IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/rjes

Examining of Air Pollutant Concentrations in Baghdad: A Semiparametric Approach Using Robust Estimation Techniques

Qutaiba N. Nayef Al-Qazaz[1], Ahmed Razzaq Abed [2]

[1] Department of Statistics / College of Administration and Economics / University of Baghdad / Iraq.
[2] Department of Statistics / College of Administration and Economics / University of Wasit / Iraq.

ARTICLE INFO

Article history:
Received 22/10/2024
Revised 22/10/2024
Accepted 14/1/2025
Available online 15/5/2025

Keywords:
Additive Partial linear Regression Model.
Generalized least-squares.
Restrictions.
Multicollinearity.
Ridge Estimators.
LTS- Estimators.
Local Polynomial Smoother.
Air Quality Index (AQI).

ABSTRACT

This article investigates the factors influencing air quality in Baghdad using a restricted partially linear additive regression model. The study addresses challenges such as multicollinearity and outliers by employing robust ridge estimates and integrating the Generalized Least Squares (GLS) method to account for heteroscedasticity. The Least Trimmed Squares (LTS) method is utilized to handle outliers by trimming the data, which enhances the accuracy of estimates. The nonparametric component of the model is smoothed using the Local Polynomial Estimator, improving the model's robustness and predictive performance.
Air quality data collected in the summer of 2023 was analyzed to assess the model's efficiency. Performance evaluation was conducted using the Mean Absolute Deviation (MAD) and Coefficient of Determination (R^2), demonstrating the model's effectiveness in estimating air quality. The results highlight significant non-linear relationships between key pollutants, particularly PM10 (particulate matter with a 10-micrometer diameter) and carbon dioxide (CO_2), and the Air Quality Index (AQI).
The findings underscore the crucial role of PM10 and CO_2 in influencing air quality in Baghdad. The study emphasizes the importance of implementing preventive measures to protect public health, given the substantial impact of these pollutants. The integration of robust statistical methods offers a comprehensive understanding of the factors affecting air quality and provides valuable insights for future air quality management and policy development in the region.

1. Introduction

The most valuable asset a person possesses is health, as it serves as the foundation for daily activities and the achievement of personal goals. With the increasing environmental challenges in modern times, air quality has become a critical factor affecting human health and well-being. Air pollution is a serious environmental problem impacting millions worldwide, linked to rising cases of respiratory and cardiovascular diseases. Therefore, this research aims to highlight the importance of studying air quality and the impact of air pollutants through the application of advanced statistical models, to provide solutions and measures to enhance public health and mitigate the negative effects of environmental pollution.

Semi-parametric regression models are characterized by their ability to integrate parametric and nonparametric components, offering greater flexibility compared to fully parametric models that require strict assumptions about the functional relationship. One of the challenges in these models is the selection of appropriate nonparametric functions, necessitating a delicate balance between bias, variance, and the complexity of the data.

When estimating the parametric component, the least squares method is the most commonly used due to its efficiency and ease of use; however, it relies on assumptions that may be violated in the presence of outliers. Another issue is multicollinearity among explanatory

* Corresponding author.E-**mail** address: dr.qutaiba@coadec.uobaghdad.edu.iq
https://doi.org/10.62933/rn535k81

variables, leading to inflated data and affecting the accuracy of estimates.

Several robust estimation methods differ in their approaches but share the commonality of utilizing weight matrices to reduce the impact of outliers. Numerous studies have demonstrated the effectiveness of combining these methods with biased estimators to address both outlier issues and multicollinearity in regression models. Therefore, we will employ an integrative approach that combines robust estimators based on the LTS method with non-random constraints imposed on the parametric part of the model to obtain high-efficiency estimates, facilitating the transition to estimating the nonparametric component of the restricted semi-parametric additive model, which has recently been studied by several researchers.

(Zhang and Huang, 2014) proposed a variable selection method for partially linear additive models using penalized spline regression to identify significant variables in longitudinal or matched data. (Gai et al., 2015) examined similar models with errors in linear variables, offering semi-parametric estimation through predictive least squares with strong spatial properties. (Emami, 2016) focused on linear interference effects in restricted estimates via semi-parametric models, proposing new diagnostic techniques to detect influential observations. (Roozbeh, 2016) suggested robust estimators for shrinkage parameters using the Least Trimmed Squares (LTS) method to address outliers. (Yang and Yang, 2017) introduced an effective method for estimating heteroscedasticity using local conditional regression in semi-parametric models. (Wu and Asar, 2017) developed a new restricted estimator based on a mixture of random estimation methods in partial linear models. (Jiang, 2017) created a robust estimator for partial models using an iterative algorithm and a generalized cross-validation method for parameter selection. (Roozbeh and Najarian, 2018) proposed a modified estimator using QR decomposition to handle multicollinearity in semi-parametric models. (Jiang et al., 2019) provided an effective method for semi-parametric models using squared exponential loss function. (El-Gohary et al., 2019) presented new estimators combining trimmed and spline estimates to tackle multicollinearity and outliers. (Abonazel and Gad, 2020) developed a robust version of the partial residual technique for estimating components of semi-parametric models. (Arzideh and Emami, 2022) proposed robust estimators using the LTS method with algorithmic enhancements to increase efficiency in the presence of outliers. (Kingsley and Fidelis, 2022) developed a new estimator combining M-estimation, principal components, and ridge estimation to resolve issues of multicollinearity and outliers. (Tang et al., 2022) discussed estimating non-parametric and semi-parametric effects in additive partial models. (Dai and Wang, 2023) introduced a generalized Liu-type estimator to address multicollinearity in partial logistic regression models. (Raad and Yousif, 2023) studied the relationship between factors affecting stock prices using partially linear additive models, focusing on efficiency in small to medium-sized samples. (Kuran and Yalaz, 2023) proposed new mixed estimators for semi-parametric models with measurement errors using kernel approximation.

(Ali & Kazem, 2023). This study applies the Weighted Least Trimmed Squares (WLTS) method to assess the impact of wastewater pollution on the Tigris River in Wasit Governorate. It focuses on Total Dissolved Solids (TDS) as the dependent variable, with covariates Sulphates, Chloride, and Phosphate. Data from 91 sites were analysed to inform environmental policies.

1.1 Partial Linear Additive Model (PLAM)

Integrates both parametric and nonparametric components, offering more flexibility in capturing complex relationships that purely parametric or nonparametric models cannot handle. This model extends the multiple linear regression framework, with unknown, one-dimensional nonparametric functions replacing linear terms to allow for both linear and nonlinear relationships.

