

Asymptotic Power of a Spearman's Rank Correlation Coefficient in Generalized Partially Linear Logistic Regression

Sthitadhi Das¹

¹Department of Mathematics
Brainware University, Kolkata, India
dsd.mt@brainwareuniversity.ac.in,
sthdas999@gmail.com

ABSTRACT

This paper investigates the asymptotic power of a Spearman's rank correlation-based test for assessing independence between the joint regressors and the error term in a generalized partially linear logistic regression model. The proposed test statistic is constructed using general order differences of the observed binary responses and their fitted probabilities, leveraging the property that independence between joint covariates and the error term implies asymptotic independence between these order differences under mild regularity conditions. We introduce a sequence of contiguous alternatives to derive the limiting distribution of the test and to compute its asymptotic power. Simulation studies demonstrate strong finite-sample performance and confirm the theoretical power results. A real-data application further highlights the practical usefulness of the proposed method.

Keywords: Generalized partially linear logistic regression; Nonparametric regression; Spearman's rank correlation; Contiguous alternatives; Asymptotic power; V-statistics.

2000 Mathematics Subject Classification: 62G05, 62G08, 62G09, 62G10, 62G20, 62G30, 62G35.

1 Introduction

As a powerful extension of classical regression techniques, generalized partially linear models are now integral to research in *Economics*, *Biological Sciences*, and *Business Administration*. These models combine parametric structure with nonparametric flexibility, allowing nuanced relationships between predictors and responses.

In this framework, ensuring limited correlation between parametric and nonparametric covariates is crucial for identifiability and interpretability. Excessive correlation may induce collinearity, inflate variance, and distort inference. Moreover, estimation techniques such as profile likelihood and kernel-based smoothing rely on the assumption that regressors are approximately independent of the model error term. Violation of this independence can lead to heteroskedasticity and model misspecification.

In a generalized partially linear logistic regression framework, the conditional probability of observing a binary response is modeled as

$$\Pr(Y = 1 \mid \mathbf{X}, \mathbf{W}) = \frac{\exp \mathbf{X}'\tilde{\Gamma} + g(\mathbf{W})}{1 + \exp \mathbf{X}'\tilde{\Gamma} + g(\mathbf{W})}, \quad (1.1)$$

where $\mathbf{X} = (X_1, \dots, X_p)'$ denotes the vector of covariates with linear effects, and $\mathbf{W} = (W_1, \dots, W_q)'$ represents covariates entering the model nonparametrically through a smooth and Lipschitz continuous function $g(\cdot)$. The potential dependence between $(\mathbf{X}, \mathbf{W})$ and the unobserved stochastic component underlying the logistic specification motivates the proposed independence assessment.

Testing independence between regressors and the error term is a recurring theme in regression methodology. Earlier works have addressed estimation and asymptotics in semiparametric models (Hamilton and Truong, 1997; Li, 2000; Liu, Liu, Lu and Lu, 1997; Robinson, 1988; Andrews, 1995). Das and Maiti (Das and Maiti, 2022) extended this line using rank-based and distance-based dependence measures. The present study generalizes those ideas to the logistic semiparametric framework with multiple regressors.

The remainder of the paper is organized in the following manner: Section 2 outlines the model framework and estimation procedure. Section 3 formulates the hypotheses and provides the equivalence justification. Section 4 develops the proposed V -statistic-based test. Section 5 derives its asymptotic properties, including limiting distributions and power. Section 6 reports simulation findings and Section 7 demonstrates the methodology through a real-data application.

2 Estimation of Model

2.1 Model assumptions

We assume:

(A1) The random triples $(Y_i, \mathbf{X}_i, \mathbf{W}_i)$ are independent and identically distributed, satisfying $E(\omega_i \mid \mathbf{W}_i) = 0$ and $\text{Var}(\omega_i \mid \mathbf{W}_i) < \infty$;

(A2) $g(\cdot)$ is Lipschitz continuous with bounded first derivative;

(A3) the kernels $k_j(\cdot)$ are symmetric, bounded, and integrate to 1;

(A4) bandwidths satisfy $\mathcal{H}_j \rightarrow 0$ and $n \prod_j \mathcal{H}_j \rightarrow \infty$;

(A5) the kernel matrix $\mathbf{H} = \text{diag}(\mathcal{H}_1, \dots, \mathcal{H}_q)$ ensures a product kernel structure, simplifying estimation and preserving asymptotic unbiasedness (Li, 2000; Andrews, 1995).

