The Collinearity Kill Chain:

Detect, Classify, Localize, Neutralize

ICEAA Conference

Atlanta, GA

• Adam JAMES • Brian FLYNN

June 2016

OUTLINE

1 Definition

2

3 Visualization

4 Symptoms

5 Consequences

6 Causation

7 Diagnosis

8 Remedies

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

DEFINITION

$$\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

$$\mu_i$$

At what point does multicollinearity become harmful?

OUTLINE

1 Definition

2

3 Visualization

4 Symptoms

5 Consequences

6 Causation

7 Diagnosis

8 Remedies

PRACTICALITIES

- Multicollinearity often occurs when different variables rise and fall together
  - Sample contains seemingly redundant information
  - Theory usually more fertile than data
  - Additional X's sometimes capture only nuances of the same, basic information
- 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 $Q_t - Q_{t-1}$
- Detection and correction of multicollinearity requires many tools and techniques

No easy solution

OUTLINE

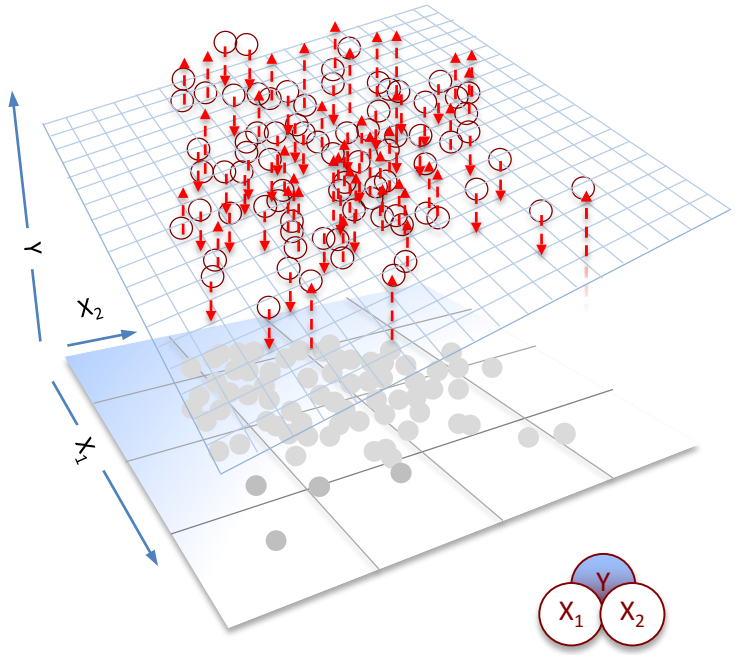
Technomics

1. Definition
2.
3. Visualization
4. Symptoms
5. Consequences
6. Causation
7. Diagnosis
8. Remedies

VISUALIZATION

VISUALIZING COLLINEARITY: CASE 1

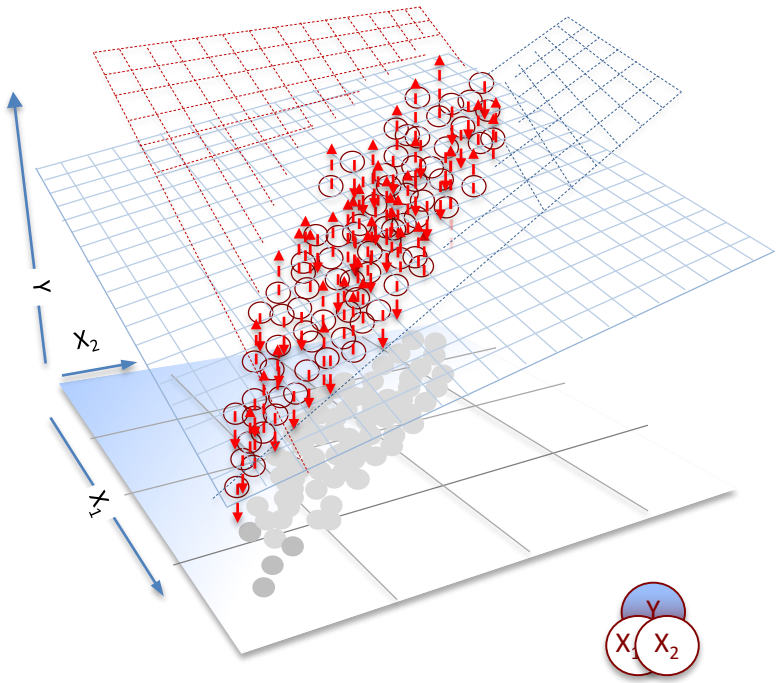
- Weak correlation between X_1 and X_2
- Regression plane is well-determined; only "slight" movement in repeated samples
- Stable OLS estimates of β_1 and β_2



