.4 in Albright, Winston and Zappe (1999)). Linear regression of annual salary on four predictors was considered: job level, education level, gender and an indicator of a computer-related job. Figure 1 presents the histogram of residuals and a kernel-based fitting for the residuals. The histogram presents large dispersion of the residuals, indicating a poor fit. Nonlinear models did not show much improvement. Part (b) of Figure 1 suggests that the residuals do not follow the same distribution, and perhaps years of experience and age have some impact on the salary. Unobserved/ignored factors are not included in the linear regression function, but in fact are absorbed into the error term.

When a classical regression model is fitted, and unobserved factors affect the distributions of the responses, we have difficulty defining a common expectation of the errors as the intercept of the regression model. In simulations in the supplement, we show that an upper expectation regression model works better even under the linear regression model of Fan and Peng (2004).

In a high-dimensional paradigm, variable selection is often required and many variables are absorbed into the error term. When a model is not sparse and variable selection is implemented, the error term may not be centered and consistent

estimation of the parameter of interest cannot be easily obtained. A relevant discussion by Lin, Zhu and Gai (2016) considers a nonsparse model and a semiparametric method to achieve estimation consistency. Here we define an upper expectation that avoids distribution randomness and allows estimation of involved parameters.

The notion of upper expectation is not new, see Huber (1981). For $Y_i = \varepsilon_i$, he proposed upper and lower expectations for data that contain gross errors. With $\mathscr{F} = \{f_t : t \in \mathcal{T}\}$ a class of distributions, where $\mathcal{T}$ is an index set, Huber (1981) defined upper and lower expectations, respectively, as $\overline{\mu} = \sup_{f_t \in \mathscr{F}} E_{f_t}[\varepsilon_t]$ and $\underline{\mu} = \inf_{f_t \in \mathscr{F}} E_{f_t}[\varepsilon_t]$. We discuss a more general model: there are covariates in a regression model, and the distribution of each error is randomly selected from a class of distributions, say $\mathscr{F}$. Thus $f \in \mathscr{F}$ can be regarded as a "conditional" distribution when the unobserved affecting factor $T = t \in \mathcal{T}$ is given, where $\mathcal{T}$ is a set of values of unobserved factors. Observations can then be written as Z_{t_i} when $T = t_i$ is given, and Z_{t_i} follows a conditional distribution $F(\cdot|T = t_i)$.

Consider the case where every distribution f_i of ε_i belongs to $\mathscr{F}$. The individual expectation $E_{f_i}(Y|X) = g(\beta, X) + E_{f_i}(\varepsilon|X)$ is difficult to estimate by the sample $\{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ because we do not know from which distribution $f_i \in \mathscr{F}$ each ε_i comes.

Under (1.1), the expectations of the ε_i are conditional expectations when the unobserved random variables $T = t_i$ are given, and are not estimable. The upper expectation can be employed. If ε has a distribution f randomly selected from $\mathscr{F}$, then

$$\mathbb{E}[Y|X] = g(\beta, X) + \overline{\mu}, \tag{1.2}$$

where $\overline{\mu}$ is the upper expectation of ε, $\overline{\mu} = \mathbb{E}[\varepsilon] = \sup_{f \in \mathscr{F}} E_f[\varepsilon]$.

In related matters, if Y is a risk measure of a financial product, the upper expectation regression can describe the relationship between the maximum risk and relevant factors in the sense of averaging; see, e.g. Chen and Epstein (2002). Under the framework of Knight uncertainty (Knight (1921)), different observations may come from different distributions randomly selected from a class of distributions and the related economic analysis is based on this uncertainty, refer to Gilboa and Schmeidler (1989).

Here interest is in consistently estimating β and $\overline{\mu}$ using observations from the model (1.1). The definition of upper expectation implies the sub-additivity:

$$\mathbb{E}[U + V] \le \mathbb{E}[U] + \mathbb{E}[V]$$

for any random variables U and V. Consequently, even if $g(\beta, X) \equiv 0$ at (1.1), the

Law of Large Numbers (LLN) under sublinear expectation, Peng (2008, 2009), has the sample mean $\overline{Y}$ of $Y_1, \cdots, Y_n$ satisfying only, with large probability,

$$\underline{\mu} \leq \overline{Y} \leq \overline{\mu},$$

where $\underline{\mu} = \inf_{f \in \mathscr{F}} E_f[Y]$ and $\overline{\mu} = \sup_{f \in \mathscr{F}} E_f[Y]$ are, respectively, the lower and upper expectations. Our problem is then to, under certain conditions, identify those observations that can be used for estimating β and $\overline{\mu}$.

As a first attempt, we consider a finite class $\mathscr{F}$. A penalized maximum least squares (PMLS) is introduced, and a two-step estimation procedure is suggested. The key feature of this method is that, for different parameters β and $\overline{\mu}$ in the model (1.2), it can identify available data for estimation. The resulting estimators are consistent and asymptotically normal in a certain sense. Moreover, the PMLS offers a potentially useful tool in data analysis when we are not sure whether all predictors/factors have been included in a working model and whether an identical distribution assumption is appropriate.

The paper is organized as follows. In Section 2, we consider the random selection of distributions, the upper expectation regression, and the motivation for an estimation procedure. Section 3 contains the methodology development, the asymptotic properties of the estimators, the tuning parameter selection, and a related algorithm. The method is extended in Section 4 to the case where the upper expectation can be attained by several distributions, and the estimator for the upper expectation is constructed. The paper concludes with some discussions in Section 5. Simulation studies, a data example, and the proofs of the theorems are given in the supplementary materials.

2. Motivation and Upper Expectation Regression

Suppose that $\mathscr{F}$ is a distribution class with a factor set $\mathcal{T}$, $\mathscr{F} = \{f(\cdot, t) : t \in \mathcal{T}\}$, and suppose the factor variable T has distribution $p(\cdot)$. Let $Z = Z(T)$ be a random variable such that for any fixed $T = t \in \mathcal{T}$, the distribution of $Z = Z(t)$ is $f_t(z) = f(z(t), t) \in \mathscr{F}$.