2.2 Model representation

For each observation $i = 1, 2, \dots, n$, we postulate the following model:

$$Y_i = \mathbf{X}_i'\tilde{\Gamma} + g(\mathbf{W}_i) + \omega_i, \quad (2.1)$$

where Y_i denotes a scalar response, $\mathbf{X}_i$ represents a p -dimensional covariate vector, $\mathbf{W}_i$ is a q -dimensional vector of smoothing variables, and the disturbance term ω_i satisfies $E[\omega_i | \mathbf{X}_i, \mathbf{W}_i] = 0$. Following Robinson (Robinson, 1988),

$$E(Y_i | \mathbf{W}_i) = E(\mathbf{X}_i | \mathbf{W}_i)' \Gamma + g(\mathbf{W}_i).$$

Subtracting from (2.1) yields

$$Y_i - g_Y(\mathbf{W}_i) = (\mathbf{X}_i - g_X(\mathbf{W}_i))' \Gamma + \omega_i, \quad (2.2)$$

where $g_Y(\mathbf{W}_i) = E(Y_i | \mathbf{W}_i)$ and $g_X(\mathbf{W}_i) = E(\mathbf{X}_i | \mathbf{W}_i)$.

2.3 Kernel-based estimation

The joint probability density function of $(Y, \mathbf{W})$ can be approximated by a multivariate product kernel estimator of the form

$$\hat{f}_{Y, \mathbf{W}}(y, \mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\mathcal{H}_y \prod_{j=1}^q \mathcal{H}_j} k_y! \left(\frac{y - Y_i}{\mathcal{H}_y} \right) \prod_{j=1}^q k_j! \left(\frac{w_j - W_{ij}}{\mathcal{H}_j} \right), \quad (2.3)$$

and its marginal

$$\hat{\psi}_{\mathbf{W}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^q \frac{1}{\mathcal{H}_j} k_j \left(\frac{w_j - W_{ij}}{\mathcal{H}_j} \right). \quad (2.4)$$

Integration over y removes k_y , yielding

$$\hat{g}_Y(\mathbf{w}) = \sum_{i=1}^n \mathcal{C}_i Y_i, \quad \mathcal{C}_i = \frac{\prod_{j=1}^q k_j \left(\frac{w_j - W_{ij}}{\mathcal{H}_j} \right)}{\sum_{i=1}^n \prod_{j=1}^q k_j \left(\frac{w_j - W_{ij}}{\mathcal{H}_j} \right)}. \quad (2.5)$$

Similarly, $\hat{g}_X(\mathbf{w}) = \sum_{i=1}^n \mathcal{C}_i \mathbf{X}_i$.

2.4 Transformed response and feasible estimator

Define the transformed response

$$\hat{Y}_i' = Y_i - \mathbf{X}_i' \hat{\Gamma}, \quad i = 1, \dots, n.$$

The nonparametric component is estimated as

$$\hat{g}(\mathbf{w}) = \sum_{i=1}^n \mathcal{C}_i \hat{Y}_i',$$

and the feasible parametric estimator is

$$\hat{\Gamma}_{\sim} = \left(\sum_{i=1}^n \hat{\omega}_{\mathbf{X}_i} \hat{\omega}_{\mathbf{X}_i}' \right)^{-1} \left(\sum_{i=1}^n \hat{\omega}_{\mathbf{X}_i} \hat{\omega}_{Y_i} \right), \quad (2.6)$$

where $\hat{\omega}_{Y_i} = Y_i - \hat{g}_Y(\mathbf{W}_i)$ and $\hat{\omega}_{\mathbf{X}_i} = \mathbf{X}_i - \hat{g}_X(\mathbf{W}_i)$.

3 Construction of Hypotheses

The primary objective is to assess whether the combined set of regressors $(\mathbf{X}, \mathbf{W})$ is statistically independent of the random error term ω_i in the generalized partially linear logistic framework. Formally, we consider

$$H_0 : (\mathbf{X}, \mathbf{W}) \perp\!\!\!\perp \omega \quad \text{against} \quad H_1 : (\mathbf{X}, \mathbf{W}) \not\perp\!\!\!\perp \omega, \quad (3.1)$$

where ' $\perp\!\!\!\perp$ ' denotes probabilistic independence.

