

Ridge Regression for Analysing the Determinants of CO₂ Emissions in Algeria under Multicollinearity (1970–2018)

Souhila LAZALI * ENSSEA, Algeria — LAMOPS Laboratory lazali.souhila@enssea.edu.dz	Fatiha MEZOUE ENSSEA, Algeria — LAMOPS Laboratory mezoued.fatiha@enssea.edu.dz
---	--

Received : 07 /05/20 25

Accepted : 01/09/ 20 25

Published : 31/12/20 25

Abstract:

This study applies Ridge regression to analyse the economic and demographic determinants of CO₂ emissions in Algeria over the period 1970–2018, in the presence of severe multicollinearity that undermines the properties of the Ordinary Least Squares (OLS) estimator. The explanatory variables include GDP per capita, energy use per capita, foreign direct investment, total population, and urbanisation rate, within a log-log specification. Multicollinearity diagnostics reveal VIF values exceeding 160 for most regressors and a condition number of 37.74, rendering OLS estimates highly unstable. The shrinkage parameter k is selected via ten-fold cross-validation; Ridge regression achieves CV-MSE = 0.0206, representing a 19.1% improvement over OLS. Energy use per capita emerges as the sole statistically significant determinant (elasticity = 0.479; $p < 0.01$), reflecting Algeria's heavy dependence on fossil fuels.

Keywords: Ridge regression, multicollinearity, CO₂ emissions, Algeria, bias-variance trade-off, cross-validation.

JEL Classification Codes : C22 ; C52 ; Q53 ;Q56.

1. Introduction

The multiple linear regression model holds a central place in applied econometrics, providing a flexible framework for quantifying relationships between economic and environmental variables. When the Gauss-Markov assumptions are satisfied, the Ordinary Least Squares (OLS) estimator is the Best Linear Unbiased Estimator (BLUE), yielding minimum-variance estimates within the class of linear unbiased estimators.

In practice, however, multicollinearity characterised by strong linear dependence among the explanatory variables, is pervasive in environmental and macroeconomic studies. Key variables such as GDP per capita, energy consumption, population, and urbanisation rate tend to move together over time (Gujarati & Porter, 2009), generating near-linear dependencies that cause the variance of OLS estimates to inflate dramatically. The statistical consequences are severe: wide confidence intervals, incorrect signs on regression coefficients, and a substantial loss of power in hypothesis tests.

To address this problem, Hoerl and Kennard (1970) introduced Ridge regression, a shrinkage estimator that trades unbiasedness for a substantial reduction in variance and mean squared error (MSE). This approach builds on the foundational contributions of Stein (1956) and James and Stein (1961), who established that shrinkage estimators dominate OLS under quadratic loss when $p \geq 3$.

The present study applies Ridge regression to the analysis of CO₂ emission determinants in Algeria over the period 1970–2018. The Algerian emissions literature (Bouznit & Pablo-Romero, 2016;

* Corresponding Author

Rouaba, 2021) documents precisely the severe multicollinearity that motivates our methodological choice. The paper is organised as follows. Section 2 presents the theoretical framework. Section 3 describes the data and multicollinearity diagnostics. Section 4 develops the bias-variance trade-off and k selection. Section 5 reports and discusses results. Section 6 concludes.

2. Theoretical Framework of Ridge Regression

2.1. The Linear Model under Multicollinearity

Consider the standard multiple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of observations on the dependent variable, $\mathbf{X}$ an $(n \times p)$ full-rank matrix of regressors, $\boldsymbol{\beta}$ a $(p \times 1)$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ an $(n \times 1)$ vector of i.i.d. errors with mean zero and variance $\sigma^2 \mathbf{I}$. The OLS estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

with covariance matrix $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. Under multicollinearity, some eigenvalues λ_j of $\mathbf{X}'\mathbf{X}$ approach zero, and the MSE of OLS:

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{MCO}}) = \sigma^2 \boldsymbol{\Sigma}_j (1/\lambda_j)$$

inflates without bound, making statistical inference unreliable. This is the precise pathology that motivates the use of biased estimators with reduced variance.