Here there exists an unobserved random factor(s) T that has impact on the distribution of the random variable Z. Under the present framework, what we can observe is just $Z_i(T_i)$ in which T_i is unobserved. Therefore, any element $f(\cdot, t)$ within the class $\mathscr{F}$ could be the distribution of $Z(T)$. We write $Z(T)$ as Z. Thus, for a random variable function $g(Z)$, the expectation $E_{f_t}[g(Z)]$ is the conditional expectation with conditional density $f_t = f(Z(t), t)$.

We mainly consider the linear regression, model $Y = \beta^\top X + \varepsilon$, where $\top$

stands for transposition, $\beta = (\beta_1, \cdots, \beta_p)^\top$ is a p-dimensional vector of unknown parameters. The extension of results to the nonlinear model (1.1) is discussed in Section 4. For the model, the error $\varepsilon = \varepsilon(T)$ is of the distribution randomness as $Z(T)$. By (1.2), the upper expectation linear regression is

$$\mathbb{E}[Y|X] = \beta^\top X + \overline{\mu}, \tag{2.1}$$

and the model is supposed to be identifiable, where $\overline{\mu} = \sup_{t\in\mathcal{T}} E_{f_t}[\varepsilon(t)]$. Note that the original model (1.1) has no constant intercept term because there is no need to consider a common constant intercept term that it is not identifiable. In model (2.1), the intercepts with all $\varepsilon(t_i)$ are absorbed in $\overline{\mu}$.

2.1. Motivation for estimating β and $\overline{\mu}$

To estimate β and $\overline{\mu}$, we start with a brief distribution of the estimation at the population level. Consider the upper expectation squared loss

$$\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right], \tag{2.2}$$

and the minimizer of this loss over β. Because $\beta^\top X + \overline{\mu}$ is identifiable and X follows a certain distribution f_X, the true β is the minimizer over all β. Next, for true β the above squared upper expectation loss is equal to

$$\mathbb{E}\left[(\varepsilon - \overline{\mu})^2\right]. \tag{2.3}$$

Suppose that there is a distribution $f_{t^*} \in \mathscr{F}$ such that

$$\mathbb{E}\left[(\varepsilon(T) - \overline{\mu})^2\right] = E_{f_{t^*}}\left[(\varepsilon(t^*) - \overline{\mu})^2\right]. \tag{2.4}$$

Then, by the projection theory, the minimizer of the loss over $\overline{\mu}$ is $E_{f_{t^*}}(\varepsilon)$. Therefore, we need a two-step procedure to estimate β and $\overline{\mu}$ separately. First, use the above criterion to estimate β and $\overline{\mu}$. The estimator $\widehat{\beta}$ of β can be consistent. After $\widehat{\beta}$ being obtained, we re-estimate $\overline{\mu}$ to obtain consistent estimation. For ease in presentation, we suppose $\mathcal{T} = \{1, \cdots, L\}$ for a positive integer L. T then follows a distribution P with unknown probability mass p_t for $t \in \{1, \cdots L\}$.

Recall that $(X_i(t_i), Y_i(t_i)), i = 1, \cdots, N$, are independent observations from the model:

$$Y_i(t_i) = \beta^\top X_i(t_i) + \varepsilon_i(t_i), \quad i = 1, \cdots, N. \tag{2.5}$$

For simplicity, we write $(X_i, Y_i(t_i))$ as (X_i, Y_i). Every $\varepsilon_i = \varepsilon_i(t_i)$ has a distribution $f_{t_i} \in \mathscr{F}$ with the unobserved factor t_i having the distribution P. For given t_is, we have the linear expectations $\mu_i = E_{f_{t_i}}[\varepsilon_i]$ and variances $\sigma_i^2 = E_{f_{t_i}}[(\varepsilon_i - \mu_i)^2]$. We consider the following treatment to get the initial estimates of β and $\overline{\mu}$.

For any given β and $\overline{\mu}$, let $\left\{G_{(j)}(\beta,\overline{\mu}) = (Y_{k_j} - \beta^\top X_{k_j} - \overline{\mu})^2 : j = 1, \cdots, N\right\}$ be the ordered statistics of $\left\{G_i(\beta,\overline{\mu}) = (Y_i - \beta^\top X_i - \overline{\mu})^2 : i = 1, \cdots, N\right\}$ satisfying

$$G_{(1)}(\beta,\overline{\mu}) \geq G_{(2)}(\beta,\overline{\mu}) \geq \cdots \geq G_{(N)}(\beta,\overline{\mu}). \tag{2.6}$$

To construct an empirical version of $\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right]$, one can use larger $G_{(i)}(\beta,\overline{\mu})$'s. The intuition is as follows. Note that $\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right]$ is the upper expectation being achieved at the distribution f_{t^*}. Although t^* is unknown, we can expect that relatively larger quantities should be close to the upper expectation. In particular, it is expected that there exists a positive number $n < N$ such that most of $G_{(j)}(\beta,\overline{\mu}), j = 1, \cdots, n$, come from the distribution f_{t^*}, or have the same expectation $\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right]$.

Based on the above explanation, to construct an empirical version of the squared upper expectation loss $\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right]$, we employ the following partial sum:

$$\frac{1}{n}\sum_{j=1}^{n} G_{(j)}(\beta,\overline{\mu}) \quad \text{for some positive integer } n \leq N. \tag{2.7}$$

An estimate $(\beta_n^\top, \overline{\mu}_n)$ of $(\beta^\top, \overline{\mu})$ is then defined as the minimizer of the partial sum:

$$(\beta_n^\top, \overline{\mu}_n) = \arg\min_{\beta\in\mathscr{B},\overline{\mu}\in\mathscr{U}} \frac{1}{n}\sum_{j=1}^{n} G_{(j)}(\beta,\overline{\mu}), \tag{2.8}$$

where $\mathscr{B}$ and $\mathscr{U}$ are, respectively, the parameter spaces of β and $\overline{\mu}$.

