Gulf Journal of Mathematics ISSN: 2309-4966
Vol 21, Issue 1 (2025) 269-287 https://doi.org/10.56947/gjom.v21i1.3254

GOODNESS-OF-FIT TESTING FOR THE PARTIALLY LINEAR MODEL

BELLY N'KÔYOSSÈ MARTIAL AHOUTOU[1], KOUAMÉ FLORENT KOUAKOU[1*] and ARMEL FABRICE EVRARD YODÉ[1]

Abstract. In this paper, we propose a test to assess the relevance of using a partially linear model (PLM). Constructing our test statistic requires estimating the error variance in the context of a partially linear model. For this purpose, we use specific estimators of error variance. The test statistic is based on the difference between two such estimators. This difference is expected to be close to zero under the null hypothesis H_0 and significantly larger under the alternative hypothesis H_1. We establish the asymptotic properties of the proposed test statistic, and through simulations on finite-sample data, we evaluate its performance.

Keywords. curse of dimensionality, partially linear regression model, test statistic, asymptotic normality
2020 Mathematics Subject Classification. 62G05; 62G10; 62G20.

1. Introduction

Assume that the observations $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are independent and identically distributed random variables such that

$$Y_i = f(X_i) + \epsilon_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where f is an unknown real-value function defined on $\mathbb{R}^d$, and the errors ϵ_i are centered random variables. We assume that the random vectors $(\epsilon_1, ..., \epsilon_n)$ and $(X_1, ..., X_n)$ are independent.

To address the "curse of dimensionality" commonly encountered in nonparametric modeling, various dimension reduction approaches have been developed. Among the most prominent are the additive model, the PLM, and the single-index model. These models aim to retain model flexibility while reducing the effective dimensionality of the covariates. In this work, we focus specifically on the PLM, which combines the interpretability of linear models with the flexibility of nonparametric components. A model is said to be partially linear when it takes the following form

$$f(X_i) = \beta' Z_i + g(U_i),$$

Date: Received: Jul 10, 2025; Accepted: Aug 19, 2025.
* Corresponding author

where the unknown parameter $\beta \in \mathbb{R}^p$, and g is an unknown real-valued function defined on $\mathbb{R}^q$ with $p+q=d$. We assume that the random vectors $(X_1, ..., X_n)$ and $\{(Z_1, U_1), ..., (Z_n, U_n)\}$ are dependent, while the random vectors $(Z_1, ..., Z_n)$ and $(U_1, ..., U_n)$ are independent. To prevent the inappropriate use of partially linear models, several statistical tests have been developed to assess the validity of such models. Notable contributions include those by [11, 10, 4, 12, 9], who proposed various testing procedures under different assumptions and frameworks. It should be noted that the construction of the test statistics proposed by [11] and [10] requires prior partitioning of the sample. The underlying idea is that, if the data are split into two subsamples Γ_1 and Γ_2, the estimators computed separately on these subsamples should be consistent under the null hypothesis H_0. In practice, identifying subsamples Γ_1 and Γ_2 on which consistent estimators can be constructed under the null hypothesis H_0 is not straightforward. Moreover, [11] and [10] did not provide simulations to confirm or refute their theoretical results. [9] highlighted that the dimension of the linear covariates negatively impacts the test statistics proposed by [4] and [12]. To address this issue, they proposed a test statistic that is not affected by the dimension of the linear covariates. It is worth noting that both [12] and [9] impose a structural assumption on the dimension of the nonlinear covariates, specifically assuming that it equals one. It should also be noted that the estimators used in the test statistics proposed in the aforementioned studies incorporate mechanisms to account for potential outliers. It is well known that outliers can adversely affect the bias and mean squared error of estimators, see [3]. In this paper, we propose test statistics whose construction does not require prior partitioning of the sample, and we impose no assumptions on the dimension of the nonlinear covariate. We develop test statistics for both heterogeneous and homogeneous error cases. In each scenario, we introduce two types of test statistics: one based on estimators constructed using the entire dataset, and another based on estimators computed after removing potential outliers. In summary, we suggest a goodness-of-fit test to evaluate the relevance of using a PLM, under both homogeneous and heterogeneous error structures. The main idea of our approach is to compare a nonparametric regression function, defined as $f(X_i) = \mathbb{E}(Y_i|X_i)$, with a partially linear regression function, defined as $k(\beta' Z_i, U_i) = \beta' Z_i + g(U_i) = \mathbb{E}(Y_i|(Z_i, U_i)), \forall i = 1, ..., n$
This comparison serves to test whether the structure imposed by the PLM is consistent with the true conditional expectation of Y_i.

$$H_0 : f(X_i) = \beta' Z_i + g(U_i)$$
against
$$H_1 : f(X_i) \neq \beta' Z_i + g(U_i),$$

Our test is motivated by the fact that if the null hypothesis H_0 is true, then the error variance of the model should be

$$\sigma_i^2 = \mathbb{E}\big[(Y_i - \mathbb{E}(Y_i|(Z_i, U_i)))^2\big].$$

Under H_0, the estimator of σ_i^2, denoted $\widehat{\sigma_i^2}$, should be very close to σ_i^2, whereas under H_1, it should be significantly larger than σ_i^2.
If the error variance σ_i^2 is known, then an obvious test statistic for the above test

hypotheses would be

$$T = \sqrt{n}\big(\widehat{\sigma_i^2} - \sigma_i^2\big)$$

where $\widehat{\sigma_i^2}$ is an estimator of the error variance in a PLM. Under H_0, using the central limit theorem, we obtain $T \to N(0, \mu)$. Under H_1, $|T|$ would take a large value.

In most practical situations, the error variance is unknown. To address this, we adopt the approach proposed by [6] for the construction of our test statistic. This method involves comparing two versions of the estimated error variance $\widehat{\sigma_i^2}$, each computed using a different weighting scheme. Under the null hypothesis H_0, both versions are consistent estimators of σ_i^2. Consequently, if the test statistic yields a value close to zero, this suggests that the PLM is appropriate. Conversely, a large absolute value of the test statistic may provide evidence against the null hypothesis. The test statistics we propose therefore rely on accurate estimation of the error variance in the PLM framework. In [1], we introduced error variance estimators under the assumption that the nonlinear covariate is univariate. In [2], we have extended these estimators to accommodate cases where the nonlinear covariate may be multivariate. This generalized formulation will be used in the construction of our test statistics.

The organization of this paper is as follows. In Section 2, we present the mathematical motivation underlying our test and introduce the proposed test statistics. We also establish their asymptotic properties under both the null hypothesis H_0 and the alternative hypothesis H_1. Section 3 reports the results of numerical simulations conducted to evaluate the performance of the test statistics under various scenarios. In Section 4, we summarize our findings and discuss potential directions for future research. The appendix contains supplementary findings.

2. Hypothesis testing problem

We aim to test

$$H_0 : \ f(X_i) = k(\beta' Z_i, U_i)$$
$$\text{against}$$
$$H_1 : \ f(X_i) \neq k(\beta' Z_i, U_i),$$

where $f(X_i) = \mathbb{E}(Y_i|X_i)$ and $k(\beta' Z_i, U_i) = \mathbb{E}(Y_i|(Z_i, U_i))$.

2.1. **Motivation.** Let us define $M(X_i, Z_i, U_i) = f(X_i) - k(\beta' Z_i, U_i)$.