VISUALIZATION

VISUALIZING COLLINEARITY: CASE 2

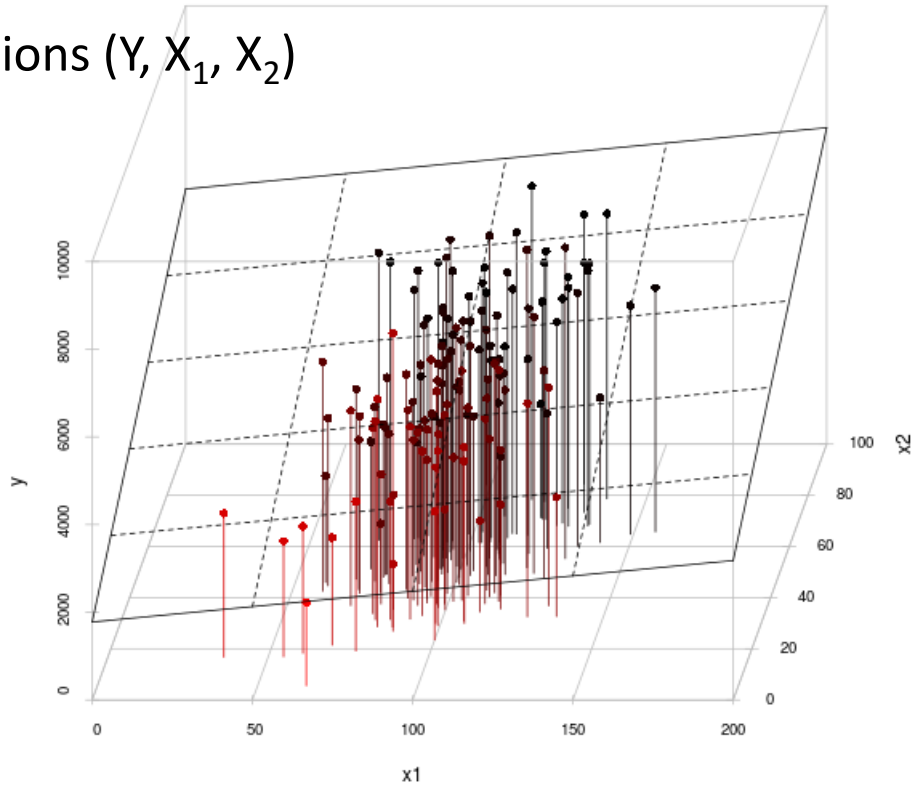
- Strong, inverse correlation between X_1 and X_2
- Observations Y_i, X_{1i}, X_{2i} in a tube
- Regression plane wobbles, with significantly different positions likely in repeated samples; but, with $E(Y_i)$ unaffected
- Unstable OLS estimates of β_1 and β_2



Simulation Example

Dashboard Demo

- Consider a model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ from some population with some normal random error
- Randomly sample a set of n observations (Y, X_1, X_2)
- Fit the linear OLS model (a 3D plane)
- Vary the amount of correlation between X_1 and X_2 and watch the plane "teeter"!



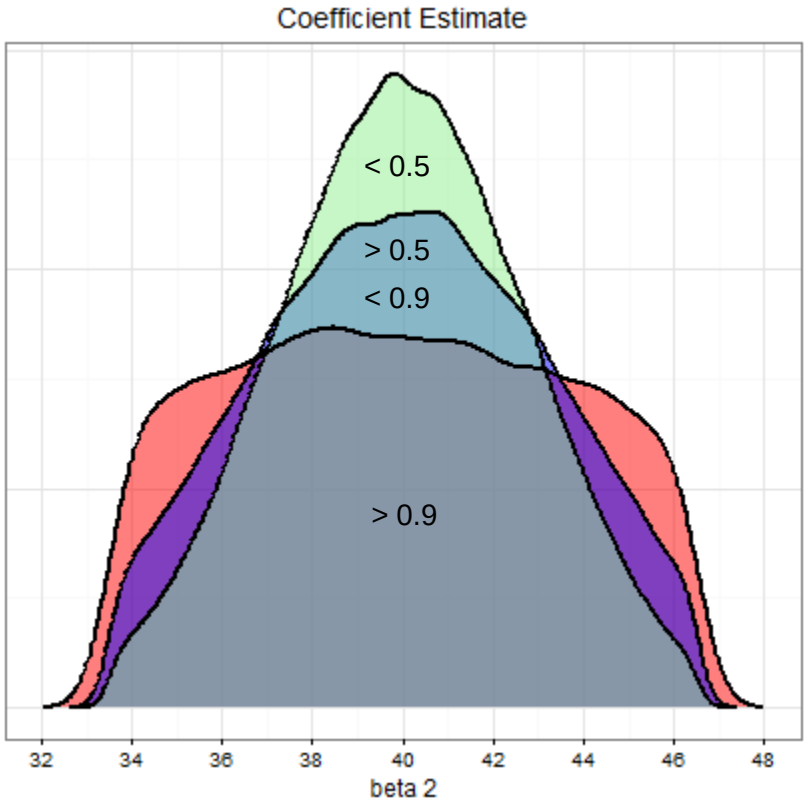
Simulation Example

- This simulation exercise was repeated 250,000 times – each with nothing changing except the degree of correlation between X_1 and X_2

- This plot separates the estimates for β_2 into three distributions:
  - $Corr(X_1, X_2) < 0.50$
  - $0.50 < Corr(X_1, X_2) < 0.90$
  - $Corr(X_1, X_2) > 0.90$

- The parameter estimate stays centered around 40, but with much higher variance as the correlation increases!

