IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Study On Effectiveness Of Spearman's Rank Correlation Coefficient Under Generalized Partially Linear Regression

[1]Sthitadhi Das

[1]Assistant Professor

[1]Department of Mathematics,

[1]Brainware University, Kolkata, India

Abstract: Under a generalized partially linear model $Y = X'\underset{\sim}{\Gamma} + g(W) + \mathcal{E}$ with $X = (X_1, \dots, X_p)' \in \mathbb{R}^p$ and $W = (W_1, \dots, W_q)' \in \mathbb{R}^q$ being the parametric and nonparametric regressors respectively and $g(.,\dots,.)$ being a Lipschitz continuous function and $\mathcal{E}$ being random error, we are motivated to test whether (X, W) and $\mathcal{E}$ are independent or not. The test statistics in this regard are developed based on a very traditional nonparametric measure of association Spearman's rank correlation coefficient ρ_s and the test statistics are constructed upon general order differences of the observed and estimated responses, since the independence of the joint regressors and error finally implies the independence between general order difference of Y and that of estimated Y under certain conditions. By defining a properly defined sequence of contiguous alternatives, the test is performed to achieve consistency, and the asymptotic powers of the test statistics are computed for rejection of the null hypothesis suggesting the independence.

Index Terms - Generalized partially linear regression model, Nonparametric regression model, Spearman's ρ_s, Contiguous alternatives, Asymptotic power, V-statistic.

Introduction

In various fields like *Economics*, *Biological sciences*, *Business administration*, the roles of semiparametric regression models are undeniable. Generally, to study a variable of interest, one may consider several variables or factors where some of the factors exhibit some tractable mathematical relationship(s) with the study variable, e.g. a study variable *Score in Mathematics* can be linearly explained by the factors *study hours*, *academic performance in previous exams*, *health condition* etc. But there could be some other factors which cannot explain the study variable in such deterministic way; e.g. the weight of a student hardly influences his/her academic performance overall. It is significantly noticeable that both types of variables, whether delineating the study variable in structural (parametric) as well as nonstructural (nonparametric), must have as much as low correlation between themselves. If the parametric and nonparametric components are highly correlated, it can be difficult to disentangle their separate effects, potentially leading to collinearity problems. Also, some semiparametric estimation techniques viz. profile likelihood, series estimation etc. may perform poorly if there is strong dependence between the parametric and nonparametric regressors. Moreover, when the set of all regressors are used to make prediction(s) about the study variable (or predictor), there should be minimal dependence between the joint regressors and the error component of the model. But in general, the situation may not arise always, as due to the presence of nonparametric regressors the noise in assessing the predictor could be sufficiently high in the underlying model.

Let us begin with a traditional generalized partially linear regression represented as

$$Y = \Gamma_1 X_1 + \cdots + \Gamma_p X_p + g(W_1, \ldots, W_q) + \mathcal{E} = X'\underset{\sim}{\Gamma} + g(W) + \mathcal{E} \qquad (1)$$

where Y is the study variable getting partly explained by p (p≥2) regressors $X = (X_1, \ldots, X_p)'$along with unknown regression coefficients (or parameters) $\underset{\sim}{\Gamma} = (\Gamma_1, \ldots, \Gamma_p)'$ as well as a continuous Lipschitz function $g(.,\ldots,.)$ of q (q≥2) nonparametric regressors $(W_1, \ldots, W_q)'$ with unknown mathematical form. The assumptions on the random error component $\mathcal{E}$ are:

(i) E $(\mathcal{E}|X, W)$=0 for all (X, W) and (ii) E $(\mathcal{E}\text{^}2|X, W)$= $\mathcal{Q}^2(X, W)$. Thereafter, we proceed to conduct a test of association between $X_1, \ldots, X_p$ and $W_1, \ldots, W_q$.

Testing independence between two or more variables is often a common topic of interest in the literature of statistics. In particular, under statistical regression setting, testing non-association between regressor and error is much sought-after issue of same regard. Various analytical aspects on a generalized partially linear model, encompassing estimation of parameters, the nonparametric regression function, and the asymptotic properties of the estimators, were extensively carried out by Robinson (1988) [12], Andrews (1995) [1], Qi Li (2000) [9], Hamilton (1997) [7], Liu et al. (1997) [10] etc. Later, Dhar et al. (2018) [6], Das et al. (2022) [4] proposed a test of association between the sole nonparametric regressor variable and random error in a simple partially linear regression model using popular nonparametric measures of association, namely Kendall's τ, Bergsma (2006) [2]'s τ^*, and Szekely et al. (2007) [13]'s *dcov* in the context of simple nonparametric regression and semiparametric model.

