MULTIPLE LINEAR REGRESSION MODEL III: Testing Multiple Linear Restrictions, F Test, LM test

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics I

Testing Hypotheses about a Single Linear Combination

- Is one year at a junior college (2-year higher education) worth one year at a university (4-year)?

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

jc: number of years attending a junior college, univ: number of years at a 4-year college, exper: experience (year)

- Null hypothesis:

$$H_0 : \beta_1 = \beta_2 \;\Leftrightarrow\; \beta_1 - \beta_2 = 0$$

- Alternative hypothesis

$$H_0 : \beta_1 < \beta_2 \;\Leftrightarrow\; \beta_1 - \beta_2 < 0$$

Testing Hypotheses about a Single Linear Combination

- Since the null hypothesis contains a single linear combination we can use t test:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

- The standard error is given by:

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{Var}(\hat{\beta}_1 - \hat{\beta}_2)}$$

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

- To compute this we need to know the covariances between OLS estimates.

Testing Hypotheses about a Single Linear Combination

- An alternative method to compute $se(\hat{\beta}_1 - \hat{\beta}_2)$ is to estimate re-arranged regression.
- Let $\theta = \hat{\beta}_1 - \hat{\beta}_2$. Now the null and alternative hypotheses are:

$$H_0 : \theta = 0, \quad H_1 : \theta < 0$$

- Substituting $\beta_1 = \theta + \hat{\beta}_2$ into the model we obtain:

$$\begin{aligned} y &= \beta_0 + (\theta + \hat{\beta}_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ &= \beta_0 + \theta x_1 + \beta_2(x_1 + x_2) + \beta_3 x_3 + u \end{aligned}$$

Example: twoyear.gdt

Estimation Results

$$\widehat{\log(\text{wage})} = \underset{(0.27)}{1.43} + \underset{(0.031)}{0.098}\,\text{jc} + \underset{(0.035)}{0.124}\,\text{univ} + \underset{(0.008)}{0.019}\,\text{exper}$$
$$n = 285 \quad R^2 = 0.243$$

Computing se

$$\widehat{\log(\text{wage})} = \underset{(0.27)}{1.43} - \underset{(0.018)}{0.026}\,\text{jc} + \underset{(0.035)}{0.124}\,\text{totcoll} + \underset{(0.008)}{0.019}\,\text{exper}$$
$$n = 285 \quad R^2 = 0.243$$

- Note: $totcoll = jc + univ$. $se(\theta) = se(\hat{\beta}_1 - \hat{\beta}_2) = 0.018$.
- t statistic: $t = -0.026/0.018 = -1.44$, p-value$= 0.075$
- There is some but not strong evidence against H_0. The return on an additional year of eduction at a 4-year college is statistically larger than than the return on an additional year at a 2-year college.

Testing Multiple Linear Restrictions: the F Test

- The t statistic can be used to test whether an unknown population parameter is equal to a given constant.
- It can also be used to test a single linear combination on population parameters as we just saw.
- In practice, we would like to test multiple hypotheses about the population parameters.
- We will use the F test for this purpose.

Exclusion Restrictions

- We want to test whether a group of variables has no effect on the dependent variable.
- For example, in the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

we want to test

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$

$$H_1 : \beta_3 \neq 0, \beta_4 \neq 0, \beta_5 \neq 0$$

- The null hypothesis states that x_3, x_4 and x_5 together have no effect on y after controlling for x_1 and x_2.
- H_0 puts 3 exclusion restrictions on the model.
- The alternative holds if at least one of β_3, β_4 or β_5 is different from zero.

Exclusion Restrictions

UnRestricted Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$
$$SSR_{ur}, \quad R^2_{ur}$$

Restricted Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$SSR_r, \quad R^2_r$$

- The restricted model is obtained under H_0.
- We can estimate both models separately and compare SSRs using the F statistic.