A specific case of this model, the Semi-Parametric Additive Partial Linear model, allows some additive functions to remain linear while others are modeled nonparametrically, as described by Opsomer and Ruppert (1998). This model, represented as

$$Y = X'\beta + \sum_{d=1}^{D} g_d(z_d) + \epsilon \qquad (1)$$

has become particularly valuable for managing datasets where explanatory variables exhibit both parametric and nonparametric behavior. For example, when D=1, the model simplifies to a partial semi-parametric model, as demonstrated by Speckman (1988), which is easier to study and apply.

The nonparametric part of PLAM is estimated using kernel regression methods, where the weight matrix determines the importance of input data for the output. This weight matrix must satisfy several conditions, ensuring that it captures the relationship between variables efficiently and that the model can handle varying data dimensions without negative effects.

2. Generalized Least Squares Estimators (GLSE)

They are employed for estimating the parametric component. GLSE is considered the best unbiased linear estimator when the model does not suffer from any issues (Kutner and et al,1996) (Roozbeh, 2016).

$$Y = X\beta + \epsilon \qquad (2)$$

$$E(\epsilon) = 0, \quad var(\epsilon) = V.$$

That is, the mean of the errors is zero, and their variance V is not necessarily an identity matrix (there may be a correlation between the errors or a difference in their variances).

$$\hat{\beta}_{GLS} = \arg\min(\tilde{Y} - \tilde{X}\beta)'\, V^{-1}(\tilde{Y} - \tilde{X}\beta) \quad (3)$$

After simplifying the expression and taking the derivative and equaling it to zero

$$\hat{\beta}_{GLS} = (\tilde{X}'V^{-1}\tilde{X})\tilde{X}'V^{-1}\tilde{y}$$

$$\hat{\beta}_{GLS} = C^{-1}\tilde{X}'V^{-1}\tilde{y} \qquad (4)$$

where:

$$C = \tilde{X}'V^{-1}\tilde{X}.$$

V: is the variance-covariance matrix of the errors.

$$V\left(\hat{\beta}_{GLS}\right) = \sigma^2 C^{-1}.$$

Where $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_n)^T$ is the vector of the response variable with dimensions n*1. It is calculated based on the weight matrix W_{ni} according to the following formula:

$$\tilde{y}_i = y_i - \sum_{i=1}^{n}\sum_{j=1}^{d} W_{nj}(z_{ij})y_j \qquad (5)$$

In the case of studying only unmeasured variables, as we mentioned earlier, the formula is as follows:

$$\tilde{y}_i = y_i - [\sum_{i=1}^{n} W_{ni}(z_{i1})y_i + \sum_{i=1}^{n} W_{ni}(z_{i2})y_i] \qquad (6)$$

As for $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_n)^T$ it represents the explanatory variable vector and is calculated according to the following formula:

$$\tilde{x}_i = x_i - \sum_{i=1}^{n}\sum_{j=1}^{D} W_{ni}(z_{ij})x_j \qquad (7)$$

$$\tilde{x}_i = x_i - [\sum_{i=1}^{n} W_{ni}(z_{i1})x_i + \sum_{i=1}^{n} W_{ni}(z_{i2})x_i] \qquad (8)$$

$$W_{ni}(z_{i1}) = \begin{bmatrix} k_h(z_{11} - z_1) & 0 & \cdots & 0 \\ 0 & k_h(z_{21} - z_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k_h(z_{n1} - z_n) \end{bmatrix}$$

$$W_{ni}(z_{i2}) = \begin{bmatrix} k_h(z_{12} - z_1) & 0 & \cdots & 0 \\ 0 & k_h(z_{22} - z_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k_h(z_{n2} - z_n) \end{bmatrix}$$

K (z) = Kernel function.

h = Is (k×1) vector of bandwidth. (by Cross Validation).

2.1. Cross Validation:

The cross-validation method, also known as "Leave-One-Out," is a widely used approach for determining the optimal value of the smoothing parameter. This parameter plays a crucial role in balancing bias and variance in estimation; bias tends to increase with a higher smoothing parameter value, while variance increases as the value decreases. The fundamental idea of this method is to exclude one data point at a time when estimating the parameters using the following formula:

$$C.V.(h_n) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}^{(-i)} - \tilde{X}^{(-i)}\hat{\beta}^{(-i)})^2 \qquad (9)$$

Where $\hat{\beta}^{(-i)}$is calculated after excluding the observation (i) using estimated weights based on a Gaussian kernel. In the presence of multicollinearity or outliers, these estimates become ineffective, adversely affecting the accuracy of the nonparametric component estimation and increasing the mean squared error of the model.

2.2 Generalized Restrictions Least Squares Estimators: (GRLS)

The existence of non-random linear restrictions imposed on the model parameters is assumed, expressed as: (Roozbeh, 2016).

$$R\beta = r$$

Where R is a known matrix of rank (q×p) with $q < p$, its rows are full rank, and the number of rows equals the number of restrictions. The number of columns is equal to the number of model parameters.

r is a known vector with dimensions q×1, and its elements represent the fixed boundaries in the restriction.

The (GRLSE) can be expressed as follows:

$$\hat{\beta}_{GRLS} = argmin_{\beta} = (\tilde{y} - \tilde{x}\beta)'\, V^{-1}\,(\tilde{y} - \tilde{x}\beta) \qquad (10)$$

s.t $R\beta = r$

It can also be represented as:

$$\hat{\beta}_{GRLS} = \hat{\beta}_{GLS} - C^{-1}R'(RC^{-1}R')^{-1}\left(R\hat{\beta}_{GLS} - r\right) \qquad (11)$$

The variance-covariance matrix of the estimated parameters is calculated according to the GLS method:

$$V\left(\hat{\beta}_{GLS}\right) = \sigma^2 C^{-1} \qquad (12)$$

From equations (11) and (12), it is clear that the GLS estimators and their variance-covariance matrix rely heavily on matrix C. If C does not meet variance minimization conditions, the estimators may become error-prone, leading to statistically insignificant parameters and wide confidence intervals. To address this, resolving multicollinearity is essential. Common methods include Principal Component Analysis (PCA) and ridge regression. Ridge regression, in particular, is a key approach to mitigating multicollinearity in restricted partial least squares regression models.

2.3 Ridge Generalized Least Squares Estimators (RGLS)

Ridge Regression, proposed by Hoerl and Kennard in 1970, addresses the issue of multicollinearity by adding a small positive constant K to the diagonal elements of matrix C. This adjustment is represented as: (Gibbons,1981) (Hoerl and Kennard, 2000) (Hoerl and et al, 1975)

$$\beta_{GLS}(k) = C_k^{-1}\tilde{X}V^{-1}\tilde{Y} \qquad (13)$$

where

$C_k = C + KI_p$

The estimator of the parameters for generalized least squares is obtained by adding a non-negative value K ≥ 0. This value is selected to minimize the mean squared error (MSE), ensuring it is lower than the mean squared error obtained with generalized least squares estimators (GLS). K is referred to as the ridge parameter and can be estimated based on the actual data.

