ECON 3740: INTRODUCTION TO ECONOMETRICS

INSTRUCTOR: CHAOYI CHEN
Department of Economics and Finance, University of Guelph

Lecture 14

Lecture outline

Last lecture, we studied the the hypothesis test with a single parameter. Today, we will

- Construct confidence interval
- Test hypotheses about a single linear combination of the parameters
- Test multiple linear restrictions
  - Test exclusion restrictions
  - Unrestricted model and restricted model
  - F test
  - Test overall significance
  - Test general linear restrictions

MLR: Computing p-Values for t Tests

- Recall that $\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1}$. Hence,

$$P\left(\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} > c_{0.05}\right) = 0.025$$

$$P\left(\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} < -c_{0.05}\right) = 0.025$$

$$P\left(c_{0.05} \leq \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \leq c_{0.05}\right) = 0.95$$

  where $c_{0.05}$ is the 5% critical value of two-sided test

- Simple manipulation of above result shows

$$P\left(\underbrace{\widehat{\beta}_j - c_{0.05} * se(\widehat{\beta}_j)}_{lower\ \ bound\ \ of\ \ the\ \ CI} \leq \beta_j \leq \underbrace{\widehat{\beta}_j + c_{0.05} * se(\widehat{\beta}_j)}_{upper\ \ bound\ \ of\ \ the\ \ CI} \right)$$

$$= P\left(\left| \frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \right| \leq c_{0.05} \right) = 0.95$$

- The confidence interval is $[\widehat{\beta}_j - c_{0.05} * se(\widehat{\beta}_j), \widehat{\beta}_j + c_{0.05} * se(\widehat{\beta}_j)]$. 0.95 is called the confidence level.

MLR: Confidence Interval - an example

- The fitted regression line is
$$\widehat{log(rd)} = \underset{(0.47)}{-4.38} \quad \underset{(0.06)}{+1.084 \log(\text{sales})} \quad \underset{(0.0128)}{+0.0217 \text{profmarg}} \quad \text{where}$$
$$rd = \text{firms spending on RD}$$
$$\text{sales} = \text{annual sales}$$
$$\text{profmarg} = \text{profits as percentage of sales}$$
- $df = n - k - 1 = 32 - 2 - 1 = 29$. Hence, $c_{0.05} = 2.045$.
- The 95% CI for $\beta_{log(sales)}$ is $[1.084 - 2.045 * (0.06),\ 1.084 + 2.045 * (0.06)] = [0.961,\ 1.21]$.The effect of log(sales) on log (rd) is relatively precisely estimated as the interval is narrow. Moreover, the effect is significantly different from zero because zero is outside the interval.
- The 95% CI for $\beta_{profmarg}$ is $[0.0217 - 2.045 * 0.0128,\ 0.0217 + 2.045 * 0.0128] = [-0.0045,\ 0.0479]$. The effect of profmarg on log (rd) is imprecisely estimated as the interval is very wide. It is not even statistically significant because zero lies in the interval.

MLR: Testing Hypotheses about a Single Linear Combination of the Parameters: An Example

- To investigate whether returns to education at 2-Year vs. at 4-Year colleges are equal or not, one propose a model as following
$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + \mu,$$
where
jc=years years of education at 2-year colleges, $univ$= at 4-year colleges
- Suppose we want to test
$$H_0 : \beta_1 - \beta_2 = 0 \ vs \ H_1 : \beta_1 - \beta_2 < 0$$
- A possible test statistic would be
$$t = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$
- However, here appear some validality problems relating to above t statistic. First, we know standardized $\widehat{\beta}_1$ follows t distribution, and standardized $\widehat{\beta}_2$ follows t distribution. But these cannot make sure that standardized $\widehat{\beta}_1 - \widehat{\beta}_2$ will also follow t distribution. Secondly, assuming standardized $\widehat{\beta}_1 - \widehat{\beta}_2 \sim t_{n-k-1}$, it is still impossible to compute such a t statistic since one cannot compute $se(\widehat{\beta}_1 - \widehat{\beta}_2)$, which relates to the $\widehat{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ and is usually not available in regression output.

