REVSTAT – Statistical Journal
Volume 7, Number 1, April 2009, 37–54

PREWHITENING-BASED ESTIMATION IN PARTIAL LINEAR REGRESSION MODELS: A COMPARATIVE STUDY

Authors: Germán Aneiros-Pérez
– Departamento de Matemáticas, Universidade da Coruña, Spain
ganeiros@udc.es

Juan Manuel Vilar-Fernández
– Departamento de Matemáticas, Universidade da Coruña, Spain
eijvilar@udc.es

Abstract:

- The problem of semiparametric modelling in time series is considered. For this, partial linear regression models are used, that is, regression models where the regression function is the sum of a linear and a nonparametric component. Two estimators for the nonparametric component are shown: one estimator takes into account the dependence structure in the errors of the regression function and the another estimator not. Both estimators are compared in practice for several real time series concerning economics and finance. In addition to the partial linear regression model, other regression models are included in the comparison.

Key-Words:

- *nonparametric regression; time series.*

AMS Subject Classification:

- 62G08, 62G20, 62M10, 91B84, 62P20.

1. INTRODUCTION

Linear regression modelling is a nice form for linking variables because in general the parameters have some kind of meaning or interpretation. Nevertheless, it is known that the main drawback of the linear regression models is their lack of flexibility. In practice, this fact causes that some interesting relationships can not be modelled by means of this class of models.

A way to avoid that drawback is to add to the linear regression function a nonparametric component. The resulting model, known as a partial linear regression (PLR) model, was introduced by Engle *et al.* (1986) to study the effect of weather on electricity demand. Another interesting feature of the PLR models is that they also avoid the "curse of dimensionality" (assuming low dimension for the explanatory variable that enters in a nonparametric way). From a theoretical point of view that dimension can be high, but usually is 1. Thus, we can say that the PLR models are flexible models that, in practice, can handle multiple variables.

Since the pioneer work of Engle *et al.* (1986), several papers have been published on this class of models in the setting of i.i.d. data (see, e.g., Speckman, 1988, Robinson, 1988, or Linton, 1995) as well as for dependent data (see, e.g., Gao, 1995, or Aneiros-Pérez *et al.*, 2004). In these papers, one can find asymptotic results (consistency, asymptotic normality) on estimators for each component in the PLR model, as well as on bandwidth selectors for those estimators, and even on testing of hypotheses. In addition, PLR models have demonstrated their usefulness in many field of applied sciences, such as economics, environmental studies, medicine... (see Härdle *et al.*, 2000, for a monograph and applications of the PLR model). A common feature in all these publications is that they study the same type of estimator, regardless of the data are independent or not.

In a recent paper, Aneiros-Pérez and Vilar-Fernández (2008) proposed a new estimator for the nonparametric component in a PLR model under random design and dependence conditions. To construct their estimator, these authors took into account the dependence structure in the errors of the model. Specifically, this dependence structure was used to transform the PLR model with dependent errors into a PLR model with uncorrelated errors (say, into a "whitened model"). Then, the estimator of the nonparametric component was based on this whitened model. Aneiros-Pérez and Vilar-Fernández (2008) obtained the asymptotic normality of both the estimator based on the original PLR model and the estimator based on the whitened model. As a consequence, they noted that the second estimator is asymptotically more efficient than the first estimator.

The aim of this paper is to illustrate, in practice, both the competitiveness of the PLR model and the usefulness of the prewhitening transformation.

Our paper is organized as follows. The PLR model is presented in Section 2, and estimators of both linear and nonparametric components of the model are motivated and defined. Then, a comparative study on the behavior of those estimators is deeply carried out in Section 3. For this, three real datasets in the context of economics and finance were analyzed. Concluding remarks are given in Section 4.

2. MOTIVATION AND CONSTRUCTION OF THE ESTIMATORS

2.1. The partial linear regression model

The class of the PLR models assumes that the regression function is the sum of a linear and a nonparametric component. This can be mathematically expressed through the regression model

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\beta + m(\mathbf{T}_i) + \varepsilon_i \quad (i = 1, ..., n) \ , \tag{2.1}$$

where $\mathbf{X}_i = (X_{i1}, ..., X_{id_0})^{\mathrm{T}}$ and $\mathbf{T}_i = (T_{i1}, ..., T_{id_1})^{\mathrm{T}}$ ($d_0 \geq 1$, $d_1 \geq 1$) are vectors of explanatory variables, $\beta = (\beta_1, ..., \beta_{d_0})^{\mathrm{T}}$ is a vector of unknown real parameters, m is an unknown smooth real function and $\{\varepsilon_i\}$ are the random errors satisfying

$$E(\varepsilon_i \,|\, \mathbf{X}_i, \mathbf{T}_i) = 0 \quad (i = 1, ..., n) \ . \tag{2.2}$$

In this paper, we focus on estimation of m in (2.1) when both $\mathbf{X}_i$ and $\mathbf{T}_i$ are random (random design) and, in addition, the errors ε_i are dependent. Specifically, we assume that these errors follow the invertible linear process

$$\varepsilon_i = \sum_{j=0}^{\infty} c_j \, e_{i-j} \ , \quad \text{where} \ \ c_0 = 1 \ \text{and} \ e_i \ \text{are i.i.d. with} \ E(e_i) = 0 \ . \tag{2.3}$$

