International Mathematical Forum, Vol. 18, 2023, no. 2, 75 - 81
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/imf.2023.912385

A Note on Orthogonalization against Multicollinearity in Linear Regression, with a Consideration of Spacetime as the Common Parameter Domain

Gregory L. Light

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

Abstract

In estimating/testing a functional relationship in Economics, one collects data -- both the dependent variable and the explanatory variables, which is not the same as an experiment in Physics with all the independent variables in full control by the analyst. This brings about the problem of multicollinearity in multiple linear regression to all fields that do not enjoy true degrees of freedom in the causal variables of a regression model. This note presents a simple example, where a pair of variables, u and v, seeks to explain y, but u(t, s) and v(t, s) share one common parameter domain, t and s, so that it becomes evident that the regression model y = a + bu + cy + e is simply invalid. We thus recommend constructing regression models based on independent variables true to their definition of independence, such as time and space, by using a spatiotemporal sample.

Mathematics Subject Classification: 62D20, 62H12, 62J05, 62J07

Keywords: Gram-Schmidt, ridge regression, Liu estimator, model misspecification

1 Introduction

There is a fundamental schism between the construct of a regression equation $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ and its algebraic underpinning $E(Y) = f(X_1, X_2)$: whereas the former entails a sampling of $\{(y_i, x_{i1}, x_{i2})\}_{i=1}^{n}$ usually with the values of $\{(x_{i1}, x_{i2})\}$ not freely set in advance, the latter casts X_1 and X_2 as quantities enjoying their own degrees of freedom. That is, there is a basic contradiction between a fixed 3-tuple (y_i, x_{i1}, x_{i2}) as from one particular observation i and the standard understanding of the formulation of a function $y = f(x_1, x_2)$.

To be sure, for implementing regression properly, one first derives a formula by theory and then experiments with alternative values of the independent variables to investigate their effects on the dependent variable. Yet, this condition of free choices of the values of the independent variables is not in the underlying assumptions of the regression model [3]. That is, the use of this methodology leans more toward observational, as in astronomy, rather than experimental, as in physical laboratories.

As such, one must exercise caution against an overreliance on "regression for a functional relationship through plain observations of a sample," wherein multi-collinearity of $\{(x_{i1}, x_{i2})\}$ can present problems, since the inseparableness of the correlated independent variables compounds the variance of the random disturbances, rendering the estimated coefficients suffering from higher errors. Accordingly, it appears that the remedy might be a reduction of the sample coefficient of correlation $|r|$ between any two independent variables through some schemes, notably, the ridge regression [1, 4, 7], wherein the variance-covariance matrix has its diagonal artificially added by a constant so as to turn the column vectors "more orthogonal" (: inner products closer to zero) - - a procedure amounting to decreasing the proportions of covariances to variances, hence lowering $\{|r_{jk}|\}$. It turns out that this logic is false: whereas high $\{|r_{jk}|\}$ are definitely detrimental to the estimation of the coefficients of the independent variables $\{X_j, X_k\}$, the converse is not true; this is demonstrated in the next section by an example. We then conclude with a summary remark in Section 3.

2. Analysis

Since the sample coefficient of correlation between any two variables depends on the inner product of their values centered at their means, we consider an application of the

Gram-Schmidt orthogonalization [5] of a pair of highly correlated independent variables (u,v), where

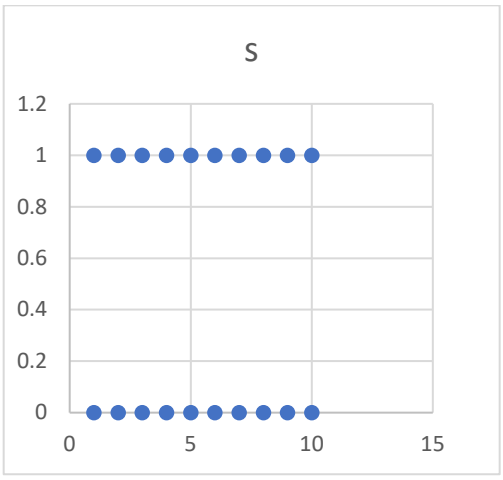
$$u = 3t + 5s,$$
$$v = 3t + 6s,$$
$$(t,s) \in \mathbb{R}^2,$$

