MLR Models: ***Estimation and Inference***

- ***Comparison of SLR and MLR analysis: What's New?... Not Much!***
- ***Variance Inflation Factors and Collinearity Regressions: VIFs and SSR$_x$***

Comparison of SLR and MLR analysis: What's New? … Not Much!

1. As you'll see in the tables below, there are remarkably few substantive differences between SLR and MLR models with respect to estimation and inference. Or put differently: Virtually all of what you learned about SLR models carries over to MLR models.
   a. The MLR conditions (MLR.1-MLR.6) will replace the SLR conditions… but they are basically the same as the SLR conditions, though SLR.3 (variation in the RHS variable) is replaced by MLR.3 (no perfect collinearity amongst the RHS variables).
   b. In the MSE formula, you now divide by n-k-1 to generate an unbiased estimator (k is the number of explanatory variables in the model).[1]
   c. A new concept, the *Variance Inflation Factor* (VIF), is now in the estimator variance and standard error formulas.
   d. The t-distributions used in Hypothesis Testing and Confidence Intervals now have n-k-1 degrees of freedom.
   e. And, *umm*, that's about it.
2. Similar to the SLR results, given MLR.1-5 and conditional on the x's,
   a. $\hat{\sigma}^2 = \frac{SSR}{n-(k+1)} = MSE$ will be an unbiased estimator of σ^2,
   b. $Var(B_x) = \frac{\sigma^2}{\left[\sum_i (x_i - \bar{x})^2\right]\left[1-R_x^2\right]} \;.= \frac{\sigma^2}{\left[\sum_i (x_i - \bar{x})^2\right]} VIF_x = \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$. Note the appearance of VIF_x.
   c. OLS will be a BLUE estimator.
3. Inference is virtually identical to the SLR case
   a. We use $se(B_x) = \sqrt{\frac{MSE}{\left[\sum_i (x_i - \bar{x})^2\right]} VIF_x} = \frac{RMSE\sqrt{VIF_x}}{S_x\sqrt{(n-1)}}$ to estimate $sd(B_x) = \sqrt{Var(B_x)}$.
   b. Given MLR.1-6, the t statistic, $\frac{B_j - \beta_j}{se(B_j)}$, has a t distribution with n-k-1 dofs (it's distributed t_{n-k-1}). We use that distribution to construct Confidence Intervals and

[1] And yes, n-2 = n-k-1 for SLR models.

conduct Hypothesis Tests… just as we did in the case of SLR models, and when looking at the Sample Mean estimator.

	SLR	**MLR**
Those (S/M)LR Assumptions	SLR.1: Linear Model $Y = \beta_0 + \beta_1 X + U$	MLR.1: Linear Model $Y = \beta_0 + \sum \beta_j X_j + U$
	SLR2: Random sampling	MLR2: Random sampling
	SLR3: Sampling variation in the independent variable	MLR3: No perfect collinearity
	SLR.4: Zero Conditional Mean $E(U \mid X = x) = 0$ for any x	MLR.4: Zero Conditional Mean $E(U \mid x_1, \ldots, x_n) = 0$ for any x_1, …
If these hold:	LUE: OLS provides unbiased parameter estimates	LUE: OLS provides unbiased parameter estimates
SRF (Sample Regression Function)	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y} = \beta_0 + \sum \beta_j \hat{x}_j$
PRF (Population Regression Function)	$E(Y \mid X = x) = \beta_0 + \beta_1 x$	$E(Y \mid x_1, \ldots, x_k) = \beta_0 + \sum \beta_j x_j$
Add one more assumption	SLR.5: Homoskedastic errors (conditional on x)	MLR.5: Homoskedastic errors (conditional on the x_j's)
If all 5 conditions hold:	$Var(B_1 \mid x's) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2}$ $= \frac{\sigma^2}{(n-1)S_{xx}}$	$Var(B_x \mid x, z\ldots) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2} VIF_x$ $= \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$
Variance Inflation Factors (VIF)		$VIF_x = \frac{1}{1 - R_x^2}$; $R_x^2 = 1 - \frac{1}{VIF_x}$
	$\hat{\sigma}^2 = \frac{SSR}{n-2} = MSE$	$\hat{\sigma}^2 = \frac{SSR}{n-(k+1)} = MSE$
	MSE unbiased estimator of σ^2	MSE unbiased estimator of σ^2

