

For comments, corrections, etc...Please contact Ahnaf Abbas: ahnaf@uaemath.com

This is an open source document. Permission is granted to copy, distribute and/or modify this document

under the terms of the GNU Free Documentation License, <http://www.gnu.org/copyleft/fdl.html>

Version 1.2 or any later version published by the Free Software Foundation.

International Institute for Technology and Management



Unit 76: Management Mathematics

Handout #12

Econometrics

Definition and Scope

Econometrics is about measuring economic relationships. It is a combination of economy theory, mathematical economics and statistics, but it is completely distinct from each one of these branches of science.

Econometrics is the overlap of economic theory, maths and statistics and, but in fact there are reasons why it should be considered a subject in its own right. Firstly, much of economic theory is qualitative – the law of demand suggests that as price increases demand will fall, but does not tell us how much. Econometrics will tell us how much the quantity demanded will fall. Secondly, mathematical economists will create mathematical models but will not empirically verify their models – Econometrics will translate models into forms that can be tested and estimated numerically. Thirdly, statistics usually use data generated from experiments but economics data is rarely generated in such a way but is instead collected by a range of public and private agencies, by questionnaire or observation, and are usually non-experimental, and thus likely to contain problems of measurement error. Econometrics has methods available to deal with “dirty” data and other data problems.

The Econometric Model : $y = \text{Constant} + \beta_2 X_2 + \beta_3 X_3 + e$

Economic relationships: is a relation between two economic variables:

Dependent variable, y , is focus of study (explain changes in y).

Example : Dependent variable, y : Product demand

Independent Variables X_2 and X_3 or predictors

Exogenous(alternatively :Endogenous):Advertising, competition,prices,competitor prices ,Economic conditions.

Exogenous variable: Its value is determined externally from the system of equations of an econometric model.

Endogenous variable: At least part of its value is determined internally in the system of equations of an econometric model.

Example:Advertising(Endogenous) ,Competition(Exogenous) ,Sales(Endogenous)

Econometric methodologies:

1. Statement of theory or hypothesis (e.g. the law of demand)
2. Specification of the mathematical model (eg $Q = \alpha + \beta P$)
3. Specification of the statistical or econometric model (eg $Q = \alpha + \beta P + u$)
4. Collection of data (eg from published sources or own survey)
5. Parameter estimation (eg $\hat{Q} = \hat{\alpha} + \hat{\beta}P$)
6. Testing of hypotheses (eg is $\beta < 0$?)
7. Forecasting/Prediction

you will see these as 5 stages later when we discuss Producing Econometric Models. A statistical method called **regression** analysis is used to estimate the relationship between various economic variables.

Purpose of regression analysis

- Model a relationship among economic variables, such as $y = f(x)$. Between independent y and dependent x .
- Measure the error RSE (*residual standard error*) in using that relationship to Forecast or predict the value of one variable, y , based on the value of another variable, x .
- Measure the degree of association(i.e. correlation) between the variables.
 r = coefficient of correlation
 R^2 = coefficient of determination.

Statistical Background

1. The covariance between two random variables, X and Y , measures the linear association between them. It is the mean of the product of the deviations of two numbers from their respective means:

$$\text{Cov}(X,Y) = \frac{\sum (X_t - \bar{X})(Y_t - \bar{Y})}{n - 1}$$

Note that variance is a special case of covariance: $\text{Var } X = \text{Cov}(X,X)$

2. The correlation between two random variables X and Y is their covariance divided by the square roots of their respective variances(product of their standard deviations).

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}X}\sqrt{\text{Var}Y}} = \frac{\text{Cov}(X,Y)}{S_X S_Y}$$

Correlation Coefficient is a pure number falling between -1 and 1.

