Presented at the 2016 ICEAA Professional Development & Training Workshop - www.iceaaonline.com/atlanta2016

THE COLLINEARITY KILL CHAIN:

Using the warfighter's engagement sequence to *detect*, *classify*, *localize*, and *neutralize* a vexing and seemingly intractable problem in regression analysis

> *"I believe that a substantial part of the regression and correlation analysis which have been made on statistical data in recent years is nonsense … If the statistician does not dispose of an adequate technique for the statistical study of the confluence hierarchy, he will run the risk of adding more and more variates [explanatory variables] in the study until he gets a set that is in fact multiple collinear and where his attempt to determine a regression equation is therefore absurd."*
>
> Ragnar Frisch, 1934, Nobel Laureate, Economics

ADAM JAMES AND BRIAN FLYNN

TECHNOMICS

1.0 INTRODUCTION

Defense analysts apply the regression model in a remarkably wide set of circumstances ranging from cost estimating to systems engineering to policy formulation. The model's frequency of use and scope of application are perhaps unparalleled in the professional's statistical toolkit. Unfortunately, Frisch's *multiple collinear* data sets are a frequent if often undetected companion in these parametric forays into national security. Whether exposed or not, they can undermine the foundations of the analysis, and, worst-case, like an IED, blowup the regression's computational algorithm.

A fundamental postulate of the classical least-squares model of equation (1) is the absence of a perfect, linear dependence among the explanatory or predictor variables.

(1) $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i.$

A serious condition may still persist in cases where perfect collinearity is avoided, but only just narrowly. And therein lies the root of the problem. To paraphrase Alain Enthoven,

Just how much linear dependence can be tolerated without undue damage to parameter estimates?

A less extreme but still often serious condition arises when this assumption is met, but perhaps only by a narrow margin. The term "multicollinearity," coined by the Nobel laureate Ragnar Frisch, captures all degrees of the problem. It's defined as the presence of a linear relationship, possibly but not necessarily perfect, between two or more of the supposedly "independent" variables in a regression equation, as exemplified in equation (2), where X_2 and X_3 from equation (1) are linearly but not perfectly related

(2) $X_{2i} = a + bX_{3i} + \mu_i.$

Multicollinearity may be weak or strong in practical applications. But, it's always present. Put another way, the explanatory variables are never orthogonal. There's always at least some small degree of linear association between them. Indeed, there's a natural tendency noted by Frisch to include in a regression equation additional variables that overlap the information content of those already present. This holds especially true in disciplines such as defense cost analysis where fresh information on potentially relevant independent variables is difficult or impossible to obtain. Good practice in CER development requires an assessment of the severity of multicollinearity since the consequences may include "bouncing β's," or unstable coefficient estimates, unreliable hypothesis testing, and wrongly specified models. The critical question becomes, from the practitioner's perspective,

At what point does multicollinearity become harmful, and what can be done about it?

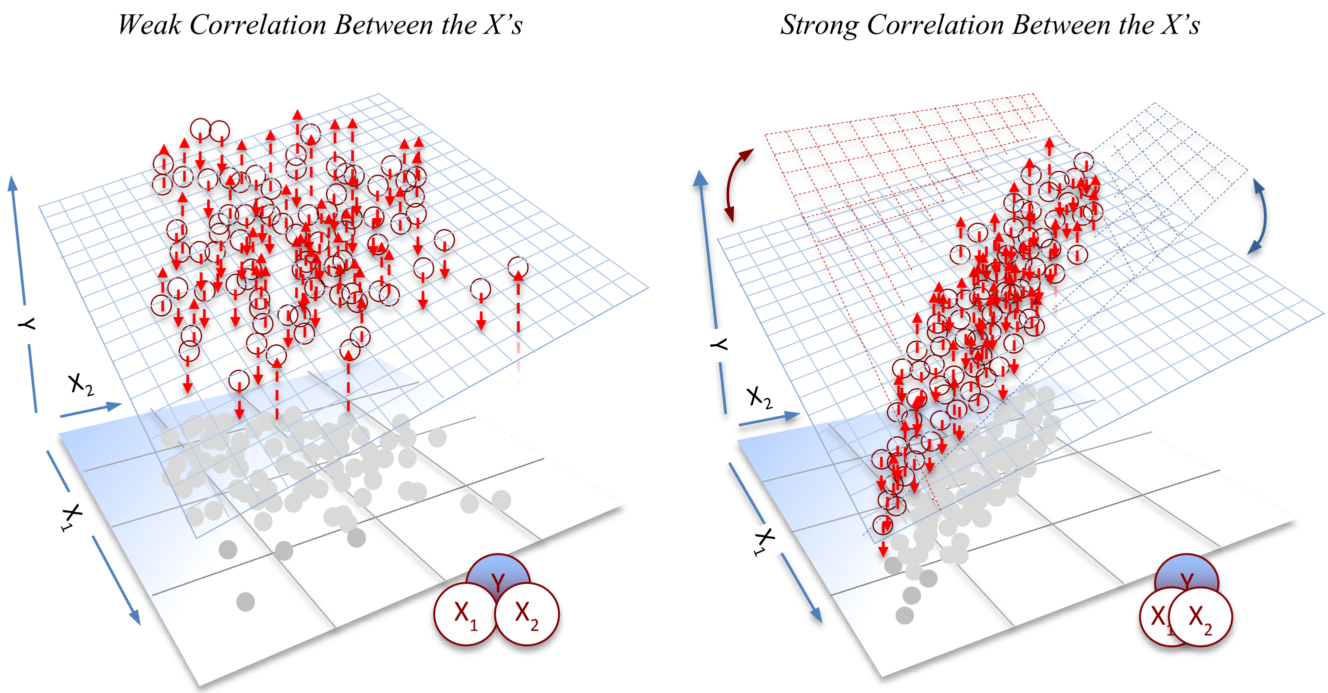
Somewhat akin to mine countermeasures in its complexity, the problem of multicollinearity defies easy solution. The statistical literature regards multicollinearity as one of the most vexing and intractable problems in all of regression analysis. Many tools and techniques are required for its detection and correction.

