Navigating the Complexities of Multicollinearity in Regression Analysis

Arturs Kalnins

Professor of Management & Entrepreneurship

Courtesy Professor of Economics

Tippie College of Business, University of Iowa

Casual observations from 25 years as an empirical business strategy scholar

- We often observe unrealistically large estimated coefficients and large t-statistics when independent variables are correlated.

Casual observations from 25 years as an empirical business strategy scholar

- We often observe unrealistically large estimated coefficients and large t-statistics when independent variables are correlated.
- The unrealistic coefficients come in pairs
  - they are of opposite sign
  - one of the signs may not align with prior expectations.

Casual observations from 25 years as an empirical business strategy scholar

- We often observe unrealistically large estimated coefficients and large t-statistics when independent variables are correlated.
- The unrealistic coefficients come in pairs
  - they are of opposite sign
  - one of the signs may not align with prior expectations.
- The more highly correlated the pair of independent variables, the larger the absolute values of their coefficients.

Casual observations from 25 years as an empirical business strategy scholar

- We often observe unrealistically large estimated coefficients and large t-statistics when independent variables are correlated.
- The unrealistic coefficients come in pairs
  - they are of opposite sign
  - one of the signs may not align with prior expectations.
- The more highly correlated the pair of independent variables, the larger the absolute values of their coefficients.
- What do these estimated coefficients mean, really?
  - surprising results worthy of publication?
  - or Type 1 errors: false positives where no legitimate result exists?

Casual observations from 25 years as an empirical business strategy scholar

- We often observe unrealistically large estimated coefficients and large t-statistics when independent variables are correlated.
- The unrealistic coefficients come in pairs
  - they are of opposite sign
  - one of the signs may not align with prior expectations.
- The more highly correlated the pair of independent variables, the larger the absolute values of their coefficients.
- What do these estimated coefficients mean, really?
  - surprising results worthy of publication?
  - or Type 1 errors: false positives where no legitimate result exists

What is being argued. Some examples drawn from several hundred empirical papers

generalized linear models multicollinearity does not bias coefficients, it only makes them unstable. In addition, multicollinearity does not reduce the predictive power or reliability of the model as a whole; it

Again, all six two-way terms (in bold) are unchanged (i.e., non-biased). This result is expected because multicollinearity does not bias estimates of regression slope parameters (unless it is extremely high) even as it inflates standard errors

tion suggests there is some potential multicollinearity between these variables, but this is not too large a concern since the presence of multicollinearity does not bias the coefficients.

Multicollinearity does not bias the parameter estimates of the effects, but may inflate standard errors and decrease statistical significance (Berry & Feldman, 1985). To some degree, this problem is countered by the large sample sizes.

So…

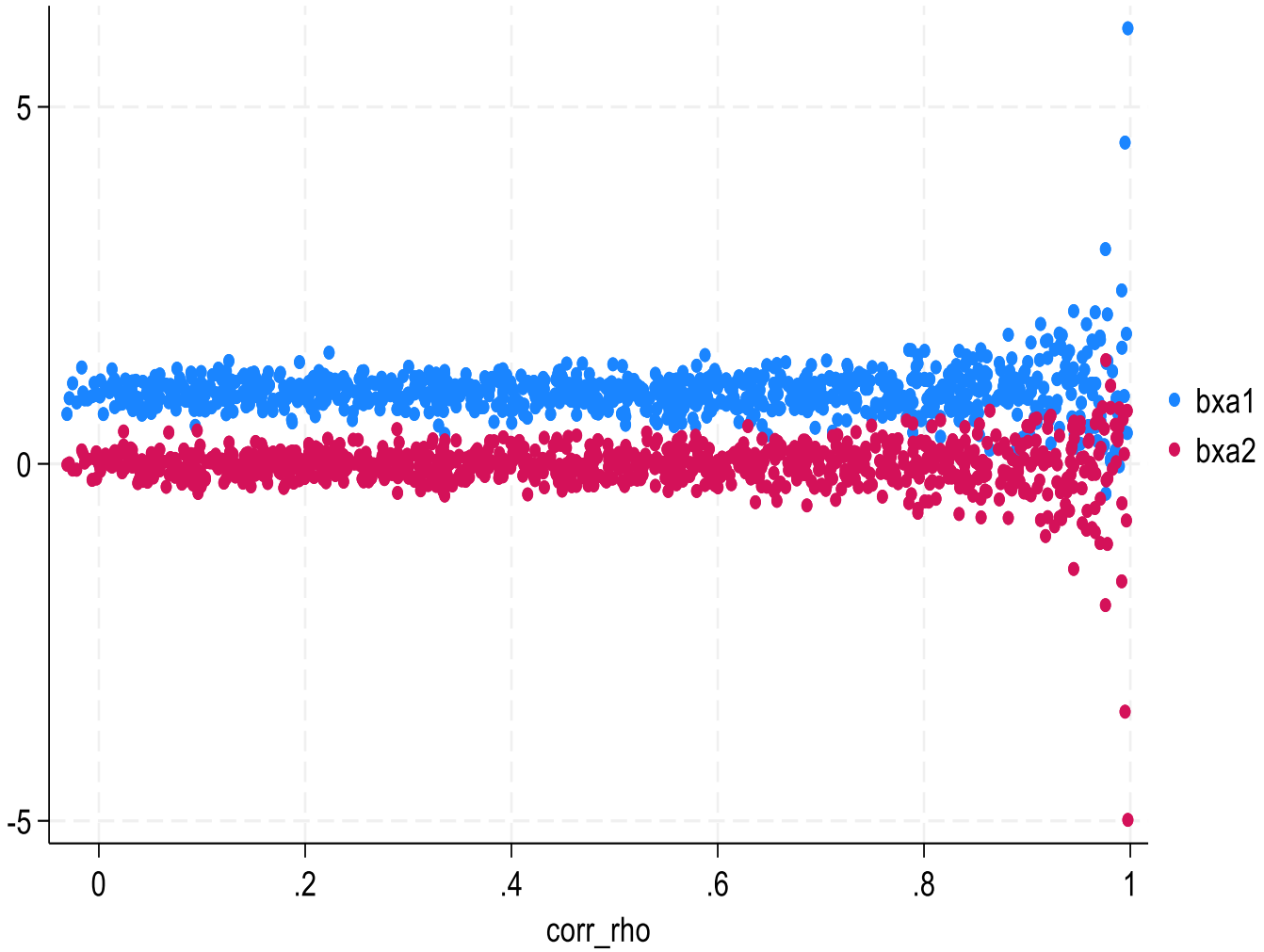
- *But multicollinearity does not bias coefficients, right? It only inflates standard errors. My textbook says so… Goldberger says so! Dozens of Youtube videos say so…*
- *Multicollinearity may make it harder to achieve statistically significant results, but if results exhibit statistically significance, they should be trusted. (Leamer, 1983)*