A certain value of $M(X_i, Z_i, U_i)$ can provide an indication of whether or not to use a PLM. To evaluate $M(X_i, Z_i, U_i)$, we can use the following distance:

$$\begin{aligned} d\big(f(X_i); k(\beta' Z_i, U_i)\big) &= \int \big(f(x) - k(\beta' z, u)\big)^2 P_X(d_x) P_Z(d_z) P_U(d_u) \\ &= \mathbb{E}\left[M(X_i, Z_i, U_i)\right]^2 \end{aligned}$$

If $f(.)$ is a nonparametric regression function that is similar to a partially linear regression function $k(.,.)$, then we have $\mathbb{E}\left[M(X_i, Z_i, U_i)\right]^2 = 0$ under H_0 and $\mathbb{E}\left[M(X_i, Z_i, U_i)\right]^2 > 0$ under H_1. Thus, if we estimate $d(f(X_i); k(\beta' Z_i, U_i))$, then a certain value of $d(f(X_i); k(\beta' Z_i, U_i))$ can indicate whether or not a PLM should be used.

An estimator of $d\big(f(X_i); k(\beta' Z_i, U_i)\big)$ is given by

$$\widehat{d}\big(f(X_i); k(\beta' Z_i, U_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \widehat{k}(\widehat{\beta}' Z_i, U_i)\big)^2, \tag{2.1}$$

where $\widehat{k}(\widehat{\beta}' Z_i, U_i) = \widehat{\beta}' Z_i + \widehat{g}(U_i)$,

$$\widehat{\beta} = \left[\sum_{i=1}^{n}(Z_i - \widehat{B}_i)(Z_i - \widehat{B}_i)'\right]^{-1}\left[\sum_{i=1}^{n}(Y_i - \widehat{A}_i)(Z_i - \widehat{B}_i)'\right] \tag{2.2}$$

$$\widehat{A}_i = \begin{cases} \dfrac{\sum_{j=1}^{n} Y_j K\left(\frac{U_j - U_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right)} & if \ \sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right) \neq 0 \\ 0 & else \end{cases} \tag{2.3}$$

$$\widehat{B}_i = \begin{cases} \dfrac{\sum_{j=1}^{n} Z_j K\left(\frac{U_j - U_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right)} & if \ \sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right) \neq 0 \\ 0 & else \end{cases} \tag{2.4}$$

and

$$\widehat{g}(U_i) = \begin{cases} \dfrac{\sum_{j=1}^{n} \big(Y_j - \widehat{\beta}' Z_j\big) K\left(\frac{U_j - U_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right)} & if \ \sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right) \neq 0 \\ 0 & else \end{cases} \tag{2.5}$$

Remark 2.1. The estimator proposed in (2.1) is nothing more than the classical estimator of the error variance in a PLM found in the literature. We observe that the process of constructing our test statistics naturally requires determining an estimator of the error variance in a PLM.

2.2. **Error variance estimators.** In this subsection, we present the estimators that we will use for constructing our test statistics. For the construction of error variance estimators and test statistics, we distinguish between two types of PLMs. For homogeneous errors, the considered PLM is

$$Y_i = \beta' Z_i + g(U_i) + \xi_i, \quad 1 \leq i \leq n \tag{2.6}$$

We assume that the random vectors $(\xi_1, ..., \xi_n)$ and $\{(Z_1, U_1), ..., (Z_n, U_n)\}$ are independent, and that (Z_i, U_i) are independent and identically distributed random variables.

For heterogeneous errors, the considered PLM is

$$Y_i = \beta' Z_i + g(U_i) + \sigma_i \kappa_i, \quad i = 1, ..., n \tag{2.7}$$

where κ_i are independent and identically distributed with zero mean and variance 1, and (Z_i, U_i) are independent and identically distributed random variables. In this work, we assume that σ_i are functions depending on the nonlinear covariates.

2.2.1. *Atypical Values Withdrawal Procedure.* We aim to suggest estimators of the error variance that are constructed without accounting for potential outliers. [7] introduced an outlier removal procedure for the covariate set in the context of a single-index model. In this work, we adopt the same procedure within the framework of the PLM. It must be pointed out that, in a partially linear model, the covariates are split into two groups: linear covariates and nonlinear covariates. The outlier removal procedure will therefore be applied to both sets. The steps are as follows.

Step 1

$Z_{ik} \in \mathbb{R}$, $k = 1, ..., p$, is the k-th component of the vector Z_i, and $U_{it} \in \mathbb{R}$, $t = 1, ..., q$, is the t-th component of the vector U_i.

Suppose that Z_{ik} has support S_{Z_k} and U_{it} has support S_{U_t}. Let c_{z_k} and r_{z_k} be the center and the radius of the set S_{Z_k}, respectively. Similarly, let c_{u_t} and r_{u_t} be the center and the radius of the set S_{U_t}, respectively. We will use a linear and a nonlinear covariate for constructing our estimators of σ_i^2 if and only if Z_{ik} and U_{it} are not too far from the centers c_{z_k} and c_{u_t}, for all $i = 1, ..., n$. To achieve this, we set a constant $\rho < 1$ as a control parameter for the radius r_{z_k} and r_{u_t}. Let $\rho_{z_k} = \rho r_{z_k}$ and $\rho_{u_t} = \rho r_{u_t}$, then we obtain the k-th component of the linear covariate as

$$\delta_k = [c_{z_k} - \rho_{z_k}; c_{z_k} + \rho_{z_k}]$$

and the t-th component of the nonlinear covariate as

$$\vartheta_t = [c_{u_t} - \rho_{u_t}; c_{u_t} + \rho_{u_t}].$$

Step 2

For each $1 \le i \le n$, we check if $(Z_i, U_i) \in \delta \times \vartheta$, where $\delta = \prod_{k=1}^{p} \delta_k$ and $\vartheta = \prod_{t=1}^{q} \vartheta_t$. If for any i, $Z_i \in \delta$ and $U_i \notin \vartheta$, or $Z_i \notin \delta$ and $U_i \in \vartheta$, or $Z_i \notin \delta$ and $U_i \notin \vartheta$, then the pair (Z_i, U_i) is not considered in the process of constructing the error variance estimator. Let us denote $(\widetilde{Z}_i, \widetilde{U}_i)$ as the pairs of covariates belonging to $\delta \times \vartheta$, then we obtain:

for homogeneous errors:

$$\widetilde{Y}_i = \beta' \widetilde{Z}_i + g(\widetilde{U}_i) + \varepsilon_i, \quad 1 \le i \le m, \tag{2.8}$$

for heterogeneous errors:

$$\widetilde{Y}_i = \beta' \widetilde{Z}_i + g(\widetilde{U}_i) + \sigma_i(\widetilde{U}_i)\kappa_i, \quad 1 \le i \le m. \tag{2.9}$$

2.2.2. *MSE-type estimators for the heterogeneous errors.* The estimator $\widehat{\sigma^2_{p'3}}(U_i)$ presented below was constructed without removing outliers from the different sets of covariates.

$$\widehat{\sigma^2_{p'3}}(U_i) = \frac{\sum_{j=1}^{n} \left(Y_j - \widehat{\beta}'_{wls} Z_j - \widehat{g}(U_i)\right)^2 K\left(\frac{U_j - U_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{U_j - U_i}{h_n}\right)},$$

where

$$K\Big(\frac{U_j - U_i}{h_n}\Big) = \prod_{t=1}^{q} K_t\Big(\frac{U_{jt} - U_{it}}{h_n}\Big). \tag{2.10}$$

K_t is a one-dimensional kernel, $h_n \longrightarrow 0$ and $nh_n \longrightarrow +\infty$, as $n \longrightarrow +\infty$.

Remark 2.2. The estimator we propose is a generalization of the estimator proposed by [5], where σ_i are functions depending on the nonlinear covariates.

