Multicollinearity in Zero Intercept Regression: They Are *Not* Who We Thought They Were

Kevin Cincotta

Presented at the Society of Cost Estimating and Analysis (SCEA) Conference

June 7-10, 2011

Albuquerque, NM

Technomics

The Science of Informed Decision Making

Acknowledgments

- Dr. Shu-Ping Hu (Tecolote), Peter Braxton (Technomics), and Andrew Busick (Technomics) for critical feedback
- Dr. David Lee (LMI) and John Wallace (AFCAA), for inspiring the research
- Former NFL Coach Dennis Green, for inspiring the title

Outline

- Background
- Summary of Earlier Findings
- "I didn't really say everything I said"
- A Better Approach
- "They are *not* who we thought they were"
- Consequences of Misspecifying the VIF
- Conclusions
- Ideas for Further Research

Background: The Importance of Individual Coefficients in CER Development

- Cost estimating relationships (CERs) are often derived using parametric approaches, including regression analysis
- It is often desirable to know the impact of one particular variable (in isolation) on cost…
  - Significance statistics: Is the variable affecting cost merely by chance?
  - Sensitivity analysis: What is the effect on cost if the variable's value is increased by 10? Doubled?
  - Engineering tradeoff analysis: What are the incremental cost implications of technically feasible design trades?

Background: Multicollinearity in Cost Estimation

- …But multicollinearity confounds the issue by making it difficult or impossible for models to separate the effects of two or more variables that tend to move together, but each of which may be reasonably argued to drive cost. Examples:
  - Length, Weight, and Crew of a ship
  - Average Power, Peak Power, aperture, and number of T/R modules of a radar
  - Number of flying hours, number of sorties, and number of landings for an aircraft
- A variety of approaches have been suggested to deal with the issue [1], but it's not uncommon for the analyst to remove one of the "offending" variables from the regression, reasoning that its effects are captured by the remaining variables

[1] Including Ridge Regression (which comes at the cost of biasing the estimate), Lemonade Methods (which are not always possible) and ignoring the problem.

Background: Statistical Consequences of Multicollinearity

- Inflates variances (and therefore standard errors) of coefficients
- Biases the estimated coefficients themselves (often manifested with very large positive and very large [in absolute value] negative numbers)
- Biases significance tests, which depend upon the coefficient's estimated value and standard error
- Does *not* bias forecasts of cost, in general, because the model remains one of "best fit" and overestimated/underestimated coefficients "offset" if all are left in model

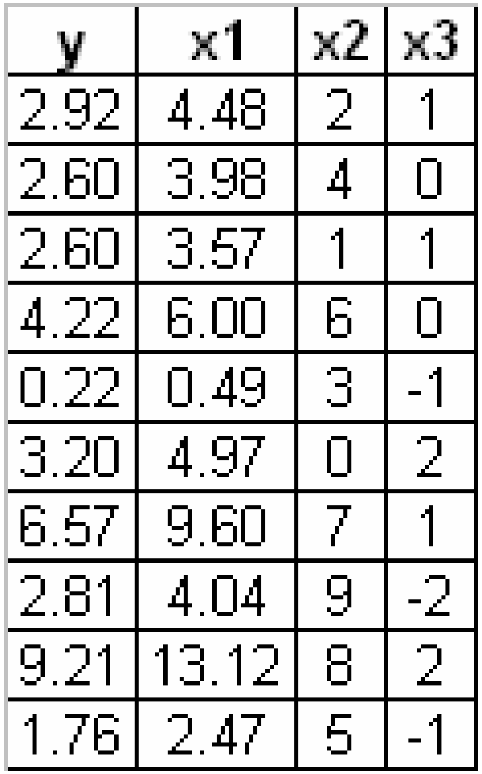
Background: Traditional Definitions of Multicollinearity

- "The situation in which two or more predictors are strongly correlated to one another…" (*Nature Reviews: Genetics*)
- "The presence of high correlations between predictor variables in a multiple regression." (Abrami et. al. *Statistical Analysis for the Social Sciences)*
- "A case of multiple regression in which the predictor variables themselves are highly correlated." (wordnet.princeton.edu)
- "In multivariate analyses, some of the independent variables may be correlated with each other. This condition is referred to as multicollinearity." (decisionanalyst.com)
- "Linear inter-correlation among variables." (*Wikipedia* 2007) updated to "a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated" (Wikipedia 2011)
- "Avoid high correlation between x variables. This is called multicollinearity and can be checked with a correlation matrix." (*CostProf*, v2, Module 8) updated to "Avoid multicollinearity, i.e., high correlation between X variables" (CEBoK, v1.1, Module 8 (63)) and "Multicollinearity occurs when there is a strong linear relationship between two or more dependent variables" (94) with citation to our prior research! [1]

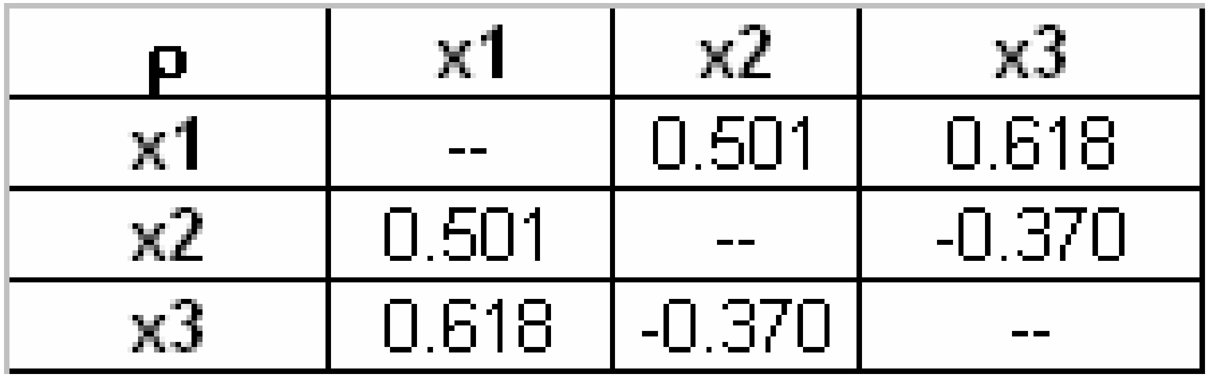
As we've said before [1], **multicollinearity is not the same as correlation**, nor even linear relationship among variables! It is inflation of the variance around an estimated coefficient due to a relationship among independent variables that is the same as the one being hypothesized in the overall model, and is revealed through variance inflation factors (VIFs).[2]

1. Cincotta, Kevin and Lee, Dr. David. *Multicollinearity: Coping With the Persistent Beast* (2007 DoDCAS).
2. As noted by Dr. Shu Ping-Hu, VIFs are an absolute measure, but ill-conditioning of the X'X matrix can occur even when no VIF is particularly high. A more suitable relative measure is the ratio of the R^2's from regressions of each X on the other X's to the overall model R^2. However, as we will see, the entire concept of R^2 must be revisited in the case of zeno-intercept regression.