But?

- *But multicollinearity does not bias coefficients, right? It only inflates standard errors. My textbook says so… Goldberger says so! Dozens of Youtube videos say so…*
- *Multicollinearity may make it harder to achieve statistically significant results, but if results exhibit statistically significance, they should be trusted. (Leamer, 1983)*
- **True only in one special, ideal case**: regression is perfectly specified as per the Classical Linear Regression Model (CLRM).
- If not…
  - If there are any omitted variables, multicollinearity inflates omitted variable bias.
  - Omitted variable bias + positive correlations will result in "beta polarization," coefficients pushed in opposite directions (Kalnins & Praitis Hill, 2025)
  - Common factor (generalized structure of correlated measurement error) will have the same polarizing effects (Kalnins, 2018, 2022)

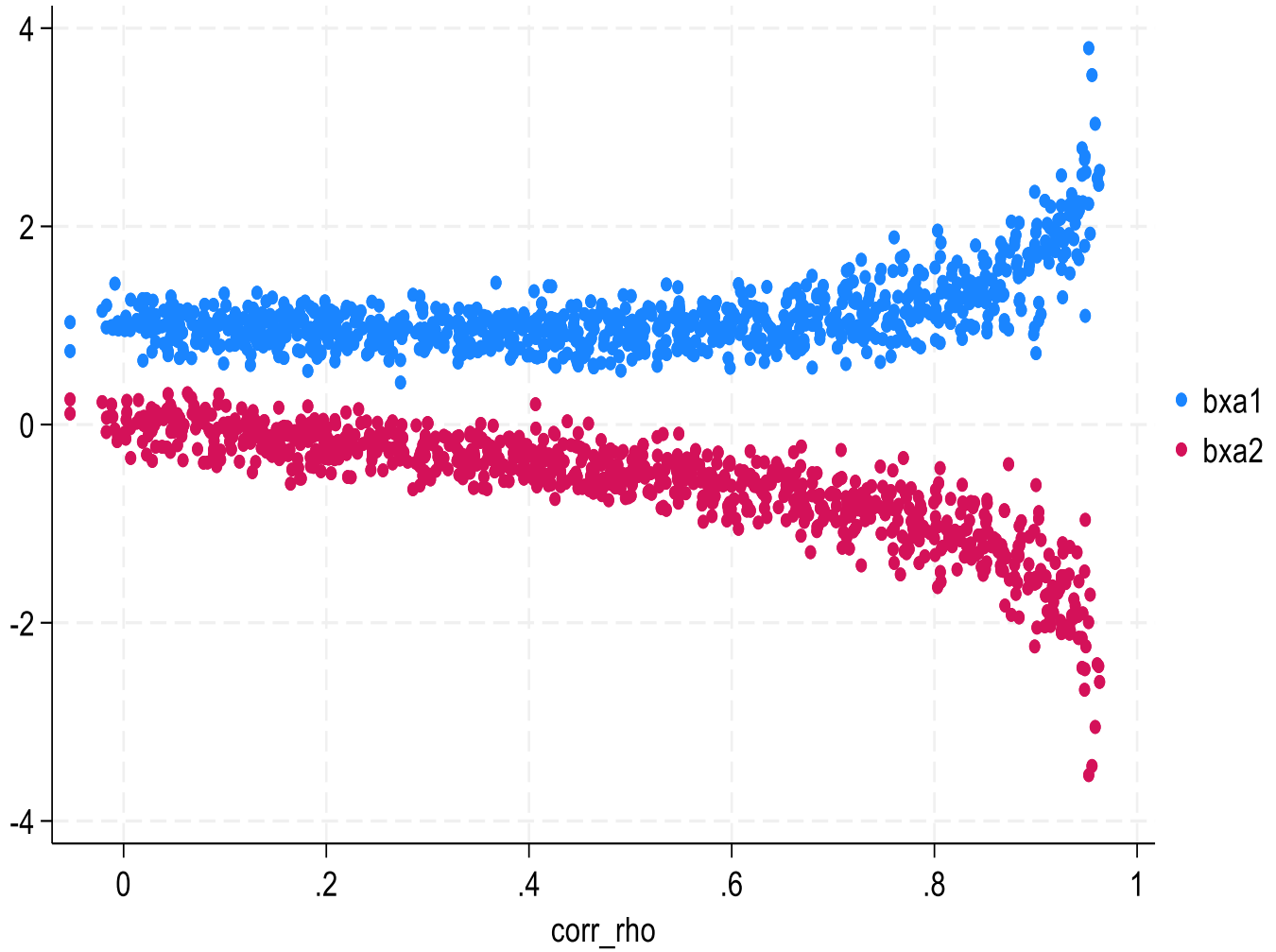
Simulation example where CLRM holds.