Two main difficulties exist. First, the integer n is unknown in practice. Second, the consistency of $1/n\sum_{j=1}^{n} G_{(j)}(\beta,\overline{\mu})$ to $\mathbb{E}\left[(Y - \beta^\top X - \overline{\mu})^2\right]$ cannot automatically result in the consistency of $\overline{\mu}_n$ to μ. Details of the estimation procedure are given next.

3. Methodology and Theoretical Properties

3.1. First-step estimators of β and $\overline{\mu}$

Assume that the distribution $f_* := f_{t^*}$ exists.

Using $G_{(j)}(\beta,\overline{\mu}) = (Y_{k_j} - \beta^\top X_{k_j} - \overline{\mu})^2$ for $j = 1, \cdots, n$, we decompose the index set $I_n = \{k_j : j = 1, \cdots, n\}$ into two subsets as $U_n = \{u_j : j = 1, \cdots, [n/2]\}$ and $L_n = \{l_s : s = n - [n/2] + 1, \cdots, n\}$ satisfying $u_j > l_s$. That is,

$$I_n = U_n \cup L_n, \text{ where } U_n \cap L_n = \emptyset, \text{ and } u_j > l_s \text{ for any } u_j \in U_n, l_s \in L_n. \tag{3.1}$$

Denote $\Delta_n = 1/[n/2]\sum_{j\in U_n} E[(Y_j - \beta^\top X_j - \overline{\mu})^2] - 1/(n - [n/2])\sum_{j\in L_n} E[(Y_j -$

$\beta^\top X_j - \overline{\mu})^2]$. Since the sums in Δ_n are based on the original indices, instead of the ordered quantities $G_{(j)}(\beta, \overline{\mu})$, it can be showed that if most of $(Y_{k_j} - \beta^\top X_{k_j} - \overline{\mu})^2, j = 1, \cdots, n$, come from the distribution f_*, or have the same expectation, then $|\Delta_n|$ should be sufficiently small. Next, assume

C0. The scatter plots of $(Y_j - \beta^\top X_j - \overline{\mu})^2, j = 1, \cdots, N$, are asymmetric.

Under this condition, $|\Delta_n| \nrightarrow 0$ if most of $(Y_{k_j} - \beta^\top X_{k_j} - \overline{\mu})^2, j = 1, \cdots, n$, do not come from the distribution f_*, or do not have the same expectation. Consequently, we choose a tuning parameter $\tau > 0$ and consider a constraint $|\Delta_{n_\tau}| < \tau$, where n_τ depends on τ. If $|\Delta_n|$ is given, the estimator of $(\beta^\top, \overline{\mu})^\top$ can be defined as

$$\left(\widehat{\beta}^\top, \widehat{\overline{\mu}}\right)^\top = \arg \min_{\beta \in \mathscr{B}, \overline{\mu} \in \mathscr{U}, n_\tau \in \mathscr{N}} \frac{1}{n_\tau} \sum_{j=1}^{n_\tau} G_{(j)}(\beta, \overline{\mu}) \quad \text{s.t.} \quad |\Delta_{n_\tau}| < \tau. \tag{3.2}$$

Because the expectation of $1/n_\tau \sum_{j=1}^{n_\tau} G_{(j)}(\beta, \overline{\mu})$ is a decreasing function of n_τ, the ideal choice of n_τ is $n_\tau = \max\{n : |\Delta_{n_\tau}| < \tau\}$. The relation between τ and n_τ implies that the optimization problem (3.2) only contains a tuning parameter τ. The condition *C4* given below can ensure that the tuning parameter τ is identifiable. By the Lagrange multiplier, the optimization problem (3.2) can be rewritten as

$$\left(\widehat{\beta}^\top, \widehat{\overline{\mu}}\right)^\top = \arg \min_{\beta \in \mathscr{B}, \overline{\mu} \in \mathscr{U}, n_\lambda \in \mathscr{N}} \frac{1}{n_\lambda} \sum_{j=1}^{n_\lambda} G_{(j)}(\beta, \overline{\mu}) + \lambda |\Delta_{n_\lambda}|. \tag{3.3}$$

Here λ is a tuning parameter, and can be thought of as a replacer of τ. Since $1/n_\lambda \sum_{j=1}^{n_\lambda} G_{(j)}(\beta, \overline{\mu})$ is a decreasing function of n_λ, and $|\Delta_{n_\lambda}|$ is not small when the value of n_λ exceeds a certain amount, the above objective function is an approximate convex function of n_λ in a certain region. Also it can be directly verified that the above objective function is a convex function of β and $\overline{\mu}$. As a result, the resulting estimator is a unique global solution of the above optimization problem.

To approximate Δ_n, consider

$$\Upsilon_n(Y, \overline{\mu}) = \frac{1}{[n/2]} \sum_{j \in U_n} (Y_j - \overline{\mu})^2 - \frac{1}{n - [n/2]} \sum_{j \in L_n}^{n} (Y_j - \overline{\mu})^2.$$

Lemma 1. *Assume that the upper expectation $\overline{\mu}$ is free of X, and the variances σ_i^2 of ε_i with distribution f_i exist for all $i = 1, \cdots, N$, then*

$$\Delta_n = \Upsilon_n(Y, \overline{\mu}) + O_p\left(\frac{1}{\sqrt{n}}\right).$$

By the lemma, when Δ_n is replaced by $\Upsilon_n(Y,\overline{\mu})$, the optimization problem (3.3) is asymptotically equivalent to

$$\left(\widehat{\beta}^\top,\widehat{\overline{\mu}}\right)^\top = \arg\min_{\beta\in\mathscr{B},\overline{\mu}\in\mathscr{U},n_\lambda\in\mathscr{N}} \frac{1}{n_\lambda}\sum_{j=1}^{n_\lambda} G_{(j)}(\beta,\overline{\mu}) + \lambda\left|\Upsilon_{n_\lambda}(Y,\overline{\mu})\right|. \tag{3.4}$$

For any given $\overline{\mu}$ and β, a choice of n_λ is $n_\tau = \max\{n : |\Upsilon_n(Y,\overline{\mu})| < \tau\}$. The above estimation method is called the penalized maximum least squares (PMLS). Under G-normal distribution (see Peng (2007)), it is a penalized maximum-maximum likelihood. The penalty used is to control the difference between the second-order moments of the random variables and to identify the available data set.