Background: Correlation Neither Necessary…



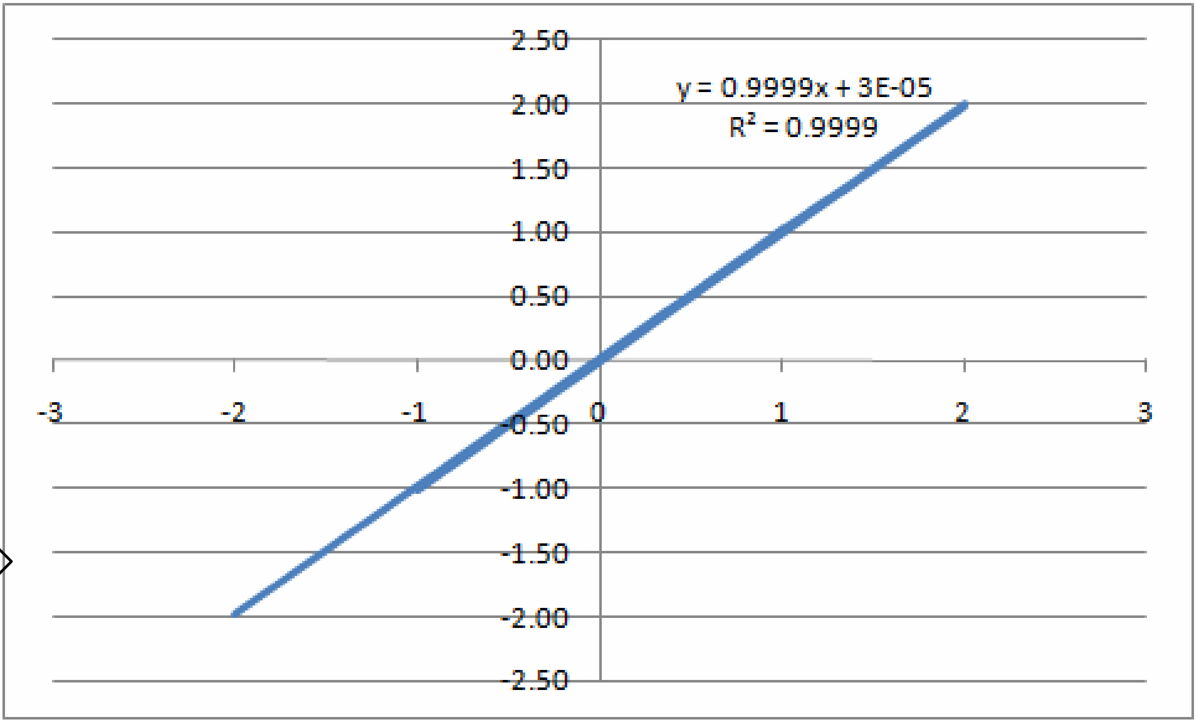
y	x1	x2	x3
2.92	4.48	2	1
2.60	3.98	4	0
2.60	3.57	1	1
4.22	6.00	6	0
0.22	0.49	3	-1
3.20	4.97	0	2
6.57	9.60	7	1
2.81	4.04	9	-2
9.21	13.12	8	2
1.76	2.47	5	-1



ρ	x1	x2	x3
x1	--	0.501	0.618
x2	0.501	--	-0.370
x3	0.618	-0.370	--

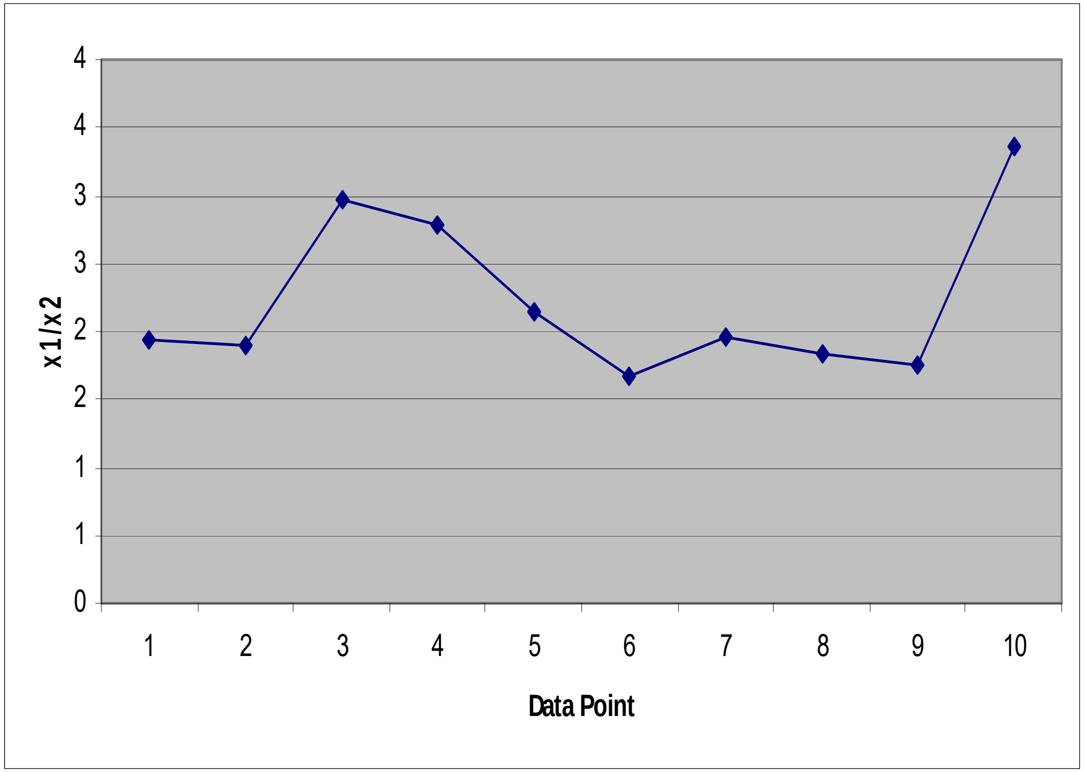
(secretly, $x_3 = 0.4x_1 - 0.4x_2 + \varepsilon$)

Plot of x3 vs. predictions of x3 based on regression on x1 and x2:



Background: …Nor Sufficient for Multicollinearity to Occur

x1	x2	x3	y
10.19	5.25	5.19	10.05
10.78	5.69	2.84	8.08
7.43	2.51	1.83	5.33
7.79	2.80	3.63	7.19
9.31	4.34	9.19	13.79
12.93	7.78	-0.34	6.66
9.97	5.11	-1.38	3.55
10.81	5.90	1.84	7.37
11.61	6.63	1.28	7.54
6.97	2.07	0.51	3.42



Note: $\rho\ (x_1,x_2) = 0.999$

If regression is performed with zero intercept, then multicollinearity *requires* that there exist constants c (not all zero) such that $c_1x_1 + c_2x_2 = 0$, i.e. ratio x_1/x_2 must be approximately constant [1]! Absence of multicollinearity in this example is confirmed by low VIFs [2].

1. Judge, George. *The Theory and Practice of Econometrics.* John Wiley and Sons: New York, NY (1980), pp. 455-505
2. Cincotta and Lee (2007)

Background: Variance Inflation Factors

- In multiple regression, **the variance inflation factor (VIF)** is the multiplicative factor by which the variance around and estimated coefficient on an independent variable is increased due to that variable's relationship [1] with other independent variables in the model [2]

- *True* multicollinearity is revealed through VIFs, where thresholds of 4, 5, and 10 [3] have been proposed as indicating a problem in the model

- You can never *reduce* variance through relationships among independent variables, so VIF >=1

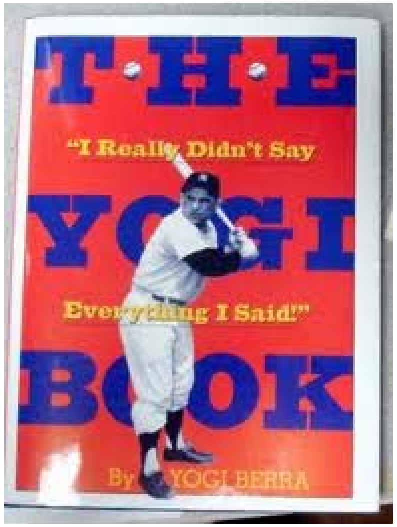
1. The relationship must be the same as the relationship being hypothesized in the general model!
2. Adapted from CEBoK, v 1.1, Module 8 (95)
3. CEBoK, v1.1, Module 8 (95), Wikipedia, and Kutner, Nachtsheim, Neter. *Applied Linear Regression Models*, 4th edition, McGraw-Hill Irwin, 2004.

Summary of Earlier Findings

- Zero-Intercept Regression (ZIR)
  - Correlation is necessary, but not sufficient, for multicollinearity
  - Multicollinearity requires *proportionality* among regressors, which is a stronger condition than the linear relationship measured by correlation
  - Results verified by VIF analysis
- Traditional OLS Regression
  - Correlation is sufficient, but not necessary, for multicollinearity
  - Extreme multicollinearity was shown when no two variables were (pairwise) highly correlated
  - Results verified by VIF analysis
- Multicollinearity *not an intrinsic property of data set*; it's relative to model form hypothesized
- How to Calculate VIFs
  - Same formula in both (ZIR and OLS) cases
  - Use "shortcut" of SE_{β}^2/MSE
  - Failed to note that this quotient is only an approximation, and is often less than 1!

Legend
Blue = We stand by these conclusions and reiterate this guidance
Red = I said *what?*

"I Didn't Really Say Everything I Said" [1]



- I "said:" VIF = SE_{β}^{2}/MSE
- The VIF statistic can be expressed as the variance around a coefficient in a regression, divided by its native [2] variance
- The square of the standard error associated with an estimated coefficient β (SE_{β}^{2}) *is* the variance around an estimated coefficient β
- However, the mean squared error of the regression (MSE), while sometimes a good proxy, is *not* the native variance about β
- You've waited 4.5 years for something better. Luckily, I'm still here.

