OLS/MLR Analytics II: *Collinearity & Co.*

- Introduction: (Multi)Collinearity
- The Collinearity Regressions
- Multicollinearity with Two RHS Variables: Correlation and R^2
- … with Three or more RHS Variables: R_j^2
- Variance Inflation Factors (VIFs): Easily generate R_j^2
- Endogeneity (Omitted Variable Bias/Impact)
- … Case I: k = 2 to k = 1
- … Case II: k to k-1: k > 2 *(What's the difference?)*
- *… … computing OVB; signing OVB*
- *Simple v. Partial Correlations*

Introduction: (Multi)Collinearity

- Multicollinearity captures the extent to which pairs or groups of RHS variables *move together*
- ***Omitted Variable Bias/Impact (Endogeneity)*:** It plays a role in driving omitted variable impact/bias (endogeneity) … recall the previous discussion.
- ***SRF interpretation of MLR coefficients*:** It can wreak havoc with the ceteris *paribus* interpretation of MLR coefficients… Does it really make sense (when interpreting coefficients) to hold other things constant when things are moving together closely?
- ***Standard Errors*:** It impacts estimated standard errors and the precision with which parameters have been estimated. Greater collinearity leads to higher standard errors (and smaller t-stats) and less precision in estimation.
- ***Wacky results*:** And perhaps most insidiously, it can lead to wacky estimated coefficients, which in turn could very well lead the researcher astray… to focus on less important explanatory factors.
- ***Explanatory power*:** Less collinear explanatory variable candidates plausibly offer more *independent* explanatory power to the RHS, perhaps making them more attractive candidates for inclusion in a MLR analysis.

The Collinearity Regressions

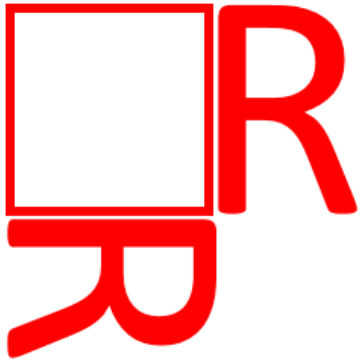
- Collinearity regs: Regress each RHS variable on the other RHS variables in the model
- There are two collinearity regs for the following MLR model: **reg wk1 wk2** and **reg wk2 wk1**

	MLR Model	Collinearity Regressions	
	(1)	(2)	(3)
	rtotgross	wk1	wk2
wk1	-0.0120		0.522***
	(-0.46)		(249.91)
wk2	4.536***	1.673***	
	(95.87)	(249.91)	
_cons	0.401	-0.178	0.889***
	(1.74)	(-1.95)	(17.75)
N	9114	9114	9114
R-sq	0.887	0.873	0.873

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Multicollinearity with Two RHS Variables: Correlation and R^2

- ***Correlation***
  - If there are just two RHS variables in the MLR model, then an obvious measure of collinearity would be the correlation of the two variables, which captures the extent to which the two variables are moving together in a linear fashion.
- ***Coefficient of Determination -*** R^2



  - Alternatively, since we know that R^2 is also correlation squared with SLR models, we could just as easily use the R^2's from the two collinearity regression above to measure the extent to which the two variables, wk1 and wk2, moved together in a linear fashion. Remember that $-1 \le \rho \le 1$ and $0 \le \rho_{xy}^2 = R^2 \le 1 \ldots$ so R^2 reflects the magnitude of the correlation, but not the sign.
- Both of these approaches will give you effectively the same metric. But they differ in one very important respect: By definition, the correlation concept can be applied only to pairs of explanatory variables. In contrast, the R^2 from the collinearity regression approach **can be easily extended** to MLR models with more than just two RHS variables. And so that's what we do to measure multicollinearity.... we employ R^2 from the collinearity regression.