2.2. The Ridge Estimator: Definition and Properties

Hoerl and Kennard (1970a, 1970b) proposed regularising $\mathbf{X}'\mathbf{X}$ by adding a diagonal perturbation $k\mathbf{I}$ ($k > 0$), yielding the Ridge estimator:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'\mathbf{y}$$

Adding $k\mathbf{I}$ guarantees invertibility of the augmented matrix even when $\mathbf{X}'\mathbf{X}$ is nearly singular. In the canonical parameterisation (after diagonalising $\mathbf{X}'\mathbf{X}$), the bias, variance and MSE simplify to:

$$\text{Bias}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)}) = k^2 \sum_{i=1}^p \left(\frac{\hat{\alpha}_i}{\lambda_i + k} \right)^2$$

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)}) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2}$$

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)}) = \text{Var}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)}) + \text{bias}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k)})^2 = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^p \left(\frac{\hat{\alpha}_i}{\lambda_i + k} \right)^2$$

It can be shown that there always exists a $k^* > 0$ such that $\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{Ridge}(k^*)}) < \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$, provided $\boldsymbol{\beta} \neq 0$ (Hoerl & Kennard, 1970a). This result establishes the theoretical superiority of Ridge over OLS under the scalar MSE criterion.

2.3. Methods for Selecting the Shrinkage Parameter k

Since k cannot be determined a priori, the literature proposes several data-driven approaches for its estimation. Table 1 below summarises the main methods.

Table 1: Main methods for selecting the Ridge shrinkage parameter k

Method	Formula / Principle	Reference	Characteristic
HK	$k = p \cdot \hat{\sigma}^2 / \hat{\beta}' \hat{\beta}$	Hoerl & Kennard (1970b)	Closed-form, simple
HKB (iterative)	Iterate HK until convergence	Hoerl et al. (1975)	More accurate
Kibria (median)	$k = \text{median}(\hat{\sigma}^2 / \hat{\beta}^2)$	Kibria (2003)	Robust to near-zero $\hat{\beta}_j$
Khalaf–Shukur	Adjusted for small n	Khalaf & Shukur (2005)	Small-sample adapted
Alkhamisi–Shukur	Simulation-optimised variant	Alkhamisi & Shukur (2007)	Improved MSE in sims
10-fold CV	$k^* = \text{argmin CV-MSE}(k)$	Kohavi (1995)	Data-driven, no assumptions

Hoerl and Kennard (1970b) proposed the closed-form estimator $k = p \cdot \hat{\sigma}^2 / \hat{\beta}' \hat{\beta}$, where $\hat{\sigma}^2$ is the OLS residual variance and $\hat{\beta}$ the OLS coefficient vector — intuitive but sensitive to outliers and near-zero coefficients. Hoerl, Kannard and Baldwin (1975) extended this to an iterative version that recalculates k until convergence, generally yielding more accurate solutions. Kibria (2003) introduced estimators based on the median and geometric mean of the ratios $\hat{\sigma}^2 / \hat{\beta}^2$, providing greater robustness when some coefficients are close to zero. Khalaf and Shukur (2005) and Alkhamisi and Shukur (2007) developed further variants specifically calibrated for small samples through extensive Monte Carlo comparisons. ten-fold cross-validation (Kohavi, 1995), the approach retained in the present study, represents the most general and assumption-free procedure: k^* is chosen to minimise the out-of-sample prediction error CV-MSE(k) over a fine grid, without any prior knowledge of β or σ^2 .