[1] Berra, Yogi. *The Yogi Book: I Didn't Really Say Everything I Said*. LTD Enterprises (1998)

[2] The variance that the coefficient would have had, in the absence of any (same-form) relationship among regressors

How to Calculate VIFs: A Better Approach

- Using the commonly accepted formula for VIF is more cumbersome, but gives more accurate results **in traditional OLS regression** (that is, with a non-fixed intercept) [1]

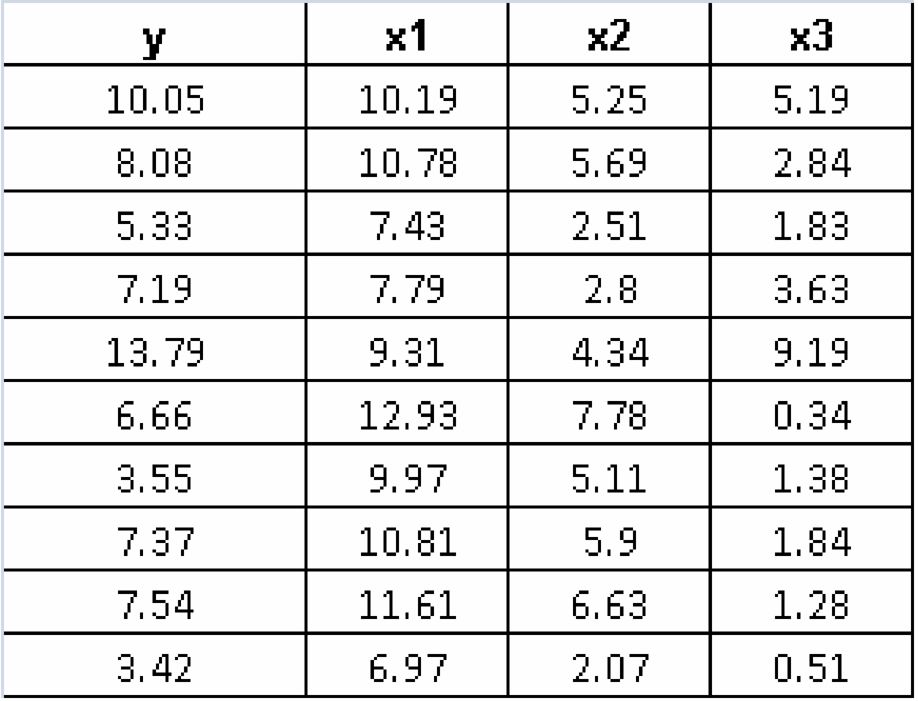
$$VIF_{\beta} = \frac{1}{(1 - R_{U-\beta}^{\ \ 2})}$$

- The VIF of an estimated coefficient β on a variable Xi is the reciprocal of the complement of the coefficient of determination obtained when Xi is regressed (with non-fixed intercept) on each of the other X's.
- Important Properties:
  - Minimum of 1 (when $R_{U-\beta}^{2} = 0$)
  - No maximum
  - Implicitly captures all relevant relationships among independent variables (not just pairwise…could be 10-way)

[1] Unfortunately, this formula is used more or less universally, which (as we will see) can have disastrous consequences in zero intercept regression.

Implied VIF Thresholds and Example

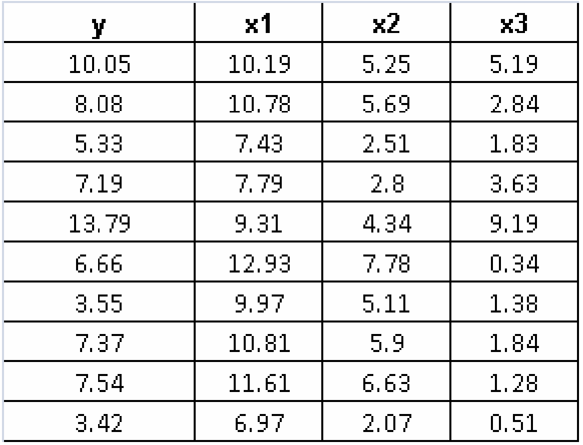
VIF Threshold	Implied Maximum R^2 from k Regressions
4	0.75
5	0.80
10	0.90



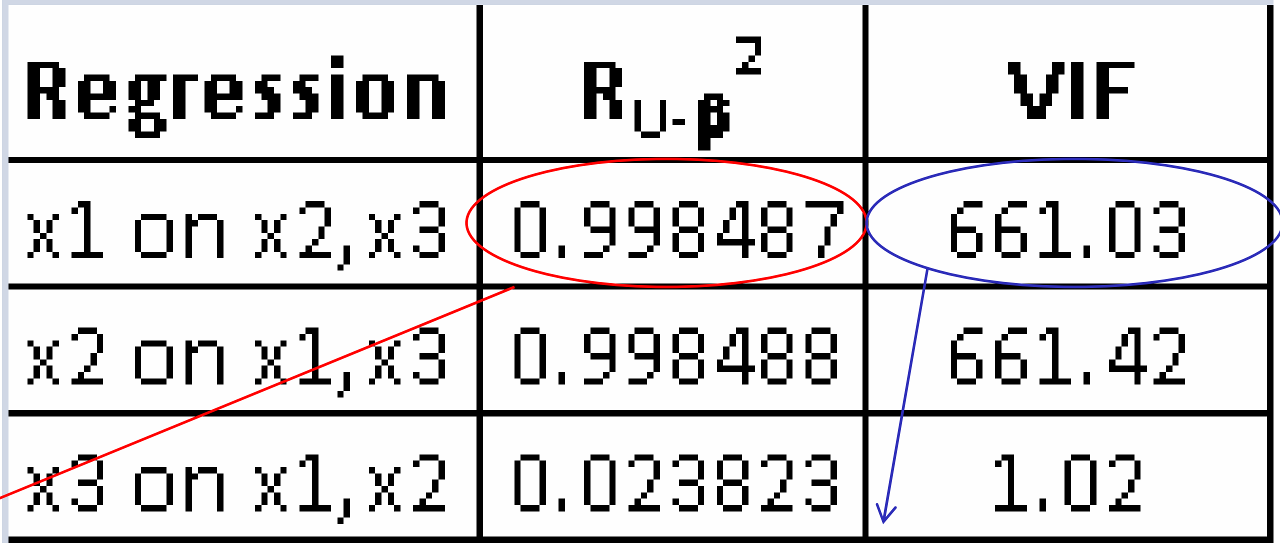
y	x1	x2	x3
10.05	10.19	5.25	5.19
8.08	10.78	5.69	2.84
5.33	7.43	2.51	1.83
7.19	7.79	2.8	3.63
13.79	9.31	4.34	9.19
6.66	12.93	7.78	0.34
3.55	9.97	5.11	1.38
7.37	10.81	5.9	1.84
7.54	11.61	6.63	1.28
3.42	6.97	2.07	0.51

Note: $\rho(x_1,x_2) = 0.999$. As we are performing traditional OLS analysis, this is sufficient (but not necessary) to conclude that multicollinearity is present, so we could stop. In fact, we expect *extreme* multicollinearity. But suppose we wish to verify this result and quantify the multicollinearity via VIFs without actually running 3 regressions…

Excel Shortcut for Calculating OLS VIFs



y	x1	x2	x3
10.05	10.19	5.25	5.19
8.08	10.78	5.69	2.84
5.33	7.43	2.51	1.83
7.19	7.79	2.8	3.63
13.79	9.31	4.34	9.19
6.66	12.93	7.78	0.34
3.55	9.97	5.11	1.38
7.37	10.81	5.9	1.84
7.54	11.61	6.63	1.28
3.42	6.97	2.07	0.51



Regression	$R_{U-\beta}^2$	VIF
x1 on x2,x3	0.998487	661.03
x2 on x1,x3	0.998488	661.42
x3 on x1,x2	0.023823	1.02

Note: $\rho\ (x_1,x_2) = 0.999$

`=INDEX(LINEST(D$4:D$13,E$4:F$13,1,1),3)`

`=1/(1-D27)`

Excel Notes

1. INDEX(array, row number, column number) Pulls the ith row jth column from an array (omitted arguments treated as 1)
2. LINEST(known y's, known x's, const, stats) returns the linear regression of [y] on [x] with an intercept term and additional regression stats
3. R^2 lies in the 3rd row and 1st column (omitted) of that array (refer to Excel Help on INDEX and LINEST for more details)

As expected, VIFs for x_1 and x_2 are very high and greatly exceed even the most generous threshold. x_3 is shown to be reasonably "independent" of the other regressors. Note: We didn't have to fit a single equation to perform this analysis.