MLR Models: ***Estimation and Inference***

Gauss-Markov Thm.	SLR.1-SLR.5 $\rightarrow$BLUE	MLR.1-MLR.5 $\rightarrow$BLUE
2. Inference	**SLR**	**MLR**
Add one last Assumption: Distribution of Y_i	SLR.6: $U_i \sim N(0,\sigma^2)$ and indept. of X	MLR.6: $U_i \sim N(0,\sigma^2)$ and indept. of the RHS variables
Degrees of Freedom	n-2	n-k-1
Distribution of Estimator:	$B_1 \sim N\left(\beta_1, \text{var}(B_1)\right)$, where $\text{var}(B_1) = \frac{\sigma^2}{(n-1)S_{xx}}$	$B_x \sim N\left(\beta_x, \text{var}(B_x)\right)$, where $\text{var}(B_x) = \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$
Standard Deviation of B_1	$sd(B_1) = \frac{\sigma}{S_x\sqrt{n-1}}$	$sd(B_1) = \frac{\sigma}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
Standard Error of B_1 (est. of $sd\left(B_1\right)$)	$se(B_1) = \frac{RMSE}{S_x\sqrt{n-1}}$	$se(B_x) = \frac{RMSE}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
Distribution of Estimator: *t statistic*	$\frac{B_1 - \beta_1}{se(B_1)} \sim t_{n-2}$	$\frac{B_x - \beta_x}{se(B_x)} \sim t_{n-k-1}$
Confidence Intervals (C is confidence level)	$\left[B_1 \pm c\ se(B_1)\right]$, $P[\|t_{n-2}\|< c) = C$	$\left[B_x \pm c\ se(B_x)\right]$, $P[\|t_{n-k-1}\|< c) = C$
Null Hypothesis:	$H_0 : \beta_1 = 0$	$H_0 : \beta_x = 0$
t stat (under H_0)	$tstat = \frac{B_1}{se(B_1)} \sim t_{n-2}$	$tstat = \frac{B_x}{se(B_x)} \sim t_{n-k-1}$
p value (for given data set and OLS estimates)	$p = P[\|t_{n-2}\| > tstat_{\hat{\beta}_1})$	$p = P[\|t_{n-k-1}\| > tstat_{\hat{\beta}_x})$
Hypothesis Test (for given…) :	Reject H_0 if $\left\|t_{\hat{\beta}_1}\right\| > c$ or if $p < \alpha$	Reject H_0 if $\left\|t_{\hat{\beta}_j}\right\| > c$ or if $p < \alpha$

What's New? Not much!

Variance Inflation Factors and Collinearity Regressions: VIFs and SSR_x

4. At first glance the SLR and MLR standard error formulas appear to differ slightly, with a $\sqrt{VIF_x}$ adjustment included in the MLR formula:

   a. SLR - $se_{\hat{\beta}_x} = \dfrac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} = \dfrac{RMSE}{S_x \sqrt{n-1}}$

   b. MLR - $se_{\hat{\beta}_x} = \dfrac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} \sqrt{VIF_x} = \dfrac{RMSE}{S_x \sqrt{n-1}} \sqrt{VIF_x}$

5. But in fact, as you'll see below, both expressions can be rewritten as: $se_{\hat{\beta}_x} = \dfrac{RMSE}{\sqrt{SSR_x}}$, where SSR_x is the SSR from the collinearity regression in which the RHS variable *x* has been regressed on the other RHS variables. (In the SLR context, and as discussed above, this means that x is regressed on the constant variable, since there are no other RHS variables in the model.)

