7. Multiple Regression Analysis: The Problem of Analysis

Three-Variable Model: Notation and Assumptions

Let us consider the following three-variable PRF as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where

Y_i is the dependent variable (regressand)

X_{2i} and X_{3i} are the regressors or the explanatory variables

u_i is the stochastic disturbance term

Remark: the subscript i is denoted the observation i from our sample data.

In case our data are time series, the subscript t will denote the t observation.

β_1 means the average value of Y when X_2 and X_3 are set equal to zero

β_2 and β_3 are called the partial regression coefficients.

We will talk about the meaning of β_1 and β_2 shortly after knowing the assumptions of the classical linear regression model (CLRM)

Under the CLRM, we assume:

1. Zero mean value of u_i

2. No serial correlation

3. Homoscedasticity

4. Zero covariance between u_i and each X variable, or

5. No specification bias or

The model is correctly specified.

6. No exact collinearity between the X variables or

By the above assumptions, we can find out the conditional expectation of Y_i:

The meaning of partial coefficients:

β_2

β_3

7.1 OLS Estimation of the Partial Regression Coefficients

In order to find the OLS estimators, we need to write down the sample regression function (SRF) corresponding to the PRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

From the FOC, we then get the normal equations:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$
$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

We therefore get:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$
$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Variance and Standard Errors of OLS Estimators

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3{}^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2} \right] * \sigma^2$$
$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_2)}$$

$$cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$

7.2 Properties of OLS Estimators

Properties of OLS Estimators (Cont:)

Properties of OLS Estimators (Cont:)

The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation R

In this section, we will study how to measure the proportion of the variation in Y explained by the variables X_2 and X_3 jointly. This is the same concept of r^2 that we have learned before.

The quantity that gives this information is known as the **the multiple coefficient of determination** and is denoted by R^2.

To derive R^2, we firstly write down the following equation:

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \tag{7.1}$$

where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i|X_{2i}, X_{3i})$.

7.1 may be written as

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \tag{7.2}$$

Squaring 7.2 on both sides and summing over the sample values, we obtain

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i{}^2 + \hat{u}_i{}^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 \end{aligned} \tag{7.3}$$

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \end{aligned} \tag{7.4}$$

The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right)$$

7.2.1 R^2 and the Adjusted R^2

It should be noted that he R^2 is a nondecreasing function of the number of explanatory variables. Thus, when the number of regressors increases, R^2 almost invariably increases and never decreases. **In other words, an additional X variable will not decrease R^2!**

To explain this fact, let we write down the definition of R^2 again:

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \end{aligned} \tag{7.5}$$

Therefore, in comparing two regression models **with the same dependent variable but differing number of X variables**, one should be very wary of choosing the model with the highest R^2.

In light of comparing two R^2 terms, we have to take into account the number of X variables present in the model. To achieve this goal , we can consider the alternative coefficient of determination, which is as follows:

k= the number of parameters in the model including the intercept term.
n= the number of observations in the sample data.

The above equation is known as **the adjusted** R^2, denoted by $\bar{R}^2$. The term adjusted means adjusted for the df associated with the sums of squares entering into 7.5.

We can rewrite the the adjusted R^2 as:

We can also get the equation which shows the relationship between $\bar{R}^2$ and R^2:

Besides R^2 and $\bar{R}^2$ as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Akaike's Information criterion and Amemiya's Prediction criteria**, which are used to select between competing models. We will discuss these criteria in greater detail later.

8. Multiple Regression Analysis: The Problem of Inference

In this chapter, we will extend the ideas of interval estimation and hypothesis testing developed there to models involving three or more variables.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

We have already known that if our objective is to do interval estimation and hypothesis testing, we need to assume that the u_i follow the normal distribution with zero mean and constant variance σ^2

With the normality assumption and the CLRM assumptions, we know that:

[1] The OLS estimations of partial regression coefficients are best linear unbiased estimators (BLUE).