… with Three or more RHS Variables: R_j^2

- We call the R-sq's in the collinearity regressions the *R-squared j* measures of collinearity, R_j^2, where the j index tells you which RHS variable, x_j, is the dependant variable in the collinearity regression.
- In the following example, *wk2* is the most collinear explanatory variable since the R^2 in Model (3) is .959 (which tells us that 95.9% of the variation in the *wk2* variable can be explained by a linear function of the other two explanatory variables, *wk1* and *wk3*).

	MLR Model	Collinearity Regressions		
	(1)	(2)	(3)	(4)
	rtotgross	wk1	wk2	wk3
wk1	0.540***		0.261***	-0.115***
	(21.36)		(111.42)	(-34.25)
wk2	0.745***	2.361***		0.792***
	(9.79)	(111.42)		(131.20)
wk3	4.778***	-1.146***	0.872***	
	(59.84)	(-34.25)	(131.20)	
_cons	-0.601**	0.110	0.0817*	0.287***
	(-2.64)	(1.07)	(2.40)	(8.91)
N	7730	7730	7730	7730
R-sq	0.921	0.886	0.959	0.908

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Variance Inflation Factors (VIFs): Easily generate R_j^2

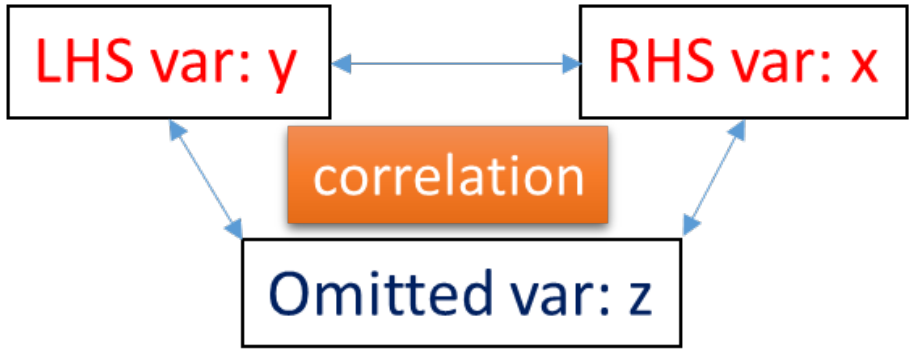
$$VIF_j = \frac{1}{1 - R_j^2}$$

$$R_j^2 = 1 - \frac{1}{VIF_j}$$

```
. reg rtotgross wk1 wk2 wk3

      Source |       SS           df       MS      Number of obs   =     7,730
-------------+----------------------------------   F(3, 7726)      =  30052.55
       Model |  27234043.1         3  9078014.37   Prob > F        =    0.0000
    Residual |  2333803.56     7,726   302.07139   R-squared       =    0.9211
-------------+----------------------------------   Adj R-squared   =    0.9210
       Total |  29567846.7     7,729  3825.57209   Root MSE        =     17.38

------------------------------------------------------------------------------
  rtotgross |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
        wk1 |   .5403215   .0253012    21.36   0.000     .4907242    .5899187
        wk2 |   .7447428   .0760905     9.79   0.000     .5955848    .8939008
        wk3 |    4.77794   .0798398    59.84   0.000     4.621432    4.934447
      _cons |  -.6009747   .2274986    -2.64   0.008    -1.046934   -.1550159
------------------------------------------------------------------------------

. vif

    Variable |       VIF        1/VIF   ||    Rsq j
-------------+----------------------------------
         wk2 |     24.62     0.040613   ||   0.959387
         wk3 |     10.88     0.091905   ||   0.908095
         wk1 |      8.79     0.113804   ||   0.886196
-------------+----------------------------------
    Mean VIF |     14.76
```

Endogeneity (Omitted Variable Bias/Impact)

- ***Endogeneity:*** As discussed previously, estimated coefficients will be *biased* (or less pejoratively, *impacted*) to the extent that those variables are correlated with omitted variables, which are themselves correlated with the dependent variable.
- ***Misleading?*** And as also discussed earlier, this is not so much a ***bias*** as a matter of ***interpretation***. The estimated coefficients reflect the average incremental relationship between changes in the particular RHS variable and changes in the LHS variable, controlling for all the other RHS variables in the model. But of course, if a RHS variable is omitted/dropped/excluded from the model, it's not the same model… and so no one should be surprised to see changes in the OLS/MLR coefficient estimates for the surviving variables.