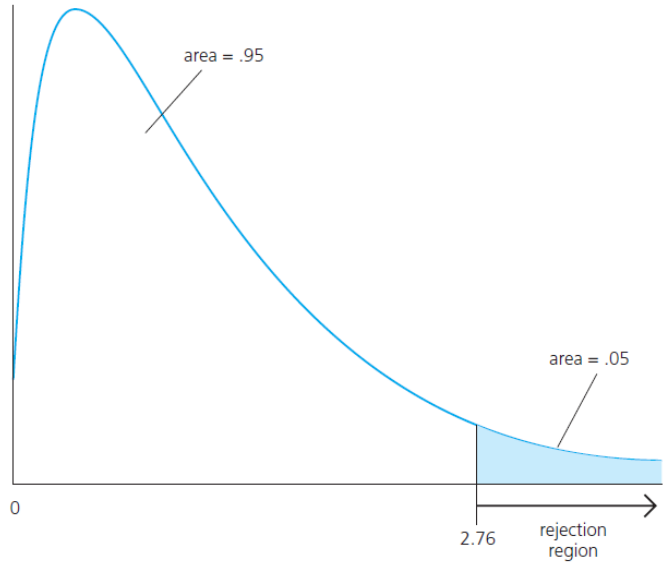
Testing Multiple Linear Restrictions

The F-test statistic

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

- SSR_r: restricted model's SSR, SSR_{ur}: unrestricted model's SSR.
- $q = df_r - df_{ur}$: total number of restrictions, also the degrees of freedom for the numerator.
- The degrees of freedom for denominator:df_{ur} obtained from the unrestricted model.
- Decision rule: If $F > c$ REJECT H_0 RED. The critical value c, is obtained from the $F_{k,n-k-1}$ distribution using $100\alpha\%$ significance level.

5% Rejection Region for the $F(3, 60)$ Distribution



The F Test

- The F test for exclusion restrictions can be useful when the variables in the group are highly correlated.
- For example, suppose we want to test whether firm performance affect salaries of CEOs. Since there are several measures of firm performance using all of these variables in the model may lead to multicollinearity problem.
- In this case individual t tests may not be helpful. The standard errors will be high due to multicollinearity.
- But F test can be used to determine whether as a group the firm performance variables affect salary.

Relationship between t and F Statistics

- Conducting an F test on a single parameter gives the same result as the t test.
- For the two-sided test of $H_0 : \beta_j = 0$ the F test statistic has $q = 1$ degrees of freedom for the numerator and the following relationship holds:

$$t^2 = F$$

- For two-sided alternatives:

$$t^2_{n-k-1} \sim F(1, n-k-1)$$

- But in testing hypotheses using a single parameter t test is easier and more flexible and also allows for one-sided alternatives.

R^2 Form of the F Statistic

- The F test statistic can be written in terms of R^2s from the restricted and unrestricted models instead of SSRs.
- Recall that

$$SSR_r = SST(1 - R_r^2), \quad SSR_{ur} = SST(1 - R_{ur}^2)$$

- Substituting into the F statistic an rearranging we obtain:

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

- R_{ur}^2: coefficient of determination from the **u**nrestricted model,
- R_r^2: coefficient of determination from the **r**estricted model
- $R_{ur}^2 \geq R_r^2$

F Test: Example

Parents' education in a birth-weight model: bwght.gdt

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

- Dependent variable: $y =$ birth weight of newly born babies, in pounds
- Explanatory variables:
  - x_1: average number of cigarettes the mother smoked per day during pregnancy,
  - x_2: the birth order of this child,
  - x_3: annual family income,
  - x_4: years of schooling for the mother,
  - x_5: years of schooling for the father.
- We want to test: $H_0 : \beta_4 = 0, \beta_5 = 0$, parents' education has no effect on birth weight, ceteris paribus.

Unrestricted Model: bwght.gdt

Model 1: OLS, using observations 1–1388 ($n = 1191$)
Missing or incomplete observations dropped: 197
Dependent variable: bwght

	Coefficient	Std. Error	t-ratio	p-value
const	114.524	3.72845	30.7163	0.0000
cigs	−0.595936	0.110348	−5.4005	0.0000
parity	1.78760	0.659406	2.7109	0.0068
faminc	0.0560414	0.0365616	1.5328	0.1256
motheduc	−0.370450	0.319855	−1.1582	0.2470
fatheduc	0.472394	0.282643	1.6713	0.0949

Mean dependent var	119.5298	S.D. dependent var	20.14124
Sum squared resid $\boldsymbol{SSR_{ur}}$	**464041.1**	S.E. of regression	19.78878
$\boldsymbol{R_{ur}^2}$	**0.038748**	Adjusted R^2	0.034692

Restricted Model: bwght.gdt

Model 2: OLS, using observations 1–1191
Dependent variable: bwght

	Coefficient	Std. Error	t-ratio	p-value
const	115.470	1.65590	69.7325	0.0000
cigs	−0.597852	0.108770	−5.4965	0.0000
parity	1.83227	0.657540	2.7866	0.0054
faminc	0.0670618	0.0323938	2.0702	0.0386

Mean dependent var	119.5298	S.D. dependent var	20.14124
Sum squared resid SSR_r	**465166.8**	S.E. of regression	19.79607
R_r^2	**0.036416**	Adjusted R^2	0.033981

Example: continued

- F statistic in SSR form

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(465167 - 464041)/2}{464041/(1191-5-1)} = 1.4377$$

- F statistic in R^2 form

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)} = \frac{(0.0387 - 0.0364)/2}{(1-0.0387)/1185} = 1.4376$$

- Critical value: F(2, 1185) distribution at 5% level, $c = 3$, at 10% level $c = 2.3$
- Decision: We fail to reject H_0 at these significance levels. Parents' education has no effect on birth weights. They are **jointly insignificant**.