Denote $\mathscr{G}_n = \{G_{(1)}(\beta,\overline{\mu}),\cdots,G_{(n)}(\beta,\overline{\mu})\}$ and suppose that there are only d_n elements $G_{(j_s)}(\beta,\overline{\mu}), s = 1,\cdots,d_n$, in the set $\mathscr{G}_n$ such that $G_{(j_s)}(\beta,\overline{\mu}), s = 1,\cdots,d_n$, do not come from f_*. Let $\mathscr{G}_{n_0}$ be the smallest set of $\mathscr{G}_n$ that contains all the elements $G_{(j)}(\beta,\overline{\mu})$ from the distribution f_*. To get the asymptotic properties, we introduce the following conditions.

C1. The intercept of regression function in model (2.5) is zero, the upper expectation $\overline{\mu}$ is free of X, $E[XX^\top]$ is a positive definite matrix, and the variances σ_i^2 of ε_i with distributions f_i exist for all i.

C2. The distribution f_* satisfying (2.4) is unique and the size n_* of the sample from f_* tends to infinity as $N\to\infty$.

C3. $\lambda = n^{\epsilon-1}$ for a constant $0<\epsilon<1$.

C4. $d_n/n^{1-\epsilon} = o(1)$ and $n^{1-\epsilon}/n_0 < C$ for a constant $C>0$.

Some remarks on the conditions are in order. The conditions in *C1* are standard. Condition *C2* is based on (2.3) and (2.4). This condition implies the second-order moment constraint: $E_{f_*}[(\varepsilon-\overline{\mu})^2] > E_f[(\varepsilon-\overline{\mu})^2]$ for all $f\neq f_*, f\in\mathscr{F}$. Based on this constraint, we can judge whether the corresponding errors $\varepsilon_{k_j}, j=1,\cdots,n_*$, come from the same distribution f_*. The use of the uniqueness assumption on f_* in *C2* is to get a simple estimation procedure. However, this uniqueness assumption may not be always true. Thus, it will be removed when an adjusted method is introduced in the next section. We need Condition *C3* to constrain the convergence rate at which $\lambda\Delta_n$ tends to zero. Condition *C4* implies that most of $G_{(j)}(\beta,\overline{\mu}), j=1,\cdots,n$, come from the distribution f_*. This also implies that approximately Δ_n has a certain distribution, and as a result, the related tuning parameters τ and λ are identifiable. In fact *C4* gives the range of n when the penalized estimation is used. Although this condition seems to be idealistic, an implementation procedure will be given later.

Denote $\mu_* = E_{f_*}[\varepsilon]$, $\sigma_*^2 = E_{f_*}[(\varepsilon - \overline{\mu})^2]$ and $\Phi(X) = \begin{pmatrix} XX^\top & X \\ X^\top & 1 \end{pmatrix}$.

Theorem 1. *Under the model* (2.5)*, suppose Conditions C1-C4 hold. Then the PMLS estimator defined in* (3.4) *satisfies*

$$\sqrt{n_*}\left[(\widehat{\beta} - \beta)^\top, \widehat{\overline{\mu}} - \mu_*\right]^\top \xrightarrow{d} N\left(0, \sigma_*^2 E^{-1}[\Phi(X)]\right) \quad (n_* \to \infty),$$

where $\xrightarrow{d}$ *stands for convergence in distribution.*

A proof of the theorem is given in the Supplement. The key of the proof is to show that most of the elements in $\mathscr{G}_n$ come from f_* via the penalty in (3.4). It can be seen from the proof that the uniqueness assumption on f_* is unnecessary. In the next section, the assumption can be removed via an additional penalty.

The theorem guarantees that the PMLS estimator $\widehat{\beta}$ is consistent and normally distributed asymptotically. However, the PMLS estimator $\widehat{\overline{\mu}}$ is not always consistent because it tends to μ_*, rather than the true parameter $\overline{\mu}$. On the other hand, compared with the properties of parameter estimation in the case of classical nonlinear regression, here the variance is enlarged and the convergence rate is reduced to $1/\sqrt{n_*}$. This is mainly because of the variability of the error terms, which comes from the distribution randomness.

3.2. Second-step estimator of $\overline{\mu}$

Similar to (2.3) and (2.4), suppose the following holds:

$$\overline{\mu} = \mathbb{E}[\varepsilon] = \sup_{f \in \mathscr{F}} E_f[\varepsilon] = E_{\widetilde{f}}[\varepsilon] \quad \text{for a } \ \widetilde{f} \in \mathscr{F}. \tag{3.5}$$