	Correlation	
Quantile	< 0.50	> 0.90
0%	27.0	-316.5
25%	38.1	33.6
50%	40.0	39.9
75%	41.9	46.5
100%	52.6	597.6



OUTLINE

Technomics

1 Definition

2

3 Visualization

4 Symptoms

5 Consequences

6 Causation

7 Diagnosis

8 Remedies

SYMPTOMS

Spacecraft Payload Example

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

- Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
- OLS Results:

$$\hat{Y} = 24.8 + 0.94X_1 - 0.04X_2$$

(1.14) (-0.52)

(t-statistics)

Insignificant t-statistics yet high R^2 and F

$R^2 = 0.96$

F-Statistic = 92

$\hat{\sigma}_\varepsilon = 6.8$

SYMPTOMS

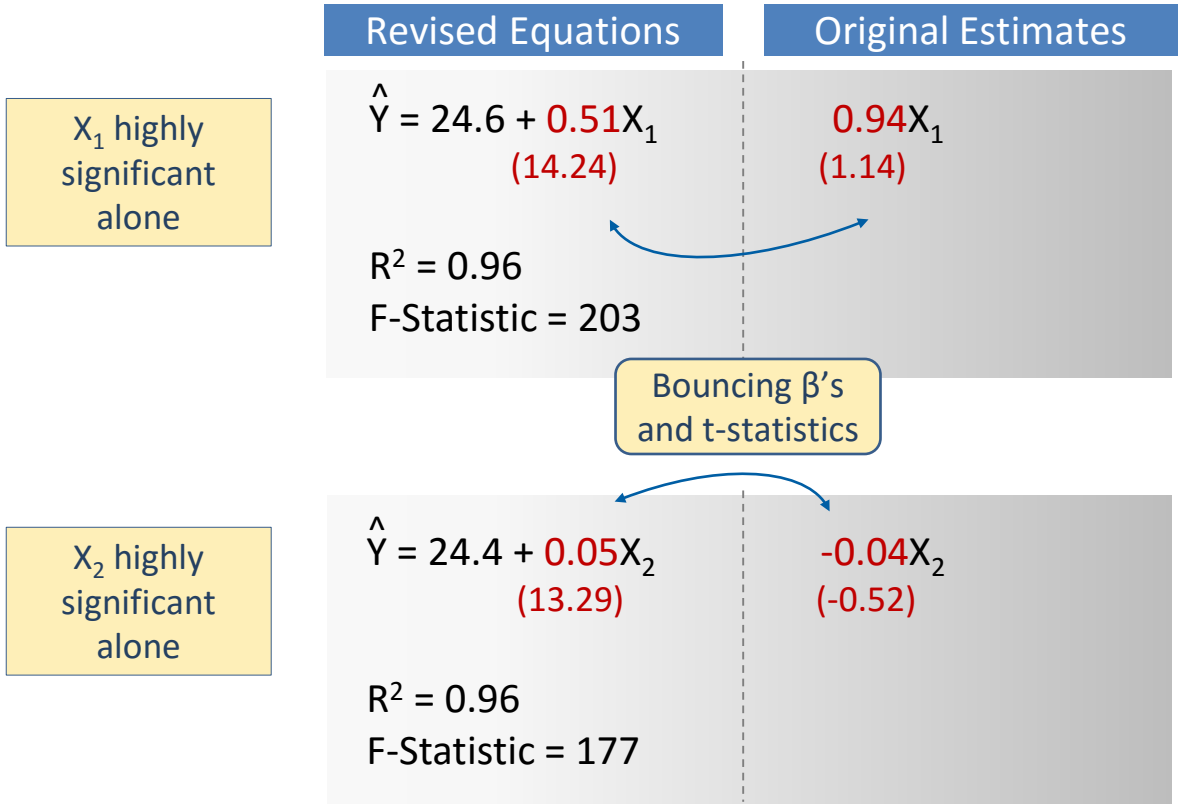
Spacecraft Payload Example

- The signs and values of estimated coefficients are inconsistent with domain of knowledge
  - Coefficient of X_2 of the wrong sign
- High pairwise correlations among X's
  - $r_{X_1,X_2} = 0.99$
- Mismatch of F and t tests
  - t's low but F and R^2 high
    - $H_o: \beta_1 = 0$; $\rho = 0.29$
    - $H_o: \beta_2 = 0$; $\rho = 0.62$
    - F-statistic $\rho < 0.00001$
- Model very unstable; significant swings in parameter estimates when:
  - One or more X's deleted
  - Observations added/deleted

SYMPTOMS

Spacecraft Payload Example

- Big swing in coefficient estimates when an X is dropped from the equation



SYMPTOMS

Spacecraft Payload Example

- Big swing in coefficient estimates when observations are deleted
  - Sample of 10 observations randomly halved
  - Regression equations estimated for each

Full sample:	$\hat{Y}_i =$	24.8	$+ 0.94X_1$	$- 0.04X_2$
Random selection → Observations 1, 4, 3, 2, 9:	$\hat{Y}_i =$	16.3	$- 0.45X_1$	$+ 0.10X_2$
Remaining Observations:	$\hat{Y}_i =$	39.1	$+ 1.44X_1$	$- 0.10X_2$