This article ventures out that earlier effort in a more general setup where more than one parametric and non-parametric regressor is involved as described in (1). If the regressors are jointly independent of $\mathcal{E}$, then the homoskedasticity of error is inevitable; otherwise $\mathcal{E}$ is heteroskedastic, as its variance depends on the observed values of the (p+q) regressors.

The entire article is organized as follows. Section 2 provides the estimation procedure of the regression coefficients $\Gamma_1, \ldots, \Gamma_p$ using Robinson's method as well as the function $g(.,\ldots,.)$ using Nadaraya-Watson kernel density estimation method, followed by the development of relevant hypotheses in Section 3. The nondegenerate V statistics based on Spearman's ρ_s are proposed in Section 4 for further progress in the construction of the test procedure. The corresponding asymptotic distributions are derived in Section 5, which leads to the computation of their asymptotic efficiencies and powers. Section 6 presents a real-life data analysis in order to check the utility of the test statistics, by computing p-values, obtained through the resampling technique for different sample sizes.

I. ESTIMATION OF MODEL

For the i-th observation $(i = 1,2, \ldots, n)$, a generalized partially linear regression model possesses the following mathematical structure:

$$Y_i = X_i'\underset{\sim}{\Gamma} + g(W_i) + \mathcal{E}_i \qquad (2)$$

where Y_i is the i-th scalar response, $X_i = (X_{i1}, \ldots, X_{ip})^T$, $W_i = (W_{i1}, \ldots, W_{iq})'$ and $\Gamma_1, \ldots, \Gamma_p$ are p parameters corresponding to the regressors $X_1, \ldots, X_p$. The nonparametric regression function $g(W_1, \ldots, W_q)$, based on nonparametric regressors $W_1, \ldots, W_q$, is a Lipschitz continuous function defined on $\mathbb{R}^q$. To estimate the parameters $\Gamma_1, \ldots, \Gamma_p$ and $g(\cdot, \ldots, \cdot)$, we apply Robinson (1988) [12]'s method. Taking $E(\cdot\,|W_i)$ to both sides of (2), we obtain $E(Y_i|W_i) = E(X_i|W_i)'\underset{\sim}{\Gamma} + g(W_i)$ and subtracting it from (2) we get

$$Y_i - g_Y(W_i) = (X_i - g_X(W_i))'\underset{\sim}{\Gamma} + \mathcal{E}_i \Longrightarrow \mathcal{E}_{Yi} = \mathcal{E}_{Xi}'\underset{\sim}{\Gamma} + \mathcal{E}_i, \text{where}$$

$$Y_i = g_Y(W_i) + \mathcal{E}_{Yi} \qquad (3)$$

$$X_i = g_X(W_i) + \mathcal{E}_{Xi}, i = 1, \ldots, n \qquad (4)$$

$$\Longrightarrow \underset{\sim}{\tilde{\Gamma}} = (\sum_{i=1}^{n} \mathcal{E}_{Xi}\mathcal{E}_{Xi}')^{-1}(\sum_{i=1}^{n} \mathcal{E}_{Xi}\mathcal{E}_{Yi}). \qquad (5)$$

Here, $\underset{\sim}{\hat{\Gamma}}$ is not a feasible estimator of $\underset{\sim}{\Gamma}$. We need to estimate the errors $\mathcal{E}_{Xi}$ and $\mathcal{E}_{Yi}$ for $i = 1, \dots, n$ so that a feasible estimator of $\underset{\sim}{\Gamma}$ can be determined. From models (3) and (4), the estimators of the regression functions $g_Y(\cdot)$ and $g_X(\cdot)$ are determined by applying the Nadaraya-Watson estimation of kernel density (NW) method. Note that,

$$g_Y(\underset{\sim}{w}) \quad = \quad E(Y|W = \underset{\sim}{w}) = \int_{-\infty}^{\infty} y \cdot \frac{\mathcal{G}_{Y,W}(y, \underset{\sim}{w})}{\psi_W(\underset{\sim}{w})} dy \qquad (6)$$

where $f_{Y|W}(\cdot\,|\underset{\sim}{w})$ is the conditional *p.d.f.* of $Y|W$ at $W = \underset{\sim}{w}$, $\mathcal{G}_{Y,W}(\cdot,\cdot)$ is the joint *p.d.f.* of (Y, W) and $\psi_W(\cdot)$ is the *p.d.f.* of W. The kernel density of W is estimated at $\underset{\sim}{w}$ as

$$\hat{\psi}_W(\underset{\sim}{w}) = \frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})\} \qquad (7)$$