[2] The estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are normally distributed with means equal to true β_1,β_2, and β_3 and variances are following:

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3{}^2 \sum x_{2i}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i}x_{3i})^2}\right] * \sigma^2$$

$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)}$$

$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1 - r_{23}^2)}$$

$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_3)}$$

Moreover, $\frac{(n-3)\hat{\sigma^2}}{\sigma^2}$ follows the χ^2 distribution with n-3 df. We can also show that, if we replace the true σ^2 by its unbiased estimator $\hat{\sigma^2}$ in the computation of the standard errors, we then get

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)}$$

follows the t distribution with n-3 df.

Example Consider the following regression:

$$\begin{array}{rcl} \widehat{\log(\text{salary})} & = & 4.32 + 0.280\ \log(\text{sales}) + 0.0174\ \text{ROE} + 0.00024\ \text{ROS} \\ se & = & (0.32)\ (0.035)\qquad\quad (0.0041)\qquad\quad (0.00054) \end{array} \tag{8.1}$$

$$R^2 = 0.283$$

where
salary =salary of CEO
sales=annual firm sales
ROE=return on equity in percent
ROS =return on firm's stock

Interpreter the partial regression coefficients

Questions What about the statistical significance of the observed results?

For the coefficient of log(sales) of 0.280, Is this coefficient statistically significant different from zero?

For the coefficient of ROE of 0.0174, Is this coefficient statistically significant different from zero?

For the coefficient of ROS of 0.00024, Is this coefficient statistically significant different from zero?

Are these three coefficients statistically significant?
To answer these questions, we have to learn the kinds of hypothesis testing.

8.1 Hypothesis Testing About Individual Regression Coefficients

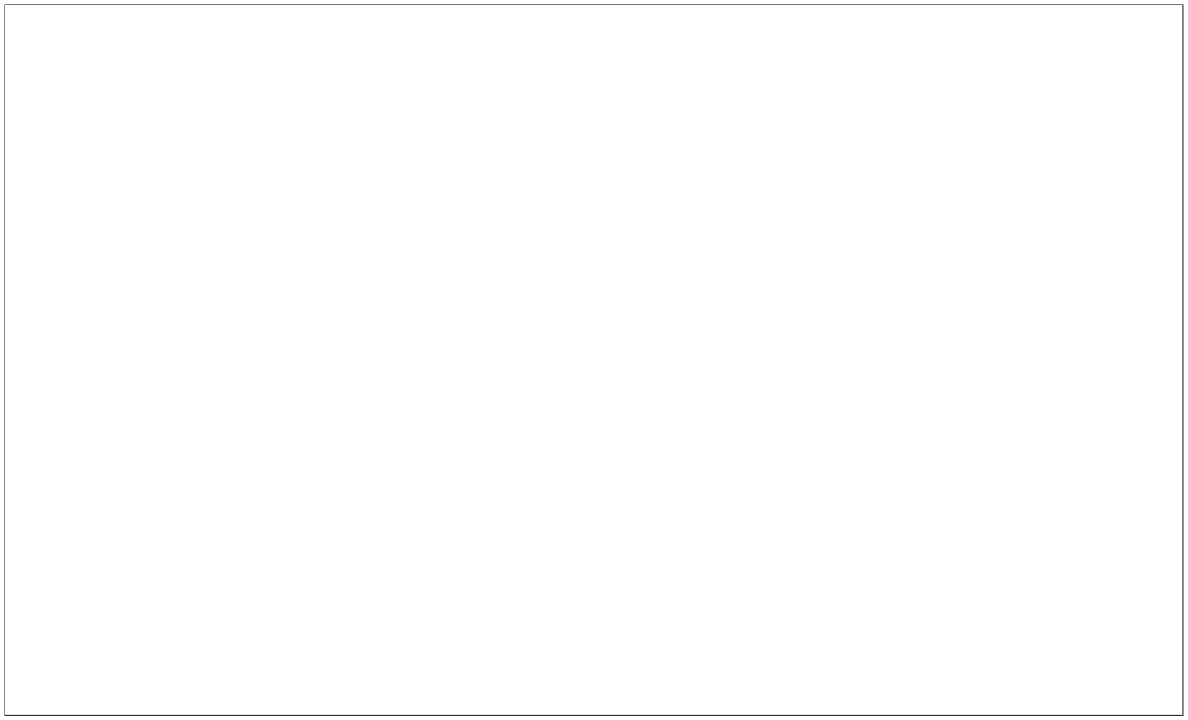
We can use the t-test to test a hypothesis about any individual partial regression coefficient.

