MULTICOLLINEARITY MAY LEAD TO ARTIFICIAL INTERACTION: AN EXAMPLE FROM A CROSS SECTIONAL STUDY OF BIOMARKERS

Pornchai Sithisarankul[1], Virginia M Weaver, Marie Diener-West, Paul T Strickland

Division of Occupational and Environmental Health, Department of Environmental Health Sciences,Johns Hopkins School of Hygiene and Public Health, Baltimore, MD 21205, USA

Abstract. Collinearity is the situation which arises in multiple regression when some or all of the explanatory variables are so highly correlated with one another that it becomes very difficult, if not impossible, to disentangle their influences and obtain a reasonably precise estimate of their effects. Suppressor variable is one of the extreme situations of collinearity that one variable can substantially increase the multiple correlation when combined with a variable that is only modestly correlated with the response variable. In this study, we describe the process by which we disentangled and discovered multicollinearity and its consequences, namely artificial interaction, using the data from cross-sectional quantification of several biomarkers. We showed how the collinearity between one biomarker (blood lead level) and another (urinary trans, trans-muconic acid) and their interaction (blood lead level* urinary trans, trans-muconic acid) can lead to the observed artificial interaction on the third biomarker (urinary 5-aminolevulinic acid).

INTRODUCTION

Collinearity is the situation which arises in multiple regression when some or all of the explanatory variables are so highly correlated with one another that it becomes very difficult, if not impossible, to disentangle their inflences and obtain a reasonably precise estimate of their [separate] effects (Kotz and Johnson, 1985). Its synonyms were multicollinear relations, collinear relations, (near) dependencies, near collinearity, and near singularity (Belsley *et al*, 1980). *Collinearity* is usually used for collinearity between 2 variables, and *multicollinearity* for more than 2; but nowadays, collinearity and multicollinearity are used interchangeably (Kotz and Johnson, 1985).

Collinearity is a data problem, not a statistical problem (Belsley *et al*, 1980), but it can cause some statistical problems especially in statistical modeling and regression diagnostics. "It is a matter of degree rather than of all or nothing; we say that it is present when some auxiliary coefficients of determination are high" (Goldberger, 1968). "All" is the exact collinearity situation, whereas "none" is the exact orthogonality situation (Scialfa and Games, 1987). Although collinearity is not an uncommon issue in statistical modeling and regression diagnostics, it does not always cause problems since "it doesn't hurt so long as it doesn't bite" (Belsey *et al*, 1980).

One will find that collinearity makes regression coefficient estimates (b_i) unstable (changed) and even become not significant (increased standard error and variance). The common situation results in the usual symptoms of collinearity such as unstable estimates and increased standard error and variance (hence, the term "variance inflation"). It is often seen that $1 \geq (R^2_{x1} + R^2_{x2}) > R^2_{x1x2} > R^2_{xi} \geq 0$. ($R^2$ = coefficient of determination; $R^2_{xi} = R^2$ of the response variable explained by x_i). However, there are unusual or extreme situations of collinearity, and one of those is the existence of "suppressor variables" (Kotz and Johnson, 1985). A suppressor variable is highly correlated with another explanatory variable but uncorrelated with the response variable. And such a variable can substantially increase the multiple correlation when combined with a variable that is only modestly correlated with the response variable. Psychological statisticians are familiar with the concept of the suppressor variable (Kotz and Johnson, 1985), but scientists in other fields may be less familiar with it.

In this study, we used data from our previous studies (Sithisarankul *et al*, in preparation; Weaver *et al*, 1996a, b) in lead-exposed children to demonstrate an example of suppressor variables causing

[1]Current address: Department of Preventive and Social Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand.

multicollinearity which consequently caused an artificial interaction.

