Summary

-Econometrics-

Inhoud

Methods of Economic Research

Chapter 1: an overview of regression analysis

Econometrics	Econometrics means measurement (the meaning of the Greek word metrics) in economic. However, econometrics includes all those statistical and mathematical techniques that are utilized in the analysis of economic data. The main aim of using those tools is to prove or disprove particular economic propositions and models.

Econometrics has three major uses

1. Describing economic reality
2. Testing hypotheses about economic theory
3. Forecasting future economic activity

Hypothesis testing	Econometrics aims primarily at validating or refuting economic laws or theories. Hypothesis testing is the evaluation of alternative theories with quantitative evidence.
Estimated regression coefficient	The ability to estimate these coefficients makes econometrics valuable. Even though the sign is positive, we first must test it: the statistical significance of that estimate would have to be investigated before such conclusions could be justified.
Forecast	It is used to predict the future value(s) of the dependent variable, on the basis of known or expected future value(s) of the explanatory expected variable. The forecasting task of econometrics is crucial as it provide the mechanism for regulating and planning future economic policies

Economics is typically an observational discipline rather than an experimental one. We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists. Different approaches make sense within the field of economics. The kind of econometric tools used depends in part on the uses of that equation.

To get a better picture of these approaches, let's look at the steps used in non-experimental quantitative research:

1. Specifying the models or relationships to be studied
2. Collecting the data needed to quantify the models
3. Quantifying the models with the data

Focus on one econometric approach: single-equation linear regression analysis. It is important to remember that regression is only one of many approaches to econometric quantification.

Econometricians use regression analysis to make quantitative estimates of economic relationships that previously have been completely theoretical in nature. To predict the direction of the change, you need a knowledge of economic theory and the general characteristics of the product in question.

To predict the amount of the change, though, you need a sample of data, and you need a way to estimate the relationship.

Regression analysis	A statistical technique that attempts to *explain* movements in one variable, the **dependent variable**, as a function of movements in a set of other variables, called the independent (or explanatory) variables, through the quantification of a single equation. It is a reliable method of identifying which variables have an impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

Regression analysis is a natural tool for economists because most economic propositions can be stated in such single-equation functional forms. Simplest single-equation linear regression model is:

$Y = \beta_0 + \beta_1 x$

The model is a single-equation model because it's the only equation specified. The model is linear because if you were to plot the equation it would be a straight line rather than a curve. The ßs are the **coefficients** that determine the coordinates of the straight line at any point.

- β_0 is the **constant** or **intercept** term: it indicates the value of Y when X equals zero.
- $\beta_1 x$ is the **slope coefficient**: it indicates the amount that Y will change when X increases one unit (much of the attention in regression analysis is on slope coefficients).

$$\beta_1 x = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{\Delta Y}{\Delta X}$$

For a linear model, the slope is constant over the entire function. If linear regression techniques are going to be applied to an equation, that equation *must be* linear. An equation is **linear** if plotting the function in terms of X and Y generates a straight line. We can redefine most nonlinear equations to make them linear.

$Z = X^2$
$Y = \beta_0 + \beta_1 Z$

Stochastic error term	A term that is added to a regression equation to introduce all the variation in Y that cannot be explained by the included Xs.

The addition of a stochastic error term (ϵ) results in a typical regression equation:

$Y = \beta_0 + \beta_1 x + \epsilon$

The expression $\beta_0 + \beta_1 x$ is called the **deterministic** component of the regression equation because it indicates the value of Y that is determined by a given value of X, which is assumed to be non-stochastic. This deterministic component can be thought of as the **expected value** of Y given X. The error term is called the stochastic component.

The regression notation needs to be extended to allow the possibility of more than one independent variable and to include reference to the numbers of observations.

$Y = \beta_0 + \beta_1 x_i + \epsilon_i$ (i = 1, 2, …, N)

These **multivariate** (more than one independent variable) **regression coefficients** serve to isolate the impact on Y of a change in one variable from the impact of Y on changes in the other variables. In the real world it is very difficult to run controlled economic experiments, because many economic factors change simultaneously, often in the opposite direction. Thus, the ability of regression analysis to measure the impact of one variable on the dependent variable, *holding constant the influence of the other variables in the equation,* is a tremendous advantage. Note that if a variable is not included in the equation, then its impact is *not* held constant in the estimation of the regression coefficients.

Dummy variable	Variable that can only take two values, 0 or 1. It is extremely useful when we want to quantify a concept that is inherently qualitative (gender).
Time series	Sample consists of a series of years and numbers.

Estimated regression equation	Quantified version of the theoretical regression equation. It is obtained from a sample of data for actual Xs and Ys.
Estimated regression coefficients ($\hat{\beta}_0$ or $\hat{\beta}_1$)	Empirical best guesses of the true regression coefficients and are obtained from data from samples of the Ys and Xs.
Residual (e_i)	The difference between the estimated value of the dependent variable ($\hat{Y}_i$) and the actual value of the dependent variable (Y_1)

Difference between the theoretical regression analysis and the estimated regression analysis

1. The theoretical regression coefficients are replaced with *estimates* of those coefficients
   - The theoretical equation is purely abstract in nature ($Y = \beta_0 + \beta_1 x_i + \epsilon_i$) and the estimated regression equation has actual numbers in it ($\hat{Y} = 103.40 + 6.38 x_i$).

The *residual* is the difference between the observed Y and the estimated regression line ($\hat{Y}$), while *error term* is the difference between the observed Y and the true regression equation (the expected value of Y). Note that the error term is a theoretical concept that can never be observed, but the residual is a real-world value that is calculated for each observation every time a regression is run.

True regression equation	**Estimated regression equation**
β_0	$\hat{\beta}_0$
$\beta_1 x_i$	$\hat{\beta}_1$
ϵ_i	e_i

Cross-sectional data set	All the observations are from the same point in time and represent different individual economic entities from that same point in time (looking at local houses that are sold in the last few weeks – building a regression model of the sales prices of the houses as a function of their size)

Chapter 2: ordinary least squares (OLS)

2.1 Introduction

The ordinary Least Square (OLS) method is used extensively in regression analysis primarily because it is intuitively appealing and mathematically much simpler than the method of maximum likelihood.

The purpose of regression analysis is to take a purely theoretical equation like

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

And use a set of data to create an estimated equation like

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Where each *hat* indicates a sample estimate of the true population value. The purpose of the estimation technique is to obtain numerical values for the coefficients of an otherwise completely theoretical regression equation. The most widely used method of obtaining these estimates is Ordinary Least Squares (OLS), which has become so standard that its estimates are presented as a point of reference even when results from other estimation techniques are used.

OLS minimizes $\sum_{i=1}^{N} e_i^2$ (i = 1, 2, …, N)

Since these residuals (e_is) are the differences between the actual Ys and the estimated Ys produced by the regression (the $\hat{Y}$ in the equation), is the equation above equivalent to saying that

OLS minimizes $\sum(Y_i - \hat{Y}_i)^2$

There are at least three important reasons for using OLS to estimate regression models:

- OLS is relatively easy to use
- The goal of minimizing $\sum_{i=1}^{N} e_i^2$ is quite appropriate from a theoretical point of view
- OLS estimates have several useful characteristics
  - The sum of the residuals is exactly zero
  - OLS can be shown to be the *best* estimator possible under a set of specific assumptions.

Conclusion:

Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data (visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line - the smaller the differences, the better the model fits the data).

Estimator	The OLS estimator is an estimator that minimizes the sum of squared residuals. The applicability of the OLS **estimator** is based on the classical assumptions of the linear regression model.

OLS is an estimator and $\hat{\beta}$ produced by OLS is an estimate (read page 39 – example on page 40)

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{N}[(X_i-\bar{X})(Y_i-\bar{Y})]}{\sum_{i=1}^{N}(X_i-\bar{X})^2}$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

Where;

- $\bar{X}$ = mean of X (or $\sum X_i /N$)
- $\bar{Y}$ = mean of Y (or $\sum Y_i /N$)

The general multivariate regression model with K independent variables:

$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + \epsilon_i$ (i = 1, 2, …, n)

The biggest difference between a single-independent-variable regression model and a multivariate regression model is in the interpretation of the latters slope coefficients. These coefficients, often called *partial* regression coefficients, are defined to allow the researcher to distinguish the impact of one variable from that of other independent variables. Specifically, a **multivariate regression coefficient** indicates the change in the dependent variable associated with a one-unit increase in the independent variable in question *holding constant the other independent variables in the equation* (but **not** holding constant any relevant variables that might have been omitted from the equation). The coefficient β_0 is the value of Y when all the Xs and the error term equal zero.

