OLS/MLR Analytics: *What's New? Not Much!*

- Introduction: MLR v. SLR Models
- Comparing MLR and SLR Results: Forecasting Box Office Revenues
- OLS: A Quick Comparison of SLR and MLR Analytics
- Interpreting MLR Coefficients I:
- … SRFs & Marginal Predicted Effects
- … Partial Correlations
- Endogeneity (Omitted Variable Bias/Impact) I: An Overview

Introduction: MLR v. SLR Models

- ***The difference***: SLR models have a single explanatory (RHS) variable… and with MLR models you can have more than just one explanatory variable. That's all there is to it!
- The MLR the world is a lot more interesting… and a lot more complicated:
  - ***In or out?*** So many RHS variable to choose from! Which RHS variables are in your analysis? …and which do you leave out? and why? How do you decide?
  - ***Interactions (w/in & w/out):*** And once you allow for multiple explanatory variables, you have to worry about how they interact with one another and with the dependent variable as well. This applies to variables in the analysis as well as those excluded/omitted.
  - ***Endogeneity:*** Explanatory variables left out of the MLR analysis could be impacting the coefficient estimates for variables in the MLR model. This is called ***Omitted Variable Bias/Impact***, or ***Endogeneity*** for short. And it's probably the second most important concept in econometrics, after ***Data Integrity***.
  - ***Find the data!*** Of course, you could always try to find data for that excluded variable … and incorporate that data into the analysis… and see what happens. So don't be lazy!
  - ***Pencils down… But when?*** When do you put your pencil down? At what point do you *call it a wrap* (Your analysis is credible, useful and worthy of attention!)? or *throw in the towel* (There's no hope; the MLR analysis just isn't *working out*!)?

Introduction: MLR v. SLR Models cont'd

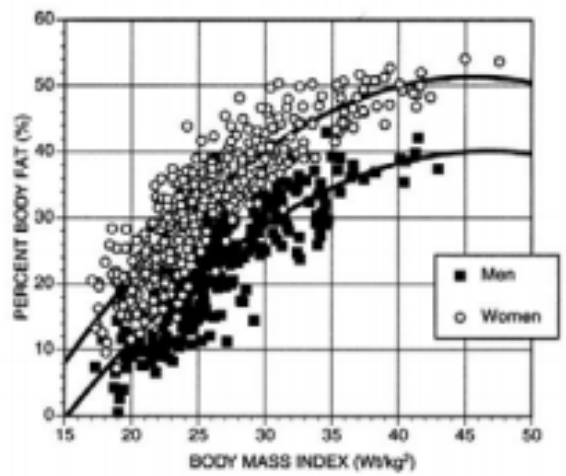
- **Bodyfat and the Body Mass Index (BMI)**

  SLR: regressed *Brozek* measure of *bodyfat* on *BMI*

  MLR: add BMI^2 to the model to allow for non-linear effects (see Figure, right)… add additional personal characteristics, or maybe break *BMI* apart… *abd*, *wgt*, *hgt*, *abd/hgt*, etc.

Figure 1

From: The effect of sex, age and race on estimating percentage body fat from body mass index: The Heritage Family Study



Non-linear plot of the relationship between BMI and measured percentage body fat of the male and female Heritage data. The quadratic regression equations are: women, $Y'_{(\%fat)}=(4.35\times BMI)-(0.05\times BMI^2)-46.24$, ($r^2=0.78$, s.e.e.=4.63%); and men, $Y'_{(\%fat)}=(3.76\times BMI)-(0.04\times BMI^2)-47.80$, ($r^2=0.68$, s.e.e.=4.90%).

- **S&P's Sovereign Debt Ratings and www.transparency.org's Corruption Perception Index**

  SLR: regressed *NSRate* on *corrupt*

  MLR: add additional macro-economic variables such as *gdp*, *inflation*, *debt/gdp*, *deficit/gdp*, etc. … add regional variables (e.g. *EU*)… etc.

- **The Pythagorean Theorem in Baseball (sort of)**

  SLR: regressed *%wins* on *RS/RA*

  MLR: add $(RS / RA)^2$ to the model to allow for non-linear effects… add variables that capture other factors the drive wins/losses… managerial quality, bullpen quality, etc.