Ridge estimators are the most suitable for solving the multicollinearity problem, but they have some drawbacks, including instability. The estimated parameters may become unstable and biased (Roozbeh, 2016).

There exists an orthogonal matrix Γ such that $C = \Gamma\,\Omega\,\Gamma^{-1}$,where $\Omega = diag\left(\lambda_1 \dots, \lambda_p\right)$ is a diagonal matrix representing the eigenvalues of matrix C.

Therefore, the model (2.6) will become in the following form:

$$\tilde{Y} = \tilde{X}^{*}\alpha + \epsilon \qquad (14)$$

$$\tilde{X}^{*} = \tilde{X}\,\Gamma \quad , \quad \alpha = \Gamma'\,\beta.$$

When the matrix C suffers from multicollinearity among its columns, the Generalized Least Squares Estimators (GLS) tend to have large variances (Hoerl and Kennard, 1970) (Hoerl, Kennard, and Baldwin, 1975).

To make the matrix C free from multicollinearity issues, it is necessary to increase the values of the eigenvalues of the matrix (i.e., magnify them) by:

$$C(k) = C + KI_P \qquad , K > 0$$

This is the same process as replacing λ_i with $(\lambda_i + K)$. This replacement helps counteract the detrimental effect of eigenvalues that are close to or equal to zero (Hoerl, Kennard, and Baldwin, 1975). Now, the parameter K can be estimated using Generalized Least Squares Estimators in the partially Restricted ridge regression model, as follows (Swamy, 1978):

$$\widehat{K}_{LS} = \frac{P\,\widehat{\sigma}^{2}_{LS}}{\widehat{\beta}'_{GRLS}\,\widehat{\beta}_{GRLS}} \qquad (15)$$

$$\widehat{\sigma}^{2}_{LS} = \frac{1}{n-(p+q)}\,(\tilde{Y} - \tilde{X}\,\widehat{\beta}_{GRLS})'\,V^{-1}(\tilde{Y} - \tilde{X}\,\widehat{\beta}_{GRLS}) \qquad (16)$$

The Ridge Regression estimator is obtained by minimizing the sum of squared residuals under linear restrictions, thereby transforming the partially restricted ridge regression model to address multicollinearity effectively. (Kaciranlar and et al ,2011).

2.4. Ridge Generalized Restricted Least Squares Estimator:(RGRLS)

Another method has been developed to address the multicollinearity problem, which relies on combining the constraints imposed on the parameters with the ridge regression method to improve the performance of the estimators, (Roozbeh, 2016) and all of the following are:

$$\min(\tilde{Y} - \tilde{X}\beta)'V^{-1}(\tilde{Y} - \tilde{X}\beta)$$

s.t

$$\beta' \top \beta \leq \emptyset^{2}$$

$$R\beta = r$$

The results of the estimators are given by the following formula:

$$\hat{\beta}_{GRLS}(K) = (C + KI)^{-1}\tilde{X}'V^{-1}\tilde{Y} - (C + KI)^{-1}\,R'(R(C + KI)^{-1}R')^{-1}(R(C + KI)^{-1}\tilde{X}V_D^{-1}\tilde{Y} - r$$

$$\hat{\beta}_{GRLS}(K) = \beta_{GLS}(k) - C_k^{-1}R'(RC_k^{-1}R')^{-1}(R\beta_{GLS}(k) - r)$$

(17)

The above estimator is referred to as the Restricted Generalized Least Squares Ridge Estimator (RGRLS), and it can be expressed in another formula as follows:

$$\hat{\beta}_{GRLS}(k) = (I - C_k^{-1}R'(R\,C_k^{-1}\,R')^{-1}R)\beta_{GLS}(k) + C_kR'(RC_k^{-1}R')'r \qquad (18)$$

since $R(C_k^{-1}R'(R\,C_k^{-1}\,R')^{-1})R = R$

So, the generalized inverse of R, denoted as R^c, can be expressed by the following formula:

$$R^c = (C_k^{-1}R'(R\,C_k^{-1}\,R')^{-1}).$$

Therefore, the equivalent equation to the equation (18) is:

$$\hat{\beta}_{GRLS}(k) = (I - R^cR)\hat{\beta}_{GLS} + R^cr \qquad (19)$$

Now it is easy to see that $\hat{\beta}_{GRLS}(k)$ *and* $\hat{\beta}_{GRLS}$ are restricted estimators with the linear constraint Rβ = r, and when K=0, the Restricted estimators of the generalized ridge least squares will be equal to the Restricted estimators of the ridge least squares.

$$\hat{\beta}_{GRLS}(0) = \hat{\beta}_{GRLS}$$

2.4.1. Properties of Ridge Generalized Restricted Least Squares Estimator:

For any estimator, the Mean Squared Error (MSE) is computed as follows (Roozbeh, 2016) (Hassanzadeh and et al.,2011):

$$MSE\left(\hat{\beta}_{GRLS}(k)\right) = E\big[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)\big]$$

If β satisfies the linear restriction $R\beta = r$, the bias, variance, and mean squared error can be expressed as follows (M. Roozbeh, 2013) respectively:

$$Bais\left(\hat{\beta}_{GRLS}(k)\right) = E\{\hat{\beta}_{GRLS}(k) - \beta\} = -kL_k\beta \qquad (20)$$

$$cov\left(\hat{\beta}_{GRLS}(k)\right) = \sigma^2 L_k C L_k \qquad (21)$$

$$MSE\left(\hat{\beta}_{GRLS}(k)\right) = \sigma^2\,tr\,(L_kCL_k) + k^2\beta'L_k^2\beta \qquad (22)$$

Where:

$L_k = C_k^{-1} - C_k^{-1}R'(RC_k^{-1}R')^{-1}RC_k^{-1}$
$var(Y) = \sigma^2$
To calculate the risk function using Restricted optimization theory, the following formula is utilized (Roozbeh, 2016).

$$R(\hat{\beta}, \beta) = E\{(\hat{\beta} - \beta)'(\hat{\beta} - \beta)\}$$

$$R(\beta_{GRLS}(k), \beta) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i(\lambda_i + k - j_{ii})^2}{(\lambda_i + k)^2} + k^2 \left[\frac{\alpha_i(\lambda_i + k - j_{ii})}{(\lambda_i + k)^2}\right]^2 \quad (23)$$

where: λ_i represents the eigenvalues of the matrix C.
$R^* = \Gamma\, R'(RC_k^{-1}R')^{-1}R\, \Gamma.$
r_{ii}^* represents the diagonal elements of R^*.

$$\alpha_i = \Gamma'\beta = (\alpha_1, \alpha_2, \dots, \alpha_p)'$$

2.5. Robust Approach:

The concept of "robustness," introduced by Box (1953) and expanded by Tukey (1960), addresses the limitations of traditional estimators like the arithmetic mean, leading to alternative robust measures. While robustness theory now applies to areas like regression, no single method excels in all cases. Key criteria for evaluating estimators include the breakdown point (the smallest proportion of outliers impacting the estimator, up to 50%), efficiency (measured by the mean squared error ratio, ideally 90%-95%), ease of computation (simplicity and convergence), and inference (testing the method's appropriateness and parameter significance).