Example: Automatic Computation in R

```
> library(wooldridge)
> library(car)
# estimate the unrestricted model
> unrestr <- lm(bwght ~  cigs + parity + faminc + motheduc + fatheduc, data=bwght)
# define linear restriction
> linearHypothesis(unrestr, c("motheduc=0", "fatheduc=0"))
Linear hypothesis test

Hypothesis:
motheduc = 0
fatheduc = 0

Model 1: restricted model
Model 2: bwght ~ cigs + parity + faminc + motheduc + fatheduc

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1187 465167
2   1185 464041  2    1125.7 1.4373  0.238
```


Overall Significance of a Regression

- We want to test the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- None of the explanatory variables has an effect on y. In other words they are jointly insignificant.
- Alternative hypothesis states that at least one of them is different from zero.
- According to the null the model has no explanatory power. Under the null hypothesis we obtain the following model

$$y = \beta_0 + u$$

- This hypothesis can be tested using the F statistic.

Overall Significance of a Regression

- The F test statistic is

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

- The R^2 is just the usual coefficient of determination from the unrestricted model.
- Standard econometrics software packages routinely compute and report this statistic.
- In the previous example

$$F - statistic(5, 1185) = 9.5535 (p - value < 0.00001)$$

- p-value is very small. It says that if we reject H_0 the probability of Type I Error will be very small. Thus, the null is rejected very strongly.
- There is **strong evidence against** the null hypothesis which states that the variables are jointly insignificant. The regression is overall significant.

R Example

```
> unrestr <- lm(bwght ~  cigs + parity + faminc + motheduc + fatheduc, data=bwght)
> summ(unrestr)
MODEL INFO:
Observations: 1191
Dependent Variable: bwght
Type: OLS linear regression

MODEL FIT:
F(5,1185) = 9.55, p = 0.00
Rsq = 0.04
Adj. Rsq = 0.03

Standard errors: OLS
-------------------------------------------------
                      Est.   S.E.   t val.      p
----------------- -------- ------ -------- ------
(Intercept)         114.52   3.73    30.72   0.00
cigs                 -0.60   0.11    -5.40   0.00
parity                1.79   0.66     2.71   0.01
faminc                0.06   0.04     1.53   0.13
motheduc             -0.37   0.32    -1.16   0.25
fatheduc              0.47   0.28     1.67   0.09
-------------------------------------------------
```


Testing General Linear Restrictions

Example: Rationality of housing valuations: hprice1.gdt

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

- Dependent variable: $y = \log(\text{price})$
- Explanatory variables:
  - x_1: log(assess), the assessed housing value (before the house was sold)
  - x_2: log(lotsize), size of the lot, in feet.
  - x_3: log(sqrft), size of the house.
  - x_4: bdrms, number of bedrooms
- We are interested in testing
  $H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$
- The null hypothesis states that *additional characteristics do not explain house prices once we controlled for the house valuations*.

Example: Rationality of housing valuations

- Unrestricted model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

- Restricted model under $H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$

$$y = \beta_0 + x_1 + u$$

- Restricted model can be estimated using

$$y - x_1 = \beta_0 + u$$

- The steps of the F test are the same.

Example: Rationality of housing valuations

Örnek: Gretl, hprice1.gdt

- Unrestricted model: SSR_{ur}=1.822

$$\widehat{\text{lprice}} = \underset{(0.56966)}{0.263745} + \underset{(0.15145)}{1.04306}\,\text{lassess} + \underset{(0.038561)}{0.00743824}\,\text{llotsize} - \underset{(0.13843)}{0.103239}\,\text{lsqrft} + \underset{(0.022098)}{0.0338392}\,\text{bdrms}$$

$$T = 88 \quad \bar{R}^2 = 0.7619 \quad F(4, 83) = 70.583 \quad \hat{\sigma} = 0.14814$$

(standard errors in parentheses)

- Restricted Model: SSR_r=1.880

$$\widehat{\text{y1}} = \underset{(0.015671)}{-0.0848134}$$

$$T = 88 \quad \bar{R}^2 = 0.0000 \quad \hat{\sigma} = 0.14701$$

(standard errors in parentheses)

Example: Rationality of housing valuations

- Test statistic:

$$F = \frac{(1.880 - 1.822)}{1.822}\frac{83}{4} = 0.661$$

- The critical value at the 5% significance level for F(4,83) distribution: $c = 2.5$
- We fail to reject H_0.
- There is no evidence against the null hypothesis that the housing evaluations are rational.