3. The Regression Model(Ordinary Least Squares OLS)

Recall that regression analysis allows us to explore relations between variables: it allows us to answer questions like is there a positive or negative relationship between X and Y? Is the relationship strong? Is it statistically significant? What happens to Y if X increases by 1%? When we move beyond the simple linear regression analysis we can also ask questions like what happens to Y if X increases, controlling for Z? This is a very powerful tool, and important for analyzing data, testing theory and for providing policy suggestions.

Example: TV advertisements on Friday night and Sales of cars the next Saturday. We plot TV ads against Sales to look first at the correlation and see that there probably exists a positive relationship between TV ads and Sales, ie the more ads on a Friday night the higher sales were the following Saturday. This plot is shown below.

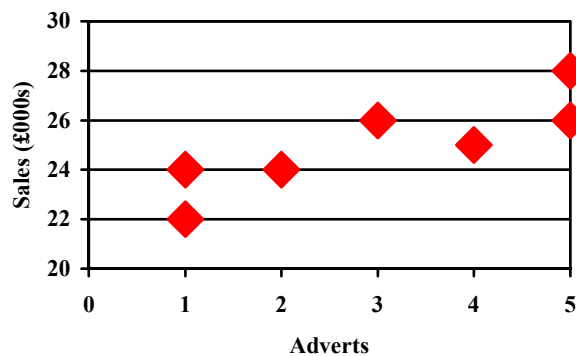


Figure 1: TV ads and Sales

Recall also that it may be possible to see that we can draw a straight line through the data, ie draw a line with the general equation $Y=a+bX$, where a is the intercept and b the slope.

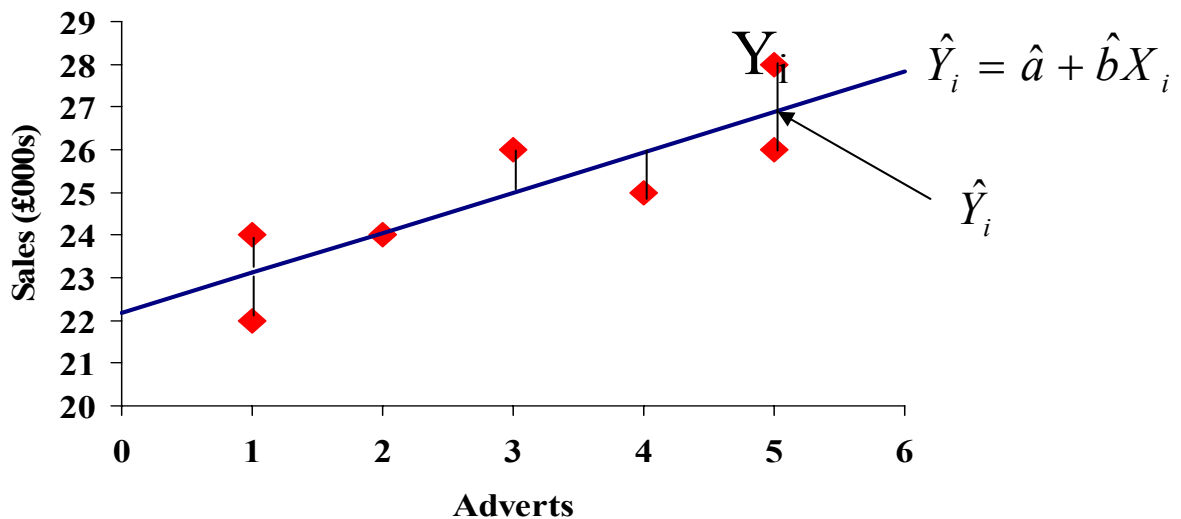


Figure 2: fitting a line through the data

Where we draw it is somewhat arbitrary: we could choose different intercepts and different slopes and draw many lines through the data: which is best?

There is however a method for drawing a line that best fits the data, and that method is called ordinary least squares or OLS for short. OLS gets its name because it finds the line that fits the data by

“minimising the sum of squared errors”.