Introduction: MLR v. SLR Models cont'd

- **Estimating *Beta* in the *CAPM***

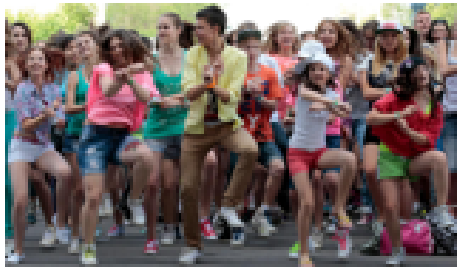
  SLR: regressed the security's returns on the market's returns

  MLR: add additional finance/macroeconomic variables to the RHS… *inflation*, *GNP*, *yield curve*, *short-term interest rates*, *oil prices*, *gold prices*, *exchange rates*, etc. … with additional RHS variables we have what is called *Arbitrage Pricing Theory* (APT) analysis

- **Election Hacking in Wisconsin**

  SLR: regressed *ClintonShift* on *paper*

  MLR: add additional control variables such as *income*, *education*, *race*, *prior voting behavior*, etc.



- ***Alexa, Take me to Funkytown***

SLR: regressed *pkstreams* on *danceability*

MLR: add in all the other EchoNest audio feature metrics, as well as genre, duration, release year, etc, etc.

MLR and SLR Results: Forecasting Box Office Revenues

See Any Differences?

SLR: wk1 on the RHS (reg rtotgross wk1)

```
      Source |       SS           df       MS      Number of obs   =     9,114
-------------+----------------------------------   F(1, 9112)      =  31044.73
       Model |    23727348         1    23727348   Prob > F        =    0.0000
    Residual |  6964261.57     9,112  764.295607   R-squared       =    0.7731
-------------+----------------------------------   Adj R-squared   =    0.7731
       Total |  30691609.6     9,113  3367.89308   Root MSE        =    27.646

------------------------------------------------------------------------------
   rtotgross |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wk1 |   2.354437   .0133627   176.20   0.000     2.328243    2.380631
       _cons |   4.432582     .32052    13.83   0.000     3.804291    5.060873
------------------------------------------------------------------------------
```

See Any Differences?

MLR: wk1 and wk2 on the RHS (reg rtotgross wk1 wk2)

```
      Source |       SS           df       MS      Number of obs   =     9,114
-------------+----------------------------------   F(2, 9111)      =  35773.10
       Model |  27224699.6         2  13612349.8   Prob > F        =    0.0000
    Residual |     3466910     9,111  380.519153   R-squared       =    0.8870
-------------+----------------------------------   Adj R-squared   =    0.8870
       Total |  30691609.6     9,113  3367.89308   Root MSE        =    19.507

------------------------------------------------------------------------------
   rtotgross |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wk1 |  -.0120343   .0264237    -0.46   0.649    -.0638307     .039762
         wk2 |   4.536046   .0473147    95.87   0.000     4.443298    4.628793
       _cons |   .4006355   .2300356     1.74   0.082     -.050286    .8515569
------------------------------------------------------------------------------
```

See Any Differences?

OLS: A Quick Comparison of SLR and MLR Analytics

Analysis	**SLR**	**MLR**
Linear Model	$y_i = \beta_0 + \beta_1 x_i + u_i$	$y_i = \beta_0 + \beta_x x_i + \beta_z z_i + u_i$ $y_i = \beta_0 + \sum_j \beta_j x_{ij} + u_i$
Residuals/Unexplained	$residual_i = y_i - (b_0 + b_1 x_i)$ $SSR = \sum residual_i^2$	$residual_i = y_i - (b_0 + b_x x_i + b_z z_i)$ $SSR = \sum residual_i^2$
OLS estimates	Min SSRs wrt b_0 and b_1	Min SSRs wrt b_0, b_x and b_z
OLS Estimates: … intercept:	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_x \bar{x} + \hat{\beta}_z \bar{z})$

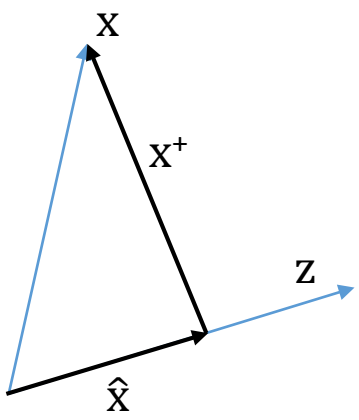
A Quick Comparison cont'd

Analysis	**SLR**	**MLR**
OLS Estimates: … slopes	$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \rho_{xy}\frac{S_y}{S_x}$	Complicated… $\hat{\beta}_x = \frac{S_{x^+y^+}}{S_{x^+x^+}} = \rho_{x^+y^+}\frac{S_{y^+}}{S_{x^+}}$ (+: effects of other RHS variables have been *partialed out*)
… sign (slope):	$sign\left(\hat{\beta}_1\right) = sign\left(\rho_{xy}\right)$, where ρ_{xy} is the correlation of the x's and y's	$sign\left(\hat{\beta}_x\right) = sign\left(\rho_{x^+y^+}\right)$, where $\rho_{x^+y^+}$ is the *partial* correlation of the x's and y's
SRF (Sample Regression Function)	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z$
… @ means	$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$	$\hat{\beta}_0 + \hat{\beta}_x \bar{x} + \hat{\beta}_z \bar{z} = \bar{y}$