MATERIALS AND METHODS

All children who were seen at the Kennedy Krieger Lead Poisoning Prevention Clinic during a 4 week study period in September and October of 1994 were eligible for enrolment. These children were originally referred to the clinic for evaluation of elevated blood lead levels; some were receiving on-going follow-up for persistently elevated blood leads or other lead related concerns such as learning disorders. Parent/guardians were invited to participate while waiting for the physician visit. Explanations to the adults and children (appropriately age-modified) were provided and informed consent was obtained from all participants. Seventy-nine children were included in the study as previously reported (Weaver *et al*, 1996a). Several blood and urinary biomarkers were cross-sectionally quantified and reported in details elsewhere (Weaver *et al*, 1996a, b; Sithisarankul *et al*, in preparation). Venous blood lead levels (PbB) were obtained as part of routine clinical care. Urine samples were obtained and stored at -80°C until analyzed for urinary trans, trans-muconic acid (MA), urinary 5-aminolevulinic acid (ALAU), urinary cotinine, and urinary creatinine (CR).

PbB was assayed by anodic stripping voltammetry (Burtis and Ashwood, 1994). MA was determined by a high performance liquid chromatography method with structural confirmation by a gas chromatography/mass spectroscopy method (Weaver *et al*, 1996a, b). ALAU was determined by chemical derivatization followed by a high performance liquid chromatography with fluorescence detection (modified from Tomokuni *et al*, 1993a, b as reported in Sithisarankul *et al*, submitted). Cotinine was determined by radioimmunoassay (Langone *et al*, 1973; Haley *et al*, 1983) at the American Health Foundation, Valhalla, New York, USA (Weaver *et al*, 1996a, b). Creatinine was determined by a modified Jaffe's reaction (Sigma creatinine kit).

Since both lead and benzene can have adverse effects on the erythropoietic system, one of our *a priori* hypotheses was whether lead exposure (as measured by PbB) and benzene exposure (as measured by MA, an open-ringed urinary metabolite of benzene) interacted with each other on the erythropoietic system (as measured by ALAU, an intermediate substrate in the heme synthetic pathway). From our previous study, we found that creatinine-adjusted ALAU (ALAUCR) was the best surrogate for plasma ALA (Sithisarankul *et al*, submitted), so ALAUCR was the response variable we used.

We explored this possible interaction by looking at a scatterplot between PbB and ALAUCR dichotomized at the median of creatinine-adjusted MA (MACR) as shown in Fig 1. It suggested an

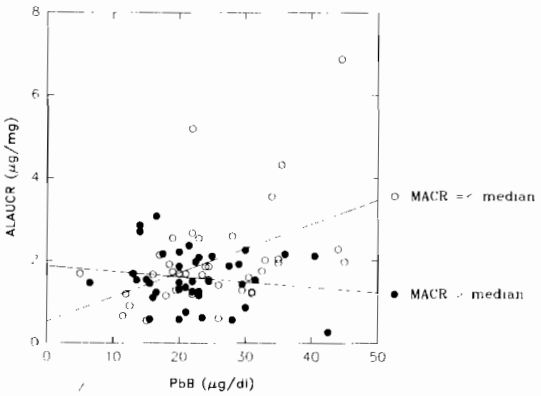

Fig 1–Association between PbB and ALAUCR dichotomized at median of MACR. Open circles: MACR = < median; y = 0.52 + 0.06x, r = 0.45, p = 0.005, n = 38. Closed circles: MACR > median; y = 1.86 - 0.01x, r = -0.15, p = 0.37, n = 39.

interaction between PbB and MACR on ALAUCR. Multiple linear regression analyses also showed that there was an interaction between PbB and MA on ALAUCR, no matter MA was put into the model as 1) unadjusted MA as a continuous variable, 2) unadjusted MA as a dummy variable (dichotomized at median: 1 > median, 0 ≤ median), 3) creatinine-adjusted MA (MACR) as a continuous variable, 4) MACR as a dummy variable (dichotomized at median: 1 > median, 0 ≤ median), 5) log-transformed MA as a continuous variable, or 6) log-transformed MACR as a continuous variable. For simplicity, we showed only the model with MACR as a continuous variable as in Table 1. The interaction term, PbB*MACR, was statistically significant (p=0.0123).

