OLS/MLR Analytics and Assessment: *Review*

This OLS/MLR review assumes that you have already taken a look at the OLS/SLR Analytics/Assessment review, and is accordingly based on what is new and different with MLR analysis. You may want to revisit the OLS/SLR *Review* to refresh your recollection.

This review is somewhat repetitive… but I hope that's a good thing!

Let's work with the *bodyfat* dataset (feel free to follow along in Stata… use *bcuse bodyfat* to access the data). In the full MLR model, *brozek* has been regressed on *hgt*, *wgt* and *hip*; the *hip* variable has been dropped in the second MLR model; and the third model is a collinearity regression in which *hip* has been regressed on the two surviving variables (*hgt* and *wgt*):

Full Model

```
      Source |       SS           df       MS      Number of obs   =       252
-------------+----------- dofs ---------------    F(3, 248)       =     71.25
       Model |      6,980         3  2326.69      Prob > F        =    0.0000
    Residual |      8,099       248   32.657      R-squared       =    0.4629
-------------+----------------------------------   Adj R-squared  $\bar{R}^2$  0.4564
       Total |     15,079       251   60.076      Root MSE        =    5.7146

------------------------------------------------------------------------------
      brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |     -.6164      .1115     -5.53   0.000       -.8360      -.3968
         wgt |      .1552      .0404      3.84   0.000        .0756       .2349
         hip |      .1314      .1601      0.82   0.412       -.1839       .4468
       _cons |     21.268      13.89      1.53   0.127       -6.087      48.624
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
         wgt |     10.85    0.0922
         hip |     10.11    0.0989
         hgt |      1.28    0.7802
-------------+----------------------
    Mean VIF |      7.41
```

… *hip* dropped from the Full Model

```
      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(2, 249)       =    106.67
       Model |      6,958         2    3479.03   Prob > F        =    0.0000
    Residual |      8,121       249     32.614   R-squared       =    0.4614
-------------+----------------------------------   Adj R-squared   =    0.4571
       Total |     15,079       251     60.076   Root MSE        =    5.7109

------------------------------------------------------------------------------
      brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |     -.6503      .1035    -6.29   0.000       -.8541      -.4466
         wgt |      .1867      .0129    14.48   0.000        .1613       .2121
       _cons |     31.155      6.913     4.51   0.000       17.539      44.771
------------------------------------------------------------------------------
```

Collinearity Regression

```
      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(2, 249)       =     1,134
       Model |     11,608          2   5804.00     Prob > F        =    0.0000
    Residual |      1,274        249    5.1175     R-squared   R_j^2    0.9011
-------------+----------------------------------   Adj R-squared   =    0.9003
       Total |     12,882        251   51.3237     Root MSE        =    2.2622

------------------------------------------------------------------------------
         hip |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |    -.2586      .0410    -6.31   0.000      -.3393      -.1779
         wgt |     .2393      .0051    46.85   0.000       .2292       .2494
       _cons |    75.231      2.738    27.47   0.000      69.838      80.625
------------------------------------------------------------------------------

. summ Brozek hgt wgt hip

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      Brozek |        252       18.94       7.751          0       45.1
         hgt |        252       70.15       3.663       29.5      77.75
         wgt |        252      178.92      29.389      118.5     363.15
         hip |        252       99.90       7.164         85     147 .7
```

1) Highlighted figures in previous regression models

   a) ***dof***: degrees of freedom are now $n-k-1=252-3-1=248$, where $n=\#obs$ and $k=\#RHS\ vars$

   b) ***adjusted R²***: $\bar{R}^2=1-\dfrac{SSR}{SST}\dfrac{n-1}{n-k-1}=1-\dfrac{8,099}{15,079}\dfrac{251}{248}=.4564$

   $=1-\dfrac{MSE}{S_{yy}}=1-\dfrac{32.657}{15,079/251}=.4564$ … R^2 is modified so that RHS variables don't get credit for *just showing up;* $\bar{R}^2<R^2\leq 1$; moves in opposition to *MSE/RMSE*

   c) ***multicollinearity (hip)*** $\left(R_j^2\right)$: R^2 from the collinearity regression; can also be calculated using the Variance Inflation Factor, $VIF_x=\dfrac{1}{1-R_x^2}=\sqrt{\dfrac{1}{1-.9011}}=1.28$

   d) ***endogeneity*** (omitted variable impact/bias): illustrated by the change in the estimated *wgt* coefficient when *hip* is dropped from the Full Model… product of the *hip* coefficient in the Full Model and the *wgt* coefficient in the collinearity regression:

   $\Delta\hat{\beta}_{wgt}=.1867-.1552=.1314\cdot.2393=.03145$

2) What's new and different since OLS/SLR Analytics and Assessment? … Not much![1] Here are the main differences:
   a) Analytics
      i) Estimated coefficients: For SLR models, the formulas for the estimated OLS coefficients are fairly simple; for MLR models, they are more complicated.
      ii) Collinearity
         (1) Impacts/factors
            (a) One of the factors in omitted variable impact/bias (endogeneity)
            (b) Affects SRF interpretation of OLS coefficients… sort of
            (c) Impacts standard errors (precision of estimation)… a concept that will arrive later
            (d) Can lead to wacky results (don't make the mistake of tossing important RHS variables just because they were highly collinear with one another)
            (e) Explanatory power: less collinear RHS variables have the potential for more independent explanatory power… because they are more independent from the other RHS variables
         (2) Metrics
            (a) R-sq from collinearity regression $\left(R_j^2\right)$
               (i) captures extent to which a particular RHS var can be explained (predicted) by the other RHS variables
               (ii) logical extension of the concept of correlation to sets of more than two variables
            (b) Variance Inflation Factor (VIF): $VIF_x = \frac{1}{1-R_x^2}$ (easier way to generate the R_j^2's.
      iii) Endogeneity (Omitted Variable Impact/Bias): extent to which OLS estimated coefficients are impacted by the exclusion of explanatory (RHS) variables from the model
         (1) What drives that impact: The product of…
            (a) OLS coefficient of the omitted variable when it's in the Full model
            (b) OLS coefficients of surviving variables (left in the model) in the collinearity regression in which the omitted variable is regressed on the surviving variables. From the notes:

[1] Warning: Some of this is a bit repetitive with the preceding… but Hey, why not? … It's a review!

- **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.: $\hat{\alpha}_x \hat{\beta}_z$)

		z coeff. in the **MLR Full Model** (SRF$_y$)		
x coeff. in 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

(2) What to do about it?

(a) Don't be lazy… grab the data and see what the impact is.

(b) If you can't get the data, maybe try using some proxy variables?

(c) And if you can't find proxy variables, maybe try the IV (Instrumental Variable) approach… but be careful, as it can be quite squishy!

(d) And if all else fails, maybe you can qualitatively evaluate the sign/direction of the impact (thinking about signs of coefficients … see above)

iv) What's New? … and What's Left?

(1) ***WhatsNew$_x$***: the residuals when the RHS variable *x* is regressed on the other RHS variables… captures the part of *x* not explained by the other RHS variables

(2) ***WhatsLeft$_y$***: the residuals when the LHS variable *y* is regressed on the RHS variables other than *x*… captures the part of *y* not explained by the other RHS variables (other than *x*)

(3) The x coefficient from the MLR model, $\hat{\beta}_x$, can also be generated by two SLR models:

(a) reg y WhatsNew$_x$ … $\hat{\beta}_x = corr(y, WhatsNew_x)\dfrac{S_y}{S_{WhatsNew_x}}$

(b) reg WhatsLeft$_y$ WhatsNew$_x$ … $\hat{\beta}_x = corr(WhatsLeft_y, WhatsNew_x)\dfrac{S_{WhatsLeft_y}}{S_{WhatsNew_x}}$

(c) And so the sign of $\hat{\beta}_x$, agrees with the sign of the two correlations just discussed.

(d) $corr(WhatsLeft_y, WhatsNew_x)$ is a *partial* correlation… where the effects of the other RHS variables have been *partialed out*, prior to calculating the correlation.

b) Assessment

i) R-sq is of limited usefulness in evaluating MLR models, since it never declines when RHS variables are added to the model (and typically increases… unless the coefficient for the new variable is zero, or the new variable is perfectly collinear with the other RHS variables)

ii) Degrees of freedom: $dofs = n - k - 1$ (n obs and k RHS vars)

iii) Adjusted R-sq doesn't merely give new RHS variables credit for just showing up… adj R-sq only increases if the drop in SSRs exceeds some minimum level:

$$\bar{R}^2 = 1 - \frac{SSR}{SST}\frac{n-1}{n-k-1} < 1 - \frac{SSR}{SST} = R^2 \leq 1$$

(1) When adding and subtracting RHS variables, $\bar{R}^2$ moves in opposite direction from MSE/RMSE (assuming S_{yy} fixed), since $\bar{R}^2 = 1 - \frac{MSE}{S_{yy}}$

iv) When dofs are changing, we often pick between models based on adj R-sq, among other factors.