This paper examines the issue of multicollinearity from the fresh perspective of a statistical warfighter. A kill-chain of steps is offered to defeat this illusive foe. Heuristics, chi-squared tests, t-tests, and eigenvalues are the *Aegis*, *Predator*, *Bradley*, and *Ma Deuce 50-caliber* statistical equivalents used to *detect*, *classify*, *localize*, and *neutralize* the problem.

1.1 BACKGROUND

Multicollinearity often occurs when different explanatory variables in a regression equation rise and fall together. The sample data set, then, contains seemingly redundant information. Theory is often more fertile than data, and additional X's sometimes capture only nuances of the same, basic information. A frequently cited example is correlation between "learn" and "rate" in a cost-improvement curve, especially when rate is measured over-simplistically as lot size, or $Q_t – Q_{t-1}$, without taking into consideration the total work flow of a fabrication facility.[1]

Figure 1 contrasts weak and strong collinearity. In the case of weak correlation between X_1 and X_2 (left side) the regression plane is well-determined, with only slight movements in repeated samples. Classical ordinary least squares, in this case, yields stable estimates of β_1 and β_2. In the case of strong correlation between X_1 and X_2 (right side), on the other hand, observations (ordered triplets, X_1, X_2, Y) are in a tube. The regression plane wobbles, with significantly different positions likely in repeated samples but with E(Y) unaffected. The net result is *unstable* although unbiased estimates of β_1 and β_2.


Weak Correlation Between the X's

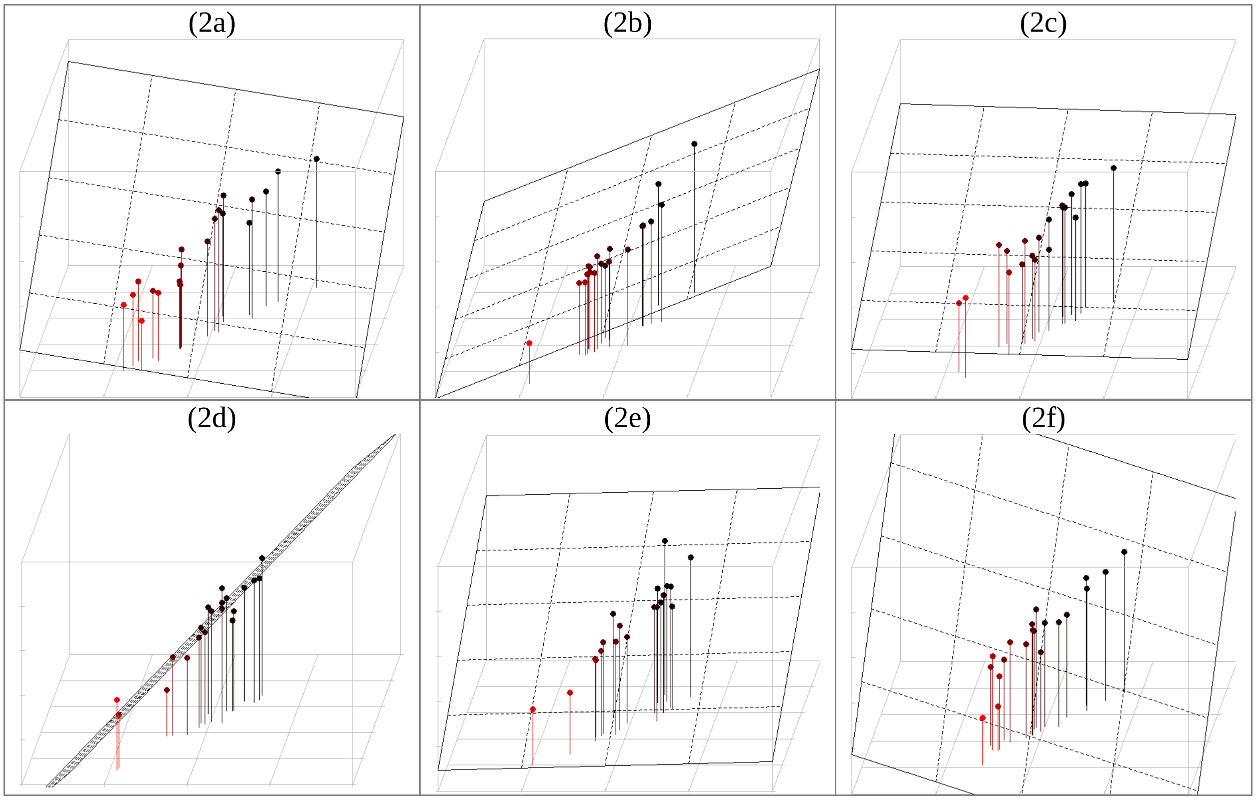
Strong Correlation Between the X's

Figure 1

[1] "Statistical Methods for Learning Curves and Cost Analysis," Goldberg, Matt, and Touw, Anduin, Center for Naval Analysis, 2003, pages 15-16.

Figures 2 and 3 demonstrate the concept on a simulated set of data from the same population regression model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Each plot holds constant the variance of ε and the number of data points. The only difference is the amount of correlation between X_1 and X_2. Each plot shows a 3D view of the data along with a best-fit plane.

Figure 2 executes the experiment with high correlation, 95%. In this scenario, the fit is highly unstable. The plane swings widely about the points. Each of the six successive samples from the population (2a to 2f) results in a drastically different regression equation. Therefore, in the practical real-world scenario with only one sample collection of data, but with high collinearity between the X's, the estimated regression equation may or may not be a reasonable model of the true, underlying variable relationships.



