Journal of Economics and Administrative Sciences
2024; 30(143), pp. 435-454

P-ISSN 2518-5764
E-ISSN 2227-703X

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

Robust Ridge-MM Estimator in Restricted Additive Partially Regression Model

Ahmed Razzaq Abed*
Department of Statistics
College of Administration and Economics
University of Wasit / Iraq
ahmedrazzaq@uowasit.edu.iq
***Corresponding author**

Qutaiba N. Nayef Al-Qazaz
Department of Statistics
College of Administration and Economics
University of Baghdad / Iraq
dr.qutaiba@coadec.uobaghdad.edu.iq

Received:15/1/2024 Accepted:4/3/2024 Published Online First: 1 /10/ 2024

Abstract:

This paper, utilized the Restricted Partially Additive Regression Model to analyze air quality data in Baghdad governorate, with a focus on addressing multicollinearity issues among independent variables and outliers in the dependent variable. Through the implementation of classical estimators, ridge estimators, robust estimators, and the imposition of non-random constraints on the parametric parts of the model Through method of Robust Ridge-MM Estimator in Restricted Additive Partially Regression Model, the study aimed to assess the model's effectiveness in dealing with air pollution challenges during the summer season. Results obtained through the use of pre-built packages and algorithms in the R programming language indicated that integrating non-random constraints with robust estimators positively impacted the accuracy of estimating functions. Furthermore, certain variables, such as PM10 (airborne particles with an aerodynamic diameter of up to 10 micrometres), were found to have a significant impact on air quality This is through the parameter values. Non-linear effects were observed for some non-parametric variables. The study highlights the importance of understanding the effects of air pollutants on public health and emphasizes the urgent need for quick solutions to mitigate these negative effects.

Paper type: Research paper.
Keywords: Additive Partial Regression Model, Generalized least-squares, Restrictions, Multicollinearity, Ridge Estimators, MM- Estimators, Local Polynomial Smoother, Air Quality Index (AQI).

1. Introduction:

When reviewing contemporary literature, it has been demonstrated that the semiparametric regression is of great value in many applications across various fields such as space science, medicine, economics, and others. This is achieved through the utilization of semiparametric regression models. the term 'semiparametric models' is attributed to (Oakes,1981; Begun et al., 1983), the parameters for the infinite-dimensional variables (the nonparametric component) and the limited-dimensional parameters of interest (the parametric component) are considered. Although this distinction is a fundamental feature of semiparametric modelling, it seems comprehensive in itself. Many issues that are typically considered 'nonparametric' or 'parametric' can be classified as 'semiparametric' according to this framework. There is a middle ground between fully parametric and nonparametric models provided by semiparametric techniques. The semiparametric approach makes assumptions about m(t) less restrictive than those of a fully parametric model, yet it is more powerful than those of nonparametric estimation.

as a result, semiparametric methods are capable of mitigating the impact of the dimensionality problem in estimation, enhancing the precision of estimation compared to nonparametric estimation while allowing for greater flexibility and lower risks of model misspecification errors than what is possible with a fully parametric model. Compared to nonparametric methods, dimensionality reduction and higher precision in estimation serve as primary justifications for semiparametric approaches.

Sometimes, the parametric component of the semiparametric model faces challenges, and one of these challenges is the presence of correlation among some explanatory parametric variables, leading to the issue of linear multicollinearity. This issue has negative effects on the model estimation process. Additionally, apart from the problem of linear multicollinearity, other issues arise in the model, especially when some observations of the response variable follow a non-normally distributed distribution, referred to as outlier observations. The existence of these outlier values in the response variable affects the estimation of both parametric and nonparametric components in the model, thereby influencing the error boundary and resulting in the problem of heteroscedasticity .other issues arise in the model, especially when some observations of the response variable follow a non-normally distributed distribution, referred to as outlier observations. The existence of these outlier values in the response variable affects the estimation of both parametric and nonparametric components in the model, thereby influencing the error boundary and resulting in the problem of heteroscedasticity.

Most prior studies have concentrated on the special case of partially additive linear models. For instance, Jean and David (1999) studied additive models that combine parametric and nonparametric terms, proposing a consistent $\sqrt{n}$ backfitting estimator for the parametric component of the model. They achieved this by developing a fast implementation algorithm for model fitting and selecting package width through simulation experiments. Li (2000) suggested an estimation of the additive partially linear regression model using a generalized sequential method based on boundary-adaptive splines, which proved to be more efficient in approximating the semiparametric estimator, ignoring the additive structure.

Liang et al. (2008) studied the standard additive model, a generalization of multiple linear regression models, extensively. To strike a balance between the interpretability of linear models and the flexibility of additive models, they developed well-behaved partially additive semiparametric models.

Heng and Hua (2013) studied into high-dimensional generalized partially additive linear models with shared variables. They attempted to identify which components (including both parametric and nonparametric components) were non-zero. They proposed employing double-penalized functions to obtain initial estimates and then using a shrinkage factor and adaptive least angle selection to identify non-zero components. Through simulation studies, they demonstrated that the proposed procedure worked well with moderate sample sizes.

Several researchers addressed the issue of imposed random constraints on semiparametric models, such as Jibo and Yasin (2017) reviewed the estimation of partially linear models when random linear constraints on the parametric components are assumed to exist. Based on the weighted mixed estimator, the least squares method for the profile, and the ridge method, they introduced the constrained weighted random estimator for the parametric component. They also discussed the properties of the new estimator. Finally, a simulation study was conducted to illustrate the performance of the new estimator.

Presenting the latest studies on ridge and robustness estimators, such as Kingsley and Fidelis (2022), suggested a new estimator to jointly address the issue of multicollinearity and extreme values by combining the following estimators: M-estimator, principal components, and ridge estimator. The new estimator is called the robust r-k estimator and is used in the partially semiparametric model. They theoretically demonstrated that the new estimator outperforms some existing estimators, supported by simulation studies and real-world applications showcasing the efficiency of the new method.

In addition, Dayang and Dabuxilatu (2023) suggested a generalized Liu-type estimator (GLTE) to address the issue of multicollinearity in the linear part of the partially logistic linear regression model. Using the maximum likelihood method, the researchers suggested GLTE as a general formula for a Liu-type estimator, including the maximum likelihood estimator, ridge estimator, Liu estimator, and Liu-type estimator as special cases. They derived the conditional superiority of the proposed GLTE over other estimators under the mean squared error matrix approximation (MSEM) criterion. Moreover, optimal choices for biasing parameters and biasing function were presented. This was illustrated through numerical simulation processes, showing that the performance of the proposed GLTE was superior to current estimators. They further demonstrated the application of a dataset arising from a study of Indian liver disease patients to clarify their theoretical results.

Talib and Hmood (2022) studied investigated the relationship between variables related to factors affecting stock prices for Pepsi Baghdad. They reviewed Partially Linear Additive Models as an effective tool for representing these relationships, whether linear or nonlinear, especially in cases that involve both types. The simulation results were analyzed using various criteria. The findings demonstrated that the Spline Approximation method was the most effective in model estimation. The Adaptive Lasso method showed high efficiency in cases of weak relationships and small sample sizes. The SCAD method proved its efficiency in cases of moderate correlation and small sample sizes. The MAVE Lasso method demonstrated effectiveness in large sample sizes. In cases of a strong relationship with explanatory variables and small sample sizes, the MAVE Elastic Net method was effective, while the MAVE Adaptive Elastic Net method showed efficacy in large sample sizes.

The problem of the research is dealing with the analysis of air quality data in Baghdad Governorate. The challenges include the presence of multicollinearity and outlier values in the data, requiring an enhancement of the model's accuracy to better understand air quality and address pollution.

