OLS/MLR Analytics I – *What’s New? Not Much!*

- ***Introduction: MLR v. SLR***
- ***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

The difference between MLR and SLR models is really very simple: with SLR models you have a single explanatory (RHS) variable… and with MLR models you can have more than just one explanatory variable. That's the difference. That's it!

But once you allow more than just one explanatory variable, the world becomes a lot more interesting… and a lot more complicated. For example:

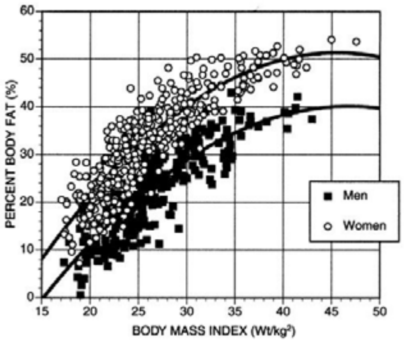
1. ***In or out?*** How do you decide which RHS variables to include in your analysis and which ones to leave out?
2. ***Interactions (w/in & w/out):*** And once you allow for multiple explanatory variables, you'll have to worry about how they interact with one another and with the dependent variable as well. And this applies to variables included in the analysis as well as the excluded/omitted variables.
3. ***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).*** 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!
4. ***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! *Just do it!*
5. ***Pencils down… But when?*** When do you put your pencil down? At what point do you *call it a wrap* (have a MLR analysis that you think is credible, useful and worthy of attention)? or *throw in the towel* (and conclude that there's no hope and the MLR analysis just isn’t working out)?

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

Earlier, we looked at a number of SLR models with single explanatory variables. Here are some examples of how you might move those models to MLR analyses, with additional RHS variables bringing additional explanatory power to the model:

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%).

- **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.

- **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.

- **Predicting Lifetime Movie Revenues**

  SLR: regressed *rtotgross* on *wk1* revenues

  MLR: add weekly revenue data from weeks 2, 3,… (we will do this below)… and perhaps add in movie ratings data (from critics as well as viewers) from *RottenTomatoes, IMDb,* etc.

- **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.

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

- ***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.

There is inevitably no shortage of candidates for RHS variables in MLR models. In the end you may return to your original SLR model as your model of choice… but you'll certainly want to explore a bunch of MLR models to see if they have more explanatory power. And they often do!

Comparing SLR and MLR Results: *Forecasting Box Office Revenues*

What better way to introduce you to MLR models than to show you MLR analysis in action? To do that, let's return to the challenge of forecasting lifetime movie revenues using weekly box office revenues. Previously you saw that if you had to focus on just one week, you'd want to pick week 3 revenue data. But what if you could as well add in revenue data from other weeks?

The following shows the impact of adding wk2 revenues to an original SLR model in which *rtotgross* was regressed on *wk1*, week 1 box office revenues

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
------------------------------------------------------------------------------
```

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
------------------------------------------------------------------------------
```

Look closely... see any differences?

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

MLR and SLR results are very very similar ... the format is basically unchanged! The only difference in format is the new line with *wk2* results (*Coef.*, *Std. Err.* etc at the bottom of the output), since *wk2* was added into the previous SLR model.

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	$residual_i = y_i - \left(b_0 + b_1 x_i\right)$ $SSR = \sum residual_i^2$	$residual_i = y_i - \left(b_0 + b_x x_i + b_z z_i\right)$ $SSR = \sum residual_i^2$
OLS estimates	Min SSRs wrt b_0 and b_1	Min SSRs wrt b_0, b_x and b_z
OLS Estimates: … intercept:	$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$	$\hat{\beta}_0 = \overline{y} - \left(\hat{\beta}_x \overline{x} + \hat{\beta}_z \overline{z}\right)$
… slopes	$\hat{\beta}_1 = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$ $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \rho_{xy} \dfrac{S_y}{S_x}$	Complicated: $\hat{\beta}_x = \dfrac{S_{x^*y^*}}{S_{x^*x^*}}$ $= \rho_{x^*y^*} \dfrac{S_{y^*}}{S_{x^*}}$
… 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 ρ_{xy} 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	$\hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y}$	$\hat{\beta}_0 + \hat{\beta}_x \overline{x} + \hat{\beta}_z \overline{z} = \overline{y}$
Predicteds, actuals and residuals	$y_i = \hat{y}_i + \hat{u}_i$; $avg(\hat{y}'s) = \overline{y}$; $avg(\hat{u}'s) = 0$; $corr(\hat{y}'s, \hat{u}'s) = 0$	$y_i = \hat{y}_i + \hat{u}_i$; $avg(\hat{y}'s) = \overline{y}$; $avg(\hat{u}'s) = 0$; $corr(\hat{y}'s, \hat{u}'s) = 0$