What About the Zero Intercept Case?

“The Bears are what we thought they were. They're what we thought they were. We played them in preseason — who the [expletive] takes a third game of the preseason like it's [expletive]? [Expletive]! We played them in the third game — everybody played three quarters — the Bears are who we thought they were! That's why we took the damn field. Now if you want to crown them, then crown their [expletive]! But they are who we thought they were! And we let them off the hook!”

–Then-Arizona Cardinals Head Coach Dennis Green, October 16, 2006, after the Cardinals blew a 20 point lead in less that 20 minutes against the Chicago Bears on Monday Night Football. Green was fired at the end of the season.

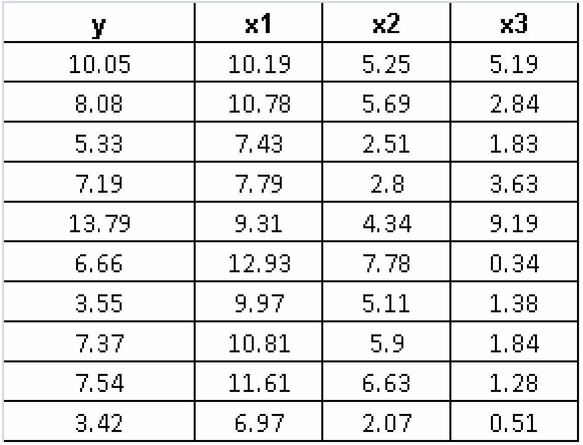
- It is not uncommon (though not a best practice) to force a CER through the origin (or to, in some other way, *constrain* the intercept)
- If multicollinearity is what we thought it was, we should be able to apply the standard formula to zero intercept regression. After all, the formula is found in numerous sources [1], and is implied by statistics given in commercial cost estimating software [2] in the case of zero intercept regression

$$VIF_{\beta} = \frac{1}{(1-R_{U-\beta}{}^{2})}$$

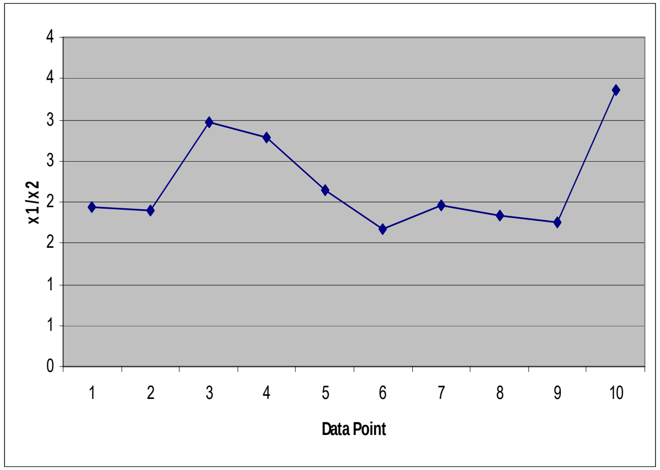
1. CEBoK, Wikipedia, and elsewhere.
2. CO$TAT in particular (though not explicitly given) in output. See later example.

Are They Who We Thought They Were?

- Recall that multicollinearity in ZIR requires *proportionality* among the involved regessors
- We have already established that the ratio (x_1/x_2) is not nearly constant in this data set
- Therefore, we expect very low VIFs and to conclude that no multicollinearity is present when the model is treated as ZIR
- We will attempt (variously) to implement the Dennis Green approach, i.e. calculate VIFs in ZIR



y	x1	x2	x3
10.05	10.19	5.25	5.19
8.08	10.78	5.69	2.84
5.33	7.43	2.51	1.83
7.19	7.79	2.8	3.63
13.79	9.31	4.34	9.19
6.66	12.93	7.78	0.34
3.55	9.97	5.11	1.38
7.37	10.81	5.9	1.84
7.54	11.61	6.63	1.28
3.42	6.97	2.07	0.51

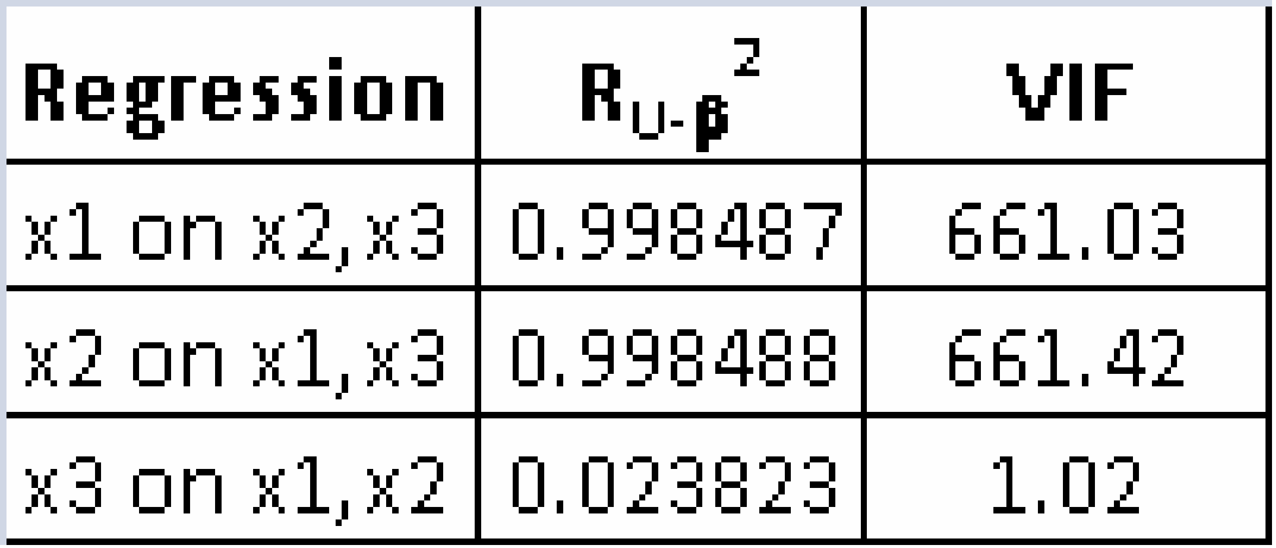


Note: $\rho\ (x_1,x_2) = 0.999$

$$VIF_\beta = \frac{1}{(1 - R_{U-\beta}{}^2)}$$

Attempt 1: Use the Standard Formula (with Excel Shortcut)

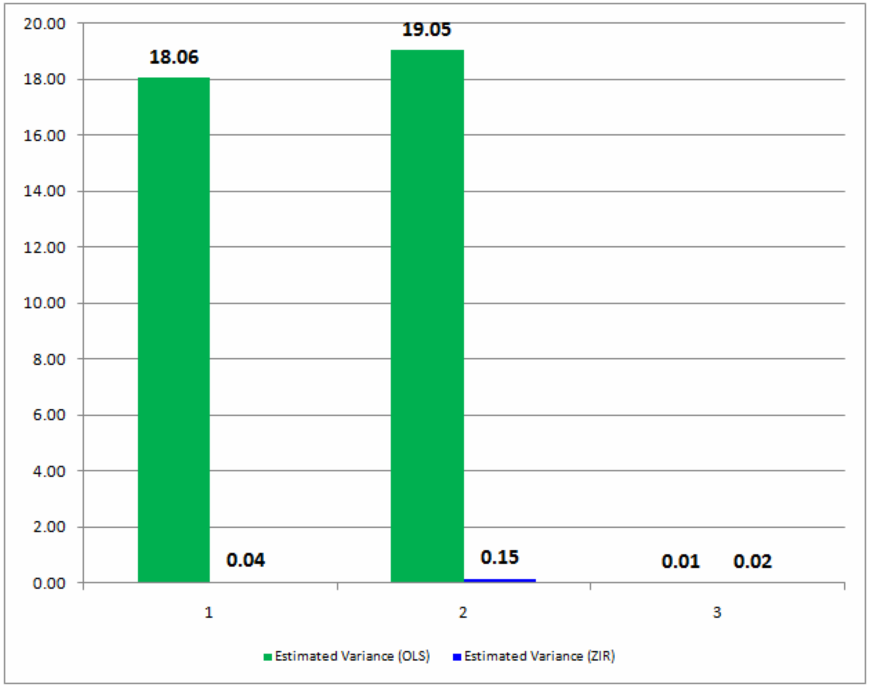
- First attempt at calculating VIF in ZIR: *Treat all coefficients (even in ZIR) as if they had a model intercept, for comparison purposes*
  - While the formula/interpretation of R^2 changes in ZIR, the standard VIF formula uses *OLS* R^2, i.e. measures strength of linear (with intercept]) relationship
  - This leads us to conclude that $VIF_{max} = 661.42$!