The objective of the research is to analyze air quality data in Baghdad Governorate using the Restricted Additive Partially Regression Model. The research focuses on addressing multicollinearity and outliers in the data, evaluating the model's performance in dealing with air pollution challenges, especially during the summer season. The results indicate that integrating non-random constraints with robust ridge positively impacts the accuracy of embedding functions. The study highlights the importance of understanding the impact of air pollutants on public health and underscores the urgent need for effective solutions.

2.Material and Methods:

2.1 Additive Partially Linear Model (APLM):

It is a generalization of multiple linear regression models that replaces one-dimensional non-parametric functions for linear components, used to investigate the complex relationship between treatment response and potential predictors (Stone, 1977). Additionally, efforts have been made to find a compromise between the flexibility of additive models and the interpretability of linear models through partially additive linear models. In these models, some of the additive component functions are linear, while the remaining functions are treated as non-parametric (Opsomer and Ruppert, 1999). The additive partially linear model is defined as follows:

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

Where $X = (x_1, x_2, \dots, x_p)'$ represents the linear variables, and $Z = (z_1, z_2, \dots, z_D)'$ represents the non-parametric variables. Additionally, $(f_1, f_2, \dots, f_D)$ are the unknown smoothing functions, and $\beta = (\beta_1, \beta_2, \dots, \beta_p)$is the vector of unknown parameters, and ϵ is the error term with a mean of zero. The conditional distribution between the error term and the variables (X, Z) is equal to zero. The model (1) can be reformulated as follows in equation (2):

$$Y = X_i'\beta + g_1(z_{i1}) + g_2(z_{i2}) + \cdots + g_L(z_{iL}) + \epsilon_i \qquad (2)$$

Where Y is a vector representing the response variable with dimensions n×1. X_i represents the matrix of explanatory variables with dimensions (n×p), where p is the number of parameters. β represents the vector of unknown parameters with dimensions p×1, and $g_1(z_{1i})$ represents the additive component for the first unmeasured variable (z_{1i}). Similarly, $g_2(z_{2i})$ represents the additive component for the second unmeasured variable (z_{2i}), and ϵ_i represents the error term. The random errors are assumed to be independent of (X_i, Z_i).

E$[y \backslash x_i, z_i] = 0$ E$[\epsilon_i] = 0$, E$[\epsilon'\epsilon] = \sigma^2 V$.

When σ^2 is unknown, and V is a known matrix with dimensions n*n as a positive definite symmetric matrix (meaning all its eigenvalues are positive), the unmeasured component is estimated as follows:

$$\hat{g}_1(z_{i1}) = \sum_{i=1}^{n} W_{ni}(z_{i1})\,(y_i - X_i'\beta)$$

$$\hat{g}_2(z_{i2}) = \sum_{i=1}^{n} W_{ni}(z_{i2})\,(y_i - X_i'\beta)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{g}_L(z_{iL}) = \sum_{i=1}^{n} W_{ni}(z_{iL})\,(y_i - X_i'\beta)$$

When $W_{ni}(z_{1i})$,$W_{ni}(z_{2i})$, …,$W_{ni}(z_{Li})$] are positive weight functions, and each function satisfies the following conditions:

1- Max $_{1\leq i\leq n}$ $\sum_{j=1}^{n} W_{ni}(z_j) = 0(1)$.

2- Max $_{1\leq i,j\leq n}$ $W_{ni}(z_j) = 0\left(n^{-\frac{2}{3}}\right)$.

3- Max $_{1\leq i\leq n}$ $\sum_{i=1}^{n} W_{ni}(z_j)\ I(|z_i - z_j| > C_n) = 0(d_n)$.

When I represent the indicator function and C_n satisfies the following condition:

$$\lim sup_{n\to\infty} nC_n^3 < \infty$$

And d_n satisfies the following condition:

$\lim sup_{n\to\infty} n\, d_n^3 < \infty$

The semi-parametric additive partial linear model studied in this paper consists of two latent variables and is formulated as follows:

$$Y = X_i'\beta + g_1(z_{i1}) + g_2(z_{i2}) + \epsilon_i \qquad (3)$$

The model (3) is estimated by estimating the parametric part, which is linear and suffers from several issues. It is then employed to estimate the non-parametric part, followed by re-estimating the parametric part using the estimated non-parametric values. This process is repeated until the difference between the new estimates and the previous ones is close to zero. Additionally, there are outlier values in the model's response variable.

2.2 Model issues:

2.2.1Multicollinearity:

Multicollinearity occurs when two or more explanatory variables are highly linearly correlated, making it difficult to separate the effect of each variable on the response variable. This issue also arises when one of the explanatory variables has the same value for all observations or when one or more explanatory variables are linearly dependent on the studied model. In the presence of multicollinearity, applying the least squares method leads to a problem of variance inflation in the estimated regression coefficients. This is due to the singularity of the information matrix (X'X), resulting in an inflation in the diagonal elements of the (X'X) matrix. To address this problem, biased methods are used (Daoud, 2017; Kazem and Muslim, 2002).

Several methods or tests exist to detect multicollinearity among explanatory variables, one of which is the Condition Number measure. The Condition Number was introduced by Muir in 1981 and is primarily based on the eigenvalues of the explanatory variables matrix (X'X). It measures the sensitivity of regression estimates to small changes in variances. Another method is calculating the Variance Inflation Factor (VIF), which quantifies the extent of multicollinearity (Adeboye et al., 2014).

2.2.2 Outlier observations:

Bross (1961) defined an outlier as an observation that deviates significantly from the other components in the sample set where this observation was found, As for **(**Freeman, 1980), an outlier is defined as any observation that did not arise in the general manner in which the vast majority of data observations were generated. (Keller, G. and Warrack, B.2000), define an outlier as an observation that deviates significantly from the regression equation and has a large error compared to the other natural observations in the data. Therefore, it will have an impact on the model and its estimates. The causes of the appearance of outlier values are often related to the data having an asymmetrical distribution. Outliers can also occur due to errors made by the researcher when recording measurements or as a result of faults in measuring devices, especially in laboratory experiments, or due to errors in calculations, leading to the emergence of outlier observations.

2.3 Generalized Least Squares Estimator:

To estimate the parametric part, we use the Generalized Least Squares Estimators (GLSE), which are the best unbiased linear estimators when the model does not suffer from any issues (Kutner and at al., 2005;
Roozbeh, 2016).

$$\hat{\beta}_{GLS} = \arg\min(\tilde{Y} - \tilde{X}\beta)'\, V^{-1}(\tilde{Y} - \tilde{X}\beta) \qquad (4)$$

$$\hat{\beta}_{GLS} = C^{-1}\tilde{X}'V^{-1}\tilde{Y} \qquad (5)$$

Where:

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

Where $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_n)$ is the vector of the dependent variable with dimensions n*1, which is calculated based on the weight matrix W_{ni}.

W_{ni} is computed using kernel functions and bandwidth, according to the following formula:

$$\tilde{y}_i = y_i - \sum_{i=1}^{n}\sum_{j=1}^{k} W_{ni}(z_{ij})y_i \qquad (6)$$

In the case of studying only non-parametric variables, as mentioned earlier, the formula is as follows:

$$\tilde{y}_i = y_i - [\sum_{i=1}^{n} W_{ni}(z_{i1})y_i + \sum_{i=1}^{n} W_{ni}(z_{i2})y_i] \qquad (7)$$

As for $\tilde{X}_i$, it represents the explanatory variable vector and is calculated according to the following formula:

$$\tilde{X}_i = X_i - \sum_{i=1}^{n}\sum_{j=1}^{k} W_{ni}(z_{ij})X_i$$

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

2.4 Generalized least-squares restricted (GRLS):

Assuming the existence of non-random linear constraints imposed on the model parameters, they can be expressed as follows:

Where:$R\beta = r$

R: is a known matrix of rank (q*p) where q<p. The rows of R are linearly independent, and the number of rows equals the number of constraints. The number of columns equals the number of model parameters.

r: is a known vector with dimensions q×1. Its elements represent the fixed bounds in the constraints.

