European Journal of Mathematics and Statistics
Vol 4 | Issue 5 | October 2023
ISSN 2736-5484

RESEARCH ARTICLE

A Cautionary Note on the Use of Linear Regression for Hypothesis Testing

Gregory L. Light*

ABSTRACT

This note draws researchers' attention to the use of linear regression for the purpose of conducting a hypothesis testing. Even when multiple explanatory variables are included in a regression equation to preclude the hazard of a simple regression of omitting other factors, multicollinearity unfortunately is inherent in multiple regression simply because the included explanatory variables can share some common parameter domain; that is, they co-vary. Here we shall show that however one transforms the variance-covariance matrix of the least-squares estimation to reduce the estimation errors, the procedure amounts to affine transformations of the explanatory variables so that despite the transformation they remain to co-vary, rendering the coefficient as a partial derivative invalid. The root of this problem originates from the fact that one receives the explanatory variables' values by observation rather than predetermining their values and then collecting the corresponding dependent variable' values. This situation becomes especially disconcerting when a transformed explanatory variable has its estimated coefficient enjoying an exceptional degree of confidence but of no mathematical status as a partial derivative, misleading engineering or medical prescriptions and public policies.

Keywords: Model misspecification, multicollinearity, orthogonalization, ridge regression.

Submitted: July 17, 2023

Published: October 16, 2023

10.24018/ejmath.2023.4.5.268

Department of Finance, Providence College, Providence, Rhode Island, 02918, USA.

Corresponding Author:
e-mail: glight@providence.edu

1. Introduction

It is a well-recognized fact that a simple linear regression can cause the problem of omitted variables or model misspecification. Thus, multiple regression serves as a better methodology; here, however, one must exercise caution about the choices of the explanatory variables and also the design of the sample. In empirical work, there are mainly two approaches: observational (as in astronomy) or experimental (as in a particle collider). In practical applications of regression, mostly it is the former: one draws a sample from the underlying population and observes the input data [1]. This presents problems, as the analyst has no control over the observed values, say, $\{(x_{i1}, x_{i2}; y_i) \mid i = 1, 2, \cdots, n\}$. This note seeks to draw researchers' attention to a highly likely situation where x_1 and x_2 are both functions of t and s (connoting time and space), i.e., $x_1(t, s)$ and $x_2(t, s)$, so that the very construct of $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$ is immediately invalid as $\beta_j = (\partial E(Y))/(\partial X_j)$, $j = 1, 2$, does not exist. In the next section we will show that all modifications of the "variance-covariance matrix" $(X^T X)$ of the least-squares estimation for the purpose of alleviating the problem of multicollinearity cannot treat this problem of explanatory variables sharing the same parameter domain. Then in Section 3 we will conclude with a summary remark.

2. Analysis

All remedial procedures against the problem of multicollinearity are based on modifications of the matrix $\left(X^T X\right)$ of the ordinary least-squares estimation [2], in the case of two explanatory variables,

$$X^T X = \begin{pmatrix} \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2 & \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) \\ \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) & \sum_{i=1}^{n} (x_{i2} - \overline{x}_2)^2 \end{pmatrix}. \tag{1}$$

Since $\left(X^T X\right)$ is a linear operator, transformations of which are in general linear or affine [3], [4]. Consider a linear transformation by $k_1 > 1$ of the $(1,1) - entry$ of the above matrix,

$$k_1^2 \sum_i (x_{i1} - \overline{x}_1)^2 = \sum_i (k_1 x_{i1} - k_1 \overline{x}_1)^2 ; \tag{2}$$

then each individual observation i of x_1 (t,s) undergoes an affine transformation,

$$k_1 \cdot x_{i1}(t_i, s_i) - (k_1 \overline{x}_1), \tag{3}$$

which still ends as a function of (t,s). Consider now an affine transformation of the $(1,1) - entry$ of the above matrix by $M_1 > 0$, i.e.,

$$M_1 + \sum_i (x_{i1} - \overline{x}_1)^2, \tag{4}$$

but then the above (4) can be re-expressed as (2),

$$\begin{aligned} &M_1 + \sum_i (x_{i1} - \overline{x}_1)^2 = k_1^2 \sum_i (x_{i1} - \overline{x}_1)^2 = \sum_i (k_1 x_{i1} - k_1 \overline{x}_1)^2 ; \\ &\text{with } k_1^2 = \frac{M_1}{\sum (x_{i1} - \overline{x}_1)^2} + 1. \end{aligned} \tag{5}$$

