ECO341: ASSIGNMENT REPORT

ACKNOWLEDGEMENT:

We would like to thank **Dr. Deep Mukherjee** for giving us this opportunity to work on this project which helped us gain a deep insight into the concepts by working on a real-life problem. We would also like to thank **Mr. Debarun Sengupta** for teaching us the R language, without which we wouldn't have been able to solve the problem. It was indeed a great learning experience.

PROBLEM DESCRIPTION:

The problem statement suggests a "cure" for multicollinearity between the columns of $\mathbf{X_1}$ and $\mathbf{X_2}$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are two sets of explanatory variables partitioned from the same dataset. First, we regress each variable of $\mathbf{X_2}$ on $\mathbf{X_1}$ and compute the residuals matrix $\mathbf{Z_2}$. Then we regress the dependent variable **y** on ($\mathbf{X_1}$ & $\mathbf{Z_2}$) and ($\mathbf{X_1}$ & $\mathbf{X_2}$). $\mathbf{c = (c_1, c_2)}$ are the coefficients of the regression of **y** on ($\mathbf{X_1}$ & $\mathbf{Z_2}$) and $\mathbf{b = (b_1, b_2)}$ are the coefficients of the regression of **y** on ($\mathbf{X_1}$ & $\mathbf{X_2}$). First, we algebraically prove that $\mathbf{c_2 = b_2}$ and $\mathbf{c_1}$ is biased and $\mathbf{c_2}$ unbiased. After that, we verify our claims using real-life data(on gasoline) by using R for computation.

THEORETICAL RESULT:

We have,

$$\mathbf{Z_2 = M_1 X_2} \quad \text{(from the regression of } \mathbf{X_2} \text{ on } \mathbf{X_1}\text{)}$$

where (i) $\mathbf{M_1 = I - X_1(X_1^T X_1)^{-1} X_1^T}$ from the regression of $\mathbf{X_2}$ on $\mathbf{X_1}$

(ii) $\mathbf{Z_2}$ denotes residuals from the regression of each variable of $\mathbf{X_2}$ on $\mathbf{X_1}$

Now,

$$\mathbf{X_1^T M_1 = X_1^T - (X_1^T X_1)(X_1^T X_1)^{-1} X_1^T = 0}$$

Thus,

$$\mathbf{X_1^T Z_2 = X_1^T M_1 X_2 = 0 \quad \ldots . (3)}$$

So, when we regress **y** on ($\mathbf{X_1}$, $\mathbf{Z_2}$), we can instead regress separately on $\mathbf{X_1}$ and $\mathbf{Z_2}$. This will hold because the moment matrix is a block diagonal, with blocks $\mathbf{X_1^T X_1}$ and $\mathbf{Z_2^T Z_2}$, its inverse will likewise be a block diagonal and hence regression of **y** on $\mathbf{X_1}$ and $\mathbf{Z_2}$ is the same as separate regressions on the two data sets (as $\mathbf{X_1}$ and $\mathbf{Z_2}$ are orthogonal using **(3)**).

Now,

$$\mathbf{c_1 = (X_1^T X_1)^{-1} X_1^T y} \quad \text{----- (1)} \quad \text{(regressing y on } X_1 \text{ separately)}$$

$$\mathbf{c_2 = (Z_2^T Z_2)^{-1} Z_2^T y} \quad \text{----- (2)} \quad \text{(regressing y on } Z_2 \text{ separately)}$$

Since $\mathbf{M_1}$ is symmetric and idempotent and $\mathbf{Z_2 = M_1 X_2}$, using both of these in **(2)**

$$\mathbf{c_2 = (X_2^T M_1 X_2)^{-1} (X_2^T M_1 y)}$$

In the original regression of **y** on $\mathbf{X_1}$ **and** $\mathbf{X_2}$, $\mathbf{b_2}$ can be calculated using Frisch–Waugh–Lovell Theorem. Using that, it comes out to be the same as $\mathbf{c_2}$, implying $\mathbf{c_2 = b_2}$.

But,

$$\begin{aligned}\mathbf{c_1} &= \mathbf{(X_1^T X_1)^{-1} X_1^T y} \\ &= \mathbf{(X_1^T X_1)^{-1} X_1^T (X_1 b_1 + X_2 b_2 + e)} \\ &= \mathbf{b_1 + (X_1^T X_1)^{-1} X_1^T X_2 b_2} \qquad \mathbf{(X_1^T e = 0} \text{ as OLS residuals)}\end{aligned}$$

From the above expression, it is clear that $\mathbf{c_1}$ is not equal to $\mathbf{b_1}$ unless $\mathbf{X_1^T X_2 = 0}$

Hence we conclude $\mathbf{c_1}$ **is biased whereas** $\mathbf{c_2}$ **is unbiased**

EMPIRICAL RESULTS:

We have taken

$\mathbf{X_1}$ matrix as [1(constant), GASP, and PCINCOME]

$\mathbf{X_2}$ matrix as [PD, PN, and PS]

$\mathbf{y}$ = GASEXP

$\mathbf{Z_2}$ = [Ed, En, Es], the residuals matrix from the regression of $\mathbf{X_2}$ on $\mathbf{X_1}$.

where PCINCOME is taken to be INCOME/POP (Per capita Income)

Models

	Dependent variable:	
	GASEXP)	
	X1-X2	X1-Z2
	(1)	(2)
INTERCEPT	-22.343***	-64.568***
	(4.466)	(3.461)
GASP	1.231***	1.537***
	(0.063)	(0.026)
PCINCOME	452.034***	768.261***
	(97.353)	(63.805)
PD	-0.697***	
	(0.068)	
PN	-0.125	
	(0.214)	
PS	0.684***	
	(0.094)	
Ed		-0.697***
		(0.068)
En		-0.125
		(0.214)
Es		0.684***
		(0.094)
Observations	52	52
R2	0.999	0.999
Adjusted R2	0.999	0.999
Residual Std. Error (df = 46)	2.319	2.319
F Statistic (df = 6; 46)	13,139.820***	13,139.820***

CONCLUSION:

If we compare the coefficients of both the regressions, we find that the coefficient of $\mathbf{X_2}$ and $\mathbf{Z_2}$ in both the cases is the same showing that **$b_2 = c_2$ or c_2 is unbiased whereas c_1 is not.**

Hence theoretical results match with the empirical results.

TEAM DETAILS:

Danish Ahmad(160219)
Sakib Malik(180652)
Shivam Goel(180714)