A Quick Comparison cont'd

Analysis	**SLR**	**MLR**
Predicteds, actuals and residuals	$y_i = \hat{y}_i + \hat{u}_i;\ avg(\hat{y}'s) = \bar{y}\ ;$ $avg(\hat{u}'s) = 0\ ;\ corr(\hat{y}'s, \hat{u}'s) = 0$	$y_i = \hat{y}_i + \hat{u}_i;\ avg(\hat{y}'s) = \bar{y}\ ;$ $avg(\hat{u}'s) = 0\ ;\ corr(\hat{y}'s, \hat{u}'s) = 0$
Estimated Impact … from changing one RHS var	$\frac{d\hat{y}}{dx} = \hat{\beta}_1$ $\Delta\hat{y} = \hat{\beta}_1 \Delta x \Leftrightarrow \frac{\Delta\hat{y}}{\Delta x} = \hat{\beta}_1$	$\frac{\partial\hat{y}}{\partial x} = \hat{\beta}_x$ *(ceteris paribus)* $\Delta\hat{y} = \hat{\beta}_x \Delta x \Leftrightarrow \frac{\Delta\hat{y}}{\Delta x} = \hat{\beta}_x$
… from changing several RHS vars		$\Delta\hat{y} = \hat{\beta}_x \Delta x + \hat{\beta}_z \Delta z$
Beta Regressions	standardize all variables: y^*, x^* regress y^* *on* x^*	standardize all variables: y^*, x^*, z^* regress y^* *on* x^* *and* z^*
	$\hat{\beta}_0 = 0;\ \hat{\beta}_1 = \rho_{x^*y^*} = \rho_{xy}$ is the correlation of x *and* y	$\hat{\beta}_0 = 0;\ \hat{\beta}_{x^*}$ is the *partial* correlation of x^* *and* y^*
Elasticities (at the means)	$\frac{d\hat{y}}{dx}\left[\frac{x}{\hat{y}}\right]_{x=\bar{x}} = \hat{\beta}_1 \frac{\bar{x}}{\bar{y}}$	$\frac{\partial\hat{y}}{\partial x}\left[\frac{x}{\hat{y}}\right]_{@means} = \hat{\beta}_x \frac{\bar{x}}{\bar{y}}$

SideTrip: *WhatsNew* and *WhatsLeft*



- $x^{+}\left(WhatsNew_{x}\right)$: x^{+}, or *What's New* about the RHS variable *x*, is the residual from the regression of *x* on the other RHS variables in the MLR model.
  - It's the part of x *not explained* by the other RHS variables in the model.
- $y^{+}\left(WhatsLeft_{y}\right)$: y^{+}, or *What's Left* of the LHS variable *y*, is the residual from the regression of the LHS variable *y* on the other RHS variables (other than *x*) in the MLR model.
  - It's the part of y *not explained* by the other RHS variables in the model.
- The new variables x^{+} *and* y^{+}, are what you have after you have *partialed out* the effects of the other RHS variables in the MLR model.
- The *partial* correlation of x and y.is the correlation between $y^{+}\left(WhatsLeft_{y}\right)$ and $x^{+}\left(WhatsNew_{x}\right)$.

Interpretation of estimated OLS/MLR slope coefficients

- **SRFs and Marginal Predicted Effects (ceteris paribus)**
  - The estimated MLR slope coefficients provide estimates of average incremental effects/relationships, *ceteris paribus (all else the same).*
- **Partial Correlations (correlations controlling for the other RHS variable in the model)**
  - The OLS/MLR coefficient for, say variable x, is defined by: $\hat{\beta}_x = \frac{S_{x^+y^+}}{S_{xvx^+}} = \rho_{x^+y^+} \frac{S_{y^+}}{S_{x^+}}$,

    where $\rho_{x^+y^+}$ is the *partial correlation* between the x's and y's.
  - Accordingly, MLR coefficients can be found in an SLR model in which the effects of the other RHS variables have been *partialed out.*
  - The sign of the x coefficient in an MLR model agrees with the sign of the partial correlation between x and y... which may or may not be the same as the sign of the <u>simple</u> correlation between x and y.

Slope coefficients in SLR models capture correlations;
Slope coefficients in MLR models capture <u>partial</u> correlations.