The assumption of full row rank for the fit is chosen and can be justified by the fact that each consistent linear equation can be transformed into an equivalent equation representing a full row rank in the matrix. Considering the imposed linear constraints, the Generalized Constrained Generalized Least Squares Estimator (GLSRE) is formulated as follows (Roozbeh, 2016):

$$\hat{\beta}_{GRLS} = argmin_{\beta} = (\tilde{Y} - \tilde{x}\beta)'\, V^{-1}\, (\tilde{Y} - \tilde{x}\beta) \qquad (9)$$

s.t $R\beta = r$

$$\hat{\beta}_{GRLS} = \hat{\beta}_{GLS} - C^{-1}R'(RC^{-1}R')^{-1}(R\hat{\beta}_{GLS} - r) \qquad (10)$$

The constrained generalized least squares estimators are inefficient when dealing with the multicollinearity issue along with the presence of outliers in the dependent variable. This is because they fail to meet the conditions that would minimize the variance. Therefore, ridge estimators will be used instead.

2.5 generalized least-squares ridge estimator (RGLS):

They proposed this method, both (Hoerl and Kennard, 1970). This approach is used to address the issue of multicollinearity, where they suggested introducing a small positive number added to the main diagonal elements of matrix C, as in the following formula: (Najm A. and Khorshid, E, 2018).

$$\beta_{GLS}(k) = C_k^{-1}\tilde{X}V^{-1}\tilde{Y} \qquad (11)$$

Where: $C_k = C + KI_p$

Matrix C is a specific positive semi-definite matrix, meaning that its eigenvalues are greater than or equal to zero (Roozbeh,2016).

There exists an orthogonal matrix Γ such that $C = \Gamma\,\Omega\,\Gamma^{-1}$, where $\Omega = diag\,(\lambda_1 \dots, \lambda_p)$ is a diagonal matrix representing the eigenvalues of matrix C. Therefore, the model (1) will become in the following form:

$$\tilde{Y} = \tilde{X}^{*}\alpha + \epsilon \qquad (12)$$

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

Now, the parameter K in the ridge-constrained semi-parametric regression model can be estimated using generalized least-squares restricted (Swamy et al., 1978).

$$\hat{\mathrm{K}}_{\mathrm{LS}} = \frac{\mathrm{P}\,\hat{\sigma}^2_{\mathrm{LS}}}{\hat{\beta}'_{\mathrm{GRLS}}\,\hat{\beta}_{\mathrm{GRLS}}} \qquad (13)$$

$$\hat{\sigma}^2_{\mathrm{LS}} = \frac{1}{\mathrm{n}-(\mathrm{p}+\mathrm{q})}\,(\tilde{\mathrm{Y}} - \tilde{\mathrm{X}}\,\hat{\beta}_{\mathrm{GRLS}})'\,\mathrm{V}^{-1}(\tilde{\mathrm{Y}} - \tilde{\mathrm{X}}\,\hat{\beta}_{\mathrm{GRLS}}) \qquad (14)$$

2.6 generalized least squares restricted ridge estimator (RGRLS):

The ridge parameter can be obtained by minimizing the sum of squared residuals with linear constraints, transforming the ridge-constrained semi-parametric regression into a model that includes the multicollinearity problem along with two constraints, as follows:

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

s.t

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

$$\mathrm{R}\beta = \mathrm{r}$$

Results of the estimators are given by the following formula:

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

$$= \beta_{GLS}(k) - C_k^{-1}R'(RC_k^{-1}R')^{-1}(R\beta_{GLS}(k) - r). \qquad (15)$$

The above estimator is referred to as the generalized least-squares restricted ridge estimator (RGRLS)
Can be expressed in another form.

$$\hat{\beta}_{GRLS}(k) = (I + C_k^{-1}R'\left(R\,C_k^{-1}\,R'\right)^{-1}R\big)\beta_{GLS}(k) + C_kR'(RC_k^{-1}R')'r \quad (16)$$

As: $R(C_k^{-1}R'\left(R\,C_k^{-1}\,R'\right)^{-1}\big)R = R$

The generalized inverse of R, denoted as R^-, can be expressed using the following formula:

$$R^- = (C_k^{-1}R'\left(R\,C_k^{-1}\,R'\right)^{-1}\big).$$

So, the equivalent equation for equation (16) is:

$$\hat{\beta}_{GRLS}(k) = (I - R^-R)\hat{\beta}_{GLS} + R^-r. \qquad (17)$$

2.7 The Robust approach:

The term "robustness" was first coined by (Box, 1953), and (Tukey, 1960a) highlighted the lack of robustness in the arithmetic mean and proposed alternative measures that are more robust. The theory has extended over the years to other applications, such as regression. Despite the abundance of methods introduced, there is no single method or approach considered the best in all aspects. Several criteria are used to determine the strengths of an estimator, and The optimal estimator is the one that possesses all the positive attributes across all criteria. Some of these important criteria include: (Leroy and Rousseeuw, 1987; Rousseeuw and Van Driessen, 2006)

i. Breakdown Point: The point at which the estimator ceases to provide reasonable results.
ii. Efficiency: How well the estimator performs in terms of precision and accuracy compared to other estimators.

iii. Computational Simplicity: The ease with which the estimator can be computed.

iv. Asymptotic Behaviour: The performance of the estimator as the sample size approaches infinity.

2.8 The Robust MM Estimator :

The Robust MM Estimator, as introduced by (Yohai, 1987) and developed by (Wilcox, 2021), is a powerful statistical estimator that aims to provide reliable estimates even in the presence of outliers. This estimator is designed to be less sensitive to extreme observations, making it suitable for situations where the data may contain outliers or outliers.

The "MM" in MM Estimator stands for "M estimate based on robust M estimators." These estimators are formulated to minimize a robust objective function, by assigning less weight to outliers.

Different functions have been proposed based on the desired properties of the estimator. The MM estimator is particularly useful in regression analysis and other statistical modelling tasks where outliers can significantly affect the results. (Rasheed, Z. H. and Abdulhafiz, A. S, 2013).

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{S_{MM}}\right) X_i' = 0 \qquad (19)$$

Where S_{MM} represents the standard deviation obtained from the robust S method. ρ represents the Tukey's square weight function.

$$\rho(u_i) = \begin{cases} \dfrac{{u_i}^2}{2} - \dfrac{{u_i}^4}{2c^2} + \dfrac{{u_i}^6}{6c^4} & -c \le u_i \le c \\ \dfrac{c^2}{6} & u_i < -c \text{ } or \text{ } u_i > c \end{cases}$$

Using the WLS method and the weight matrix W, we obtain the MM estimators.

$$\hat{\beta}_{MM} = (\tilde{X}'W\tilde{X})^{-1}\, \tilde{X}'W\tilde{Y} \qquad (20)$$

Algorithm MM:

1. Estimate regression coefficients on the data using the Generalized Least Squares (GLS) equation (5).

2. Detect the presence of outliers in the data.

3. Determine the preliminary estimated estimator, which has a breakdown point of 50%, usually using the S-estimator or the Least Trimmed Squares (LTS) estimator.

4. Calculate the residual values $e_i = y_i - \hat{y}_i$ for the S-estimator.

5. Calculate the value $\hat{\sigma}_i$.

$$\hat{\sigma}_i = \begin{cases} \dfrac{med|e_i - med(e_i)|}{0.6745} & t = 1 \\ \sqrt{\dfrac{1}{nk}\sum_{i=1}^{n} w_i e_i^2} & t > 1 \end{cases}$$

where *t* represents the number of iterations.