To see this, imagine a line plotted through the data series. Let the line be called $\hat{Y}_i = \hat{a} + \hat{b}X_i$, that is we predict sales (\hat{Y}_i) is a linear function of TV Ads, X_i , by estimating what \hat{a}, \hat{b} are. **See Figure 3.** Look at the gaps between the line (\hat{Y}_i) and each data point (Y_i) – these are called the regression errors or the regression residuals. Some of these residuals or errors are positive (ie the data point lies above the line, so $Y_i > \hat{Y}_i$) and others are negative (ie the data point lies below the line, so $Y_i < \hat{Y}_i$).

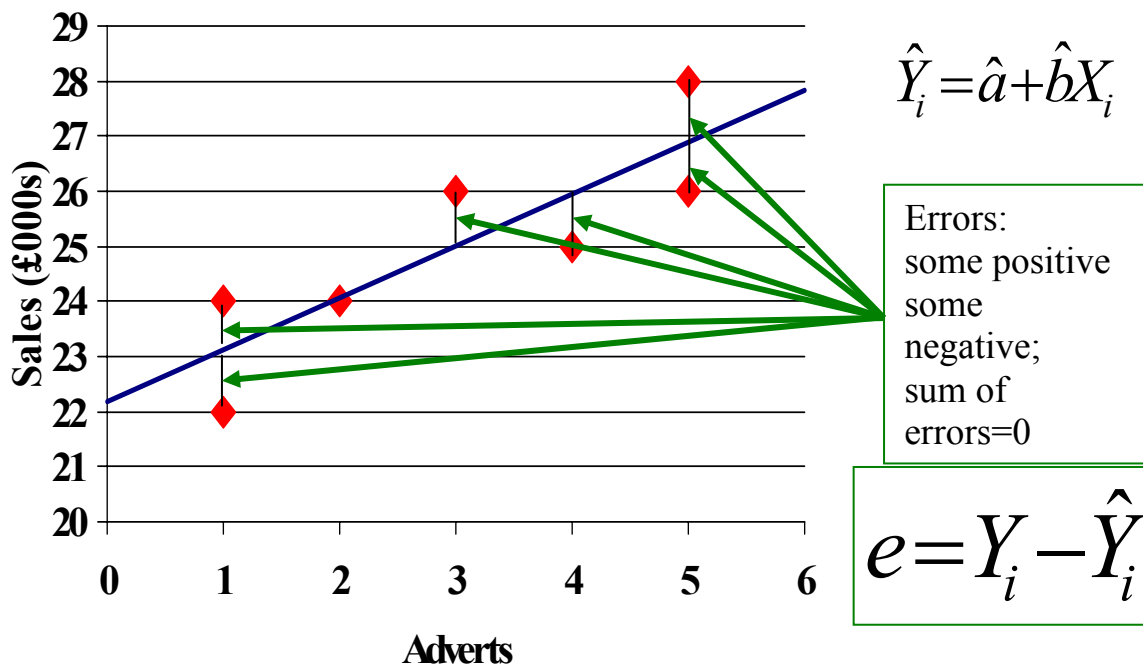


Figure 3: minimising the regression errors

OLS does two things:

- it finds the line that makes the sum of these errors of residuals equal to zero – so they cancel each other out
- it finds the line that minimizes the sum of the squared errors or residuals – i.e. squares all the gaps (so they all become positive values), adds them up and finds the values of the slope and intercept that minimizes this sum.

Sometimes the OLS line will not be a very good fit – there might not be an identifiable relationship between X and Y, perhaps for theoretical reasons or because our data contains some values that don't fit very well with the rest of the data (perhaps the weather has been too bad for people to go and buy cars) .

We can use the simple linear case again to illustrate. We want to know how well the line fits the data. One way of doing this is to consider the diagram in **Figure4**, and in particular the gap between actual Sales on a each day and mean Sales for the period. If we add up all the squares of these gaps we have the Total Sum of Squares, **TSS**, and we can see that this can be broken into two parts. The first part is the sum of squared errors or residuals which we have already looked at and which we have tried to minimize – call this **RSS**. The second part is called the explained sum of squares, **ESS**, and is the sum of the squares of the gap between the regression line and mean sales. (some text books confusingly call the residual sum of squares the error sum of squares – ie ESS for short, and the explained sum of squares the regression sum of squares, ie RSS for short: there is no convention, so be careful !)

