UPPER EXPECTATION PARAMETRIC REGRESSION

Lu Lin¹, Ping Dong¹, Yunquan Song² and Lixing Zhu³

¹Shandong University, ²China University of Petroleum
and ³Hong Kong Baptist University

Abstract: In regression analysis, it is often the case that some predictors/factors are
either unobservable or at least not observed or ignored, and they are not included in a working model. However, these factors actually affect the response randomly.

Every observation in a sample may thus follow a conditional distribution when these factors are given, that is, every observation may follow a distribution randomly selected from a class of distributions. This phenomenon is called the distribution randomness. To well-estimate parameter of interest in such a working model, an upper expectation regression is proposed and a two-step penalized maximum least squares procedure is suggested to estimate the mean function and the upper expectation of the error. The resulting estimators are consistent and asymptotically normal in a certain sense. Simulation studies and a real data example are conducted to show that the classical least squares estimation does not work but the new penalized maximum least squares 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 such that a regression model can be well fitted. In this case, the relevant statistical analyses have been very intensive and mature. Under some other situations, however, the problems are much more complicated and difficult. A very typical scenario is that some predictors are either unobservable or unobserved or intentionally ignored such that the working model can be parsimonious enough for the consideration of interpretation and estimability. In high-dimensional paradigms, this is typically the case because a selected working model is often parsimonious and does not (or simply cannot) include all the predictors.

To better understand the problem, we start with a description about regression modelling. Suppose that a sample \( \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \) is available in a regression modelling. When there is a predictor(s) \( T_i \) that is paired with \( (X_i, Y_i) \) and has impact on the response \( Y_i \), but is unobservable or unobserved or ignored, we then cannot include it in the regression function. For simplicity, we call it the unobserved factor throughout the present paper. To be specific, suppose a parametric regression model is fitted as

\[
Y_i = g(\beta, X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, N, \tag{1.1}
\]
where \( g(\cdot, \cdot) \) is a given function, \( Y_i \)'s are the scalar response variables, \( X_i = (X_i^{(1)}, \cdots, X_i^{(p)})' \)'s are the associated \( p \)-dimensional predictors with a probability density \( f_X(\cdot) \). The parameters of interest are \( \beta \) and some others such as the expectations of \( \varepsilon_i \)'s. Note that in the case with unobserved \( T_i \), \( \varepsilon_i \) can be written as a function \( \varepsilon_i(T_i) \) of \( T_i \). Under the independent identically distributed (IID) case, the observations follow a common distribution, the errors \( \varepsilon_i(T_i) \)'s are then IID with a common expectation as the intercept of the model.

To relax the IID assumption, there are a number of efforts in the literature to deal with heteroscedasticity and other complicated model structures. However, a more serious situation is that the unobserved factor(s) \( T_i \) may affect the response randomly from observation to observation. Thus, \( \varepsilon_i \) is affected by \( T_i \) randomly from observation to observation. For any \( 1 \leq i \leq n \), the distribution of \( \varepsilon_i \) when \( T_i \) is given is of a random nature, meaning that for every \( T_i \), \( \varepsilon_i \) follows a conditional distribution when \( T_i \) is given, rather than an unconditional distribution. We call this the distribution randomness. This is obviously different from heteroscedasticity. Its strict definition will be given in the next section.

An relevant example is contained in Huber (1981) whose model can be regarded as a special case in which \( Y_i = \varepsilon_i \). When data contain gross
errors in the true random variables, the identical distribution assumption of sample is no longer satisfied. This causes the difficulty to define a common expectation of all $\varepsilon_i$’s. Further, it is often the case in practice that one only uses a part of predictors/factors to build a parsimonious working model as an approximation to the true model or, theoretically, any working model could be regarded as an approximation to the underlying model such that the modelling and follow-up analysis can proceed efficiently. However, those unobserved or ignored factors would have impact on the response. We now use a simple example that has been used in Fan and Peng (2004) to explain the problem that some ignored factors may effect to the response and thus the distribution randomness would be considered to appear. We also see whether modelling could be done better when distribution randomness is taken into consideration. Fan and Peng (2004) used a linear regression model, together with OLS, to fit the data set of the Fifth National Bank of Springfield from 1995 (see examples 11.3 and 11.4 in Albright et al. (1999)). The regression relation of the annual salary on four predictors are considered: job level, education level, gender and a dummy variable with value 1 if the employee’s current job is computer related and value 0 otherwise. Now we present the histogram of residuals and a kernel-based fitting for the residuals that are given in Figure 1. The histogram presents
large dispersion of the residuals and particularly there is a small cluster on the left hand side of the main cluster. In other words, the distribution would have more than one mode. It seems that the data are not well fitted.