The observed interaction was biologically plausible. However, we statistically explored it as follows: First, we compared PbB and ALAUCR in

Table 1

Predictors of ALAUCR identified by multiple linear regression (n = 79-2 with missing data on PbB = 77).

Variable	b	SE	T for b = 0	p-value
Age	-0.1299	0.0646	-2.010	0.0482
PbB	0.0474	0.0147	3.217	0.0019
MACR	0.0025	0.0011	2.251	0.0274
PbMACR[a]	-0.0001	0.00005	-2.567	0.0123

Model F = 5.029, p = 0.0012, R^2 = 0.2184
[a]interaction term = PbB*MACR

both groups to exclude the possibility that the difference in distribution in either of these variables caused this artificial interaction. Second, we excluded the 4 highest points (with ALAUCR > 3.28) and re-analyzed in the same manner as the previous 2 steps to exclude if this observed interaction was driven by these 4 data points. Third, we performed collinearity diagnostics to investigate the existence and magnitude of multicollinearity. Fourth, we performed all the possible auxiliary regressions (modeling one explanatory variable on the others) to address the nature of multicollinearity in this data set.

All analyses were performed on SAS 6.04.

RESULTS

The correlation analyses of relevant variables were computed. ALAUCR was correlated with PbB (r = 0.28, p = 0.01) and age (r = -0.31, p = 0.006), but not with MACR (r = -0.08, p = 0.5) or PbMACR (r = -0.11, p = 0.33). MACR was strongly correlated with PbMACR (r = 0.94, p = 0.0001).

First, to investigate if this was a false interaction (Jaccard *et al*, 1990) simply caused by the difference in the range of PbB and/or ALAUCR in both groups, we compared PbB and ALAUCR in both groups by Wilcoxon rank sum test. There were no differences in PbB (p = 0.17 by Wilcoxon rank sum test) and ALAUCR (p = 0.31 by Wilcoxon rank sum test) in the 2 groups.

Second, by looking at Fig 1, it might be argued that this interaction was driven by the 4 highest points (those with ALAUCR > 3.28) which were exclusively in the MACR ≤ median group. We addressed this issue by excluding these 4 points and re-analyzed as first and second steps. We also re-analyzed the model. There were no significant differences, implying that this interaction was not driven by the 4 highest data points.

Third, the collinearity diagnostics in Table 2 showed that MACR and PbMACR had high variance inflation factor (VIF), low tolerance, and that condition indexes of PbB, MACR, and PbMACR (in bold), were high (> 0.5), indicating that they were highly correlated (*ie* collinear).

Fourth, in order to address the nature of this multicollinearity, we performed all the possible auxiliary regression analyses by modeling each explanatory variable on the other explanatory variables. We found that PbMACR and MACR were highly correlated. Each of them alone was not correlated with PbB, but together they were highly correlated with PbB. This is the situation of suppressor variables as mentioned earlier. In a similar fashion, PbB-MACR-PbMACR multicollinearity made each of them more significant as explanatory variables to ALAUCR in the model than any single one of them as emphasized in Table 3.

DISCUSSION

We have presented the data and the process by which we disentangled one of the problems of multiple regression: multicollinearity. We demonstrated that a high correlation between PbMACR and MACR made them together have a high corre-

Table 2

Collinearity diagnostics of the model in Table 1.