6. Calculate the value $u_i = e_i \backslash s_i$.

7. Compute the diagonal weight matrix.

$$8. w_i = \begin{cases} \begin{pmatrix} [1-(\frac{u_i}{4.685})^2]^2 & |u_i| \le 4.685 \\ 0 & |u_i| > 4.685 \end{pmatrix} \end{cases}$$

8. Find $\hat{\beta}_S$, Or $\hat{\beta}_{LTS}$.

9. Repeat steps 5 to 8 using the new estimate $\hat{\beta}_{MM}$ until obtaining a close value for $\hat{\beta}_{MM}$.

To express robust MM estimators in the constrained semi-parametric regression model RSRM (Nyquist, 1992):

$$\min_{\beta,w} \varphi(\beta,w) = (\tilde{Y} - \tilde{X}\beta)' V^{-\frac{1}{2}} W V^{-\frac{1}{2}} (\tilde{Y} - \tilde{X}\beta) \qquad (21)$$

s.t

$R\beta = r$

The resulting estimate is the Restricted robust estimator (MMRSRE) in the semi-parametric regression model, expressed in the following formula:

$$\hat{\beta}_{MMR} = \hat{\beta}_{MM}(w) - C(w)^{-1} R' (R C(w)^{-1} R')^{-1} (R\hat{\beta}_{MM}(w) - r) \quad (22)$$

Where:

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

$$\hat{\beta}_{MM}(w) = C(w)^{-1} \tilde{X}' V^{-\frac{1}{2}} W V^{-\frac{1}{2}} \tilde{Y}$$

2.9 Ridge estimators based on the robust approach:

To estimate the ridge parameter relying on the robust MM in the Restricted semi-parametric regression model (RSRM), this method is referred to as Robust Ridge MM Estimation (RRMME). (Qazaaz, Q. N. and Saleh, R. A,2015).

$$\hat{K}_S = \frac{P\, \hat{\sigma}^2_{MM}}{\hat{\beta}'_{MMR}(w)'\, \hat{\beta}_{MMR}(w)} \qquad (23)$$

$$\hat{\sigma}^2_{MM} = \frac{1}{n-(p+q)} (\tilde{Y} - \tilde{X}\, \hat{\beta}_{MMR}(w))' V^{-\frac{1}{2}} W\, V^{-\frac{1}{2}} (\tilde{Y} - \tilde{X}\, \hat{\beta}_{MMR}(w)) \quad (24)$$

$$\hat{\beta}_{MMR}\left(\hat{k}_{MM}, w\right) = \hat{\beta}_{MM}\left(\hat{k}_{MM}, w\right) - C\left(\hat{k}_{MM}, w\right)^{-1} R' (R C(\hat{k}_{MM}, w)^{-1} R')^{-1} (R\hat{\beta}_{MM}\left(\hat{k}_{MM}, w\right) - r) \quad (25)$$

Where:

$$C\,(\hat{k}_{MM}, w) = C(w) + \hat{k}_{MM} I$$

$$\hat{\beta}_{MM}(\hat{k}_{MM}, w) = C(\hat{k}_{MM}, w)^{-1} \tilde{X}' V^{-\frac{1}{2}} W V^{-\frac{1}{2}} \tilde{Y} \qquad (26)$$

2.9 Non-parametric estimation methods in Restricted Additive Partially Regression Models:

Model (3) is estimated by initially estimating the parametric part, which is linear and suffers from various issues. This estimate is then utilized in estimating the non-parametric part. Subsequently, the parametric part is re-estimated using the non-parametric estimates, and this process is iterated until the difference between the new estimates and the previous ones becomes close to zero. The non-parametric part is estimated using the Local Polynomial estimation method (Speckman, 1988; Li and Yin, 2008; Yang, 2010; Hmood, M. and Muslim, A, 2012).

2.10 Local polynomial estimator :

The local linear regression is considered a good smoothing method because it has high efficiency compared to other smoothing methods. Assuming (d=2) in model (3), i.e.

The additive functions can be written as follows: (Talib.H and Hmood.M, 2022; Hamood, K. and Qais, S. ,2013; Talib, H. R. and Hmood, M. Y,2022).

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

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

The backfitting algorithm is utilized for the model (3), assuming that s_{2,Z_2}^T, s_{1,Z_1}^T represent equivalent kernel functions for the local linear regression at Z_2, Z_1 respectively. (Lexin Li and yin 2008)

$$s_{1,Z_1}^T = e_1^T \left(Z_1^T \omega_1 Z_1\right)^{-1} Z_1^T \omega_1 \qquad (27)$$

$$s_{2,Z_2}^T = e_1^T \left(Z_2^T \omega_2 Z_2\right)^{-1} Z_2^T \omega_2 \qquad (28)$$

$e_1 = (1, 0)^T$

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

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

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:

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

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

$$S_1 = \begin{bmatrix} S_1 Z_{11} \\ \cdot \\ \cdot \\ \cdot \\ S_1 Z_{n1} \end{bmatrix} \quad ; \quad S_2 = \begin{bmatrix} S_2 Z_{12} \\ \cdot \\ \cdot \\ \cdot \\ S_2 Z_{n2} \end{bmatrix}$$

and $\{S_1^C = (I - 11^T/n) S_1\}$ denotes the centered smoothing matrix for S_1 . and $\{S_2^C = (I - 11^T/n) S_2\}$ 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:

$$\left.\begin{aligned} \hat{g}_1^{(m)} &= S_1^C \left(Y - X\hat{\beta}_P - g_2^{(m-1)}\right) \\ \hat{g}_2^{(m)} &= S_2^C \left(Y - X\hat{\beta}_P - g_1^{(m-1)}\right) \end{aligned}\right\} \qquad (29)$$

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

$$\hat{\beta}_o = \{X^T(I - S_{12})X\}^{-1} X^T(I - S_{12})Y \qquad (30)$$

$$S_{12} = \left\{I - \left(I - S_1^C S_2^C\right)^{-1} \left(I - S_1^C\right)\right\} + \left\{I - \left(I - S_2^C S_1^C\right)^{-1} \left(I - S_2^C\right)\right\} \quad (31)$$

To ensure that $\hat{\beta}_o$ is a consistent estimate of the root of n within the necessary bootstrap 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 (29), except replacing $\hat{\beta}$ by $\beta. and$ $(\hat{g}_1, \hat{g}_2)$can be expressed as follows:

$$\left.\begin{aligned}\hat{g}_1(\beta) &= \left\{ I - \left(I - S_1^C S_2^C \right)^{-1} \left(I - S_1^C \right) \right\} (Y - X\beta) \\ \hat{g}_2(\beta) &= \left\{ I - \left(I - S_2^C S_1^C \right)^{-1} \left(I - S_2^C \right) \right\} (Y - X\beta) \end{aligned}\right\} \qquad (32)$$

Now, substituting $\hat{g}_1(\beta)$, $\hat{g}_2(\beta)$ into model (3) and using the least square method, we obtain estimates based on the β formula of the form:

$$\hat{\beta}_f = \{X^T(I - S_{12})(I - S_{12})^T X\}^{-1} X^T(I - S_{12})(I - S_{12})^T Y \qquad (33)$$

As discussed by (Hastie and Tibshirani,1990; Opsomer and Ruppert,1999), centering each $\left[S_1^C\ , S_2^C\right]$ 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$\left(n^{-1/5}\right)$. This means that the estimator $\hat{\beta}$ is consistent for $\sqrt{n}$.

to apply the mentioned methods in The theoretical aspect of studying air quality standards in Baghdad province for a period of 46 days during the summer season, from June 1, 2023, to September 1, 2023.

3.Data Sources:

Daily averages of particle concentrations (PM10, PM2.5, CO2, CO, NO2, O3, Temp) were utilized, derived from daily fixed measurements or data aggregatable into daily averages. This index relies on the following:

i. PM10: Airborne particles with an aerodynamic diameter of up to 10 Micrometres, encompassing both fine and coarse particles.

ii. PM2.5: Airborne particles with an aerodynamic diameter of up to 2.5 Micrometres, also referred to as fine particulate matter.

iii. CO2: Emissions of carbon dioxide gas.

iv. CO: Emissions of carbon monoxide gas.

v. NO2: Emissions of nitrogen dioxide gas.

vi. O3: Emissions of ozone gas.

vii. Temp: Temperature.