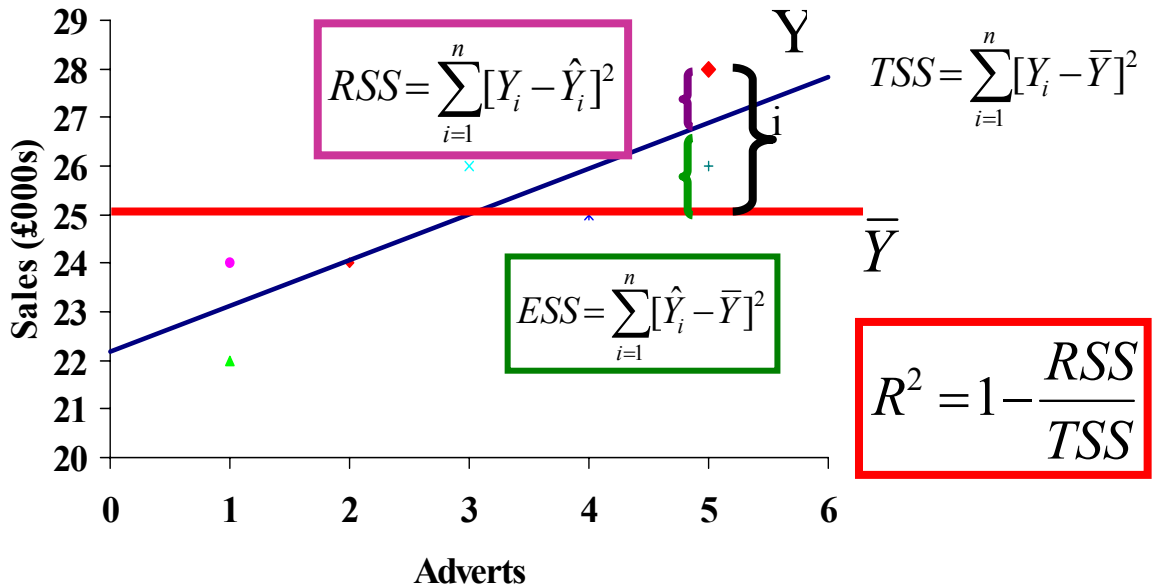


Figure 4: Goodness of fit

The better we have done at **minimizing RSS the better the goodness of fit**

So $TSS = RSS + ESS$, ie total = residual + explained

We can divide everything through by TSS, to give: $1 = RSS/TSS + ESS/TSS$

and re-arrange to give us : $ESS/TSS = 1 - RSS/TSS$

If the line is a really good fit then RSS will be very small compared to TSS, so the ratio RSS/TSS will be *very small*, hence ESS/TSS will be close to one. The closer it is to 1 the better the fit. ESS/TSS is better known as the **R²** coefficient, or the *coefficient of determination*. In technical terms the R^2 measures the percentage of variation in the Y variable that is explained by the independent variables.

4. Autocorrelation: known also as **serial correlation**

An autocorrelation measures the association between two sets of observations separated by a lag.

Example: The demand of a product (e.g. beer) related to the demand for the same day last week (this Saturday versus Saturday last week).

The most common cause of autocorrelation errors is that the model is mis-specified by omitting a variable. If this omitted variable is itself auto correlated then the error term will mirror this behavior.

The Durbin-Watson test is used to detect auto correlated errors. (More on this later)

5. Multicollinearity

When the independent variables in a multi regression model are highly correlated with each other.

Producing an Econometrics model

Application of econometric methods consists of five Stages :

1. Formulation :

- Choice of the variables: for example demands & advertising expenditure.
- Specifications of variables: exogenous & endogenous.
- Mathematical form of the equation : linear & non-linear (usually called *Functional Forms*).