$\beta_{X_1} = 1, \beta_{X_2} = 0$
$R^2 = 0.04$



Simulation example where CLRM does not hold.

$\beta_{X_1} = 1$, $\beta_{X_2} = 0$
$R^2 = 0.04$



But?

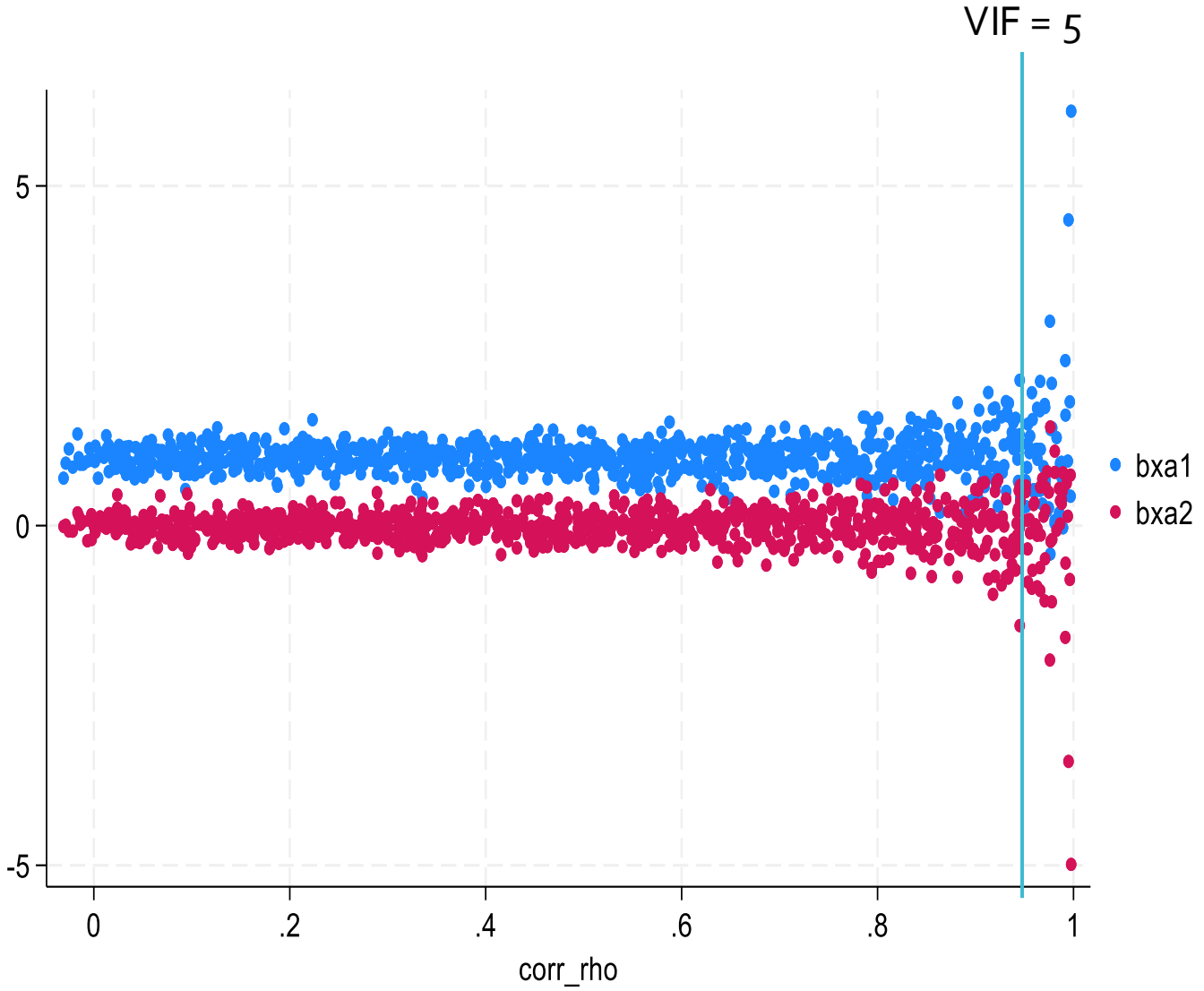
- *But we have calculated VIF scores. They are below 5! Or 10! No multicollinearity concerns now, right?*