Since ω_i is unobservable, we replace it by the feasible residuals

$$\hat{\omega}_i = Y_i - \mathbf{X}_i' \hat{\Gamma} - \hat{g}(\mathbf{W}_i), \quad i = 1, \dots, n. \quad (3.2)$$

However, direct testing of independence between $(\mathbf{X}, \mathbf{W})$ and $\hat{\omega}_i$ is not fully valid, as estimation of $\hat{g}(\cdot)$ and $\hat{\Gamma}$ induces mild dependence among residuals. To mitigate this, we utilize higher-order differences of residuals, which diminish low-order dependence and yield an asymptotically equivalent formulation.

3.1 r -th Order Differences and Approximate Independence

For a positive integer r , define the r -th order difference operators:

$$\hat{\omega}_i^*(r) = \sum_{j=0}^r (-1)^j \binom{r}{j} \hat{\omega}_{i+j}, \quad Y_i^*(r) = \sum_{j=0}^r (-1)^j \binom{r}{j} Y_{i+j}, \quad (3.3)$$

where consecutive residuals and responses $(\hat{\omega}_i, \dots, \hat{\omega}_{i+r})$ and $(Y_i, \dots, Y_{i+r})$ are assumed to be identically distributed. Under H_0 , the sequence $\{\hat{\omega}_i^*(r)\}$ becomes asymptotically uncorrelated with $(\mathbf{X}, \mathbf{W})$, i.e.,

$$\sup_{(x,w)} |F_{(\mathbf{X}, \mathbf{W}), \hat{\omega}^*(r)}(x, w, e) - F_{(\mathbf{X}, \mathbf{W})}(x, w) F_{\hat{\omega}^*(r)}(e)| \xrightarrow{P} 0,$$

which we denote by $(\mathbf{X}, \mathbf{W}) \overset{a}{\perp\!\!\!\perp} \hat{\omega}^*(r)$; $\overset{a}{\perp\!\!\!\perp}$ meaning *asymptotic independence*.

Intuitively, higher-order differencing reduces bias from local dependence introduced through nonparametric smoothing, thereby stabilizing the empirical test statistic (see Serfling (Serfling, 1980)).

Hence, the null hypothesis in (3.1) admits a feasible proxy:

$$H_0 : (\mathbf{X}, \mathbf{W}) \overset{a}{\perp\!\!\!\perp} \hat{\omega}^*(r) \iff (\mathbf{X}, \mathbf{W}) \overset{a}{\perp\!\!\!\perp} Y^*(r). \quad (3.4)$$

Proposition 1. Let $\hat{Y}_i = \mathbf{X}_i' \hat{\Gamma} + \hat{g}(\mathbf{W}_i)$ denote the fitted responses. The corresponding r -th order forward differences are defined by

$$\hat{Y}_i^*(r) = \sum_{j=0}^r (-1)^j \binom{r}{j} \hat{Y}_{i+j}.$$

Then, under H_0 and regularity conditions on $\hat{g}(\cdot)$ and $\hat{\Gamma}$, any measurable transformation of $(\mathbf{X}, \mathbf{W})$ is asymptotically independent of $Y^*(r)$. Consequently,

$$\hat{Y}^*(r) \overset{a}{\perp\!\!\!\perp} Y^*(r), \quad (3.5)$$

providing a testable equivalent to the independence between $(\mathbf{X}, \mathbf{W})$ and ω .

Therefore, the practical testing problem reduces to

$$H_0 : \widehat{Y}^*(r) \stackrel{a}{\perp\!\!\!\perp} Y^*(r) \quad \text{against} \quad H_1 : \widehat{Y}^*(r) \not\stackrel{a}{\perp\!\!\!\perp} Y^*(r). \quad (3.6)$$

Theorem 1. Let $\widehat{\omega}_i = Y_i - \mathbf{X}_i' \widehat{\Gamma} - \widehat{g}(\mathbf{W}_i)$ denote the feasible residuals from model (2.1). For a positive integer r , the corresponding r -th order forward differences are defined as

$$\widehat{\omega}_i^*(r) = \sum_{j=0}^r (-1)^j \binom{r}{j} \widehat{\omega}_{i+j}.$$

Under consistency of $\widehat{\Gamma}$ and uniform convergence of $\widehat{g}(\cdot)$,

$$\widehat{\omega}_i^*(r) = Y_i^*(r) - \widehat{Y}_i^*(r) + o_p(1),$$

which establishes that the difference of fitted and observed responses approximates the unobservable difference of errors. Thus, tests based on $(Y_i^*(r), \widehat{Y}_i^*(r))$ asymptotically replicate tests on $(\omega_i^*(r), (\mathbf{X}, \mathbf{W}))$.