The estimator $\widehat{\sigma^2_{p3}}(\widetilde{U}_i)$ presented below was constructed after removing outliers from the different sets of covariates.

$$\widehat{\sigma^2_{p3}}(\widetilde{U}_i) = \frac{\sum_{j=1}^{m} \big(\widetilde{Y}_j - \widetilde{\beta}'_{wls}\widetilde{Z}_j - \widetilde{g}(\widetilde{U}_i)\big)^2 K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)}{\sum_{j=1}^{m} K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)},$$

where

$$\widetilde{g}(\widetilde{U}_i) = \begin{cases} \dfrac{\sum_{j=1}^{m} \big(\widetilde{Y}_j - \widetilde{\beta}'_{wls}\widetilde{Z}_j\big) K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)}{\sum_{j=1}^{m} K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)} & if \ \sum_{j=1}^{m} K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big) \neq 0 \\ 0 & else \end{cases} \tag{2.11}$$

and

$$K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big) = \prod_{t=1}^{q} K_t\Big(\frac{\widetilde{U}_{jt} - \widetilde{U}_{it}}{h_m}\Big). \tag{2.12}$$

K_t is a one-dimensional kernel, $h_m \longrightarrow 0$ and $mh_m \longrightarrow +\infty$, as $m \longrightarrow +\infty$.

Remark 2.3. The estimator we propose is a generalization of the estimator proposed by [1].

Remark 2.4. The asymptotic properties of the above estimators of the error variance were established in [2].

2.2.3. *U-statistic estimators for the homogeneous errors.* In this subsection, the proposed estimators are constructed based on Neumann's difference-based method [8].
For the first estimator, we consider the model (2.8). We obtain

$$\widetilde{Y}_i - \beta'\widetilde{Z}_i = g(\widetilde{U}_i) + \varepsilon_i.$$

Let $S_i = \widetilde{Y}_i - \beta'\widetilde{Z}_i$, then the model becomes:

$$S_i = g(\widetilde{U}_i) + \varepsilon_i.$$

We define $\widetilde{S}_i = \widetilde{Y}_i - \widetilde{\beta}'\widetilde{Z}_i$, where $\widetilde{\beta}$ is similar to $\widehat{\beta}$ defined in (2.2). We order the values $\widetilde{S}_i$, $1 \le i \le m$, such that $\widetilde{S}_1 < \widetilde{S}_2 < ... < \widetilde{S_m}$. We propose the following U-statistic estimator:

$$\widehat{\nu_1} = \frac{\sum_{i=1}^{m} \sum_{i \neq j} \big(\widetilde{S}_i - \widetilde{S}_j\big)^2 K\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)}{2\sum_{i=1}^{m} \sum_{i \neq j} K\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)} \tag{2.13}$$

where $\widetilde{S}_l = \widetilde{Y}_l - \widetilde{\beta}'\widetilde{Z}_l, \quad l \in \{i, j\}$ and $K(.)$ is a multidimensional kernel defined in (2.12).

The second U-statistic estimator is a generalization of the one proposed by [1]. We define the estimator as:

$$\widehat{\nu_2} = \frac{\sum_{i=1}^{n}\sum_{i\neq j}\left(\widehat{S_i}-\widehat{S_j}\right)^2 K\left(\frac{U_i-U_j}{h_n}\right)}{2\sum_{i=1}^{n}\sum_{i\neq j}K\left(\frac{U_i-U_j}{h_n}\right)} \tag{2.14}$$

where $\widehat{S_l} = Y_l - \widehat{\beta}' Z_l, \quad l \in \{i,j\}$, and $K(.)$ is a multidimensional kernel defined in (2.10).

Remark 2.5.

- $\widehat{\nu_1}$ is the U-statistic estimator obtained by removing atypical values from the sets of covariates.
- $\widehat{\nu_2}$ is the U-statistic estimator obtained by not having removed atypical values from the sets of covariates.
- The asymptotic properties of these estimators were established in [2].

2.2.4. *Assumptions.* We shall make the following technical assumptions, which will be used to establish the results of the above theorems.

- (A_1) The weights functions $Q_{mi}(\widetilde{U}_j) = \dfrac{K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)}{\sum_{i=1}^{m}K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)}$ and $Q_{ni}(U_j) = \dfrac{K\left(\frac{U_i-U_j}{h_n}\right)}{\sum_{i=1}^{n}K\left(\frac{U_i-U_j}{h_n}\right)}$, satisfying $\sup_{1\leq i\leq m}|Q_{mi}(\widetilde{U}_j)| = O(b_m)$ and $\sup_{1\leq i\leq n}|Q_{ni}(U_j)| = O(b_n)$.
- (A_2) $g(.)$ is Lipschit continuous, $\sup_{1\leq i\leq m}|g(\widetilde{U}_i) - \widetilde{g}(\widetilde{U}_i)| = O(c_m)$ and $\sup_{1\leq i\leq n}|g(U_i) - \widehat{g}(U_i)| = O(c_n)$
- (A_3) *i)* The errors ε_i and ξ_i have moments of at least fourth order.
  ii) $\max_{1\leq i\leq m}|\varepsilon_i| < m^s$, and $\max_{1\leq i\leq n}|\xi_i| < n^s$, $s\in\mathbb{R}$
  iii) $\sum_{i=1}^{m}|\varepsilon_i| = O(\log\log(m))$ and $\sum_{i=1}^{n}|\xi_i| = O(\log\log(n))$
- (A_4) *i)* $\max_{1\leq i\leq m}\|\widetilde{Z}_i\| \leq \lambda_0$ and $\max_{1\leq i,j\leq m}\|\widetilde{Z}_i - \widetilde{Z}_j\| \leq \lambda$, where $\lambda_0, \lambda \in \mathbb{R}_+$
  ii) $\max_{1\leq i\leq n}\|Z_i\| \leq \lambda_1$ and $\max_{1\leq i,j\leq n}\|Z_i - Z_j\| \leq \lambda_2$, where $\lambda_1, \lambda_2 \in \mathbb{R}_+$
- (A_5) ψ and ϕ are the density functions of U_i and $U_i - U_j$ respectively.
- (A_6) The domains S_Z of Z and S_U of U are compact.
- (A_7) $K(.)$ is Lipschitz continuous of $[-1:1]^q$, $\sup_u|K(u)| \leq K_{max}$ where K_{max} is a positive real number, $\|K\|_2^2 < \infty$

$b_m = m^{-2k/(2k+q)}\log^{-1/(2k+q)}(m)$, $c_m = m^{-k/(2k+q)}\log(m)$, and $b_n = n^{-2k/(2k+q)}\log^{-1/(2k+q)}(n)$, $c_n = n^{-k/(2k+q)}\log(n)$ where k is the order of differentiability of $g(.)$.

2.3. **Test.** Our test hypotheses are

$$H_0 : f(X_i) = \beta' Z_i + g(U_i)$$

against

$$H_1 : f(X_i) \neq \beta' Z_i + g(U_i).$$

As mentioned in the introduction, if σ_i^2 is known, our test statistic would be in the form

$$T = \sqrt{n}\left(\widehat{\sigma_i^2} - \sigma_i^2\right)$$

where $\widehat{\sigma_i^2}$ is an estimator of the error variance in a PLM.

For homogeneous errors, the error variance is constant. Let us denote $\sigma_i^2 = \sigma^2$.