2.5.1. Least Trimmed Squares Estimator (LTS)

To address outliers, estimators like Least Trimmed Squares (LTS) minimize the sum of the first h ordered squared residuals, with $h = [\,n(1-\alpha) + 1]$, where α\alphaα is the trimming proportion, as proposed by Rousseeuw (1984) and Jung (2005). Typically, h is restricted between n and n/2, giving LTS a high breakdown points of up to 50% (Rousseeuw and Leroy, 1987). However, it is criticized for being computationally difficult, especially with large samples, and can have low efficiency, sometimes as low as 8%. In the semi-parametric restricted regression model (SRRM), Z_i can be used as an indicator to assess observation quality and express trimmed squared residuals.

$$\min_{\beta,z} \varphi(\beta, z) = (\tilde{Y} - \tilde{X}\beta)'V^{-\frac{1}{2}}ZV^{-\frac{1}{2}}(\tilde{Y} - \tilde{X}\beta) \quad (24)$$

s.t
$R\beta = r$
$e'z = h$
$z_i \in \{0,1\}$, i= 1, 2…, n.
when Z is a diagonal matrix with elements $z = (z_1, z_2, \dots, z_n)'$ and $e = (1,1,\dots,1)'_{n*1}$
The resulting estimator is the Restricted Least Trimmed Squares Robust Estimator (RLTS) in the semi-parametric regression model, and it is expressed by the following formula:

$$\hat{\beta}_{GRLTS}(z) = \hat{\beta}_{GLTS}(z) - C(z)^{-1}R'(RC(z)^{-1}R')^{-1}(R\hat{\beta}_{GLTS}(z) - r) \quad (25)$$

where:

$$C(z) = \tilde{X}'V^{-\frac{1}{2}}\,Z\,V^{-\frac{1}{2}}\tilde{X}$$

$$\hat{\beta}_{GLTS}(z) = C(z)^{-1}\tilde{X}'V^{-\frac{1}{2}}ZV^{-\frac{1}{2}}\tilde{Y} \quad (26)$$

Algorithm LTS:

1- Specify the trimming percentage (α) between 0 and 1. This determines the percentage of values to be trimmed.
2- Determine the number of observations to be trimmed (h) using the formula: h = [n(1-α) + 1] where n is the sample size.
3- Calculate the ordered data in descending order.
4- Identify the data to be trimmed based on the computed value of h.
5- Use only the trimmed data to estimate the parameters using any appropriate estimation method (e.g., least squares estimation).
6- Calculate the estimated parameters using only the trimmed data.
7- Repeat steps 1-6 iteratively using different sets of estimated values (using different values for α) to evaluate the robustness and effectiveness of the LTS estimator across various datasets.

2.5.2 Ridge estimates based on the robust approach

To address both outliers and multicollinearity in the semi-parametric restricted regression model, it is essential to find robust regression estimates. The first step in applying the ridge regression method is to estimate the ridge parameter K using robust methods, which involves replacing σ^2 in the model to estimate K (Roozbeh, 2016; Qazaz and Saleh, 2015).

The estimates $\hat{\beta}$ are obtained in two steps: first, by using the Truncated Least Trimmed Squares (LTS) approach to estimate K in the semi-parametric restricted regression model (RSRM), referred to as RRLTS. The equations are as follows:

$$\hat{K}_{LTS} = \frac{P\,\hat{\sigma}^2_{LTS}}{\hat{\beta}'_{RLTS}(z)'\,\hat{\beta}_{RLTS}(z)} \quad (27)$$

$$\hat{\sigma}^2_{LtS} = \frac{1}{n-(p+q)}\,(\tilde{Y} - \tilde{X}\,\hat{\beta}_{RLTS}(z))'\,V^{-\frac{1}{2}}ZV^{-\frac{1}{2}}(\tilde{Y} - \tilde{X}\,\hat{\beta}_{RLTS}(z)) \quad (28)$$

$$\hat{\beta}_{RRLTS}\left(\hat{k}_{LTS}, z\right) = \hat{\beta}_{GLTS}\left(\hat{k}_{LTS}, z\right) - C\left(\hat{k}_{LTS}, z\right)^{-1}R'(RC(\hat{k}_{LTS}, z)^{-1}R')^{-1}(R\hat{\beta}_{GLTS}\left(\hat{k}_{LTS}, z\right) - r) \quad (29)$$

where:

$$C\,(\hat{k}_{LTS}, z) = C(z) + \hat{k}_{LTS}I$$

$$\hat{\beta}_{GLTS}(\hat{k}_{LTS}, z) = C(\hat{k}_{LTS}, z)^{-1}\tilde{X}'V^{-\frac{1}{2}}ZV^{-\frac{1}{2}}\tilde{Y} \quad (30)$$

2.6. Local Polynomial Estimator:

The local linear regression is considered a good smoothing method because it has high efficiency compared to other smoothing methods. we take the model (1). (Li, 2007) (Speckman, 1988) (Liu and Yang,2010) (Raad, and Yousif, 2023).

$$Y = X^T\beta + g_1(Z_1) + g_2(Z_2) + \epsilon$$

the additive functions can be written as follows:

$$g_1 = \{g_1(Z_{11}), g_1(Z_{21}) \ldots g_1(Z_{n1})\}^T$$
$$g_2 = \{g_2(Z_{12}), g_1(Z_{22}) \ldots g_1(Z_{n2})\}^T$$

the backfitting algorithm is utilized for the model (1), assuming that s^T_{1,Z_2}, s^T_{2,Z_1}represent equivalent kernel functions for the local linear regression at Z_2 , Z_1 respectively. (Lexin Li and yin 2008).