3. Data and Multicollinearity Diagnostics

3.1. Data Description

The study draws on annual data for Algeria covering the period 1970–2018 ($n = 49$ observations), sourced from the World Bank's World Development Indicators (WDI) and the International Energy Agency (IEA). The dependent variable is CO₂ emissions per capita, measured in metric tons. Following the literature on the Environmental, five explanatory variables are included:

- GDP per capita (GDP) in US dollars — proxy for economic development.
- Energy use per capita (EG) in kg of oil equivalent — the primary technical driver of emissions.
- Net inflows of foreign direct investment (FDI) in current US dollars.
- Total population (POP) — to control for scale effects.
- Urbanisation rate (URB) in percent — share of population residing in urban areas.

All variables are transformed to natural logarithms and standardised prior to estimation, yielding the following log-log specification in which each coefficient is directly interpretable as an elasticity:

$$\ln(\text{CO}_{2t}) = \beta_0 + \beta_1 \ln(\text{POP}_t) + \beta_2 \ln(\text{GDP}_t) + \beta_3 \ln(\text{URB}_t) + \beta_4 \ln(\text{FDI}_t) + \beta_5 \ln(\text{EG}_t) + \varepsilon_t$$

for $t = 1970, \dots, 2018$. A static time-series specification is retained because the central objective is to compare estimators under multicollinearity rather than to model long-run dynamics.

3.2. Multicollinearity Diagnostics

As show in figure 1 of the correlation matrix reveals strong positive correlations among all regressors, particularly between population, urbanisation, and EG. This pattern is characteristic of the structural co-evolution of macroeconomic and demographic variables in a rapidly developing economy.

Figure 1. Correlation Heatmap among the Study Variables

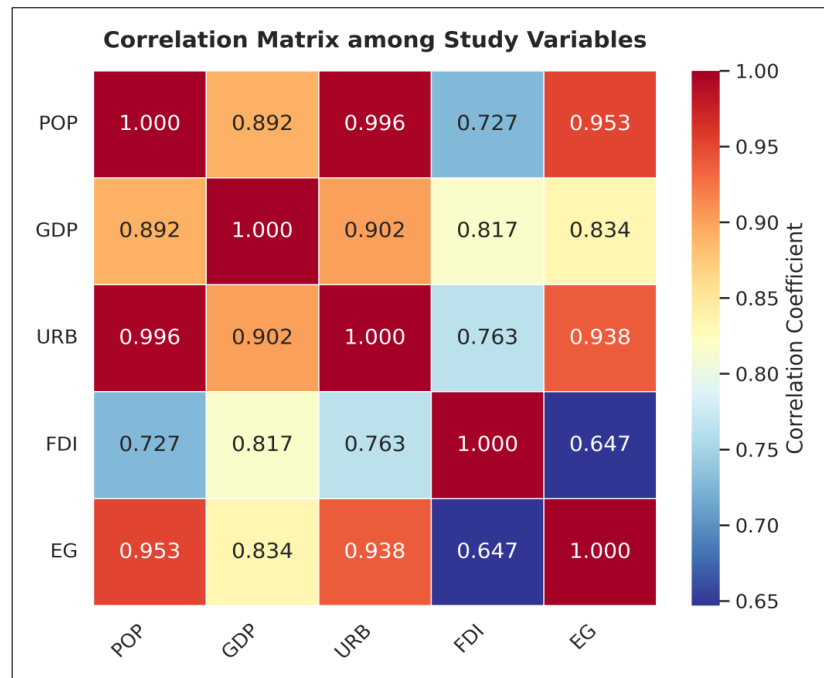


Table 2 quantifies multicollinearity through the Variance Inflation Factor (VIF), defined as $\text{VIF}_j = 1/(1 - R_j^2)$, where R_j^2 is the coefficient of determination from regressing x_j on the remaining $p-1$ regressors. A VIF exceeding 10 is conventionally taken as evidence of severe multicollinearity.

Table 2: Variance Inflation Factors (VIF) — Algeria, 1970–2018

Variable	ln(POP)	ln(GDP)	ln(URB)	ln(FDI)	ln(EG)
VIF	229.31	80.40	162.57	1.82	16.77

Four of the five regressors display VIF values far above the threshold of 10: $\ln(\text{POP})$ reaches 229.31 and $\ln(\text{URB})$ 162.57. Only $\ln(\text{FDI})$ presents an acceptable VIF (1.82).