- If we have removed outliers from the covariate sets, the test statistic will be
$$T_1 = \sqrt{m}\big(\widehat{\nu_1^2} - \sigma^2\big).$$
- If we have not removed outliers from the covariate sets, the test statistic will be
$$T_2 = \sqrt{n}\big(\widehat{\nu_2^2} - \sigma^2\big).$$

For heterogeneous errors,

- If we have removed outliers from the covariate sets, the test statistic will be
$$T_3 = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\frac{T_{3i}}{\sqrt{\omega}_i}$$
where
$$T_{3i} = \sqrt{m}\big[\widehat{\sigma_{p3}^2}(\widetilde{U}_i) - \sigma_j^2(\widetilde{U}_i)\big].$$
- If we have not removed outliers from the covariate sets, the test statistic will be
$$T_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{4i}}{\sqrt{\omega}_i}$$
where
$$T_{4i} = \sqrt{n}\big[\widehat{\sigma_{p'3}^2}(U_i) - \sigma_j^2(U_i)\big].$$

In the theorems below, we establish that under the null hypothesis H_0, when m or n tends to infinity, the test statistics converge in distribution to a mean-zero normal distribution. Under the alternative hypothesis H_1, we establish that, in absolute value, the test statistics take on increasingly large values as m or n tends to infinity.

Theorem 2.6. *Let assumptions (A_1) - (A_7) be satisfied, we have under H_0*
$$T_1 \longrightarrow N\left(0, \pi\right),$$
where
$$\pi = \frac{1}{\mathbb{E}^2\big[\sum_{i\neq j} K\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)\big]} Var(\varepsilon_i^2)\mathbb{E}\Big[\Big(\sum_{i\neq j} K\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)\Big)^2\Big]$$
and, under H_1, $|T_1|$ diverges to infinity at the rate of $\sqrt{m}$.

Theorem 2.7. *Let assumptions (A_1) - (A_7) be satisfied, we have under H_0*
$$T_2 \longrightarrow N\left(0, \pi\right)$$
where
$$\pi = \frac{1}{\mathbb{E}^2\big[\sum_{i\neq j} K\Big(\frac{U_i - U_j}{h_n}\Big)\big]} Var(\xi_i^2)\mathbb{E}\Big[\Big(\sum_{i\neq j} K\Big(\frac{U_i - U_j}{h_n}\Big)\Big)^2\Big]$$
and, under H_1, $|T_2|$ diverges to infinity at the rate of $\sqrt{n}$.

Theorem 2.8. *Let* $\tau_{ij} = \sigma_j(\widetilde{U}_i)\kappa_j$, $s = \mathbb{E}\big[\tau_{ij}^2 K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)\big]$, $t = \mathbb{E}\big[K\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)\big]$, *and* $\zeta = \mathbb{E}(\tau_{ij}^4)$.
Suppose that assumptions (A_1) - (A_7) *are fulfilled. Then, we have under* H_0,

$$T_3 \longrightarrow N(0,1)$$

where

$$\omega_i = \frac{1}{t^4}\left(t^2\zeta\mathbb{E}\big[K^2\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)\big] + s^2\mathbb{E}\big[K^2\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)\big] - 2st\sigma_j^2(\widetilde{U}_i)\mathbb{E}\big[K^2\Big(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\Big)\big]\right)$$

and, under H_1, $|T_3|$ *diverges to infinity .*

Theorem 2.9. *Let us define* $\upsilon_{ij} = \sigma_j(U_i)\kappa_j$, $s' = \mathbb{E}\big[\upsilon_{ij}^2 K\Big(\frac{U_j - U_i}{h_n}\Big)\big]$, $t' = \mathbb{E}\big[K\Big(\frac{U_j - U_i}{h_n}\Big)\big]$, *and* $\varsigma = \mathbb{E}(\upsilon_{ij}^4)$.

Suppose that assumptions (A_1) - (A_7) *hold. Then, we have under* H_0,

$$T_4 \longrightarrow N(0,1)$$

where

$$\omega_i = \frac{1}{t'^4}\left(t'^2\varsigma\mathbb{E}\big[K^2\Big(\frac{U_j - U_i}{h_n}\Big)\big] + s'^2\mathbb{E}\big[K^2\Big(\frac{U_j - U_i}{h_n}\Big)\big] - 2s't'\sigma_j^2(U_i)\mathbb{E}\big[K^2\Big(\frac{U_j - U_i}{h_n}\Big)\big]\right)$$

and, under H_1, $|T_4|$ *diverges to infinity.*

In most cases, σ_i^2 is unknown. In such cases, we cannot use $T = \sqrt{n}\big(\widehat{\sigma_i^2} - \sigma_i^2\big)$ as a test statistic. We therefore propose test statistics based on the method used by [6] for the single-index model. We propose the following construction. Let $K_1(.)$ and $K_2(.)$ be two arbitrary kernels chosen for constructing the two estimators of the error variance that compose the test statistic.

When the errors are homogeneous, we have

$$\widehat{\nu_{1l}^2} = \frac{\sum_{i=1}^m \sum_{i\neq j}\big(\widetilde{Y}_i - \widetilde{\beta}'\widetilde{Z}_i - \widetilde{Y}_j + \widetilde{\beta}'\widetilde{Z}_j\big)^2 K_l\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)}{2\sum_{i=1}^m \sum_{i\neq j} K_l\Big(\frac{\widetilde{U}_i - \widetilde{U}_j}{h_m}\Big)}$$

or

$$\widehat{\nu_{2l}^2} = \frac{\sum_{i=1}^n \sum_{i\neq j}\big(Y_i - \widehat{\beta}'Z_i - Y_j + \widehat{\beta}'Z_j\big)^2 K_l\Big(\frac{U_i - U_j}{h_n}\Big)}{2\sum_{i=1}^n \sum_{i\neq j} K_l\Big(\frac{U_i - U_j}{h_n}\Big)},$$

$l \in \{1;2\}$.

For heterogeneous errors, we have

$$\widehat{\sigma_{p'3l}^2}(U_i) = \frac{\sum_{j=1}^n \big(Y_j - \widehat{\beta}'_{wls}Z_j - \widehat{g}(U_i)\big)^2 K_l\Big(\frac{U_j - U_i}{h_n}\Big)}{\sum_{j=1}^n K_l\Big(\frac{U_j - U_i}{h_n}\Big)},$$

or

$$\widehat{\sigma^2_{p3l}}(\widetilde{U}_i) = \frac{\sum_{j=1}^{m}\left(\widetilde{Y}_j - \widetilde{\beta}'_{wls}\widetilde{Z}_j - \widetilde{g}(\widetilde{U}_i)\right)^2 K_l\left(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\right)}{\sum_{j=1}^{m} K_l\left(\frac{\widetilde{U}_j - \widetilde{U}_i}{h_m}\right)},$$

$l \in \{1;2\}$.

The test statistics we propose for the case of homogeneous errors are

- $$T_1 = \sqrt{m}\left(\widehat{\nu^2_{11}} - \widehat{\nu^2_{12}}\right),$$

  when the estimators that compose the test statistic have been constructed by removing atypical values from the sets of covariates,

- $$T_2 = \sqrt{n}\left(\widehat{\nu^2_{21}} - \widehat{\nu^2_{22}}\right),$$

  when the estimators that compose the test statistic have been constructed without removing atypical values from the sets of covariates.

The test statistics we propose for the case of heterogeneous errors are

- $$T_3 = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\frac{T_{3i}}{\sqrt{\omega}_i}$$

  where

  $$T_{3i} = \sqrt{m}\left[\widehat{\sigma^2_{p31}}(\widetilde{U}_i) - \widehat{\sigma^2_{p32}}(\widetilde{U}_i)\right],$$

  when the estimators that compose the test statistic have been constructed by removing atypical values from the sets of covariates.

- $$T_4 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{T_{4i}}{\sqrt{\omega}_i}$$

  where

  $$T_{4i} = \sqrt{n}\left[\widehat{\sigma^2_{p'31}}(U_i) - \widehat{\sigma^2_{p'32}}(U_i)\right],$$

  when the estimators that compose the test statistic have been constructed without removing atypical values from the sets of covariates.