Thus, each diagonal entry of $\left(X^T X\right)$ remains as a function of (t,s) following any linear/affine transformations, and thereof each transformed explanatory variable a′ la (3) is still a function of (t,s). Next, consider the off-diagonal entries of $\left(X^T X\right)$: Denote the transformed x_1 and x_2 by x_1^* and x_2^* respectively; then one has

$$\begin{aligned} &\sum_i \left(x_{i1}^* - \overline{x}_1^*\right)\left(x_{i2}^* - \overline{x}_2^*\right) = \left\langle k_1 x_{1,n\times 1} - k_1 \overline{x}_{1,n\times 1},\, k_2 x_{2,n\times 1} - k_2 \overline{x}_{2,n\times 1} \right\rangle \\ &= k_1 k_2 \cdot (\langle x_1, x_2 \rangle - \langle x_1, \overline{x}_2 \rangle - \langle \overline{x}_1, x_2 \rangle + \langle \overline{x}_1, \overline{x}_2 \rangle), \end{aligned} \tag{6}$$

also still a function of (t,s).

As such, any linear or affine transformation of the matrix $\left(X^T X\right)$ amounts to a transformation of an explanatory variable x_j into $k_j \cdot x_j(t,s) - m_j = f_j(t,s)$. Consequently, the transformed regression equation remains as

$$\hat{y}^* = b_1 x_1^*(t,s) + b_2 x_2^*(t,s) + \cdots + b_k x_k^*(t,s) ; \tag{7}$$

i.e., the root problem of multicollinearity remains: while the t-statistics of $b_1, b_2, \cdots, b_k$ can be increased to high (absolute) values to the extent that the 99% confidence intervals around $b_1, b_2, \cdots, b_k$ are extremely narrow, one still has the problem

$$E\left(b_j\right) = \beta_j = \frac{\partial E(Y)}{\partial X_j^*}, j = 1, 2, \cdots, k, \text{do not exist}, \tag{8}$$

by the simple fact that $x_j^* = f_j(t,s), \forall j = 1, 2, \cdots, k$; i.e., all the explanatory variables co-vary. Therefore, the engineered highly significant estimates of the coefficients can only lead to a false confidence in their values [5]–[7].

3. Summary Remark

From the above analysis, we see that hypothesis testing by a regression equation without an underlying mathematical model as based on theories can be problematic, unless one fixes the explanatory variables' input values $\{(x_{i1}, x_{i2}, \cdots, x_{ik}) \mid i = 1, 2, \cdots, n\}$ in advance and then collects the dependent variable's values $\{y_i \mid i = 1, 2, \cdots, n\}$. Otherwise, no matter how one transforms the matrix $\left(X^T X\right)$, the problem of $x_j^*(t,s)\, \forall j$ persists; then the transformation not only is futile but can also be dangerous—considering a wrong sign of a coefficient (in the sense of $\partial X_j / \partial t > 0, \partial X_j / \partial s > 0, \partial Y / \partial t > 0, \partial Y / \partial s > 0$ but $b_j < 0$) that yet carries a high degree of confidence in engineering or medical applications.

Conflict of Interest

Author declares that he does not have any conflict of interest.

References

[1] Kmenta J. *Elements of Econometrics*. New York: Macmillan; 1971.
[2] Huffel SV, Vandewalle J. *Algebraic Connections between Total Least Squares Estimation and Classical Linear Regression in Multicollinearity Problems*. Philadelphia, PA: SIAM; 1991, ch. 9. doi: 10.1137/1.9781611971002.
[3] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12:55–67. doi: 10.2307/1267351.JSTOR1267351.
[4] Liu K. A new class of biased estimate in linear regression. *Commun Stat Theory Methods*. 1993;22:393–402. doi: 10.1080/03610929308831027.
[5] Liu R, Wang H, Wang S. Functional variable selection via Gram-Schmidt orthogonalization for multiple functional linear regression. *J Stat Comput Simul*. 2018;88:3664–80. doi: 10.1080/00949655.2018.1530776.
[6] Oman SD. A confidence bound approach to choosing the biasing parameter in ridge regression. *JASA*. 1981;76:452–61. doi: 10.2307/2287849.
[7] Smith G, Campbell F. A critique of some ridge regression methods. *JASA*. 1980;75:74–81. doi: https://doi.org/10.1080/01621459.1980.10477428.