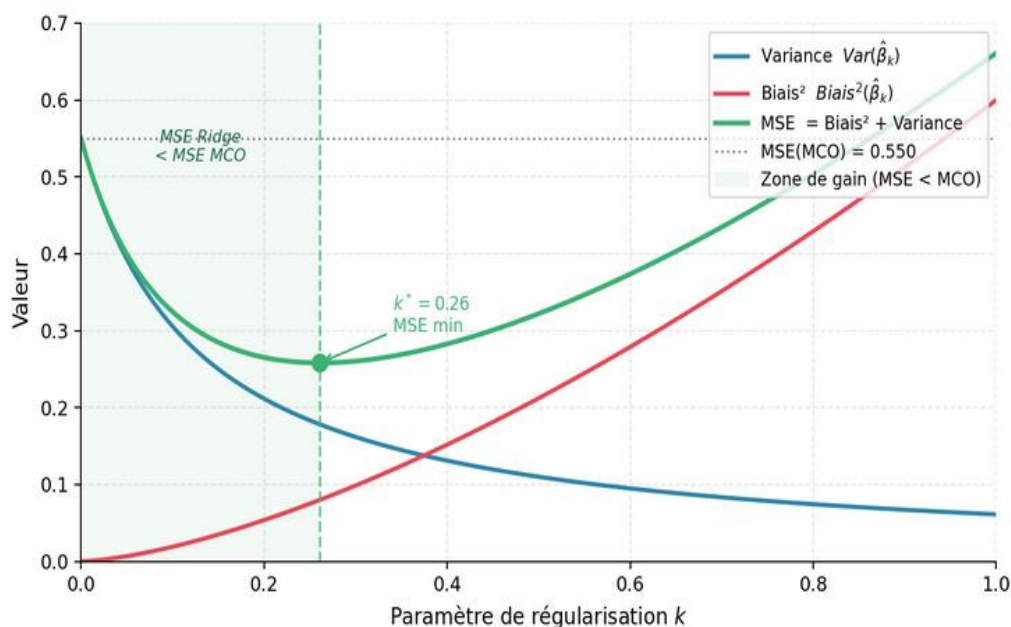
The condition number of $X'X$ is 37.74, exceeding the threshold of 30 associated with moderate-to-severe multicollinearity. These diagnostics confirm that OLS estimates will be highly unstable, fully justifying the use of Ridge regression.

4. Bias–Variance Trade-off and Selection of k

4.1. The Bias–Variance Trade-off Illustrated

The essence of Ridge regression lies in a fundamental trade-off between bias and variance, illustrated in Figure 1 for our Algerian application. At $k = 0$, Ridge coincides with OLS: zero bias but very high variance due to the severe multicollinearity documented in Table 1. As k increases, variance decreases monotonically (the effective eigenvalues $\lambda_j + k$ move away from zero), while the squared bias grows. The total MSE — the sum of both components — traces a U-shaped curve with a global minimum at k^* .

Figure 2. . Bias²–Variance trade-off as a function of the regularisation parameter k



Ridge MSE (green) reaches its minimum at k^* , well below the OLS MSE (grey dotted line). The green shaded area indicates the range of k over which Ridge outperforms OLS.

Figure 2 makes clear that Ridge improves upon OLS over a range of k values around its optimum. For k too large, the introduced bias becomes excessive and Ridge MSE exceeds that of OLS. This motivates the need for a rigorous selection of k^* , described in the next section.