But?

- *But we have calculated VIF scores. They are below 5! Or 10! No multicollinearity concerns now, right?*
- **Wrong!** Low VIF scores may be associated with multicollinearity-induced type 1 errors (Kalnins & Praitis Hill, 2025).
- Because:
  - Omitting variables decreases VIF scores and also causes beta polarization.
  - VIF scores imply nothing about type 1 errors: Statistically significant coefficients are not made more legitimate by VIF scores below thresholds.

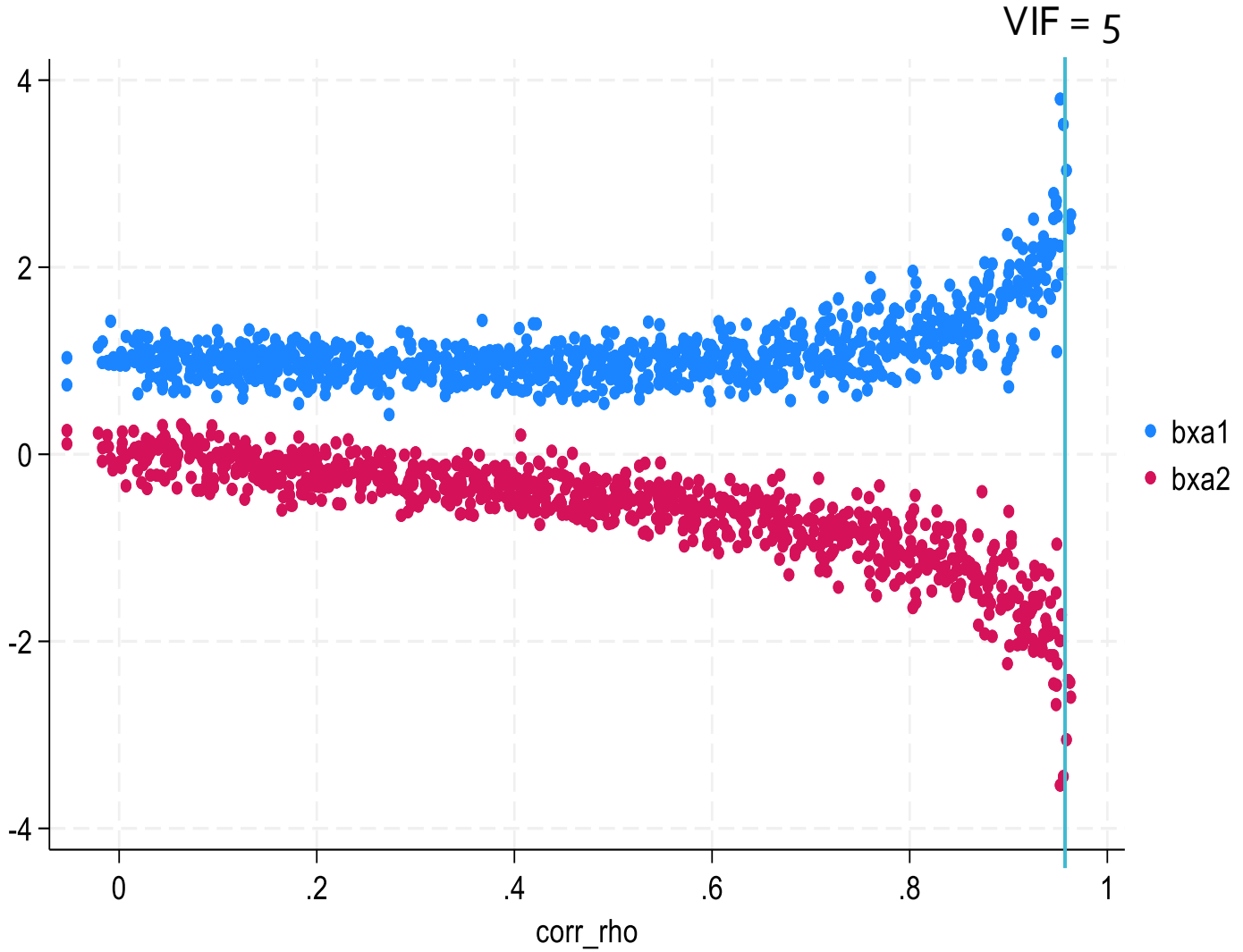
Simulation example where CLRM holds.