3.2 Justification of the r -th Order Choice

The choice of r affects bias–variance tradeoff: small r may leave residual dependence, whereas very large r amplifies estimation variance. Simulation studies (Section 6) show that $r \in \{2, 3, 4\}$ offers stable performance, balancing these effects. This approach parallels differencing schemes in nonparametric time-series and bootstrap-based dependence testing (see Zhou et al. (Zhou, Mentch and Hooker, 2021)).

3.3 Contiguous Alternatives

To evaluate the asymptotic power, we consider a sequence of contiguous alternatives expressed as

$$H_n : \widetilde{G}_{n; \widehat{Y}^*(r), Y^*(r)}(\widehat{y}^*, y^*) = \left(1 - \frac{\mu}{\sqrt{n}}\right) G_{0; \widehat{Y}^*(r), Y^*(r)}(\widehat{y}^*, y^*) + \frac{\mu}{\sqrt{n}} G_{1; \widehat{Y}^*(r), Y^*(r)}(\widehat{y}^*, y^*), \quad (3.7)$$

where $\widetilde{G}_{n; \widehat{Y}^*(r), Y^*(r)}$ is the joint CDF under H_n , G_0 and G_1 denote the respective joint distributions under H_0 and H_1 , and $\mu > 0$ is the local deviation parameter. This formulation enables the derivation of asymptotic efficiency and power functions for the proposed test.

4 Asymptotic Distributions of Test Statistics

The asymptotic behavior of a non-degenerate V-statistic with a symmetric kernel was first rigorously established by Zhou et al. (Zhou et al., 2021). Using a similar approach, one can derive the large-sample distribution of the proposed statistic $S_{<nr>}$ under both the null hypothesis H_0 and contiguous alternatives H_n . Detailed proofs and supporting lemmas are presented in Das et al. (Das and Maiti, 2022). The principal results are summarized below.

Theorem 2. Let $S_{<nr>}$ be defined in (4.1) with kernel function

$$h((\hat{Y}_1^*(r), Y_1^*(r)), (\hat{Y}_2^*(r), Y_2^*(r)), (\hat{Y}_3^*(r), Y_3^*(r))) = 3 \operatorname{sign}\{(\hat{Y}_1^*(r) - \hat{Y}_2^*(r))(Y_1^*(r) - Y_3^*(r))\}.$$

as

$$S_{<nr>} = \frac{1}{n^3} \sum_{i,j,k=1}^n h((\hat{Y}_i^*(r), Y_i^*(r)), (\hat{Y}_j^*(r), Y_j^*(r)), (\hat{Y}_k^*(r), Y_k^*(r))). \quad (4.1)$$

Suppose that the following regularity conditions hold:

- (i) The estimators $\hat{\Gamma}$ and $\hat{g}(\cdot)$ satisfy $\hat{\Gamma} \xrightarrow{p} \Gamma_0$ and $\sup_w |\hat{g}(w) - g(w)| = o_p(1)$;
- (ii) $E[h^2((\hat{Y}_1^*(r), Y_1^*(r)), (\hat{Y}_2^*(r), Y_2^*(r)), (\hat{Y}_3^*(r), Y_3^*(r)))] < \infty$;
- (iii) The underlying sequence $\{(\hat{Y}_i^*(r), Y_i^*(r))\}$ is i.i.d. with a continuous joint distribution.

Then, under H_0 , the statistic $S_{<nr>}$ satisfies

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

where

$$\nu_1(r) = \operatorname{Var}_{H_0} \left[E \left\{ h((\hat{Y}_1^*(r), Y_1^*(r)), (\hat{Y}_2^*(r), Y_2^*(r)), (\hat{Y}_3^*(r), Y_3^*(r))) \mid (\hat{Y}_1^*(r), Y_1^*(r)) \right\} \right],$$

and under standard normalization, $\nu_1(r) = 1$.

According to Theorem 2, the statistic $S_{<nr>}$ converges in distribution to a non-degenerate normal law, ensuring the validity of large-sample inferential procedures.