In the theorems below, we have also established that under the null hypothesis H_0, when m or n tends to infinity, the test statistics converge in distribution to a mean-zero normal distribution. Under the alternative hypothesis H_1, we have also established that in absolute value, the test statistics take on increasingly large values as m or n tends to infinity.

Theorem 2.10. *Let assumptions (A_1) - (A_7) be satisfied, we have under H_0,*

$$T_1 \longrightarrow N(0, \pi)$$

where

$$\begin{aligned}\pi &= \frac{1}{\mathbb{E}^2\big[\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]}Var(\varepsilon_i^2)\mathbb{E}\big[\big(\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big)^2\big]\\ &+ \frac{1}{\mathbb{E}^2[\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)]}Var(\varepsilon_i^2)\mathbb{E}\big[\big(\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big)^2\big]\end{aligned}$$

and, under $H_1, |T_1|$ diverges to infinity at the rate of $\sqrt{m}$.

Proof. To show the asymptotic normality, we will rely on the following result used by [7]. Suppose that the random variables S_j have mean s, T_j are independent with mean t and with finite variances respectively. We have:

$$\Big(\sum_{j=1}^m S_j\Big)\,/\,\Big(\sum_{j=1}^m T_j\Big)-s/t=\frac{1}{mt^2}\sum_{j=1}^m\big(tS_j-sT_j\big)+O\big(m^{-1}\big).$$

$$\widehat{\nu_{1l}^2}=\frac{\sum_{i=1}^m\sum_{i\neq j}\big(\widetilde{Y}_i-\widetilde{\beta}'\widetilde{Z}_i-\widetilde{Y}_j+\widetilde{\beta}'\widetilde{Z}_j\big)^2K_l\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)}{2\sum_{i=1}^m\sum_{i\neq j}K_l\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)},$$

let $S_{1ij}=\varepsilon_i^2\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)$, $T_{1ij}=\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)$, $S_{2ij}=\varepsilon_i^2\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)$, $T_{2ij}=\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)$, $s_1=\mathbb{E}\big[\varepsilon_i^2\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]$, $t_1=\mathbb{E}\big[\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]$, $s_2=\mathbb{E}\big[\varepsilon_i^2\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]$, $t_2=\mathbb{E}\big[\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]$, $\sigma^2=\frac{\mathbb{E}\big[\varepsilon_i^2\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]}{\mathbb{E}\big[\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]}$ or $\sigma^2=\frac{\mathbb{E}\big[\varepsilon_i^2\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h}\Big)\big]}{\mathbb{E}\big[\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)\big]}$, $Q_{mi1}(\widetilde{U}_j)=\frac{K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)}{\sum_{i=1}^m\sum_{i\neq j}K_1\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)}$ and $Q_{mi2}(\widetilde{U}_j)=\frac{K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)}{\sum_{i=1}^m\sum_{i\neq j}K_2\Big(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\Big)}$

From proposition (4.1), we obtain:

$$\begin{aligned}T_1 &= \sqrt{m}\left[\sum_{i=1}^m\varepsilon_i^2Q_{mi1}(\widetilde{U}_j)-\sigma^2-\sum_{i=1}^m\varepsilon_i^2Q_{mi2}(\widetilde{U}_j)+\sigma^2\right]\\ &= \sqrt{m}\left[m^{-1}\sum_{i=1}^m\left(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}-\frac{S_{2ij}}{t_2}+\frac{s_2T_{2ij}}{t_2^2}\right)\right]\end{aligned}$$

Let $X_{ij}=\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}$ and $G_{ij}=\frac{S_{2ij}}{t_2}-\frac{s_2T_{2ij}}{t_2^2}$.

$$Var(X_{ij}-G_{ij})=Var(X_{ij})+Var(G_{ij})$$

Applying the central limit theorem, we obtain:

$$\sqrt{m}[m^{-1}\sum_{i=1}^{m}(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}-\frac{S_{2ij}}{t_2}+\frac{s_2T_{2ij}}{t_2^2})] \longrightarrow N(0,Var(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2})+Var(\frac{S_{2ij}}{t_2}-\frac{s_2T_{2ij}}{t_2^2}))$$

hence the result under H_0.
From proposition (4.1), we obtain under H_1

$$\begin{aligned}
T_1 &= \sqrt{m}\big[\sum_{i=1}^{m}\sum_{i\neq j}\varepsilon_i^2 Q_{mi1}(\widetilde{U}_j)-\sum_{i=1}^{m}\sum_{i\neq j}\varepsilon_i^2 Q_{mi2}(\widetilde{U}_j)+\sum_{i=1}^{m}\sum_{i\neq j}\big(f(\widetilde{X}_i)-\widetilde{Z}_i'\beta-g(\widetilde{U}_i)\\
&- f(\widetilde{X}_j)+\widetilde{Z}_j'\beta+g(\widetilde{U}_j)\big)^2 Q_{mi1}(\widetilde{U}_j)-\sum_{i=1}^{m}\sum_{i\neq j}\big(f(\widetilde{X}_i)-\beta'\widetilde{Z}_i-g(\widetilde{U}_i)-f(\widetilde{X}_j)\\
&+ \beta'\widetilde{Z}_j+g(\widetilde{U}_j)\big)^2 Q_{mi2}(\widetilde{U}_j)\big]
\end{aligned}$$

$\sum_{i=1}^{m}\sum_{i\neq j}\varepsilon_i^2 Q_{mi1}(\widetilde{U}_j)\longrightarrow\sigma^2$ and $\sum_{i=1}^{m}\sum_{i\neq j}\varepsilon_i^2 Q_{mi2}(\widetilde{U}_j)\longrightarrow\sigma^2$
Finally we have

$$\begin{aligned}
T_1 &= \sqrt{m}\big[\sum_{i=1}^{m}\sum_{i\neq j}\big(f(\widetilde{X}_i)-\beta'\widetilde{Z}_i-g(\widetilde{U}_i)-f(\widetilde{X}_j)+\beta'\widetilde{Z}_j+g(\widetilde{U}_j)\big)^2 Q_{mi1}(\widetilde{U}_j)\\
&- \sum_{i=1}^{m}\sum_{i\neq j}\big(f(\widetilde{X}_i)-\beta'\widetilde{Z}_i-g(\widetilde{U}_i)-f(\widetilde{X}_j)+\beta'\widetilde{Z}_j+g(\widetilde{U}_j)\big)^2 Q_{mi2}(\widetilde{U}_j)\big]
\end{aligned}$$

hence the fact that under H_1, $|T_1|$ diverges to infinity at the rate of $\sqrt{m}$. □

Theorem 2.11. *Let assumptions (A_1) - (A_7) be satisfied, we have under H_0,*

$$T_2 \longrightarrow N(0,\pi)$$

where

$$\begin{aligned}
\pi &= \frac{1}{\mathbb{E}^2\big[\sum_{i\neq j}K_1(\frac{U_i-U_j}{h_n})\big]}Var(\xi_i^2)\mathbb{E}\big[\big(\sum_{i\neq j}K_1\big(\frac{U_i-U_j}{h_n}\big)\big)^2\big]\\
&+ \frac{1}{\mathbb{E}^2\big[\sum_{i\neq j}K_2(\frac{U_i-U_j}{h_n})\big]}Var(\xi_i^2)\mathbb{E}\big[\big(\sum_{i\neq j}K_2\big(\frac{U_i-U_j}{h_n}\big)\big)^2\big]
\end{aligned}$$

and, under H_1, $|T_2|$ diverges to infinity at the rate of $\sqrt{n}$.