4.2. Selection of k via Ten-fold Cross-Validation

Since the Ridge MSE depends on unknown parameters (β and σ^2), k cannot be determined analytically. As detailed in Section 2.3, we adopt ten-fold cross-validation (Kohavi, 1995): the sample is partitioned into 10 disjoint folds; the model is fitted on 9 folds and evaluated on the remaining fold, yielding the out-of-sample prediction error:

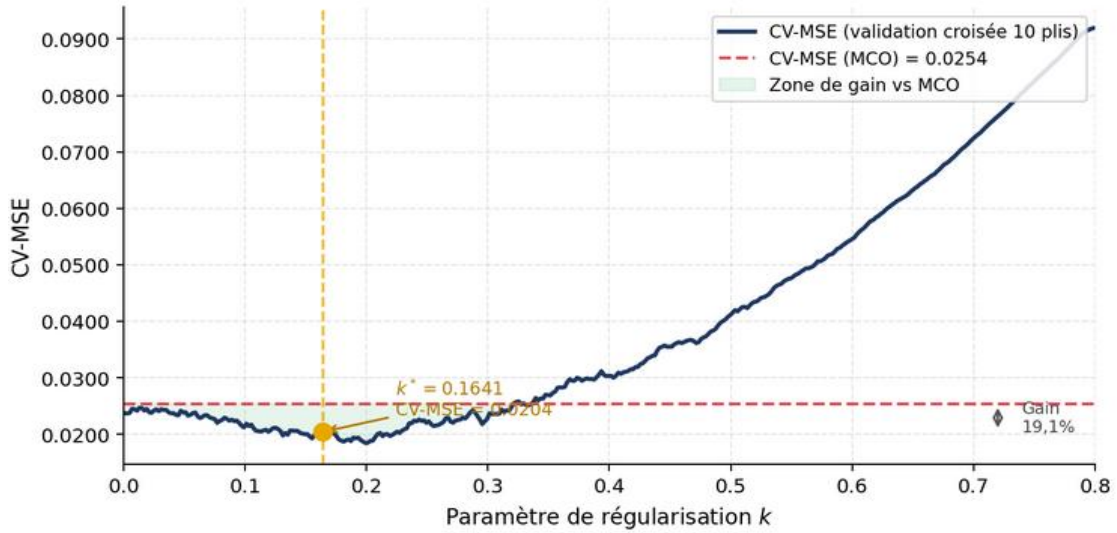
$$CV_MSE(\mathbf{k}) = \left(\frac{1}{n}\right) \sum_i (y_i - \hat{y}_{-i}(\mathbf{k}))^2$$

The optimal parameter minimises this criterion over a fine grid:

$$k^* = \operatorname{argmin}_{\{k \geq 0\}} \text{CV-MSE}(k)$$

This procedure is entirely data-driven, requires no prior assumptions on β or σ^2 , and is well-suited to small samples such as ours ($n = 49$). Figure 3 presents the resulting CV-MSE curve

Figure 3. CV-MSE curve (ten-fold cross-validation) as a function of k .



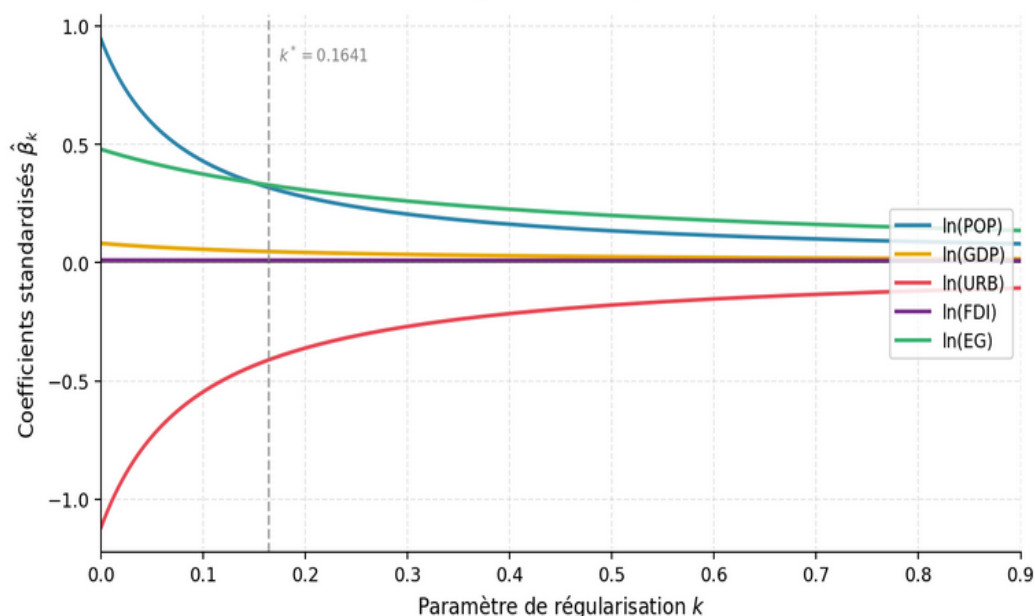
The global minimum is reached at $k^* = 0.1641$ ($\text{CV-MSE} = 0.0206$), representing a 19.1% improvement over OLS (red dotted line, $\text{CV-MSE} = 0.0254$). The green shaded area indicates the region of gain relative to OLS.