$\beta_{X_1} = 1, \beta_{X_2} = 0$
$R^2 = 0.04$

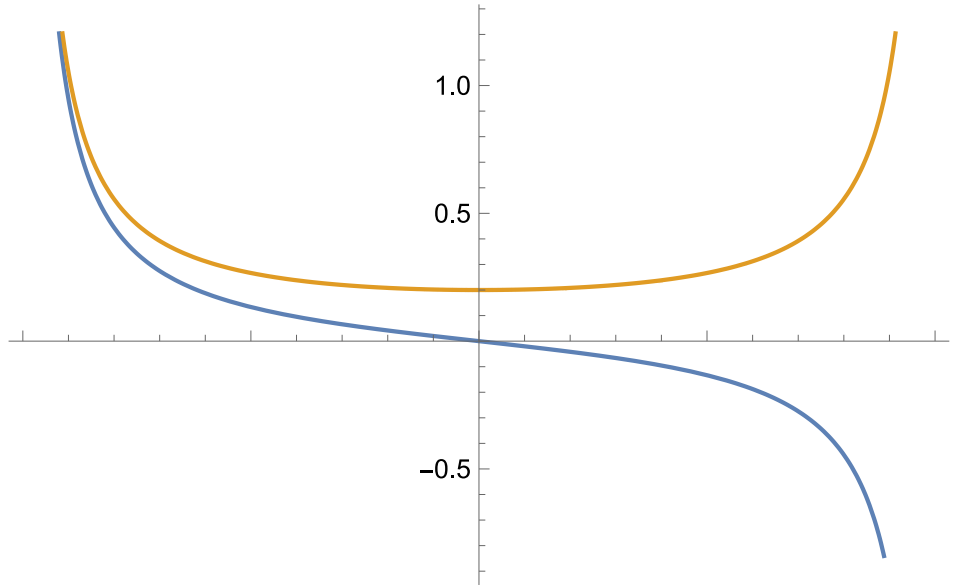


Simulation example where CLRM does not hold.

$\beta_{X_1} = 1, \beta_{X_2} = 0$
$R^2 = 0.04$



Stylized version derived mathematically from
(1) omitted variable
(2) common factor – common measurement error



But?

- *If two variables are correlated and one's sign flips when you add the other in a regression, is that totally ok? It is a suppressor variable.*

But?

- *If two variables are correlated and one's sign flips when you add the other in a regression, is that totally ok? It is a suppressor variable.*
- *A suppressor variable increases the predictive validity of another variable or variables in the model.*
- *Essentially, suppressor variables can help reveal the true relationship between the independent variable (predictor) and the dependent variable (outcome), by accounting for certain variance that would otherwise obscure this relationship.*
- *Without the suppressor: The predictor might show a weak or no relationship with the outcome, or even an opposite relationship, due to confounding noise.*
- *With the suppressor: The suppressor variable "cleans up" this noise, allowing the predictor to reveal its true strength in relation to the outcome.*

But?

- *If two variables are correlated and one's sign flips when you add the other in a regression, is that totally ok? It is a suppressor variable.*
- **Often Wrong!** Signs can often flip from true direction of the effect to the opposite when beta coefficients are polarized by multicollinearity.

An example from Medicine

Vatcheva et al.

corr(BMI, WC) ~= 0.7

Variables	Estimates	Models* for Systolic Blood Pressure			Models* for Diastolic Blood Pressure		
		Model 1†	Model 2‡	Model 3^	Model 1†	Model 2‡	Model 3^
BMI (body mass index)	Coeff. estimate	0.4		0.53	0.34		0.43
	SE	0.04		0.09	0.03		0.06
	p-value	<0.0001		<0.0001	<0.0001		<0.0001
	VIF	1.01		4.48	1.01		4.48
Waist Circumf.	Coeff. estimate		**0.13**	**−0.08**		**0.12**	**−0.05**
	SE		0.02	0.04		0.01	0.03
	p-value		<0.0001	**0.0526**		<0.0001	**0.0679**
	VIF		1.06	**4.67**		1.06	**4.67**

The first scholar to raise concerns about this issue

- First Economics Nobel Laureate (1969) Ragnar Frisch agrees…
- *"[The Statistician] 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.*
- *In practice these cases are apt to arrive more frequently than is usually recognized. As a matter of fact I believe that a substantial part of the regression and correlation analyses which have been made on economic data in recent years is nonsense for this very reason" (Frisch, 1934: p. 6).*
- His explanation? Multiple variables are included in a single regression contain a common unobservable variable, with orthogonal measurement errors.
- He was on the right track about the common unobservable variable but not quite correct. Orthogonal measurement errors are not sufficient to generate nonsense results.