Proof. Similar to that of Theorem (2.10). □

Theorem 2.12. *Let $\tau_{ij}=\sigma_j(\widetilde{U}_i)\kappa_j$, $s_l=\mathbb{E}[\tau_{ij}K_l(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m})]$, $t_l=\mathbb{E}[K_l(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m})]$, with $l=1,2,$ and $\zeta=\mathbb{E}(\tau_{ij}^4)$.*
Suppose that assumptions (A_1) - (A_7) are fulfilled. Then, we have under H_0,

$$T_3 \longrightarrow N(0,1)$$

where

$$\omega_i = \frac{1}{t_1^4}\Bigg(t_1^2\zeta\mathbb{E}\big[K_1^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]+s_1^2\mathbb{E}\big[K_1^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]-2s_1t_1\sigma_j^2(\widetilde{U}_i)\mathbb{E}\big[K_1^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]\Bigg)$$
$$+ \frac{1}{t_2^4}\Bigg(t_2^2\zeta\mathbb{E}\big[K_2^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]+s_2^2\mathbb{E}\big[K_2^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]-2s_2t_2\sigma_j^2(\widetilde{U}_i)\mathbb{E}\big[K_2^2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]\Bigg)$$

and, under $H_1, |T_3|$ *diverges to infinity.*

Proof. Let $S_{1ij}=\tau_{ij}^2K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)$, $T_{1ij}=K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)$, $S_{2ij}=\tau_{ij}^2K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)$ $T_{2ij}=K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)$, $s_1=\mathbb{E}\big[\tau_{ij}^2K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]$, $t_1=\mathbb{E}\big[K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]$, $s_2=\mathbb{E}\big[\tau_{ij}^2K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]$, $t_2=\mathbb{E}\big[K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)\big]$,,

$$Q_{mj1}(\widetilde{U}_i)=\frac{K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)}{\sum_{j=1}^m K_1\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)} \text{ and } Q_{mj2}(\widetilde{U}_i)=\frac{K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)}{\sum_{j=1}^m K_2\Big(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\Big)}$$

From proposition (4.1), we obtain:

$$\begin{aligned} T_{3i} &= \sqrt{m}\big[\sum_{j=1}^m \tau_{ij}^2Q_{mj1}(\widetilde{U}_i)-\sigma_i^2-\sum_{j=1}^m\tau_{ij}^2Q_{mj2}(\widetilde{U}_i)+\sigma_i^2\big]\\ &= \sqrt{m}\big[\frac{1}{m}\sum_{j=1}^m\big(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}-\frac{S_{2ij}}{t_2}+\frac{s_2T_{2ij}}{t_2^2}\big)\big]\end{aligned}$$

Let $X_j=\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}$ and $G_j=\frac{S_{2ij}}{t_2}-\frac{s_2T_{2ij}}{t_2^2}$.

$$Var(X_j-G_j)=Var(X_j)+Var(G_j)$$

Applying the central limit theorem, we obtain

$$\sqrt{m}\Big[\frac{1}{m}\sum_{j=1}^m\big(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}-\frac{S_{2ij}}{t_2}+\frac{s_2T_{2ij}}{t_2^2}\big)\Big]\longrightarrow N\Big(0,Var\big(\frac{S_{1ij}}{t_1}-\frac{s_1T_{1ij}}{t_1^2}\big)+Var\big(\frac{S_{2ij}}{t_2}-\frac{s_2T_{2ij}}{t_2^2}\big)\Big)$$

$$T_{3i}\longrightarrow N(0,\omega_i)$$

hence the result under H_0.
From proposition (4.1), we obtain under H_1

$$\begin{aligned} T_{3i} &= \sqrt{m}\big[\sum_{j=1}^m\tau_{ij}^2Q_{mj1}(\widetilde{U}_i)-\sum_{j=1}^m\tau_{ij}^2Q_{mj2}(\widetilde{U}_i)+\sum_{j=1}^m\big(f(\widetilde{X}_j)\\ &- \widetilde{Z}_j'\beta-g(\widetilde{U}_j)\big)^2Q_{mj1}(\widetilde{U}_i)-\sum_{j=1}^m\big(f(\widetilde{X}_j)-\beta'\widetilde{Z}_j-g(\widetilde{U}_j)\big)^2Q_{mj2}(\widetilde{U}_i)\big]\end{aligned}$$

$\sum_{j=1}^m\tau_{ij}^2Q_{mj1}(\widetilde{U}_i)\longrightarrow\sigma_i^2$ and $\sum_{j=1}^m\tau_{ij}^2Q_{mj2}(\widetilde{U}_i)\longrightarrow\sigma_i^2$
Finally we have

$$T_{3i}=\sqrt{m}[\sum_{j=1}^m(f(\widetilde{X}_j)-\beta'\widetilde{Z}_j-g(\widetilde{U}_j))^2Q_{mj1}(\widetilde{U}_i)-\sum_{j=1}^m(f(\widetilde{X}_j)-\beta'\widetilde{Z}_j-g(\widetilde{U}_j))^2Q_{mj2}(\widetilde{U}_i)]$$

hence the fact that under H_1, $|T_3|$ diverges to infinity. $\square$

Theorem 2.13. *Let* $\upsilon_{ij} = \sigma_j(U_i)\kappa_j$, $s'_l = \mathbb{E}[\upsilon_{ij} K_l(\frac{U_j - U_i}{h_n})]$, $t'_l = \mathbb{E}[K_l(\frac{U_j-U_i}{h_n})]$, *with* $l = 1, 2$, *and* $\varsigma = \mathbb{E}(\upsilon_{ij}^4)$.
Suppose that assumptions (A_1) - (A_7) *are fulfilled. Then, we have under* H_0,

$$T_4 \longrightarrow N(0,1)$$

where

$$\begin{aligned}\omega_i &= \frac{1}{t_1'^4}\Big(t_1'^2\varsigma\mathbb{E}\big[K_1^2\Big(\frac{U_j-U_i}{h_n}\Big)\big] + s_1'^2\mathbb{E}\big[K_1^2\Big(\frac{U_j-U_i}{h_n}\Big)\big] - 2s_1't_1'\sigma_j^2(U_i)\mathbb{E}\big[K_1^2\Big(\frac{U_j-U_i}{h_n}\Big)\big]\Big) \\ &+ \frac{1}{t_2'^4}\Big(t_2'^2\varsigma\mathbb{E}\big[K_2^2\Big(\frac{U_j-U_i}{h_n}\Big)\big] + s_2'^2\mathbb{E}\big[K_2^2\Big(\frac{U_j-U_i}{h_n}\Big)\big] - 2s_2't_2'\sigma_j^2(U_i)\mathbb{E}\big[K_2^2\Big(\frac{U_j-U_i}{h_n}\Big)\big]\Big)\end{aligned}$$

and, under $H_1, |T_4|$ *diverges to infinity.*

Proof. As in Theorem (2.12). $\square$

Thus, in the case of homogeneous errors, our test would be to reject the null hypothesis if $\left\{\dfrac{|T_k|}{\sqrt{\widehat{\pi}}} > q_{\alpha/2}\right\}$ where $q_{\alpha/2}$ denote the quantile of order $\alpha/2$ of the standard normal distribution, $\widehat{\pi}$ is an estimator of π, and $k = 1, 2$.
For the case of heterogeneous errors, our test would be to reject the null hypothesis if $\left\{\dfrac{1}{\sqrt{m}}\sum_{i=1}^m \dfrac{|T_{3i}|}{\sqrt{\widehat{\omega}_i}} > q_{\alpha/2}\right\}$ or $\left\{\dfrac{1}{\sqrt{n}}\sum_{i=1}^n \dfrac{|T_{4i}|}{\sqrt{\widehat{\omega}_i}} > q_{\alpha/2}\right\}$, where $q_{\alpha/2}$ denote the quantile of order $\alpha/2$ of the standard normal distribution, $\widehat{\omega}_i$ is an estimator of ω_i.