This is either because an inappropriate model is used or because there are some other factors not in the regression function. We then also tried some nonlinear models and the results did not show much improvement. It seems that if we just use these factors as the covariates to fit a model, linear model would be good enough. However, the residual plot and model fit show that the regression errors $e_i$ may not be identically distributed as many residual values are deviated from the values around zero such that the fitted curve is significantly far away from a flat straight line. On the other hand, it is reasonably believed that both the year of working experience and age would also have impact on the salary. The unobserved/ignored factors are not included in the linear regression function, but in fact are absorbed into the error term.

This simple example just wants to show that when there are some unobserved factors, the results of statistical analysis would not be well interpretable. In other words, there are some other potential factors that would affect the response. Therefore, if a classical regression model is fitted, the error terms $e_i$'s may not be centered when we cannot determine
Figure 1: Figure for OLS fitting.

exactly whether and how unobserved factors make the distributions of the responses different. Consequently, we have difficulty to define a common expectation as the intercept as all individual observations would have their corresponding “intercepts”. In Section 5, we will see that an upper expectation regression model works better to fit the data when we still use the linear regression function as Fan and Peng (2004) used.

It would be argued that to well fit data, a natural way is including all possible factors into a working model. However, in practice this strategy makes the working model neither parsimonious nor possible to well estimate the parameters in the model in some scenarios. For instance in a high-dimensional paradigm, to establish a parsimonious working model, variable selection is often required and some or many variables are absorbed into the error term. When a model is not sparse and variable selection is imple-
Upper Expectation Regression

The error term may not be centered and the consistent estimation of the parameter of interest cannot be easily obtained. The distribution of the error may also be affected by other variables removed from the regression part. A relevant issue was discussed by Lin et al. (2016) in which a nonsparse model was considered and a semiparametric method was proposed to achieve estimation consistency. Therefore, instead of the classical expectation, a realistic consideration is to define an upper expectation that is the supreme of the classical expectations over a class of distributions of errors. This can avoid the distribution randomness and make the involved parameters estimable.

It is worth pointing out that the notion of upper expectation is not new, see Huber (1981). For the simplest case $Y_i = \varepsilon_i$, he proposed upper and lower expectations for the data that contain gross errors. To be precise, suppose that $\mathcal{F} = \{f_t : t \in \mathcal{T}\}$ is a class of distributions where $\mathcal{T}$ is an index set. Huber (1981) defined upper and lower expectations respectively as $\bar{\mu} = \sup_{f \in \mathcal{F}} E_{f}[\varepsilon_i]$ and $\underline{\mu} = \inf_{f \in \mathcal{F}} E_{f}[\varepsilon_i]$. But in his book, there was no method to consistently estimate the upper and lower expectation.

The present paper discusses a more general model: there are covariates in the regression model, and the distribution of each error is an element that is randomly selected from a class of distributions, say $\mathcal{F}$. Thus, all
of the elements in \( \mathcal{F} \) can be seen as possible scenarios in this random selection of distribution. More specifically, every element \( f \in \mathcal{F} \) can be regarded as a “conditional” distribution when unobserved affecting factor(s) \( T = t \in \mathcal{T} \) is given, where \( \mathcal{T} \) is a set of values of unobserved factors. Thus, the observations can be rewritten as \( Z_i \), and when \( T = t_i \) is given, \( Z_i \) follows a conditional distribution \( \mathcal{F}(\cdot | T = t_i) \). We will give a formal definition in the next section.