(2a) (2b) (2c)

(2d) (2e) (2f)

Figure 2

Figure 3, on the other hand, has a low correlation, 20%. In this case, the regression plane is much more stable. While it experiences random variation, the planes – or regression models – are similar to each other (3a to 3f). Regardless of the sample collected, the result is likely a reasonable estimate of the true, underlying relationship between the dependent and explanatory variables.

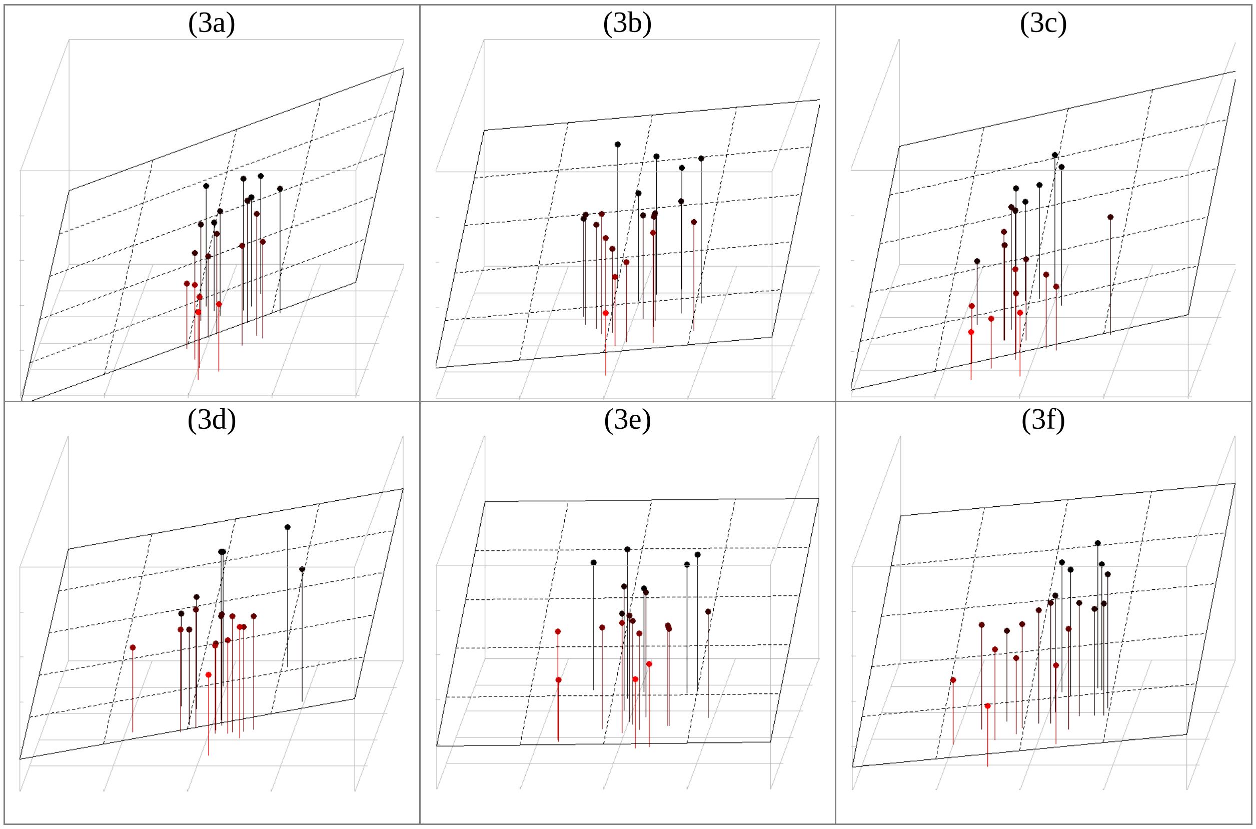


Figure 3

1.2 SYMPTOMS

Symptoms of multicollinearity include:

- The signs and values of estimated coefficients are inconsistent with the domain of knowledge
- High pairwise correlations among X's
- Mismatch of F and t tests
- Significant swings in parameter estimates when observations are added or deleted, and when one or more explanatory variables are deleted.

To illustrate, equation (3) proffers a linear relationship between spacecraft payload costs and payload weight and volume, with Table 1 presenting a sample of historical observations.

(3) $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_{i,}$ where

Y = payload cost in FY14K$
X_1 = payload weight (kg)
X_2 = payload volume (cm^3).

Spacecraft Payload Data		
Y	X_1	X_2
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

Table 1

OLS yields equation (4), with t-statistics in parentheses.

$$(4)\ \ \hat{Y} = 24.8 + \underset{(1.14)}{0.94X_1} - \underset{(-0.52)}{0.04X_2}$$

$R^2 = 0.96$
F-Statistic = 92
Standard Error of Estimate = 6.8

These symptoms of multicollinearity emerge:

- Wrong Sign of a Coefficient
  The estimate of B_2 is negative while a positive value is expected. Normally, cost should increase with volume, not decrease.

- High Correlation Coefficient
  Pairwise correlation between X_1 and X_2 = 0.99.

- Mismatch of t's and F
  t's are low (*statistically insignificant by a wide margin*) but F and R^2 are high.

H_o: $\beta_1 = 0$; $\rho = 0.29$
H_o: $\beta_2 = 0$; $\rho = 0.62$
F-statistic $\rho < 0.00001$

- Model Instability
    - Big swing in coefficient estimates (*bouncing β's*) when an X is dropped from the equation. X_1 and X_2 alone are highly significant

Revised Equations	Original Estimates
$\hat{Y} = 24.6 + 0.51X_1$ (14.24)	$0.94X_1$ (1.14)
$\hat{Y} = 24.4 + 0.05X_2$ (13.29)	$-0.04X_2$ (-0.52)

-
    - Big swing in coefficient estimates (*bouncing β's*) when observations are deleted. That is, the sample of 10 observations is randomly halved. Regression equations are then estimated for each subsample, with these results:

Sample	OLS Estimates		
	β_0	β_1	β_2
Full; n = 10	24.80	0.94	-0.04
Random; n = 5	16.30	-0.45	0.10
Other half; n = 5	39.10	1.44	-0.10

The weight of evidence of these symptoms clearly suggests a problem of damaging multicollinearity in the payload equation.