3. Simulations

The simulations were carried out on the assumption that the errors were homogeneous. The estimators that make up the test statistic were constructed by taking all data into account.
The simulations were based on finite sample sizes (n=50, n=100). The linear covariates are of dimension 2, 4, 6 and 10, each component of these covariates is generated according to the uniform distribution $U[0;1]$. The nonlinear covariates are of dimension 1 and are also generated according to the uniform distribution $U[0;1]$. For dimension 2, 4, 6 and 10 we have taken respectively $\beta = (1, 1.5)'$, $\beta = (1, 1.5, 0.75, \sqrt{2})'$, $\beta = (1, 1.5, 0.75, \sqrt{2}, 1, 0.5)'$ and $\beta = (1, 1.5, 0.75, \sqrt{2}, 1, 0.5, \sqrt{3}, 1.75, 1.25, 0.25)'$. The ξ_i errors have zero mean and variance $\sigma^2 = 0.09$. For the test statistic, we used the Gaussian kernel defined by $K_1(x) = \exp(-(x)^2/2)/\sqrt{2\pi}$ and the Epanechnikov kernel defined by $K_2(x) = 0.75(1 - x^2)I(|x| \leq 1)$.
In Tables (1) and (2), under H_0 we have used $g_1(u) = \sin(2u)$, $g_2(u) = \sin(3u)$, $g_3(u) = \sin(\pi u)$ and $g_4(u) = \sin(0.5\pi u)$, for the nonparametric part. In Tables (3) and (4) under H_1 we have used $f_1(u) = \beta' z + \exp(2\pi u) + \sin(u)$, $f_2(u) = \beta' z + \exp(2\pi u) + \cos(u)$, $f_3(u) = \beta' z + 7\exp(\pi u)$ and $f_4(u) = \beta' z + 7\exp(\pi u) + sin(u)$.

Table 1. Values of the test statistic under H_0

n	dim	T.S	g_1	g_2	g_3	g_4
50	2	$\|T\|$	0.0008	0.0008	0.0008	0.0014
50	4	$\|T\|$	0.0014	0.0017	0.0015	0.0015
50	6	$\|T\|$	0.0024	0.0016	0.0011	0.0022
50	10	$\|T\|$	0.0036	0.0038	0.0044	0.0042

Table 2. Values of the test statistic under H_0

n	dim	T.S	g_1	g_2	g_3	g_4
100	2	$\|T\|$	0.0010	0.0018	0.0031	0.0028
100	4	$\|T\|$	0.0010	0.0026	0.0014	0.0011
100	6	$\|T\|$	0.0001	0.0009	0.0012	0.0029
100	10	$\|T\|$	0.0028	0.00008	0.0006	0.0001

Table 3. Values of the test statistic under H_1

n	dim	T.S	f_1	f_2	f_3	f_4
50	2	$\|T\|$	3635.01	3891.58	323.05	340.08
50	4	$\|T\|$	3812.13	3877.82	348.59	311.32
50	6	$\|T\|$	3507.62	3646.12	319.04	335.25
50	10	$\|T\|$	3836.64	3794.58	318.58	328.80

Table 4. Values of the test statistic under H_1

n	dim	T.S	f_1	f_2	f_3	f_4
100	2	$\|T\|$	5117.00	5155.42	432.81	415.45
100	4	$\|T\|$	4981.27	5026.14	417.04	431.80
100	6	$\|T\|$	5321.38	5202.54	430.68	425.17
100	10	$\|T\|$	5260.69	5121.33	424.20	439.76

In Tables (1) and (2) under H_0 the values of the test statistic are close to zero despite the fact that the dimension of the linear covariates is increases. Therefore, the dimension of the linear covariate does not adversely affect the performance of the proposed test statistic. The choice of kernels does not affect the consistency of the estimators of the error variance that make up the test statistic. We can affirm that under H_0, whatever the dimension of the linear covariates, and the kernels chosen for the construction of the estimators of the error variance that make up the test statistic, the test statistic will provide us with sufficient information regarding the use of a PLM.
In Tables (3) and (4), under H_1, the dimension of the linear covariates does not affect the performance of the test statistic. In fact, despite that the dimension of the covariates increases, the test statistic takes on a large value. Under H_1 the choice of kernel has an adverse effect on the values taken by the estimators of the error variance that make up the test statistic, hence the fact that the test statistic takes on a large value.
The test statistic detects whether or not a PLM is used, whatever the dimension of the linear covariates.

Remark 3.1.
- Simulation results confirm that, under the null hypothesis H_0, the test statistique T converges in distribution to a centered normal distribution as n tends to infinity, whereas under the alternative hypothesis H_1, $|T|$ diverges to infinity.
- For our simulations we used the PLRModels package of the R software.

4. Conclusion

In this paper, we proposed a test for assessing the adequacy of fitting data to a PLM. The construction of our test statistics relies on an estimate of the error variance, for which we used a generalization of the estimators proposed by [1]. Under the null hypothesis H_0, we established the asymptotic normality of our test statistics. Under the alternative hypothesis H_1, we showed that they tend to take large absolute values, which makes it possible to consistently detect deviations from the PLM structure. The simulation results presented in tables 1 through 4 support our theoretical findings. They also show that the dimension of the linear covariates has no impact on the performance of our test statistics under either H_0 or H_1.

Based on the results of the test, the data appears to follow a PLM. However, it is possible that the dimension of the nonlinear covariates is greater than or equal to three. In such cases, it would be relevant to develop a test to assess whether the nonlinear component can be approximated by a single-index or an additive regression function. This remains an open problem.

.

References

1. Ahoutou. B, N, M., Yodé, A, F, E. & Kouakou K, F., (2023). Error variance estimation for the partially linear model. Afrika Statistika, 18(1), 3399- 3428. https://doi.org/10.16929/as/2023.3399.311
2. Ahoutou. B, N, M., Yodé, A, F, E. & Kouakou K, F., (2024). Comparative study of residual variance estimators constructed without or with the presence of atypical values: case of the partially linear model. Far East Journal of Theoretical Statistics, 68(3), 335- 372. https://doi.org/10.17654/0972086324019
3. Fan, J. (1992). Design-adaptive nonparametric regression. J. Amer. Statist. Assoc. 87(420), 998-1004. https://doi.org/10.2307/2290637
4. Fan, Y. & Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms, Econometrica, 64(4), 865-890. https://doi.org/10.2307/2171848
5. Härdle, W., Liang, H., & Gao,J.(2000). Partially linear models. Contributions to Statistics. Physica-Verbag, Heidelberg. https://doi.org/10.1007/978-3-642-57700-0
6. Kulasekera, K.B & Lin, W. (2005). Error variance estimation and testing for the single-index model. Technical Report TR2005_04_LK, Department of Mathematical Sciences, Clemson University. https:// www.math.clemson.edu/reports/TR2005_04_LK.pdf.
7. Kulasekera, K.B & Lin, W. (2010). Error variance estimation for the single-index model, Australian and New Zealand Journal of Statistics. 52(2), 201-219. https://doi.org/10.1111/j.1467-842X.2010.00575.x
8. Neumann, V, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. Ann. Mathemat. Statist, 12(4), 367-395. https://doi.org/10.1214/aoms/1177731677
9. Sun, Z. Chen, F. Liang, H & Ruppert, D. (2021). A projection-based consistent test incorporating dimension-reduction in partially linear models. Statistica Sinica, 31, 1489-1508. https://doi.org/10.5705/ss.202019.0238
10. Whang, Y-J. & Andrews, W, K, D., (1993). Tests of specification for parametric and semiparametric models, Journal of Econometrics, 57, 277-318. https://doi.org/10.1016/0304-4076(93)90068-G
11. Yatchew, A., (1997). An elementary estimator of the partial linear model. Economics Letters, 57(2), 135-143. https://doi.org/10.1016/S0165-1765(97)00218-8
12. Zhu, L, X. & Ng, K, W., (2003). Checking the adequacy of a partial linear model. Statistica Sinica, 13(3), 763-781. https://www.jstor.org/24307123

