EC 228: Topics Review for Mid Term Examination #1

Comparison of Simple Linear and Multiple Linear Regression Analysis

Analytics	**SLR**	**MLR**
Data Generation Model	SLR.1: Linear Model $y_i = \beta_0 + \beta_1 x_i + u_i$	MLR.1: Linear Model $y_i = \beta_0 + \beta_x x_i + \beta_z z_i + u_i$ $y_i = \beta_0 + \sum_j \beta_j x_{ij} + u_i$
Residuals/ Unexplained	$u_i = y_i - \left(\beta_0 + \beta_1 x_i\right)$	$u_i = y_i - \left(\beta_0 + \beta_x x_i + \beta_z z_i\right)$, etc
OLS estimates	Min SSRs	Min SSRs
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	$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$ $= r_{xy} \frac{S_y}{S_x} = wgtd.\ avg\ of\ slopes$	Complicated… but similar: $\hat{\beta}_x = \frac{S_{x^+y^+}}{S_{x^+x^+}} = r_{x^+y^+} \frac{S_{y^+}}{S_{x^+}}$, where $y^+ = WhatsLeft_y$; $x^+ = WhatsNew_x$
sign(slopes)	$sign\left(r_{xy}\right)$, where r_{xy} is the correlation of the x's and the y's	$sign\left(r_{x^+y^+}\right)$, where $r_{x^+y^*}$ is the <u>partial</u> correlation of x's and y's
SRF (Sample Regression Function)	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_x x_i + \hat{\beta}_z z_i, etc ...$ $\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_j x_j$
SRF @ the means	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y}$	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_x \overline{x} + \hat{\beta}_z \overline{z} = \overline{y}$
Predicteds, actuals and residuals	$y_i = \hat{y}_i + \hat{u}_i$; $avg(\hat{y}'s) = \overline{y}$; $avg(\hat{u}'s) = 0$; $corr(\hat{y}'s, \hat{u}'s) = 0$	$y_i = \hat{y}_i + \hat{u}_i$; $avg(\hat{y}'s) = \overline{y}$; $avg(\hat{u}'s) = 0$; $corr(\hat{y}'s, \hat{u}'s) = 0$

MT #1 Topics Review

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*) $\Delta\hat{y} = \hat{\beta}_x \Delta x \Leftrightarrow \frac{\Delta\hat{y}}{\Delta x} = \hat{\beta}_x$
Analytics	**SLR**	**MLR**
… from changing several RHS vars		$\Delta\hat{y} = \hat{\beta}_x \Delta x + \hat{\beta}_z \Delta z$
Elasticities (at the means)	$\left[\frac{x}{\hat{y}}\frac{d}{dx}\hat{y}\right]_{x=\bar{x}} = \hat{\beta}_1 \frac{\bar{x}}{\bar{y}}$	$\frac{\partial\hat{y}}{\partial x}\left[\frac{x}{\hat{y}}\right]_{@\,means} = \hat{\beta}_x \frac{\bar{x}}{\bar{y}}$
Beta Regressions (standardized variables)	$\hat{\beta}_0 = 0$; $\hat{\beta}_1 = r_{xy}$	$\hat{\beta}_0 = 0$; $\hat{\beta}_x = ?$, $\hat{\beta}_z = ?$

Assessment	**SLR**	**MLR**
Sum Squares	SST = SSE + SSR	SST = SSE + SSR
R^2 (Coefficient of Determination) (w/ intercept term)	$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$	$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$
	$R^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = r_{xy}^2 = r_{\hat{y}y}^2$	$R^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = r_{\hat{y}y}^2$
Degrees of freedom (dofs)	$dofs = n - 2$	$dofs = n - k - 1$
MSE	$MSE = \frac{SSR}{dofs} = \frac{SSR}{n-2}$	$MSE = \frac{SSR}{dofs} = \frac{SSR}{n-k-1}$
RMSE	$RMSE = \sqrt{MSE}$	$RMSE = \sqrt{MSE}$
Adjusted R^2		$\bar{R}^2 = 1 - \frac{SSR}{SST}\frac{n-1}{n-k-1} = 1 - \frac{MSE}{S_{yy}}$
Collinearity		R_x^2

MT #1 Topics Review

VIF (Variance Inflation Factor)		$VIF_x = \frac{1}{1-R_x^2}$

Further discussion

1) Sample statistics:

Sample mean: $\bar{x} = \frac{1}{n}\sum x_i$

Sample variance: $S_{xx} = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Sample standard deviation: $S_x = \sqrt{S_{xx}}$