The goal of OLS is to choose those $\hat{\beta}$s that minimize the summed square residuals. The application of OLS to an equation with more than one independent variable is quite like its application to a single-independent-variable model.

- $\hat{\beta}_1 = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$
- $\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$

Where;

- $\bar{X}$ = mean of X (or $\sum X_i /N$)
- $\bar{Y}$ = mean of Y (or $\sum Y_i /N$)

econometricians use the squared variations of Y around its mean as a measure of the amount of variation to be explained by the regression. This computed quantity is usually called the **total sum of squares (TSS)** and is written as

$TSS = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$

For OLS, the TSS has two components, variation that can be explained by the regression and variation that cannot:

$\sum_i (Y_i - \bar{Y})^2$	$= \sum_i (\hat{Y}_i - \bar{Y})^2$	+	$\sum_i e_i^2$
Total sum of squares (TSS)	= Explained Sum of Squares (ESS)	+	Residual Sum of Squares (RSS)

This is usually called the **decomposition of variance (read page 49).**

Explained Sum of Squares	Measures the amount of the squared deviation of Y_i from its mean that is explained by the regression line. ESS is attributable to the fitted regression line
Residual Sum of Squares	The unexplained portion of TSS is called the RSS. The smaller the RSS is relative to the TSS, the better the estimated regression line fits the data

2.4 Describing the overall fit of the estimated model

OLS is the estimating technique that minimizes the RSS and therefore maximizes the ESS for a given TSS. We expect that a good, estimated regression equation will explain the variation of the dependent variable in the sample accurately. If it does, we say that the estimated model fits the data well.

R^2

The simplest commonly used measure of fit is R^2 or the coefficient of determination. R^2 is the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$$

The higher the R^2 is, the closer the estimated regression equation fits the sample data. Measures of this type are called **goodness of fit** measures.
When the R^2 is 0.95, this means that the relationship between X and Y can be *explained* quite well by a linear regression equation. This kind of result is typical of a time-series regression with a good fit. Most of the variation has been explained (ESS), but there remains a portion of the variation that is essentially random or unexplained by the model (RSS).

In time-series data, we often get a very high R^2 because there can be significant time trends on both sides of the equation. In cross-sectional data, we often get low R^2s because the observations (say, countries) differ in ways that are not easily quantified. So, there is no simple method of determining how high R^2 must be for the fit to be considered satisfactory (matter of experience). It should be noted that a high R^2 does not imply that changes in X lead to changes in Y, as there may be an underlying variable whose changes lead to changes in both X and Y.

The simple correlation coefficient, r

Simple correlation coefficient (r)	It is a measure of the strength and direction of the linear relationship between two variables
r = + 1	If two variables are perfectly positively correlated
r = – 1	If two variables are perfectly negatively correlated
r = 0	If two variables are totally uncorrelated

The closer the absolute value of r is to 1, the stronger the correlation between the two variables. We will use the simple correlation coefficient to describe the correlation between two variables. Interestingly, it turns out that r and R^2 are related if the estimated equation has exactly one independent variable. The square of r equals R^2 for a regression where one of the two variables is the dependent variable, and the other is the only independent variable

$\bar{R}^2$, *adjusted* R^2

a major problem with R^2 is that adding another independent variable to a particular equation can never decrease R^2. The equation with the greater number of independent variables will always have a better (or equal) fit as measured by R^2. To see this, recall:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Adding one variable cannot change TSS, but in most cases, the added variable will reduce RSS, so R^2 will rise. RSS will never increase because the OLS program could always set the coefficient of the added variable equal to zero, thus giving the same fit as the previous equation. The coefficient of the newly added variable being zero is the only circumstance in which R^2 will stay the same when a variable is added. Otherwise, R^2 will always increase when a variable is added to an equation. The lower the degrees of freedom, the less reliable the estimates are likely to be.

R^2 is little help is we are trying to decide whether adding a variable to an equation improves our ability to meaningfully explain the dependent variable. Because of this problem, econometricians have developed another measure of the quality of the fit of an equation. That measure is $\bar{R}^2$, which is R^2 **adjusted for the degrees of freedom**:

$$\bar{R}^2 = 1 - \frac{\frac{\sum e_i^2}{(N-K-1)}}{\frac{\sum (Y_i - \bar{Y})^2}{(N-1)}} = 1 - \frac{\frac{RSS}{(N-K-1)}}{\frac{TSS}{(N-1)}}$$

Alternative: the **adjusted R^2** can also be written as:

$$\bar{R}^2 = R^2 - \frac{K}{(N-K-1)} \; x \; (1 - R^2)$$

$\bar{R}^2$ measures the percentage of the variation of Y around its mean that is explained by the regression equation, *adjusted for degrees of freedom*. $\bar{R}^2$ will increase, decrease or stay the same when a variable is added to an equation, depending on whether the improvement in fit caused by the addition of the new variable outweighs the loss of the degrees of freedom. $\bar{R}^2$ can be used to compare the fits of the equation with the same dependent variable and a different number of independent variables. Because of this property, most researchers automatically use $\bar{R}^2$ instead of R^2 when evaluating the fit of their estimated regression equations.

Important: always remember that the quality of the fit of an estimated equation is only one measure of the overall quality of that regression. The degree to which the estimated coefficients conform to economic theory and the researcher's previous expectations about those coefficients are just as important as the fit itself.

Chapter 3: learning to use regression analysis

3.1 steps in applied regression analysis

the relative emphasis and effort expended on each step will vary, but normally all the steps are necessary for successful research.

Steps

1. review the literature and develop the theoretical model.
2. specify the model: select the independent variables and the functional form.
3. hypothesize the expected signs of the coefficients.
4. collect the data. Inspect and clean the data.
5. Estimate and evaluate the equation.
6. Document the results.

Step 1 review the literature and develop the theoretical model

The first step in any applied research is to get a good theoretical grasp of the topic to be studied. You should start your investigation where earlier researchers left off. The most convenient approaches to reviewing the literature are to obtain several recent issues of the *Journal of Economic Literature* or a business-oriented publication of abstracts, or to run an Internet search or an *EconLit* search on your topic. When are topic is so new or obscure that you won't be able to find any articles on it, there are two recommended possible strategies:

1. Try to transfer theory from a similar topic to yours
2. If all else fails, pick up the telephone and call someone who works in the field you're investigating.

Step 2 specify the model: select the independent variables and the functional form

After selecting the dependent variables, the **specification** of a model involves choosing the following components:

1. The independent variables and how they should be measured
2. The functional (mathematical) form of the variables
3. The properties of the stochastic error term

A regression equation is specified when each of those elements has been treated appropriately. A mistake in any of the three elements results in a **specification error** (most disastrous to the validity of the estimated equation). Thus, the more attention paid to economic theory at the beginning of a project, the more satisfying the regression results are likely to be.

An explanatory variable is chosen because it is a theoretical determinant of the dependent variable. It is expected to explain at least part of the variation in the dependent variable. Our goal should be to specify only relevant explanatory variables, those expected theoretically to assert a substantive influence on the dependent variable. Variables suspected of having little effect should be excluded unless their possible impact on the dependent variable is of some particular (i.e. policy) interest.

When researchers decide that prices of only two other goods need to be included, they are said to impose their **priors** (i.e. previous theoretical beliefs) or their working hypotheses on the regression equation. The danger is that a prior may be wrong and could diminish the usefulness of the estimated regression equation. Each of the priors should be explained and justified in detail.

Dummy variable (gender)	Takes on the values of one or zero depending on whether a specified condition holds.

Step 3 Hypothesize the expected signs of the coefficients

Once the variables are selected, it's important to hypothesize the expected signs of the regression coefficients. The signs above the variables indicate the hypothesized sign of the respective regression coefficients in a linear model (page 75). In many cases, the basic theory is general knowledge, so the reason for each sign need not be discussed. However, in any doubt surrounds the selection of an expected sign, you should document the opposing forces at work and the reason for hypothesizing a positive or negative coefficient.

Step 4 Collect the data. Inspect and clean the data.

A general rule regarding sample size is *the more observations the better*, as long as the observations are from the same general population. In regression analysis, all the variables must have the same number of observations. They also should have the same frequency and time period. Often, the frequency is determined by the availability of data.

The reason there should be as many observations as possible concerns the statistical concept of *degrees of freedom*. Estimation of a line (page 76) takes place only when a straight line is fitted to three or more points that were generated by some process that is not exact. The excess of the number of observations (three) over the number of coefficients to be estimated (in this case two, the intercept and slope) is the **degree of freedom**. All that is necessary for estimation is a single degree of freedom, but the more degrees of freedom there are, the better. This is because when the number of degrees of freedom is large, every positive error is likely to be balances by a negative error. When degree of freedom are low, the random element is likely to fail to provide such offsetting observations.