$$s^T_{1,Z_1} = e^T_1\,(Z^T_1\,\omega_1\,Z_1)^{-1}Z^T_1\,\omega_1 \quad (31)$$

$$s^T_{2,Z_2} = e^T_1\,(Z^T_2\,\omega_2\,Z_2)^{-1}Z^T_2\,\omega_2 \quad (32)$$

$$e^T_1 = (1\,,0)$$

$$\omega_1 = \mathrm{diag}\left\{\frac{1}{h_1}K\left(\frac{Z_{11}-Z_1}{h_1}\right),\ldots\frac{1}{h_1}K\left(\frac{Z_{n1}-Z_1}{h_1}\right)\right\}$$

$$\omega_2 = \mathrm{diag}\left\{\frac{1}{h_2}K\left(\frac{Z_{12}-Z_2}{h_2}\right),\ldots\frac{1}{h_2}K\left(\frac{Z_{n2}-Z_2}{h_2}\right)\right\}$$

where $K(.)$ represents the kernel function, h_2, h_1 are the bandwidths, and Z_2, Z_1 are design matrices with dimensions $(n \times 2)$defined as follows:

$$Z_1 = \begin{bmatrix} 1 & Z_{11} - Z_1 \\ . & . \\ . & . \\ . & . \\ 1 & Z_{n1} - Z_1 \end{bmatrix}\ ;\ Z_2 = \begin{bmatrix} 1 & Z_{12} - Z_2 \\ . & . \\ . & . \\ . & . \\ 1 & Z_{n2} - Z_2 \end{bmatrix}$$

S_2, S_1 are smoother matrices representing or equating the kernel functions at the observations $(Z_{12}, \ldots Z_{n2})^t$, $(Z_{11}, \ldots Z_{n1})^t$, respectively.

$$S_1 = \begin{bmatrix} S_1Z_{11} \\ . \\ . \\ . \\ S_1Z_{n1} \end{bmatrix}\ ;\quad S_2 = \begin{bmatrix} S_2Z_{12} \\ . \\ . \\ . \\ S_2Z_{n2} \end{bmatrix}$$

When $\left\{S^C_1 = (I - 11^T/n)S_1\right\}$ denotes the centered smoothing matrix for $S_1$and $\left\{S^C_2 = (I - 11^T/n)S_2\right\}$ denotes the centered smoothing matrix for S_2.

1 a unit vector of dimension (n×1)" refers to a vector consisting of n rows, each having a value of 1.

Using the backfitting algorithm for the partially linear additive linear model to estimate both the parametric and non-parametric components is as follows:

$$\left.\begin{aligned} \hat{g}^{(m)}_1 &= S^C_1\left(Y - X\hat{\beta}_P - g^{(m-1)}_2\right) \\ \hat{g}^{(m)}_2 &= S^C_2\left(Y - X\hat{\beta}_P - g^{(m-1)}_1\right) \end{aligned}\right\} \quad (33)$$

$\hat{g}^{(m)}_1$ and $\hat{g}^{(m)}_2$ represent the estimators in the m^{th} stage of the backfitting algorithm. as a result, the non-iterative estimators for β take the form:

$$\hat{\beta}_0 = \{X^T(I - S_{12})X\}^{-1}X^T(I - S_{12})Y \quad (34)$$

Where:

$$S_{12} = \left\{I - \left(I - S^C_1\,S^C_2\right)^{-1}\left(I - S^C_1\right)\right\} + \left\{I - \left(I - S^C_2S^C_1\right)^{-1}\left(I - S^C_2\right)\right\} \quad (35)$$

To ensure that $\hat{\beta}_o$ is a consistent estimate of the root of n within the necessary smother by removing the restriction using the likelihood form procedure, the basic idea can be described as follows:

Let $\hat{g}_2(\beta, Z_2)$, $\hat{g}_1(\beta, Z_1)$be the backfitting estimates for $g_2(Z_2)$, $g_1(Z_1)$ respectively, as in formula (33), except replacing $\hat{\beta}$ by β. and ($\hat{g}_1$, $\hat{g}_2$)can be expressed as follows:

$$\left.\begin{aligned} \hat{g}_1(\beta) &= \left\{\begin{matrix} I - \left(I - S_1^C S_2^C\right)^{-1} \\ \left(I - S_1^C\right) \end{matrix}\right\}(Y - X\beta) \\ \hat{g}_2(\beta) &= \left\{\begin{matrix} I - \left(I - S_2^C S_1^C\right)^{-1} \\ \left(I - S_2^C\right) \end{matrix}\right\}(Y - X\beta) \end{aligned}\right\} \quad (36)$$

Now, substituting $\hat{g}_1(\beta)$, $\hat{g}_2(\beta)$ into model (1) and using the least square method, we obtain estimates based on the β formula of the form:

$$\hat{\beta}_f = \{X^T(I - S_{12})(I - S_{12})^T X\}^{-1} X^T(I - S_{12})(I - S_{12})^T Y \quad (37)$$

as discussed by Hastie and Tibshirani (1990), Opsomer and Ruppert (1999), centering each $[S_1^C , S_2^C]$ is necessary to ensure the convergence of the algorithm and the estimator $\hat{\beta}_f$, and is well defined by the assumption that $\sum_{i=1}^{n} g_1(Z_{i1}) = \sum_{i=1}^{n} g_2(Z_{i2}) = 0$

usually, the optimal bandwidth is$(n^{-1/5})$. This means that the estimator $\hat{\beta}$ is consistent for $\sqrt{n}$.

2.7 Real Data: Air Quality and Public Health Impact

Air quality is deteriorating globally, especially in Iraq, due to rising emissions. The WHO reports that 98% of the global population is exposed to air pollution, causing around 7 million deaths annually. Air pollution is linked to major health issues like stroke, heart disease, respiratory disorders, and cancer, while also harming ecosystems and economies. (Reports from the world health organization (WHO).

Air quality data for Baghdad was collected over 46 days in the summer of 2023 from platforms like AccuWeather, Tomorrow.io, and IQAir. These measurements, including PM_{10} and $PM_{2.5}$, were sourced from the Global Burden of Disease project and other regional networks. Government of the United Arab Emirates (2021).

2.8. Description of data

- PM_{10}: Particles up to 10 micrometers in diameter.
- $PM_{2.5}$: Fine particles up to 2.5 micrometers.
- CO_2: Carbon dioxide.
- CO: Carbon monoxide.
- NO_2: Nitrogen dioxide.
- O_3: Ozone.
- Temperature: Ambient temperature.

These pollutants are essential for calculating the Air Quality Index and assessing health impacts, as noted by the UAE National Air Quality Agenda 2031.

2.9. Data modelling:

To model the data in a semi-parametric regression model, model (1). the variables are defined as follows: The Air Quality Index (AQI) is the response variable (Y), and the explanatory variables include (PM_{10}, $PM_{2.5}$, CO_2, CO, NO_2, O_3, Temp), because the difference in measurement units between the data of the parametric variables, we converted the data to the standard format and to determine which explanatory variables are parametric variables or non-parametric, we draw the variables to see if their relationship is linear with the response variable. if the relationship is linear, it means that the variable is a parametric variable, as shown in Figure (1) indicates that each of the variables (PM_{10}, $PM_{2.5}$, CO_2, CO, Temp) has a somewhat linear relationship with the response variable, and therefore, they are considered parametric variables.