Sample covariance: $S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Sample correlation: $r_{xy} = \rho_{xy} = \frac{S_{xy}}{S_x S_y}$ (note that r and ρ notation are both used)

2) Ordinary Least Squares (OLS)
   a) *Residuals*: *Actuals* - *Predicteds*
   b) Min SSRs - First Order Conditions (FOCs) and Second Order Conditions (SOCs)
   c) Excel *trendline* generates OLS/SLR coefficients
   d) Actuals = Predicteds + Residuals; corr (Predicteds, Residuals) = 0;

      Var(Actuals) = var(Predicteds) +Var(Residuals)
3) Assessing responsiveness/meaningfulness of predicted (SRF) effects: derivatives, elasticities and *beta* regressions
4) Assessment I: *Goodness-of-Fit*
   a) How well does the model fit the data? How close are *predicteds* to *actuals*? (R^2, MSE / $RMSE$ and $\bar{R}^2$)
5) MLR Model building: Now the SRF controls for all other RHS variables/effects, and we worry about what other RHS variables should be in the model ... and which ones we should take out … and when to put our pencil down … and can brag about the model?
6) Choosing between Models (same LHS variable)
   a) SLR models: max R^2, max $|r_{xy}|$, max $|r_{\hat{y}y}|$, min *MSE* and min *RMSE* all lead to the same preferred Model
   b) MLR models: max $adj\ R^2$, min *MSE* and min *RMSE* all lead to the same preferred Model (R^2 not an attractive metric since it cannot decrease with more RHS vars)
7) Collinearity regression: Regress one RHS variable on the other RHS variables. Useful in regards to:
   a) Multicollinearity and VIFs in MLR models
   b) Omitted Variable Bias/Impact (Endogeneity)

c) Interpretation of MLR estimated coefficients: *What's New?* (The residual from the collinearity regression).

8) Multicollinearity and VIFs: the extent to which the values of any one RHS variable can be predicted as a linear function of the other RHS variables. This is typically measured using R_j^2, the R squared in the regression of explanatory variable x_j on the other RHS variables.

   Multicollinearity impacts the Variance Inflation Factor (VIF), since $VIF_j = \frac{1}{\left[1 - R_j^2\right]}$ …. and

   $R_j^2 = 1 - \frac{1}{VIF_j}$. Multicollinearity can lead to wacky coefficient estimates, so beware!

9) Omitted variable bias/impact (Endogeneity): The extent to which parameter estimates are biased/impacted by the exclusion of other RHS variables from the model. In the case one one omitted/dropped variable, it is driven by two factors:

   - the coefficient of the omitted variable when it is in the Full Model,
   - the collinearity of the omitted variable with the other explanatory variables in the model

   Often times, we're happy to be able to just sign the bias, and get a sense of whether we have under- or over-estimated slope coefficients.

10) Interpretation of MLR estimated coefficients:

    a) *SRF*: As implemented with the SRF, the MLR coefficients capture the relationships between incremental changes in a RHS variable *ceteris paribus*, and changes in the predicted values of the dependent variable.

    b) *What's New*: $WhatsNew_x$ about RHS variable *x* is the residual from the collinearity regression of *x* on the other RHS variables. The MLR estimated coefficient for a RHS variable *x* can be derived by regressing the dependent variable on $What'sNew_x$.

    c) *What's Left*: $WhatsLeft_y$ about the LHS variable *y* is the residual from the regression of the LHS variable *y* on the RHS variables other than *x*. The MLR estimated coefficient for RHS variable *x* can be derived by regressing $WhatsLeft_y$ on $WhatsNew_x$.

       i) The correlation between $WhatsLeft_y$ and $WhatsNew_x$ is a *partial* correlation

11) Adjusted R-squared: Adjusted R-squared is an attempt to adjust the coefficient of determination (R^2) for the fact that R^2 cannot decline (since SSR cannot increase) when you add RHS variables to a model. It basically adjusts R^2 for changes in the degrees of freedom (*dofs*) in a model, and will increase or decrease depending on whether the percentage decrease in SSRs is greater than or less than the percentage change in dofs. $\bar{R}^2$ is always less than R^2, and accordingly, bounded below 1…. and moves in the opposite direction from MSE/RMSE.

12) Endogeneity and correlations: In signing OVB, it is tempting to consider the correlation between *y* and the omitted RHS variable, as well as the correlations between the omitted and surviving RHS variables. But that would be a mistake! You need to focus on the *partial* correlations (not simple correlations)… which can be conceptually far more challenging to sign/evaluate.