Degree of freedom	The excess of the number of observations over the number of coefficients to be estimated
Units of measurement of the variable	All conclusions about signs, significance, and economic theory are independent of units of measurement.

The final step before estimating your equation is to inspect and clean the data. You should make it a point always to look over your data set to see if you can find any errors. To inspect the data, obtain a printout and a plot (graph) of the data and look for outliers. In addition, it is a good habit to look at the mean, maximum, and minimum of each variable and then think about possible inconsistencies in the data.

Outlier	An observation that lies outside the range of the rest of the observations and looking for outliers is an easy way to find data entry errors.

Typically, the data can be cleaned of these errors by replacing an incorrect number with the correct one. In extremely rare circumstances, an observation can be dropped from the sample, but only if the correct number cannot be found or if that particular observation clearly is not from the same population as the rest of the sample.

- **But:** a regression needs to be able to explain all the observations in a sample, not just the well-behaved ones.

Step 5 Estimate and evaluate the equation

Typically, estimation is done using OLS, but if another estimation technique is used the reason for that alternative technique should be carefully explained and evaluated. Once this evaluation is complete, do not automatically go to step 6. Regression results are rarely what one expects, and additional model development often is required. If you are missing an important variable, you need to go back to step 1. You'd then go through each of the steps in order until you had estimated your new specification in step 5. Finally, it is often worthwhile to estimate additional specifications of an equation in order to see how stable your observed results are (**sensitivity analysis**).

Step 6 Document the results

A standard format usually is used to present estimated regression results (**page 79**). For time series data sets, the documentation also includes the frequency (quarterly, annually) and the time period of the data. Most computer programs present statistics to eight or more digits, but it is important to recognize the difference between the number of digits computed and the number of *meaningful digits*, which may be as low as two or three.

Example of the whole regression analysis process on page 80 – 88 (important to read).

Chapter 4: The classical model

The term *classical* refers to a set of basic assumptions required to hold in order for OLS to be considered the *best* estimator available for regression models. When one or more of these assumptions do not hold, other estimation techniques (such as Generalized Least Squares) sometimes may be better than OLS. As a result, one of the most important jobs in regression analysis is to decide whether the classical assumptions hold for a particular equation.

4.1 The classical assumptions

The **classical assumptions** must be met for OLS estimators to be the best available. Because of their importance in regression analysis, the assumptions are presented here in tabular form as well as in words.

The classical assumptions

1. The regression model is linear, is correctly specified, and has an additive error term
2. The error term has a zero population mean
3. All explanatory variables are uncorrelated with the error term
4. Observations of the error term are uncorrelated with each other (no serial correlation)
5. The error term has a constant variance (no heteroskedasticity)
6. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity)
7. The error term is normally distributed (this assumption is optional but usually is invoked)

Classical error term	Error term satisfying assumptions I through V
Classical normal error term	Error term satisfying assumptions I through V and VII is added

<u>1) the regression model is linear, is correctly specified, and has an additive error term</u>

The regression model is assumed to be linear:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_{ki} + \epsilon_i$

The assumption that the regression model is linear does not require the underlying theory to be linear (could be exponential)

$Y_i = e^{\beta_0} X_1^{\beta_1} e^{\epsilon_1}$

$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \epsilon_1$

if the variables are relabelled as $Y_i^* = \ln(Y_i)$ and $X_i^* = \ln(X_i)$, then the form of the equation becomes linear

$Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_1$

<u>2) The error term has a zero population mean</u>

Econometricians add a stochastic (random) error term to regression equations to account for variation in the dependent variable that is not explained by the model. The specific value of the error term for each observation is determined purely by chance. When the entire population of possible values for the stochastic error term is considered, the average value of that population is zero. For a small sample, it is not likely that the mean is exactly zero, but as the size of the sample approached infinity, the mean of the sample approached zero.

To compensate for the chance that the mean of ϵ might not equal zero, the mean of ϵ_1 for any regression is forced to be zero by the existence of the constant term in the equation. In essence, the

constant term equals the fixed portion of Y that cannot be explained by the independent variables, whereas the error term equals the stochastic portion of the unexplained value of Y. although it's true that the error term can never be observed, it's instructive to pretend that we can do so to see how the existence of a constant term forces the mean of the error term to be zero in a sample.

3) all explanatory variables are uncorrelated with the error term

if an explanatory variable and the error term were instead correlated with each other, the OLS estimates would be likely to attribute to the X some of the variation in Y that came from the error term. As a result, it's important to ensure that the explanatory variables are uncorrelated with the error term. One of the major components of the stochastic error term is omitted variables, so if a variable has been omitted, then the error term will change when the omitted variables change. If this omitted variable is correlated with an included independent variable, then the error term is correlated with that independent variable as well (violation assumption III).

4) observations of the error term are uncorrelated with each other

The observations of the error term are drawn independently from each other. If a systematic correlation exists between one observation of the error term and another, then it will be more difficult for OLS to get accurate estimates of the standard errors of the coefficients. This assumption is most important in time series models.

Serial correlated	If, over all the observations of the sample, ϵ_{t+1} is correlated with ϵ_t. This is also called *autocorrelated* and this violates assumption IV.

5) the error term has a constant variance (no heteroskedasticity)

the observations of the error term are assumed to be drawn continually from identical distributions. The alternative would be for the variance of the distribution of the error term to change for each observation or range of observations (figure 2 violates assumption V). The lack of a constant variance for the distribution of the error term causes OLS to generate inaccurate estimates of the standard error of the coefficients (cross-sectional data).

Heteroskedasticity	The violation of assumption V

6) no explanatory variable is a perfect linear function of any other explanatory variable(s)

Perfect collinearity	Between two independent variables implies that they are really the same variable, or that one is a multiple of the other, and/or that a constant has been added to one of the variables.

Perfect collinearity is that the relative movements of one explanatory variable will be matched exactly by the relative movements of the other even though the absolute size of the movements might differ. Because every movement is matched exactly by a relative movement in the other, the OLS estimation procedure will be incapable of distinguishing one variable from the other.

Many instances of perfect collinearity (or **multicollinearity** if more than two independent variables are involved) are the result of the researcher not accounting for identities (definitional equivalences) among the independent variables. This problem can be corrected easily by dropping one of the perfectly collinear variables from the equation. Perfect multicollinearity also can occur when two independent variables always sum to a third or when one of the explanatory variables does not change within the sample. With perfect multicollinearity, the OLS computer program (or any other estimation technique) will be unable to estimate the coefficients of the collinear variables (unless there is a rounding error).

7) the error term is normally distributed

although we have already assumed that observations of the error term are drawn independently (assumption IV) from a distribution that has a zero mean (assumption II) and that has a constant variance (assumption V), we have said little about the shape of that distribution. Assumption VII states that the observations of the error term are drawn from a distribution that is normal.

This assumption of normality is not required for OLS estimation. Its major application is in **hypothesis testing**, which used the estimated regression coefficient to investigate hypothesis about economic behaviour. Even though this assumption is optional, it's usually advisable to add the assumption of normality to the other six assumptions for two reasons:

1. The error term ϵ_i can be thought of as the sum of several minor influences or errors. As the number of these minor influences gets larger, the distribution of the error term tends to approach the normal distribution.
2. The *t*-statistics and the *F*-statistics are not truly applicable unless the error term is normally distributed (or the sample is quite large).

Standard normal distribution

- Mean is 0
- Variance is 1

4.2 The sampling distribution of $\hat{\beta}$

Each different sample of data typically produces a different estimate of β.

Sampling distribution of $\hat{\beta}$	The probability distribution of these $\hat{\beta}$ values across different samples
Estimator	A formula, such as the OLS formula
Estimate	The value of $\hat{\beta}$ computed by the formula for a given sample

The collection of all the possible samples has a distribution, with a mean and a variance, and we need to discuss the properties of this sampling of distribution of $\hat{\beta}$, even though in the most real applications we will encoder only a single draw from it. Sampling distribution refers to the distribution of different values of $\hat{\beta}$ across different samples, not within one. These $\hat{\beta}$s usually are assumed to be normally distributed because the normality of the error term implies that the OLS estimates of beta are normally distributed as well.

Unbiasedness	For a *good* estimation technique, we'd want the mean of the sampling distribution $\hat{\beta}$s to be equal to our true population β.

Properties of the mean

A desirable property of a distribution of estimates is that its mean equals the true mean of the variable being estimated. An estimator that yields such estimates is called an unbiased estimator.