In general, estimators proposed in the statistical literature to estimate m in (2.1) do not have into account the dependence structure in the errors. However, it seems natural to think that incorporate such information in the construction of the estimator can be helpful. Aneiros-Pérez and Vilar-Fernández (2008) proposed to use that dependence structure in the following way. Let us denote $c(L) = \sum_{j=0}^{\infty} c_j L^j$, where L is the lag operator, and

$$a(L) = c(L)^{-1} = a_0 - \sum_{j=1}^{\infty} a_j L^j \qquad \text{with} \ \ a_0 = 1 \ . \tag{2.4}$$

Applying $a(L)$ to the PLR model (2.1) and rewriting the corresponding equation, we obtain the new PLR model

$$\underline{Y}_i = \mathbf{X}_i^{\mathrm{T}}\beta + m(\mathbf{T}_i) + e_i \quad (i = 1, ..., n) \ , \tag{2.5}$$

where $\underline{Y}_i = Y_i - \sum_{j=1}^{\infty} a_j\big(Y_{i-j} - \mathbf{X}_{i-j}^{\mathrm{T}}\,\beta - m(\mathbf{T}_{i-j})\big) = Y_i - \sum_{j=1}^{\infty} a_j\,\varepsilon_{i-j}$. As we see, the regression function in the PLR models (2.1) and (2.5) is the same, but in (2.5) the errors are i.i.d. Now, it should be noted that an estimator for m based on the PLR model (2.5) takes into account the dependence structure of the errors ε_i (through $\underline{Y}_i$). From now on, we will name "original PLR model" and "whitened PLR model" to the models (2.1) and (2.5), respectively.

2.2. The estimators

Aneiros-Pérez and Vilar-Fernández (2008) studied and compared (from an asymptotic point of view) two estimators for $m(\mathbf{t})$, one (say $\widehat{m}(\mathbf{t})$) based on the original PLR model (2.1) and the other (say $\underline{\widehat{m}}(\mathbf{t})$) based on the whitened PLR model (2.5). Specifically, these authors proved that, under suitable conditions, the asymptotic distribution of these estimators (properly normalized) is Gaussian. In summary, from that result one can observe that both estimators asymptotically have the same bias but different variances, the variance of $\widehat{m}(\mathbf{t})$ relative to the variance $\underline{\widehat{m}}(\mathbf{t})$ being $\sigma_\varepsilon^2/\sigma_e^2 = \sum_{j=0}^{\infty} c_j^2 \geq 1$ (the equality holding if and only if $\{\varepsilon_i\}$ is i.i.d.). Thus, we have that the estimator based on the whitened PLR model is asymptotically more efficient than the estimator based on the original PLR model.

Now, we motivate and construct both estimators $\widehat{m}(\mathbf{t})$ and $\underline{\widehat{m}}(\mathbf{t})$. We begin with $\widehat{m}(\mathbf{t})$. Let us assume that we have a root-n consistent estimator for β (say $\widehat{\beta}_{h_0}$). Then, from the original PLR model (2.1) we can write

$$Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\beta}_{h_0} \approx m(\mathbf{T}_i) + \varepsilon_i \quad (i = 1, ..., n) \ ,$$

and, assuming that m is a smooth function, we proceed to estimate $m(\mathbf{t})$ by means of the nonparametric estimate

$$\widehat{m}_{h_0,h_1}(\mathbf{t}) = \sum_{j=1}^{n} w_{h_1,j}(\mathbf{t})\,\big(Y_j - \mathbf{X}_j^{\mathrm{T}}\widehat{\beta}_{h_0}\big)\,, \tag{2.6}$$

where $w_{h_1,j}(\mathbf{t})$ are weight functions depending on $\mathbf{t}$ and the design points $\mathbf{T}_i$, and both h_0 and h_1 are smoothing parameters or bandwidths that typically appear in any setting of nonparametric or semiparametric estimation. The weight functions considered in Aneiros-Pérez and Vilar-Fernández (2008) were local p-order polynomial type weights (for local polynomial estimation see, e.g., Fan and Gijbels, 1996, or Francisco-Fernández and Vilar-Fernández, 2001).

As we see, the estimator (2.6) is based on the original PLR model (2.1), but similar steps could be used to construct an estimator for $m(\mathbf{t})$ based on the whitened PLR model (2.5). Because in practice the response variable $\underline{Y}_i$ in (2.5) is unknown (depends on a_i, β and m), the first step in constructing such an estimator should be to propose a "reasonable" approximation for $\underline{Y}_i$. In this way, Aneiros-Pérez and Vilar-Fernández (2008) proposed to use the residuals $\widehat{\varepsilon}_i = Y_i - \mathbf{X}_i^{\mathrm{T}}\widehat{\beta}_{h_0} - \widehat{m}_{h_0,h_0}(\mathbf{T}_i)$ of the original PLR model to construct an estimate of $A_{\mathcal{T}} = (a_1, ..., a_{\mathcal{T}})^{\mathrm{T}}$, $\mathcal{T}$ being a truncation parameter large enough to avoid problems with the bias. Specifically, this estimator for $A_{\mathcal{T}}$ is constructed by means of the ordinary least squares (OLS) method applied to the model

$$\widehat{\varepsilon}_i \;=\; a_1\widehat{\varepsilon}_{i-1} + \cdots + a_{\mathcal{T}}\,\widehat{\varepsilon}_{i-\mathcal{T}} + \mathit{residual}_i \quad (i = \mathcal{T}+1, ..., n)\ . \tag{2.7}$$