8.1.1 Two-tail test:

Let us postulate that

$H_0: \ \beta_2 = 0$

$H_1: \ \beta_2 \neq 0$



8.1.2 One-tail test:

Let us postulate that

$H_0: \quad \beta_2 \leq 0$

$H_1: \quad \beta_2 > 0$

8.2 Testing The Overall Significance of the Sample Regression

In the previous section, we test the significance of the estimated partial regression coefficients individually, that is under the separate hypothesis that each true population partial regression coefficient was zero. But now we are interested in testing β_2 , β_3 and β_4 are jointly or simultaneously equal to zero. In other words, we would like to test the following hypothesis:

$$H_0 \;\; \beta_2 = \beta_3 = \beta_4 = 0$$

In order to reach this goal, we have to learn the following test.

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F-Test

The joint hypothesis can be tested by the **Analysis of Variance (ANOVA)** which can be demonstrated as follows:

Table 8.1: ANOVA Table for the three-variable regression model

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

Decision Rule Given the k- variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0 : \quad \beta_2 = \beta_3 = ... = \beta_k = 0$$

(i.e ., all slope coefficients are simultaneously zero) versus

H_1 Not all slope coefficients are simultaneously zero

If $F > F_\alpha(k-1, n-k)$, we reject H_o ; otherwise we cannot reject it, where $F_\alpha(k-1, n-k)$ is the critical F value at the α level of significance and (k-1) numerator df and (n-k) denominator df.

An important Relationship between R^2 and F

Table 8.2: ANOVA Table in Terms of R^2

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

Decision Rule Testing the overall significance of a regression in terms of R^2

Given the k- variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0 : \quad \beta_2 = \beta_3 = ... = \beta_k = 0$$

(i.e ., all slope coefficients are simultaneously zero) versus

H_1 Not all slope coefficients are simultaneously zero

Compute

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

If $F > F_\alpha(k-1, n-k)$, we reject H_o ; otherwise we cannot reject it, where $F_\alpha(k-1, n-k)$ is the critical F value at the α level of significance and (k-1) numerator df and (n-k) denominator df.

8.3 The "Incremental" or "Marginal" Contribution of an Explanatory Variable

Let consider the following regression:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_i$$

Having run the above regression, let us suppose we decide to add the additional variable, X_{3i}, to the model and obtain the multiple regression as follow:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Comparing between these two regressions, we might need to answer the below questions:

[1]. What are the marginal, or incremental, contribution of X_{3i}, knowing that X_{2i} is already in the model and that it is significantly related to Y_i.

[2]. Is the incremental contribution of X_{3i} statistically significant?

[3]. What is the criterion for adding variables to the model?

By contribution we mean whether the additional of the variable, X_{3i}, to the model increases ESS (and hence R^2) "significantly" in relation to the RSS. This contribution is called **the incremental, or marginal** contribution of an additional variable.

To assess the incremental contribution of X_3 after allowing for the contribution of X_2, we form

$$\begin{aligned} F &= \frac{Q_2/\text{df}}{Q_4/\text{df}} \\ &= \frac{(ESS_{new} - ESS_{old})/\text{number of new regressors}}{RSS_{new}/df(\text{=n-number of parameters in the new model})} \end{aligned} \tag{8.2}$$

Under the normality assumption of u_i and CLRM assumptions, this F value follows the F distribution with 1 and n-number of parameters in the new model.