To present air quality data representing human exposure, urban measurements were primarily used, including urban background, residential areas, commercial areas, mixed areas, and industrial areas near urban settlements.

Data from fixed measurements were included, excluding mobile stations. Air quality stations covering specific "hotspots" and exclusive industrial areas were not included in the analysis, as these measurements often represent areas with higher exposure rather than the average exposure of the population.

"Hotspots" were defined in the original reports or categorized as such because they were, for example, near exceptionally congested roads. However, it should be noted that omitting these measurements may have led to underestimations of average air pollution in the city. Data has been obtained through monitoring the following websites:

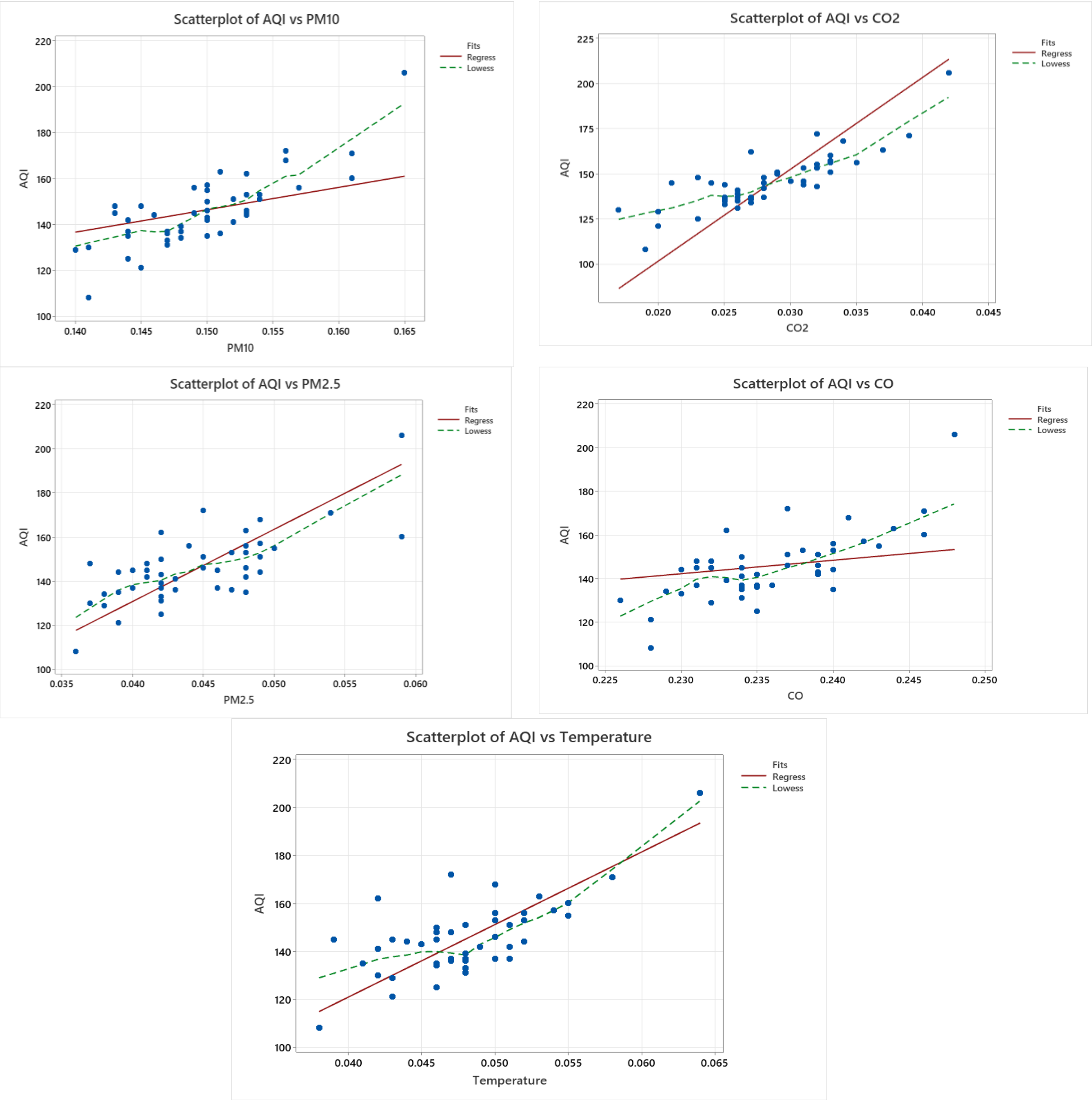
1-[AccuWeather] (https://www.accuweather.com, 2023).

2-[Tomorrow.io] (https://www.tomorrow.io/weather, 2023).

3-[IQ Air] (https://www.iqair.com, 2023).

3.1 Data Modelling:

To model the data in a semi-parametric regression model (46 observations), the variables are defined as follows: The Air Quality Index (AQI) is the dependent variable (Y), and the independent variables include (PM10, PM2.5, CO2, CO, NO2, O3, Temp). To determine which independent variables are labeled as parametric or not, we plot the variables to ascertain whether their relationship is linear with the dependent variable. If the relationship is linear, it indicates that the variable is parametric, as shown in the following figure: (Rawya. E. and Mohammed. J, 2023).



Scatterplot of AQI vs PM10

Scatterplot of AQI vs CO2

Scatterplot of AQI vs PM2.5

Scatterplot of AQI vs CO

Scatterplot of AQI vs Temperature

Figure 1: Type of Relationship and Scatterplots Between Each of (PM10, PM2.5, CO2, CO, Temp) and the Dependent Variable (Y).

It is evident from Figure (1) that each of the variables (PM10, PM2.5, CO2, CO, Temp) has a somewhat linear relationship with the dependent variable, indicating that they are parametric variables. As for the variables (NO2, O3), they are non-parametric variables, as shown in Figure (2): the type of relationship between them and the dependent variable.

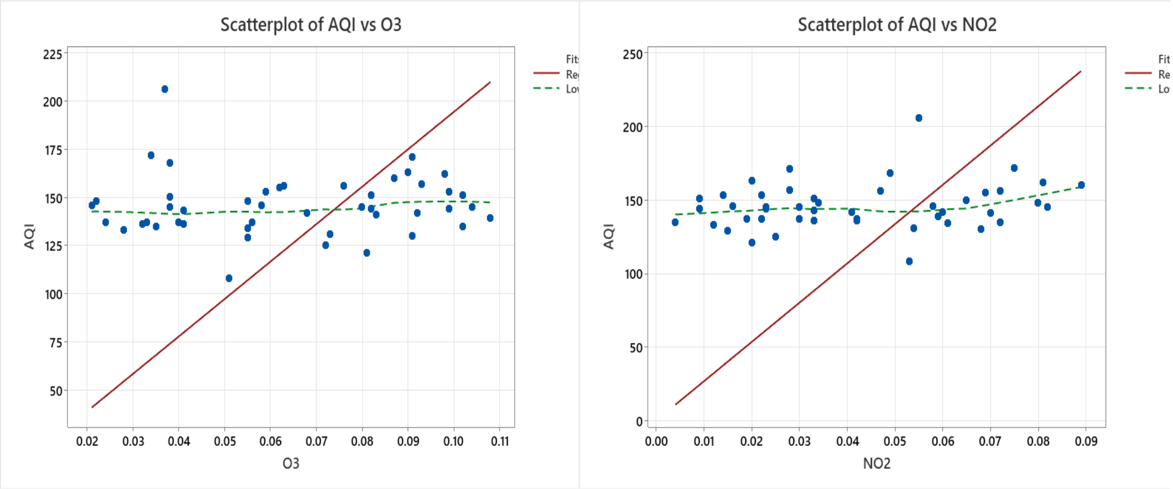


Figure 2: Type of Relationship and Scatterplots Between Each of (NO2, O3) and the Dependent Variable (Y).