The curve displays a well-defined minimum at $k^* = 0.1641$, confirming the numerical stability of the solution. This minimum is well below the OLS CV-MSE (0.0254), attesting to the predictive superiority of Ridge for this dataset.

4.3. Ridge Coefficient Paths (Ridge Trace)

Figure 3 presents the evolution of the Ridge coefficients $\beta_{\text{Ridge}}(k)$ as a function of k — the so-called ridge trace — for all five regressors. This representation complements the CV-MSE criterion by visualising how regularisation redistributes coefficient mass across collinear regressors as k increases.

Figure 4. Ridge coefficient paths (ridge trace) as a function of k .



The coefficients most affected by multicollinearity notably $\ln(\text{POP})$ and $\ln(\text{URB})$, which carry the highest VIF values in Table 1 undergo the strongest contractions as k increases. By contrast, the coefficient of $\ln(\text{EG})$ remains relatively stable throughout, confirming its robust identification as the primary determinant of emissions regardless of the degree of regularization.

5. Estimation Results and Discussion

5.1. Estimated Coefficients and Model Performance

Table 3 reports the standardized estimated coefficients, shrinkage parameters, and performance indicators for OLS and Ridge. It should be noted that statistical inference standard errors, t-statistics, and p-values is reported exclusively for OLS: since Ridge estimators are biased by construction, classical test theory based on the normal distribution does not apply directly (Kibria, 2003; Saleh et al., 2019). Comparison across estimators is therefore conducted solely on the basis of the scalar MSE criterion.

Table 3: Estimated Coefficients and Performance — OLS vs Ridge (Algeria, 1970–2018)

Variable	OLS (β)	Std. Error	p-value	Ridge (β)
Intercept (β_0)	-11.568	10.291	0.267	-0.019
$\ln(\text{POP})$ (β_1)	0.943	0.864	0.281	-0.040
$\ln(\text{GDP})$ (β_2)	0.082	0.091	0.372	0.121
$\ln(\text{URB})$ (β_3)	-1.118	1.187	0.352	-0.040
$\ln(\text{FDI})$ (β_4)	0.011	0.007	0.112	0.007
$\ln(\text{EG})$ (β_5)	0.479***	0.167	0.007	0.568
R^2	0.934	—	—	0.932
CV-MSE (out-of-sample)	0.0254	—	—	0.0206
Shrinkage parameter k	—	—	—	0.1641

5.2. Discussion

Predictive superiority of Ridge

Ridge regression achieves an out-of-sample CV-MSE of 0.0206, representing a 19.1% improvement over OLS (0.0254). This gain is directly attributable to the variance reduction enabled by regularisation at $k^* = 0.1641$, identified in Section 4. The well-defined minimum of the CV-MSE curve (Figure 3) guarantees the numerical stability of the Ridge solution.