Table 8.3: ANOVA Table To Assess Incremental Contribution of A Variable(s)

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
ESS due to X_2 alone	$Q_1 = \hat{\alpha_2}^2 \sum x_2^2$	1	$\frac{Q_1}{1}$
ESS due to the addition of X_3	$Q_2 = Q_3 - Q_1$	1	$\frac{Q_2}{1}$
ESS due to both X_2,X_3	$Q_3 = \hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i$	2	$\frac{Q_3}{2}$
RSS	$Q_4 = Q_5 - Q_3$	n-3	$\frac{Q_4}{n-3}$
TSS	$Q_5 = \sum y_i^2$	n-1	

As usual method, we can re write 8.2 in term of R^2 only. Thus the F ratio of 8.2 is equivalent to the following F ratio:

$$\begin{aligned} F &= \frac{R^2_{new} - R^2_{old}/\text{df}}{(1 - R^2_{new})/\text{df}} \\ &= \frac{(R^2_{new} - R^2_{old})/\text{number of new regressors}}{1 - R^2_{new}/df(\text{=n-number of parameters in the new model})} \end{aligned} \tag{8.3}$$

This F ratio follows the F distribution with 1 and n-number of parameters in the new model.

Example

Consider the child mortality example. We considered the behavior of child mortality (CM) in relation to per capita GNP (PGNP). There we found that PGNP has a negative impact on CM, as one would expect. Now let us bring in female literacy as measured by the female literacy rate (FLR). A priori, we expect that FLR too will have a negative impact on CM. Our sample consists of 64 countries.

In model 1, we regressed child mortality (CM) on per capita GNP (PGNP) and female literacy rate (FLR).

Model 1:

$$\begin{aligned} \widehat{CM_i} &= 263.6416 - 0.0056PGNP_i - 2.2316FLR_i \\ se &= (11.5932) \quad (0.0019) \qquad (0.2099) \quad R^2 = 0.7077 \end{aligned} \tag{8.4}$$

Now we extend this model to model 2 by including total fertility rate (TFR):

Model 2:

$$\begin{aligned} \widehat{CM_i} &= 168.3067 - 0.00555GNP_i - 1.7680FLR_i + 12.8686TFR_i \\ se &= (32.8916) \quad (0.0018) \qquad (0.2480) \qquad (?) \quad R^2 = 0.7474 \end{aligned} \tag{8.5}$$

Questions

1.How would you choose between models 1 and 2? Which statistical test would you use to answer this question? Show the necessary calculations.
2. We have not given the standard error of the coefficient of TFR. Can you find it out? (Hint: Recall the relationship between the t and F distributions.

8.4 Testing the Equality of Two Regression Coefficients

Suppose we have the following model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + ... + \beta_k X_{ki} + u_i$$

We would like to test the hypotheses:

$$H_0 : \ \beta_3 = \beta_4 \ \text{ or } \ (\beta_3 - \beta_4) = 0$$
$$H_1 : \ \beta_3 \neq \beta_4 \ \text{ or } \ (\beta_3 - \beta_4) \neq 0$$

Under the classical assumptions, it can be shown that:

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{se(\hat{\beta}_3 - \hat{\beta}_4)}$$

where the t follows the t distribution with (n-k) df because the above equation is a k-variable model, where k is the total number of parameters estimated, including the constant term.

The $se(\hat{\beta}_3 - \hat{\beta}_4)$ is calculated from the following formula:

$$se(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{var(\hat{\beta}_3) + var\hat{\beta}_4 - 2cov(\hat{\beta}_3, \hat{\beta}_4)}$$

Example

among other things, you were asked to consider the following demand function for chicken:

$$\begin{aligned} \widehat{\ln Y_t} &= 2.0328 + 0.4515 \ln X_{2t} - 0.3772 \ln X_{3t} \\ se &= (0.1162) \quad (0.0247) \qquad (0.0635) \quad R^2 = 0.9801 \end{aligned} \tag{8.6}$$

where Y = per capita consumption of chicken, lb, X_2 = real disposable per capita income, \$, X_3 = real retail price of chicken per lb.