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

$$(AQI)_i = \beta_0 + \beta_1(PM10)_i + \beta_2(PM2.5)_i + \beta_3(CO2)_i + \beta_4(CO)_i + \beta_5(Temp)_i + f_1(NO2)_i + f_2(O3)_i + \varepsilon_i \qquad (34)$$

After clarifying both the parametric and non-parametric variables, it is essential to understand and verify the correlations between the parametric variables. This can be achieved by examining the correlation matrix as in Table (1):

Table 1: Correlation Matrix

	AQI	Temp	PM2.5	PM10	CO2	CO
AQI	1	0.69379	0.71231	0.80195	0.83737	0.73008
Temp	0.69379	1	0.83213	0.65980	0.80386	0.82244
PM2.5	0.71231	0.83213	1	0.86159	0.77529	0.90963
PM10	0.80195	0.65980	0.86159	1	0.84527	0.77484
CO2	0.83737	0.80386	0.77529	0.84527	1	0.87506
CO	0.73008	0.82244	0.90963	0.77484	0.87506	1

Upon investigation, it becomes evident that there are multiple strong linear relationships between almost all variables. Therefore, it is necessary to calculate the eigenvalues of the information matrix $(X^T X)$, which are as in table (2):

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

λ_1	λ_2	λ_3	λ_4	λ_5
3.8278249	**0.0019443**	**0.0004060**	**0.0002665**	**0.0001093**

Based on the eigenvalues provided in Table (2), the Condition Number (C.N) was calculated to detect multicollinearity issues, relying on the following formula:

$$C.N = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{3.8278249}{0.0001093}} = 187.149$$

For further verification and confirmation of the presence of multicollinearity issues, we calculate the Variance Inflation Factor (VIF) for each variable as in table (3):

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

Temp	PM2.5	PM10	CO2	CO
11.22382	**49.22341**	**25.42113**	**31.33400**	**28.43729**

It has become evident that the data suffers from multicollinearity issues. Consequently, the information matrix will be highly problematic. Therefore, utilizing ridge regression or imposing non-random constraints on the model would be an appropriate solution to address the multicollinearity problem. To combine both approaches, we need to impose the following constraint matrix, assuming it achieves the least pollution ratio obtainable. The constraints are as follows:

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

To test the hypothesis $H_0: R\beta = 0$, we use the Chi-square test employing the estimators of the least squares method and comparing them to the tabulated value with degrees of freedom (2) for two constraints at a significance level of 0.05. The formula for the test is as follows:

$$\chi^2_{Cal} = \left(R\widehat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{r}\right)' \left(R\hat{\Sigma}R'\right)^{-1} \left(R\widehat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{r}\right) = 5.317$$

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

$\chi^2_{tab} = \chi^2_{(0.05,2)} = 5.991$. Since the calculated value is smaller than the tabulated value, H_0 is accepted.

After identifying the presence of multicollinearity in the data and discussing the methods to address it, the next step is to test for outliers in the data. We use the Studentized Deleted Residuals (SDR) method to detect outliers in the dependent variable. It has been found that there are four outliers, representing 8.7% of the total sample. Therefore, robust methods should be employed for estimation, utilizing ridge regression with imposed constraints to achieve the best estimators and the most representative model of the phenomenon. The estimation process using the programming (R) will be as in table (4):

Table 4: Model Estimates and Comparison Criteria

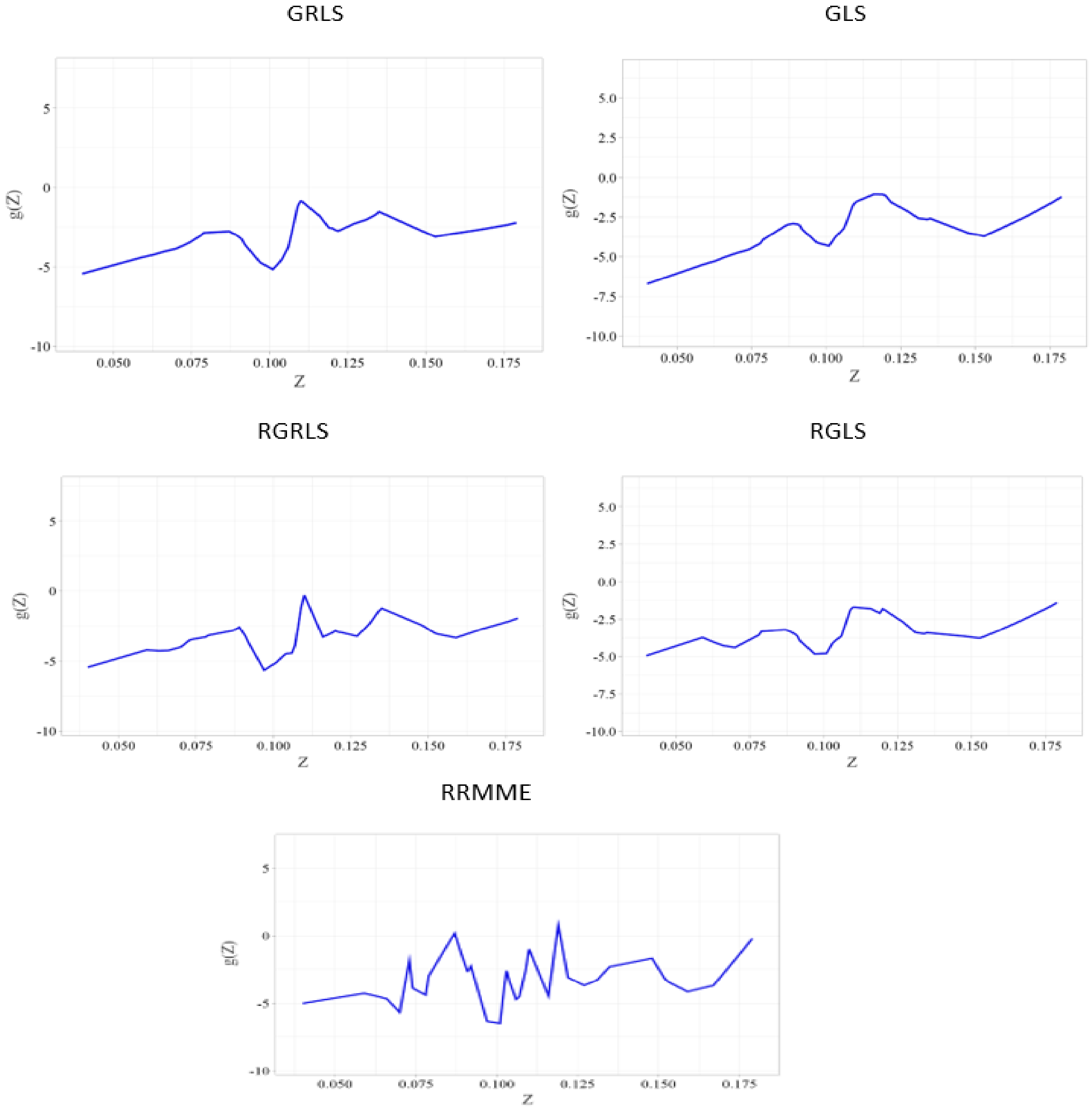
Methods	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$	$\boldsymbol{\beta_5}$	MAD	R2
GLS	0.905752	-0.484036	1.074623	1.580387	0.333326	8.3008	0.4131
GRLS	1.662499	-0.764781	1.988307	0.547714	0.88804	7.5807	0.6175
RGLS	0.708632	-0.1457	0.872612	1.454519	0.213629	7.9228	0.5828
RGRLS	1.597204	-0.830483	1.879708	0.549923	0.793004	7.3198	0.6606
RRMME	1.575838	-0.848313	1.846967	0.550328	0.765097	5.0606	0.8650

Table (4) displays the parameter estimates of the Restricted Additive Partially Regression Model. The parametric part was estimated using Generalized Least Squares (GLS), Generalized Ridge Least Squares (GRLS), Ridge Generalized Least Squares (RGLS), Ridge Generalized Least Squares with Constraints (RGRLS), and the Robust Method (MM) for constrained generalized least squares. The non-parametric part was estimated using Local polynomial estimator, employing the Epanechnikov kernel function. The bandwidth was selected using Cross-Validation. Model estimation methods were compared using the Coefficient of Determination and Mean Absolute Deviation. This criterion was chosen due to its sensitivity to outliers in the dependent variable. The results indicate that the method combining constraints and ridge estimates outperformed the classical method and individual ridge estimates, as well as individual non-random constraints. The best method for estimating of the Restricted Additive Partially Regression Model, considering both multicollinearity and outliers, is achieved by integrating non-random constraints with ridge estimates using the MM method. Examining the estimated parameters for the parametric part reveals that the second independent variable, PM2.5, has an inverse relationship with the dependent variable.