Functional Forms:

-Linear Form : $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + e$

Example : Sales = $\alpha + \beta_1(\text{Advertising}) + \beta_2(\text{Personal Income}) + \beta_3(\text{age}) + e$

-log-log : Consider the following 'exponential' regression model: $Y_i = \alpha X_i^{\beta_1} e^{\varepsilon_i}$ which we can express as a linear (in logs) regression model by taking natural logarithms of both sides: $\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$ where $\beta_0 = \ln \alpha$

Example: Coffee demand function: $\ln Y_i = -0.7774 - 0.2530 \ln X_i$ where Y_i : Coffee consumption in cups per day & X_i = Coffee price per pound.

-linear-log: Take an example from labour economics. The theory of human capital investment says that individuals will invest in education because it raises their productivity, and higher productivity raises their potential wages in the labour market.

$$W_i = Y_0 e^{\beta_1 S_i} e^{\varepsilon_i}$$

where W is income or earnings, and S is the number of years of schooling (education). Y_0 represents earnings in the absence of all education.

Taking the logs of both sides: $\ln W_i = \beta_0 + \beta_1 S_i + \varepsilon_i$ where $\beta_0 = \ln Y_0$

2. Estimation:

- Data collection.
- Select an estimator: Ordinary least square method (OLS).
- Estimate the regression model using the chosen estimator.
- Test whether the assumptions made are valid (in which case the regression model is statistically well-specified) and the estimator will have the desired properties.

Regression output:

The output from an OLS estimation procedure will give the following information:

-A constant : α

-Regression Coefficients: β_1, β_2, \dots

-Standard error of **Y** estimate (residual error): $\sqrt{\frac{e^2}{n-2}}$; $e = Y - \hat{Y}$

-Standard Error of Regression Coefficients **SE**(β) : is used to measure how much error there is in these estimates.

- Coefficient of Determination: **R**²(we want this close to 1)

-Adjusted **R**² : \bar{R}^2 used to judge the relative accuracy of any model.

for e.g. if $\bar{R}^2 = 0.9375$ then the model has explained 93.75% of the original variance.

-Number of Observations : **n**

-Degrees of freedom: **k**

- Durbin-Watson : **DW** used to detect auto correlated errors;should be close to 2. (More on this later)

- The F-value : used to measure the overall statistical significance of the relationship between the variables(joint testing; More on this later).

Estimation Details: Testing the coefficients for being non zero

Common procedure :

H₀ : $\beta = 0$; the true regression coefficient is not statistically significantly different than 0.(we want to be able to reject this one)

H₁ : $\beta \neq 0$; the true regression coefficient is statistically significantly different than 0.

A test statistic, *t* , is calculated and compared to a ***t*-table(for less than 30 observations)** value assuming the null hypothesis is true.

The statistic t = coefficient /Standard error of coefficient = β / SE(β)

If $|t| < t$ -table $\alpha = 0.05$ or less : Accept H₀

If $|t| > t$ -table $\alpha = 0.05$ or less : Reject H₀

Use the following **for observations > 30 : $t > Z_{\alpha/2}$ to reject H₀** where Z is the tabulated **Normal-value.**

Example:

In order to model the demand for motor vehicles, an econometrician proposes the general linear regression model

$$Y_t = \beta_0 + \beta_P P_t + \beta_E E_t + \beta_B B_t + u_t \quad ; t = 1965, 1966, \dots, 1986$$

where

Y is the logarithm of an index of consumer expenditure on motor vehicles, spares and accessories at constant prices,

P is the logarithm of a relative price index of motor vehicles,

E is the logarithm of real total household expenditure,

B is the logarithm of a relative price index of public road transport,

u is the error term.