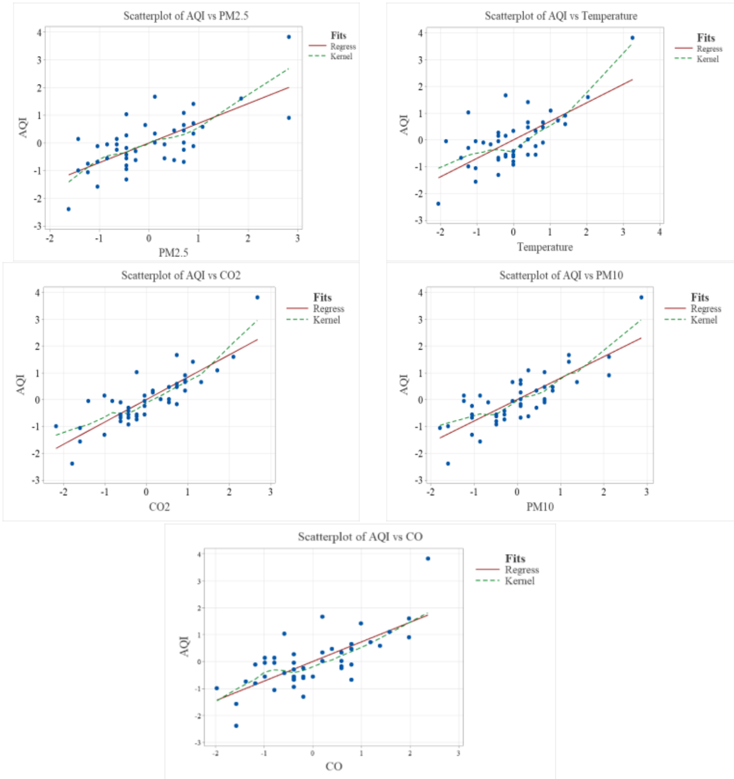

Figure (1): illustrates the type of relationship and correlation between each of (PM_{10}, $PM_{2.5}$, CO_2, CO, Temp) with the response variable (Y).

As for the variables (NO_2, O_3), they are non-linear variables, as shown in figure (2), indicating the type of relationship between them and the response variable, they are considered non-parametric variables.

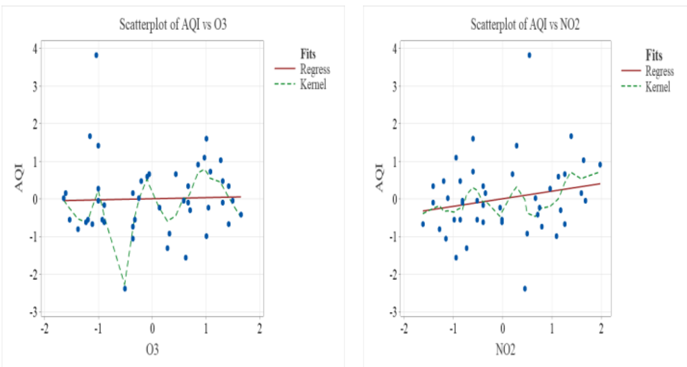

Figure (2) shows the type of relationship and dispersion between both (NO_2, O_3) variables and the response variable (Y).

Therefore, the dataset is modeled using the partial least squares regression model.

$$(AQI)_i = \beta_1(PM10)_i + \beta_2(PM2.5)_i + \beta_3(CO2)_i + \beta_4(CO)_i + \beta_5(Temp)_i + g_1(NO2)_i + g_2(O3)_i + \varepsilon_i \quad (38)$$

2.10. Multicollinearity and Outliers test:

The correlation matrix revealed strong multicollinearity among variables, as in table (1), confirmed by the eigenvalues of the information matrix, with a Condition Number of 33.49**,** in table (2) indicating a significant issue.

Table 1: Correlation Matrix

	Temp	$PM_{2.5}$	PM_{10}	CO_2	CO
Temp	1	0.83213	0.65980	0.80386	0.82244
$PM_{2.5}$	0.83213	1	0.86159	0.77529	0.90963
PM_{10}	0.65980	0.86159	1	0.84527	0.77484
CO_2	0.80386	0.77529	0.84527	1	0.87506
CO	0.82244	0.90963	0.77484	0.87506	1

Table 2: Eigenvalues of the Information Matrix ($X^T X$)

λ_1	λ_2	λ_3	λ_4	λ_5
192.018496	15.7154723	10.1684286	6.7552805	0.1711613

$$C.N = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{192.018496}{0.1711613}} = 33.49413$$

The Variance Inflation Factor (VIF) values also showed high multicollinearity, with PM2.5 having a VIF of 49. as in table (3), To address this, ridge regression and non-random restrictions to improve estimator efficiency.

Table 3: Values of the Variance Inflation Factor (VIF) for Independent Variables

Temp	$PM_{2.5}$	PM_{10}	CO_2	CO
11.22382	49.22341	25.42113	31.33400	28.43729

An F-test was conducted to verify the imposed restrictions, with the calculated F value of 3.0084, which is lower than the tabulated value of 3.23, leading to the acceptance of the null hypothesis ($H_0: R\beta = 0$).

$$R = \begin{bmatrix} 1 & 2 & 0 & 3 & 2 \\ 1 & 3 & 2 & 2 & 5 \end{bmatrix}; r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$F = \frac{[(R\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{r})' (\sigma^2_{GLS} R\tilde{X}'\tilde{X}R')^{-1} (R\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{r})]/q}{\hat{\sigma}^2}$$

where: $\hat{\Sigma} = \hat{\sigma}^2_{GLS} \, (\tilde{X}'\tilde{X})^{-1}$

$$F = \frac{7.521/2}{0.8} = 3.0084$$

$F_{tab} = F_{(2,40,0.05)} = 3.23$. Since the calculated value is smaller than the tabulated value, H_0 is accepted.

Outlier detection using Studentized Deleted Residuals (SDR) identified four outliers, representing 8.7% of the sample. Therefore, robust estimation methods, such as partial least squares regression with imposed restrictions, should be used to achieve optimal estimates and a well-fitted model.

2.11. Estimation:

Perform the smoother process after finding the parameter estimates for the parametric part of the model, we use Local polynomial estimator (LPE). Which is the best smoother because it combines the flexibility of non-parametric methods with the properties of least squares method parametric. All this is done using ($R_{4.3.2}$) programming with ready-made functions and packages**.**

Table (4) Model estimates using parametric methods, the Local polynomial method, and comparison criteria.