Bouncing β's

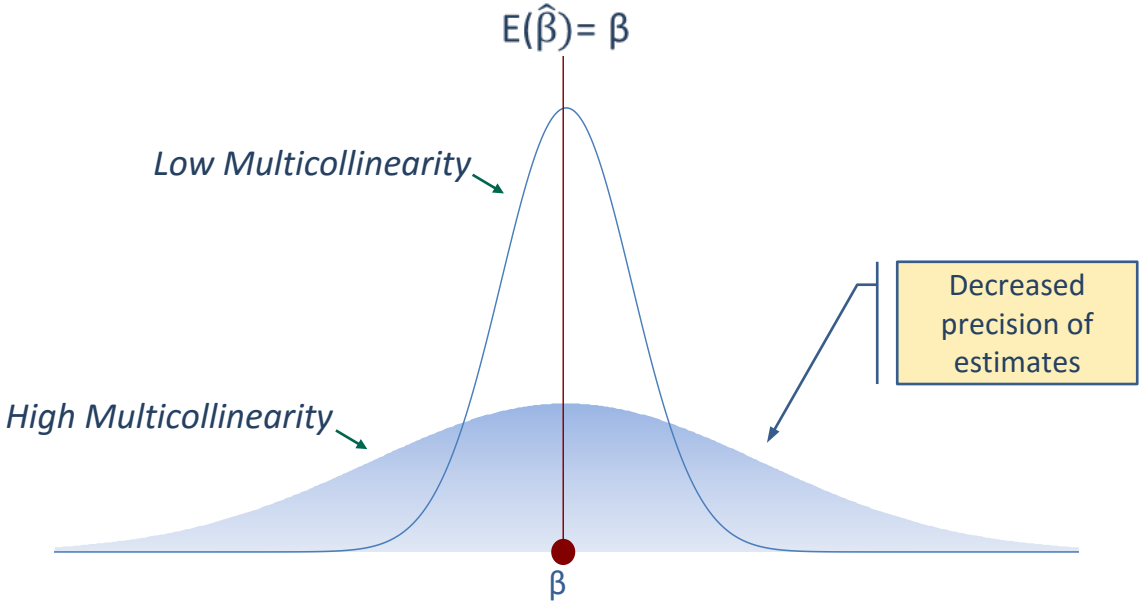
OUTLINE

Technomics

1. Definition
2.
3. Visualization
4. Symptoms
5. **Consequences**
6. Causation
7. Diagnosis
8. Remedies

CONSEQUENCES

- OLS estimates ($\hat{\beta}$'s) of coefficient parameters (β's) remain BLU and consistent in the face of multicollinearity[1]
- But cold comfort since variances and covariances increase!

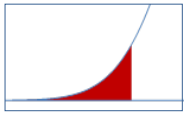


[1]BLU: Best, Linear Unbiased [estimator]
Consistent: sampling distribution collapses on β as sample size →∞

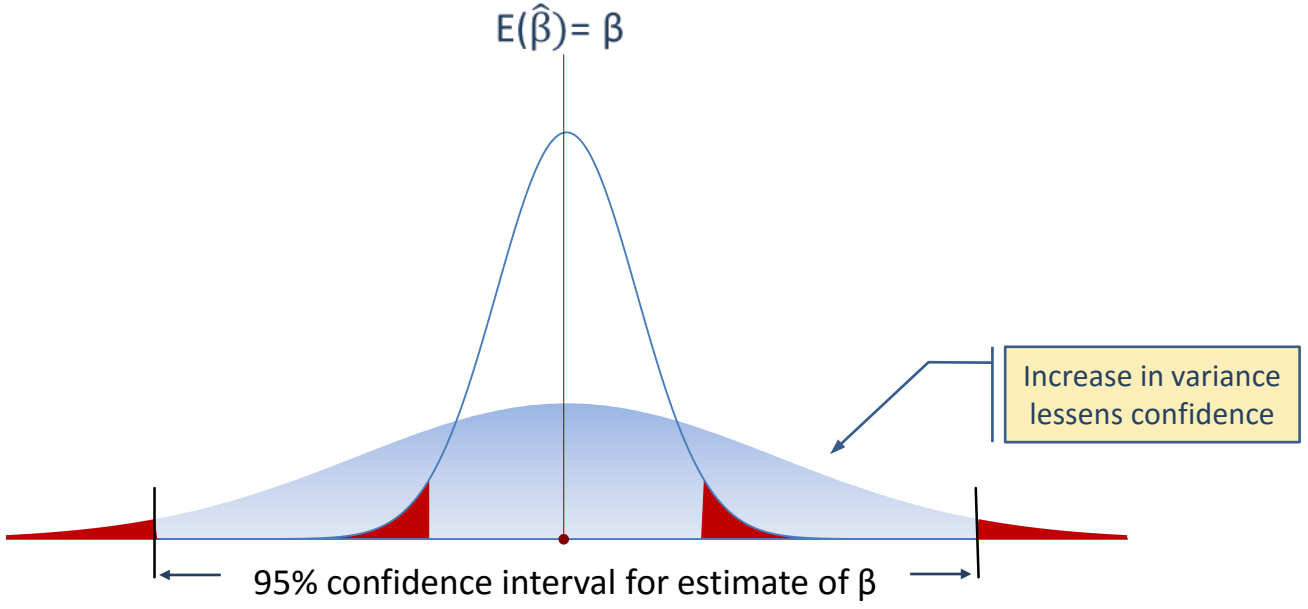
CONSEQUENCES

- Widens confidence intervals for estimates of the β's

$$(3)\ \hat{\beta} - t * \sigma_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t * \sigma_{\hat{\beta}}$$



2.5% of area



COLLATERAL DAMAGE

- Imparts a bias toward incorrect model specification
  - Makes it difficult to disentangle the relative influences of the X's
  - Increased tendency for relevant X's to be discarded incorrectly from regression equations:

H_o: $\beta=0$

H_A: $\beta\neq0$

$$(5)\quad -t*\sigma_{\hat{\beta}} \leq \hat{\beta} \leq +t*\sigma_{\hat{\beta}}$$