Asymptotic Normality and Large Sample Inference

- The last classical assumption (MLR.6) states that, conditional on x variables, the error term has a normal distribution. This implies that the conditional distribution of y is also normal because linear combinations of normal random variables also follow the normal distribution.
- We do not need the normality assumption for the unbiasedness of OLS estimators. The normality assumption is required to derive the exact (finite sample) sampling distributions of OLS estimators (which are also normal).
- If the normality assumption fails, does this imply that we cannot carry out t and F tests?
- The answer is NO! If the sample size is large enough, we may be able to rely on the **Central Limit Theorem** to conclude that OLS estimators are **asymptotically normal**.
- Asymptotic = large sample = we collect more data (hence more information, $n \to \infty$)

Asymptotic t Test

- As n gets larger, the t statistic converges to the standard normal distribution (asymptotic t statistic):

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim^a \ t_{n-k-1}$$

- We may be able to relax MLR.6 Normality.
- The only requirement is that the error variance is finite: $\sigma^2 > 0$.
- But constant variance and zero conditional mean assumptions are required.
- The standard errors of the OLS estimators ($se(\hat{\beta}_j)$) shrink to zero at the rate $1/\sqrt{n}$.

A Large Sample Test: The Lagrange Multiplier (LM) Test Statistic

- In large samples, we can use the Lagrange Multiplier (LM, or score test) statistic to test linear restrictions.
- The LM statistic relies only on the estimation of the restricted model. After the restricted model is estimated an auxiliary regression is run to get the LM statistic.
- Example: Exclusion restrictions - let the unrestricted model be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

- $H_0 : \beta_3 = \beta_4 = 0$, H_1: at least one of them is not zero.
- The LM test statistic is computed by multiplying the sample size n by R^2 which is obtained from the regression of the residuals from the restricted model on all explanatory variables.

LM Test

- Under $H_0 : \beta_3 = \beta_4 = 0$ the unrestricted model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Let $\tilde{u}$ be the residuals from this regression. Solve the following auxiliary regression by running the regression of these residuals on all xs:

$$\tilde{u} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + error$$

- Let $R^2_{\tilde{u}}$ be the coefficient of determination of this regression. Then the LM test statistic is:

$$LM = nR^2_{\tilde{u}} \sim \chi^2_q$$

- Under the null hypothesis the LM statistic follows a chi-squared distribution with q degrees of freedom.
- Decision Rule: If $LM > c$ then reject H_0.

LM Test: Example

Newborn weight and parents' education: bwght.gdt

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

- Dependent variable: $y =$ weights of newly born babies (ounces)
- Explanatory variables:
  - x_1: average number of cigarettes smoked per day during pregnancy
  - x_2: parity (birth order)
  - x_3: family income
  - x_4: mother's education level, year
  - x_5: father's education level, year.
- We are interested in: $H_0 : \beta_4 = 0, \beta_5 = 0$, parents' education level does not have any effect on weights of newborns.

LM Test: Example

Step 1: **Unrestricted** Model

$$\widehat{\text{bwght}} = \underset{(1.656)}{115.470} - \underset{(0.1088)}{0.5979}\,\text{cigs} + \underset{(0.6575)}{1.8323}\,\text{parity} + \underset{(0.0324)}{0.0671}\,\text{faminc}$$

$$n = 1191 \quad R^2 = 0.036 \quad F(3, 1187) = 14.953 \quad \hat{\sigma} = 19.796$$

Save the residuals from this regression: uhat

Step 2: Run the regression of uhat on all xs

$$\widehat{\text{uhat}} = \underset{(3.729)}{-0.9456} + \underset{(0.110)}{0.0019}\,\text{cigs} - \underset{(0.659)}{0.0447}\,\text{parity} - \underset{(0.0366)}{0.011}\,\text{faminc}$$
$$- \underset{(0.319)}{0.370}\,\text{motheduc} + \underset{(0.283)}{0.472}\,\text{fatheduc}$$

$$n = 1191 \quad \boldsymbol{R^2 = 0.00242} \quad F(5, 1185) = 0.57491 \quad \hat{\sigma} = 19.789$$

LM Test: Example

Step 3: calculate the LM test statistic

$$LM = nR^2_{\tilde{u}} \sim \chi^2_q$$
$$LM = (1191)(0.00242) \sim \chi^2_2$$
$$LM = 2.88$$

5% critical value at 2 degrees of freedom is $c = 5.99$. Thus, we fail to reject H_0. Also note that p-value $= 0.24$

Reporting Regression Results

- The estimated OLS coefficients should always be reported. The key coefficient estimates should be interpreted taking into account the functional forms and units of measurement.
- Individual t statistics and F statistic for the overall significance of the regression should also be reported.
- Standard errors for the coefficient estimates can be given along with the estimates. This allows us to conduct t tests for the values other than zero and to compute confidence intervals.
- R^2 and n should always be reported. One may also consider reporting SSR and the standard error of the regression ($\hat{\sigma}$).