1.3 CONSEQUENCES

OLS estimates of coefficient parameters (β's) remain BLU and consistent in the face of multicollinearity.[2] But that's cold comfort since variances and covariances increase, as Figure 4 illustrates.

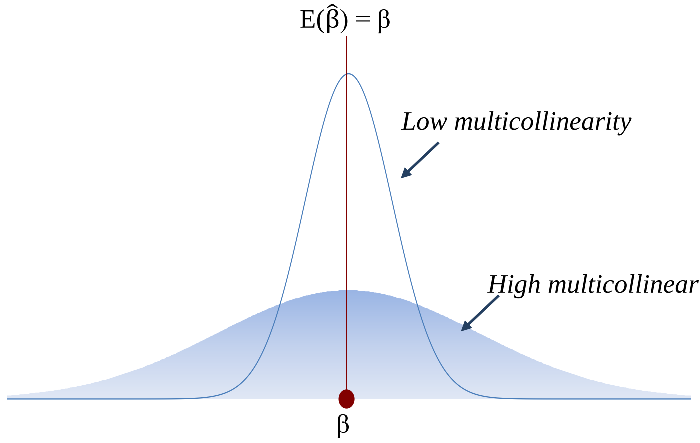


Figure 4

The increase in the sampling distribution variance, $\sigma^2_{\hat{\beta}}$, is directly related to degree of multicollinearity as shown in equation (5).

If 0, no impact (X's orthogonal)

$$(5)\ \sigma^2_{\hat{\beta}} = \frac{\sigma^2_{\varepsilon}}{\left(1 - R^2_{X_i\,|other\ X's}\right)(n-1)\sigma^2_{X_i}},\ where\ 0 \le \mathrm{R}^2{}_{\mathrm{X}_i}|\mathrm{other\ X's} \le 1$$

If 1, then **X'X** matrix is singular

As the denominator of the equation demonstrates, the variance of the estimate of the regression parameter will decrease as sample size, *n*, increases and as the variance in the explanatory variable increases, all else equal. This suggests that improved data collection might increase the precision of the estimates.

[2] BLU: Best, Linear Unbiased [estimator]
Consistency: sampling distribution collapses on β as sample size increases to infinity.

In any event, the increase in variance due to multicollinearity degrades precision of the estimates of the β’s as confidence intervals widen, as shown in equation (6) and Figure 5.

$$(6)\ \ \hat{\beta} - t * \sigma_{\hat{\beta}}\ \ \leq \beta\ \leq \hat{\beta} - t * \sigma_{\hat{\beta}}$$

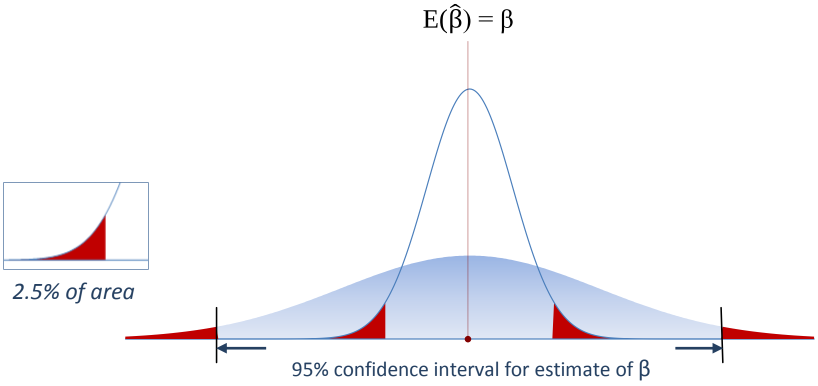


Figure 5

Additionally, and in summary, severe cases of multicollinearity

- Impart a bias toward incorrect model specification by
  - Making it difficult to disentangle the relative influences of the X's
  - Increasing the tendency for relevant X's to be discarded incorrectly from regression equations since the acceptance region for the null hypothesis is widened (*type II error*), as shown in equation (7):

    H_o: $\beta = 0$
    H_A: $\beta \neq 0$

    $$(7)\ \ -t * \sigma_{\hat{\beta}}\ \ \leq \hat{\beta}\ \leq +t * \sigma_{\hat{\beta}}$$

- Generate unstable models where small changes in the data produce big changes in parameter estimates [bouncing β's]

- Jeopardize accuracy of predictions, which require a perpetuation of a
  - Stable relationship between Y and the X's
  - Stable *interdependency* amongst the X's, which is often not the case

- Widen confidence and prediction intervals for Y, given a set of X’s.

1.4 HIDDEN EXTRAPOLATION

Besides the direct implications to the model and related statistics, multicollinearity leads to hidden extrapolation. Consider Figure 6, below. In this example, there are two moderately correlated predictor variables, X1 and X2. The range of X1 is (13, 49) and the range of X2 is (46, 187). A new point to be predicted (X1=17, X2=130), in red, falls well within these bounds. Intuitively, it does not seem that this point is extrapolating from the model. However, this point does in fact suffer from hidden extrapolation. The blue ellipse provides a conceptual view of the independent variable hull (IHV). This is the range of the data used to fit the model. Attempting to predict points outside of this range leads to extrapolation and all of the problems that come with it. The higher the correlation between X1 and X2, the narrower this ellipse becomes. The lower the correlation between X1 and X2, the wider this range becomes – almost to the point of covering the entire range of both X1 and X2!