where $(W_{i1}, \dots, W_{iq})' \equiv W_i$ for $i = 1, \dots, n$, $k_j(\cdot)$'s are the kernel density functions of W_j's, $j = 1, \dots, q$; $\mathcal{H}_1, \dots, \mathcal{H}_q$ are the bandwidths (>0) for estimation of kernel density functions of $W_1, \dots, W_q$. In similar manner, the joint *p.d.f.* of (Y, W) is estimated as

$$\hat{\mathcal{G}}_{Y,W}(y, \underset{\sim}{w}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathcal{H}_y}k_y(\frac{y - Y_i}{\mathcal{H}_y})\{\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})\} \qquad (8)$$

where $\mathcal{H}_y$ is the bandwidth for estimating the *p.d.f.* of Y. Therefore, $\hat{g}_Y(\cdot)$ is estimated as

$$\hat{g}_Y(\underset{\sim}{w}) \quad = \quad \int_{-\infty}^{\infty} y \cdot \frac{\hat{\mathcal{G}}_{Y,W}(y, \underset{\sim}{w})}{\hat{\psi}_W(\underset{\sim}{w})} dy = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})\}Y_i}{\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})}. \qquad (9)$$

Similarly, $g_X(\underset{\sim}{w}) = E(X|W = \underset{\sim}{w})$ is estimated as

$$\hat{g}_X(\underset{\sim}{w}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})\}X_i}{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q}\frac{1}{\mathcal{H}_j}k_j(\frac{w_j - W_{ij}}{\mathcal{H}_j})\}}.$$

Next, the errors are estimated as $\hat{\mathcal{E}}_{Yi} = Y_i - \hat{g}_Y(W_i)$ and $\hat{\mathcal{E}}_{Xi} = X_i - \hat{g}_X(W_i)$. Then, a feasible estimator of $\underset{\sim}{\Gamma}$ is obtained as

$$\underset{\sim}{\hat{\Gamma}} = (\sum_{i=1}^{n}\hat{\mathcal{E}}_{Xi}\hat{\mathcal{E}}_{Xi}')^{-1}(\sum_{i=1}^{n}\hat{\mathcal{E}}_{Xi}\hat{\mathcal{E}}_{Yi}).$$

Now, we have to estimate the nonparametric regression function $g(\cdot)$. Note that, the semiparametric model can be transformed to a nonparametric regression model as

$$Y_i - X_i'\underset{\sim}{\Gamma} = g(W_i) + \mathcal{E}_i \Rightarrow Y_i' = g(W_i) + \mathcal{E}_i, i = 1, \dots, n$$

where Y_i' is the *transformed response* as defined $(Y_i - X_i'\underset{\sim}{\Gamma})$, $i = 1, \dots, n$. Now, based on the *i.i.d.* observations $(Y_i', W_i)'s, i = 1, \dots, n$, the expression of $g(\underset{\sim}{w})$ is derived as

$$g(\underset{\sim}{w}) = E(Y'|W) = \int_{-\infty}^{\infty} y' \cdot \frac{f_{Y',W}(y', \underset{\sim}{w})}{\psi_W(\underset{\sim}{w})} dy' \qquad (11)$$

where $\phi_{Y'|W}(\cdot\,|\underset{\sim}{w})$ is the *p.d.f.* of $(Y'|W)$ and $f_{Y',W}(\cdot,\cdot)$ is the joint *p.d.f.* of (Y', W). Then,

$$\hat{g}(\underset{\sim}{w}) = \int_{-\infty}^{\infty} y' \cdot \frac{\frac{1}{n}\sum_{i=1}^{n}\{k_{y'}(\frac{y'-Y_i'}{\mathcal{H}_{y'}})\prod_{j=1}^{q} k_{j;W_i}(\frac{w_j-W_{ij}}{\mathcal{H}_j})\}}{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q} k_{j;W_i}(\frac{w_i-W_{ij}}{\mathcal{H}_j})\}} dy'$$

$$= \sum_{i=1}^{n} \mathcal{C}_i Y_i' \text{ where } \mathcal{C}_i = \frac{\{\prod_{j=1}^{q} k_{j;W_i}(\frac{w_j-W_{ij}}{\mathcal{H}_j})\}}{\sum_{i=1}^{n}\{\prod_{j=1}^{q} k_{j;W_i}(\frac{w_j-W_{ij}}{\mathcal{H}_j})\}}, i = 1, \dots, n \qquad (12)$$

which is finally estimated as

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q} k_{j;W_i}(\frac{w_j-W_{ij}}{\mathcal{H}_j})\}\hat{Y}_i'}{\frac{1}{n}\sum_{i=1}^{n}\{\prod_{j=1}^{q} k_{j;W_i}(\frac{w_j-W_{ij}}{\mathcal{H}_j})\}} = \sum_{i=1}^{n} \mathcal{C}_i(Y_i - X_i'\hat{\underset{\sim}{\Gamma}}).$$