Acceptance region for H_o widens

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

- Jeopardizes accuracy of predictions, which requires perpetuation of:
  - Stable relationship between Y and the X's, and
  - Stable interdependency amongst the X's [often not the case]

- Widens confidence and prediction intervals for $Y_o|_{\text{set of x values}}$

OUTLINE

1 Definition

2

3 Visualization

4 Symptoms

5 Consequences

6 Causation

7 Diagnosis

8 Remedies

CAUSATION

- Increase in the sampling distribution variance, $\sigma^2_{\hat{\beta}}$, directly related to degree of multicollinearity

$$(4)\ \sigma^2_{\hat{\beta}} = \frac{\sigma^2_{\varepsilon}}{(1 - R^2_{X_i}|_{\text{other X's}})(n-1)\sigma^2_{X_i}}$$

If 0, no impact (X's orthogonal)

where, $0 \leq R^2_{X_i}|_{\text{other X's}} < 1$

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

- Interestingly, $\sigma^2_{\hat{\beta}}$ will decrease as
  - Sample size, n, increases
  - Variance of the explanatory variable, $\sigma^2_{X_i}$, increases

Improved collection of data might increase precision of estimates

OUTLINE

1 Definition

2

3 Visualization

4 Symptoms

5 Consequences

6 Causation

7 Diagnosis

8 Remedies

DIAGNOSIS OVERVIEW

PERSPECTIVES

- Computational
  1) Determinant of the simple correlation matrix of the X's
  2) Condition number of the matrix, based on eigenvalues

- Heuristic, or Rules of Thumb
  1) Simple correlation coefficients: r_{X_i,X_j}
  2) Mismatch between F and t statistics
  3) Klein's r_{X_i,X_j} relative to over-all degree of multiple correlation R_y
  4) R-squared values: $R^2(X_i \mid \text{all other X's})$
  5) Variance Inflation Factors

- Testable Hypotheses
  1) Bartlett's χ^2 for the presence of multicollinearity
  2) Farrar-Glauber F-test and t-tests for the location and pattern of multicollinearity

DIAGNOSIS

SOFTWARE SIZING DATA

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

Y = ESLOC (millions)
X_1 = Function points (thousands)
X_2 = Manmonths of effort
X_3 = Productivity index (base =100)
X_4 = Number of system interfaces

DIAGNOSIS

RULE OF THUMB: r_{X_i, X_j}

Software Sizing Example

- Model

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

$R^2 = 0.95$

$F = 24$

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

- Metric: Collinearity harmful if $|r_{X_i, X_j}| > 0.90$
  - Measures pairwise interdependence only
  - No pretense of theoretical validity

Simple correlation coefficient between X_1 and X_2

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

DIAGNOSIS

RULE OF THUMB: F & t MISMATCH

Software Sizing Example

- Low t's but high F and R^2

- Metric: multicollinearity harmful if F significant ($\rho < 0.05$) but each coefficient insignificant ($\rho > 0.05$)
  - 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
  - t-values
    - Probability that an individual X has no influence on Y; i.e., $\beta_i = 0$

$$Y = \underset{(0.969)}{\alpha} + \underset{(0.197)}{\beta_1 X_1} + \underset{(0.084)}{\beta_2 X_2} + \underset{(0.688)}{\beta_3 X_3} + \underset{(0.571)}{\beta_4 X_4} + \varepsilon \quad \text{(p-values)}$$

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

Only X_2 comes close to 5% level

DIAGNOSIS

RULE OF THUMB: KLEIN'S

Software Sizing Example

- Metric: collinearity harmful if any $r_{ij} \geq$ [or $\approx$] R_Y,
  where
  r_{ij} = simple correlation between X_i and X_j
  R_Y = multiple correlation coefficient between Y and all the X's

- Calculations:
  $R^2_{Y|X_1, X_2, X_3, X_4} = 0.95$
  $R_{Y|} = \sqrt{R^2} = 0.97$

- Result:
  - In all cases, $r_{ij} <= R_Y$
  - r_{X_1,X_4} closest to R_Y

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

Source: Lawrence Klein, Nobel Laureate in Economics, 1980. "Intercorrelation ... is not necessarily a problem unless it is high relative to the over-all degree of multiple correlation …"; *An Introduction to Econometrics*, Prentice-Hall, 1962, p.101.

RULE OF THUMB: R^2 VALUES

Software Sizing Example

- Metric: Multicollinearity $R^2(X_i | \text{all other X's}) > 0.90$
- Sample calculation:
  - $X_1 = 48.2 + 0.298X_2 - 0.271X_3 + 0.301X_4$
    (2.012) (1.159) (-1.290) (4.149) (t-statistics)

    $R^2 = 0.92$
    $F = 23$
- Results:
  - $R^2_{X_1} | X_2, X_3, X_4 = 0.92$ ← Harmful multicollinearity
  - $R^2_{X_2} | X_1, X_3, X_4 = 0.61$
  - $R^2_{X_3} | X_1, X_2, X_4 = 0.35$
  - $R^2_{X_4} | X_1, X_2, X_3 = 0.89$ ← Potentially Harmful

DIAGNOSIS

RULE OF THUMB: VIFS

Software Sizing Example

- Previous rule translatable into Variance Inflation Factors
  - Re-grouping terms from equation (4) [slide 21]

$$(4)\ \sigma^2_{\hat{\beta}} = \left[\frac{1}{1 - R^2_{X_i}|_{\text{other X's}}}\right]\frac{\sigma^2_{\varepsilon}}{(n-1)\sigma^2_{X_i}}$$