Question

For the above demand function, how would you test the hypothesis that the income elasticity is equal in value but opposite in sign to the price elasticity of demand? Show the necessary calculations. [Note: $\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00142$. and the sample data = 23 observations]

8.5 Restricted Least Squares: Testing Linear Equality Restriction

In economic theories, the coefficients in a regression model need to satisfy some linear equality restrictions, For example, in microeconomics, consider the Cobb-Douglas production function:

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$$

where Y=output, X_2 = labor input, and X_3=capital input. We can transform the above equation to be the log form as:

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

where $\beta_0 = \ln \beta_1$

Now, if there are the constant returns to scale, economic theory would suggest that

$$\beta_2 + \beta_3 = 1$$

which is an example of a linear equality restriction.
In order to test the above linear equality restriction, we can follow two approaches which are:

[1]. The t-test approach

[2]. The F-test approach: Restricted Least Squares.

First Approach: The t-Test
A test of the hypothesis or restriction can be conducted by the t-test:

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$

where the t follows the t distribution with (n-k) df for a k-variable model, where k is the total number of parameters estimated, including the constant term. In this case, df=n-3.

The $se(\hat{\beta}_2 + \hat{\beta}_3)$ is calculated from the following formula:

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{var(\hat{\beta}_2) + var\hat{\beta}_3 + 2cov(\hat{\beta}_2, \hat{\beta}_3)}$$

Example
Consider the Cobb-Douglas production function to the Mexican economy (1955-1974: n=20):

$$\begin{aligned} \widehat{\ln GDP_t} &= -1.6524 + 0.3397 \ln Labor_t + 0.8460 \ln Capital_t \\ t &= (-2.7259) \quad (1.8295) \qquad\quad (9.0625) \quad R^2 = 0.9951 \;\; RSS_{UR} = 0.0136 \end{aligned} \tag{8.7}$$

where GDP = Real GDP, Millions of 1960 pesos, *Labor* = Employment, Thousands of People, *Capital* = Fixed Capital, Millions of 1960 pesos.

Question
As you can see, the output/labor elasticity is about 0.34 and the output/capital elasticity is about 0.85. If we add these coefficients, we obtain 1.19, suggesting that perhaps the Mexican economy during the stated time period was experiencing increasing returns to scale. However, we do not know if 1.19 is statistically different from 1.

Therefore, we have to test this linear equality restriction.

8.6 The F-Test Approach: Restricted Least Squares

From the Cobb-Douglas production function:

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \tag{8.8}$$

if there are the constant returns to scale, economic theory would suggest that

$$\beta_2 + \beta_3 = 1$$

We can rewrite it as:

$$\beta_2 = 1 - \beta_3$$

or

$$\beta_3 = 1 - \beta_2$$

Using either of these equalities, we can eliminate one of the β coefficients. Therefore, we can rewrite the Cobb-Douglas production function as:

$$\ln (Y_i/X_{2i}) = \beta_0 + \beta_3 \ln (X_{3i}/X_{2i}) + u_i \tag{8.9}$$

where $\frac{Y_i}{X_{2i}}$=output/labor ratio
$\frac{X_{3i}}{X_{2i}}$= capital labor ratio.

It should be noted that:
8.8 is known as **unrestricted Least Squares (URLS)**
8.9 is known as **restricted Least Squares (RLS)**

We can compare the unrestricted and restricted least-squares regressions by applying the F-test as follows:

$\sum \hat{U}_{UR}^2$=RSS of the unrestricted regression 8.8

$\sum \hat{U}_{R}^2$ = RSS of the restricted regression 8.9

m=number of linear restrictions (in this example, we have 1 restriction)

k= number of parameters in the unrestricted regression

n= number of observations

Then, we have

$$\begin{aligned} F &= \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} \\ &= \frac{(\sum \hat{U}_R^2 - \sum \hat{U}_{UR}^2)/m}{\sum \hat{U}_{UR}^2/(n-k)} \end{aligned} \tag{8.10}$$

follows the F-distribution with m, (n-k) df.

