複迴歸中的 F 檢定

(取自 Wooldridge Chap 4)

- 請先復習「六大檢定」中，那個檢定是 t 檢定，那個檢定是 F 檢定? 請說明此兩檢定的本質不同之處。
- 在 ANOVA 檢定中，明明是檢定平均值，確為什麼用 F 檢定?
- 在複迴歸中，F 檢定扮演了怎樣的角色?

  簡答: Test for multiple restrictions (參考投影片 p.31 以後)

- **Testing multiple linear restrictions: The F-test**
- **Testing exclusion restrictions**

Salary of major league baseball player | Years in the league | Average number of games per year

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr$$

$$+\beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Batting average | Home runs per year | Runs batted in per year

$\boxed{H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0}$ against $H_1 : H_0$ is not true

Test whether performance measures have no effect/can be excluded from regression.

- **Estimation of the unrestricted model**

$$\widehat{\log}(salary) = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689}\ years + \underset{(.0026)}{.0126}\ gamesyr$$

$$+ \underset{(.00110)}{.00098}\ bavg + \underset{(.0161)}{.0144}\ hrunsyr + \underset{(.0072)}{.0108}\ rbisyr$$

None of these variabels is statistically significant when tested individually

$n = 353,\ SSR = 183.186,\ R^2 = .6278$

Idea: How would the model fit be if these variables were dropped from the regression?

- **Estimation of the restricted model**

$$\widehat{\log}(salary) = \underset{(0.11)}{11.22} + \underset{(.0125)}{.0713}\ years + \underset{(.0013)}{.0202}\ gamesyr$$

$$n = 353,\ SSR = 198.311,\ R^2 = .5971$$

The sum of squared residuals necessarily increases, but is the increase statistically significant?

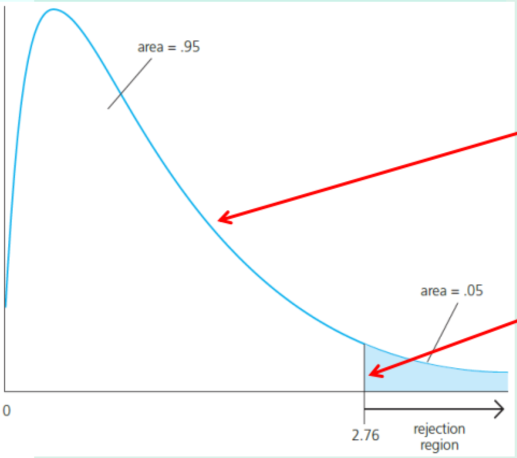
- **Test statistic**

Number of restrictions

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

The relative increase of the sum of squared residuals when going from H_1 to H_0 follows a F-distribution (if the null hypothesis H_0 is correct)

- **Rejection rule**



A F-distributed variable only takes on positive values. This corresponds to the fact that the sum of squared residuals can only increase if one moves from H_1 to H_0.

Choose the critical value so that the null hypothesis is rejected in, for example, 5% of the cases, although it is true.

- **Test decision in example**

Number of restrictions to be tested

$$F = \frac{(198.311 - 183.186)/3}{183.186/\boxed{(353-5-1)}} \approx 9.55$$

Degrees of freedom in the unrestricted model

$$F \sim F_{3,347} \Rightarrow c_{0.01} = 3.78$$

$$P(F - statistic > 9.55) = 0.000$$

The null hypothesis is overwhelmingly rejected (even at very small significance levels).