Let $\{H_{(j)} = Y_{s_j} - \widehat{\beta}^\top X_{s_j} : j = 1, \cdots, N\}$ be the order statistics of $\{H_j = Y_j - \widehat{\beta}^\top X_j : j = 1, \cdots, N\}$ satisfying $H_{(1)} \geq H_{(2)} \geq \cdots \geq H_{(N)}$. Similar to the decomposition in (3.1), the index set $I_n = \{s_j : j = 1, \cdots, n\}$ is decomposed as $I_n = U_n \cup L_n$. Then, by the same argument used in the first-step estimation, the second-step estimator of $\overline{\mu}$ is defined by

$$\widehat{\overline{\mu}}_{Sec} = \arg \min_{\overline{\mu} \in \mathscr{U}, n_{\widetilde{\tau}} \in \mathscr{N}} \frac{1}{n_{\widetilde{\tau}}} \sum_{j=1}^{n_{\widetilde{\tau}}} \left(H_{(j)} - \overline{\mu}\right)^2 \quad \text{s.t. } |\Gamma_{n_{\widetilde{\tau}}}| < \widetilde{\tau},$$

where $\Gamma_n = 1/[n/2] \sum_{j \in U_n} (Y_j - \widehat{\beta}^\top X_j) - 1/(n - [n/2]) \sum_{j \in L_n} (Y_j - \widehat{\beta}^\top X_j)$ and $\widetilde{\tau}$ is a tuning parameter. Equivalently,

$$\widehat{\overline{\mu}}_{Sec} = \arg \min_{\overline{\mu} \in \mathscr{U}, n_{\widetilde{\lambda}} \in \mathscr{N}} \frac{1}{n_{\widetilde{\lambda}}} \sum_{j=1}^{n_{\widetilde{\lambda}}} \left(H_{(j)} - \overline{\mu}\right)^2 + \widetilde{\lambda} |\Gamma_{n_{\widetilde{\lambda}}}|. \tag{3.6}$$

Here the tuning parameter $\widetilde{\lambda} \geq 0$ may be different from that in (3.4), but also satisfies Condition *C3*. The objective function in (3.6) is a convex function of $\overline{\mu}$, and the estimator of (3.6) is a PMLS estimator as well. Comparing with the estimation procedure in (3.4), the data set $\{(X_{s_j}, Y_{s_j}) : j = 1, \cdots, \widetilde{n}\}$ used here should be different from the data set $\{(X_{k_j}, Y_{k_j}) : j = 1, \cdots, n_*\}$ used in (3.4).

Let $\widetilde{n}$ be the size of the sample from $\widetilde{f}$. The following conditions are required to establish the estimation consistency for the second-step estimator of $\overline{\mu}$.

C5. $\widetilde{n} \to \infty$ and $n_*/\widetilde{n} \to c \neq 0$ as $N \to \infty$.

C6. Condition *C4* holds when the notations are replaced by the above accordingly.

Unlike *C2*, here the uniqueness assumption on $\widetilde{f}$ is not required. It is because the penalty for Γ_n in (3.6) ensures that most of $\varepsilon_{s_j}, j = 1, \cdots, \widetilde{n}$, have the common mean $\overline{\mu}$.

Denote $\widetilde{\sigma}^2 = E_{\widetilde{f}}[(\varepsilon - \overline{\mu})^2]$, $c = 1 - E[X^\top](E[XX^\top])^{-1}E[X]$ and

$$\Omega^{-1}(X, \theta) = (E[XX^\top])^{-1} + (E[XX^\top])^{-1}E[X]E[X^\top]\frac{(E[XX^\top])^{-1}}{c}.$$

Theorem 2. *Under the conditions in Theorem 1, Conditions C5 and C6, when* $\widetilde{\lambda}$ *satisfies the same condition of* λ *as given in Condition C3, and* $\{\varepsilon_{k_j}, j = 1, \cdots, n_*\}$ *and* $\{\varepsilon_{s_j}, j = 1, \cdots, \widetilde{n}\}$ *are not overlapped, then the second-step estimator in* (3.6) *satisfies*

$$\sqrt{\widetilde{n}}\left(\widehat{\overline{\mu}}_{Sec} - \overline{\mu}\right) \xrightarrow{d} N\left(0, \widetilde{\sigma}^2 + \sigma_*^2 E[X^\top]E[\Omega^{-1}(X)]E[X]\right) \quad (\widetilde{n} \to \infty).$$

Here the constraint of non-overlapping between $\{\varepsilon_{k_j}, j = 1, \cdots, n_*\}$ and $\{\varepsilon_{s_j}, j = 1, \cdots, \widetilde{n}\}$ is only for the simplicity of proof and representation. The condition can be replaced by $f_* \neq \widetilde{f}$ and can be further reduced to that the number n^o of overlapping elements in these two sets $\{\varepsilon_{s_j}, j = 1, \cdots, \widetilde{n}\}$ and $\{\varepsilon_{k_j}, j = 1, \cdots, n_*\}$ satisfies $n^o/\widetilde{n} = o(1)$. After n_* and $\widetilde{n}$ being determined, the condition can be checked by the methods of testing distributions to be equal; the details are omitted here. By the theorem, the second-step PMLS estimator $\widehat{\overline{\mu}}_{Sec}$ is consistent and normally distributed asymptotically.

3.3. A summary of the algorithm

The above estimation procedures involve two tuning parameters: τ and $\widetilde{\tau}$ or λ and $\widetilde{\lambda}$. We use the tuning parameters τ and $\widetilde{\tau}$ to design the algorithm. The parameters can be chosen by the cross-validation. But as n_τ and $n_{\widetilde{\tau}}$ are the functions of τ and $\widetilde{\tau}$ respectively, the cross-validation algorithm used needs to take

this fact into consideration. If the discrete function $s(n) = 1/n \sum_{j=1}^{n} G_{(j)}(\beta, \overline{\mu})$ is approximated by a continuously differentiable function, and a prior distribution $\pi(\beta, \overline{\mu}, n)$ for $(\beta, \overline{\mu}, n)$ is assumed, then, criterion for the Bayesian cross-validations for τ and $\widetilde{\tau}$ can be defined, respectively, as

$$CV(\tau) + \frac{(p+2)\log n_\tau}{n_\tau} \quad \text{and} \quad CV(\widetilde{\tau}) + \frac{(p+2)\log n_{\widetilde{\tau}}}{n_{\widetilde{\tau}}},$$

where $CV(\cdot)$ is the cross-validation criterion defined by Fan and Li (2001). The above Bayesian cross-validations do not depend on the prior distribution, and they are in fact the large-sample criteria. Combining the above estimation procedure with the cross-validation for tuning parameter selection, the whole algorithm can be summarized into the following steps:

Step 1. Initial estimator of $(\beta, \overline{\mu})$. Let $(\beta^1, \overline{\mu}^1)$ be an initial selection of $(\beta, \overline{\mu})$, and $\{(Y_{k_j} - X_{k_j}^\top \beta^1 - \overline{\mu}^1)^2 : j = 1, \cdots, N\}$ be the order quantities of the original squared quantities $\{(Y_j - X_j^\top \beta^1 - \overline{\mu}^1)^2 : i = 1, \cdots, N\}$ in descending order. For each tuning parameter τ, the full data set $T = \{(X_{k_j}, Y_{k_j}) : i = 1, \cdots, n_\lambda\}$ is divided at random into cross-validation training sets $T - T^\nu$ and test sets T^ν, $\nu = 1, \cdots, 5$, and then the initial estimator $(\widehat{\beta}^{(\nu)}(\tau), \widehat{\overline{\mu}}^{(\nu)}(\tau))$ is obtained by the training set $T - T^\nu$ via the method given in the previous subsection.

Step 2. Selection of τ. Write $G_i^{(\nu)}(\tau) = (Y_i - X_i^\top \widehat{\beta}^{(\nu)}(\tau) - \widehat{\overline{\mu}}^{(\nu)}(\tau))^2$ and let

$$\left\{ G_{(j)}^{(\nu)}(\tau) = \left(Y_{k_j} - X_{k_j}^\top \widehat{\beta}^{(\nu)}(\tau) - \widehat{\overline{\mu}}^{(\nu)}(\tau) \right)^2 : (X_{k_j}, Y_{k_j}) \in T^\nu \right\}$$

be the order statistic of $\{G_i^{(\nu)}(\tau) : (X_{k_j}, Y_{k_j}) \in T^\nu\}$ in descending order. Define a Baysian cross-validation criterion as

$$CV(\tau) = \frac{1}{n_\lambda} \sum_{\nu=1}^{5} \sum_{(X_{k_j}, Y_{k_j}) \in T^\nu, 1 \le j \le n_\lambda} G_{(j)}^{(\nu)}(\tau) + \frac{(p+2)\log n_\tau}{n_\tau}.$$

We then get an estimator $\widehat{\tau}$ by minimizing $CV(\tau)$.

Step 3. Final estimator of β. With the selected estimator $\widehat{\tau}$, we estimate β as the first component $\widehat{\beta}$ of the following vector:

$$\left(\widehat{\beta}^\top, \widehat{\overline{\mu}} \right)^\top = \arg \min_{\beta \in \mathscr{B}, \overline{\mu} \in \mathscr{U}} \frac{1}{\widehat{n}_\tau} \sum_{j=1}^{\widehat{n}_\tau} G_{(j)}(\beta, \overline{\mu}) + \widehat{\lambda} \left| \Upsilon_{\widehat{n}_\tau}(Y, \overline{\mu}) \right|. \tag{3.7}$$

Step 4. Initial estimator of $\overline{\mu}$. With the estimator $\widehat{\beta}$ obtained above, and for each tuning parameter $\widetilde{\tau}$ and the training set $T - T^\nu$, we find the estimator $\widehat{\overline{\mu}}^{(\nu)}(\widetilde{\tau})$ by the method given in the previous subsection.

Step 5. Selection of $\widetilde{\tau}$. Write $G_i^{(\nu)}(\widetilde{\tau}) = (Y_i - X_i^\top \widehat{\beta} - \widehat{\overline{\mu}}^{(\nu)}(\widetilde{\tau}))^2$ and let

$$\left\{G_{(j)}^{(\nu)}(\widetilde{\tau}) = \left(Y_{k_j} - X_{k_j}^\top \widehat{\beta} - \widehat{\overline{\mu}}^{(\nu)}(\widetilde{\tau})\right)^2 : (X_{k_j}, Y_{k_j}) \in T^\nu\right\}$$

be the order statistic of $\{G_i^{(\nu)}(\widetilde{\tau}) : (X_{k_j}, Y_{k_j}) \in T^\nu\}$ in descending order. Define the Bayesian cross-validation criterion as

$$CV(\widetilde{\tau}) = \frac{1}{n_{\widetilde{\tau}}} \sum_{\nu=1}^{5} \sum_{(X_{k_j}, Y_{k_j}) \in T^\nu, 1 \le j \le n_{\widetilde{\tau}}} G_{(j)}^{(\nu)}(\widetilde{\tau}) + \frac{2 \log n_{\widetilde{\tau}}}{n_{\widetilde{\tau}}}.$$

We then get an estimator $\widehat{\widetilde{\tau}}$ by minimizing $CV(\widetilde{\tau})$.

Step 6. Final estimator of $\overline{\mu}$. With the selected estimator $\widehat{\widetilde{\tau}}$, we estimate $\overline{\mu}$ by

$$\widehat{\overline{\mu}}_{Sec} = \arg\min_{\overline{\mu} \in \mathscr{U}} \frac{1}{\widehat{n}_{\widetilde{\tau}}} \sum_{j=1}^{\widehat{n}_{\widetilde{\tau}}} \left(H_{(j)} - \overline{\mu}\right)^2 + \widehat{\widetilde{\tau}} \, |\Gamma_{\widehat{n}_{\widetilde{\tau}}}|. \tag{3.8}$$

4. Extension and Discussions

It is known that there may be more than one distribution in $\mathscr{F}$ that can attain the upper expectation. In this section we first recommend an extended method to remove the uniqueness assumption on f_* in *C2*. It can be seen from the proof of Theorem 1 that the uniqueness assumption is only to guarantee that most of $\varepsilon_{k_j}, j = 1, \cdots, n_*$, have the same mean μ_*. Therefore, all we need in the data selection procedure is to identify the data that satisfy the first two order moment conditions: $\varepsilon_{k_j}, j = 1, \cdots, n_*$, have the same mean μ_* and the variance σ_*^2.