Consider the case where every distribution \( f_i \) of \( \varepsilon_i \) belongs to \( \mathcal{F} \). The individual expectation \( E_{f_i}(Y|X) = g(\beta, X) + E_{f_i}(\varepsilon|X) \) is difficult to estimate by the sample \( \{(X_1, Y_1), \ldots, (X_N, Y_N)\} \) because we cannot observe the random variable \( T_i \) and do not know from which distribution \( f_i \in \mathcal{F} \) every \( \varepsilon_i \) comes, even when this class of distributions only has finite elements. This is a much more complicated model structure than the common case of which all predictors are observable.

In this situation, all expectations of the errors \( \varepsilon_i \) are actually conditional expectations when the unobserved random variables \( T = t_i \) are given. Thus, we cannot define a common expectation for all observations because the data possibly follow different distributions \( f_i \) that are randomly selected from the class \( \mathcal{F} \). These conditional expectations are not estimable.

Similarly as shown in Huber (1981), the upper expectation can also be em-
employed in regression model, which is the maximum over the expectations of the errors. At the population level, we assume that \( \varepsilon \) follows a distribution \( f \) randomly selected from \( \mathcal{F} \). We then have

\[
\mathbb{E}[Y|X] = g(\beta, X) + \bar{\mu},
\]

where \( \bar{\mu} \) is the upper expectation of \( \varepsilon \) defined as \( \bar{\mu} = \mathbb{E}[\varepsilon] = \sup_{f \in \mathcal{F}} \mathbb{E}_f[\varepsilon] \).

This is an upper expectation regression deduced from the model (1.1).

The upper expectation regression (1.2) can describe some realistic situations. Huber's upper expectation is a case as it has no regression component and the upper expectation is a special case of constant \( \bar{\mu} \). Also, 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). Their work is about financial risk theory. Further, a famous example in financial risk theory is the Knight uncertainty (Knight (1921)). Under the framework of Knight uncertainty, the different observations may come from different distributions randomly selected from a class of distributions and the related economic analysis is based on this uncertainty with such a class. The reader can refer to Gilboa and Schmeidler (1989) for more details.

Bayesian statistics also deals with distribution randomness. But its main difference from the distribution randomness described herein is as
follows. Bayesian models still assume that all observations are from a certain distribution when the parameter that has a given prior distribution is given. See, for example, the review of Clyde and George (2004) and the references therein. However, for all the above mentioned problems, no similar or comparable researches in the literature give us valuable information as references, to the best of our knowledge.

In the present paper, the primary target is consistently estimating $\beta$ and $\mu$ by the observations from the model (1.1). The following problems have to be solved in any estimation procedure:

Upper expectation estimability: The definition of upper expectation implies the basic feature of nonlinearity, or more precisely, the subadditivity:

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

for any random variables $U$ and $V$. Consequently, the sample mean when a sample is available cannot be guaranteed to converge to a fixed value such as the classical expectation under IID cases. For example, if the regression function $g(\beta, X) \equiv 0$ in the model (1.1), then we want to consistently estimate the upper expectation $\mu = \mathbb{E}[Y]$ of $Y$.

By Law of Large Numbers (LLN) under sublinear expectation such as Peng (2008, 2009), the sample mean $\bar{Y}$ of $Y_1, \cdots, Y_n$ would only satisfy
that with large probability,

\[ \mu \leq \bar{Y} \leq \bar{\mu}, \]

where \( \mu = \inf_{f \in \mathcal{F}} E_f[Y] \) and \( \bar{\mu} = \sup_{f \in \mathcal{F}} E_f[Y] \) are respectively the lower and upper expectations. It presents an obvious evidence that even under very simple models, existing methods have difficulty to consistently estimate the upper expectation \( \bar{\mu} \). Thus, the estimation consistency is a very challenging issue under distribution randomness.

