AREC422 Notes

Youpei Yan

February 21, 2019

In the MLR, GMAs also change/upgrade to the following version:
1. "linearity".
2. Random sampling.
3. No perfect collinearity. (None of the Xs is constant, no exact linear relationship among Xs.)
4. Zero conditional mean. $E(u|x_1, x_2, ...x_k) = 0$
5.Homoscedasticity. $Var(u|x_1, x_2, ...x_k) = \sigma^2$

From GMAs, we can derive 3 Theorems:
1. Unbiasedness of OLS: $E(\hat{\beta}_j) = \beta_j$, $j = 1, 2, ..., k$.
2. Unbiased estimation of σ^2. $E(\hat{\sigma}^2) = \sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{u_i}^2$$

3. (**Gauss-Markov Theorem**): The OLS estimator $\hat{\beta}_j$ for β_j is the redbest redlinear redunbiased redestimator (BLUE).

- "best": have the smallest variance for any estimator $\tilde{\beta}_j$ that is linear and unbiased. $Var(\hat{\beta}_j) < Var(\tilde{\beta}_j)$
- "linear": the estimator $\tilde{\beta}_j$ can be expressed as a linear function of the dependent variable: $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$

Normality Assumption $u \sim N(0, \sigma^2)$+ GMA1-5: Classical linear model assumptions.

In MLR: Adding more variables will change the estimated coefficient(s) and may cause issues. (Recall that adding another X, R^2 will increase, but low $R^2 \neq$ unreliable estimates, high $R^2 \neq$ reliable estimates.

If adding irrelevant variable(s), the model is **overspecified**:
True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, but we specify the model as: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ (x_3 has no effect, $\beta_3 = 0$, but we did not know this. Instead, we collect the corresponding data and run regression with x_3.)

Consequences:
(1) No effect in terms of $\hat{\beta}_1$ and $\hat{\beta}_2$'s unbiasedness.
(2) Variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be higher, which leads to less precise estimators. This is undesirable.

Overspecifying a model, although undesired, is allowed. However, if we underspecify a model by missing a relevant variable, we may have more serious consequences.

This time, say the true model is $y = \beta_0+\beta_1 x_1+\beta_2 x_2+u$, but we specify the model as: $y = \beta_0+\beta_1 x_1+e$ (Then we will get a line $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ rather than $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$).

Consequences:
(1) $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$, where $\tilde{\delta}_1$ is the coefficient if we run regression of x_2 on x_1: $x_2 = \delta_0 + \delta_1 x_1 + e$. We have $\tilde{\beta}_1 = \hat{\beta}_1$ only if x_2 has no impact on y and x_1, x_2 are uncorrelated. Since x_2 is relevant, $\beta_2 \neq 0$. we have a biased estimator. Also, since x_1 and x_2 are likely to be correlated, e includes $\beta_2 x_2 + u$, $E(x_1 e) \neq 0$, which violates GMA4, we can still conclude that $\tilde{\beta}_1$ is a biased estimator.
(2) $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$

More about the biased estimator: $E(\tilde{\beta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = E(\hat{\beta}_1) + E(\hat{\beta}_2)\tilde{\delta}_1 = \beta_1 + \beta_2\tilde{\delta}_1$
Omitted Variable Bias formula:
OVB= $Bias(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1$

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

For example, if we have an equation: $log(price) = \beta_0 + \beta_1 log(NO_2) + \beta_2 \#Schools + u$.

Q1: What's the expected signs of β_1 and β_2? (Answer: -, +)
Q2: What is β_1? (Answer: elasticity of price with respect to NO_2.)
Q3: Suppose that we omit $\#Schools$ and got the following estimated line:

$\widehat{log(price)} = 11.71 - 1.043 log(NO_2)$, with $R^2 = 0.264$, and the true model is supposed to be $\widehat{log(price)} = 9.23 - 0.718 log(NO_2) + 0.306 \#Schools$, with $R^2 = 0.514$.
What is the correlation between $\#Schools$ and nearby NO_2 density in this example?
Answer: they are negatively correlated, because $-1.043 = -0.718 + 0.306 + \tilde{\sigma_1}$. Also, this negative bias is caused by ignoring a positive impact ($\#Schools$ to the price) that is negatively correlated to the observed variable NO_2.

Similar to the SLR, when we had our t-test for each $\hat{\beta}$, we have the same test in the MLR. More than that, we can test if a group of variables' impact on the dependent variable.

(Feb 26, Tuesday)
Testing Hypotheses: the F-test.

If we specify our hypotheses as a group of equations, such as $H_0 : \beta_3 = 0, \beta_4 = 0$ (i.e., both x_3 and x_4 have no effect on y) vs. $H_1 : H_0$ is not true, we need an F-test rather than a t-test.

For instance, with $n = 353$ observations, we run the regression and get the regression line as:
$\widehat{log(pollution)} = \underset{(0.29)}{11.19} + \underset{(0.0121)}{0.0689} firm\_num + \underset{(0.0026)}{0.0126} car\_den + \underset{(0.0011)}{0.00098} production - \underset{(0.0161)}{0.0144} abate\_invest - \underset{(0.0072)}{0.0108} fine\_level$
with $SSR = 183.186$, $R^2 = 0.6278$.

We are testing if $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ (vs. $H_1 : H_0$ is not true.)

If H_0 is true, the model becomes $log(pollution) = \beta_0 + \beta_1\ firm\_num + \beta_2\ car\_den + u$. This is called the "restricted model" (vs. the original "unrestricted model").

We estimate the restricted model, suppose that we get the new regression line as:

$\widehat{log(pollution)} = \underset{(0.11)}{11.22} + \underset{(0.0125)}{0.0713} firm\_num + \underset{(0.0013)}{0.0202} car\_den$ with $SSR = 198.311$, $R^2 = 0.5971$.

To examine if H_0 is true, is the same as examining if the two models are the same.

Define F-value

$$F \equiv \frac{(SSR_r - SSR_{ur})/(df_r - df_{ur})}{SSR_{ur}/df_{ur}}$$

In the above example, $df_r = 353 - 2 - 1 = 350$ and $df_{ur} = 353 - 5 - 1 = 347$.

$$F = \frac{(198.311 - 183.186)/3}{183.186/347} = 9.55$$

We can reject the hypothesis that the production level, the abatement investment, and the level of pollution fine have no effect on pollution.

However, note that if we perform the t-test on each of the three variables, none of them are statistically significant: $t_{prod} = 0.891$, $t_{abate} = 0.894$, $t_{fine} = 1.5$, all of them are smaller than 1.645 (critical value at 10%).

Why this is the case? This is because 2 variables *abate_invest* and $fine\_level$ may be strongly linked (multicollinearity). they are jointly significant, but their partial effects are covered by each other. That is why we need F-test as a complement of t-test to examine the effects.

Another formula for the F-statistics (because $SSR = SST(1 - R^2)$):

$$F = \frac{(R^2_{ur} - R^2_r)/(df_r - df_{ur})}{(1 - R^2_{ur})/df_{ur}}$$

This is called the R^2-form of the F-statistics.

Note that if H_0 : all the slope parameters are 0, the corresponding F stat. becomes the one for "overall significance of a regression". This is a term reported in the R result in case you are curious what that is.