Let $\{D_{(j)} = Y_{l_j} - X_{l_j}^\top \widehat{\beta}_{LS}, j = 1, \cdots, N\}$ be the order statistics of $\{D_j = Y_j - X_j^\top \widehat{\beta}_{LS}, j = 1, \cdots, N\}$ in descending order. Similar to (3.1), the index set $I_n = \{l_j : j = 1, \cdots, n\}$ is decomposed as $I_n = U_n \cup L_n$. Write

$$\Lambda_n(X, Y) = \frac{1}{[n/2]} \sum_{j \in L_n} D_j - \frac{1}{n - [n/2]} \sum_{j \in L_n} D_j.$$

The proof of Lemma 1 given in the Supplement shows

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \widehat{\beta}_{LS}) = \frac{1}{n} \sum_{i=1}^{n} \mu_i - \frac{1}{N} \sum_{j=1}^{N} \mu_j E[X^\top] E^{-1}[XX^\top] E[X] + O_p\left(\frac{1}{\sqrt{N}}\right).$$

Thus, we can use $\Lambda_n(X, Y)$ to measure the difference among the means $\mu_{l_j}, j = 1, \cdots, n$, and then use $|\Lambda_n(X, Y)| < \tau_1$ to control the difference among the means μ_{l_j} for all j. Then an improved estimator of β is defined as the first component of the following solution:

$$(\widehat{\beta}_I^\top, \widehat{\overline{\mu}}_I)^\top = \arg \min_{\beta \in \mathscr{B}, \overline{\mu} \in \mathscr{U}, n_\lambda \in \mathscr{N}} \Big\{ \frac{1}{n_\lambda} \sum_{j=1}^{n_\lambda} G_{(j)}(\beta, \overline{\mu}) + \lambda |\Upsilon_{n_\lambda}(Y, \overline{\mu})| + \lambda_1 |\Lambda_{n_\lambda}(X, Y)| \Big\}, \tag{4.1}$$

where $\lambda \geq 0$ and $\lambda_1 \geq 0$ are two tuning parameters. We now use two penalties Λ_n and $\Upsilon_n(Y, \overline{\mu})$ to make sure that the selected data satisfy the first and second order moment conditions. Here a possible choice of n_λ is

$$n_\tau = \max \{ n : |\Upsilon_n(Y, \overline{\mu}) < \tau, |\Lambda_n(X, Y)| < \tau_1 \}.$$

Without the uniqueness assumption on f_*, Condition *C2* is replaced by

C7. The number n_c of the errors satisfying the first two order moment conditions tends to infinity as $N \to \infty$.

Theorem 3. *Under the model* (2.5)*, suppose Conditions C1, C4 and C7 hold,* λ *and* λ_1 *satisfy condition C3. Then the PMLS estimator defined in* (4.1) *satisfies*

$$\sqrt{n_c} \Big[(\widehat{\beta}_I - \beta)^\top, \widehat{\overline{\mu}}_I - \mu_* \Big]^\top \xrightarrow{d} N \left(0, \sigma_*^2 E^{-1}[\Phi(X)] \right) \quad (n_c \to \infty),$$

where μ_* *and* σ_*^2 *are defined in the previous section.*

A proof of the theorem is given in the Supplement.

Also we can use the second-step estimation procedure given in the previous section to construct the consistent estimator for $\overline{\mu}$. Let $\{H_{(j)}^I = Y_{m_j} - \widehat{\beta}_I^\top X_{m_j} : j = 1, \cdots, N\}$ be the order statistic of $\{H_j^I = Y_j - \widehat{\beta}_I^\top X_j : j = 1, \cdots, N\}$ in descending order, and $I_n = U_n \cup L_n$ be the decomposition of the index set $I_n = \{m_j : j = 1, \cdots, n\}$ as in (3.1). The second-step estimator of $\overline{\mu}$ is then defined by

$$\widehat{\overline{\mu}}_{Sec}^I = \arg \min_{\overline{\mu} \in \mathscr{U}, n_{\widetilde{\lambda}} \in \mathscr{N}} \frac{1}{n_{\widetilde{\lambda}}} \sum_{j=1}^{n_{\widetilde{\lambda}}} \Big(H_{(j)}^I - \overline{\mu} \Big)^2 + \widetilde{\lambda} \, |\Gamma_{n_{\widetilde{\lambda}}}^I|, \tag{4.2}$$

where $\Gamma_n^I = (1/[n/2]) \sum_{j \in U_n} H_j^I - (1/(n - [n/2])) \sum_{j \in L_n} H_j^I$. Then, this second-step estimator is consistent.

Theorem 4. *Under the conditions of Theorem 3, Conditions C5 and C6, when* $\widetilde{\lambda}$ *satisfies condition C3 and* $\{\varepsilon_{l_j}, j = 1, \cdots, n_c\}$ *and* $\{\varepsilon_{m_j}, j = 1, \cdots, \widetilde{n}\}$ *are not overlapped, then the second-step estimator in* (4.2) *satisfies*

$$\sqrt{\widetilde{n}} \left(\widehat{\overline{\mu}}_{Sec}^I - \overline{\mu} \right) \xrightarrow{d} N \left(0, \widetilde{\sigma}^2 + \sigma_*^2 E[X^\top] E[\Omega^{-1}(X)] E[X] \right) \quad (\widetilde{n} \to \infty),$$

where $\widetilde{\sigma}^2$*,* σ_*^2 *and* $\Omega(X, \theta)$ *are defined in the previous section.*

The difficulty we are facing now is the computational complexity because

there are three tuning parameters: λ, λ_1 and $\widetilde{\lambda}$. The computational steps are similar to those in the previous section. Because of the complexity, if the prior information on the uniqueness of the distribution f_* is available, we prefer to use the method given in the previous section to construct the estimators.