II. Construction of Hypotheses

Here, the hypotheses of interest to check independence between (X, W) and $\mathcal{E}$ are considered as follows.

$$H_0: (X, W) \perp\!\!\!\perp \mathcal{E} \text{ against } H_1: (X, W) \not\perp\!\!\!\perp \mathcal{E} \qquad (13)$$

where '$\perp\!\!\!\perp$' stands for independence. However, development of any test procedure under (13) is quite inconvenient, as $\mathcal{E}$ is an unobservable quantity. Therefore, a bonafide substitute of $\mathcal{E}$ is required in (13) to re-frame the null hypothesis. Let us define *r*-th order difference of $\mathcal{E}$ and Y as

$$\mathcal{E}^*(r) = \sum_{j=1}^{r+1} (-1)^{j-1} \binom{r}{j-1} \mathcal{E}_j \text{ and } Y^*(r) = \sum_{j=1}^{r+1} (-1)^{j-1} \binom{r}{j-1} Y_j.$$

where $\mathcal{E}_1, \dots, \mathcal{E}_{r+1}$ and $Y_1, \dots, Y_{r+1}$ are $(r+1)$ *i.i.d.* errors and responses respectively. Hence, by defining a general r-th order difference of $\mathcal{E}^*$, as denoted $\mathcal{E}^*(r)$, we only verify if $\mathcal{E}^*(r) \approx Y^*(r)$, where $Y^*(r)$ is the r-th order difference of Y^*.

Theorem 3.1. $\mathcal{E}^*(r)$, which has maximum *k*-th order absolute moment among all possible linear functions $\sum_{j=1}^{r+1} u_j \mathcal{E}_j$ with real coefficients u_j's, is approximated as $Y^*(r)$.

Under a simple partially linear regression model, **Das et al.** (2022) [4] considered third order difference of $\mathcal{E}^*$, to develop a test of independence between nonparametric regressor X and random error $\mathcal{E}$. The proof of Theorem 3.1. is elaborately discussed in Appendix-II. We are in quest of a general *r*-th order difference $\mathcal{E}^*(r)$. Therefore, we transform null hypothesis as $H_0: (X, W) \perp\!\!\!\perp \mathcal{E}^*(r)$ which approximately implies that $(X, W) \perp\!\!\!\perp Y^*(r)$. Furthermore, any function of (X, W) is independent to $Y^*(r)$.

Proposition 1. $\hat{Y}^*(r)$, the *r*-th order difference of $\hat{Y}$, can be approximated as a function of (X, W).

Therefore, H_0 further implies that $\hat{Y}^*(r) \perp\!\!\!\perp Y^*(r)$ and the new hypotheses are formulated as

$$H_0: \hat{Y}^*(r) \perp\!\!\!\perp Y^*(r) \text{ against } H_1: \hat{Y}^*(r) \not\perp\!\!\!\perp Y^*(r). \qquad (14)$$

Since the objective is to carry out a consistent test of (14), a conventional approach is to define a contiguous sequence of all possible alternative hypotheses with the aid of theory of contiguity due to **Le Cam** (1960) [3]. Such a sequence converges to the null hypothesis of interest as the sample size increases. Using Le Cam's first lemma, we construct the following sequence of contiguous alternatives (**Das et al.** (2022) [4])

$$H_n: \quad \tilde{G}_{n;\hat{Y}^*(r),Y^*(r)}(\hat{y}^*, y^*) = (1-\frac{\mu}{\sqrt{n}})G_{0;\hat{Y}^*(r),Y^*(r)}(\hat{y}^*, y^*) + \frac{\mu}{\sqrt{n}}G_{\hat{Y}^*(r),Y^*(r)}(\hat{y}^*, y^*) \qquad (15)$$

where $\tilde{G}_{n;\hat{Y}^*(r),Y^*(r)}(\cdot,\cdot)$ is the joint cumulative distribution function of $(\hat{Y}^*(r), Y^*(r))$ under H_n, $G_{0;\hat{Y}^*(r),Y^*(r)}(\cdot,\cdot)$ and $G_{\hat{Y}^*(r),Y^*(r)}(\cdot,\cdot)$ are the joint cumulative distribution functions of $(\hat{Y}^*(r), Y^*(r))$ under H_0 and H_1 respectively. $\mu(>0)$ is the mixing parameter. H_n is constructed as a sequence of contiguous alternatives provided that the corresponding joint densities of $(\hat{Y}^*(r), Y^*(r))$ under H_0 and H_n and their marginal densities exist.