Variance Inflation Factor

VIF = 1 when X's orthogonal

- Results:
  - VIF for $R^2_{X_1}|_{X_2, X_3, X_4} = 12.5$
  - VIF for $R^2_{X_2}|_{X_1, X_3, X_4} = 2.6$
  - VIF for $R^2_{X_3}|_{X_1, X_2, X_4} = 1.5$
  - VIF for $R^2_{X_4}|_{X_1, X_2, X_3} = 9.1$

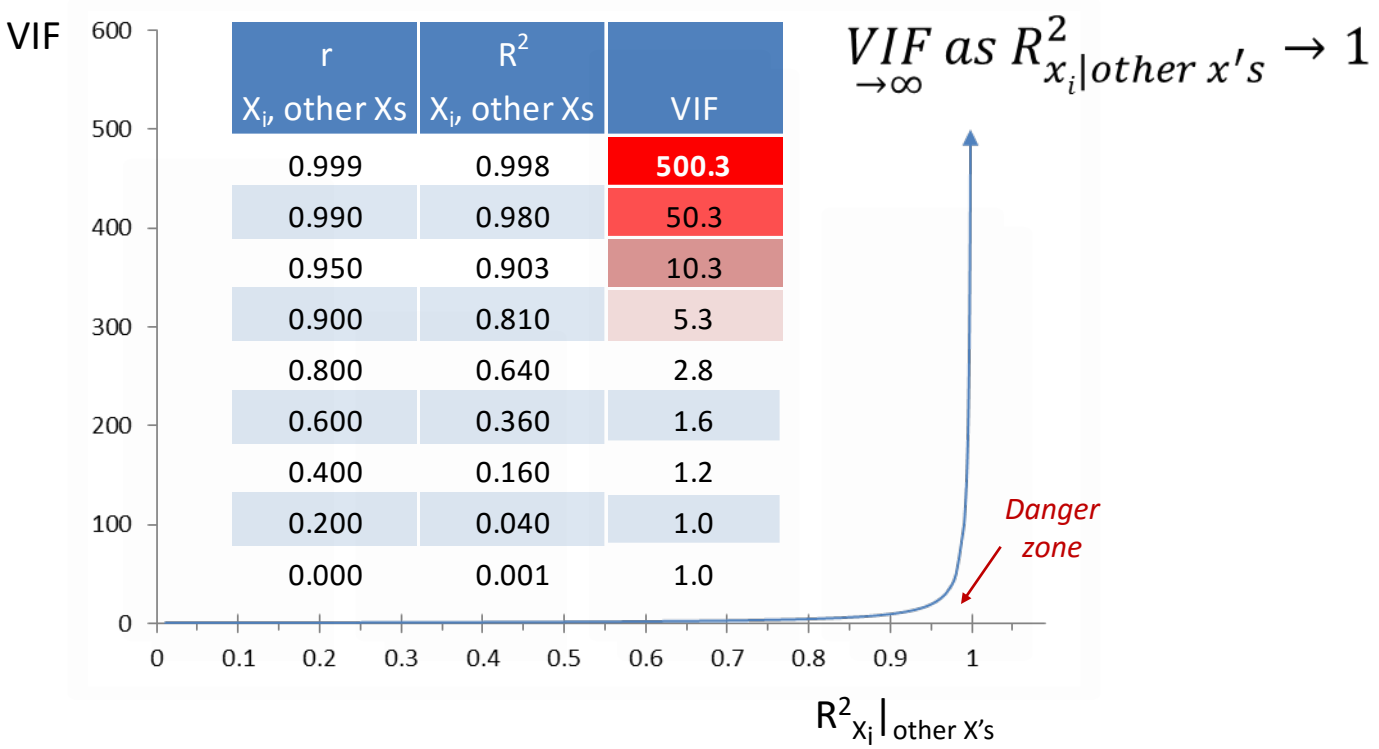
DIAGNOSIS

RULE OF THUMB: VIFS

Software Sizing Example

- Metric:
  - No formal criteria for determining when when a VIF harmful
  - Suggested threshold values from lit review: 4, 5, 10, 20, and 30

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



FORMAL TESTS

Technomics

Presence of Multicollinearity Using Software Sizing Example

Assumption: X's are joint normally distributed

- Bartlett's χ^2 test:

  H_0: X's are orthogonal

  H_A: X's not orthogonal

  |R| = 1 — *determinant of simple correlation matrix of X's*

- Test statistic: $\chi_v^2 = -\left[n - 1 - \frac{(2k+5)}{6}\right] ln|\boldsymbol{R}|, where$ $v = degrees\ of\ freedom = 0.5k(k-1)$

  n = sample size; *k = number of explanatory variables*

- Test criterion:

  $If\ observed\ \chi_v^2 > \chi_{critical}^2$, reject H_0

  ρ = 0.0006; heuristic measure of severity

- Result:
  - 23.8 (sample value) > 12.6 (critical value at 5% level of significance)

- Conclusion: X's linearly interdependent

FORMAL TESTS

Technomics

Location of Multicollinearity Using Software Sizing Example

Assumption: X's are joint normally distributed

- Farrar-Glauber F-test:

  H_0: X_i is not affected by multicollinearity ($i = 1, 2, \ldots, k$)

  H_A: X_i is affected

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

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

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

- Test criterion:

  $If\ observed\ F_v > F_{critical=8.9\ at\ 5\%\ level\ of\ signficance}$, reject H_0