Endogeneity (Omitted Variable Bias/Impact) I: An Overview

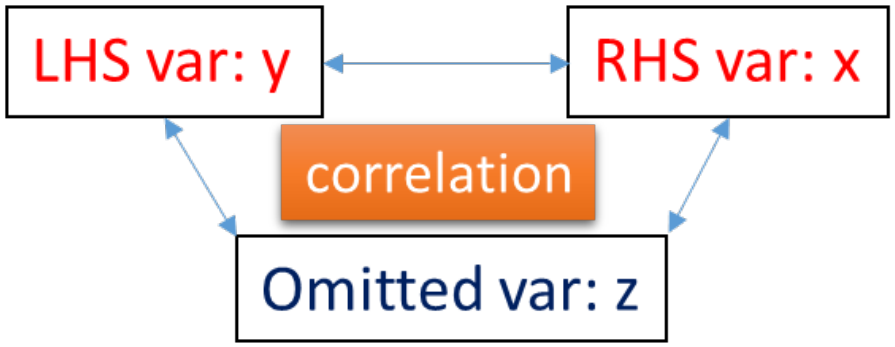
Explanatory variables left out of the MLR analysis could be impacting the coefficient estimates for variables in the MLR model. ***This is called Omitted Variable Bias/Impact, or Endogeneity for short. And it's probably the second most important concept in econometrics (after Data Integrity).***

You should lose sleep worrying about how excluded variables might be impacting your coefficient estimates and biasing your conclusions. If you're lucky, you might be able to sign the bias… and say whether or not your estimated coefficient is biased up or down. Let's hope you are so lucky!

Yes, endogeneity really is that important. You never really know whether or not your estimated coefficients have been biased (or less pejoratively, *impacted*) by omitted variables. So don’t be lazy! Bring lots of potential explanatory variables to the analysis and see what happens. It's the best you can do.

Endogeneity Overview cont'd

For the moment, assume that the estimated SLR model has just one explanatory variable, x, and that potential RHS variable z has been excluded (omitted) from the estimated model.



The omitted variable bias/impact (endogeneity) associated with the exclusion of z from the estimated model is typically thought to be driven by two factors:

- The correlation of the excluded variable z with the RHS variable in the model, x.
- The correlation of the excluded variable z with the LHS variable in the model, y.

And the direction of the impact/bias is **determined by the signs of these correlations**:

- *Positive Omitted Variable Bias/Impact:* If both correlations are positive, then OLS estimated coefficient will be biased upwards by the omission of the excluded variable from the analysis; the OLS estimated coefficient for x will be greater than it would be otherwise.
- *Negative Omitted Variable Bias/Impact:* If one correlation is positive and the other negative, then the bias will be downward; the OLS estimated coefficient for x will be less than it would be otherwise.

What to do if you fear endogeneity (omitted variable bias/impact)

- ***Don't be lazy!*** Get the data and include it in your model… and see what happens.
- ***Proxy variable?*** But maybe you can't get the data. Then maybe use an available proxy variable which is highly correlated with the omitted variable. Or try several proxy variables and see if it matters.
- ***IV's?*** And if you are really lazy and don't want to find proxies, try the *oh so sophisticated* ***Instrumental Variables*** approach… which we'll discuss later in the semester. But only if you are really really lazy! (Yes, you see my bias!)
- ***Sign the impact/bias?*** And if you can't do any of the above, then as a last resort you might try to *sign* the bias and determine whether the estimated model over- or under-estimates the MLR coefficient estimates (relative to a model in which the omitted variable(s) is included in the analysis).

OLS/MLR Analytics: *TakeAways*

- MLR analysis looks a lot like SLR analysis; the one major difference being that you can now have many RHS explanatory variables… or put differently, control for the impacts of additional explanatory factors
- The OLS estimation is as before, the one major difference being that while SLR slope coefficients reflect simple correlations, the MLR slope coefficients now reflect partial correlations
- The partial correlation between, say, x and y, is a simple correlation in which the effects of the other RHS variables have been partialed out. Put differently, it's the simple correlation between WhatsNew about x (after the effects of the other RHS vars have been partialed out) and WhatsLeft of y (after accounting for what is explained by the other RHS variables).
- MLR coefficients also capture the average incremental relationship between a RHS variable and y, holding all of the other variables in the model constant (ceteris paribus).
- Endogeneity (omitted variable impact/bias) refers to the change in estimated MLR coefficients when RHS variables are excluded/omitted for the model. The direction and magnitude of the effect is typically driven by the correlation between the excluded variable and LHS and RHS variables left in the model.
- Possible endogeneity remedies include: 1) finding the missing data; 2) using proxy variables; 3) using Instrumental Variables (IVs), and perhaps 4) signing the bias/impact.

onwards… to *MLR Assessment*