III. TEST STATISTICS

Next, we construct non-degenerate test statistics based on Spearman's ρ_s. The concept of V-statistic is utilized to propose the test statistics, as provided below.

$$S_{<nr>} = 3n^{-3}\sum_{u_1=1}^{n}\sum_{u_2=1}^{n}\sum_{u_3=1}^{n} sign\{(\hat{y}^*_{u_1}(r) - \hat{y}^*_{u_2}(r))(y^*_{u_1}(r) - y^*_{u_3}(r))\} \qquad (16)$$

where $(\hat{y}^*_1(r), y^*_1(r)), \ldots, (\hat{y}^*_n(r), y^*_n(r))$ are *i.i.d.* samples on $(\hat{Y}^*(r), Y^*(r))$. It is noteworthy that as an implication of H_0, $\rho_s(\hat{Y}^*(r), Y^*(r)) = 0$. The measures take nonzero values under dependence of $(\hat{Y}^*(r), Y^*(r))$. Note that the kernel of ρ_s is $3sign\{(\hat{y}^*_{u_1}(r) - \hat{y}^*_{u_2}(r))(y^*_{u_1}(r) - y^*_{u_3}(r))\}$, $1 \le u_1 \ne u_2 \ne u_3 \le n$.

Proposition 2. Under H_0, the kernel of $S_{<nr>}$ is nondegenerate.

Hence, the asymptotic distributions of (16) under both H_0 and H_n are asymptotically normal, as deduced in the next section.

IV. ASYMPTOTIC DISTRIBUTIONS OF TEST STATISTICS

The asymptotic distribution of a nondegenerate V-statistic was determined by **Zhou et al.** (2021) [14]. In similar way, the asymptotic distributions of $S_{<nr>}$ under H_0 and H_n can be derived. Proofs are available in **Das et al.** (2022) [4].

Theorem 5.1. Under H_0, provided that $E[h^2((\hat{Y}^*_1(r), Y^*_1(r)), (\hat{Y}^*_2(r), Y^*_2(r)), (\hat{Y}^*_3(r), Y^*_3(r)))] < \infty$,

$\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) \xrightarrow{L} N(0, 4\nu_1(r))$,

where $\nu_1(r) = Var_{H_0}[E\{h((\hat{Y}^*_1(r), Y^*_1(r)), (\hat{Y}^*_2(r), Y^*_2(r)), (\hat{Y}^*_3(r), Y^*_3(r)))|(\hat{Y}^*_1(r), Y^*_1(r))\}] = 1$.

The asymptotic power of $S_{<nr>}$ is derived as

$$P_{H_n}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > s) \quad = \quad \Phi(\frac{\Delta^{(r)} - s}{\sqrt{4\nu_1(r)}})$$

where s satisfies $P_{H_0}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > s)$ for $0 < \alpha < 1$, α being the level of significance of the test. μ=0 implies $\Delta^{(r)} = 0$, yielding asymptotic power as the size of the test. Here, $\Delta^{(r)}$=$3P'_c - 3P'_d$ where P'_c is the probability of concordance of three arbitrary paired observations among n *i.i.d.* samples from $(\hat{Y}^*(r), Y^*(r))$ and P'_d denotes the probability of discordance of those three paired observations for computation of ρ_s, expressed as

$$P_c' = P(\hat{Y_1^*}(r) < \hat{Y_2^*}(r), Y_1^*(r) < Y_3^*(r)) + P(\hat{Y_1^*}(r) > \hat{Y_2^*}(r), Y_1^*(r) > Y_3^*(r)),$$

and

$$P_d' = P(\hat{Y_1^*}(r) > \hat{Y_2^*}(r), Y_1^*(r) < Y_3^*(r)) + P(\hat{Y_1^*}(r) < \hat{Y_2^*}(r), Y_1^*(r) > Y_3^*(r)).$$

The next proposition is presented in order to establish consistency of $S_{<nr>}$.

Proposition 3. For $n^* > n$,

$$P_{H_n}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > q_\alpha) < P_{H_n}(\sqrt{n}(S_{<n^*r>} - E_{H_0}(S_{<n^*r>})) > q_\alpha)$$

and $P_{H_n}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > q_\alpha) \uparrow 1$ as $\mu \uparrow$ and $n \to \infty$.

We consider a pair of examples to assess the power performance of the test statistics against μ by taking sample size n=1000 and r=2,3,4,5,10 as the orders of difference. Also, $\mathcal{E} \sim N(0,0.02)$ and $Y \sim t_2$ (t-distribution with 2 *d.f.*) under H_0.