Endogeneity - Case I: k = 2 to k = 1

- The Omitted Variable Bias/Impact (OVB) on the estimated *wk1* coefficent (when *wk2* is dropped from the *Full Model*) is the product of:
  - ***Collinearity Regression (SLR)***: the estimated wk1 coefficient in the collinearity regression of the omitted variable, *wk2*, on the surviving variable, *wk1*, and
  - ***Full Model (MLR)***: the estimated wk2 coefficient in the full model.

	Full Model (1) rtotgross	wk2 dropped (2) rtotgross	Collinearity Reg (3) wk2
wk1	-0.0120	2.354***	0.522***
	(-0.46)	(176.20)	(249.91)
wk2	4.536***		
	(95.87)		
cons	0.401	4.433***	0.889***
	(1.74)	(13.83)	(17.75)
N	9114	9114	9114
R-sq	0.887	0.773	0.873

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Endogeneity - Case I cont'd

Full Model - SRF$_y$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z$

Collinearity Regression - SRF$_z$: $\hat{z} = \hat{\alpha}_0 + \hat{\alpha}_x x$ (the omitted variable, z, is regressed on the surviving variable, x)

Omitted Variable Bias (dropping z; impact on the x coeff.: $\hat{\alpha}_x \hat{\beta}_z$)

		$\hat{\beta}_z$ from the **MLR Full Model** (SRF$_y$)		
$\hat{\alpha}_x$ from the **SLR Collinearity Regression** (SRF$_z$)		$\hat{\beta}_z > 0$	$\hat{\beta}_z = 0$	$\hat{\beta}_z < 0$
$\hat{\alpha}_x > 0$		positive	0	negative
$\hat{\alpha}_x = 0$		0	0	0
$\hat{\alpha}_x < 0$		negative	0	positive

Endogeneity - Case II: k > 2 to k-1 *(What's the difference?)*

Full Model: This model includes the third explanatory variable, *w*, which will be dropped:

- **Full Model - SRF$_y$:** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_w w$

Collinearity Regression: As before, you also run the collinearity regression, regressing the omitted variable, *w*, on the two surviving/remaining variables in the model, *x* and *z*:

- **Collinearity Regression - SRF$_w$:** $\hat{w} = \hat{\alpha}_0 + \hat{\alpha}_x x + \hat{\alpha}_z z$

Then the omitted variable biases/impacts from excluding *w* from the model are the same sorts of products of coefficients that you saw before:

$OVB_x = \hat{\alpha}_x \hat{\beta}_w$ (the product of the SRF$_w$ ***x* coeff** and the SRF$_y$ ***w* coeff**)

$OVB_z = \hat{\alpha}_z \hat{\beta}_w$ (the product of the SRF$_w$ ***z* coeff** and the SRF$_y$ ***w* coeff**)

Endogeneity - Case II: *Computing OVB*

	Full Model	wk3 dropped	Collinearity Reg
	(1)	(2)	(3)
	rtotgross	rtotgross	wk3
wk1	0.540***	-0.00941	-0.115***
	(21.36)	(-0.33)	(-34.25)
wk2	0.745***	4.528***	0.792***
	(9.79)	(88.37)	(131.20)
wk3	4.778***		
	(59.84)		
_cons	-0.601**	0.772**	0.287***
	(-2.64)	(2.82)	(8.91)
N	7730	7730	7730
R-sq	0.921	0.884	0.908

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Applying the formulas above, we estimate the omitted variable biases using the product of the *wk3* coefficient in Model (1), 4.778, and the respective RHS variable coefficients in the collinearity regression, Model (3):

- *wk1* OVB: 4.778 * (-0.115) = -.549, as advertised
- *wk2* OVB: 4.778 * (0.792) = 3.784, almost as advertised… blame rounding error