- Results:

F-Statistics

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

FORMAL TESTS

Technomics

Pattern of Multicollinearity Using Software Sizing Example

Assumption: X's are joint normally distributed

- **Farrar-Glauber t-test:**

  $H_0: r_{ij.g} = 0$ [*partial correlation between* X_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:** $t_v = r_{ij.g} \frac{(n-k-1)^{0.5}}{\sqrt{(1-r_{ij.g}^2)}}$, *where* v = *degrees of freedom* $= n-k-1$

- **Test criterion:**

  If $t_{lower} < t_{observed} < t_{upper}$, *then do NOT reject* H_0 *(two tail test)*

  -2.571 $+2.571$

- **Results:**
  - Partial correlation between X_1 and X_4 significant
  - All others insignificant, with $r_{X_1 X_3 . X_4}$ less so than the rest

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

OUTLINE

Technomics

1. Definition
2.
3. Visualization
4. Symptoms
5. Consequences
6. Causation
7. Diagnosis
8. **Remedies**

REMEDIES

- Additional primary data collection [first choice]
  - Increase in *n* will decrease variance & may break pattern of collinearity
- Variable respecification; e.g.,
  - Using first differences or percentages
  - Replacing two variables with one

E.g., Density = Weight/Volume

- Variable selection
  - Suitable subset of important X's
  - Minimum sum of squares; or max $\overline{R}^2$

Frisch's confluence analysis

- Use of prior information about regression coefficients

Bayesian techniques

- Biased estimation [last resort]
  - Ridge regression
  - Ridge constant, *k*, requires knowledge about the unknown β's that we're trying to estimate!

VARIABLE SELECTION

"If the statistician does not dispose of [use] an adequate technique for the statistical study of the confluence hierarchy, he will run the risk of adding more and more variates 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"

Frisch's Confluence Analysis

- Regress Y on the X_i which has the highest simple correlation with Y
- Gradually insert additional X's into the regression equation
- Examine effects on
  - Regression coefficients
  - Standard errors
  - R^2, or better yet, $\bar{R}^2$
- Retain variable in the regression equation only if it's "useful"

Source: 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.

VARIABLE SELECTION

Software Sizing Example

$Max\ r_{Y,Xi}$

Equation	Parameter	Estimate	t-Statistic	R^2	$\bar{R}^2$	$\hat{\sigma}_\varepsilon$
$Y = f(X_1, \varepsilon)$	α	0.053	0.055	0.8877	0.8736	0.4386
	β_1	0.138	7.951			
2nd round alternatives:						
$Y = f(X_1, X_2, \varepsilon)$	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			

Best Choice

VARIABLE SELECTION

Technomics

Software Sizing Example

Equation	Parameter	Estimate	t-Statistic	R^2	$\bar{R}^2$	$\hat{\sigma}_\varepsilon$
3rd round alternatives:						
$Y = f(X_1, X_2, X_3, \varepsilon)$	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			

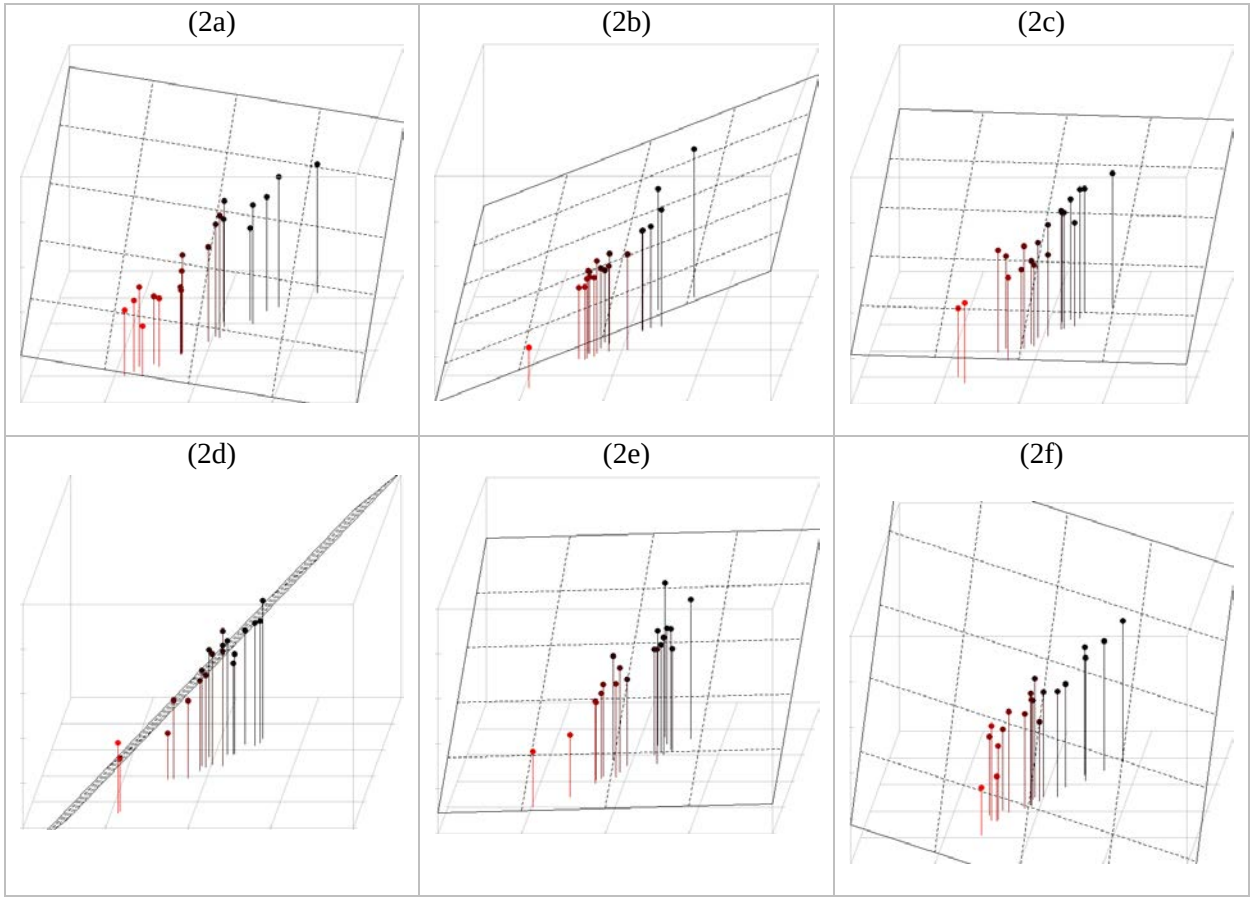
With X_4 included, X_1 now statistically insignificant