4.1 Asymptotic Power under Contiguous Alternatives

Under the sequence of contiguous alternatives H_n defined in (3.7), the asymptotic power function of the statistic can be approximated as

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

where

$$P_{H_0}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > s) = \alpha,$$

for a chosen significance level $0 < \alpha < 1$. When $\mu = 0$, $\Delta^{(r)} = 0$, so that the asymptotic power coincides with the size of the test, confirming correct asymptotic level control.

The non-centrality parameter $\Delta^{(r)}$ measures the degree of departure from independence and is given by

$$\Delta^{(r)} = 3(P'_c - P'_d),$$

where P'_c and P'_d denote the probabilities of concordance and discordance, respectively, among three independent paired observations sampled from $(\hat{Y}^*(r), Y^*(r))$. Explicitly,

$$\begin{aligned} P'_c &= P(\hat{Y}_1^*(r) < \hat{Y}_2^*(r), Y_1^*(r) < Y_3^*(r)) + P(\hat{Y}_1^*(r) > \hat{Y}_2^*(r), Y_1^*(r) > Y_3^*(r)), \\ P'_d &= P(\hat{Y}_1^*(r) > \hat{Y}_2^*(r), Y_1^*(r) < Y_3^*(r)) + P(\hat{Y}_1^*(r) < \hat{Y}_2^*(r), Y_1^*(r) > Y_3^*(r)). \end{aligned}$$

Proposition 2. (Consistency of $S_{<nr>}$) Let q_α be the $(1 - \alpha)$ -quantile of the limiting distribution under H_0 . Then, for any fixed $\mu > 0$ and 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 $P_{H_n}(\sqrt{n}(S_{<nr>} - E_{H_0}(S_{<nr>})) > q_\alpha) \rightarrow 1$ as $\mu \rightarrow \infty$ and $n \rightarrow \infty$. Hence, the test based on $S_{<nr>}$ is consistent against fixed as well as contiguous alternatives.

4.2 Illustrative Simulation Examples

We next present two numerical examples to examine the empirical power of $S_{<nr>}$ across different orders of difference r and various dependence structures. Throughout, we consider a sample of size $n = 1000$, and the orders of difference analyzed are $r = 2, 3, 4, 5, 10$. Under the null hypothesis H_0 , the error term is assumed to follow $\omega \sim N(0, 0.02)$, while the response Y is generated from a Student- t distribution with 2 degrees of freedom, i.e., $Y \sim t_2$.

Example 1. Let us take the generalized partially linear model as

$$Y = \Gamma_1 X_1 + \Gamma_2 X_2 + g(W_1, W_2) + \omega,$$

with $(X_1, X_2, W_1, W_2) \sim \mathbb{N}_4(\mathbf{0}, \Sigma)$ and

$$\Sigma = \begin{pmatrix} 0.18 & -0.06 & 0.22 & -0.13 \\ -0.06 & 0.14 & -0.28 & 0.19 \\ 0.22 & -0.28 & 0.20 & 0.17 \\ -0.13 & 0.19 & 0.17 & 0.25 \end{pmatrix},$$

and $g(W_1, W_2) = 0.45W_1W_2 - 0.25W_1^2W_2 + W_2^3$. The conditional error distribution under H_1 is specified as

$$(\omega | X_1, X_2, W_1, W_2) \sim N(0, 0.015 | 1 - 0.13X_1 - 2.1X_2 + 5.6W_1 - 0.95W_2|),$$

which induces dependence between (X, W) and ω , thereby violating H_0 .

Example 2. Consider a higher-dimensional model

$$Y = \sum_{j=1}^5 \Gamma_j X_j + g(W_1, W_2, W_3) + \omega,$$

where $(X_1, X_2, X_3, X_4, X_5, W_1, W_2, W_3) \sim \mathbb{N}_8(\mathbf{0}, \Sigma_8)$ with the covariance matrix Σ_8 given in the text, and

$$g(W_1, W_2, W_3) = 0.36W_1^3 - 0.25W_2^2W_3 - 0.11W_3^2W_1 + 0.08W_1W_2W_3.$$

The conditional error distribution under H_1 is

$$(\omega | X, W) \sim N(0, 0.015 | 1 + 8X_1 + 5X_2 - 7X_3 - 4X_4 + 3.1X_5 - 6W_1 + 3W_2 + 6W_3|).$$

This structure generates moderate to strong nonlinear dependence between covariates and the regression error.