6. And so in this sense, there is no change in the standard error formula in moving from SLR to MLR models.

7. Here are the details:

   a. ***SLR:*** If you regress x on the constant variable, you know from before that the OLS predicted value will be the mean of x, so $\hat{x} = \bar{x}$. In this case, residuals are defined by $\hat{u}_i = x_i - \hat{x} = x_i - \bar{x}$, and so $SSR_x = \sum \left(x_i - \bar{x}\right)^2$, and as advertised, we have:

      i. SLR standard error: $se_{\hat{\beta}_1} = \dfrac{RMSE}{\sqrt{\sum (x_i - \bar{x})^2}} = \dfrac{RMSE}{\sqrt{SSR_x}}$.

   b. ***MLR***: We know that $se_{\hat{\beta}_x} = \dfrac{RMSE}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}} \sqrt{VIF_x}$.

      i. Focusing on the collinearity regression for *x*, $R_x^2 = 1 - \dfrac{SSR_x}{SST_x}$, where $SST_x = \sum \left(x_i - \bar{x}\right)^2$.

      ii. And so $\dfrac{1}{1-R_x^2} = \dfrac{SST_x}{SSR_x}$ and $\sqrt{VIF_x} = \sqrt{\dfrac{1}{1-R_x^2}} = \sqrt{\dfrac{SST_x}{SSR_x}}$.

      iii. Accordingly, we have: $se_{\hat{\beta}_x} = \dfrac{RMSE}{\sqrt{SST_x}} \sqrt{VIF_x} = \dfrac{RMSE}{\sqrt{SST_x}} \sqrt{\dfrac{SST_x}{SSR_x}} = \dfrac{RMSE}{\sqrt{SSR_x}}$.

   c. And so in both cases, for SLR and for MLR models, the standard error is *RMSE* divided by the square root of the SSRs from the respective collinearity regression.

8. ***Example:*** Let's return to the *movierevs* dataset and working through a couple examples, which are built around the MLR model in which *rtotgross* is regressed on *wk1*, *wk2* and *wk3* revenues. Here are the regression results from the full model and the collinearity regressions. Note that Standard Errors are reported below the estimated coefficients (that's an *esttab* option: *se*).

```
. esttab, se r2 scalar(rss rmse) compress

-------------------------------------------------------------
           Full Model          Collinearity Regressions
                  (1)          (2)          (3)          (4)
           rtotgross          wk1          wk2          wk3
-------------------------------------------------------------
wk1             0.540***                 0.261***    -0.115***
             (0.0253)                (0.00234)    (0.00336)

wk2             0.745***     2.361***                 0.792***
             (0.0761)     (0.0212)                (0.00603)

wk3             4.778***    -1.146***     0.872***
             (0.0798)     (0.0334)    (0.00664)

_cons          -0.601**      0.110       0.0817*      0.287***
              (0.227)      (0.102)     (0.0340)     (0.0323)
-------------------------------------------------------------
N                7730         7730         7730         7730
R-sq            0.921        0.886        0.959        0.908
rss (SSR)  2,333,804      471,875       52,173       47,388
rmse            17.38        7.815        2.598        2.476
-------------------------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

9. Let's confirm the standard errors in the Full Model using the formula above:

   a. $wk1: \; se_1 = \frac{RMSE}{\sqrt{SSR_1}} = \frac{17.38}{\sqrt{471{,}875}} = .02530$

   b. $wk2: \; se_2 = \frac{RMSE}{\sqrt{SSR_2}} = \frac{17.38}{\sqrt{52{,}173}} = .07609$

   c. $wk3: \; se_3 = \frac{RMSE}{\sqrt{SSR_3}} = \frac{17.38}{\sqrt{47{,}388}} = .07984$

10. In both SLR and MLR models, the standard errors are defined by: $se_{\hat{\beta}_x} = \frac{RMSE}{\sqrt{SSR_x}}$.