Variable	DF	Tolerance	Variance inflation
Intercept	1		0.0000
Age	1	0.9424	1.0611
PbB	1	0.6873	1.4549
MACR	1	0.0764	13.0918
PbMACR	1	0.0759	13.1692

Collinearity diagnostics

Number	Eigenvalue	Condition Number	Var prop intercept	Var prop PbB	Var prop age	Var prop MACR	Var Prop PbMACR
1	3.5503	1.0000	0.0028	0.0050	0.0073	0.0026	0.0027
2	1.2316	1.6979	0.0033	0.0062	0.0112	0.0171	0.0163
3	0.1467	4.9201	0.0000	0.2550	0.3782	0.0124	0.0091
4	0.0522	8.2467	0.2401	0.0350	0.3543	0.2824	0.3097
5	0.0193	13.5724	0.7538	**0.6988**	0.2490	**0.6854**	**0.6622**

lation with PbB, and a high correlation between PbMACR-MACR and PbB made them together have a high correlation with ALAUCR. So to summarize technically: PbMACR and MACR were suppressor variables in predicting PbB; and PbB, MACR, and PbMACR were suppressor variables in predicting ALAUCR. Since the interaction term is a product term of 2 or more main terms, it is not uncommon for a main term and its interaction term to be correlated. In this data set, PbMACR was correlated with MACR but not with PbB. Parallel analyses replacing MA with urinary cotinine yielded similar results (not shown) but not as dramatic as MA.

Cross-sectionally looking at several biomarkers is subject to this issue of multicollinearity. Caution should be exercised in interpreting the data, the correlation, and the interaction between these biomarkers. The presence of collinearity, either the usual type or the suppressor one, makes it difficult to distinguish the effect of one variable from the others. In this example, it makes us unable to interpret the effects of PbB, PbMACR, and MACR on ALAUCR. Consequently, it is impossible for us to answer whether there is an interaction between lead exposure (as PbB) and benzene exposure (as MACR) on heme synthesis (as ALAUCR), based on this cross-sectional data and the presence of two levels of multicollinearity as simplified in Fig 2.

As shown in Table 3, suppressor variables do not seem to "deflate" the variance (as the opposite to the usual type of collinearity which inflates the variance) as one might expect. In fact, they seem to "inflate" or increase the variance as well, but they also increase the regression coefficient estimates (b_i) to a greater extent than they do to the variance or standard error (SE). This can be seen for PbB and MACR. For PbB, SE increases from 0.0130 to 0.0147, and b_i increases from 0.0332 to 0.0474. For MACR, SE increases from 0.0003 to 0.0011, and b_i increases from -0.0002 to 0.0025. However, they may decrease b_i as for PbMACR. PbMACR's SE increases from 0.00001 to 0.00005, whereas its b_i decreases from -0.00001 to -0.0001. In either case, it makes the absolute T-value b_i/standard error larger and more significant.

Also, of note, is the increase in R^2 in this situation of suppressor variables. In Table 3 with ALAUCR as the response variable, $R^2_{age} + R^2_{PbB} + R^2_{MACR} + R^2_{PbMACR} = 0.0857 + 0.0802 + 0.0064 +$

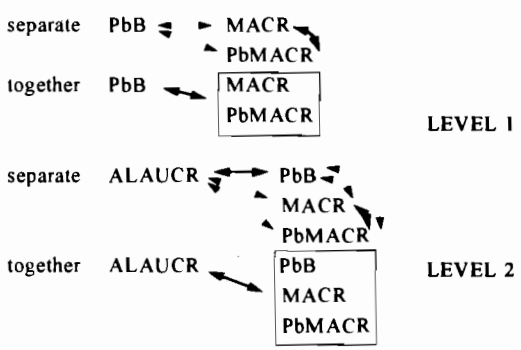

Fig 2–Diagrams showing correlation between ALAUCR, PbB, MACR, and PbMACR; dotted lines represent weak correlation, solid lines represent strong correlation.

$0.0125 = 0.1848$, which is lower than $R^2_{age, PbB, MACR, PbMACR}$ (0.2184). So we propose that one of the clues for suppressor variables should be that $1 \geq R^2_{x1x2} > (R^2_{x1} + R^2_{x2}) > R^2_{xi} \geq 0$, which is opposite to the usual collinearity that $(R^2_{x1} + R^2_{x2}) > R^2_{x1x2}$.