Example 1. Consider a generalized partially linear model $Y = \Gamma_1 X_1 + \Gamma_2 X_2 + g(W_1, W_2) + \mathcal{E}$ with error assumptions $(i) E(\mathcal{E}|X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2) = 0$ for all (x_1, x_2, w_1, w_2),
$(ii) E(\mathcal{E}^2|X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2) = Q^2(x_1, x_2, w_1, w_2)$.

Moreover,

$$(X_1, X_2, W_1, W_2) \sim \mathbb{N}_4\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix} 0.18 & -0.06 & 0.22 & -0.13 \\ -0.06 & 0.14 & -0.28 & 0.19 \\ 0.22 & -0.28 & 0.20 & 0.17 \\ -0.13 & 0.19 & 0.17 & 0.25 \end{pmatrix}\right).$$

The nonparametric regression function $g(W_1, W_2) = 0.45W_1W_2 - 0.25W_1^2W_2 + W_2^3$. The conditional error is distributed as $(\mathcal{E}|X_1, X_2, W_1, W_2) \sim N(0,0.015|1 - 0.13X_1 - 2.1X_2 + 5.6W_1 - 0.95W_2|)$ under H_1.

Example 2. Another generalized partially linear model is considered as $Y = \Gamma_1 X_1 + \Gamma_2 X_2 + \Gamma_3 X_3 + \Gamma_4 X_4 + \Gamma_5 X_5 + g(W_1, W_2, W_3) + \mathcal{E}$ with assumptions on $\mathcal{E}$ as

$(i)\ E(\mathcal{E}|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5, W_1 = w_1, W_2 = w_2, W_3 = w_3) = 0,$

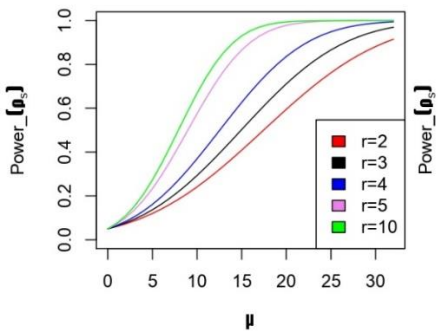
$(ii) E(\mathcal{E}^2|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5, W_1 = w_1, W_2 = w_2, W_3 = w_3) = Q^2(x_1, x_2, x_3, x_4, x_5, w_1, w_2, w_3)$ for all $(x_1, x_2, x_3, x_4, x_5, w_1, w_2, w_3)$.

The joint distribution of $(X_1, X_2, X_3, X_4, X_5, W_1, W_2, W_3)^T$ is

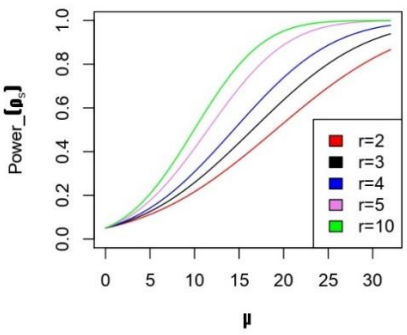
$$\mathbb{N}_8\left(\begin{pmatrix}0\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}, \begin{pmatrix} 0.18 & -0.06 & 0.09 & -0.13 & 0.16 & -0.17 & 0.31 & -0.22 \\ -0.06 & 0.14 & -0.17 & 0.26 & -0.14 & 0.18 & 0.25 & -0.33 \\ 0.09 & -0.17 & 0.25 & 0.33 & -0.18 & -0.24 & -0.15 & 0.15 \\ -0.13 & 0.26 & 0.33 & 0.32 & 0.05 & -0.14 & -0.07 & 0.12 \\ 0.16 & -0.14 & -0.18 & 0.05 & 0.24 & 0.26 & -0.18 & -0.03 \\ -0.17 & 0.18 & -0.24 & -0.14 & 0.26 & 0.11 & -0.21 & -0.19 \\ 0.31 & 0.25 & -0.15 & -0.07 & -0.18 & -0.21 & 0.27 & 0.14 \\ -0.22 & -0.33 & 0.15 & 0.12 & -0.03 & -0.19 & 0.14 & 0.40 \end{pmatrix}\right)$$

We take

$g(W_1, W_2, W_3) = 0.36W_1^3 - 0.25W_2^2W_3 - 0.11W_3^2W_1 + 0.08W_1W_2W_3$.
$(\mathcal{E}|X_1, X_2, X_3, X_4, X_5, W_1, W_2, W_3) \sim N(0,0.015|1 + 8X_1 + 5X_2 - 7X_3 - 4X_4 + 3.1X_5 - 6W_1 + 3W_2 + 6W_3|)$ under H_1.