MLR: Testing Hypotheses about a Single Linear Combination of the Parameters: An Example Continue

- An Alternative Method: Define $\theta_1 = \beta_1 - \beta_2$. Therefroe, the hypothesis test can be rewritten as

$$H_0 : \theta_1 = 0 \ vs \ H_1 : \theta_1 < 0$$

- Now, $\beta_1 = \theta_1 + \beta_2$. Inserting it into the original regression, we have

$$\begin{aligned} log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 univ + \beta_3 exper + \mu \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + univ) + \beta_3 exper + \mu \end{aligned}$$

  where $jc + univ$ is a new regressor, representing total years of college.

MLR: Testing Hypotheses about a Single Linear Combination of the Parameters: An Example Continue

- After estimation, the fitted regression line is

$$\log(\widehat{wage}) = \underset{(0.021)}{1.472} \quad \underset{(0.0069)}{\text{-0.0102jc}} \quad \underset{(0.0023)}{+0.0769(\text{jc}+\text{univ})} \quad \underset{(0.0002)}{+0.0049\text{expe}}$$

  where $n = 6763$ and $R^2 = 0.222$
- Hence, the t statistic is

$$t = \frac{-0.0102}{0.0069} = -1.48$$

- Recall that the critical values with significance level 5% and 10% are -1.645 and -1.282 respectively for the left tail one sided test.
- Thus, the null is rejected at 10% level but not at 5% level.
- You can also compute the p value, which gives you $P(T < -1.48) = 0.07 \in (0.05, 0.10)$, and the 95% CI for θ_1 is $-0.102 + / - 1.96 * (0.0069) \longrightarrow (-0.0237, 0.0003)$, which covers zero.
- Note that this method works always for single linear hypotheses.

MLR: Testing Multiple Linear Restrictions: Testing Exclusion Restrictions

- The estimated restricted model is

  $$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr +$$

  where

  $$\begin{aligned} salary &= \text{the 1993 total salary} \\ years &= \text{years in the league} \\ gamesyr &= \text{average games played per year} \\ bavg &= \text{career batting average} \\ hrunsyr &= \text{home runs per year} \\ rbisyr &= \text{runs batted in per year} \end{aligned}$$

- Now, we consider following hypothesis,

  $$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \; vs \; H_1 : H_0 \; is \; not \; true$$

  where H_0 is not true means "at least one of $\beta_3, \beta_4, \beta_5$ is not zero.
- This is to test whether performance measures have no effects or can be excluded from regression.

MLR: Testing Multiple Linear Restrictions: Estimation of the Unrestricted Model

- The estimated unrestricted model is

$$\begin{array}{llll} \widehat{\log(salary)} = & 11.19 & -0.0689\text{years} & +0.012\text{gamesyr} \\ & (0.29) & (0.0121) & (0.0026) \\ & +0.00098\text{bavg} & +0.0144\text{hrunsyr} & +0.0108\text{rbisyr} \\ & (0.0011) & (0.0161) & (0.0072) \end{array}$$

  where $n = 353$, $SSR_{ur} = 181.186$, and $R^2 = 0.6278$

- Note that none of these three variables is statistically significant when tested individually. (Why? compute the corresponding t statistic). However, the individual insignificance may not imply together they are insignificant.
- Idea: How would the model fit (measured in SSR) be if these variables were dropped from the regression?

MLR: Testing Multiple Linear Restrictions: Estimation of the restricted Model

- By dropping those variables from the model, The estimated restricted model is
$$\widehat{\log(salary)} = \underset{(0.11)}{11.22} \; \underset{(0.0125)}{+0.0713 years} \; \underset{(0.0013)}{+0.202 gamesyr}$$
where $n = 353$, $SSR_r = 198.311$, and $R^2 = 0.5971$
- The sum of squared residuals (SSR) necessarily increases in the restricted model. But, is this increase statistically significant?
- To figure out the problem, we consider to use a rigorous test statistic
$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$
where $q = df_r - df_{ur}$ is the number of restrictions, and $n - k - 1 = df_{ur}$.
- The relative increase of the sum of squared residuals when going from H_1 (unrestricted model) to H_0 (restricted model) follows a F distribution.