I will show why multicollinearity does more than just inflate errors when variables are omitted. It inflates biases

- True data generating process $y = \delta_X^l X + \delta_Z^l Z + e^l$
  - Z is not observable
- Regression estimates misspecified equation:
  - $y = \beta_X^s X + e$ where $e = \delta_Z^l Z + e^l$.
- Omitted Variable Bias Equation
  - $\beta_X^s = \delta_X^l + \left[\frac{X^T Z}{X^T X}\right] \delta_Z^l$
- This derivation originally published in: Kalnins, A., & Praitis Hill, K. (2025). The VIF score. What is it good for? Absolutely nothing. *Organizational research methods*, *28*(1), 58-75.

Simplifying Assumptions (still generalizable)

- We assume there are two standardized variables x_1 and x_2 in the "short" regression matrix *X*,
- Omitted important third variable, also standardized, and designated by *Z*.
- For even more simplicity, we assume true effects are $\delta_X^l = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ but that $\delta_Z^l = 1$.
- In other words, the true DGP $y = Z + e$, where *e* is a spherical error term with variance σ^2.
- Yet we can at best estimate $y = \beta_X^s X + e$ because of the unavailability of Z.
- Based on assumptions, the Omitted Variable Bias Equation can be simplified:
  - $\beta_X^s = \delta_X^l + \left[\frac{X^T Z}{X^T X}\right] \delta_Z^l$ becomes $\beta_X^s = \left[\frac{X^T Z}{X^T X}\right]$

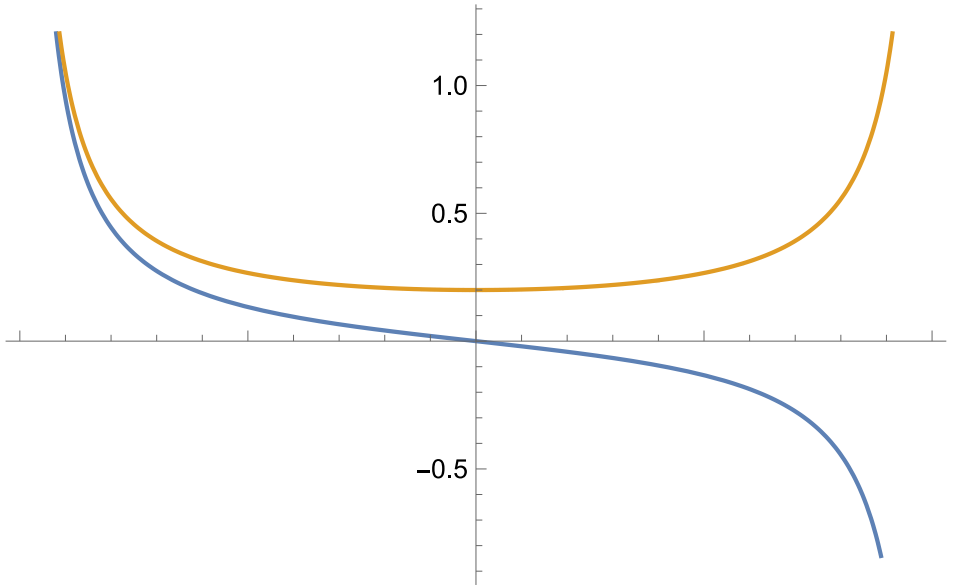
Biases to beta coefficients where their true effects are zero!

- Given our standardization assumptions, $X^T Z = n \begin{bmatrix} corr(x_1, Z) \\ corr(x_2, Z) \end{bmatrix}$.
- Correlation matrix $X^T X = n \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$ with correlation θ between x_1 and x_2.
- The inverse $[X^T X]^{-1} = \frac{1}{n(1-\theta^2)} \begin{bmatrix} 1 & -\theta \\ -\theta & 1 \end{bmatrix}$.
- VIF statistic $= \frac{1}{1-\theta^2}$ in a regression with $k_X = 2$ (only two regressors)
- We can write $[X^T X]^{-1} = \frac{VIF}{n} \begin{bmatrix} 1 & -\theta \\ -\theta & 1 \end{bmatrix}$ with the VIF now treated as a scalar.
- The expected value of the estimated coefficients can be written as:
- $\beta_X^S = \begin{bmatrix} X^T Z \\ X^T X \end{bmatrix} = \frac{VIF}{n} \begin{bmatrix} 1 & -\theta \\ -\theta & 1 \end{bmatrix} n \begin{bmatrix} corr(x_1, Z) \\ corr(x_2, Z) \end{bmatrix} = VIF \begin{bmatrix} corr(x_1, Z) - \theta corr(x_2, Z) \\ corr(x_2, Z) - \theta corr(x_1, Z) \end{bmatrix}$

What happens when correlation θ between the two x variables approaches one?