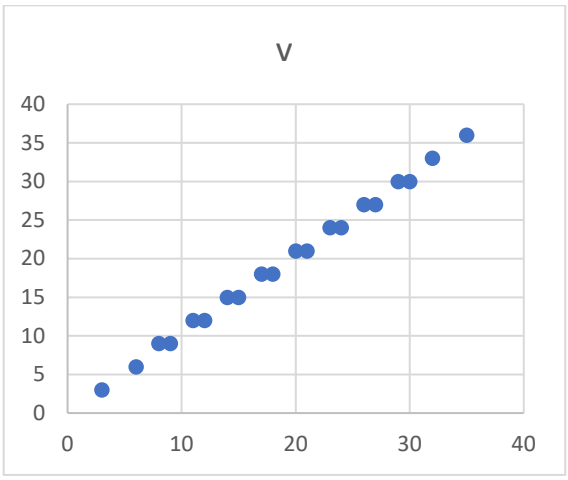
over a sample as based on:

t	s	u	v
1	0	3	3
1	1	8	9
2	0	6	6
2	1	11	12
3	0	9	9
3	1	14	15
4	0	12	12
4	1	17	18
5	0	15	15
5	1	20	21

t	s	u	v
6	0	18	18
6	1	23	24
7	0	21	21
7	1	26	27
8	0	24	24
8	1	29	30
9	0	27	27
9	1	32	33
10	0	30	30
10	1	35	36

where $r(t,s) = 0$, and $r(u,v) = 0.999$, as graphed below:





Set - -

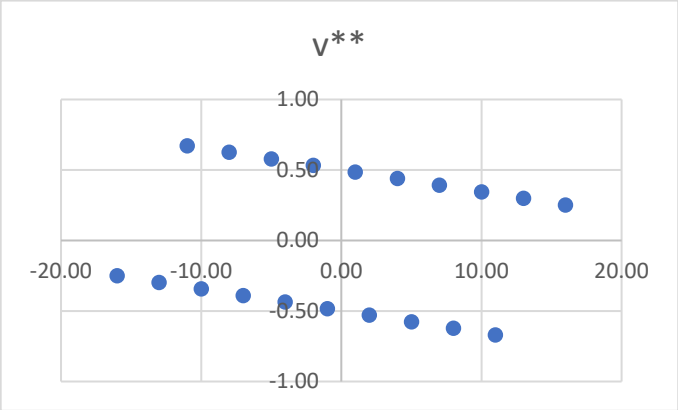
$$u_i^* \equiv u_i - \overline{u},\ v_i^* \equiv v_i - \overline{v},\ \cos\theta := \left.\frac{\langle u^*, v^* \rangle}{\|u^*\|\|v^*\|}\right|_{l^2},\ \text{and}$$

$$v_i^{**} := v_i^* - \frac{\|v^*\|\cos\theta}{\|u^*\|} u_i^*,\ i = 1, 2, \cdots, 20,$$

or in values - -

u*	v**		u*	v**
-16.00	**-0.25**		**-1.00**	**-0.48**
-11.00	**0.67**		**4.00**	**0.44**
-13.00	**-0.30**		**2.00**	**-0.53**
-8.00	**0.62**		**7.00**	**0.39**
-10.00	**-0.34**		**5.00**	**-0.58**
-5.00	**0.58**		**10.00**	**0.34**
-7.00	**-0.39**		**8.00**	**-0.62**
-2.00	**0.53**		**13.00**	**0.30**
-4.00	**-0.44**		**11.00**	**-0.67**
1.00	**0.48**		**16.00**	**0.25**

Then $r\left(u^*, v^{**}\right) = 0$, as shown in the following graph:



Consider the following population regression equation,

$$Y = 1 + 2t + 3s + \varepsilon$$
$$\equiv 1 + u - \frac{1}{3}v + \varepsilon,$$
$$\varepsilon \sim N\left(0, \sigma^2 = 0.25\right).$$

Upon the ordinary least-squares estimation of
$y_i^* \equiv y_i - \overline{y} = b_1 u_i^* + b_2 v_i^{**} + e_i$
with the following simulated Y with their associated Y**

y	**y***		**y**	**y***
3.27	**-10.17**		**12.81**	**-0.63**
6.16	**-7.28**		**16.01**	**2.56**
4.85	**-8.59**		**14.65**	**1.21**
8.76	**-4.68**		**17.71**	**4.26**
6.42	**-7.02**		**16.55**	**3.11**
10.35	**-3.10**		**20.45**	**7.01**
8.49	**-4.95**		**18.82**	**5.38**
12.56	**-0.88**		**22.11**	**8.67**
10.92	**-2.52**		**20.92**	**7.47**
13.09	**-0.35**		**23.95**	**10.51**

we obtained

$$\hat{y}_i^* = 0.66u_i^* + 0.06v_i^{**}, \textit{ with the } t-\textit{statistics}:$$
$$t_{u^*} = 71.18, \textit{ and } t_{v^{**}} = 0.36,$$