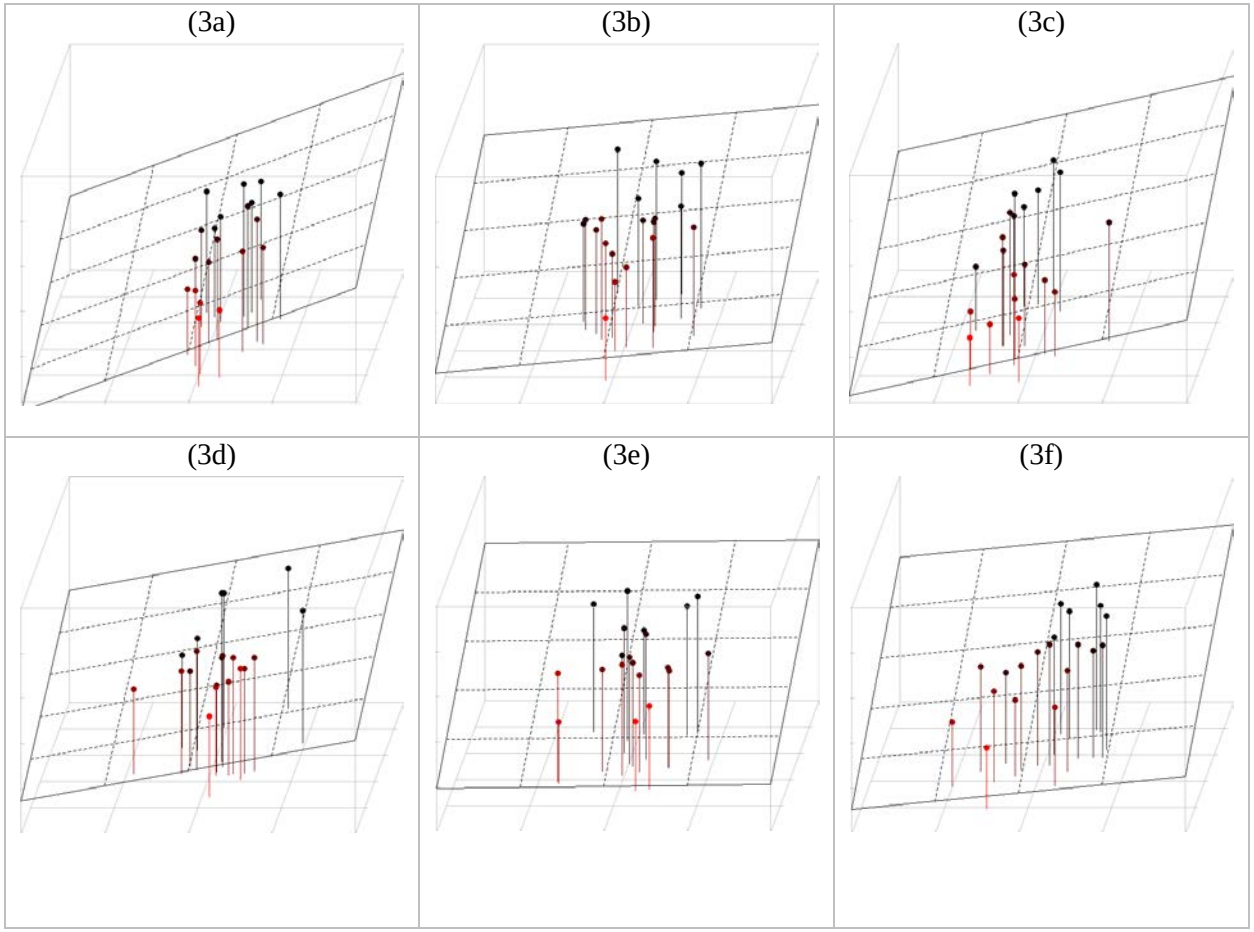
- Conclusions:
  - X_1 and X_2 appear to be the most appropriate explanatory variables to use
  - X_3 superfluous, and possibly detrimental [changes the sign of α]
  - X_4 detrimental

BACKUP

Simulation Example

95% Correlation



Simulation Example

20% Correlation