In this way, the estimator

$$\widehat{A}_{\mathcal{T}} = \left(\widehat{\varepsilon}_{\mathcal{T}}^{\mathrm{T}}\,\widehat{\varepsilon}_{\mathcal{T}}\right)^{-1}\widehat{\varepsilon}_{\mathcal{T}}^{\mathrm{T}}\,\widehat{\varepsilon} \tag{2.8}$$

is obtained, where $\widehat{\varepsilon} = (\widehat{\varepsilon}_{\mathcal{T}+1}, ..., \widehat{\varepsilon}_n)^{\mathrm{T}}$ and $\widehat{\varepsilon}_{\mathcal{T}} = (\widehat{\varepsilon}_{i,j})_{\substack{1\le i\le n-\mathcal{T}\\ 1\le j\le \mathcal{T}}}$ with $\widehat{\varepsilon}_{i,j} = \widehat{\varepsilon}_{i-j+\mathcal{T}}$. Now, using $\widehat{A}_{\mathcal{T}}$ together with $\widehat{\beta}_{h_0}$ and $\widehat{m}_{h_0,h_0}$, we define

$$\underline{\widehat{Y}}_{\mathcal{T},i} \;=\; Y_i - \sum_{j=1}^{\mathcal{T}} \widehat{a}_j\Big(Y_{i-j} - \mathbf{X}_{i-j}^{\mathrm{T}}\,\widehat{\beta}_{h_0} - \widehat{m}_{h_0,h_0}(\mathbf{T}_{i-j})\Big) \quad (i = \mathcal{T}+1, ..., n)\ . \tag{2.9}$$

Finally, from (2.5) and (2.9), we can write

$$\underline{\widehat{Y}}_{\mathcal{T},i} - \mathbf{X}_i^{\mathrm{T}}\widehat{\beta}_{h_0} \approx\; m(\mathbf{T}_i) + \varepsilon_i \quad (i = \mathcal{T}+1, ..., n)\ ,$$

and, as in (2.6), we construct the estimator

$$\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t}) \;=\; \sum_{i=\mathcal{T}+1}^{n} w_{h_1,i}(\mathbf{t})\left(\underline{\widehat{Y}}_{\mathcal{T},i} - \mathbf{X}_i^{T}\widehat{\beta}_{h_0}\right). \tag{2.10}$$

In summary, the steps taken to construct the estimator (2.10) are:

Step 1: Construct a root-n consistent estimator $\widehat{\beta}_{h_0}$ for β.

Step 2: Construct the residuals $\widehat{\varepsilon}_i$.

Step 3: Use $\widehat{\varepsilon}_i$ $(i = \mathcal{T}+1, ..., n)$ to construct an estimator $\widehat{a}_j$ for a_j $(j = 1, ..., \mathcal{T})$.

Step 4: Use $\widehat{a}_j$ $(j = 1, ..., \mathcal{T})$, $\widehat{\beta}_{h_0}$ and $\widehat{m}_{h_0,h_0}$ to construct an approximation $\underline{\widehat{Y}}_{\mathcal{T},i}$ for $\underline{Y}_i$ $(i = \mathcal{T}+1, ..., n)$.

Step 5: Use $\widehat{\beta}_{h_0}$ and $\underline{\widehat{Y}}_{\mathcal{T},i}$ to construct the estimator (2.10).

Finally, we motivate the root-n consistent estimator $\widehat{\beta}_{h_0}$ for β used in Aneiros-Pérez and Vilar-Fernández (2008). If we subtract $E(Y_i \,|\, \mathbf{T}_i)$ on both sides of equality (2.1), we get the linear regression model

$$Y_i - E(Y_i \,|\, \mathbf{T}_i) \,=\, \big(\mathbf{X}_i - E(\mathbf{X}_i \,|\, \mathbf{T}_i)\big)^{\mathrm{T}} \beta + \varepsilon_i \quad (i = 1, ..., n) \,. \tag{2.11}$$

Then, replacing the response and explanatory variables in (2.11) (that is, $Y_i - E(Y_i \,|\, \mathbf{T}_i)$ and $\mathbf{X}_i - E(\mathbf{X}_i \,|\, \mathbf{T}_i)$, respectively) by the corresponding residuals obtained when Y and $\mathbf{X}$ are nonparametrically adjusted for $\mathbf{T}$, we can write (in matricial form)

$$\widetilde{\mathbf{Y}}_{h_0} \approx \, \widetilde{\mathbf{X}}_{h_0}^{T} \, \beta + \varepsilon_i \quad (i = 1, ..., n) \,, \tag{2.12}$$

where, for both the n-dimensional vector $\mathbf{A} = \mathbf{Y}$ and the $(n \times d_0)$-matrix $\mathbf{A} = \mathbf{X}$, and for any real number $h_0 > 0$, we have denoted $\widetilde{\mathbf{A}}_{h_0} = (\mathbf{I} - \mathbf{W}_{h_0})\, \mathbf{A}$ with $\mathbf{W}_{h_0} = \big(w_{h_0,j}(\mathbf{T}_i)\big)_{i,j}$. Now, using OLS in (2.12), one obtains

$$\widehat{\beta}_{h_0} = \big(\widetilde{\mathbf{X}}_{h_0}^{\mathrm{T}} \, \widetilde{\mathbf{X}}_{h_0}\big)^{-1} \, \widetilde{\mathbf{X}}_{h_0}^{\mathrm{T}} \widetilde{\mathbf{Y}}_{h_0} \,.$$