- $\beta_X^S = VIF \begin{bmatrix} corr(x_1, Z) - \theta corr(x_2, Z) \\ corr(x_2, Z) - \theta corr(x_1, Z) \end{bmatrix}$
- VIF statistic $= \frac{1}{1-\theta^2}$
- When $\theta \rightarrow \pm 1$, VIF approaches ∞ and thus causes the bias inflation.
- If $corr(x_1, Z) \neq corr(x_2, Z)$ and the difference is larger than smallest value e, while VIF approaches ∞.
  - [until matrix of correlations among $corr(x_1, Z) \neq corr(x_2, Z)$ loses its positive definiteness.]
- The bias approaches ±∞.
- Further, if $\theta \rightarrow 1$ then there will always exist a θ close enough to 1 such that the quantities $(corr(x_1,Z) - \theta\ corr(x_2,Z))$ and $(corr(x_2,Z) - \theta\ corr(x_1,Z))$ will necessarily have opposite signs.
- This combination yields $\beta_X^S \rightarrow \begin{bmatrix} +\infty \\ -\infty \end{bmatrix}$. Beta Polarization

Corr(x1,Z) = .2
Corr(x2,Z) = 0
When
Corr(x1,x2) = 0,
the Corrs = βs



$$\beta_X^s = VIF \begin{bmatrix} corr(x_1, Z) - \theta corr(x_2, Z) \\ corr(x_2, Z) - \theta corr(x_1, Z) \end{bmatrix}$$

The technical contribution of my work

- I developed analytic theory that when two collinear variables share correlations with an unobservable common factor (e.g., common measurement error, Kalnins, 2018) or an omitted variable (Kalnins and Praitis Hill, 2025), their estimated beta coefficients become polarized and highly misleading.
- No matter how small their real effects, when correlation between two regressor variables approaches one:
  - Estimated beta coefficients of these regressor variables on any dependent variable y will tend towards infinite magnitudes in opposite directions.
  - The standard errors at high correlations will grow as well (Goldberger is right about that!)
    - But often not enough to eliminate the false appearance of statistical significance of the near-infinite beta coefficients!

The implications of my work

- First, multicollinearity does bias estimates—can create Type 1 errors.

The implications of my work

- First, multicollinearity does bias estimates—can create Type 1 errors.
- Second, the frequently expressed perspective that multicollinearity is strictly a problem of small data set size, i.e., "micronumerosity," is not correct.
  - In the case of common-factor multicollinearity, results will be misleading even if an infinite population were to be analyzed.
  - Misleading results will be more likely to be viewed as meaningful with large data sets because of large t-statistics.

The implications of my work

- First, multicollinearity does bias estimates—can create Type 1 errors.
- Second, the frequently expressed perspective that multicollinearity is strictly a problem of small data set size, i.e., "micronumerosity," is not correct.
  - In the case of common-factor multicollinearity, results will be misleading even if an infinite population were to be analyzed.
  - Misleading results will be more likely to be viewed as meaningful with large data sets because of large t-statistics.
- Third, the conventional wisdom that exogenous control variables are harmless (e.g., Angrist and Pischke, 2009) is wrong.
  - Even a fully exogenous control variable can bias a variable of theoretical interest if the two are correlated via a common factor. Adding more control variables may exacerbate the problem.

The implications of my work

- First, multicollinearity does bias estimates—can create Type 1 errors.
- Second, the frequently expressed perspective that multicollinearity is strictly a problem of small data set size, i.e., "micronumerosity," is not correct.
  - In the case of common-factor multicollinearity, results will be misleading even if an infinite population were to be analyzed.
  - Misleading results will be more likely to be viewed as meaningful with large data sets because of large t-statistics.
- Third, the conventional wisdom that exogenous control variables are harmless (e.g., Angrist and Pischke, 2009) is wrong.
  - Even a fully exogenous control variable can bias a variable of theoretical interest if the two are correlated via a common factor. Adding more control variables may exacerbate the problem.
- Fourth, bivariate correlation as low as 0.3 can be problematic.

Multi-collinearity can be a problem. What can be done about it?

- Report full bivariate correlation tables, including dependent variables, interaction terms and polynomial terms.
  - No good reason to exclude interactions from correlation tables.
- If you use a lot of fixed effects, report full bivariate correlation tables of residuals after regression on only fixed effects.
- If two collinear variables, report multiple specifications:
  - Neither variable
  - Variable 1 only
  - Variable 2 only
  - Both variables
- Note I am *not* advocating simply dropping a correlated variable. The point is to compare specifications (with and without control) and to compare effects of hypothesized variable.