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

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		$\Delta\hat{y}=\hat{\beta}_x\Delta x+\hat{\beta}_z\Delta z$

Analysis	**SLR**	**MLR**
Elasticities (at the means)	$\frac{d\hat{y}}{dx}\left[\frac{x}{\hat{y}}\right]_{x=\bar{x}}=\hat{\beta}_1\frac{\bar{x}}{\bar{y}}$	$\frac{\partial\hat{y}}{\partial x}\left[\frac{x}{\hat{y}}\right]_{@\,means}=\hat{\beta}_x\frac{\bar{x}}{\bar{y}}$

So what's new with MLR?... Not much, really!

Interpreting MLR Coefficients I

SRFs and Marginal Predicted Effects (ceteris paribus)

With SLR models, the estimated slope coefficient in the SRF captures the estimated/predicted marginal impact (*ceteris paribus*) from changes in the RHS variable, and provides estimates of changes in the predicted values associated with given changes in the RHS variable. That interpretation of estimated coefficients extends to MLR models, with one modification. In MLR models, the estimated slope coefficients provide estimates of average incremental effects/relationships, *ceteris paribus (all else the same).*

Recall the SRF from Model (2) above: $\boxed{\hat{y}=.401-.012wk1+4.536wk2}$.

Effects at the margin: Given the SRF interpretation of the estimated MLR coefficients, the estimated coefficents tell us how predicted *rtotgross* will change (on average) given incremental changes in individual RHS variables, *ceteris paribus*:

- $\frac{\partial\hat{y}}{\partial x_1}=-0.012<0$ ... So holding *wk2* revenues fixed, an increase in *wk1* revenues of $1M is on average associated with a reduction in lifetime revenues of $12K.

  Does that makes sense to you? ... did you expect to find that the incremental/marginal effects is negative?

- $\frac{\partial\hat{y}}{\partial x_2}=4.536>0$ ... So holding *wk1* revenues fixed, an increase in *wk2* revenues of $1M is on average associated with an increase in lifetime revenues of $4.56M.

Again: Does that sound right to you?

Discrete effects: The MLR coefficients also tell you something about the average predicted effects associated with discrete changes in the RHS variables.

Continuing with the example above: Suppose that *wk1* and *wk2* revenues were each higher by $1M. Then the estimated SRF predicts that lifetime revenues would increase of -$12K+$4.56M = $4.55M.

Partial Correlations

With SLR models, the estimated slope coefficient is defined by $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \rho_{xy}\frac{S_y}{S_x}$, where ρ_{xy} is the correlation between the x's and y's. We have almost identical definitions for the case of MLR models, with one very important and significant difference:

The OLS/MLR coefficient for, say variable x, is now defined by: $\hat{\beta}_x = \frac{S_{x^*y^*}}{S_{x^*x^*}} = \rho_{x^*y^*}\frac{S_{y^*}}{S_{x^*}}$, where $\rho_{x^*y^*}$ is the *partial correlation* between the x's and y's.

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.

And so: MLR coefficients capture the correlation between x and y after the effects of the other RHS variables have been removed from those two variables, x and y.

We will go through this in much greater detail later, but for now:

- $x^*\left(WhatsNew_x\right)$: x*, or *What's New* about the RHS variable *x*, is the residual from the collinearity 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 *partial* correlation of x and y.is the correlation between $WhatsLeft_y$ and $WhatsNew_x$.

And the estimated OLS/MLR slope coefficients:

- The MLR estimated slope coefficient for any RHS variable, say *x*, can be derived by regressing the dependent variable y on *What'sNew*$_x$.
- It can also be derived by regressing $WhatsLeft_y$ on $WhatsNew_x$.
- Notice that in both cases, you are estimating an SLR model… and so all MLR coefficients can be generated by an appropriate SLR model.

Bottom Line: *Slope coefficients in SLR models capture correlations; slope coefficients in MLR models capture partial correlations.*

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

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

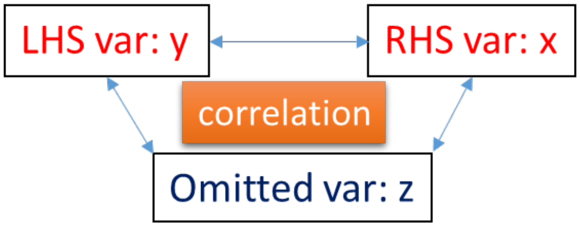
From the start of this section:

> 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.

We'll review this topic in much greater detail in the next section. At the moment, though, it's useful to develop some intuition for what drives endogeneity. The following is not precisely correct, but close enough to the truth to be useful.