Unbiased estimator	An estimator $\hat{\beta}$ is an unbiased estimator if its sampling distribution has as its expected value the true value of β → $E(\hat{\beta}) = \beta$
Biased estimator	If an estimator produces $\hat{\beta}$s that are not centered around the true β

Only one value of $\hat{\beta}$ is obtained in practice, but the property of unbiasedness is useful because a single estimate drawn from an unbiased distribution is more likely to be near the true value (assuming identical variances) than one taken from a distribution not centred around the true value. Without any information about the distribution of the estimates, we would always rather have an unbiased estimate rather than a biased one.

Properties of the variance

Just as we would like the distribution of the $\hat{\beta}$s to be centred around the true population of β, so too would we like that distribution to be as narrow (or precise) as possible. A distribution centred around the truth but with an extremely large variance might be of very little use because any given estimate would quite likely be far from the true β value. For a $\hat{\beta}$ with small variance, the estimates are likely to be close to the mean of the sampling distribution.

The variance of the distribution of the $\hat{\beta}$s can be decreased by increasing the size of the sample. This also increases the degrees of freedom, since the number of degrees of freedom equals the sample size minus the number of coefficients or parameters estimated. One method of deciding whether this decreased variance in the distribution of the $\hat{\beta}$s is valuable enough to offset the bias is to compare different estimation techniques by using a measure called the **mean square error (MSE)**.

Mean Square Error	Equal to the variance plus the square of the bias. The lower MSE, the better.

A final item of importance is that as the variance of the error term increases, so too does the variance of the distribution of $\hat{\beta}$.

- **Reason:** with the larger variance of ϵ_i, the more extreme values of ϵ_i are observed with more frequency, and the error term becomes more important in determining the values of Y_i

The standard error of $\hat{\beta}$

Since the standard error of the estimated coefficient, SE ($\hat{\beta}$), is the square root of the estimated variance of the $\hat{\beta}$s, it is similarly affected by the size of the sample and the other factors mentioned. The larger the sample, the more precise our coefficient estimates will be (smaller standard error).

4.3 The Gauss-Markov Theorem and the properties of OLS estimators

The Gauss-Markov theorem proves two important properties of OLS estimators. The **Gauss-Markov theorem** states that:

Given Classical assumptions I through VI (assumption VII, normality, is not needed for this theorem), the Ordinary Least Squares estimator of β_k is the minimum variance estimator from among the set of all linear unbiased estimators of β_k, for k = 0, 1, 2, …, K.

The Gauss-Markov theorem (requires that only six out of seven assumptions are met) is most easily remembered by stating that **OLS is BLUE**

- *Best (meaning minimum variance) Linear Unbiased Estimator*

If an equation's coefficient estimation is unbiased (that is, if each of the estimated coefficients is produced by an unbiased estimator of the true population coefficient), then:

$E(\hat{\beta}_k) = \beta_k$ (k = 0, 1, 2, …, K)

Best means that each $\hat{\beta}_k$ has the smallest variance possible. An unbiased estimator with the smallest variance is called **efficient**, and that estimator is said to have the property of efficiency. If all seven assumptions are met in the Gauss-Markov theorem, the OLS is *BUE* (page 111).

Given all seven classical assumptions, the OLS coefficient estimators can be shown to have the following properties:

1. They are unbiased
   - $E(\hat{\beta}) = \beta$
2. They are minimum variance
3. They are consistent
   - As the sample size gets larger, the variance gets smaller, and each estimate approaches the true value of the coefficient being estimated.
4. They are normally distributed
   - The $\hat{\beta}$s are N (β, VAR[$\hat{\beta}$]).

4.4 Standard Econometric Notation

important (notation) see table page 112

Chapter 5: hypothesis testing

Hypothesis testing determines what we can learn about the real world from a sample.

t-test	Statistical tool typically used for hypothesis rest of individual regression coefficients

Our approach will be classical in nature, since we assume that the sample data are our best and only information about the population. An alternative, **Bayesian statistics**, used a completely different definition of probability and does not use the sampling distribution concept.

5.1 what is hypothesis testing?

Classical null and alternative hypotheses

The first step in hypothesis testing is to state the hypotheses to be tested. This should be done *before* the equation is estimated because hypotheses developed after estimation run the risk of being justifications of results rather than tests of the validity of those.

Null hypothesis	A statement of the values that the researcher does not expect (H_0)
Alternative hypothesis	A statement of the values that the researcher expects (H_A)

One-sided test
H_0: $\beta \geq 0$
H_A: $\beta < 0$

The above hypotheses are for a **one-sided test** because the alternative hypotheses have values on only one side of the null hypotheses. Another approach is to use a **two-sided test** (or a **two-tailed test**) in which the alternative hypothesis has values on both sides of the null hypothesis:

Two-sided test
H_0: $\beta = 0$
H_A: $\beta \neq 0$

Classical hypothesis testing requires that the null hypothesis contains the equal sign in some form (whether it be =, $\leq$ or $\geq$). This requirement means that researchers are forced to put the value they expect in the null hypothesis if their expectations include an equal sign. Economists always put what they expect in the alternative hypothesis. This allows us to make rather strong statements when we reject a null hypothesis. However, we can never say that we *accept* the null hypothesis. We must always say that we *cannot reject* the null hypothesis.

Type I and type II errors

The typical testing technique in econometrics is to hypothesize an expected sign (or value) for each regression coefficient (except the constant term) and then to determine whether they reject the null hypothesis. Since the regression coefficients are only estimates of the true population parameters, it would be unrealistic to think that conclusions drawn from regression analysis will always be right. There are two kinds of errors we can make in such hypothesis testing (**see figures on page 131**):

Type I error	We reject a true null hypothesis (we have rejected the truth): Type I error consists of rejecting the null hypothesis when it is true. This is a very serious error that we want to seldomly make. We don't want to be very likely to conclude the experiment had an effect when it didn't.
Type II error	We do not reject a false null hypothesis (we have failed to reject a false null hypothesis):

	Type II error consists of failing to reject the null hypothesis when it is false. This error has less grievous implications, so we are willing to err in this direction.

Decision rules of hypothesis testing

Decision rule	A method of deciding whether to reject a null hypothesis
Critical value	A value that divides the *acceptance* region from the *rejection* region when testing a null hypothesis (graphed in figures on page 133)

Typically, a decision rule involves comparing a sample statistic with a preselected *critical value* found in tables. A decision rule should be formulated before regression estimates are obtained. The range of possible values of $\hat{\beta}$ is divided into two regions where the terms are expressed relative to the null hypothesis.

1. Acceptance region
2. Rejection region

To define these regions, we must determine a *critical value* (or two for a two-tailed test) of $\hat{\beta}$. If the observed $\hat{\beta}$ is greater than the critical value, we can reject the null hypothesis that β is zero or negative. This can be seen in figure 3: any $\hat{\beta}$ above 1.8 can be seen to fall into the rejection region, whereas any $\hat{\beta}$ below 1.8 can be seen to fall into the acceptance region.

Decreasing the chance of a Type I error means increasing the chance of a Type II error (not rejecting a false null hypothesis). This is because if you make the rejection region so small that you almost never reject a true null hypothesis, then you're going to be unable to reject almost every null hypothesis, whether they are true or not. As a result, the probability of a type II error will rise.

5.2 the *t*-test

the *t*-test is the test that econometricians usually use to test hypotheses about individual regression slope coefficients. Tests of more than one coefficient at a time (joint hypotheses) are typically done with the *F*-test. The t-test is easy to use because it accounts for differences in the units of measurement of the variables and in the standard deviation of the estimated coefficients. More important, the *t*-statistics is the appropriate test to use when the stochastic error term is normally distributed and when the variance of that distribution must be estimated.

The *t*-statistics

For a statistical multiple regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \in_i$$

We can calculate t-values for each of the estimated coefficients in the equation. The t-tests are usually done only on the slope on the coefficients. For these, the relevant form of the **t-statistics** for the *k*th coefficient is

$$t_k = \frac{(\hat{\beta}_k - \hat{\beta}_{H_0})}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

where;

- $\hat{\beta}_k$ = the estimated regression coefficient of the *k*th variable
- $\hat{\beta}_{H_0}$ = the border value (usually zero) implied by the null hypothesis for $\hat{\beta}_k$
- $SE(\hat{\beta}_k)$= the estimated standard error of $\hat{\beta}_k$ (square root of the estimated variance of $\hat{\beta}_k$). There is no *hat* attached to SE, because SE is already an estimate.

Since most regression hypotheses test whether a particular regression coefficient is significantly different from zero, $\hat{\beta}_{H_0}$ is typically zero, and the most-used form of the *t*-statistics becomes

$$t_k = \frac{(\hat{\beta}_k - 0)}{SE(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \qquad (k = 1, 2, \ldots, K)$$

this is the estimated coefficient divided by the estimate of its standard error. Note that the sign of the t-value is always the same as that of the estimated coefficient of P, the population variable. The larger the absolute value of the *t*-value is, the greater the likelihood that the estimated regression coefficient is significantly different from zero.

