AREC422 Notes

Youpei Yan

February 28, 2019

About the exam: 75 mins, 3 major questions, 10 small questions. No calculator is needed. Use formulas with numbers plugged in to asnwer all questions.

1. OLS estimates.
2. Read the results in R (MLR analysis)
- coefficient interpretation
- t-stat/Hypothesis Testing
- CI
3. MLR-specification & F-test
- under/over-specification
— OVB & compare E() or Var()
—Intuitive explanation
- F-test: single hypothesis/group of hypotheses
— It's relationship with t-test if $H_0 : \beta_j = 0$

Go through the HW2 questions:
(1) 1B: "explain the coefficient are statistically significant or not". You need to use t-value/p-value in your answer during the explanation.

(2) 1C & 2B: "follow your expectation or not". You should say, increasing x will increase/decrease y, such as, "increasing the number of firms will boost local pollution level".

(3) 3B: "OVB's intuitive interpretation". Use the formula $E(\tilde{\beta}_1) = E(\hat{\beta}_1) + E(\hat{\beta}_2\tilde{\delta}_1)$, read $\hat{\beta}_1$ and $\hat{\beta}_2$ in the regression. Find the relationship between x_1 and x_2 to get the sign of $\tilde{\delta}_1$. Specify the positive/negative bias and provide the intuitive interpretation:
In this specific case, x_2 decreases y, so missing x_2 will let the coefficient $\tilde{\beta}_1$ absorb this negative effect. Because x_1 and x_2 are positively correlated, the true impact is less negative than the impact estimated in the wrong model.

(4) 3E: Just note that, if the F-test for a joint hypotheses suggests that a group of variables are jointly significant, but the variables' t-statistics suggest that they're not significant, it's very likely that the group of variables are highly correlated with the multi-collinearity issue. We may not uncover the partial effect for each of them.

(5) F-test in general. It can perform a single hypothesis testing: $H_0 : \beta_j = 0$ or a group of hypotheses testing: $H_0 : \beta_3 = \beta_4 = 0$. The F-statistic reported at the end of the R result is testing the overall significance with $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.

The exam will be very similar to HW questions. Let's extend the HW questions here.

(1) See HW2. 1A's R result:

F-stat: 419.8 on 1 and 24528 DF, means that $F_{1,24528} = 419.8$ when we test the hypothesis: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neg 0$.

Check the t-value fo the same hypothesis: $t_{24528} = 20.49$. Note that $t^2_{n-k-1} = F_{1,(n-k-1)}$, this works when the 1st DF of F is 1.

What is the number of observations in this dataset? $n - k - 1 = 24528$ and $k = 1$, then $n = 24530$.

(2) See HW2. 2A's R result:

F-stat: 1691 on 3 and 24526 DF, means that $F_{3,24526} = 1691$ when we test the hypothesis:$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1 : H_0$ is not true.

We can also write the CI given the regression result for each of the variables. Say the 95% CI for $frac$ is $[41.04 \pm 1.96 \cdot 0.749]$. If a value is inside the range, we cannot reject the corresponding $H_0 : \beta_j =$ this value.

(3) See HW2. 3A:

Coefficient interpretation can be shown in all kinds of models: level-level, level-log, log-level, and log-log. When we interpret a coefficient, we should say:

Holding everything else fixed, 1 unit increase in x_i will increase/decrease $\hat{\beta}_i$ units in y. Or:
(level-log): Holding everything else fixed, 1% increase in x_i will increase/decrease $\hat{\beta}_i$ units in y.
(log-level): Holding everything else fixed, 1 unit increase in x_i will increase/decrease $100\hat{\beta}_i\%$ in y.
(log-log): Holding everything else fixed, 1% increase in x_i will increase/decrease $\hat{\beta}_i\%$ in y.

(4) See HW2. 3C/D:

When performing t/F-tests, you should always write the formula down first:

$$t_{n-k-1} = \frac{\hat{\beta}_i - a}{se(\hat{\beta}_i)}$$

if $H_0 = \beta_i = a$ vs.$H_1 : \beta_i \neq a$. I'll provide the critical values, and you should write your conclusion as: If $t_{n-k-1} >$ critical value, we reject H_0 at the 5% or 1% level, and if $t_{n-k-1} <$ critical value, we fail to reject H_0 at the 5% or 1% level.

Similarly, write F-statistic down with R^2 form formula or SSR form formula, conclude with critical values and the corresponding significance levels as well.

Now, several things un-covered in the HW2, but will show up in the exam:

OLS estimates
1.1 Derivation (SLR)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{sample\ covariance}{sample\ variance\ of\ x_i}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$sample\_mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = E(x) = \mu$$

$$sample\_variance = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

What would happen if we try to fit a model like $y = \beta_0 + \beta_1 x^2 + u$?
Treat $x^2 = z$, so we get a SLR: $y = \beta_0 + \beta_1 z + u$.

1.2 GMAs
Gauss Markov Theorem: OLS estimates are BLUE if 5 GMAs are satisfied. What is BLUE and what are the GMAs?

“Best”: $\hat{\beta}_j$ estimated using OLS has the smallest variance.
“Linear”: $\hat{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$
“Unbiased”: $E(\hat{\beta}_j) = \beta_j$

GMAs: linearity, random sampling (e.g.: if we observe only the rich people’s income when we run the regression income∼education, we will get biased estimator), no perfect collinearity, zero conditional mean (e.g., OVB issue), homoscedasticity.