where the estimation of the effect of $u\left(or\ u^*\right)$ *on* $y\left(or\ y^*\right)$ by 0.66 deviated from the true coefficient 1 by -0.34, yet with $t = 71.18$. We thus see that the above reduction of r by the Gram-Schmidt orthogonalization can result in false confidence in an estimated coefficient that greatly deviated from its true value, the same drawback as in the ridge regression [2, 6]. Two remarks here are in order: (a) the two perpendicular vectors u^*, v^{**} in $\mathbb{R}^{20}$ do not lend to orthogonality in $\mathbb{R}^2$ as in $\left\langle (x_1, 0), (0, x_2) \right\rangle = 0, (X_1, X_2) \in \mathbb{R}^2$, but nevertheless, (b) the variance-covariance matrix is diagonal:

$$\left(\frac{1}{19}\right)\begin{pmatrix} \left\|u^*\right\|_{l^2}^2 & \left\langle u^*, v^{**} \right\rangle \\ \left\langle u^*, v^{**} \right\rangle & \left\|v^{**}\right\|_{l^2}^2 \end{pmatrix} = \left(\frac{1}{19}\right)\begin{pmatrix} 1610 & 0 \\ 0 & 4.61 \end{pmatrix} = \begin{pmatrix} 84.7 & 0 \\ 0 & 0.24 \end{pmatrix}.$$

Now, a straightforward regression of Y on u, v also led to the following unsatisfactory results:

$\hat{y}_i = 0.93 + 0.59u_i + 0.06v_i$, *with*

$t_{\text{constant}} = 4.60, t_u = 3.29, t_v = 0.35.$

This led to the following regression as based on the parameter domain, $t \times s$:

$$\hat{y}_i = 0.93 + 1.97t_i + 3.35s_i, \text{ with}$$
$$t_{\text{constant}} = 4.60, t_t = 66.34, t_s = 19.61,$$
$$R^2 = 0.996,$$

which suggested that in treating multi-collinearity one simply regresses the dependent variable y directly against time and some spatially oriented dummy variables, treating which as the common parameter domain for both the dependent variable y and all the correlated independent variables. From this perspective, one sees that the attempt of estimating $\beta_1 := \dfrac{\partial y}{\partial u(t,s)}$ as implied in the regression model $y = \beta_0 + \beta_1 u(t,s) + \beta_2 v(t,s) + \varepsilon$ be fundamentally invalid. It is clear from the above example that the only way to transform u and v into two orthogonal variables would be

$$\begin{pmatrix} 3 & 5 \\ 3 & 6 \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} t \\ s \end{pmatrix},$$

in which case, we come to the same conclusion that regressors ought to be time and space oriented.

3. Summary Remark

From the above analysis, we see that correlated independent variables in samples may share certain unspecified parameter domain so that the posited regression model per se is invalid; thereof, a reformulation of the model becomes necessary. In the case where a model is theoretically derived, we contend that regressing over a time-series cross-section sample as patterned after the above $\{(t_i, s_i)\}$ appears to be the most promising against multicollinearity. Otherwise, we suggest a multiplicative modeling followed by a log-linear transformation, since proportional relationships abound in Nature.

References

[1] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12** (1970), 55–67.
https://doi.org/10.1080/00401706.1970.10488634

[2] S.V. Huffel, J. Vandewalle, Algebraic connections between total least squares estimation and classical linear regression in multicollinearity problems (Chap. 9), *Frontiers in Applied Mathematics, The Total Least Squares Problem,* SIAM, Philadelphia, 1991. https://doi.org/10.1137/1.9781611971002

[3] J. Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971.

[4] K. Liu, A new class of biased estimate in linear regression, *Communications in Statistics-Theory and Methods*, **22** (1993), 393-402. https://doi.org/10.1080/03610929308831027

[5] R. Liu, H. Wang, S. Wang, Functional variable selection via Gram–Schmidt orthogonalization for multiple functional linear regression, *Journal of Statistical Computation and Simulation*, **88** (2018), 3664-3680. https://doi.org/10.1080/00949655.2018.1530776

[6] S.D. Oman, A confidence bound approach to choosing the biasing parameter in ridge regression, *Journal of the American Statistical Association*, **76** (1981), 452-461. https://doi.org/10.1080/01621459.1981.10477667

[7] G. Smith, F. Campbell, A critique of some ridge regression methods, *Journal of the American Statistical Association*, **75** (1980), 74-81. https://doi.org/10.1080/01621459.1980.10477428

Received: May 17, 2023; Published: June 6, 2023