At this time, four facts should be clear. First, we dispose of two estimators for m in (2.1): $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$. Second, $\widehat{m}_{h_0,h_1}(\mathbf{t})$ does not take into account the dependence structure in the errors of the PLR model (2.1). Third, $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$ takes into account the dependence structure in the errors of the PLR model (2.1). Fourth, Aneiros-Pérez and Vilar-Fernández (2008) proved that $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$ is asymptotically more efficient than $\widehat{m}_{h_0,h_1}(\mathbf{t})$.

3. APPLICATIONS TO REAL DATA

The main goal of this section is to compare the behavior of the estimators $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$ when they are applied to real data. In addition, in order to make more general the study and not only confined to the PLR model, in a first attempt we will consider a set of regression models together with their conventional estimators. Then, the accuracy of these models/estimators will be compared with that of the PLR model/estimators $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$. In a second attempt, we will take into account the fact that the prewhitening transformation (2.4) can also be applied to that set of regression models. Thus, we will include in the study the estimators based on the corresponding whitened regression models. Then, when the study is completed, we will have shown both the competitiveness of the PLR model and the usefulness of the prewhitening transformation.

Three real datasets will be analyzed, all related to the field of economics and finance. Specifically, the first example deals with market shares and prices of two dentifrices, while the second dataset is related to the building industry. Finally, in the third study we consider relationships between stock indices.

3.1. Models

In the three datasets we will study, we have a response variable (say Y) and two explanatory variables (say X and T). Thus, we will consider four classes of regression models linking Y with X and/or T. Now, we quickly present each one of these classes and give a short motivation of them:

Linear models. Maybe, the first class that comes to mind is that of the classical linear regression models. These models allow easy interpretation of the effect of each explanatory variable on the response variable. Nevertheless, it is known than its main handicap is the lack of flexibility.

Nonparametric models. In order to avoid the handicap named in the previous kind of models, the second class to be considered is that of nonparametric models. A problem of this class is the known as "curse of dimensionality", which is based on the fact that, when the number d of real explanatory variables is greater than 1, large sample sizes are required to obtain good estimates (these sample sizes increase exponentially with d). In view of this problem, we will restrict to the class of nonparametric models with only a real explanatory variable.

Partial linear models. This class was presented in a general setting in Section 2. In this practical study, we will consider PLR models that include one real variable in a linear form and another one in a nonparametric form (that is, $d_0 = d_1 = 1$ in (2.1)). Note that this class overcomes the curse of dimensionality.

Additive models. The last class is composed by the additive models with two explanatory variables, that is, nonparametric models whose regression function is the sum of two nonparametric components. It is interesting to observe that, as the previous class, this class avoids the curse of dimensionality (in fact, this is achieved in both classes by means of the same method: to express the regression function as the sum of two components).

The wide range of models named above can be seen in Table 1.

Table 1: Regression models.

Models	**Notation**
Linear models	
$Y = \alpha + X\beta + \varepsilon$	L1
$Y = \alpha + T\beta + \varepsilon$	L2
$Y = \alpha + X\beta_1 + T\beta_2 + \varepsilon$	L3
Nonparametric models	
$Y = m(X) + \varepsilon$	NP1
$Y = m(T) + \varepsilon$	NP2
Partial linear models	
$Y = X\beta + m(T) + \varepsilon$	PL1
$Y = T\beta + m(X) + \varepsilon$	PL2
Additive model	
$Y = \mu + m_1(X) + m_2(T) + \varepsilon$	ADD

3.2. Estimators

Now we indicate, for each regression model in Table 1, the kind of estimator considered. OLS estimators were used to estimate the parameters corresponding to the linear models, while the local linear polynomial estimator was considered to estimate the regression function in the nonparametric models. Both the parametric and the nonparametric component in the PLR models were estimated by means of the estimators presented in Section 2. Finally, for the additive model, a backfitting algorithm was considered (see Hastie and Tibshirani, 1990). In the last two classes of models, the weight functions were local linear polynomial type weights (as for the nonparametric case).

3.3. Choosing the parameters of the estimates

In practice, as usual in both nonparametric and semiparametric settings, it is necessary to choose some parameters related to the estimates. Specifically, we refer to the kernel and the bandwidths. In addition, for the cases where the prewhitening transformation is considered, we must give a value for the truncation parameter $\mathcal{T}$.

On the one hand, the Epanechnikov Kernel $K(u) = 0.75(1-u)^2 I_{[-1,1]}(u)$ and the truncation parameter $\mathcal{T} = 2$ were used in the estimates. On the other hand, the bandwidths were chosen by means of the cross-validation procedure.