- **Discussion**
  - The three variables are "jointly significant"
  - They were not significant when tested individually
  - The likely reason is **multicollinearity** between them

Multicollinearity → variables are not significant in individual tests
but they are significant in a joint test

- **Test of overall significance of a regression**

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u$$

$$\boxed{H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

The null hypothesis states that the explanatory variables are not useful at all in explaining the dependent variable

$$y = \beta_0 + u$$

Restricted model (regression on constant)

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

- **The test of overall significance is reported in most regression packages; the null hypothesis is usually overwhelmingly rejected**

- 注意：在「單迴歸」問題中(即 k=1)，當 testing for overall significance 時，會有 $F = t^2$ 的關係式成立，請自行驗證。

以下為一特殊案例研討 (參考用)

- **Testing general linear restrictions with the F-test**
- **Example: Test whether house price assessments are rational**

Actual house price

The assessed housing value (before the house was sold)

Size of lot (in square feet)

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

Square footage

Number of bedrooms

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

In addition, other known factors should not influence the price once the assessed value has been controlled for.

If house price assessments are rational, a 1% change in the assessment should be associated with a 1% change in price.

- **Regression output for the unrestricted regression**

$$\widehat{\log}(price) = \underset{(.570)}{.264} + \underset{(.151)}{1.043} \log(assess) + \underset{(.0386)}{.0074} \log(lotsize) - \underset{(.1384)}{.1032} \log(sqrft) + \underset{(.0221)}{.0338}\, bdrms$$

$$n = 88, SSR = 1.822, R^2 = .773$$

When tested individually, there is also no evidence against the rationality of house price assessments

- Question: 在下例中，作者用 F 檢定作怎樣的推論? 其結果如何?

EXAMPLE 4.9 **Parents' Education in a Birth Weight Equation**

As another example of computing an F statistic, consider the following model to explain child birth weight in terms of various factors:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u, \quad [4.42]$$

where

$bwght$ = birth weight, in pounds.
$cigs$ = average number of cigarettes the mother smoked per day during pregnancy.
$parity$ = the birth order of this child.
$faminc$ = annual family income.
$motheduc$ = years of schooling for the mother.
$fatheduc$ = years of schooling for the father.

Let us test the null hypothesis that, after controlling for *cigs*, *parity*, and *faminc*, parents' education has no effect on birth weight. This is stated as H_0: $\beta_4 = 0$, $\beta_5 = 0$, and so there are $q = 2$ exclusion restrictions to be tested. There are $k + 1 = 6$ parameters in the unrestricted model (4.42); so the *df* in the unrestricted model is $n - 6$, where n is the sample size.

We will test this hypothesis using the data in BWGHT. This data set contains information on 1,388 births, but we must be careful in counting the observations used in testing the null hypothesis. It turns out that information on at least one of the variables *motheduc* and *fatheduc* is missing for 197 births in the sample; these observations cannot be included when estimating the unrestricted model. Thus, we really have $n = 1{,}191$ observations, and so there are $1{,}191 - 6 = 1{,}185$ *df* in the unrestricted model. We must be sure to use these *same* 1,191 observations when estimating the restricted model (not the full 1,388 observations that are available). Generally, when estimating the restricted model to compute an F test, we must use the same observations to estimate the unrestricted model; otherwise, the test is not valid. When there are no missing data, this will not be an issue.

The numerator *df* is 2, and the denominator *df* is 1,185; from Table G.3, the 5% critical value is $c = 3.0$. Rather than report the complete results, for brevity, we present only the R-squareds. The R-squared for the full model turns out to be $R^2_{ur} = .0387$. When *motheduc* and *fatheduc* are dropped from the regression, the R-squared falls to $R^2_r = .0364$. Thus, the F statistic is $F = [(.0387 - .0364)/(1 - .0387)](1{,}185/2) = 1.42$; since this is well below the 5% critical value, we fail to reject H_0. In other words, *motheduc* and *fatheduc* are jointly insignificant in the birth weight equation. Most statistical packages these days have built-in commands for testing multiple hypotheses after OLS estimation, and so one need not worry about making the mistake of running the two regressions on different data sets. Typically, the commands are applied after estimation of the unrestricted model, which means the smaller subset of data is used whenever there are missing values on some variables. Formulas for computing the F statistic using matrix algebra—see Appendix E—do not require estimation of the retricted model.

註: 上述 F 統計量的計算是採 R^2 式，但這並非必要!