We can also rewrite the F-test in terms of R^2 as follows:

$$F = \frac{R_{UR}^2 - R_R^2/\text{m}}{(1-R_{UR}^2)/\text{n-k}} \tag{8.11}$$

Example
Consider the Cobb-Douglas production function to the Mexican economy(1955-1974: n=20):

$$\begin{aligned} \widehat{\ln GDP_t} &= -1.6524 + 0.3397 \ln Labor_t + 0.8460 \ln Capital_t \\ t &= (-2.7259) \quad (1.8295) \qquad (9.0625) \quad R^2 = 0.9951 \;\; RSS_{UR} = 0.0136 \end{aligned} \tag{8.12}$$

where GDP = Real GDP, Millions of 1960 pesos, *Labor* = Employment, Thousands of People, *Capital* = Fixed Capital, Millions of 1960 pesos.

The restriction of constant return to scale, which gives the following regression:

$$\begin{aligned} \widehat{\ln (GDP/Labor)_t} &= -0.4947 + 1.0153 \ln (Capital/Labor)_t \\ t &= (-4.0612) \quad (28.1056) \quad R_R^2 = 0.9777 \;\; RSS_R = 0.0166 \end{aligned} \tag{8.13}$$

8.7 Testing for Structural or Parameter Stability of Regression Models: The Chow Test

Sometime when we estimate the regression model, it may happen that there is a **Structural Change** in the relationship between the regressand Y and the regressors X's, especially the model involving time series data. The structural change may be due to the external forces (i.e the financial crisis of 2007-2008) or due to policy changes (such as the switch from a fixed exchange rate system to a flexible exchange rate system in 1997).

The question is **"How do we figure out that there is a structural change in our sample data?"**

To answer this question, consider the following example.

Based on the sample data, we found out that in 1982 the United State suffers its worst peacetime regression. This event might disturb the relationship between savings and DPI.

To see this effect, we can divide our sample data into two time periods: 1970-1981 (Pre-1982 crisis) and 1982-1995 (Post-1982 crisis).

Therefore we have three possible regressions:

Time period 1970-1981: $Y_t = \beta_1 + \beta_2 X_t + u_1 t$ where $n_1 = 12$

Time period 1982-1995: $Y_t = \gamma_1 + \gamma_2 X_t + u_2 t$ where $n_2 = 14$

Time period 1970-1995: $Y_t = \alpha_1 + \alpha_2 X_t + u_t$ where $n = n_1 + n_2 = 26$

For our sample data, we can get the following results:

Time period 1970-1981:

$$\begin{aligned} \hat{Y}_t &= 1.0161 + 0.0803 X_t \\ t &= (0.00873) \quad (9.6015) \end{aligned} \tag{8.14}$$

$R^2 = 0.9021 \quad RSS_1 = 1785.032 \quad df = 10$

Time period 1982-1995:

$$\begin{aligned} \hat{Y}_t &= 153.4947 + 0.0148 X_t \\ t &= (4.6922) \quad (1.7707) \end{aligned} \tag{8.15}$$

$R^2 = 0.2971 \quad RSS_2 = 10,005.22 \quad df = 12$

Time period 1970-1995:

$$\begin{aligned} \hat{Y}_t &= 62.4226 + 0.0376 X_t \\ t &= (4.8917) \quad (8.8937) \end{aligned} \tag{8.16}$$

$R^2 = 0.7672 \quad RSS_3 = 23,248.30 \quad df = 24$

We can apply **the Chow test** to investigate the structural changes that may be caused by differences in the intercept or the slope coefficient or both.

The chow test assumes that:

[1] $u_{1t} \sim N(0, \sigma^2)$ and $u_{2t} \sim N(0, \sigma^2)$

[2] The two error terms u_{1t} and u_{2t} are independently distributed.

Chow Test

$$H_0: \text{ There is no structural change in the model}$$

$$H_1: \text{ There is structural change in the model}$$

Then, we need to construct the F-ratio:

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1 + n_2 - 2k)} \tag{8.17}$$

where the F ratio follows the F distribution with k and $(n_1 + n_2 - 2k)$ df in the numerator and denominator, respectively.

We do not reject the null hypothesis of parameter stability (i.e no structural change) if the computed F value does not exceed the critical value F value obtained from the F table.