In short, this bandwidth selector proposes to choose the value $\widehat{h}_{CV}$ that minimizes to the function

$$CV(h) \,=\, n^{-1} \sum_{i=1}^{n} \left(Y_i - \widehat{r}_h^i(X_i, T_i)\right)^2 \omega(T_i) \;,$$

where $\widehat{r}_h^i(\cdot,\cdot)$ denotes the estimator of the regression function (of each model considered) constructed without using the i-th observation, and $\omega(\cdot)$ is a weight function included to avoid boundary effects. Note that for both the PLR and the additive models h is a vector (say $h = (h_0, h_1)^{\mathrm{T}}$ and $h = (h_1, h_2)^{\mathrm{T}}$, respectively). The good (asymptotic) properties of the cross-validation selector are based on the fact that $CV(h)$ is (asymptotically) equivalent (except for a constant) to the average squared error (see Aneiros-Pérez and Quintela-del-Río 2001 for some results on this selector in the setting of PLR models).

3.4. Measure of accuracy

To compare the accuracy of the different models/estimators, we considered the Relative Cross-Validation (RCV)

$$RCV = \frac{CV\left(\widehat{h}_{CV}\right)}{\mathrm{Var}\{Y_i\}} \;,$$

where $\mathrm{Var}\{Y_i\}$ denotes the variance corresponding to the sample of responses. Observe that RCV gives a global measure of the accuracy of each model/estimator.

3.5. The dentifrice data

The first dataset analyzed consists of weekly market shares of Crest and Colgate dentifrice, together with the price of Crest dentifrice, during the period January 1, 1958 to April 30, 1963 (276 data). This dataset was used in Wichern and Jones (1977) to asses the impact of market disturbances, and can be found on the website `http://www.alianzaeditorial.es` (Book title: Análisis de Series Temporales; Author: Peña, D., Section: Ejercicios prácticos). The graphics of these time series are shown in Figure 1.

From Figure 1, we clearly observe the presence of trend in the three series. These trends were eliminated by differentiating. Then, using the transformed data, we seek in Table 1 an adequate regression model to link the data corresponding to the market share of Crest dentifrice (Y) with those of market share of Colgate dentifrice (X) and price of Crest dentifrice (T).

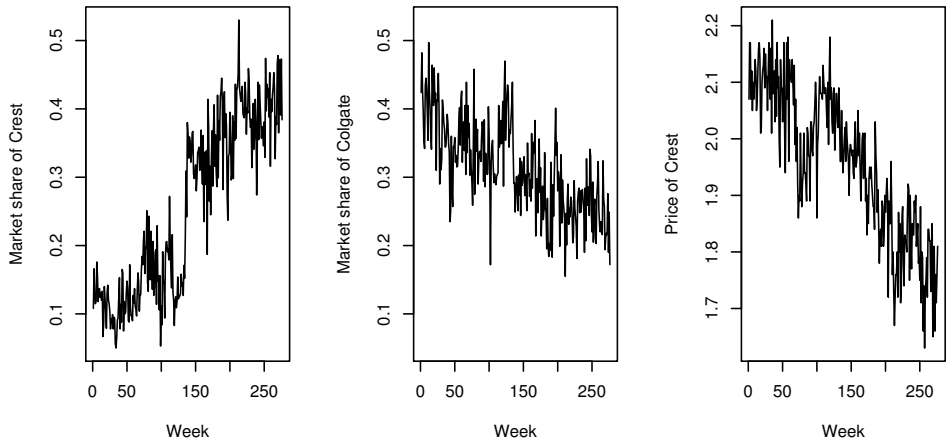

Figure 1: The market shares of Crest (left) and Colgate (middle) dentifrice, and the price of Crest dentifrice (right).

The RCV's of the models/estimates considered, as well as comparisons between them, are displayed in Table 2. From the third column in this table, we can say that the prewhitening transformation has proved useful in all the models. In addition, the fourth column indicates that the best model is the PLR model that includes the effect of market share of Colgate dentifrice and price of Crest dentifrice in a linear and a nonparametric way, respectively (that is, the PLR model PL1).

Table 2: Values of the criterion error and ratios (^{a}original model, ^{b}whitened model).

Model	**RCV**	$\mathbf{RCV}^b/\mathbf{RCV}^a$	$\mathbf{RCV}/\min \mathbf{RCV}$
L1	^{a}0.8347 ^{b}0.8313	0.9959	1.0370 1.0328
L2	0.9793 0.9748	0.9953	1.2167 1.2110
L3	0.8161 0.8108	0.9935	1.0139 1.0073
NP1	0.8359 0.8308	0.9939	1.0385 1.0321
NP2	0.9802 0.9751	0.9948	1.2177 1.2114
PL1	0.8090 0.8049	0.9950	1.0050 1.0000
PL2	0.8118 0.8077	0.9949	1.0086 1.0034
ADD	0.8219 0.8147	0.9912	1.0211 1.0121

Finally, we give some information on the estimates of the parameter β and the function m in the best model (PL1). The estimates of β using the conventional and the prewhitened-based estimators were -0.4146 and -0.4220, respectively. The corresponding estimates of m are shown in Figure 2 (from now on, in the graphics corresponding to the estimates of m, "Estimator 1" and "Estimator 2" are referred to as $\widehat{m}_{h_0,h_1}$ and $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}$, respectively).