This model and a restricted version of the model were fitted using ordinary least squares (OLS) to annual data covering the period 1965-1986 and the following results were obtained

$$\hat{Y}_t = 6.27 - 0.705 P_t \quad (A)$$

(0.56) (0.067)

$$R^2 = 0.0738, \text{RSS} = 0.636;$$

$$\hat{Y}_t = -2.05 - 0.926 P_t + 1.78 E_t + 0.0608 B_t \quad (B)$$

(3.03) (0.347) (0.644) (0.310)

$$R^2 = 0.720, \text{RSS} = 0.192,$$

where R^2 is the coefficient of determination, RSS denotes residual sum of squares and estimated standard errors are given in parentheses.

(a) Test the hypothesis $H_0 : \beta_p = 0$ in both fitted models (A) and (B). Comment on your results.

(b) Test the individual hypotheses $H_0 : \beta_E = 0$ and $H_0 : \beta_B = 0$ and the joint hypothesis $H_0 : \beta_E = \beta_B = 0$.

(a) $H_0^p : \beta_p = 0$ vs $H_1^p : \beta_p < 0$ (Note the one-tail alternative hypothesis)

(A) $T = 22, k = 2, t_{\text{calc}} = -0.705/0.0673 = -10.48, t_{20}(0.05) = -1.725$. Reject.

(B) $T = 22, k = 2, t_{\text{calc}} = -0.926/0.346 = -2.67, t_{18}(0.05) = -1.734$. Reject.

(b) $H_0^e : \beta^e = 0$ vs $H_1^e : \beta^e > 0$, (once again note the one-tail alternative) $t_{\text{calc}} = 1.78/0.64 = 2.76$. Reject.

$H_0^t : \beta^t = 0$ vs $H_1^t : \beta^t > 0, t_{\text{calc}} = 0.0608/0.310 = 0.196$. Do not reject.

$H_0^{te} : \beta_e = \beta_t = 0$ vs $H_1^{et} : \beta^t \neq 0$ and/or $\beta^e \neq 0$.

$$F_{\text{calc}} = \frac{(0.636 - 0.192)/2}{0.192/18} = 20.81$$

$F_{2,18}(0.05) = 3.49$. Reject.

The F-Value : for joint coefficient testing :

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

where n is the sample size and k is the number of parameters in the regression.

If $F_{\text{calculated}} > F_{\text{table}}$, reject H_0 (as in the above example).

3. Validation: Checking the model against econometric criteria (*Assumptions*).

OLS Assumptions: $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \dots + e_i$

For the residuals e_i

A1 : The residuals e_i are normal random variables.

A2 : The residuals have *constant variance* (this is called homoscedasticity)

A3 : The expected value of the residuals is always zero

A4 : The residuals are *independent* from one another (not auto correlated)

For the independent variable

A5 : The X values are *precise*

A6 : The independent variables are not too strongly *collinear*

Further assumptions:

A7: The independent variables and the residuals are uncorrelated.

A8: The model is true : no mis-specification, no missing variables.

The Durbin-Watson test for autocorrelation errors:

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}; d \approx 2(1 - r); r : \text{Coefficient of correlation between } e_t \text{ \& } e_{t-1}$$

-The value of d is between 0 : perfect positive autocorrelation and 4: perfect negative autocorrelation. .

- d = 2 indicates no autocorrelation.

-dl and du are the lower and the upper values used to make inference.

-The following illustrates the conclusions associated with different values of d :

	Positive	Inconclusive	No autocorrelation	Inconclusive	Negative	
	reject	H_0	Accept	H_0	reject	
0	dl		du		4 - dl	4

Checking violations:

A2: Errors are homoskedastic; they have the same variance for all values of the independent variable(X_i).

A Violation: Errors are heteroskedastic; the variance changes as the independent variable changes.

Standard errors of the regression coefficients are **biased** (either too big or too small).

-If too small, then t-statistics will be too large.

-If too large, then t-statistics will be too small.

-This causes inaccurate significance tests.

A4: Errors Are Independent – No Autocorrelation

Standard errors of the regression coefficients are biased.