Asymptotic powers of $S_{<nr>}$ in Example 1



Asymptotic powers of $S_{<nr>}$ in Example 2

We observe the following from above examples:

1. The asymptotic power increases as we increase the order of difference (r).
2. Moreover, under lower order Gaussian set-up (Example 1), $S_{<nr>}$ delivers comparatively abrupt sloping in higher and lower order of Gaussian distributions.

V. Concluding remarks

In this article, we have constructed some consistent tests, pertinent for generalized partially linear regression setup. The test statistics are based on Spearman's ρ_s and they are nondegenerate V-statistics. Such tests are evident to check whether the assumptions on error $\mathcal{E}$ are relevant or not. If independence between the regressors and error holds further we proceed on the validity of homoscedastic error. Both tests based on $s_{\langle nr \rangle}$ are efficient and consistent with the improvement order of difference r. One may perform this test by considering a degenerate test statistic. The order of degeneracy indeed affects the power of a nonparametric test statistic. It is noteworthy that incorporation of more regressors in underlying model enhances the asymptotic powers of the concerning test statistics as shown by Example 1 and Example 2. As a conclusion, the association between the higher-order difference in the estimated response and the observed response causes monotonicity in the power performances of the test statistics in this regard. Furthermore, as a prospective future introspection, r can be improved from 10 to achieve more powerful tests.

REFERENCES

[1] Andrews, D. W. 1995. Nonparametric kernel estimation for semiparametric models. Econometric Theory, 11(3): 560-586.

[2] Bergsma, W. P. 2006. A new correlation coefficient, its orthogonal decomposition and associated tests of independence. https://doi.org/10.48550/arXiv.math/0604627.

[3] Cam Le, L. 1960. Locally asymptotically normal families of distributions. University of California Publications in Statistics, 3: 37-98.

[4] Das, S. & Maiti, S. I. 2022. On the Test of Association Between Nonparametric Covariate and Error in Semiparametric Regression Model. Journal of the Indian Society for Probability and Statistics, 23(2): 541-564.

[5] Das, S. 2025. Study on efficacy of Kendall's τ based test statistic in generalized partially linear regression model. American Journal of Theoretical and Applied Statistics. Accepted for publication.

[6] Dhar, S. S., Dassios, A., and Bergsma, W. 2018. Testing Independence of Covariates and Errors in Nonparametric regression. Scandinavian Journal of Statistics, 45: 421-443.

[7] Hamilton, S. A., & Truong, Y. K. 1997. Local linear estimation in partly linear models. Journal of Multivariate Analysis, 60(1): 1-19.

[8] Lévy, P. 1939. Sur la division d'un segment par des points choisis au hasard. CR Acad. Sci. Paris, 208: 147-149.

[9] Li, Q. 2000. Efficient estimation of additive partially linear models. International Economic Review, 41(4): 1073-1092.

[10] Liu, Z., Liu, Z., Lu, X., & Lu, X. 1997. Root-n-consistent semiparametric estimation of partially linear models based on k-nn method. Econometric Reviews, 16(4): 411-420.

[11] Pyke, R. 1965. Spacings. Journal of the Royal Statistical Society: Series B (Methodological), 27(3): 395-436.
[12] Robinson, P. M. 1988. Root-N-consistent semiparametric regression. Econometrica, 56(4): 931-954.
[13] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. 2007. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6): 2769-2794.
[14] Zhou, Z., Mentch, L., & Hooker, G. 2021. V-statistics and variance estimation. Journal of Machine Learning Research, 22(287): 1-48.

APPENDIX-I: TABLES

μ	POWERS OF $S_{<nr>}$				
	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 10$
0	0.05	0.05	0.05	0.05	0.05
1	0.0605	0.0624	0.0651	0.072	0.0754
5	0.1203	0.1369	0.1614	0.2335	0.2728
10	0.2413	0.2934	0.3696	0.5753	0.6686
15	0.4088	0.5029	0.6267	0.8659	0.9301
20	0.5952	0.7115	0.8362	0.9786	0.9941
25	0.7619	0.8662	0.949	0.9984	0.9998
30	0.8817	0.9515	0.989	0.9999	1

TABLE 1: ASYMPTOTIC POWERS OF $S_{<nr>}$ FOR r = 2,3,4,5,10 IN **EXAMPLE 1**

μ	POWERS OF $S_{<nr>}$				
	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 10$
0	0.05	0.05	0.05	0.05	0.05
1	0.0595	0.0612	0.0629	0.0666	0.0695
5	0.1123	0.1258	0.1415	0.1763	0.2066
10	0.2165	0.2585	0.3076	0.415	0.5032
15	0.3617	0.4406	0.5272	0.6916	0.798
20	0.5305	0.6364	0.7387	0.8879	0.9516
25	0.6939	0.8016	0.8869	0.9732	0.9936
30	0.8257	0.9108	0.9626	0.9959	0.9995

TABLE 2: ASYMPTOTIC POWERS OF $S_{<nr>}$ FOR r = 2,3,4,5,10 IN **EXAMPLE 2**

APPENDIX-II: PROOFS

Proof of Theorem 3.1.