MLR: Testing Multiple Linear Restrictions: F test

- Therefore, the F statistic of our example is

$$F = \frac{(198.311 - 181.186)/3}{181.186/(353 - 5 - 1)} \approx 9.55$$

- Check with the F table, we find if $q = 3$, $df = 347$, the critical value with 1% significance level is 3.78.
- 9.55¿3.78. Therefore, we reject the null.
- Alternatively, you can compare p value. $P(F_{3,347} > 9.55) = 0.0000$, which implpies the null hypothesis is overwhelmingly rejected (even at very small significance levels).
- Remarks:
  - If H_0 is rejected, we say that the three variables are "jointly significant".
  - They were not significant when tested individually.
  - The possible reason is multicollinearity between them.

MLR: Testing Multiple Linear Restrictions: The R-Squared Form of the F Statistic

- Recall that

$$R^2 = -\frac{SSR}{TSS} \Longrightarrow SSR = TSS(1 - R^2)$$

- Hence,

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{[TSS(1-R_r^2) - TSS(1-R_{ur}^2)/q]}{TSS(1-R_{ur}^2)/(n-k-1)}$$
$$= \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}$$

- With previous example, since $R_{ur}^2 = 0.6278$, $R_r^2 = 0.5971$, therefore,

$$F = \frac{(0.6278 - 0.5971)/3}{(1 - 0.6278)/347} \approx 9.54$$

which is very close to the result based on SSR (difference due to rounding error).

MLR: Testing Multiple Linear Restrictions: The F Statistic for Overall Significance of a Regression

- Consider a typical population regression model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \mu$$

- Now, suppose we would like to conduct a special hypothesis test

$$H_0 : \beta_1 = ... = \beta_k = 0 \ vs \ H_1 : H_0 \ is \ not \ true$$

- As a result, the restricted model is

$$y = \beta_0 + \mu$$

  which is a regression on constant. Clearly, $\widehat{\beta}_0 = \bar{y}$ and $R_r^2 = 0$ from the knowledge in SLR.
- Thus, the F statistic of this special hypothesis test is

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)} = \frac{R_{ur}^2/k}{(1 - R_{ur}^2)/(n-k-1)} \sim F_{k,n-k-1}$$

- The test of overall significance is reported in most regression packages (also in 'lm' package in R). The null hypothesis is usually overwhelmingly rejected.

MLR: Testing Multiple Linear Restrictions: Testing General Linear Restrictions

- Suppose you and your group member would like to test whether house price assessments are rational, where the population regression model is

$$\begin{aligned} log(price) = \beta_0 + \beta_1 log(assess) + \beta_2 log(lotsize) \\ + \beta_3 log(sqrft) + \beta_4 bdrms + \mu \end{aligned}$$

where

$$\begin{aligned} price &= \text{house price} \\ assess &= \text{the assessed housing value (before the house was sold)} \\ lotsize &= \text{size of the lot, in feet} \\ sqrft &= \text{square footage} \\ bdrms &= \text{number of bedrooms} \end{aligned}$$

- Now, suppose we focus on the following hypothesis test

$$H_0 : \beta_1 = 1, \beta_2 = \beta_3 = \beta_4 = 0$$

MLR: Testing Multiple Linear Restrictions: Testing General Linear Restrictions

- Under the null, $\beta_1 = 1$, which means that if house price assessments are rational, a 1% change in the assessment should be associated with a 1% change in price. (log-log model). Also, $\beta_2 = \beta_3 = \beta_4 = 0$, which means that in addition, other known factors should not influence the price once the assessed value has been controlled for.
- Since it is a test involving multiple linear restrictions, we consider to use F test.
- The restricted model is

$$y = \beta_0 + x_1 + \mu$$

- However, if you just regress y on x_1, we cannot ensure the restriction $\beta_1 = 1$ hold. Hence, we consider to change the resticted model as $y - x_1 - \beta_0 + \mu$, which means The restricted model is actually a regression of $y - x1$ on a constant, and the resulting $\widehat{\beta}_0$ is the sample mean of $y - x_1$.

MLR: Testing Multiple Linear Restrictions: Testing General Linear Restrictions Continue

- Suppose that after estimation with both unrestricted model restricted model, you have the following results

$$SSR_r = 1.88,\ \ SSR_{ur} = 1.822,\ \ n = 88$$

- Hence, the F statistic is

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(1.88 - 1.822)/4}{1.822/(88-4-1)} \approx 0.661$$

- Checking the F table, you find the critical value with 5% significance level of a $F_{4,83}$ distribution is 2.5, which is greater than 0.661. Therefore, we cannot reject the null