The critical t-value and the t-test decision rule

Critical t-value	The value that distinguished the acceptance region from the rejection region
Degrees of freedom	The number of observations minus the number of coefficients estimated (N – K – 1)

The critical t-value (t_c) is selected from a *t*-table depending on whether the test is one-sides or two-sided, on the level of Type I error you specify and on the degrees of freedom. The level of type I error is also called the *level of significance* of that test. A critical t-value t_c is thus a function of the probability of Type I error that the researcher wants to specify.

Once you have obtained a calculated t-value t_k and a critical t-value t_c, you can reject the null hypothesis if the calculated t-value is greater in absolute value than the critical t-value and if the calculated t-value has the sign implied by H_A. Thus, the rule to apply when testing a single regression coefficient is that you should:

Reject H_0 if $|t_k| > t_c$ and if t_k also has the sign implied by H_A. do not reject H_0 otherwise.
The decision rule is the same: reject the null hypothesis if the appropriately calculated t-value t_k is greater in absolute value than the critical t-value t_c, if the sign of t_k is the same as the sign of the coefficient implied in H_A. Otherwise, do not reject H_0. Note from statistical table B-1 that the critical t-value for a one-tailed test at a given level of significance is exactly equal to the critical t-value for a two-tailed test at twice the level of significance as the one-tailed test.

Choosing a level of significance

The words *statistically positive* usually carry the statistical interpretation that H_0 was rejected in favour of H_A according to the pre-established decision rule, which was set up with a given level of significance.

Level of significance	Indicates the probability of observing an estimated t-value greater than the critical t-value if the null hypothesis were correct.

An extremely low level of significance dramatically increases the probability of making a Type II error. Therefore, unless you are in the unusual situation of not caring out mistakenly *accepting* a false null hypothesis, minimizing the level of significance is **not** good standard practice. Instead, we recommend using a 5-percent level of significance except in those circumstances when you know something unusual about the relative costs of making Type I and Type II errors.

Some researchers avoid choosing a level of significance by simply stating the lowest level of significance possible for each estimated regression coefficient. The use of the resulting significance levels, called *p*-values, is an alternative approach to the t-test.

Other researchers produce tables of regression results, typically without hypothesized signs for their coefficients, and then mark *significant* coefficients with asterisks.

Asterisks	Indicate when the lowest t-score is larger in absolute value than the two-sided 10-percent critical value (*), the two-sided 5-percent value (**), or the one-sided 1-percent critical value (***)

Now and then researchers will use the phrase *degree of confident* or *level of confidence (has similar meaning as level of significance)* when they rest hypotheses.

Level of confidence	100 percent minus the level of significance *(t-test 5% significance → 95% confidence)*

Confidence intervals

Confidence interval	A range that contains the true value of an item a specified percentage of the time; Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter; however, the interval computed from a particular sample does not necessarily include the true value of the parameter.

This percentage is the level of confidence associated with the level of significance used to choose the critical t-value in the interval. For an estimated regression coefficient, the confidence interval can be calculated using the two-sided critical t-value and the standard error of the estimated coefficient

Confidence interval = $\hat{\beta} \pm t_c \, x \, SE(\hat{\beta})$ (example on page 141)

p-values
there is an alternative approach to the t-test. A p-value (or marginal significance level) for a t-score is the probability of observing a t-score that size or larger (in absolute value) if the null hypothesis were true. Graphically, it is the area under the curve of t-distribution between the actual t-score and infinity. A *p*-value is a probability, so it runs from 0 to 1.

p-value	Tells us the lowest level of significance at which we could reject the null hypothesis (assuming that the estimate is in the expected direction)

A small p-value casts doubt on the null hypothesis, so to reject a null hypothesis, we need a low *p*-value. You can read *p*-values off your regression output just as you would your $\hat{\beta}$s (mostly presented for two-sided alternative hypotheses). If your test is one-sided, you need to divide the p-value in your regression output by 2 before doing any tests. The p-value decision rule (page 142):

Reject H_0 if p-value$_K$ < the level of significance and if $\hat{\beta}_K$ has the sign implied by H_A.

p-values have several advantages
- easy to use
- allow readers of research to choose their own level of significance instead of being forced to use the level chosen by the original researcher.
- Convey information to the reader about the relative strength with which we can reject a null hypothesis

If you know how to use the standard t-test approach, it is easy to switch to the p-value approach, but the reverse is not necessarily true. Therefore, beginning researchers should use t-tests.

5.3 examples of t-tests

Examples of one-sided t-tests

The most common use of one-sided t-tests is to determine whether a regression coefficient is significantly different from zero in the direction predicted by theory. The four steps to use when working with the t-test are

1. Set up the null and alternative hypotheses
   - Remember that a t-test typically is not run on the estimate of the constant term β_0
2. Choose a level of significance and therefore a critical t-value
   - Note that the level of significance does not have to be the same for all the coefficients in the same regression equation
3. Run the regression and obtain an estimated t-value (or t-score)
   - Note that since standard errors are always positive, a negative estimated coefficient implies a negative t-value.
4. Apply the decision rule by comparing the calculated t-value with the critical t-value in order to reject or not reject the null hypothesis.

Examples of two-sided t-tests

Although most hypotheses in regression analysis should be tested with one-sided t-tests, two-sided t-tests are appropriate situations. The kinds of circumstances that call for a two-sided test fall into two categories

1. Two-sided tests of whether an estimated coefficient is significantly different from zero.
   - A two-sided test implies two different rejection regions surrounding the acceptance region. There is an advantage to testing hypotheses with a one-sided test if the underlying theory allows because, for the same t-values, the possibility of type I error is half as much for a one-sided test as for a two-sided test.
2. Two-sided tests of whether an estimated coefficient is significantly different from a specific nonzero value.
   - Since the hypothesized β value is no longer zero, the formula with which to calculate the estimated t-value is now $t_k = \frac{(\widehat{\beta}_k - \widehat{\beta}_{H_0})}{SE(\widehat{\beta}_k)}$.

5.4 Limitations of the t-test

Problems with the t-test

- It is easy to misuse
- The usefulness of the t-test diminishes rapidly as more and more specifications are estimated and tested.

The t-test does not test theoretical validity

Recall that the purpose of the t-test is to help the researcher make inferences about a particular population coefficient based on an estimate obtained from a sample of a particular population. Some beginning researchers conclude that any *statistically* significant result is also a *theoretically* correct one. This is dangerous because such a conclusion confuses statistical significance with theoretical validity.

The t-test does not test *importance*

One possible use of a regression equation is to help determine which independent variable has the largest relative effect (importance) on the dependent variable. Some beginning researchers draw the unwarranted conclusion that the most statistically significant variable in their estimated regression is also the most important in terms of explaining the largest portion of the movement of the dependent variable. Statistical significance indicates the likelihood that a particular sample result could have been obtained by chance, but it says little – if anything – about which variables determine

the major portion of the variation in the dependent variable. To determine importance, a measure such as the size of the coefficient multiplied by the average size of the independent variable or standard error of the independent variable would make much more sense.

The t-test is not intended for tests of the entire population

The t-test helps make inferences about the true value of a parameter from an estimate calculated from a sample of the *population* (the group from which the sample is being drawn). **Read page 154.** The standard error will approach zero as the sample size approaches infinity. Thus, the t-score will eventually become

$$t = \frac{\beta}{0} = \infty$$

the mere existence of a large t-score for a huge sample has no real substantive significance, because of the sample size is large enough, you can reject almost any null hypothesis.

5.6 Appendix: The F-test

although the t-test is invaluable for hypotheses about individual regression coefficients, it cannot be used to test multiple hypotheses simultaneously. To test multiple hypotheses, most researchers would use the F-test.

What is the F-test?

F-test	Formal hypothesis test that is designed to deal with a null hypothesis that contains multiple hypotheses or a single hypothesis about a group of coefficients

The way in which the F-test works is ingenious

1. Translate the null hypothesis in question into constraints that will be placed on the equation. The resulting constrained equation can be thought of as what the equation would look like if the null hypothesis were correct.
   - In F-test, the H_0 always leads to a constrained equation, even if this violates our standard practice that the alternative hypothesis contains what we expect is true
2. Estimate the constrained equation with OLS and compare the fit of this constrained equation with the fit of the unconstrained equation.
   - Of the fits of the constrained equation and the unconstrained equation are not significantly different, the null hypothesis should **not** be rejected.
   - If the fit of the unconstrained equation is significantly better than that of the constrained equation, then we reject the null hypothesis.