Proof is similar to Combined Proofs of Proposition 3.1 and Theorem 3.1 in **Das (2025)** [5]. The proofs are furnished using **Lévy (1939)** [8] and **Pyke (1965)** [11].

Proof of Proposition 1.

Proof is similar to Proof of Proposition 3.2 in **Das (2025)** [5].

Proof of Theorem 5.1.

Here, the kernel is $\psi((a_1,b_1),(a_2,b_2),(a_3,b_3)) = 3sign\{(a_1-a_2)(b_1-b_3)\}$ for three arbitrary bivariate points $(a_1,b_1),(a_2,b_2),(a_3,b_3)$.
Define,

$$f_1(\hat{Y_1^*}(r), Y_1^*(r))$$
$$= E[3sign\{(\hat{Y_1^*}(r) - \hat{Y_2^*}(r))(Y_1^*(r) - Y_3^*(r))\}|(\hat{Y_1^*}(r), Y_1^*(r))]$$
$$= 3\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} sign(\hat{Y_1^*}(r) - \hat{y_2^*}(r))sign(Y_1^*(r) - y_3^*(r)) \times \prod_{i=2,3} dG_{\hat{Y^*(r)},Y^*(r)}(\hat{y_i^*}(r), y_i^*(r))$$
$$= 3\int_{\mathbb{R}^2} sign(\hat{Y_1^*}(r) - \hat{y_2^*}(r))I(Y_1^*(r))dG_{\hat{Y^*(r)},Y^*(r)}(\hat{y_2^*}(r), y_2^*(r))$$

where

$I(Y_1^*(r))$

$= \int_{\mathbb{R}^2} sign(Y_1^*(r) - y_3^*(r)) dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_3^*(r), y_3^*(r))$

$= \int_{\mathbb{R}} \int_{\mathbb{A}_1} sign(Y_1^*(r) - y_3^*(r)) dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_3^*(r), y_3^*(r)) + \int_{\mathbb{R}} \int_{\mathbb{A}_1^c} sign(Y_1^*(r) - y_3^*(r)) dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_3^*(r), y_3^*(r)),$

where $\mathbb{A}_1 = \{y_3^*(r): y_3^*(r) < Y_1^*(r)\}$

$= \int_{\mathbb{R}} \int_{-\infty}^{Y_1^*(r)} dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_3^*(r), y_3^*(r)) - \int_{\mathbb{R}} \int_{Y_1^*(r)}^{\infty} dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_3^*(r), y_3^*(r))$

$= H_{Y^*(r)}(Y_1^*(r)) - [1 - H_{Y^*(r)}(Y_1^*(r))] = 2H_{Y^*(r)}(Y_1^*(r)) - 1.$

$\therefore f_1(\hat{Y}_1^*(r), Y_1^*(r)) = 3[2H_{Y^*(r)}(Y_1^*(r)) - 1] \int_{\mathbb{R}^2} sign(\hat{Y}_1^*(r) - \hat{y}_2^*(r)) dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_2^*(r), y_2^*(r)).$

In similar way of deduction of $I(Y_1^*(r))$, one can compute $\int_{\mathbb{R}^2} sign(\hat{Y}_1^*(r) - \hat{y}_2^*(r)) dG_{\hat{Y}^*(r),Y^*(r)}(\hat{y}_2^*(r), y_2^*(r))$ as $[2G_{\hat{Y}^*(r)}(\hat{Y}_1^*(r)) - 1]$.

Here, $G_{\hat{Y}^*(r)}(\cdot)$ and $H_{Y^*(r)}(\cdot)$ are the marginal *c.d.f.'s* of $\hat{Y}^*(r)$ and $Y^*(r)$ respectively and they are uniformly distributed on [0,1]. Therefore, $f_1(\hat{Y}_1^*(r), Y_1^*(r)) = 3[2G_{\hat{Y}^*(r)}(\hat{Y}_1^*(r)) - 1][2H_{Y^*(r)}(Y_1^*(r)) - 1]$, which further implies $\nu_1 = Var(f_1(\hat{Y}_1^*(r), Y_1^*(r))) = 9 \cdot \frac{1}{3} \cdot \frac{1}{3} = 1 > 0$ under H_0.