3) Estimated OLS/MLR coefficients, SRFs and elasticities

(Even more repetitive of the prior material… but again, maybe helpful.)

a) *OLS*: Minimize $SSR = \sum (u_i)^2 = \sum \left(brozek_i - \left(b_0 + b_{hgt} hgt_i + b_{wgt} wgt_i + b_{hip} hip_i \right) \right)^2$ *wrt* $b_0,\ b_{hgt},\ b_{wgt}\ and\ b_{hip}$ (FOCs and SOCs)

b) slope coefficients (*hgt*, *wgt* and *hip*):

i) $\hat{\beta}_{hgt} = -.616,\ \hat{\beta}_{wgt} = .155\ and\ \hat{\beta}_{hip} = .131$

ii) formulas are complicated; but coefficients can be generated by regressing *y*'s (or *WhatsLeft* of *y*'s) on *WhatsNew* about *x*'s

c) Intercept coefficient (_*cons*): $\hat{\beta}_0 = \bar{y} - \left(\hat{\beta}_{hgt} \overline{hgt} + \hat{\beta}_{wgt} \overline{wgt} + \hat{\beta}_{hip} \overline{hip} \right)$

$$= 18.94 + \left(-.616(70.15) + .115(178.92) + .131(99.90) \right) = 21.27$$

d) SRF (Sample Regression Function; *predicteds*): $\hat{y} = \hat{\beta}_0 + \left(\hat{\beta}_{hgt} hgt + \hat{\beta}_{wgt} wgt + \hat{\beta}_{hip} hip \right)$

$$\hat{y} = 21.27 + \left(-.616\, hgt + .155\, wgt + .131\, hip \right)$$

i) average marginal effects: $\frac{\partial \hat{y}}{\partial hgt} = \hat{\beta}_{hgt} = -.616$; $\frac{\partial \hat{y}}{\partial wgt} = \hat{\beta}_{wgt} = .155$; $\frac{\partial \hat{y}}{\partial hip} = \hat{\beta}_{hip} = .131$

ii) elasticity @means:[2] $\varepsilon_x = \frac{\partial \hat{y}}{\partial x}\frac{\overline{x}}{\overline{y}}$, and so…

(1) $\varepsilon_{hgt} = \frac{\partial \hat{y}}{\partial hgt}\frac{\overline{hgt}}{\overline{y}} = \hat{\beta}_{hgt}\frac{\overline{hgt}}{\overline{y}} = -.616\frac{70.15}{18.94} = -2.28$

(2) $\varepsilon_{wgt} = \frac{\partial \hat{y}}{\partial wgt}\frac{\overline{wgt}}{\overline{y}} = \hat{\beta}_{wgt}\frac{\overline{wgt}}{\overline{y}} = .155\frac{178.92}{18.94} = 1.47$

(3) $\varepsilon_{hip} = \frac{\partial \hat{y}}{\partial hip}\frac{\overline{hip}}{\overline{y}} = \hat{\beta}_{hip}\frac{\overline{hip}}{\overline{y}} = .131\frac{99.90}{18.94} = .69$

(4) … can also generate using the Stata *margins* command:

margins, eyex(_all) atmeans

4) *Goodness of Fit* metrics: MSE/RMSE, R^2 and $\overline{R}^2$

a) Degrees of freedom (dofs): $dofs = n - k - 1 = 252 - 3 - 1 = 248$

b) (Root) Mean Squared Error: $MSE = \frac{SSR}{n-k-1} = \frac{8,099}{248} = 32.657$, and

$RMSE = \sqrt{MSE} = \sqrt{\frac{SSR}{dofs}} = \sqrt{32.657} = 5.7146$

c) Coefficient of Determination:

i) $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{8,099}{15,079} = 0.4629$

ii) $R^2 = \frac{SSE}{SST} = \frac{6,980}{15,079} = 0.4629$

iii) $R^2 = \rho^2_{\hat{y}y}$ (square of correlation between predicted and actuals)