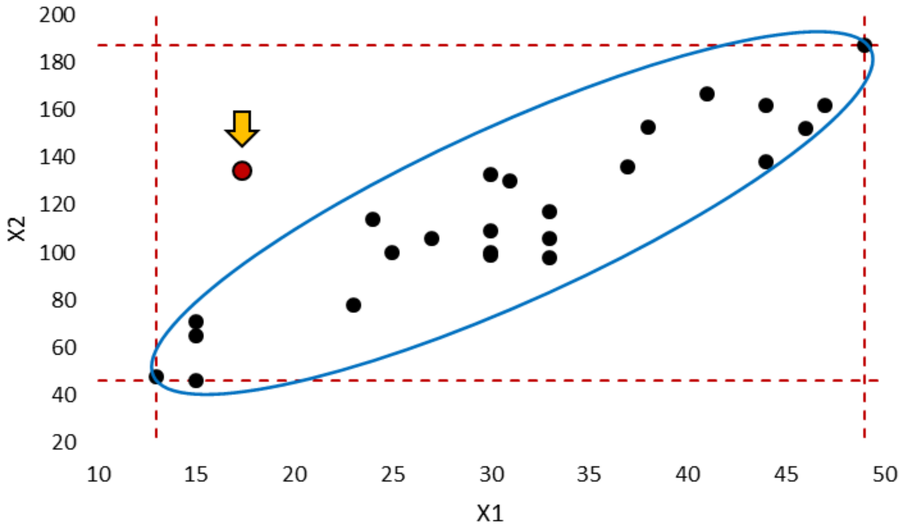


Figure 6

1.5 DIAGNOSIS

1.5.1 OVERVIEW

A review of the literature reveals different perspectives for diagnosing the problem of severe multicollinearity.

- Heuristic, or Rules of Thumb
  - Simple correlation coefficients: r_{X_i,X_j}
  - Mismatch between F and t statistics
  - Klein's r_{X_i,X_j} relative to over-all degree of multiple correlation R_y
  - R-squared values: $R^2 for\ X_i|all\ other\ X's$
  - Variance Inflation Factors
- Computational
  - Condition number of the matrix, based on eigenvalues
  - Determinant of the scaled $\mathbf{X'X}$ matrix

- Testable Hypotheses
  - Bartlett's $\chi 2$ for the presence of multicollinearity
  - Farrar-Glauber F-test and t-tests for the location and pattern of multicollinearity

Unfortunately, there's no unanimous view of a single best measure of when "severity" precisely kicks in. Various numerical rules of thumb are suggested, but they sometimes vary to an alarming and confusing degree. A good practice therefore seems to use several highly-regarded heuristics coupled with the formal statistical tests to obtain better insights into the sample at hand.[3] Use of the formal tests, however, does require an assumption that the X's are joint normally distributed. This is a departure from the usual assumption in the classical normal linear regression model that the X's of equation (1) are fixed in repeated samples.

Use of the heuristics and tests are demonstrated using the software sizing data of Table 2 for ten observations on Computer Software Configuration Items (CSCI)

CSCI	Y	X_1	X_2	X_3	X_4
1	6.0	40.1	5.5	108	63
2	6.0	46.3	4.7	94	72
3	6.5	47.5	5.2	108	86
4	7.1	49.2	6.8	100	100
5	7.2	52.3	7.3	99	107
6	7.6	58.0	8.7	99	111
7	8.0	61.3	10.2	101	114
8	9.0	62.5	10.1	97	116
9	9.0	64.7	17.1	93	119
10	9.3	61.8	21.3	102	121

Table 2

Y = Software development effort in
thousands of person months
X_1 = Function points in thousands
X_2 = Measure of number and complexity
of interfaces
X_3 = Productivity index (base =100)
X_4 = Environment index (base = 100)

The application of ordinary least squares yields the following estimated relationship between software development effort (Y) and the proffered explanatory variables (X's) of Table 2.

$$(8)\ \hat{Y} = \underset{(0.041)}{0.161} + \underset{(1.487)}{0.078X_1} + \underset{(2.157)}{0.079X_2} + \underset{(0.426)}{0.013X_3} + \underset{(0.606)}{0.011X_4} \quad (t - statistics)$$

$$R^2 = 0.95; F = 24$$

1.5.2 RULES OF THUMB

1.5.2.1 Value of a simple correlation coefficient r_{X_i,X_j}

Metric: Collinearity harmful if $\left|r_{X_i,X_j}\right| > 0.90$

- Measures pairwise interdependence only
- No pretense of theoretical validity

r_{X_i,X_j}
Simple correlation coefficient between X_i and X_j

[3] Use of eigenvalues will be covered in future updates to the Guide.

Calculations:

Simple Correlation Matrix of X's

	X_1	X_2	X_3	X_4
X_1	1.00	0.77	-0.52	0.94
X_2		1.00	-0.26	0.73
X_3			1.00	-0.41
X_4				1.00

Result: $r_{X_1,X_4} = 0.94$

1.5.2.2 *Comparison of t-statistics with F and R^2*

Metric: multicollinearity harmful if F significant ($\rho < 0.05$) but each coefficient insignificant ($\rho > 0.05$)

Calculations:

p-value for F-statistic

- Probability that the X's, evaluated together, have no influence on Y; i.e., $\alpha = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$; p-value = 0.002

p-values for t-statistics

- Probability that an individual X has no influence on Y; i.e., $\beta_i = 0$

$$(9)\ \hat{Y} = 0.161 + 0.078X_1 + 0.079X_2 + 0.013X_3 + 0.011X_4$$
$$(0.969)\quad (0.197)\quad (0.084)\quad (0.688)\quad (0.571)\quad (p - values)$$

Only X_2 comes close to threshold 5% level of significance

Result:

- F-statistic highly significant ($p < 1\%$)
- But each X insignificant (three of the p's > 50%!)