To assess the suitability of the estimation methods used for the parametric part in determining the non-parametric part, this was illustrated through Figure (1). The figure depicts the dispersion relationship between the combined non-parametric variables (O3, NO2) and the dependent variable (AQI).

Figure 3: Represents the behavior of the combined non-parametric variables with the dependent variable.

Figure 3 shows a drawing of the behavior curve of the two nonparametric variables together with the dependent variable y. By knowing the behavior of the curve, it can be compared with the nonparametric functions estimated in the nonparametric part of the model.



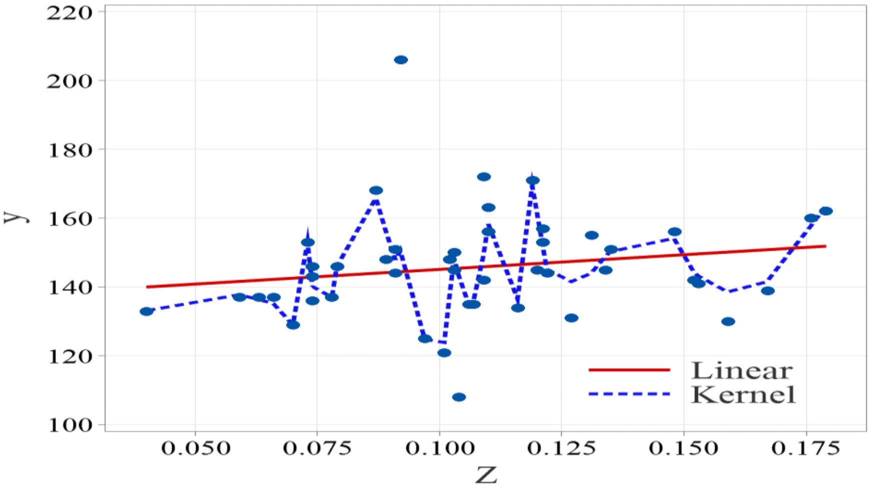


Figure 4: represents the behavior of the combined non-parametric variables with the estimated smoothing functions.

Figures (3) and (4) show that the method that combined both non-random constraints and the robust ridge estimator was more suitable than other methods when using its estimates to find the smoothing function estimates, Figure 4 also shows that the behavior curve of the two nonparametric variables when using the Robust approach MM was closer to their behavior with the dependent variable shown in Figure 3.

4. Discussion of Results:

In this paper, we utilized restricted additive partially regression model to model air quality data in Baghdad Governorate, as all the independent variables used collectively influence the dependent variable in an aggregate manner. Furthermore, we addressed the issues of multicollinearity and outliers in the dependent variable by employing classical estimators. Subsequently, non-random constraints were imposed on the parameters of the parametric part of the model and the robust ridge estimators obtained through the robust approach, specifically the MM estimators, which exhibit a high breakdown point and efficiency. It was evident that the method combining non-random constraints with robust ridge estimators was superior to other approaches when used to estimate the smoothing functions in a Local polynomial estimator, as indicated by the comparison criteria employed in model estimation. additionally, we demonstrated that both variables (O3, NO2) exhibit a non-linear relationship with the dependent variable (AQI). The latter, with exceptionally high rates during the summer season, adversely affects public health and human life in Iraq. Urgent solutions are required to address the prevalent diseases resulting from air pollution. Moreover, based on the estimated parameter values, we find that the variable (PM10), representing airborne particles with an aerodynamic diameter up to 10 micrometers, has a more significant impact on air quality compared to other variables. Additionally, there are non-linear effects observed for the non-parametric variables.

5. Conclusion:

In summary, our analysis of air quality data in Baghdad Governorate using the Restricted Additive Partially Regression Model has provided valuable insights into the dynamics of air pollution and its implications for public health. By effectively addressing issues of multicollinearity and outliers in the dependent variable through the application of classical estimators and robust modeling techniques, we have gained a deeper understanding of the complex relationships among various factors influencing air quality.

Our findings underscore the urgent need for interventions to mitigate the adverse effects of air pollution, particularly during the summer season when pollutant levels are notably high. The non-linear relationship observed between certain variables and air quality indicators highlights the complexity of the underlying mechanisms driving pollution levels in the region. moreover, our analysis emphasizes the significant impact of airborne particles, particularly PM10, on air quality, indicating the importance of targeted measures to reduce their emissions and mitigate their harmful effects on public health.

Overall, our study contributes to the growing body of knowledge on air quality management and underscores the importance of adopting robust modeling approaches to better understand and address the challenges posed by air pollution in urban environments like Baghdad Governorate. Future research in this area should focus on refining modeling techniques and exploring additional factors that may influence air quality dynamics, ultimately informing more effective policy interventions and public health strategies.

Acknowledgments:
Gratitude and appreciation are extended to the scientific supervisor and all those who contribute to enhancing the quality of this paper. Your guidance and support are invaluable in improving the overall content and presentation of this work.

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, Which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved By The Local Ethical Committee in The University.

References:

1.AccuWeather. (2023). Weather Forecast. Retrieved from https://www.accuweather.com.

2.Adeboye, N. O., Fagoyinbo, I. S., and Olatayo, T. O. 2014. Estimation of the effect of multicollinearity on the standard error for regression coefficients. *Journal of Mathematics*. 10(4), pp. 16-20.

3.Begun et al. 1983. Information and Asymptotic Efficiency in Parametric-Nonparametric Models. *The Annals of Statistics*. 11(2), pp. 432-452. doi:10.1214/aos/1176346151

4.Box, G. E. 1953. Non-normality and tests on variances. *Biometrika*. *40*(3/4), pp. 318-335.

5.Bross, I. D. 1961. Outliers in patterned experiments: A strategic appraisal. *Technometrics*. pp. 91-102.

6.Dai, D. and Wang, D. 2023. A Generalized Liu-Type Estimator for Logistic Partial Linear Regression Model with Multicollinearity. *AIMS Mathematics*. 5, pp. 11851–11874. doi:10.3934/math.2023600.

7.Daoud, J. I. 2017, December. Multicollinearity and regression analysis. In Journal of Physics: *Conference Series*. (Vol. 949, No. 1, p. 012009). IOP Publishing.

8.Freeman, P. R. 1980. On the number of outliers in data from a linear model. *Trabajos de estadística y de investigación operativa*. 31, pp. 349-365.

9.Hamood, K. Y. and Qais, S. 2013. Comparison of some Non-parametric Additive regression methods. *Al-Nahrain Journal for Sciences*. 16(3), pp. 62-77.

10. Hmood, M. Y. and Muslim, A. 2012. A comparison Of Some Semiparametric Estimators For consumption function Regression. Journal of Economic and Administrative Science. 12 (67), pp. 273-288.