Consider the special case of $\beta = 0$. The model is simplified to

$$Y = \varepsilon. \tag{4.3}$$

We can see how the upper expectation $\overline{\mu} = \mathbb{E}[\varepsilon] = \mathbb{E}[Y]$ is estimated consistently whereas the existing result only derives $\underline{\mu} \leq \overline{Y} \leq \overline{\mu}$. Although the methods proposed can be used, estimation of this simple model, becomes much simpler. Let $\{Y_{(j)} = Y_{t_j}, j = 1, \cdots, N\}$ be the order statistics of $\{Y_j, j = 1, \cdots, N\}$ in descending order. For the index set $I_n = \{t_j : j = 1, \cdots, n\}$, we define the decomposition as $I_n = U_n \cup L_n$ as (3.1). Write $\Delta_n(Y) = 1/[n/2] \sum_{j \in U_n} Y_j - 1/(n - [n/2]) \sum_{j \in L_n} Y_j$. Then, the estimator for $\overline{\mu}$ is defined by

$$\widehat{\overline{\mu}} = \arg \min_{\overline{\mu} \in \mathscr{U}, n_\lambda \in \mathscr{N}} \frac{1}{n_\lambda} \sum_{j=1}^{n_\lambda} (Y_{(j)} - \overline{\mu})^2 + \lambda |\Delta_{n_\lambda}(Y)|, \tag{4.4}$$

where $\lambda \geq 0$ is a tuning parameter.

Let $\widetilde{n}$ be the sample size from $\widetilde{f}$ given in (3.5). We need the following simpler conditions than before:

C8. The variances σ_i^2 of ε_i exist for all $i = 1, \cdots, N$.

C9. $\widetilde{n} \to \infty$ as $N \to \infty$.

C10. Condition *C4* holds when the notations are replaced by the above accordingly.

Theorem 5. *Suppose that Conditions C3, and C8-C10 hold. Then the PMLS estimator* $\widehat{\overline{\mu}}$ *defined in* (4.4) *satisfies*

$$\sqrt{\widetilde{n}}(\widehat{\overline{\mu}} - \overline{\mu}) \xrightarrow{d} N\left(0, \sigma_*^2\right) \quad (\widetilde{n} \to \infty).$$

5. Concluding Remarks

This paper studied regression analysis with distribution randomness, and proposed some estimation methods for structing consistent estimators. Under the framework of distribution randomness, we define an upper expectation regression and construct consistent estimators for model parameters. Some issues are of importance.

First, the conditional upper expectation of error, given X, is required to be free of X. This condition is used to guarantee the identifiability of the upper

expectation regression so that we can estimate the parameters β and $\overline{\mu}$. As discussed in Section 1, the model considered in the paper can be regarded as an extension of Huber's upper expectation (Huber (1981)), and parallels to the classical regression setting with independence between ε and X. It is naturally to ask whether this condition can be relaxed. For instance, consider the model

$$\mathbb{E}[Y|X] = g(\beta, X) + \overline{\mu}[X], \tag{5.1}$$

where both $g(\beta, X)$ and $\overline{\mu}[X]$ are functions of the variable X. Clearly, without further constraints on their structures, the model is unidentifiable. Furthermore, even if the model is identifiable, both parameter and function may not be estimable. To explain the two difficulties, consider the identifiability issue first. Similar to the classical partially parametric models, for model identifiability, the function $\overline{\mu}[X]$ should have a structure such that it can be separated from $g(\beta, X)$. A possible scenario is that $\overline{\mu}[X] = m(\eta^\top X)$ for an unknown function $m(\cdot)$ and an unknown parameter vector η that is orthogonal to β such that $g(\beta, X) + \overline{\mu}[X]$ is identifiable. A parallel in the classical setting is the partially linear/parametric single-index model. However, even with this would-be-identifiable structure, estimating the function $\overline{\mu}[X]$ is still difficult because, from the data grouping approach described in Section 3, each group might not have enough data for different X_i to construct consistent estimators.

Second, Condition *C4* gives the range of the initial choice of n when the penalized estimation is used. This condition is mainly for technical purpose in the proof of estimation consistency and the identifiability of tuning parameter τ. It is also worth of a further investigation.

Third, our methodology is computationally intensive because it involves the choices of three tuning parameters in different scenarios.

Fourth, because of the lack of the information on the underlying distribution for every observation, the convergence rate of the estimator is slower than the typical rate of $1/\sqrt{n}$ in the classical parametric regression setting. Due to its difficulties, our study is regarded as a first attempt, while provides good opportunities for further study.

Supplementary Materials

Proofs and numerical studies can be found in the online supplementary materials.

Acknowledgment

The authors thank the Editor, the AE and two referees for their construc-

tive comments and suggestions that led to an improved presentation of an early manuscript. The research described here was supported by NNSF projects (11571204, 11231005 and 11671042) of China, NSF project (ZR2014AQ017) of Shandong Province of China, the Fundamental Research Funds for the Central Universities (15CX02083A), and a grant from the University Grants Council of Hong Kong.

References

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.

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.

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

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

Lin, L., Zhu, L. and Gai, Y. (2016). Inference for Biased models: a quasi-instrumental variable approach. *Journal of Multivariate Analysis*, **145**, 22–36.

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.

Zhongtai Securities Institute for Financial Studies, Shandong University, 27 Shanda Nanlu, Jinan, P.R. China 250100

E-mail: linlu@sdu.edu.cn

Zhongtai Securities Institute for Financial Studies, Shandong University, 27 Shanda Nanlu, Jinan, P.R. China, 250100

E-mail: dongping35@outlook.com

College of Science,China University of Petroleum, 66 Changjiang West Road, Huangdao District, Qingdao, P.R. China, 266580

E-mail: syqfly1980@163.com

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

E-mail: lzhu@hkbu.edu.hk

(Received May 2016; accepted June 2016)