1.5.2.3 *Klein's r_{X_i,X_j} relative to over-all degree of multiple correlation R_y*

Metric: Multicollinearity harmful if any $r_{X_i,X_j} \geq R_y$, where
R_y = multiple correlation coefficient between Y and all the X's

Calculations:

$$R^2 = 0.95; R_y = \sqrt{R^2} = 0.97$$

Simple Correlation Matrix of X's

	X_1	X_2	X_3	X_4
X_1	1.00	0.77	-0.52	0.94
X_2		1.00	-0.26	0.73
X_3			1.00	-0.41
X_4				1.00

Result:

- In all cases, $r_{X_i,X_j} \leq R_y$

- $r_{X_1,X_4}\ closest\ to\ R_y$

1.5.2.4 R² values amongst the X's

Metric: Multicollinearity harmful if any $R^2 for\ X_i|all\ other\ X's > 0.90$

Calculation: *example*

$$(10)\ \hat{X}_1 = 48.2 + 0.298X_2 - 0.271X_3 + 0.301X_4\,;\ R^2 = 0.92$$

Result:

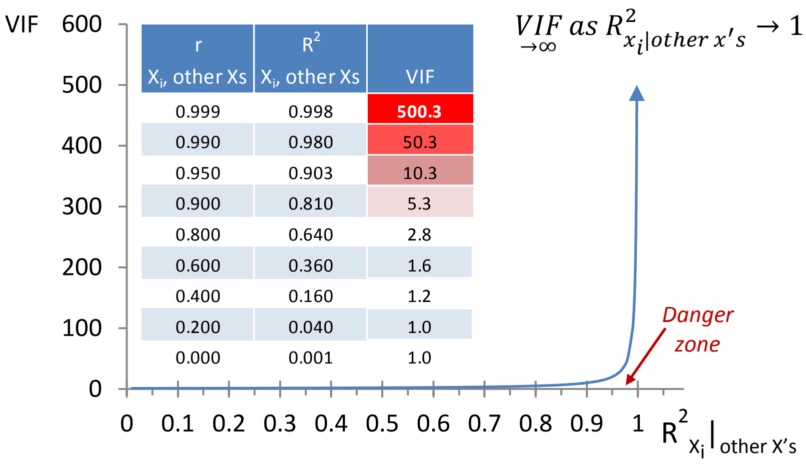
- $R^2 for\ X_1|X_2, X_3, X_4 = 0.92$
- $R^2 for\ X_2|X_1, X_3, X_4 = 0.61$
- $R^2 for\ X_3|X_1, X_2, X_4 = 0.35$
- $R^2 for\ X_4|X_1, X_2, X_3 = 0.89$

Harmful multicollinearity

Potentially Harmful

1.5.2.5 Variance Inflation Factors (VIFs)

Metric: No formal criteria for determining when a VIF harmful. Suggested threshold values cited in the literature include 4, 5, 10, 20, and 30. However, one recommendation is that VIF > 5 suggests a closer examination and VIF > 10 indicates the presence of likely harmful multicollinearity.[4]



r X_i, other Xs	R^2 X_i, other Xs	VIF
0.999	0.998	500.3
0.990	0.980	50.3
0.950	0.903	10.3
0.900	0.810	5.3
0.800	0.640	2.8
0.600	0.360	1.6
0.400	0.160	1.2
0.200	0.040	1.0
0.000	0.001	1.0

[4] CEBoK uses a value of 10. Module 8, "Regression Analysis," page 106.

Calculation:

Re-grouping terms in equation (5) gives

Variance Inflation Factor; = 1 when X's orthogonal

$$(11)\ \sigma_{\hat{\beta}}^2 = \left[\frac{1}{1 - R^2_{X_i\,|other\ X's}}\right]\frac{\sigma_\varepsilon^2}{(\mathrm{n}-1)\sigma_{X_i}^2}$$

$$= \left[\frac{1}{1 - 0.92}\right]\frac{\sigma_\varepsilon^2}{(\mathrm{n}-1)\sigma_{X_i}^2} for\ R^2_{X_1\,|other\ X's}$$

Potentially Harmful

Result:

- VIF $for\ X_1|X_2, X_3, X_4 = 12.5$
- VIF $for\ X_2|X_1, X_3, X_4 = 2.6$
- VIF $for\ X_3|X_1, X_2, X_4 = 1.5$
- VIF $for\ X_4|X_1, X_2, X_3 = 9.1$

1.5.3 COMPUTATIONAL

Many computational metrics rely on the use of the *centered and scaled* model. Remember, the traditional model has been defined as:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

The centered and scaled model simply transforms each predictor $\boldsymbol{X}_i$ into a corresponding $\boldsymbol{Z}_i$,

$$\boldsymbol{Z}_i = \frac{\boldsymbol{X}_i - \bar{X}_i}{\sqrt{\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2}}$$

And the new model becomes:

$$Y_i = \bar{Y}_i + \beta_1^* Z_{1i} + \cdots + \beta_k^* Z_{ki} + \varepsilon_i$$

Under this construct, the correlation matrix[5] is easy to calculate,

$$corr(\boldsymbol{X}) = \boldsymbol{Z}'\boldsymbol{Z}$$

1.5.3.1 Condition number of the matrix, based on eigenvalues

Metric: Collinearity harmful if $\psi > 30$

- Taken as the maximum condition index of corr(**X**)
- Based on the eigenvalue / eigenvector decomposition

[5] This rewrite of the model is also useful for many other reasons. For example, the inverse of the correlation matrix yields the variance inflation factors (VIFs) on the diagonals.

Calculations: *it is best to let software calculate at least the eigenvalues for you!*

$$corr(\boldsymbol{X}) = \boldsymbol{Z}^{'}\boldsymbol{Z}$$

$$\boldsymbol{\lambda} = vector\ of\ eigenvalues\ from\ the\ correlation\ matrix$$
$$= (2.884, 0.782, 0.286, 0.048)$$

$$\theta_i = \sqrt{\frac{\max(\lambda)}{\lambda_i}}$$
$$= (1.000, 1.920, 3.176, 7.751)$$

$$\psi = \max(\theta_i)$$

Result:

- Condition number is $\psi = 7.751 < 30$
- Does not suggest harmful multicollinearity

1.5.3.2 Determinant of the scaled X'X matrix

Metric: Collinearity harmful if det($\boldsymbol{Z'Z}$) is very small

- Solving the OLS model requires inverting the **X**'**X**, or $\boldsymbol{Z'Z}$ matrix
- A very small determinant makes computer algorithms unstable – the $\boldsymbol{X'X}$ matrix cannot be calculated accurately, or even at all
- Known as "singularity" problem – indicates an matrix that cannot be inverted

Calculations: *it is best to let software calculate this for you!*

$$\det(\boldsymbol{Z'Z}) = 0.03085$$

or (note that eigenvalues are rounded to 3 decimals)

$$\det(\boldsymbol{Z'Z}) = product\ of\ eigenvalues$$
$$= 2.884 \cdot 0.782 \cdot 3.176 \cdot 7.751$$
$$= 0.03085$$

Result: Determinant is small – but not "very small" in terms of computer precision – so does not suggest harmful multicollinearity.