Appendix

Proposition 4.1. *Let assumptions (A_1) - (A_4) and (A_6) be satisfied, we have under H_0,*

$$\begin{aligned}\widehat{\sigma^2_{p3}}(\widetilde{U}_i) &= \frac{\sum_{j=1}^m \tau_{ij}^2 K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)}{\sum_{j=1}^m K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)} + O(b_m) + O(c_m^2 b_m) + O(m^{-1/2}c_m b_m)\\ &+ +O(\log\log(m)c_m b_m) + O(m^{-1/2}c_m),\end{aligned}$$

and

$$\begin{aligned}\widehat{\sigma^2_{p'3}}(U_i) &= \frac{\sum_{j=1}^n \upsilon_{ij}^2 K\left(\frac{U_j-U_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{U_j-U_i}{h_n}\right)} + O(b_n) + O(c_n^2 b_n) + O(c_n b_n n^{-1/2})\\ &+ +O(c_n b_n \log\log(n)) + O(c_n n^{-1/2}),\end{aligned}$$

under H_1,

$$\begin{aligned}
\widehat{\sigma^2_{p3}}(\widetilde{U}_i) &= \frac{\sum_{j=1}^m \tau_{ij}^2 K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)}{\sum_{j=1}^m K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)} + \frac{\sum_{j=1}^m \left[f(\widetilde{X}_j) - \beta'\widetilde{Z}_j - g(\widetilde{U}_j)\right]^2 K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)}{\sum_{j=1}^m K\left(\frac{\widetilde{U}_j-\widetilde{U}_i}{h_m}\right)} \\
&+ O(b_m) + O(c_m^2 b_m) + O(c_m b_m) + O(b_m \log\log(m)) + O(m^{-1/2} b_m) + O(b_m c_m) \\
&+ O(\log\log(m) c_m b_m) + O(m^{-1/2} c_m)
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\sigma^2_{p'3}}(U_i) &= \frac{\sum_{j=1}^n \upsilon_{ij}^2 K\left(\frac{U_j-U_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{U_j-U_i}{h_n}\right)} + \frac{\sum_{j=1}^n \left[f(X_j) - \beta' Z_j - g(U_j)\right]^2 K\left(\frac{U_j-U_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{U_j-U_i}{h_n}\right)} \\
&+ O(b_n) + O(c_n^2 b_n) + O(c_n b_n) + O(\log\log(n) b_n) + O(b_n n^{-1/2}) + O(\log\log(n) c_n b_n) \\
&+ O(n^{-1/2} c_n) + O(b_n c_n)
\end{aligned}$$

Proof. Similar to that of [1]. □

Proposition 4.2. *Let* $\boldsymbol{M} = \sum_{i=1}^m \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)$, $a_m = \sigma\sqrt{2Ch_m^q m} + 2/3 K_{max} m^{s+1}\varsigma$ *and* $d_m = \sqrt{2C_2 h_m^q m} + 2/3 m^2 B\zeta$ *where* $C = \| \prod_{l=1}^q K_l\|_2^2 \|\phi_l\|_\infty$, $B = K_{max} + h_m^q |C_1|$, $C_2 = \int \left(\sum_{i\neq j} \prod_{l=1}^q K_l(v_l)\right)^2 \phi_l(v_l h_m) dv_l$ *and* $C_1 = \int \prod_{l=1}^q K_l(v_l)\phi_l(v_l h_m) dv_l$
Suppose that assumptions (A_2) - (A_7), *are fulfilled, then*
under H_0, *we have*

$$\begin{aligned}
\widehat{\nu_1} &= \frac{\sum_{i=1}^m \varepsilon_i^2 \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)}{\sum_{i=1}^m \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)} + [O(m^2 h_m^q) + o(d_m)]^{-1}[O(a_m \log\log(m)) + O(c_m^2 d_m) \\
&+ O(m^{-1} d_m) + O(a_m c_m) + O(m^{-1/2} a_m) + O(m^{-1/2} c_m d_m)],
\end{aligned}$$

under H_1, *we have*

$$\begin{aligned}
\widehat{\nu_1} &= \frac{\sum_{i=1}^m \varepsilon_i^2 \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)}{\sum_{i=1}^n \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)} + \frac{\sum_{i=1}^m \sum_{i\neq j} \left(D(T_i) - D(T_j)\right)^2 \sum_{i\neq j} K\left(\frac{\widetilde{U}_i-\widetilde{U}_j}{h_m}\right)}{2\boldsymbol{M}} \\
&+ [O(m^2 h_m^q) + o(d_m)]^{-1}[O(a_m \log\log(m)) + O(c_m^2 d_m) + O(m^{-1} d_m) + O(a_m c_m) \\
&+ O(a_m m^{-1/2}) + O(c_m d_m m^{-1/2}) + O(c_m) + O(m^{-1/2} d_m) + O(a_m)]
\end{aligned}$$

Proof. Similar to that of [1]. □

Proposition 4.3. *Let* $\boldsymbol{M} = \sum_{i=1}^n \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)$, $a_n = \sigma\sqrt{2Ch_n^q n} + 2/3 K_{max} n^{s+1}\varsigma$ *and* $d_n = \sqrt{2C_2 h_n^q n} + 2/3 n^2 B\zeta$ *where* $B = K_{max} + h_n^q |C_1|$ $C = \| \prod_{l=1}^q K_l\|_2^2 \|\phi_l\|_\infty$, $C_1 = \int \prod_{l=1}^q K_l(v_l)\phi_l(v_l h_n) dv_l$ *and* $C_2 = \int \left(\sum_{i\neq j} \prod_{l=1}^q \mathsf{K}(v_l)\right)^2 \phi_l(v_l h_n) dv_l$
Suppose that assumptions (A_2) - (A_7), *then*

under H_0, we have

$$\begin{aligned}
\widehat{\nu_2} &= \frac{\sum_{i=1}^n \xi_i^2 \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)}{\sum_{i=1}^n \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)} + [O(n^2h_n^q)+o(d_n)]^{-1}[O(\log\log(n)a_n)+O(c_n^2 d_n)\\
&+ O(n^{-1}d_n)+O(a_nc_n)+O(n^{-1/2}a_n)+O(n^{-1/2}c_nd_n)],
\end{aligned}$$

under H_1, we obtain

$$\begin{aligned}
\widehat{\nu_2} &= \frac{\sum_{i=1}^n \xi_i^2 \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)}{\sum_{i=1}^n \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)} + \frac{\sum_{i=1}^n \sum_{i\neq j} \left(D(T_i)-D(T_j)\right)^2 \sum_{i\neq j} K\left(\frac{U_i-U_j}{h_n}\right)}{2\boldsymbol{M}}\\
&+ [O(n^2h_n^q)+o(d_n)]^{-1}[O(a_n\log\log(n))+O(c_n^2d_n)+O(d_nn^{-1})+O(a_nc_n)\\
&+ O(a_nn^{-1/2})+O(c_nd_nn^{-1/2})+O(a_n)+O(d_nn^{-1/2})+O(c_n)]
\end{aligned}$$

Proof. Similar to that of [1]. □

[1] Department of Mathematics and Computer Science, Félix Houphouët Boigny University , 22 BP 582 Abidjan 22, Ivory Coast.

Email address: ahoutoumartial1@gmail.com kouameflorentk@gmail.com yafevrard@yahoo.fr