ACKNOWLEDGEMENTS

The authors acknowledge Grant No. ES03819.

REFERENCES

Belsey DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of colli-

Table 3

Comparison of b_i, SE, T for $b_i = 0$, p-value, and R^2 of each explanatory variable, separate and together, in predicting the response variable, ALAUCR (n = 79-2 with missing data on PbB = 77).

Explanatory variable	Estimate	Separate	Together
Age	b_i	-0.1763	-0.1299
	SE	0.0665	0.0646
	T for $b_i = 0$	-2.652	-2.010
	p-value	0.0098	0.0482
	R^2	0.0857	-
PbB	b_i	0.0332	0.0474
	SE	0.0130	0.0147
	T for $b_i = 0$	2.558	3.217
	p-value	0.0126	0.0019
	R^2	0.0802	-
MACR	b_i	-0.0002	0.0025
	SE	0.0003	0.0011
	T for $b_i = 0$	-0.694	2.251
	p-value	0.4898	0.0274
	R^2	0.0064	-
PbMACR	b_i	-0.00001	-0.0001
	SE	0.00001	0.00005
	T for $b_i = 0$	-0.974	-2.567
	p-value	0.3332	0.0123
	R^2	0.0125	-
Age, PbB, MACR, PbMACR (together)	R^2	-	0.2184

nearity. Wiley Series in Probability and Mathematical Statistics, New York, USA. 1980.

Burtis CA, Ashwood ER. Teitz Textbook of Clinical Chemistry. 2nd ed. Philadelphia: WB Saunders, 1994, pp. 1221-2, 2020-1, 2096-7.

Goldberger AS. Topics in Regression Analysis. Macmillan, New York, USA. 1968.

Haley NJ, Axelrad CM, Tilton KA. Validation of self-reported smoking behavior: Biochemical analyses of cotinine and thiocyanate. *Am J Public Health* 1983; 73 : 1204-7.

Jaccard J, Turrisi R, Wan CK. Interaction Effects in Multiple Regression. Series 72 in Quantitative Applications in the Social Sciences. Sage Publications, CA, USA. 1990.

Kotz S, Johnson NL. eds. Encyclopedia of Statistical Sciences. New York: John Wiley and Sons. 1985; 2: pp. 44, 5 : pp 639-43.

Langone JJ, Gjika HB, Van Vunakis H. Nicotine and its metabolites. Radioimmunoassays for nicotine and cotinine. *Bichemistry* 1973; 12 : 5025-30.

Scialfa CT, Games PA. Problems with step-wise regression in research on aging and recommended alternatives. *J Gerontol* 1987; 42 : 579-83.

Sithisarankul P, Weaver VM, Davoli CT, Strickland PT. Urinary 5-aminolevulinic acid in lead-exposed children. *Environ Health Perspect* (in preparation)

Sithisarankul P, Schwartz BS, Lee BK, Strickland PT. Urinary 5-aminolevulinic acid (ALA) adjusted by creatinine: A surrogate for plasma ALA. *Am Indust Hyg Assoc J* (submitted)

Tomokuni K, Ichiba M, Hirai Y. HPLC micro-method for determining δ-aminolevulinic acid in plasma. *Clin Chem* 1993; 39 : 169-170.

Tomokuni K, Ichiba M, Fujishiro K. Interrelation between urinary δ-aminolevulinic acid (ALA), serum ALA, and blood lead in workers exposed to lead. *Industrial Health* 1993; 31 : 51-7.

Weaver VM, Davoli CT, Heller PJ, *et al.* Benzene exposure, assessed by urinary trans, trans-muconic acid, in urban children with elevated blood lead levels. *Environ Health Perspect* 1996a; 104 : 318-23.

Weaver VM, Davoli CT, Murphy SE, *et al.* Environmental tobacco smoke exposure in inner-city children. *Cancer Epidemiol Biomakers Preven* 1996b; 5 : 135-7.