Methods	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$
GLS	0.357	0.369	0.513	0.350	0.056
GRLS	0.618	0.532	0.689	0.275	0.190
RGLS	0.236	0.107	0.336	0.442	0.022
RGRLS	0.558	0.395	0.644	0.225	0.222
RRLTS	0.579	0.441	0.662	0.243	0.213

Table (4) presents the parameter estimates for the semi-parametric restricted partially additive regression model using various methods. The parametric part was estimated using several approaches: Generalized Least Squares (GLS), Generalized Restricted Least Squares (GRLS), Ridge Generalized Least Squares (RGLS), Ridge Generalized Restricted Least Squares (RGRLS), and Ridge Restricted Robust Least Trimmed Squares (RRLTS). For example, the estimates for β_1 range from 0.23663 (RGLS) to

0.61821 (GRLS), illustrating how different methods impact the parameter estimates.

2.12. Comparison Criteria:

For the non-parametric part, Local polynomial estimator (LPE) was used, the Gaussian kernel function was employed, and the bandwidth was chosen using cross-validation. (Härdle,1994).

Gaussian kernel function: $k(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2) \quad I\,(|z| < \infty)$.

Model estimation methods were compared using the coefficient of determination (R-squared) and Mean Absolute Deviation (MAD). MAD was utilized to assess the sensitivity of the mean square error metric in the presence of outliers in the response variable. The formulas are as follows: (Rousseeuw,1987).

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \qquad (39)$$

Table (5): Represents the comparison criteria for the model using the LPE approach.

Methods	GLS	GRLS	RGLS	RGRLS	RRLTS
MAD	0.53069	0.51278	0.51551	0.48117	0.42279

The results in Table (5) demonstrate the effectiveness of the integrative approach that combines the robust ridge estimator based on the LTS method with the non-random constraints imposed on the parameters. The positive characteristics of the parameters were reflected in the smoothing results, making the calculations more efficient and reducing the time required to obtain the results.

By examining the estimated parameters for the parametric part, it is evident that the second independent variable, $PM_{2.5}$, has an inverse relationship with the response variable. Additionally, the variable PM_{10}, representing particulate matter with an aerodynamic diameter of up to 10 Micrometer, has a more pronounced impact on air quality than other variables. Furthermore, there are noticeable nonlinear effects of non-parametric variables.

To ensure the suitability of the estimation methods used for the parametric part in finding the parameters of the non-parametric part, this was illustrated through Figure (3), which shows the dispersion between the two non-parametric variables (O_3, NO_2) and the response variable (AQI).

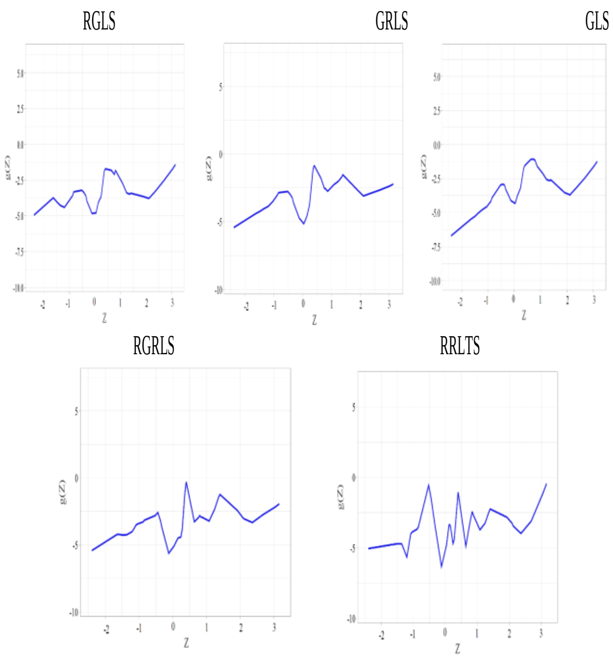

Figure (3) Represents the behaviour of the two non-parametric variables (O_3, NO_2) with the response variable.

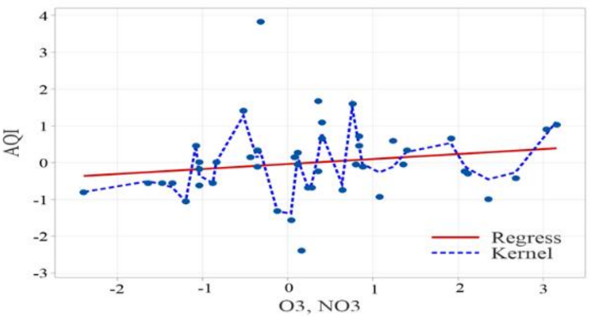

Figure (4) Represents the behaviour of the two non-parametric variables (O_3, NO_2) with the estimated imputation functions in (LPE)

Figure (3) and (4) demonstrate that the method combining non-random restriction and robust ridge regression was more suitable than the other approaches when using its estimates to find smoother function estimates. The model estimates was also suitable for the observed phenomenon data, as shown by the estimated values of the model and their consistency with actual observations, as depicted in Figure (5).

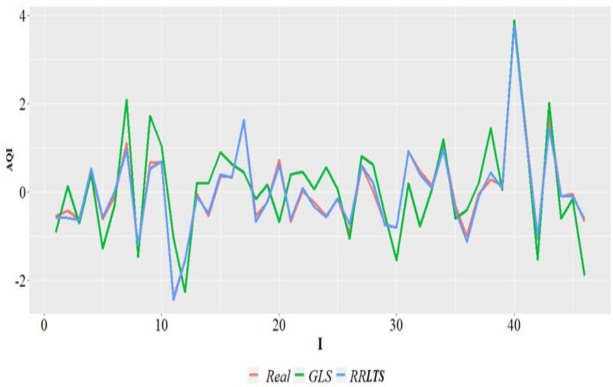

Figure (5) illustrates the estimated values of the model using the Local polynomial estimator (LPE).

3. Discussion of results

The results indicate that the Ridge Restricted Robust Least Trimmed Squares (RRLTS) method outperformed all other approaches in parameter estimation, achieving the lowest Mean Absolute Deviation (MAD) of 0.42279. These findings suggest that the RRLTS method effectively addresses the issues of multicollinearity and outliers, leading to greater accuracy in the estimates.

Conversely, the Generalized Least Squares (GLS) method exhibited the poorest performance, with the highest MAD of 0.53069, indicating its inadequacy in handling complex data characterized by multicollinearity and outlier effects.

Other methods, such as Ridge Generalized Restricted Least Squares (RGRLS), demonstrated acceptable performance with a MAD of 0.48117, but they did not reach the efficiency levels of RRLTS. This underscores the advantage of employing robust estimators with non-random restrictions, as evidenced by the superior results obtained from the RRLTS approach in this context.