2) Data availability: From problem 1) above, we need to explore the conditions and then select the data that can be used to estimate the parameters of interest. This raises the issue of data availability. First, intuitively, for an index if \( E_{\bar{f}}(\varepsilon) = \bar{\mu} \) holds or at least approximately holds in a certain sense, the corresponding observation \((X_i, Y_i)\) could then be used for estimation purpose. Thus, we have to, under certain conditions, identify those observations. Second, an embedded issue in point estimation is about data availability for different parameters of interest. The observations that can be used for estimating the mean function \( g(\beta, \cdot) \) may not be feasible for estimating \( \bar{\mu} \).

These problems, particularly the second problem, have not yet been explored in the literature. The classical statistical methodologies such as
the least squares and the maximum likelihood are no longer applicable and thus a new method is highly demanded.

As a first attempt, in this paper, we consider the class $F$ contains finite members. A penalized maximum least squares (PMLS) is introduced and then a two-step estimation procedure is suggested. The key feature of this method is that for different parameters $\beta$ and $\mu$ in the model (1.2), the method 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 identical distribution assumption is appropriate.

The remainder parts of the paper are organized in the following way. In Section 2, the definition about how to randomly select distribution is given, the upper expectation regression is reexamined, and the motivation for an estimation procedure is discussed. Section 3 contains the methodology development, the asymptotic properties of the estimators, the tuning parameter selection, and a related algorithm. The method is further extended in Section 4 to the case where the upper expectation can be attained by several distributions. As a special case, the estimator for the upper expectation is constructed in Section 4. Simulation studies and a real data example
are presented in Section 5. The paper concludes with some discussions in Section 6. The proofs of the theorems are given in the supplementary materials.

2. Motivation and upper expectation regression

2.1 Definition

We first give the formal definition about a model that randomly selects a distribution for every variable. Suppose that $\mathcal{F}$ is a distribution class with a factor set $\mathcal{T}$ such that $\mathcal{F} = \{f(\cdot, t) : t \in \mathcal{T}\}$. On the sample space $\Omega_T$ of $\mathcal{T}$, there is a probability measure $P(\cdot)$ such that $p(\cdot)$ is the distribution with respect to $P(\cdot)$. The factor variable $T$ defined on $\Omega_T$ follows the distribution $p(\cdot)$.

**Definition 2.1.** Let $Z = Z(T)$ be a random variable whose distribution satisfies the following property: for any fixed $t = t \in \mathcal{T}$, the distribution of $Z = Z(t)$ is $f_t(z) = f(z(t), t) \in \mathcal{F}$, and $T$ is an unobserved variable following the distribution $p(\cdot)$. For convenience, we call it the distribution randomness. Two random variables are called independent identically distributed under the above distribution structure if they are independent and each satisfy the above property.

This definition can be explained as follows. There is an unobserved random factor(s) $T$ that has impact on the distribution of the random variable.
$Z(T): t \in T$ is a stochastic process/random variable sequence. Consider the pair of random variables $(Z(T), T)$. The corresponding joint distribution is $f(Z(t), t)p(t)$, where $f(Z(t), t)$ can be regarded as a conditional distribution of $Z(t)$ when $T = t \in T$ is given. Because the randomness and unobservability of $T$, $Z(T)$ is different from the random variable defined in the classical stochastic process. If $T$ were observable, the problem would reduce to the classical functional data framework where all observations $(Z(T), T)$ were functional data. However, under the present framework, what we can observe is just $Z(T)$ in which $T$ is unobserved. Therefore, any element $f(., t)$ within the class $F$ could be the distribution of $Z(T)$ in the above random manner. Without notional confusion, we then simply write $Z(T)$ as $Z$ sometimes. Thus, for a random variable function $g(Z)$, the expectation $E[f(Z)]$ is actually the conditional expectation with conditional density $f_t = f(Z(t), t)$ given above.

2.2 Upper Expectation Linear Regression

For ease of exposition, we mainly consider the linear regression, i.e., at the population level, the regression model has the form of $Y = \beta^T X + \epsilon$, where $T$ stands for transposition, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional vector of unknown parameters. A brief discussion on the extension of the results to the nonlinear model (1.1) will be given at the end of Section 4.