Endogeneity - Case II: *Signing OVB*

$$OVB_x = \hat{\alpha}_x \hat{\beta}_w$$

- $\hat{\alpha}_x$ is the *x* coeff in the collinearity reg when the omitted variable *w* is regressed on the other RHS vars
  - … the slope coeff when $WhatsLeft_w$ is regressed on $WhatsNew_x$
  - And $sign(\hat{\alpha}_x) = sign(\rho_{w^+x^+})$, where $\rho_{w^+x^+}$ is the partial correlation of *w* and *x*
- $\hat{\beta}_w$ is the *w* coeff. in the ***Full Model***, when *y* is regressed on all of the RHS variables
  - … the slope coeff when $WhatsLeft_y$ is regressed on $WhatsNew_w$
  - And $sign(\hat{\beta}_w) = sign(\rho_{y^+w^+})$, where $\rho_{y^+w^+}$ is the partial correlation of *y* and *w*

		$\hat{\beta}_w$ from the **MLR** Full Model		
$\hat{\alpha}_x$ (from the **MLR** Collinearity Regression with *w* on the LHS)		$\hat{\beta}_w > 0; \rho_{y^+w^+} > 0$	$\hat{\beta}_w = 0; \rho_{y^+w^+} = 0$	$\hat{\beta}_w < 0; \rho_{y^+w^+} < 0$
$\hat{\alpha}_x > 0; \rho_{w^+x^+} > 0$		positive	0	negative
$\hat{\alpha}_x = 0; \rho_{w^+x^+} = 0$		0	0	0
$\hat{\alpha}_x < 0; \rho_{w^+x^+} < 0$		negative	0	positive

Simple v. Partial Correlations

	(1) Brozek	(2) Brozek
wgt	0.162***	-0.136***
	(12.27)	(-7.08)
abd		0.915***
		(17.42)
_cons	-9.995***	-41.35***
	(-4.18)	(-17.14)
N	252	252

	(1) rtotgross	(2) rtotgross
wk1	2.354***	-0.0120
	(176.20)	(-0.46)
wk2		4.536***
		(95.87)
_cons	4.433***	0.401
	(13.83)	(1.74)
N	9114	9114

Note that the signs of the slope coefficients in the SLR models will agree with the signs of the pairwise correlations. Blame multicollinearity!

- **Box Office Revenues: *wk1* and *wk2***

  wk1 revenues are positively correlated with total box office revenues but *wk1* has a negative coefficient in the MLR model that includes wk2 (left).

- **Bodyfat: *wgt* and *abd***

  wgt is positively correlated with the *Brozek* measure of bodyfat (so *wgt* has a positive slope coeffcient in the SLR model), but *wgt* has a negative coeffcient in the MLR model that includes *abd* (waist size) (right).

OLS/MLR Analytics II: *TakeAways*

- Collinearity of RHS variables can cause problems in MLR models, including: endogeneity (omitted variable bias/impact), misinterpretation of estimated effects, increased standard errors and less precise estimation, and misleading wacky coefficients
- **Collinearity regressions**: When one RHS variable is regressed on the other RHS variables. The R-sq in that regression provides a measure of collinearity, which we often label R_j^2 and which is a logical extension of the pairwise concept of correlation.
- R_j^2s can easily be generated using Variance Inflation Factors (vifs): R_j^2 = 1-1/vif$_j$
- OVB when dropping **one** RHS variable from the Full Model. The magnitude of OVB wrt a surviving variable is the product of two OLS coefficients:
  - the estimated coefficient for the omitted variable in the Full Model, and
  - the estimated coefficient for the surviving variable in the collinearity regression (with the omitted variable on the LHS)
- Those MLR coefficients can also be derived using SLR models in which *WhatsLeft* is regressed on *WhatsNew*
- Knowing partial correlations will enable us to sign those estimated (MLR) coefficients and OVB
- OVB when dropping **more than one** RHS variable from the Full Model: *Complicated*

onwards… to *Stats Review*