-SE too small (positive autocorrelation):t-statistics too large and coefficients too significant.

-SE too large (negative autocorrelation):t-statistics too small and coefficients not significant enough.

-To determine whether autocorrelation exists : Use Durbin-Watson Test:

d = 0 when r=1 (perfect positive correlation).

d = 4 when r=-1 (perfect negative correlation).

d = 2 when r=0 (no correlation).

A7: Errors Are Uncorrelated with Independent Variables

Results in biased regression coefficients.

Causes

- 1) Specification Bias (Omitted Variables): Variable left out of model, Error term picks up variation of left-out variable, Due to bad theory or data is not available.
- 2) Measurement Error: Data entry errors, Use of bad data
- 3) Simultaneous Equation Bias: Regression equation is part of a simultaneous equations system.

Multicollinearity Problems

Multicollinearity occurs when the X values, the predictors, are themselves highly correlated (intercorrelated) **(see definition above)**.

Solutions:

- Keep variables in equation but understand interpretation.
- Drop one or more variables, but understand interpretation.
- Combine variables.
- Collect more data.

4. Forecasting: Prediction & estimating errors.

5. Implementation.

Explain the importance of R^2 (coefficient of determination) and \bar{R}^2 (R^2 adjusted for degrees of freedom). What advantages does \bar{R}^2 have over R^2 ?

Explain what you understand by autocorrelation of the disturbance term in a regression model? What are the causes of autocorrelation?

1. (a) R^2 is the proportion of variance explained by the model, i.e. ESS/TSS where ESS is explained sum of squares and TSS is total sum of squares. \bar{R}^2 is defined as

$1 - \frac{n-1}{n-k}(1 - R^2)$ where n is the sample size and k is the number of parameters. R^2 is

important because it gives a measure of the overall 'fit' of the model but it suffers from the deficiency that as more variables are included in the regression the value will increase. \bar{R}^2 , on the other hand, has a penalty in the form of $(n-1)/(n-k)$ which will reduce the value as k increases. The problem is that \bar{R}^2 will increase if the absolute value of the t statistic of the included variable is greater than 1, i.e. it does not mean that the specification of the model has improved.

(b) If the error term u_t in the model $Y_t = \beta_0 + \beta_1 X_t + u_t$ is such that $E(u_t u_s) \neq 0$ then the error term is said to be serially correlated. This is a very general condition and it is generally necessary to assume the more restrictive condition that $u_t = \rho u_{t-1} + v_t$ where $|\rho| < 1$. The consequence is that ordinary least squares (OLS) parameter estimates are unbiased but inefficient and that their standard errors, and hence t -values, are incorrect.

The most common cause of serially correlated errors is that the model is mis-specified by omitting a variable. If this omitted variable is itself serially correlated then the error term will mirror this behaviour. Other causes of serial correlation are 'cobweb' type behaviour and mis-specified dynamic behaviour.

Explain what you understand by omitted variable bias?

Let a regression equation be:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad ; \quad t = 1, 2, \dots, T.$$

Outline briefly, how would you test

(i) $\beta_2 = 1$

(ii) jointly β_2 and β_3 are significantly different from zero.

(i) is a standard two-tail t test of the form $t = \frac{(\hat{\beta}_2 - 1)}{se_{\hat{\beta}_2}}$. The degrees of freedom for the

t test are T-3.

(ii) The test of the null hypothesis $H_0: \beta_2 = \beta_3 = 0$ against the alternative hypothesis $H_A: \beta_2 \neq 0$ and/or $\beta_3 \neq 0$ is achieved by applying the F test where

$F = \frac{R^2 / (k - 1)}{(1 - R^2) / (T - k)}$ where T is the sample size and k is the number of parameters in the regression, i.e. 3. The F test has 2, T-3 degrees of freedom.

Read the case study :

Lydia Pinkham's Vegetable Compound.

study guide pp 78 – 83.