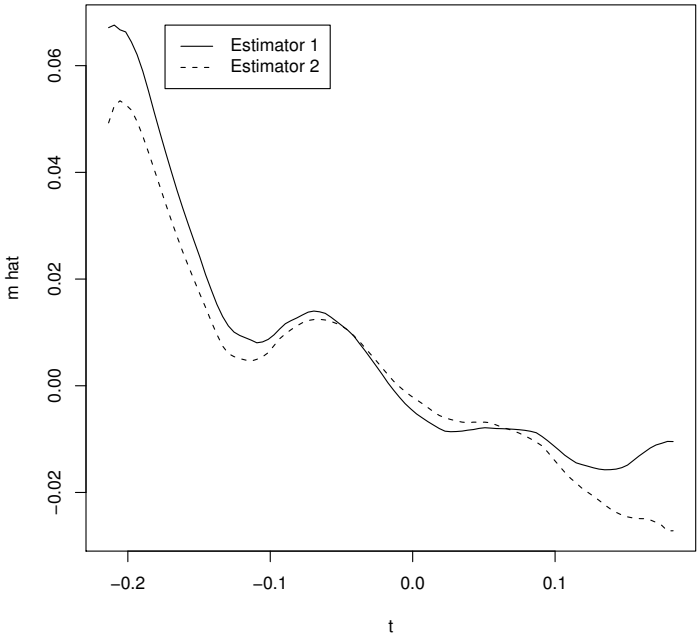

Figure 2: Estimates of the nonparametric component m in the PLR model PL1.

3.6. The building industry data

The building industry data is the second example we analyze. We have monthly observations corresponding to the number of buildings started, quantity of cement produced and number of buildings completed in Galicia (an autonomous community located in northwestern Spain) during the period January 1987 to December 2000 (168 data). These time series are available on the website `http://www.ige.eu`. Figure 3 displays these data.

Our goal is to get a model to analyze the effect of both the quantity of cement produced (X) and the number of buildings completed (T) on the number of buildings started (Y). Because, as in the previous example, our time series contain trend (see Figure 3) and therefore they are not stationary, we have worked with the differenced data.

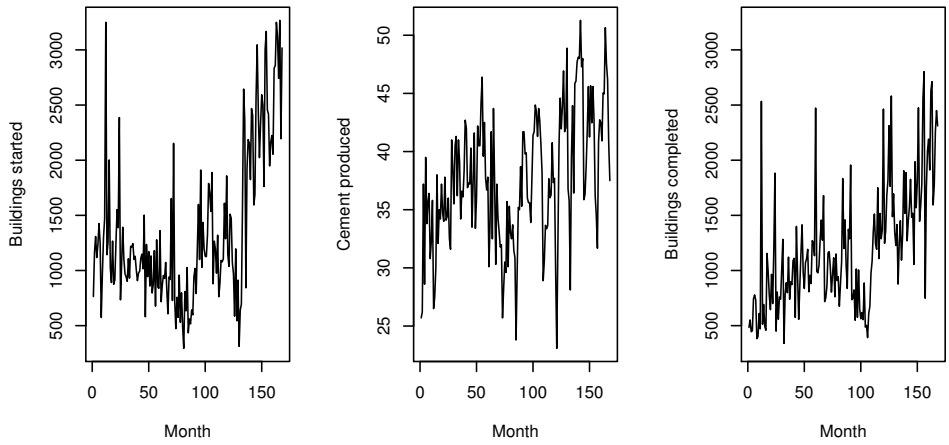

Figure 3: Number of buildings started (left), quantity of cement produced (middle) and number of buildings completed (right) in Galicia.

Table 3 shows interesting information on the accuracy of the models considered, as well as on the behavior of the different estimators. From this table, we note that the prewhitening transformation does not always improve the original model (see column 3), but there are improvements for the best two models (PL1 and ADD). Finally, the prewhitened-based estimator applied on the PLR model PL1 gives the best accuracy (see column 4).

Table 3: Values of the criterion error and ratios (^{a}original model, ^{b}whitened model).

Model	**RCV**	$\mathbf{RCV}^b/\mathbf{RCV}^a$	$\mathbf{RCV}/\min\mathbf{RCV}$
L1	^{a}0.9613 ^{b}0.9616	1.0003	1.0704 1.0707
L2	0.9872 0.9879	1.0007	1.0992 1.0999
L3	0.9488 0.9458	0.9969	1.0564 1.0531
NP1	0.9409 0.9472	1.0067	1.0476 1.0547
NP2	0.9610 0.9492	0.9877	1.0700 1.0569
PL1	0.9071 0.8981	0.9901	1.0100 1.0000
PL2	0.9170 0.9176	1.0006	1.0210 1.0216
ADD	0.9055 0.9001	0.9940	1.0082 1.0022

Focusing on the PLR model PL1, we have that the estimates of β using the conventional and the prewhitened-based estimators were 21.91 and 21.16, respectively. The corresponding estimates of m are shown in Figure 4.