Since…

```
. corr yhat brozek
(obs=252)

             |     yhat   brozek
-------------+------------------
        yhat |   1.0000
      brozek |   0.6804   1.0000


. di .6804^2
.46294
```

$R^2 = \rho^2_{\hat{y}y} = .6804^2 = 0.4629$

[2] Elasticities are not required to be evaluated at the means… but they have to be evaluated somewhere… and why not start @ the means?

d) Adjusted R-squared: $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{SST} = 1 - \frac{251}{248}\frac{8,099}{15,079} = .4564$

5) Collinearity Regressions

a) Collinearity metric: $R_j^2 = .9011$

b) Variance Inflation factor (*VIF*): $VIF_{hip} = \frac{1}{1-R_{hip}^2} = \frac{1}{1-.9011} = 10.11$

6) MLR coefficients: What'sNew? … What'sLeft?

Full Model

```
. reg brozek hgt wgt hip

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(3, 248)       =     71.25
       Model |  6980.06726         3  2326.68909   Prob > F        =    0.0000
    Residual |  8098.94937       248  32.6570539   R-squared       =    0.4629
-------------+----------------------------------   Adj R-squared   =    0.4564
       Total |  15079.0166       251  60.0757635   Root MSE        =    5.7146

------------------------------------------------------------------------------
      brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hgt |  -.6163599   .1114903    -5.53   0.000    -.8359486   -.3967713
         wgt |   .1552489   .0404222     3.84   0.000     .0756344    .2348635
         hip |   .1314181   .1600891     0.82   0.412    -.1838896    .4467257
       _cons |   21.26829   13.88907     1.53   0.127    -6.087274    48.62386
------------------------------------------------------------------------------
```

Generate WhatsNew about hip [regress hip on hgt and wgt and capture residuals]

```
. reg hip hgt wgt
. predict whatsnew, resid

. reg brozek whatsnew
[slope coeff. agrees with MLR coeff.]

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(1, 250)       =      0.37
       Model |  22.0071353         1  22.0071353   Prob > F        =    0.5461
    Residual |  15057.0095       250   60.228038   R-squared       =    0.0015
-------------+----------------------------------   Adj R-squared   =   -0.0025
       Total |  15079.0166       251  60.0757635   Root MSE        =    7.7607

------------------------------------------------------------------------------
      brozek |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    whatsnew |   .1314181   .2174066     0.60   0.546    -.2967638       .5596
       _cons |   18.93849   .4888764    38.74   0.000     17.97565    19.90133
------------------------------------------------------------------------------
```

```
. summm whatsnew Brozek

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    whatsnew |        252    5.19e-09    2.253148  -8.390721   9.494614
      Brozek |        252    18.93849    7.750856          0       45.1

. corr Brozek whatsnew
(obs=252)

             |   Brozek whatsnew
-------------+------------------
      Brozek |   1.0000
    whatsnew |   0.0382   1.0000


Check:  . di .0382*7.750856/2.253148
.13140846
```

Generate WhatsLeft with brozek [regress brozek on hgt and wgt and capture residuals]

```
. reg brozek hgt wgt
. predict whatsleft, resid
. reg whatsleft whatsnew
[slope coeff., SSRs agree with MLR; MSE, RMSE, se and t are close (dof difference)]

      Source |       SS           df       MS      Number of obs   =       252
-------------+----------------------------------   F(1, 250)       =      0.68
       Model |  22.0071338         1  22.0071338   Prob > F        =    0.4106
    Residual |  8098.94924       250  32.3957969   R-squared       =    0.0027
-------------+----------------------------------   Adj R-squared   =   -0.0013
       Total |  8120.95637       251  32.3544078   Root MSE        =    5.6917

------------------------------------------------------------------------------
   whatsleft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    whatsnew |   .1314181   .1594475     0.82   0.411    -.1826135    .4454496
       _cons |   9.87e-09   .3585453     0.00   1.000    -.7061544    .7061545

Here's partial correlation between the brozek and hip … the correlation between
whatsnew and whatsleft:

. corr whatsleft whatsnew
(obs=252)

             | whatsleft whatsnew
-------------+------------------
   whatsleft |   1.0000
    whatsnew |   0.0521   1.0000


. summ whatsnew whatsleft

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    whatsnew |        252    5.19e-09    2.253148  -8.390721   9.494614
   whatsleft |        252    1.05e-08    5.688094  -18.54253   14.68069


Check:  . di .0521 *  5.688094/2.253148
.13152696
```