The fit of the constrained equation is never superior to the fit of the unconstrained equation. The fits of the equations are compared with the general F-statistics

$$F = \frac{\frac{(RSS_M - RSS)}{M}}{\frac{RSS}{(N-K-1)}}$$

Where;

- RSS = residual sum of squares from the unconstrained equation
- RSS_M = residual sum of squared from the constrained equation
- M = number of constraints placed on the equation (usually equal tot eh number of βs eliminated from the unconstrained equation)
- (N – K – 1) = degrees of freedom in the unconstrained equation

RSS_M is always greater than or equal to RSS. Imposing constraints on the coefficients instead of allowing OLS to select their values can never decrease the summed squared residuals. Recall that OLS selects that combination of values of the coefficients that minimizes RSS). As the difference between the constrained coefficients and the unconstrained coefficients increases, the data indicates that the null hypothesis is less likely to be true. The decision rule to use the F-test is to reject the null hypothesis if the calculated F-value (F) from the equation above is greater than the appropriate critical F-value (F_c).

Reject **H_0 if $F > F_c$**
Do not reject **H_0 if $F \leq F_c$**

The critical F-value F_c is determined from the tables and depends on a level of significance chosen by the researcher and on the degrees of freedom. The F-statistic has two types of degrees of freedom:

1. degrees of freedom for the numerator (M, the number of constraints implies by H_0)
2. degrees of freedom for the denominator (N – K – 1, the degrees of freedom in regression equation)

The F-test of overall significance

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they do not provide a formal hypothesis test of that overall fit. Such a test is provided by the F-test. The null hypothesis in an F-test of overall significance is that all the slope coefficients in the equation equal zero simultaneously.

H_0: $ß_1 = ß_2 = \ldots = ß_K = 0$
H_A: H_0 is not true

To show that the overall fit of the estimated equation is statistically significant, we must be able to reject this null hypothesis using the F-test. For the F-test of overall significance, the equation above simplifies to:

$$F = \frac{\frac{ESS}{K}}{\frac{RSS}{(N-K-1)}} = \frac{\frac{\sum(\hat{Y}_i - \bar{Y})^2}{K}}{\frac{\sum e_i^2}{(N-K-1)}}$$

In this case, the constrained equation to which we are comparing the overall fit is:

$Y_i = ß_0 + \epsilon_i$

Which is nothing more than saying $\hat{Y}_i = \bar{Y}$. Thus, the F-test of the overall significance is testing the null hypothesis that the fit of the equation is not significantly better than that provided by using the mean alone. Just as p-values provide an alternative approach to the t-test, so too can p-values provide an alternative approach to the F-test of overall significance.

Other uses of the F-test

There are many other uses of the F-test besides the test of overall significance. For example, let's look at a Cobb-Douglas production function (**see page 169 – 172**):

$Q_t = ß_0 + ß_1 L_t + ß_2 K_t + \epsilon_t$

Chapter 7: specification: choosing a functional form

7.2 Alternative functional forms

there are different functional forms (**summed on page 234**)

- linear functional form
- double-log functional form
- semi log functional form
- polynomial functional form
- inverse functional form

before we can talk about functional forms, we need to make a distinction between an equation that is linear in the coefficients and one that is linear in the variables. An equation is **linear in the variables** if plotting the function in terms of X and Y generates a straight line.

$Y = ß_0 + ß_1X + \epsilon$ (**linear in the variables** and **linear in the coefficients**)
$Y = ß_0 + ß_1X^2 + \epsilon$ (**not linear in the variables** → not a straight line)
$Y = ß_0 + X^{ß1}$ (**not linear in the coefficients**)

An equation is **linear in the coefficients** only if the coefficients (the ßs) appear in their simplest form

- they are not raised to any powers (other than one)
- they are not multiplied or divided by other coefficients
- they do not themselves include some sort of function (like logs or exponents)

in fact, of all possible equations for a single explanatory variable, *only* function of the general form:

$f(Y) = ß_0 + ß_1f(X)$

are linear in the coefficients $ß_0$ and $ß_1$. Linear regression analysis can be applied to an equation that is nonlinear in the variables as long as the equation is linear in the coefficients. When econometricians use the phrase *linear regression,* they usually mean *regression that is linear in the coefficients.*

The choice of a functional form almost always should be based on the underlying theory and only rarely on which form provides the best fit. The logical form of the relationship between the dependent variable and the independent variable in question should be compared with the properties of various functional forms, and the one that comes closest to that underlying theory should be chosen.

Linear form

The linear regression model assumes that the slope of the relationships between the independent variable and the dependent variable is constant:

$$\frac{Y}{\Delta X_k} = \beta_K \qquad (k = 1, 2, \ldots, K)$$

if the hypothesized relationship between Y and X is such that the slope of the relationship can be expected to be constant, then the linear function form should be used. Since the slope is constant, the **elasticity** of Y concerning X can be calculated with:

$$\text{Elasticity}_{Y,X_K} = \frac{\Delta y/y}{\Delta X_k/X_k} = \frac{\Delta Y}{\Delta X_k} x \frac{X_k}{Y} = \beta_k \frac{X_k}{Y}$$

Double-log form

The double-log form is the most common functional form that is nonlinear in the variables while still being linear in the coefficients. In a **double-log functional form**, the natural log of Y is the dependent variable, and the natural log of X is the independent variable:

$\ln Y = ß_0 + ß_1 \ln X_1 + ß_2 \ln X_2 + \epsilon$

the double-log form, sometimes called the log-log form, often is used because a researcher has specified that the elasticities of the model are constant, and the slopes are not. This is in contrast to the linear model, in which the slopes are constant, but the elasticities are not. In a double-log equation, an individual regression coefficient can be interpreted as an elasticity because:

$ß_k = \frac{\Delta(\ln Y)}{\Delta(\ln X_k)} = \frac{\Delta y/y}{\Delta X_k/X_k} = \text{elasticity}_{Y,X_K}$

since regression coefficients are constant, the condition that the model has a constant elasticity is met by the double-log equation. The way to interpret $ß_k$ in a double-log equation is that if X_k increases by 1 percent while the other Xs are held constant, then Y will change by $ß_k$ percent. Since elasticities are constant, the slopes are now no longer constant. **See figure 2 on page 227.**

Double-log models should be run only when the logged variables take on positive values. Dummy variables, which can take on the value of zero, should not be logged but still can be used in a double-log equation if they are adjusted.

Semi log form

The semi-log **functional form** is a variant of the double-log equation in which some but not all the variables (dependent and independent) are expressed in terms of their natural logs:

$Y = ß_0 + ß_1 \ln X_1 + ß_2 X_{2i} + \epsilon_i$

Read page 228. Not all semi log functions have the log on the right-hand side of the equation. The alternative semi-log form is to have the log on the left-hand side of the equation. This would mean that the natural log of Y would be a function of unlogged values of the Xs

$\ln Y = ß_0 + ß_1 X_1 + ß_2 X_2 + \epsilon$

this model has neither a constant slope nor a constant elasticity, but the coefficients do have a very useful interpretation. If X_i increases by one *unit*, then Y will change in *percentage* terms. Specifically, Y will change by $ß_1 * 100$ percent, holding X_2 constant, for every unit that X_1 increases.

Polynomial form

In most functions, the slope of the cost curve changes sign as output changes. If the slopes of a relationship are expected to depend on the level of the variable itself, then a polynomial model should be considered. **Polynomial functional forms** express Y as a function of independent variables, some of which are raised to powers other than 1.

$Y_1 = ß_0 + ß_1 X_{1i} + ß_2 (X_{1i})^2 + ß_3 X_{2i} + \epsilon_i$

The slope of Y with respect to X_1 is

$\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1$

note that the slope depends on the level of X_1. With polynomial regressions, the interpretation of the individual regression coefficients become difficult, and the equation may produce unwanted results for ranges of X. Great care must be taken when using a polynomial regression equation to ensure that the functional form will achieve what is intended by the researcher and no more.

Inverse form

The **inverse functional form** expresses Y as a function of the reciprocal (or inverse) of one or more of the independent variables (in this case X_1)

$Y_1 = ß_0 + ß_1(1/X_{1i}) + ß_2X_{2i} + \epsilon_i$

The inverse functional form should be used when the impact of a particular independent variable is expected to approach zero as that independent variable approaches infinity. In the equation above, X_1 cannot equal zero, since if X_1 equalled zero, dividing it into anything would result in infinite or undefined values. The slope with respect to X_1 is:

$$\frac{\Delta Y}{\Delta X_1} = \frac{-\beta_1}{X_1^2}$$

the slopes for X_1 fall into two categories, both shown in figure 5 (page 233):

1. when $ß_1$ is positive, the slope with respect to X_1 is negative and decreases in absolute value as X_1 increases. As a result, the relationship between Y and X_1 holding constant X_2 approaches $ß_0 + ß_2X_2$ as X_1 increases (ignoring the error term).
2. When $ß_1$ is negative, the relationship intersects the X_1 axis at $\frac{-\beta_1}{(\beta_0 + \beta_2 X_2)}$ and slopes upward toward the same horizontal line (called asymptote) that is approaches when $ß_1$ is positive.