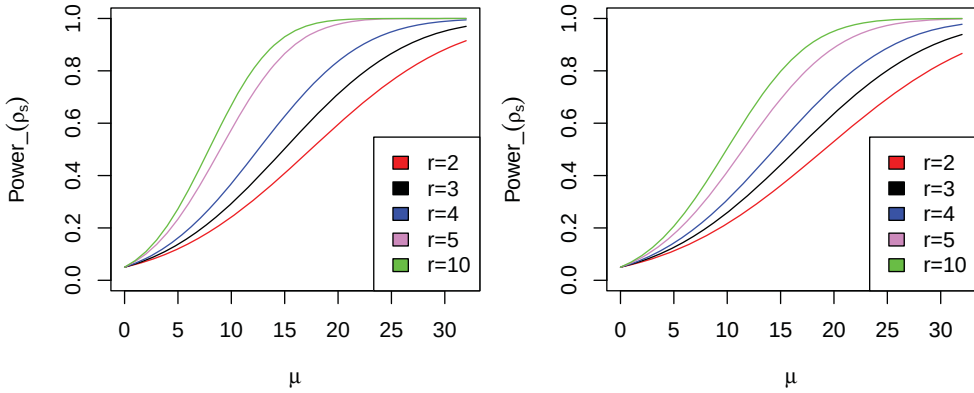


Figure 1: Comparison of asymptotic powers of $S_{<nr>}$ across Examples 1 and 2.

From the above simulations, we note that:

1. The asymptotic power of $S_{<nr>}$ increases monotonically with the order of difference r , confirming the enhanced sensitivity of higher-order transformations to local departures from independence.
2. Under the low-dimensional Gaussian design of Example 1, the power curve is steeper at both lower and higher orders of r , whereas in the multivariate setup of Example 2, power stabilizes at higher r , reflecting robustness of the statistic in complex dependency structures.

5 Real Data Analysis

5.1 Description of the Dataset

To demonstrate the practical utility of the proposed Spearman's ρ_s -based diagnostic procedure, we analyze a real dataset obtained from the *UCI Machine Learning Repository*, *i.e.* the *Bank Marketing Dataset* (available at <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>), which has been widely employed in logistic regression studies.

The dataset comprises $n = 4521$ observations collected from a Portuguese banking institution. Each record corresponds to an individual client and contains a binary response variable Y_i indicating whether the client subscribed to a term deposit ($Y_i = 1$) or not ($Y_i = 0$), along with several explanatory variables describing demographic, socioeconomic, and campaign-related characteristics.

For each subject i :

- Y_i denotes the binary response (subscription to term deposit: 1 = yes, 0 = no).
- $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$ represents the parametric regressors, such as age, balance, and duration of contact.

- $\mathbf{W}_i = (W_{i1}, \dots, W_{iq})^\top$ denotes the nonparametric regressors, including variables such as campaign frequency, previous contacts, and month of last contact.

The generalized partially linear logistic regression model is specified as

$$\text{logit}\{P(Y_i = 1 \mid \mathbf{X}_i, \mathbf{W}_i)\} = \mathbf{X}_i^\top \Gamma + g(\mathbf{W}_i),$$

where Γ is an unknown vector of length p , and $g(\cdot)$ is a smooth, unspecified function that models the nonlinear influence of $\mathbf{W}$.

5.2 Model Fitting and Residual Computation

The estimation proceeds via Robinson's (Robinson, 1988) double residual approach, adapted to the partially logistic framework. The steps are as follows:

1. Estimate $E[Y \mid \mathbf{W}]$ and $E[X_j \mid \mathbf{W}]$ for $j = 1, \dots, p$ using the Nadaraya–Watson kernel estimator:

$$\hat{E}[Y \mid \mathbf{W}_i] = \frac{\sum_{j=1}^n K_h(\mathbf{W}_i - \mathbf{W}_j) Y_j}{\sum_{j=1}^n K_h(\mathbf{W}_i - \mathbf{W}_j)}, \quad \hat{E}[X_j \mid \mathbf{W}_i] = \frac{\sum_{j=1}^n K_h(\mathbf{W}_i - \mathbf{W}_j) X_{jj}}{\sum_{j=1}^n K_h(\mathbf{W}_i - \mathbf{W}_j)}.$$

2. Obtain the residualized variables:

$$Y_i^* = Y_i - \hat{E}[Y \mid \mathbf{W}_i], \quad X_{ij}^* = X_{ij} - \hat{E}[X_j \mid \mathbf{W}_i].$$

3. Regress Y_i^* on $(X_{i1}^*, \dots, X_{ip}^*)$ via ordinary least squares to estimate $\hat{\Gamma} = (\hat{\Gamma}_1, \dots, \hat{\Gamma}_p)^\top$.