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

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

Positive Omitted Variable Bias/Impact: If both correlations are positive, then OLS estimated coefficients will be biased upwards (by the omission of the excluded variable from the analysis), so that the estimated coefficients will be greater than they would be otherwise (had the excluded RHS variable been in the model). (If both correlations are negative then, as discussed later, the bias is also positive.)

Negative Omitted Variable Bias/Impact: If one correlation is positive and the other negative, then the bias will be downward... so that the OLS estimated coefficients will be less than they would be otherwise.

Examples: Endogeneity (Omitted Variable Bias/Impact)

Here are a couple examples using the *movierevs* dataset, and the MLR models discussed above. Weekly revenues *wk1* and *wk2* are both included in the *Full Model*, Model (1). In Model (2), *wk2* revenues have been dropped/omitted from the Full Model, Model (1), and in Model (3), *wk1* revenues have been dropped/omitted from Model (1).

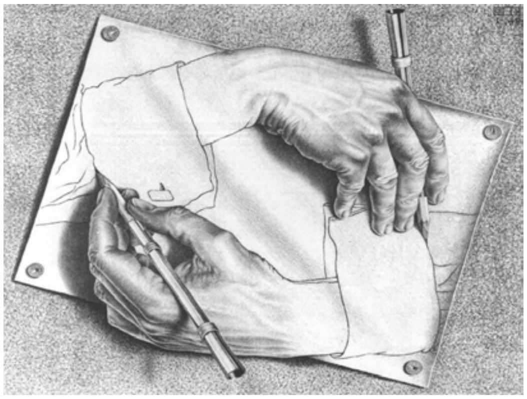
```
------------------------------------------------------------
                      (1)             (2)             (3)   
                rtotgross       rtotgross       rtotgross   
------------------------------------------------------------
wk1               -0.0120           2.354***                
                  (-0.46)        (176.20)                   

wk2                 4.536***                        4.516***
                  (95.87)                        (267.49)   

_cons               0.401           4.433***        0.403   
                   (1.74)         (13.83)          (1.75)   
------------------------------------------------------------
N                    9114            9114            9114   
R-sq                0.887           0.773           0.887   
------------------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

- When *wk2* is omitted/dropped from Model (1), the estimated *wk1* coefficient increases from -.012 (Model (1)) to 2.354 (Model (2)), an increase of 2.366.

  In this case we say that there is ***positive*** omitted variable bias/impact, since the estimated *wk1* coefficient *increased* when *wk2* was dropped/omitted from Model (1), the *Full Model*. Or put differently: the *wk1* coefficient was biased upwards by the omission/exclusion of *wk2* from the MLR model.

- When *wk1* is omitted/dropped from Model (1), the estimated *wk2* coefficient now decreases slightly from 4.536 (Model (1)) to 4.516 (Model (3)), a decrease of 0.020.

  In this case we say that there is ***negative*** omitted variable bias/impact, since the estimated *wk1* coefficient decreased when *wk1* was dropped/omitted from Model (1), the *Full Model*. Or put differently: the *wk2* coefficient was biased downwards by the omission/exclusion of *wk1* from the MLR model (1).



Misleading? It's perhaps misleading to say that endogeneity leads to *biased* estimated coefficients. The estimated coefficients reflect the incremental average relationship between changes in the particular explanatory variable and changes in the dependent variable, controlling for all of the other variables in the model. But of course, the omitted variable is not in the model. When the RHS variables change, so do the *ceteris paribus* conditions… what is being held fixed is changing. So no one should be surprised to see estimated coefficients change when explanatory variables are added to, or subtracted from, MLR models. And don't pejoratively call it *bias*… just call it a *different model*.

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

1. ***Don't be lazy!*** Get the data and include it in your model… and see what happens.
2. ***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. (Example: If you don't have data on disposable personal income by MSA, use median per capita income as a proxy, or maybe median housing sales prices, or median monthly rent data, or …)
3. ***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!)
4. ***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).

Stay tuned for more details… but in the meanwhile: If you're lucky, then you might be able to say something like:

> *I estimated a positive effect/coefficient for RHS variable x in my MLR analysis. I know that I have an issue with omitted variable bias… but since I'm confident that that bias is negative, I know that the true x effect is even larger than what I've estimated… and so I'm confident that there really is a positive relationship, and if anything, I've underestimated its magnitude!*

But of course, if the omitted variable bias is positive, then you know that you've overestimated the effect, and now maybe you aren't so sure that the actual effect is positive, or very sizable… it could just be omitted variable bias/impact driving the result.

We'll have a more complete treatment of endogeneity in the next section.