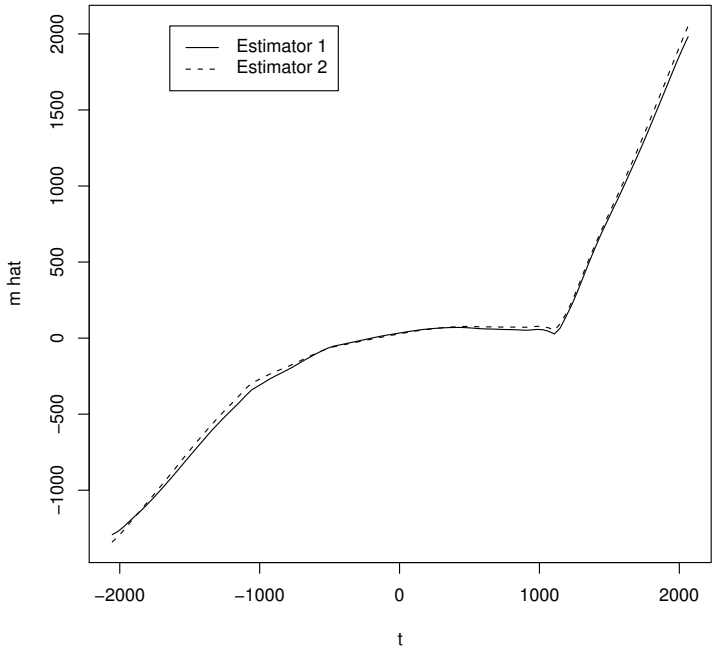

Figure 4: Estimates of the nonparametric component m in the PLR model PL1.

3.7. The stock data

Finally, we present an analysis on stock data. Specifically, our time series collect Banca Commerciale Index (Milan), FT-SE 100 Index (London) and General Index (Madrid) for each month during the period January 1988 to December 2000 (156 data). These data, which can be obtained on the website `http://www.ec.europa.eu/eurostat`, are shown in Figure 5.

From a first analysis of the data, we found the presence of both heteroscedasticity and trend. Thus, the data have been transformed using logarithms and then differentiated to achieve stationarity. Now, using the transformed data, we are interested in the construction of an adequate regression model to link the Banca Commerciale Index (Y) with the FT-SE 100 Index (X) and the General Index (T).

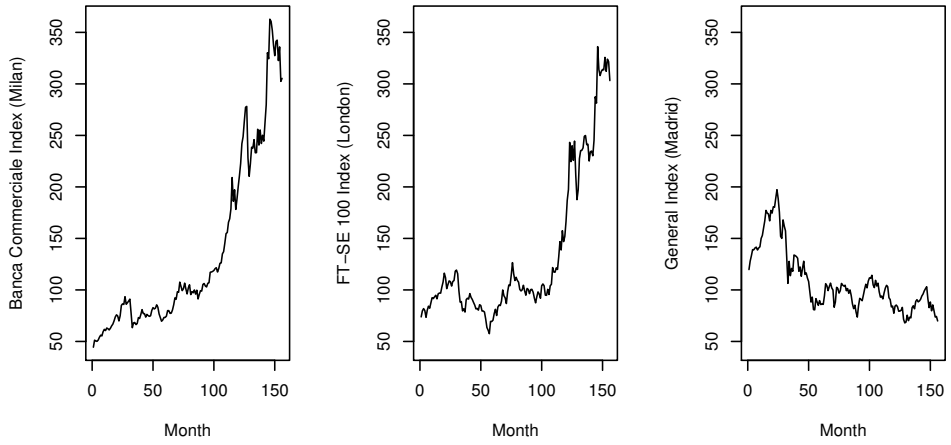

Figure 5: Banca Commerciale Index (left), FT-SE 100 Index (middle) and General Index (right).

The RCV's obtained for the different models considered, as well as comparisons between them, are given in Table 4. Some conclusions can be obtained from this table. First, only in the models L1 and NP2 the prewhitening transformation does not improve the original model (see column 3). Second, the best model is the PLR model that includes the effect of FT-SE 100 Index and General Index in a linear and a nonparametric way, respectively (that is, the PLR model PL1). Third, the prewhitened-based estimator applied on this PLR model gives the best accuracy.

Table 4: Values of the criterion error and ratios (^{a}original model, ^{b}whitened model).

Model	**RCV**	$\mathbf{RCV}^b/\mathbf{RCV}^a$	$\mathbf{RCV}/\min\mathbf{RCV}$
L1	^{a}0.7027 ^{b}0.7041	1.0019	1.0226 1.0246
L2	0.9859 0.9843	0.9984	1.4347 1.4324
L3	0.7129 0.7119	0.9986	1.0374 1.0360
NP1	0.7058 0.7044	0.9980	1.0271 1.0251
NP2	0.9768 0.9797	1.0029	1.4215 1.4257
PL1	0.6911 0.6872	0.9944	1.0057 1.0000
PL2	0.7004 0.6994	0.9986	1.0192 1.0178
ADD	0.7040 0.6978	0.9912	1.0245 1.0155

We complete the analysis showing the estimates of the parameter β and the function m in the best model (PL1). The estimates of β using the conventional and the prewhitened-based estimators were 0.4713 and 0.4666, respectively. The corresponding estimates of m are displayed in Figure 6.