Regression	$R_{U\text{-}\beta}^{2}$	VIF
x1 on x2,x3	0.998487	661.03
x2 on x1,x3	0.998488	661.42
x3 on x1,x2	0.023823	1.02

Why Don’t We Believe The Results of Attempt 1?

- Because the variances of estimated coefficients in our ZIR example are *much* lower:



Up to 99.8% reduction in variance of estimated coefficient when intercept is removed

Coefficient	$SE_{\beta\ ols}^2$	$SE_{\beta\ zir}^2$	VIF_{ols}	Implied ZIR N.V. if OLS formula used
β_1	18.06	0.04	661.03	0.0001
β_2	19.05	0.15	661.42	0.0002
β_3	0.01	0.02	1.02	0.0151

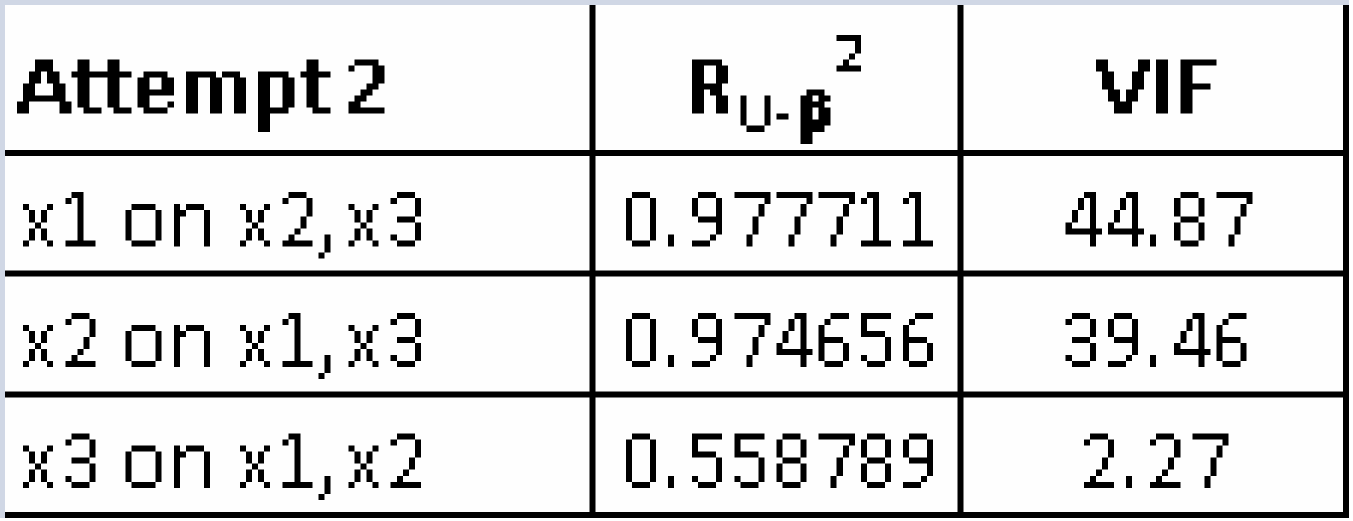
These values are *implausibly* low!

VIFs using the standard (with-intercept) formulas imply implausible results about the native variances of these ZIR coefficients!

Technomics

Attempt 2: Use the Standard Formula, but Apply ZIR Formula for R^2 when Calculating VIF

- Second attempt at calculating VIF in ZIR: Account for different definition of R^2 in ZIR
- This is as simple as changing the "const" argument in our LINEST(.) formula: =INDEX(LINEST(D$4:D$13,E$4:F$13,0,1),3)
- This leads us to conclude that VIF_{max} = 44.87!



Attempt 2	$R_{U-\beta}^2$	VIF
x1 on x2,x3	0.977711	44.87
x2 on x1,x3	0.974656	39.46
x3 on x1,x2	0.558789	2.27

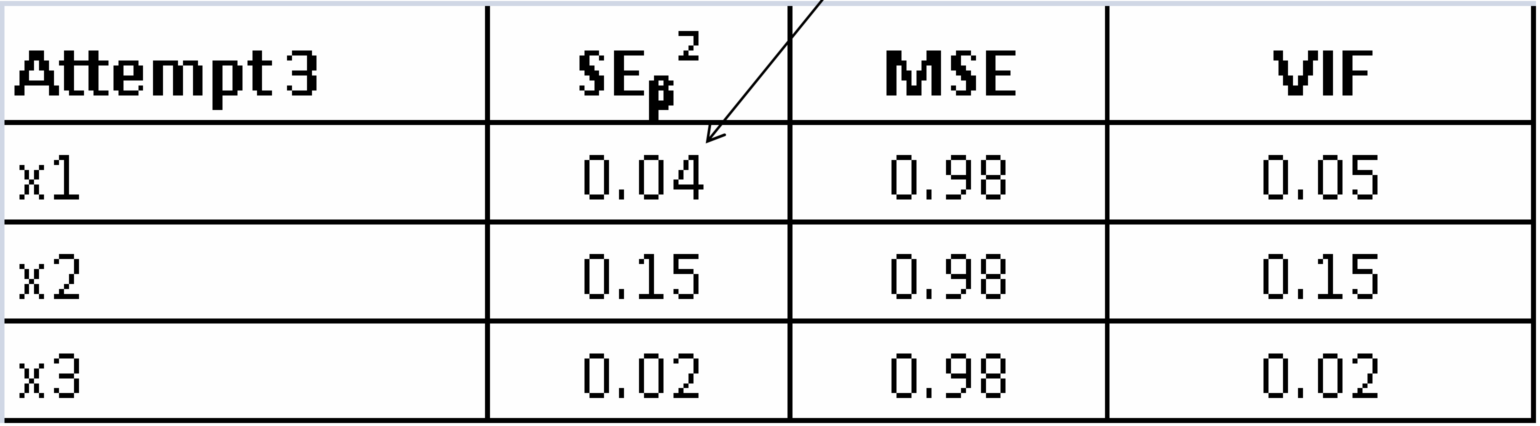
This is better, but we still reach the *spurious* conclusion that severe multicollinearity is present in the model. Let's press on, though…

Attempt 3: Resort to the Old "Approximation"

- Third attempt at calculating VIF in ZIR: Use the approximation: SE_{β}^{2}/MSE
- We know that this formula is imprecise and sometimes gives implausible results, but we are getting desperate…

=INDEX(LINEST(C$4:C$13,D$4:F$13,0,1),2,3))

- This leads us to conclude that $VIF_{max} = 0.15$!



Attempt 3	SE_{β}^{2}	MSE	VIF
x1	0.04	0.98	0.05
x2	0.15	0.98	0.15
x3	0.02	0.98	0.02

These results are untenable because they show VIFs < 1, which is impossible. However, they lead us to the opposite conclusion (that approximate VIFs are small) and therefore multicollinearity is not present. Let's keep going…

Attempt 4: Consider the *Nature* of the VIF

- Another attempt at calculating VIF in ZIR: Consider the nature of the VIF statistic
- It is the multiplicative amount by which the (native) variance about an estimated coefficient is increased due to multicollinearity in the model
- $VIF_j = SE_\beta^2$/native variance, but it can be shown that native variance = $SEE^2/[(n-1)Var(X_j)]$[1] where:
  - SEE= standard error of the estimate (a noisier estimate implies more *native* variance around coefficients within the estimate)
  - n = number of data points (a greater number of data points implies proportionately less native variance in estimated coefficients)
  - $Var(X_j)$ = sample variance of the observations of the jth regressor (variance in the sample data varies *inversely* with variance of the estimated coefficient)
- In other words:

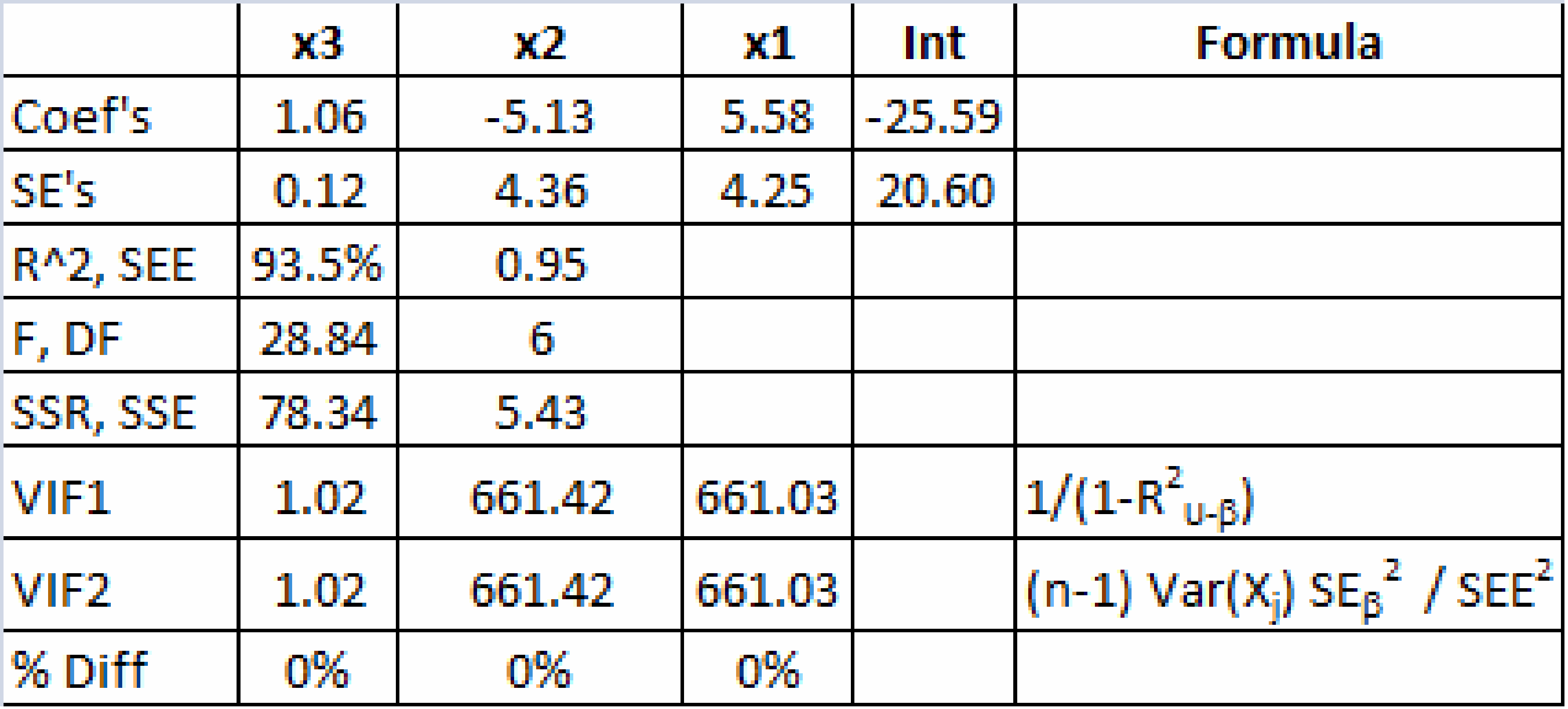
$$VIF = SE_\beta^2 / (SEE/[(n-1)Var(X_j)] = (n-1)\ Var(X_j)\ SE_\beta^2\ /\ SEE^2$$

$$= \mathbf{DEVSQ(X)\ (SE_\beta^2\ \ /\ SEE^2)}$$

[1] http://en.wikipedia.org/wiki/Variance_inflation_factor

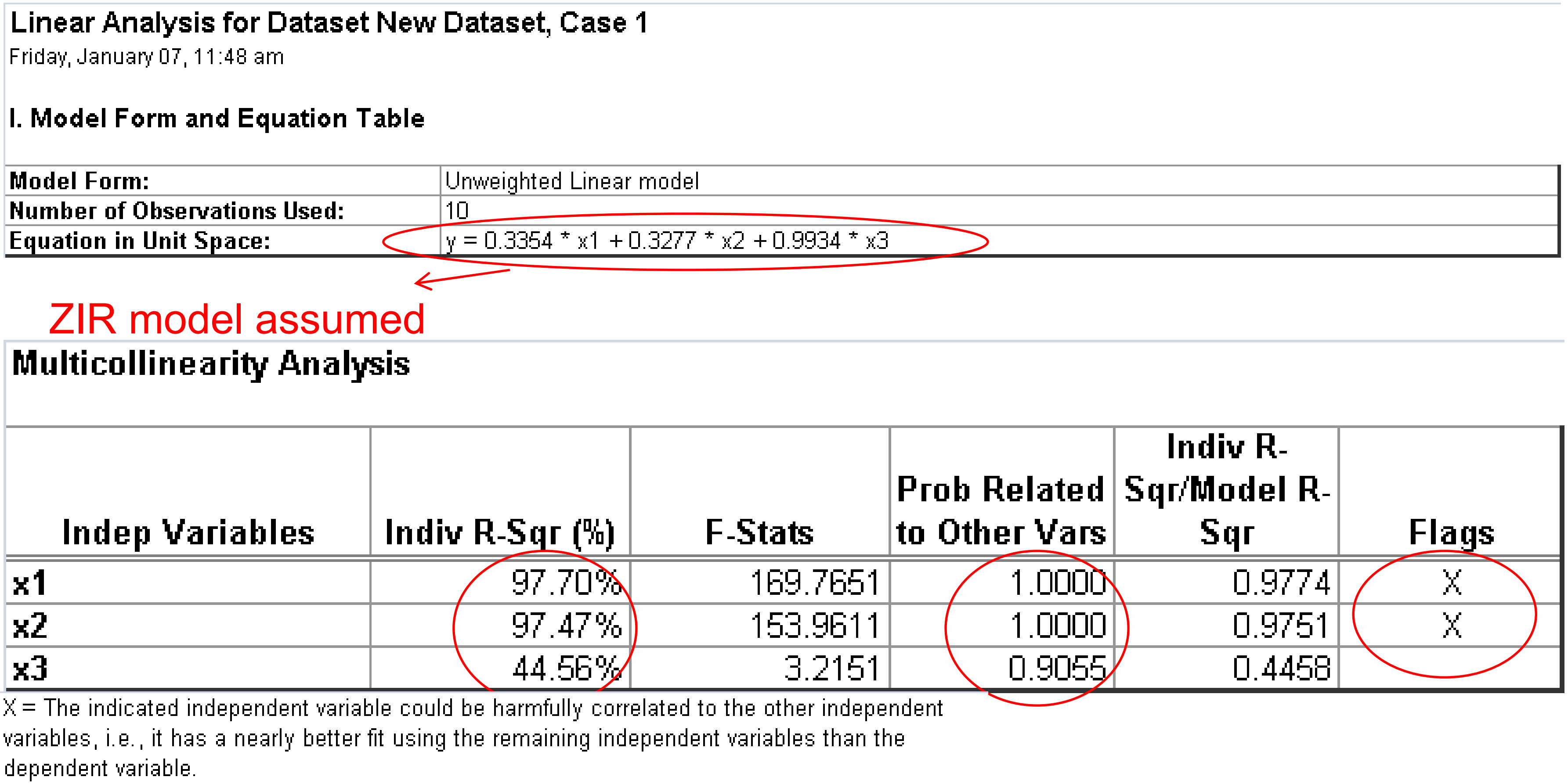
Relationship Between "Native Variance" Method and True VIF

- Regressing our example data *with* an intercept gives us a test case



	x3	x2	x1	Int	Formula
Coef's	1.06	-5.13	5.58	-25.59	
SE's	0.12	4.36	4.25	20.60	
R^2, SEE	93.5%	0.95			
F, DF	28.84	6			
SSR, SSE	78.34	5.43			
VIF1	1.02	661.42	661.03		$1/(1-R^2_{u-\beta})$
VIF2	1.02	661.42	661.03		$(n-1)\ Var(X_j)\ SE_{\beta}^2\ /\ SEE^2$
% Diff	0%	0%	0%		

CO$TAT Output for ZIR Version of Same Problem



Linear Analysis for Dataset New Dataset, Case 1

Friday, January 07, 11:48 am

I. Model Form and Equation Table

Model Form:	Unweighted Linear model
Number of Observations Used:	10
Equation in Unit Space:	y = 0.3354 * x1 + 0.3277 * x2 + 0.9934 * x3