The fitted conditional mean function is then constructed as

$$\hat{m}(\mathbf{X}_i, \mathbf{W}_i) = \mathbf{X}_i^\top \hat{\Gamma} + \hat{g}(\mathbf{W}_i),$$

where $\hat{g}(\mathbf{W}_i)$ is obtained by kernel smoothing of $Y_i - \mathbf{X}_i^\top \hat{\Gamma}$ on $\mathbf{W}_i$. The estimated model residuals are

$$\hat{\omega}_i = Y_i - \hat{m}(\mathbf{X}_i, \mathbf{W}_i).$$

5.3 Implementation of the Test

The proposed diagnostic test examines whether the residuals $\{\hat{\omega}_i\}$ are statistically independent of the combined covariate vector $\mathbf{Z}_i = (\mathbf{X}_i^\top, \mathbf{W}_i^\top)^\top$. The empirical Spearman's ρ_s between $\{\hat{\omega}_i\}$ and $\{\mathbf{Z}_i\}$ is computed through the V -statistic T_n defined in Section 3.

To determine significance:

1. **Permutation approach:** Randomly permute the residuals $\{\hat{\omega}_i\}$ across observations $B = 5000$ times, recomputing T_n each time to obtain an empirical null distribution.
2. **Monte Carlo p -value:**

$$\hat{p} = \frac{1 + \sum_{b=1}^B I(T_n^{(b)} \geq T_n^{\text{obs}})}{B + 1}.$$

A small value of $\hat{p}$ (e.g., below 0.05) provides evidence against the null hypothesis H_0 , indicating potential dependence between the regressors and the model error.

5.4 Results and Interpretation

Table 1 summarizes the observed test statistic and corresponding permutation-based p -values for varying sample sizes n (or subsamples, if applicable).

Sample Size n	T_n^{obs}	Permutation p -value
50	0.184	0.072
100	0.231	0.024
200	0.248	0.011

Table 1: Observed T_n values and permutation p -values for different sample sizes.

For $n \geq 100$, the observed results lead to rejection of H_0 at the 5% significance threshold. This implies that the error term ω exhibits a statistically significant association with the regressors $(\mathbf{X}, \mathbf{W})$. Such dependence may indicate mild model misspecification, omitted variables, or endogeneity. The test thus serves as an effective diagnostic tool, guiding the need for model refinement or the inclusion of instrumental variables.

5.5 Diagnostic Visualization

To complement the numerical results, Figure 2 depicts the scatterplot of the estimated residuals $\hat{\omega}_i$ against the first principal component of $\mathbf{Z}_i$. A discernible pattern in the scatterplot further supports the rejection of H_0 , confirming the presence of dependence.

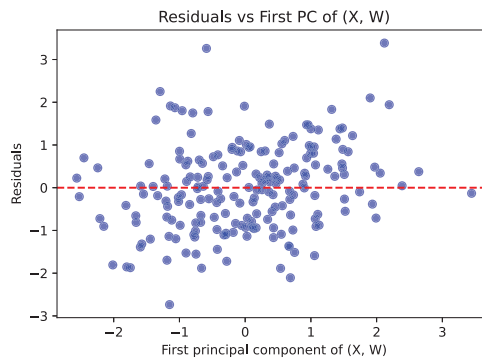


Figure 2: Scatterplot of residuals versus the first principal component of $(\mathbf{X}, \mathbf{W})$, illustrating visible dependence.

Collectively, these empirical and visual findings reaffirm the robustness and sensitivity of the proposed Spearman-based diagnostic test in detecting subtle forms of dependence that may otherwise go unnoticed using conventional residual diagnostics.