4. Conclusions

The study on air quality in Baghdad, utilizing the restricted linear partial additive model, revealed that the air quality index (AQI) contained outliers, with 4 out of 46 observations classified as outliers, constituting 8.7% of the sample.

Graphical analysis demonstrated that PM_{10}, $PM_{2.5}$, CO_2, CO, and Temp followed a linear relationship with AQI, while NO_2 and O_3 exhibited non-linear behavior.

Using Mean Absolute Deviation (MAD) as evaluation criteria, the study identified the integrated approach combining the robust estimator based on the Least Trimmed Squares (LTS) with non-random restrictions imposed on the parametric portion of the model as the most effective.

Furthermore, the PM_{10} variable had the most significant impact on air quality, followed by CO_2 emissions from vehicles. A non-linear relationship was observed between O_3 and NO_2 with AQI, particularly during the summer, posing significant health risks.

These findings enhance the understanding of the complex dynamics affecting air quality in Baghdad and underscore the importance of employing robust modeling techniques that integrate robust estimators with non-random restrictions for accurate analysis and policy recommendations.

References:

1. Abonazel, M. R, and Gad, A. A. E. (2020). Robust partial residuals estimation in semiparametric partially linear models. *Communications in Statistics - Simulation and Computation*, 49(5), 1223-1236.
2. Ali, O. A., & Kazem, K. J. (2023). Weighted Least Squares Estimation of the Effect of Wastewater Pollution of Tigris River / Wasit Governorate. *Journal of Engineering and Applied Sciences*, 24(109). *https://doi.org/10.33095/jeas.v24i109.1573.*
3. Arashi, M, and Roozbeh, M. (2013). Feasible ridge estimator in partially linear models. *Journal of Multivariate Analysis*, 114, 35-44.
4. Arzideh, K, and Emami, H. (2022). Robust ridge estimator in censored semiparametric linear models. *Communications in Statistics - Theory and Methods*. *https://doi.org/10.1080/03610926.2021.2023573.*
5. Dai, D, and Wang, D. (2023). A generalized Liu-type estimator for logistic partial linear regression models with multicollinearity. *AIMS Mathematics*, 5, 11851–11874. *https://doi.org/10.3934/math.2023600.*
6. El-Gohary, M, Abonazel, R., Helmy, M. M., and Azazy, R. A. (2019). New robust-ridge estimators for partially linear models. *International Journal of Applied Mathematical Research*, 8(2), 46-52. *https://doi.org/10.14419/ijamr.v8i2.29932.*
7. Emami, H. (2016). Local influence in ridge semiparametric models. *Journal of Statistical Computation and Simulation*. *https://doi.org/10.1080/00949655.2016.1152273.*
8. Gai, Y., Zhang, J, Li, G, and Luo, X. (2015). Statistical inference on partial linear additive models with distortion measurement errors. *Statistical Methodology*, 20-38. *https://doi.org/10.1016/J.STAMET.2015.05.004*
9. Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373), 131–139. *https://doi.org/10.2307/2287058.*
10. Härdle, W, Linton, O. (1994). Applied nonparametric methods. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics* (pp. 2295-2339).
11. Härdle, W, Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4), 1926-1947.
12. Härdle, W, Müller, M, Sperlich, S, and Werwatz, A. (2004). *Nonparametric and semiparametric models*.
13. Härdle, W., Liang, H, and Gao, J. (2000). *Partially linear models*. Springer Science and Business Media.
14. Hoerl, A. E, and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12(1), 69-82.
15. Hoerl, A. E, and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1), 80–86. *https://doi.org/10.2307/1271436.*
16. Hoerl, A. E, Kennard, R. W, and Baldwin, K. F. (1975). Ridge regression: Some simulation. *Communications in Statistics - Simulation and Computation*, 4, 105-123.
17. Jiang, Y, Tian, G. L, and Fei, Y. (2019). A robust and efficient estimation method for partially nonlinear models via a new MM algorithm. *Statistical Papers*, 60, 2063–2085. *https://doi.org/10.1007/s00362-017-0909-5.*
18. Jiang, Y. (2017). S-estimator in partially linear regression models. *Journal of Applied Statistics*, 6, 968-977. *https://doi.org/10.1080/02664763.2016.1189523.*
19. Kaciranlar, S, Sakallioglu, S, Ozkale, M. R., and Guler, H. (2011). More on the restricted ridge regression estimation. *Journal of Statistical Computation and Simulation*, 81, 1433–1448.
20. Keller, G, and Warrack, B. (2000). Statistics for Managing and Economics (5th ed.).
21. Kingsley, C. A, and Fidelis, I. U. (2022). Combining principal component and robust ridge estimators in linear regression models with multicollinearity and outliers. *Concurrency and Computation: Practice and Experience*. *https://doi.org/10.1002/cpe.6803.*
22. Kuran, Ö, and Yalaz, S. (2023). Kernel mixed and kernel stochastic restricted ridge predictions in partially linear mixed measurement error models: An application to COVID-19. *Journal of Applied Statistics*, 1-25.
23. Li, L, and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64(1), 124-131.
24. Liu, R, and Yang, L. (2010). Spline-backfitted kernel smoothing of additive coefficient models. *Econometric Theory*, 26(1), 29-59.
25. Liu, R. Y, and Yang, L. (2009). Efficient estimation of a semiparametric partially linear varying-coefficient model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 447-466.

26. Opsomer, J. D, and Kauermann, G. (2004). Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models. *Journal of Computational and Graphical Statistics*, 13(1), 66-89.
27. Opsomer, J. D, and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, 93(443), 605-619.
28. Opsomer, J. D, and Ruppert, D. (1999). A root-inconsistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8(4), 715-732.
29. Rousseeuw, P. J, and Leroy, A. M. (2005). *Robust regression and outlier detection*. John Wiley and Sons.
30. Rousseeuw, P. J, and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.
31. Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871-880.
32. Swamy, P. A. V. B, and others. (1978). Two methods of evaluating Hoerl and Kennard’s ridge regression. *Communications in Statistics - Theory and Methods*, 7(11), 1133-1155.
33. Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 448-485).
34. World Health Organization - Air quality guidelines (2023).
35. Wu, J. and Asar, Y. (2017). A weighted stochastic restricted ridge estimator in partially linear models. *Communications in Statistics - Theory and Methods*. *https://doi.org/10.1080/03610926.2016.1206936*.
36. Yang, J. and Yang, H. (2017). Robust modal estimation and variable selection for single-index varying-coefficient models. *Communications in Statistics - Simulation and Computation*, 4, 2976-2997. https://doi.org/10.1080/03610918.2015.1069346.
37. Zhang, W, and Huang, J. (2014). Variable selection in partially linear additive models for longitudinal/clustered data. *Journal of the Royal Statistical Society*.