Choosing a functional form

The best way to choose a functional form for a regression model is to choose a specification that matches the underlying theory of the equation.

7.4 Using dummy variables

A dummy variable is one that takes on values of 0 or 1 (gender).

Intercept dummy	We can use dummy variables as an **intercept dummy**, a dummy variable that changes the constant or intercept term, depending on whether the qualitative condition is met

These take the general form

$Y_i = ß_0 + ß_1X_i + ß_2D_i + \epsilon_i$

Where;

- D_i = 1 if it observation meets a particular condition and 0 otherwise.

The intercept dummy does change the intercept depending on the value of D, but the slopes remain constant no matter what value D takes. The event not explicitly represented by a dummy variable, the **omitted condition**, forms the basis against which the included conditions are compared.

What would happen if you used two dummy variables to describe the two conditions?
Suppose: $X_1 = 1$ is a person is a male and $X_2 = 1$ if a person is female. In such a situation, X_1 plus X_2 would always add up to 1. Thus, X_1 would be perfectly, linearly correlated with X_2 and the equation would violate Classical Assumption VI. If you were to make this mistake, sometimes called a *dummy*

variable trap, you'd have perfect multicollinearity and OLS almost surely would fail to estimate the equation.

Create one less dummy variable than there are alternatives and to use each dummy to represent just one of the possible conditions. Dummy variables can be used as *dependent* variables. **Read page 238.**

7.5 slope dummy variables

$Y_i = ß_0 + ß_1X_i + ß_2D_i + \epsilon_i$

In this equation X is multiplied only by $ß_1$, and D is multiplied only by $ß_2$, and there are no other factors involved. This restriction does not apply to the **interaction term**.

Interaction term	An independent variable in a regression equation that is the *multiple* of two or more other independent variables.

Each interaction term has its own regression coefficient, so the result is that the interaction term has three or more components, as in $ß_3X_iD_i$. interaction terms can involve two quantitative variables ($ß_3X_1X_2$) or two dummy variables ($ß_3D_1D_2$), but the most frequent application of interaction terms involves one quantitative variable and one dummy variable ($ß_3X_1D_1$), a combination that is typically called a **slope dummy.**

Slope dummy variables	Allow the slope of the relationship between the dependent and an independent variable to be different depending on whether the condition specified by a dummy variable is met

This contrasts with an intercept dummy variable, which changes the intercept but does not change the slope when a general condition is met. In general, a slope dummy is introduced by adding to the equation a variable that is the multiple of the independent variable that has a slope you want to change and the dummy variable that you want to cause the changed slope. The general form of a slope dummy equation is:

$Y_i = ß_0 + ß_1X_i + ß_2D_i + ß_3X_iD_i + \epsilon_i$

When

- D = 0 $\Delta Y/\Delta X = ß_1$
- D = 1 $\Delta Y/\Delta X = (ß_1 + ß_3)$

In essence, the coefficient of X *changes* when the condition specified by D is met.

Chapter 8: multicollinearity

Perfect multicollinearity	Violation of Classical Assumption VI – that no independent variable is a perfect linear function of one or more other independent variables.

In essence, the more highly correlated two (or more) independent variables are, the more difficult it becomes to accurately estimate the coefficients of the true model.

8.1 Perfect versus imperfect multicollinearity

Perfect multicollinearity	Violates Classical Assumption VI, which specifies that no explanatory variable is a perfect linear function of any other explanatory variables.

The *perfect* in this context implies that the variation in one explanatory variable can be *completely* explained by movements in another explanatory variable:

$X_{1i} = \alpha_0 + \alpha_1 X_{2i}$

Where the αs are constants and the Xs are independent variables in:

$Y_i = ß_0 + ß_1 X_{1i} + ß_2 X_{2i} + \epsilon_i$

Notice that there is no error term in the first equation. This implies that X_1 can be exactly calculated given X_2 and the equation. With perfect multicollinearity, OLS is incapable of generating estimates of the regression coefficients, and most OLS computer programs will print out an error message in such a situation. Perfect multicollinearity ruins our ability to estimate the coefficients because the two variables cannot be distinguished. A special case related to perfect multicollinearity occurs when a variable that is definitionally related to the dependent variable is included as an independent variable in a regression equation. Such a **dominant variable** is so highly correlated with the dependent variable that it completely masks the effect of all other independent variables in the equation.

imperfect multicollinearity	A linear functional relationship between two or more independent variables that is so strong that it can significantly affect the estimation of the coefficients of the variables.

Imperfect multicollinearity occurs when two (or more) explanatory variables are imperfectly linearly related, as in:

$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + u_i$

Where u_i is a stochastic error term. Imperfect multicollinearity is a strong linear relationship between the explanatory variables. The stronger the relationship between the two (or more) explanatory variables, the more likely it is that they'll be considered significantly multicollinear. **Read page 265**.

8.2 the consequences of multicollinearity

OLS estimators are BLUE if the Classical Assumptions hold. This means that OLS estimates can be thought of as being unbiased and having minimum variance possible for unbiased linear estimators.

What are the consequences of multicollinearity?

The major consequences of multicollinearity are (**page 266 – 268**):

1. Estimates will remain unbiased
2. The variances and standard errors of the estimates will increase

3. The computed t-score will fall
   - $t_k = \frac{(\hat{\beta}_k - \hat{\beta}_{H_0})}{SE(\hat{\beta}_k)}$
   - multicollinearity increases the standard error of the estimated coefficients, and if the standard error increases, then the *t*-score must fall.
4. Estimates will become very sensitive to changes in specification
5. The overall fit of the equation and the estimation of the coefficients of non-multicollinear variables will be largely unaffected.
   - The overall fit is measured by $\bar{R}^2$ and this will not fall much, if at all, in the face of significant multicollinearity.

8.3 The detection of multicollinearity

a first step is to recognize that some multicollinearity exists in every equation (determine *how much* multicollinearity exists in an equation, not *whether* any multicollinearity exists). A second key point is that the severity of multicollinearity in each equation can change from sample to sample depending on the characteristics of the sample. As a result, the theoretical underpinnings of the equation are not quite as important in the detection of multicollinearity as they are in the detection of an omitted variable or an incorrect functional form. Because multicollinearity is a sample phenomenon, and the level of damage of its impact is a matter of degree, many of the methods used to detect it are informal tests without critical values or levels of significance.

High simple correlation coefficients

One way to detect severe multicollinearity is to examine the simple correlation coefficients between the explanatory variables. If an *r* is high in absolute value, then we know that these two Xs are quite correlated, and that multicollinearity is a potential problem. r is high is it causes unacceptably large variances in the coefficient estimates in which we are interested. But be careful. The use of simple correlation coefficients as an indication of the extent of multicollinearity involves a major limitation if there are more than two explanatory variables. As a result, simple correlation coefficients must be sufficient but not necessary tests for multicollinearity. Although a high r does indeed indicate the probability of severe multicollinearity, a low r by no means proves otherwise.

High variance inflation factors (VIFs)

The use of tests to give an indication of the severity of multicollinearity in a particular sample is controversial. One measure of severity of multicollinearity that is easy to use and that is gaining popularity is the variance inflation factor.

Variance inflation factor	A method of detecting the severity of multicollinearity by looking at the extent to which a given explanatory variable can be explained by all the other explanatory variables in the equation

The VIF is an index of how much multicollinearity has increased the variance of an estimated coefficient. A high VIF indicates that multicollinearity has increased the estimate variance of the estimated coefficient by quite a bit, yielding a decreased *t*-score.

Suppose you want to use the VIF to attempt to detect multicollinearity in an original equation with K independent variables:

$Y = ß_0 + ß_1X_1 + ß_2X_2 + \ldots + ß_kX_k + \epsilon$

Calculating the VIF for a given X_i involves two steps:

1. Run an OLS regression that has X_i as a function of all the other explanatory variables in the equation. For i = 1, this equation would be:
   - $X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + v$
   - Thus there are K auxiliary regressions, one for each independent variable in the original equation.
2. Calculate the variance inflation factor for $\hat{\beta}_i$
   - $\text{VIF}(\hat{\beta}_i) = \frac{1}{(1-R_i^2)}$

Where;

- v: stochastic error term
- R^2: coefficient of determination of the auxiliary regression in step one
- R_i^2: separate auxiliary regression for each independent variable in the original equation

the higher the VIF, the more severe the effects of multicollinearity.