Determinants of CO₂ emissions

Under OLS, energy use per capita ($\ln(EG)$) is the sole statistically significant determinant, with an estimated elasticity of 0.479 (SE = 0.167; $t = 2.860$; $p < 0.01$): a one percent increase in per-capita energy consumption is associated, *ceteris paribus*, with a 0.479 percent increase in CO₂ emissions. This finding reflects Algeria's heavy reliance on fossil fuels — hydrocarbons and natural gas — throughout the sample period. The remaining regressors, population (0.943), GDP per capita (0.082), urbanisation (−1.118), and FDI (0.011) carry economically plausible signs but fail to reach statistical significance, as multicollinearity inflates their standard errors and reduces the power of individual t-tests.

Ridge coefficient behaviour

As illustrated by the ridge trace in Figure 4, the coefficient of $\ln(POP)$ reverses from +0.943 (OLS) to −0.040 (Ridge), and that of $\ln(URB)$ moves from −1.118 to −0.040. These sign reversals do not reflect genuine changes in the underlying economic relationships, but rather a redistribution of coefficient mass among highly collinear regressors: regularisation disentangles the individual contribution of each variable by reducing their mutual confounding. The coefficient of $\ln(EG)$ remains positive and stable, ranging from 0.479 to 0.568, confirming its robust identification as the dominant driver of Algerian CO₂ emissions irrespective of the estimation method employed.

6. Conclusion and Policy Recommendations

This study has provided compelling empirical evidence of the superiority of Ridge regression over OLS for analysing the determinants of CO₂ emissions in Algeria. The multicollinearity diagnostic (Table 1) with VIF values reaching 229.31 and a condition number of 37.74 first established the instability of OLS estimates. The bias-variance trade-off (Figure 2) and the CV-MSE curve (Figure 3) then showed that $k^* = 0.1641$, identified through ten-fold cross-validation, constitutes the optimal balance point, reducing out-of-sample prediction error by 19.1% relative to OLS. The ridge trace (Figure 4) provided a visual account of how regularisation redistributes coefficient mass under collinearity.

On the substantive side, energy use per capita emerges as the primary and robust determinant of Algerian CO₂ emissions, with an elasticity of 0.479. This result calls for ambitious energy policies: improving energy efficiency across the industry, transport, and residential sectors; diversifying the energy mix towards renewable sources (solar and wind), for which Algeria possesses considerable potential; and progressively reforming fossil fuel subsidies to align domestic prices with their true social costs.

These findings invite applied researchers in environmental econometrics to systematically diagnose multicollinearity via VIF and condition number, and to adopt Ridge regression with cross-validated parameter selection. This approach is particularly recommended for time series and models featuring structurally correlated macroeconomic and demographic variables characteristics common to environmental studies in developing economies.

7. Bibliography List

- Alkhamisi, M., & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communications in Statistics — Simulation and Computation*, 36(3), 535–547.
- Bouznit, M., & Pablo-Romero, M. P. (2016). CO₂ emission and economic growth in Algeria. *Energy Policy*, 96, 93–104.
- Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill/Irwin.
- Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.
- Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Application to nonorthogonal problems. *Technometrics*, 12(1), 69–82.
- Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4(2), 105–123.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.
- Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics — Theory and Methods*, 34(5), 1177–1182.
- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics — Simulation and Computation*, 32(2), 419–435.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI*, 2, 1137–1143.
- Rouaba, S. (2021). Determinants of CO₂ emissions in Algeria: An application of the ARDL model. *Al-Iqtisad Al-Jadid*, 12(4), 234–254.
- Saleh, A. K. M. E., Arashi, M., & Kibria, B. M. G. (2019). *Theory of Ridge Regression Estimation with Applications*. John Wiley & Sons.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206.
- Younici, K., & Belhadi, A. (2021). Analyse de la relation entre les émissions de CO₂ et la croissance du PIB en Algérie. *Economic Researcher Review*, 9(2), 205–216.