ZIR model assumed

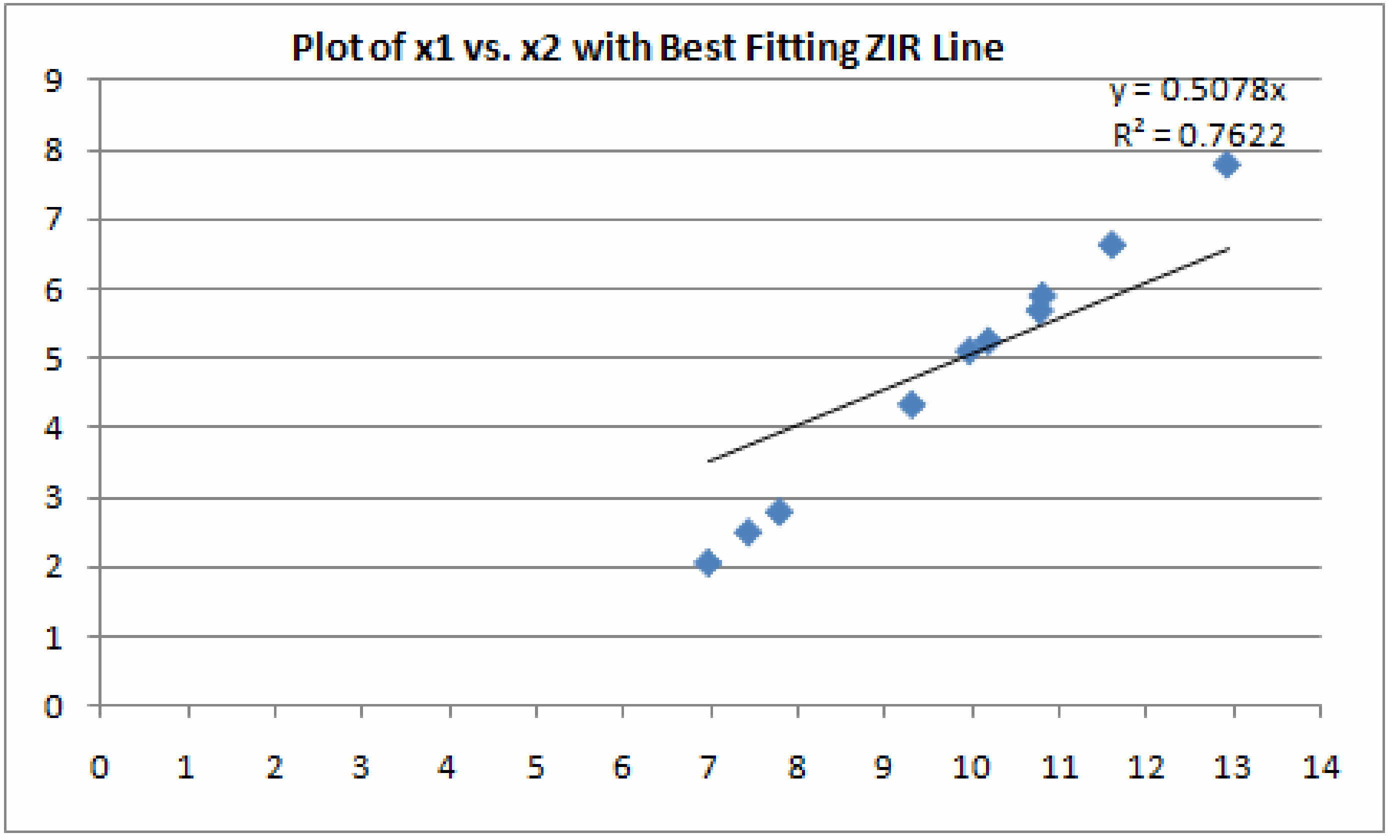
Multicollinearity Analysis

Indep Variables	Indiv R-Sqr (%)	F-Stats	Prob Related to Other Vars	Indiv R-Sqr/Model R-Sqr	Flags
x1	97.70%	169.7651	1.0000	0.9774	X
x2	97.47%	153.9611	1.0000	0.9751	X
x3	44.56%	3.2151	0.9055	0.4458	

X = The indicated independent variable could be harmfully correlated to the other independent variables, i.e., it has a nearly better fit using the remaining independent variables than the dependent variable.

CO$TAT correctly uses the ZIR formula for R^2 (calculates explained variation in terms of comparison to the x-axis, rather than $y = \mu_y$). However, this formula does not apply for our purposes. The explained variation in x_2 due to x_1 (relative to the x-axis) is *not* the same as a measure of the proportionality of the two. *Approximate proportionality* is required for multicollinearity in ZIR.

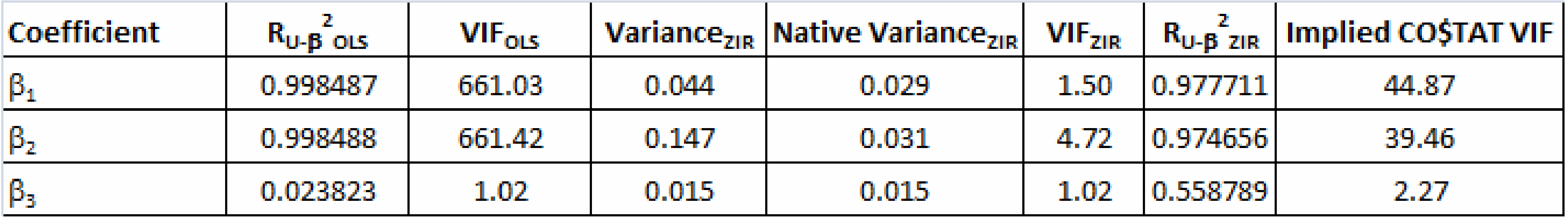
Another View of the Issue



When x_2 is regressed on x_1 with no intercept, the resulting R^2 is only 76%. The points do not nearly lie on any line that passes through the origin. Misuse of R^2 in VIF formulas leads to overstated VIFs and misguided conclusions about multicollinearity in ZIR. As the line of "best fit" shows, the two regressors are actually not all that "correlated" when ZIR is assumed. The line that we "want" to draw violates the zero-intercept constraint.

The *Nature* Formula: Bringing It Home…

- We assert that this formula gives true VIFs, but unlike all of the others we tried, it can be faithfully applied to ZIR[1]