1.5.4 TESTABLE HYPOTHESES

Formal statistical tests for the presence, location, and pattern of multicollinearity are available upon relaxation of the classical assumption of non-stochastic X's in a regression equation. Assuming now that the explanatory variables follow a joint-normal distribution, as in Figure 7, Bartlett's χ^2 test and the Farrar-Glauber t and F tests are useful complements to the rules of thumb in diagnosing problems of multicollinearity in data samples. The ellipses in the figure represent iso-probabilities, such as a constant one-sigma or two-sigma deviations from the joint mean. The greater the correlation between X_1 and X_2, the narrower the ellipses. On the other hand, the ellipses become concentric circles in the case of orthogonality.

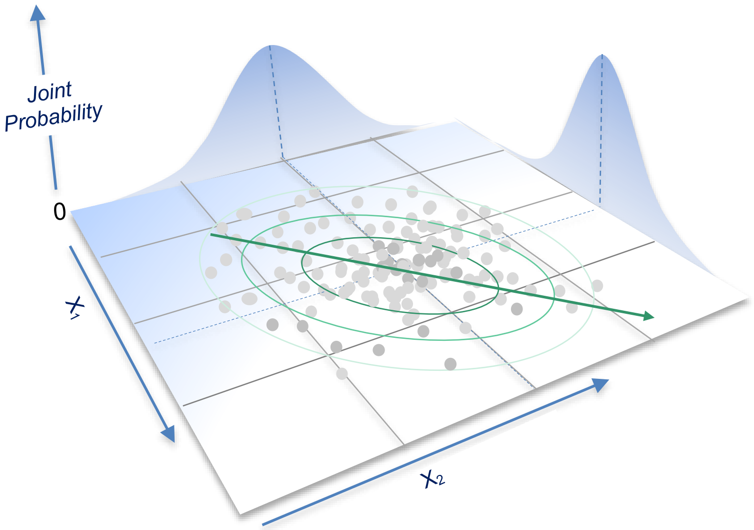


Figure 7

1.5.4.1 Test for the PRESENCE of multicollinearity

Bartlett's χ^2 test:
H$_0$: X's are orthogonal
H$_A$: X's not orthogonal

$|\boldsymbol{R}| = 1$ *Determinant of the simple correlation matrix of X's*

Test statistic: $\chi_v^2 = -\left[n - 1 - \frac{(2k+5)}{6}\right] ln|\boldsymbol{R}|, \; where$

n = sample size
k = number of explanatory variables
v = number of degrees of freedom = 0.5*k*(*k*-1)

Test criterion: $If \; observed \; \chi_v^2 > \chi_{critical}^2 \,, then \; reject \; H_0$

Test result: $\chi_{observerd}^2 = 23.8 > \chi_{critical=12.6 \; at \; 5\% \; level \; of \; sigficance}^2$

Test conclusion: X's harmfully interdependent, linearly

1.5.4.2 Test for the LOCATION of multicollinearity

Farrar-Glauber F test:

H$_0$: X_i is not affected by multicollinearity (i = 1, 2, …, k)
H$_A$: $X_i$is affected

i^th diagonal element of the inverse matrix of simple correlations

Test statistic: $F_v = \left[(r^{*i} - 1)\frac{(n-k)}{(k-1)}\right]$, $where$

v = number of degrees of freedom = $n - k$, k - 1

Test criterion: $If\ observed\ F > F_{critical=8.9\ at\ 5\%\ level\ of\ signficance}, then\ reject\ H_0$

Test result and conclusion: X_1 and X_4 are significantly affected by multicollinearity

F-Statistics

X_1	X_2	X_3	X_4
23.39	3.18	1.09	16.38
Affected	*Unaffected*	*Unaffected*	*Affected*

1.5.4.3 Test for the PATTERN of multicollinearity

Farrar-Glauber t test:

H$_0$: $r_{ij.g} = 0\ [partial\ correlation\ betweenX_i and\ X_j]$
H$_A$: $r_{ij.g} \neq 0$

g = set of all explanatory variables excluding X_i and X_j

Test statistic: $t_v = r_{ij.g}\frac{(n-k-1)^{0.5}}{\sqrt{\left(1-r_{ij.g}^2\right)}}$, $where$

v = number of degrees of freedom = n - k - 1

Test criterion:

$If\ t_{lower\,=\,-2.571} < t_{observed} < t_{upper\,=\,+2.571}, then\ do\ NOT\ reject\ H_0$

Test results:

Observed t-Statistics

	X_2	X_3	X_4
X_1	1.058	-1.177	*3.788*
X_2		0.564	-0.103
X_3			0.580

Test conclusion:

- Partial correlation between X_1 and X_4 significant; both terms, then, are seriously impacted by multicollinearity
- All other partial correlations are statistically insignificant, but with $r_{13.g}$ somewhat suspect.

1.5.5 DIAGNOSTIC CONCLUSION

Heuristic indicators and formal statistical tests result in a diagnosis of harmful multicollinearity in the sample of Table 2. Remedial action, therefore, seems in order to improve the precision of least-squares estimates.

1.6 REMEDIES

1.6.1 OVERVIEW

In order of preference, the following are recommended courses of action for treating the problem of multicollinearity in a regression equation.