11. Hoerl, A. E, and Kennard, R. W. 1970. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*. 12(1970),pp. 69-82.

12. IQAir. (2023). Air Quality Index. Retrieved from https://www.iqair.com.

13. Kazem, A. H., and Muslim, B. S. 2002 . *Advanced Economic Measurement: Theory and App.lication*. Baghdad: Donia Al Amal Library.

14. Keller, G. and Warrack, B. 2000. *Statistics for Management and Economics* Fifth Edition. USA: South -Western College Pub.

15. Kingsley, C. A., and Fidelis, I. U. 2022. Combining Principal Component and Robust Ridge Estimators in Linear Regression Model with Multicollinearity and Outlier. *Concurr. Computat. Pract. Exper*. doi: https://doi.org/10.1002/cpe.6803

16. Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. 2005. *Applied linear statistical models*. USA New York: McGraw-hill.

17. Leroy, A. M. and Rousseeuw, P. J. 1987. *Robust Regression and Outlier Detection*. Wiley, New York.

18. Li, L. and Yin, X. 2008. Sliced inverse regression with regularizations. *Biometrics*. (641), pp. 124-131.

19. Li, Q. 2000. Efficient estimation of additive partially linear models. *International Economic Review*. (414), pp. 1073-1092.

20. Lian, H., and Liang, H. 2013. Generalized additive partial linear models with high-dimensional covariates.*EconometricTheory*.29(6),pp.1136-1161.
doi https://doi.org/10.1017/S0266466613000029.

21. Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T., and Hauser, R. 2008. Additive partial linear models with measurement errors. *Biometrika*. (953), pp. 667-678.

22. Liu, R., and Yang, L. 2010. Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory*. (261), pp29-59.

23. Muir, W. W. 1981. Regression Diagnostics*: Identifying Influential Data and Sources of Collinearity*.

24. Najm.A. S and Khorshid. E. S. 2018. A Comparison Between the Lasso and Liu-Type Methods in Estimating Parameters of the Negative Binomial Regression Model in the Presence of Multicollinearity Using Simulation. *Journal of Economics and Administrative Sciences*. 24109, pp. 515-515.

25. Nyquist, H. 1992. Restricted M-estimation. *Computational statistics and data analysis*. (144), pp. 499-507.

26. Oakes, R. 1981. Existence Across Possible Worlds: An Epistemological Resolution. *The Southern Journal of Philosophy*. 19 (2), 205.

27. Opsomer, J. D., and Ruppert, D. 1999. A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*. 8(4), pp.715-732

28. Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *Ann. Math. Statist*. (351), pp.73-101. https://doi.org/10.1214/aoms/1177703732.

29. Qazaaz, Q. N. N., & Saleh, R. A. R. 2015. Analysis of Robust Principal Components Depends on Some Methods of Projection-Pursuit. *Journal of Economics and Administrative Sciences*. (2183), pp. 317-327.

30. Rasheed, Z. H. and Abdulhafiz, A. S. 2013. Estimation of robust two-stage and local polynomial kernel approximation for time-varying coefficient model with balanced longitudinal data. *Journal of Economics and Administrative Sciences*. (1970), pp. 297-297.

31. Rawya. E. K and Mohammed. J. M. 2023. Estimate the Factors Affecting Air Pullution in Iraq Using Fuzzy Regression Models. *Journal of Al-Rafidain University College For Sciences*. 1 (54), pp. 279-291.

32. Roozbeh, M. 2016. Robust ridge estimator in restricted semiparametric regression models. *Journal of Multivariate Analysis*. 147, pp.127-144.

33. Rousseeuw, P. J., and Van Driessen, K. 2006. Computing LTS regression for large data sets. *Data mining and knowledge discovery*. 12, pp. 29-45.

34. Speckman, P. 1988. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 50(3), pp.413-436.

35. Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*. 39(1), pp.44-47.

36. Swamy, P. A. V. B., Mehta, J. S., and Rappoport, P. N. 1978. Two methods of evaluating Hoerl and Kennard's ridge regression. *Communications in Statistics-Theory and Methods*. 7(12), pp. 1133-1155.

37. Talib, H. R. and Hmood, M. Y. 2022. Using Spline Approximation and Local Polynomial Methods to Estimate the Additive Partial Linear Model. *Mathematical Statistician and Engineering Applications*. 71(4), pp. 5046-5059.

38. Talib, H. R., & Hmood, M. Y. 2022. Comparison between some methods for Estimation and Variables Selection for Semi–Parametric Additive model with practical application. Al Kut Journal of Economics and Administrative Sciences. 14(44), pp. 405-426.

39. Tomorrow.io. 2023. Weather Forecast. Retrieved from https://www.tomorrow.io/weather.

40. Tukey, J. W. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*. pp.448-485.

41. Wilcox, R. R. 2021. *Introduction to Robust Estimation and Hypothesis Testing*. 5th Edition. Netherlands: Elsevier.

42. Wu, J. and Asar, Y. 2017. A weighted stochastic restricted ridge estimator in partially linear model. *Communications in Statistics-Theory and Methods*. 46(18), pp. 9274-9283.

43. Yohai, V. J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*. pp.642-656.

Journal of Economics and Administrative Sciences
2024; 30(143), pp. 435-454

P-ISSN 2518-5764
E-ISSN 2227-703X

مقدر Ridge-MM الحصين في نموذج الانحدار التجميعي الجزئي المقيد

احمد رزاق عبد رمضان[1]
جامعة واسط/ كلية الإدارة والاقتصاد/ قسم الإحصاء
العراق
ahmed.razak1101a@coadec.uobaghdad.edu.iq

قتيبة نبيل نايف القزاز[2]
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
العراق
dr.qutaiba@coadec.uobaghdad.edu.iq

Received:15/1/2024 Accepted:4/3/2024 Published Online First: 1 /10/ 2024

مستخلص البحث:

في هذه الورقة، تم استخدام نموذج الانحدار الإضافي المقيد جزئيا لتحليل بيانات جودة الهواء في محافظة بغداد، مع التركيز على معالجة قضايا العلاقة الخطية المتعددة بين المتغيرات المستقلة والقيم المتطرفة في المتغير التابع. من خلال تطبيق المقدرات الكلاسيكية ومقدرات الحرف والمقدرات الحصينة وفرض قيود غير عشوائية على الأجزاء المعلمية للنموذج من خلال طريقة مقدر Ridge-MM Robust في نموذج الانحدار الجزئي التجميعي المقيد، هدفت الورقة إلى تقييم مدى نجاح النموذج وفعاليته في التعامل مع تحديات تلوث الهواء خلال فصل الصيف. أشارت النتائج التي تم الحصول عليها من خلال استخدام الحزم والخوارزميات المعدة مسبقًا في لغة برمجة R إلى أن دمج القيود غير العشوائية مع المقدرات الحصينة أثر بشكل إيجابي على دقة وظائف التقدير. علاوة على ذلك، وجد أن بعض المتغيرات، مثل PM10 (الجزيئات المحمولة بالهواء والتي يصل قطرها الديناميكي الهوائي إلى 10 ميكرومتر)، لها تأثير كبير على جودة الهواء، وذلك من خلال قيم المعلمات. وقد لوحظت تأثيرات غير خطية لبعض المتغيرات غير المعلمية. وتسلط الدراسة الضوء على أهمية فهم تأثيرات ملوثات الهواء على الصحة العامة، وتؤكد على الحاجة الملحة إلى حلول سريعة للتخفيف من هذه الآثار السلبية.

نوع البحث: تصنيف الورقة الخاصة بك تحت أحد هذه التصنيفات: ورقة بحثية.

المصطلحات الرئيسة للبحث: انموذج الانحدار التجميعي الجزئي شبه المعلمي، المربعات الصغرى المعممة، القيود، التعدد الخطي، مقدر الحرف، مقدر MM الحصين، ممهد متعدد الحدود، مؤشر جودة الهواء.

(2)

(3)