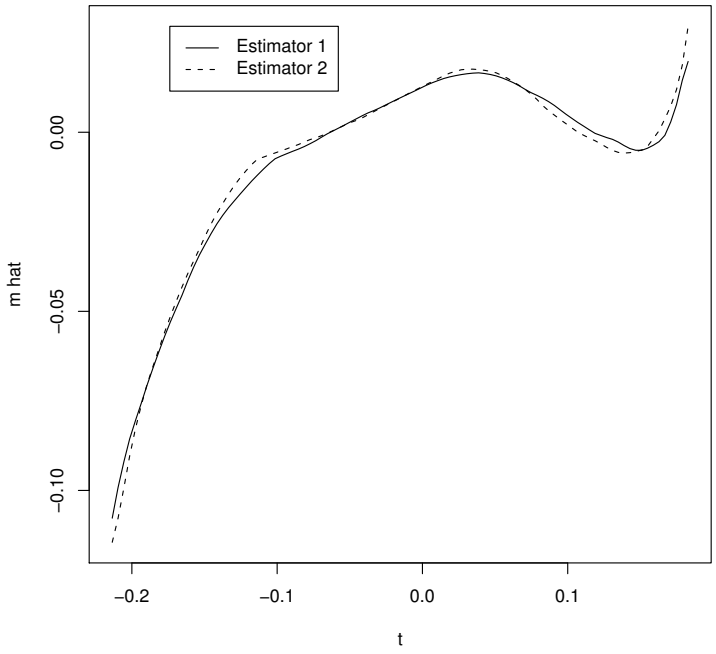

Figure 6: Estimates of the nonparametric component m in the PLR model PL1.

4. CONCLUDING REMARKS

In this paper, partial linear regression modelling in time series was dealt from a practical point of view. For this, we divided the paper into two parts. In the first part, some theory was shown. Specifically, we motivated and presented the PLR model. Then, we carefully constructed the estimator proposed in Aneiros-Pérez and Vilar-Fernández (2008), which is based on a whitened version of the original PLR model. By this motive, the estimator takes into account the dependence structure in the random errors (this fact is crucial for its good asymptotic behavior). The second part contains the main contribution of our work. It analyzes several real time series concerning economics and finance. Specifically, these time series were modelled by means of a wide range of regression models, including PLR models. Then, the corresponding regression functions were estimated. For this, both conventional and whitened-model based estimators were used. Finally, the performance of the corresponding estimators was measured.

In all the time series studied, the PLR model (estimated using the estimator proposed in Aneiros-Pérez and Vilar-Fernández, 2008) gave the best results.

We are aware that the improvement on the point-estimates is small. In fact, from the theoretical results, it is expected that a greater improvement is obtained on comparing confidence intervals (as noted in Subsection 2.2, the asymptotic difference between the conventional and the whitened-model based estimators is in their variances). Nevertheless, it should be noted that to construct confidence intervals one needs to estimate those variances, and the variability of the corresponding estimators could mask the theoretical result. For this reason, we have preferred to compare the point-estimates.

ACKNOWLEDGMENTS

This work has been supported by the grants numbers PGIDIT07PXIB105259PR and 07SIN012105PR from Xunta de Galicia (Spain), and by the grant number MTM2008-00166 from Ministerio de Ciencia e Innovación (Spain). We also acknowledge the valuable suggestions from the editor and the referees.

REFERENCES

[1] Aneiros-Pérez, G.; González-Manteiga, W. and Vieu, P. (2004). Estimation and testing in a partial linear regression model under long-memory dependence, *Bernoulli*, **10**, 49–78.

[2] Aneiros-Pérez, G. and Quintela-del-Río, A. (2001). Modified cross-validation in semiparametric regression models with dependent errors, *Communications in Statistics: Theory and Methods*, **30**, 289–307.

[3] Aneiros-Pérez, G. and Vilar-Fernández, J.M. (2008). Local polynomial estimation in partial linear regression models under dependence, *Computational Statistics and Data Analysis*, **52**, 2757–2777.

[4] Engle, R.; Granger, C.; Rice, J. and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association*, **81**, 310–320.

[5] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman and Hall.

[6] Francisco-Fernández, M. and Vilar-Fernández, J.M. (2001). Local polynomial regression estimation with correlated errors, *Communications in Statistics: Theory and Methods*, **30**, 1271–1293.

[7] Gao, J.T. (1995). Asymptotic theory for partly linear models, *Communications in Statistics: Theory and Methods*, **24**, 1985–2009.

[8] Härdle, W.; Liang, H. and Gao, J.T. (2000). *Partially Linear Models*, Physica Verlag.

[9] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman and Hall, New York.

[10] Linton, O. (1995). Second order approximation in the partially linear regression model, *Econometrica*, **63**, 1079–1112.

[11] Robinson, P. (1988). Root-n-consistent semiparametric regression, *Econometrica*, **56**, 931–954.

[12] Speckman, P. (1988). Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society: Series B*, **50**, 413–436.

[13] Wichern, D.W. and Jones, R.H. (1977). Assessing the impact of market disturbances using intervention analysis, *Management Science*, **24**, 329–337.