- Collect additional observations

  An increase in sample size, as equation (11) suggests, will likely decrease the variance of estimates of the regression coefficients and could break the pattern of multicollinearity in the data. In defense cost analysis, however, data collection is difficult and often expensive. Nevertheless, the marginal benefit of adding even on or two more observations to a small sample may outweigh the marginal cost of collection.

- Re-specify variables

  In some cases, use of first differences or percent changes between successive observations in a time-series sequence may eliminate the multicollinearity problem. Combining two explanatory variables into one, such as weight and volume into density, may help, too.

- Select a subset of X's

  Given the practical problems of implementing the first two remedies, selecting a subset of X's for retention in the regression equation may be the next-best choice. As explained in the next section, Frisch's confluence analysis seems a robust methodology for the selection, based on a careful examination of the impact on the regression coefficients, t-statistics, sum of squared residuals, and R-bar squared.[6]

- Use prior information

  Prior information on the values of one or more coefficient parameters can be invaluable. To be useful, however, the information needs to be accurate. In practical applications in defense cost analysis, the pedigree of prior knowledge is usually somewhat suspect.

[6] Ragnar Frisch, Nobel Laureate in Economics, 1969. "Statistical Confluence Analysis by Means of Complete Regression Systems," Publication No. 5, University Institute of Economics, Oslo, 1934.

- Apply ridge regression

  Ridge regression introduces bias into the regression equation in exchange for a decrease in variance of the coefficient estimates. In its bare essence, ridge regression is a form of Bayesian statistical inference where choice of the ridge constant, *k*, requires prior knowledge about the unknown β's in the equation under consideration.

- Apply principal component regression

  Principal component regression performs a dimension reduction on the explanatory variables. The correlation matrix is decomposed and the X's are broken down into a smaller subset of principal components, Z's, which explain some percentage of the overall variance. When high multicollinearity is present, there is much redundant information that can be explained in fewer dimensions. In the extreme case of orthogonal X's (i.e., zero correlation), this dimension reduction results in significant loss of information.

1.6.2 CONFLUENCE ANALYSIS

In noting the natural tension in regression analysis between the urge to fully explain changes in Y by adding more explanatory variables to an equation and the inevitable need for at least some degree of parsimony, Ragnar Frisch coined the phrase "multiple collinear"

> "If the statistician does not … use an adequate technique for the statistical study of the confluence hierarchy, he will run the risk of adding more and more variates [variables] in the study until he gets a set that is in fact multiple collinear and where his attempt to determine a regression equation is therefore absurd."[7]

In Frisch's confluence analysis, the bugaboo of multicollinearity is overcome by following these steps

- Regress Y on that $X_j, j = 1, 2, \ldots, k$ which has the highest simple correlation with Y
- Gradually insert additional X's into the regression equation
- Examine effects on
  - Regression coefficients and their standard errors,
  - Standard error of the estimate, and
  - $R^2 or\ \overline{R}^2$
- Retain a variable in the regression equation only if it is "useful."

[7] Ibid.

Table 3 applies confluence analysis to the software sizing example. In the first round of Frisch's procedure, X_1 is selected as the "best" explanatory variable of the group.[8] In the second round, X_2 clearly performs better than the alternatives, either X_3 or X_4, according to all three measures of goodness-of-fit: $R^2, \overline{R}^2, and \widehat{\sigma}_{\varepsilon}$. In the third and final round, the inclusion of X_4 significantly damages the regression equation. It renders X_1 statistically insignificant and increases rather than decreases the standard error. Likewise, X_3 increases the standard error and also changes the statistical sign of the constant term.

In summary, then, Frisch's analysis yields these conclusions

- X_1 and X_2 appear to be the most appropriate explanatory variables to use;
- X_3 appears superfluous and possibly detrimental; and
- X_4 is detrimental.

[8] As shown on page 10, X_1 has the highest VIF. But, contrary to some guidance, rather than summarily dropping the variable from further consideration in the regression equation, it is retained in Frisch's procedure.

Frisch's Confluence Analysis

Equation	Parameter	Estimate	t-Statistic	R^2	$\overline{R}^2$	$\hat{\sigma}_\varepsilon$
Initial estimate						
Y = $f(X_1, X_2, X_3, X_4, \varepsilon)$	α	0.161	0.041	0.9499	0.9099	0.3704
	β_1	0.078	1.487			
	β_2	0.079	2.157			
	β_3	0.013	0.426			
	β_4	0.011	0.606			
1st round:						
Y = $f(X_1, \varepsilon)$	α	0.053	0.055	0.8877	0.8736	0.4386
	β_1	0.138	7.951			
2nd round:						
Y = $f(X_1, X_2, \varepsilon)$	α	1.493	1.627	0.9427	0.9263	0.3349
	β_1	0.097	4.682			
	β_2	0.083	2.592			
Y = $f(X_1, X_3, \varepsilon)$	α	-3.775	-0.922	0.9008	0.8724	0.4407
	β_1	0.148	7.272			
	β_3	0.033	0.962			
Y = $f(X_1, X_4, \varepsilon)$	α	0.374	0.335	0.8937	0.8633	0.4562
	β_1	0.107	2.014			
	β_4	0.014	0.629			
3rd round:						
Y = $f(X_1, X_2, X_3, \varepsilon)$	α	-0.657	-0.186	0.9463	0.9194	0.3503
	β_1	0.105	4.178			
	β_2	0.078	2.253			
	β_3	0.018	0.632			
Y = $f(X_1, X_2, X_4, \varepsilon)$	α	1.791	1.765	0.9481	0.9222	0.3442
	β_1	0.067	1.567			
	β_2	0.082	2.509			
	β_4	0.013	0.792			

$Max\ r_{Y,X_i}$

Best Choice

$\hat{\sigma}_\varepsilon$ *increases in both cases*

With X_4 included, X_1 now statistically insignificant

Table 3