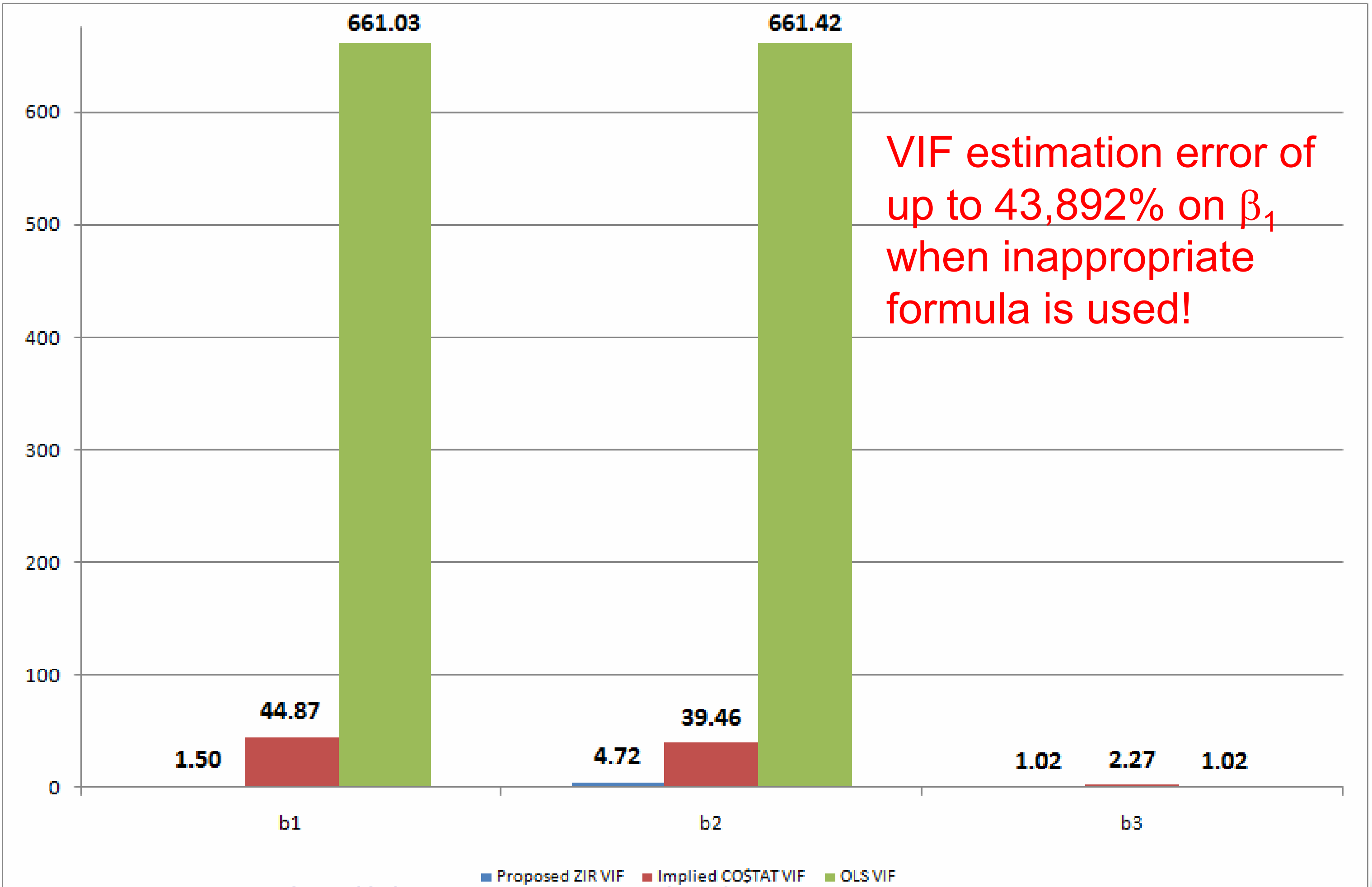
Coefficient	$R^2_{U\text{-}\beta\ OLS}$	VIF_{OLS}	$Variance_{ZIR}$	Native $Variance_{ZIR}$	VIF_{ZIR}	$R^2_{U\text{-}\beta\ ZIR}$	Implied CO$TAT VIF
β_1	0.998487	661.03	0.044	0.029	1.50	0.977711	44.87
β_2	0.998488	661.42	0.147	0.031	4.72	0.974656	39.46
β_3	0.023823	1.02	0.015	0.015	1.02	0.558789	2.27

- Our conclusions about multicollinearity change markedly when ZIR is assumed
  - This is expected because, as we have seen, *multicollinearity is not an intrinsic property of a data set*, but is relative to the model form being hypothesized
  - The linear relationship between x_1 and x_2 is very strong when a constant term is allowed, but not as strong when a constant term is disallowed (as in ZIR)
  - This allows us to keep both variables in the model if even a moderate threshold (VIF_{max} <=5) is used. x_2 is eliminated (perhaps needlessly) or the estimate is biased through Ridge Regression (again, perhaps needlessly) if the OLS VIF formula is used in the ZIR case. When a variable is needlessly eliminated, *explanatory power and cost driver visibility are lost.*

[1] Where n replaces (n-1) due to lack of intercept term.

Alternative Views of VIF for Same Coefficient in Same Data Set

Coefficient	$R^2_{U\text{-}\beta\ OLS}$	VIF_{OLS}	$Variance_{ZIR}$	Native $Variance_{ZIR}$	VIF_{ZIR}	$R^2_{U\text{-}\beta\ ZIR}$	Implied CO$TAT VIF
β_1	0.998487	661.03	0.044	0.029	1.50	0.977711	44.87
β_2	0.998488	661.42	0.147	0.031	4.72	0.974656	39.46
β_3	0.023823	1.02	0.015	0.015	1.02	0.558789	2.27



VIF estimation error of up to 43,892% on β_1 when inappropriate formula is used!

Conclusions

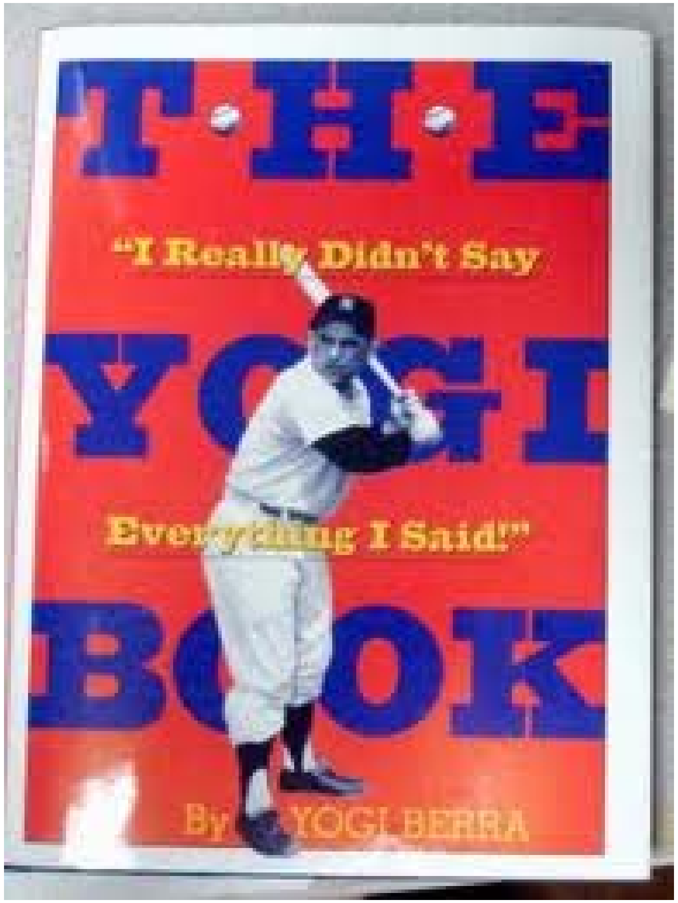
- Multicollinearity is not the same thing as correlation among regressors, but pairwise correlation can be a useful indicator:
  - OLS: Sufficient, but not necessary
  - ZIR: Necessary, but not sufficient
- Multicollinearity in ZIR requires *proportionality*, not just *correlation*
- Multicollinearity is *not* an intrinsic property of a data set: it is *relative* to the model you specify
- Ambiguity about the meaning of R^2 contributes to multicollinearity confusion: Using R^2-based formulas to calculate VIFs can be misleading
- **"I didn't really say everything I said"**
  - SE_β^2/MSE is not a precise formula for VIFs
- **"They are *not* who we thought they were**"
  - Even well-intentioned use of standard VIF formulas can lead to severe overstatement of multicollinearity in ZIR.
  - If you have a genuine OLS multicollinearity problem (without proportionality), the variable you need to drop may be the intercept; you can keep the T/R modules!
  - I propose The *Nature (Boy)* VIF in all cases:

$$\mathrm{VIF} = (n-1)\,\mathrm{Var}(X_j)\,SE_\beta^2\ /\ SEE^2$$

$$\mathbf{= DEVSQ(X)\ (SE_\beta^2\ \ /\ SEE^2)}$$

Ideas for Further Research

- Proportionality coefficient for ZIR that serves analogous role to correlation coefficient in OLS
- Equivalent VIF formulas for nonlinear cases, including General Error Regression Models (GERM)
- Automated software reporting of VIFs (with appropriate formulas) in all cases
  - With recommendations on variables to drop (potentially including the intercept) *and* when to resort to other methods (e.g. Ridge regression)
  - With options so that method of VIF calculation can be directly specified
- A way to directly calculate the VIF of the *intercept* term in OLS
  - Can't be calculated using either formula proposed here because doesn't have a sample variance and can't be regressed on the x-variables
  - Yet our example suggests that sometimes the presence of the intercept *is* the problem



And that is all I have to say about that!

References

- Berra, Yogi. *I Didn't Really Say Everything I Said*. LTD Enterprises (1998)
- Cincotta, Kevin and Lee, Dr. David. Multicollinearity: *Coping With The Persistent Beast* (2007 DoDCAS)
- Judge, George. *The Theory and Practice of Econometrics*. Wiley & Sons (1980)
- Kutner, Nachtsheim, Neter. *Applied Linear Regression Models*, 4th edition. McGraw-Hill Irwin (2004)
- Society of Cost Estimating and Analysis (SCEA). *Cost Estimating Body of Knowledge (CEBoK), v1.1. Module 8: Regression*
- http://en.wikipedia.org/wiki/Variance_inflation_factor