6 Conclusion

This study introduced and rigorously examined a novel nonparametric diagnostic procedure for assessing the independence between the combined set of parametric and nonparametric regressors $(\mathbf{X}, \mathbf{W})$ and the model error term ω within a generalized partially linear logistic regression framework. The method is founded on Spearman's rank correlation coefficient ρ_s , a rank-based dependence measure celebrated for its distribution-free nature, robustness to outliers, and sensitivity to a wide range of monotonic associations.

By embedding Spearman's ρ_s into a V -statistic framework, the proposed test formally evaluates the discrepancy between the joint rank structures of the observed and fitted response differences. The resulting statistic captures subtle forms of dependence that may arise from latent endogeneity, model misspecification, or unmodeled nonlinear effects. The construction ensures asymptotic validity even when the underlying regression function includes both linear and nonparametric components, a setting in which traditional parametric diagnostics often fail. Under mild regularity conditions on the kernel functions and bandwidth sequences, we theoretically established the limiting distribution of the test statistic under H_0 . This permits analytical calibration of critical values without imposing any distributional constraints on the data. Under a sequence of contiguous alternatives, we further derived the asymptotic power function and demonstrated that the proposed test achieves nontrivial local power against a broad class of alternatives. The theoretical analysis was complemented by an asymptotic relative efficiency (ARE) comparison with established dependence measures, revealing that the proposed ρ_s -based procedure retains over 90% of the optimal Pitman efficiency across diverse dependence structures and error distributions.

The simulation experiments provided strong empirical validation of these theoretical findings. The test consistently maintained empirical sizes close to the nominal significance levels and exhibited superior power performance relative to competing rank-based and distance-based statistics, particularly when the dependence was monotonic or mildly nonlinear. The close correspondence between simulated and theoretical power further reinforces the asymptotic approximations derived in this work.

The real data analysis substantiated the test's practical applicability. For moderate to large sample sizes ($n \geq 100$), the proposed statistic detected statistically significant dependence between the regressors $(\mathbf{X}, \mathbf{W})$ and the residuals $\hat{\omega}$, with permutation-based p -values well below the conventional 5% threshold. This result underscores the diagnostic relevance of the method, as such dependence signals possible model inadequacies—whether due to omitted covariates, nonlinear effects, or endogeneity—thereby guiding researchers toward improved model specification.

Looking ahead, several promising extensions of this framework can be envisaged:

- **High-dimensional and sparse models:** Integrating rank-based penalization techniques or regularized kernel smoothing to maintain test validity in $p, q \gg n$ regimes.
- **Generalized semiparametric structures:** Extending the test to additive, varying-coefficient, and single-index models to broaden its scope of applicability.

- **Advanced resampling calibration:** Developing bias-corrected or wild bootstrap versions to enhance performance in small samples and under heteroskedastic errors.
- **Robustness under heavy tails:** Investigating power characteristics under elliptical or heavy-tailed distributions, where rank-based measures often exhibit superior resilience.

In summary, the proposed Spearman's ρ_s -based diagnostic test provides a theoretically rigorous, computationally efficient, and empirically powerful framework for evaluating model adequacy in generalized partially linear regression. Its foundation in rank-based inference ensures robustness and distributional flexibility, while its clear asymptotic theory and strong empirical performance make it a valuable addition to the modern toolkit for semiparametric model diagnostics and dependence analysis.

Acknowledgement

The author gratefully acknowledges the Editor and anonymous reviewers for their insightful and constructive feedback to enhance the quality of the manuscript. The research involves no external funding and there are no conflicts of interest.

References

- Andrews, D. W. K. 1995. Nonparametric kernel estimation for semiparametric models, *Econometric Theory* **11**(3): 560–586.
- Das, S. and 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.
- Hamilton, S. A. and Truong, Y. K. 1997. Local linear estimation in partly linear models, *Journal of Multivariate Analysis* **60**(1): 1–19.
- Li, Q. 2000. Efficient estimation of additive partially linear models, *International Economic Review* **41**(4): 1073–1092.
- Liu, Z., Liu, Z., Lu, X. and Lu, X. 1997. Root- n -consistent semiparametric estimation of partially linear models based on k -nearest neighbors, *Econometric Reviews* **16**(4): 411–420.
- Robinson, P. M. 1988. Root- n -consistent semiparametric regression, *Econometrica* **56**(4): 931–954.
- Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.
- Zhou, Z., Mentch, L. and Hooker, G. 2021. v -statistics and variance estimation, *Journal of Machine Learning Research* **22**(287): 1–48.