What to do if you have hypotheses attempting to discriminate effects of correlated variables

- Authors should present clear, straightforward theory-based hypotheses.
- Data with multicollinearity should not be used to test tricky, overly clever theory.
    - Counter-intuitive theory may be simply be "predicting" what are really type 1 errors
    - "Horse-race" opposing hypotheses are subject to type 1 error being the winning horse.
    - HARKing is particularly dangerous here.
    - Type 1 errors may "replicate" if the same controls are included by future authors in future studies.

Pay attention to your control variables

- Authors should present clear evidence that the correlated control's coefficient (when focal variable is included) is:
  - consistent in sign with existing theory
  - consistent in sign and magnitude existing results from other sources.
  - Many papers ignore control results that make no sense (who cares? They are only controls!). Bad practice.
- If correlated control variable appears to have the "wrong sign" (omitted variable bias in wrong direction, multicollinearity inflates it), focal regressor's coefficient size is also likely to be inflated by multicollinearity. This may be a type 1 error.

What not to do: Previously Discredited mitigation approaches

- Mean-centering the regressors (Echambadi and Hess, 2007),
- Residualizing one independent variable by regressing it on the other ones (Kennedy, 2003)
- Orthogonalization of the variables via principal components (Mitchell, 1991).

What not to do: I discredit a few approaches to "deal with" multi-collinearity

- Penalized regression/Ridge regression:
  - Can only reduce the size of the beta coefficients.
  - It cannot flip them back to the correct sign with a meaningful t-statistic
- Collecting additional data/Replicating
  - If CLRM does not hold, additional data will just yield more Type 1 errors.
  - Replicating could just yield the same Type 1 errors as the original.
- Partial Least Squares – Does not in any way "solve" the multicollinearity problem. It just assumes it away. Attributes effects to both variables equally.
- Do nothing – several texts advocate ignoring the problem. Based on supposed unbiasedness.

Some more flaws of VIF derived in Kalnins & Praitis Hill, 2025

- When even one single relevant independent variable is unavailable and therefore omitted (like in every regression ever estimated by anyone anywhere) a VIF close to its minimum of 1 may be associated with amplified t-statistics that are sufficiently large to generate type 1 error.
  - The commonly accepted idea that low VIF scores provide sufficient information to dismiss concerns related to multicollinearity is not valid. Pure myth.
- Related observation: insufficiently conservative VIF thresholds. Even if VIF scores could be used, hypothetically, to dismiss multicollinearity concerns, the currently accepted cutoffs of 10 or 5 are far too high.
  - For example, correlation $\theta = 0.7$ implies a VIF of 1.96 in a two-variable regression model. While research routinely recognizes a bivariate correlation of 0.7 as highly problematic, a VIF < 2 is inappropriately viewed as perfectly acceptable using even the most stringent cutoffs.

What not to do: Previously Discredited mitigation approaches

- Mean-centering the regressors (Echambadi and Hess, 2007),
- Residualizing one independent variable by regressing it on the other ones (Kennedy, 2003)
- Orthogonalization of the variables via principal components (Mitchell, 1991).

What not to do: I discredit a few approaches to "deal with" multi-collinearity

- Penalized regression/Ridge regression:
  - Can only reduce the size of the beta coefficients.
  - It cannot flip them back to the correct sign with a meaningful t-statistic
- Collecting additional data/Replicating
  - If CLRM does not hold, additional data will just yield more Type 1 errors.
  - Replicating could just yield the same Type 1 errors as the original.
- Partial Least Squares – Does not in any way "solve" the multicollinearity problem. It just assumes it away. Attributes effects to both variables equally.
- Do nothing – several texts advocate ignoring the problem. Based on supposed unbiasedness.

Some more flaws of VIF derived in Kalnins & Praitis Hill, 2025

- When even one single relevant independent variable is unavailable and therefore omitted (like in every regression ever estimated by anyone anywhere) a VIF close to its minimum of 1 may be associated with amplified t-statistics that are sufficiently large to generate type 1 error.
  - The commonly accepted idea that low VIF scores provide sufficient information to dismiss concerns related to multicollinearity is not valid. Pure myth.
- Related observation: insufficiently conservative VIF thresholds. Even if VIF scores could be used, hypothetically, to dismiss multicollinearity concerns, the currently accepted cutoffs of 10 or 5 are far too high.
  - For example, correlation $\theta = 0.7$ implies a VIF of 1.96 in a two-variable regression model. While research routinely recognizes a bivariate correlation of 0.7 as highly problematic, a VIF < 2 is inappropriately viewed as perfectly acceptable using even the most stringent cutoffs.

Thank you for listening!

- This presentation was an overview of my published papers:
- Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal*, *39*(8), 2362-2385.
- Kalnins, A. (2022). When does multicollinearity bias coefficients and cause type 1 errors? A reconciliation of Lindner, Puck, and Verbeke (2020) with Kalnins (2018). *Journal of International Business Studies*, *53*(7), 1536-1548.
- Kalnins, A., & Praitis Hill, K. (2025). The VIF score. What is it good for? Absolutely nothing. *Organizational research methods*, *28*(1), 58-75.