- R_i^2 = 1: indicating perfect multicollinearity, produces a VIF of infinity
- R_i^2 = 0: indicating no multicollinearity at all, produces a VIF of 1.

While there is no table of formal critical VIF values, a common **rule of thumb** is that if VIF($ß_i$) > 5, the multicollinearity is severe. As the number of independent variables increases, it makes sense to increase this number slightly.

Some authors and statistical software programs replace the VIF with its reciprocal, (1 - R_i^2), called **tolerance** (or TOL). Whether calculating VIF or TOL is a matter of personal preference. Unfortunately, there are a couple of problems with using VIFs.

1. There is no hard-and-fast VIF decision rule.
2. It is possible to have multicollinear effects in an equation that has no large VIFs

So, there is no test that allows a researcher to reject the possibility of multicollinearity with any real certainty.

Chapter 9: serial correlation

9.1 pure versus impure serial correlation

Pure serial correlation	Occurs when classical assumption IV, which assumes uncorrelated observations of the error term, is violated in a *correctly specified* equation

Assumption IV implies that:

$E(r_{\epsilon_i \epsilon_j}) = 0$ and $(I \neq J)$

If the expected value of the simple correlation coefficient between any two observations of the error term is not equal to zero, then the error term is said to be serially correlated. When econometricians use the term serial correlation without any modifier, they are referring to pure serial correlation. The most assumed kind of serial correlation is **first-order serial correlation**, in which the current value of the error term is a function of the previous value of the error term:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

Where;

- ϵ: the error term of the equation in question
- ρ: the first-order autocorrelation coefficient
- u: a classical (not serially correlated) error term

ρ measures the functional relationship between the value of an observation of the error term and the value of the previous observation of the error term. The magnitude of ρ indicates the strength of the serial correlation in an equation:

- ρ = 0: then there is no serial correlation (because ϵ would equal u, a classical error term)
- ρ = |1| the value of the previous observation of the error term becomes more important in determining the value of ϵ_t, and a high degree of serial correlation exists

for ρ to be greater than one in absolute value is unreasonable because it implies that the error term has a tendency to continually increase in absolute value over time (*explode*). We can state that:

$-1 < \rho < +1$

the sign of ρ indicates the nature of the serial correlation in an equation.

Positive serial correlation	A positive value for ρ implies that the error term tends to have the same sign from one period to the next
Negative serial correlation	A negative value for ρ implies that the error term has a tendency to switch sign from negative to positive and back again in consecutive observations

In most time-series applications however, negative pure serial correlation is much less likely than positive pure serial correlation. As a result, most econometricians analysing pure serial correlation concern themselves primarily with positive serial correlation. Serial correlation can take on many forms other than first-order serial correlation.

Seasonally based serial correlation	In a quarterly model, for example, the current quarter's error term observation may be functionally related to the observations of the error term in the previous year
Second-order serial correlation	It is possible that the error term in an equation might be a function of more than one previous observation of the error term

Seasonally based serial correlation:

- $\epsilon_t = \rho\epsilon_{t-4} + u_t$

second-order serial correlation:

- $\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + u_t$

Impure serial correlation

Impure serial correlation	Correlation that is caused by a specification error such as an omitted variable or an incorrect functional form

While pure serial correlation is caused by the underlying distribution of the error term of the true specification of an equation (which cannot be changed by the researcher), impure serial correlation is caused by a specification error that often can be corrected.

The error term can be thought of as the effect of omitted variables, nonlinearities, measurement errors, and pure stochastic disturbances on the dependent variable. The error term for an incorrectly specified equation thus includes a portion of the effect of any omitted variables and/or a portion of the effect of the difference between the proper functional form and the one chosen by the researcher.

The proper remedy for serial correlation depends on whether the serial correlation is likely to be pure or impure. Not surprisingly, the best remedy for impure serial correlation is to attempt to find the omitted variable (or at least a good proxy) or the correct functional form of the equation. Both the bias and the impure serial correlation will disappear if the specification error is corrected. As a result, most econometricians try to make sure they have the best specification possible before they spend too much time worrying about pure serial correlation. **Example page 327 – 329.**

9.2 The consequences of serial correlation

serial correlation is more likely to have internal symptoms. It affects the estimated equation in a way that is not easily observable from an examination of just the results themselves. The existence of serial correlation in the error term of an equation violates Classical Assumption IV, and the estimation of the equation with OLS has at least three consequences:

1. pure serial correlation does not cause bias in the coefficient estimates
   - if the serial correlation is impure bias may be introduced using an incorrect specification
   - unbiased in this case is that the distributions of the $\hat{\beta}$ is still centered around the true ß
2. serial correlation causes OLS to no longer be the minimum variance estimator (of all the linear unbiased estimators)
   - the serially correlated error term causes the dependent variable to fluctuate in a way that the OLS estimation procedure sometimes attributes to the independent variables
3. serial correlation causes the OLS estimates of the SE($\hat{\beta}$)s to be biased, leasing to unreliable hypothesis testing
   - biased SE($\hat{\beta}$)s cause biased t-scores and unreliable hypothesis testing in general
   - typically, the bias in the estimate of SE($\hat{\beta}$) is negative, meaning that OLS underestimates the size of the SE of the coefficients. This comes about because serial correlation usually results in a pattern of observations that allows a better fit than the actual (not serially correlated) observations would otherwise justify.
   - This tendency of OLS to **underestimate** SE($\hat{\beta}$) means that OLS typically **overestimates** t-scores of the estimated coefficients, since: t = $\frac{(\hat{\beta}-\beta_{H_0})}{\mathrm{SE}(\hat{\beta})}$. this makes it more likely that we will reject a null hypothesis when it is in fact true (Type I error).

Thus, the t-scores printed out by a typical software regression package in the face of serial correlation are likely to be too high.

9.3 The Durbin-Watson *d* Test

the test for serial correlation that is most widely used is the Durbin-Watson *d* test.

The Durbin-Watson *d* Statistics

Durbin-Watson *d* statistics	Used to determine if there is first-order serial correlation in the error term of an equation by examining the *residuals* of a particular estimation of that equation.

It is important to use the Durbin-Watson *d* statistics only when the assumptions that underlie its derivation are met:

1. The regression model includes an intercept term
2. The serial correlation is first order in nature
3. The regression model does not include a lagged dependent variable as an independent variable

The equation for the *Durbin-Watson d statistics* for T observations is:

$d = \sum_2^T \frac{(e_t - e_{t-1})^2}{\sum_1^T e_t^2}$

where;

- e_t: the OLS residuals

The Durbin-Watson *d* statistics equals 0 if there is extreme positive serial correlation, 2 if there is no serial correlation, and 4 is there is extreme negative serial correlation. **See page 334.**

Using the Durbin-Watson *d* test

The Durbin-Watson *d* test is unusual in two respects:

1. econometricians almost never test the one-sided null hypothesis that there is a negative serial correlation in the residuals because negative serial correlation is quite difficult to explain theoretically in economic or business analysis. Its existence unusually means that impure serial correlation has been caused by some error of specification.
2. The Durbin-Watson test is sometimes inconclusive. Whereas previously explained decision rules always had only *acceptance* regions and *rejection* regions, the Durbin-Watson test has a third possibility, called the **inconclusive region**.

To test for positive serial correlation, the following steps are required:

1. Obtain the OLS residuals from the equation to be tested and calculate the *d* statistics using
   - $d = \sum_2^T \frac{(e_t - e_{t-1})^2}{\sum_1^T e_t^2}$
2. determine the sample size and the number of explanatory variables and then consult Statistical Tables B-4, B-5 or B-6 to find the upper critical d value (d_U) and the lower critical d value (d_L).
3. given the null hypothesis of no positive serial correlation and a one-sided alternative hypothesis:
   - H_0: $\rho \leq 0$ (no positive serial correlation)
     H_A: $\rho > 0$ (positive serial correlation)

   The appropriate decision rule is:
   - If $d < d_L$ reject H_0
   - If $d > d_U$ do not reject H_0
   - If $d_L \leq d \leq d_U$ inconclusive

   Read page 335 for two-sided *d* test and read 336 – 337 for example Durbin-Watson test.

Chapter 11: time-series models

11.4 Spurious correlation and non-stationarity

one problem with time-series data is that independent variables can appear to be more significant than they are if they have the same underlying trend as the dependent variable.

Spurious correlation	A strong relationship between two or more variables that is not caused by a real underlying causal relationship

If you run a regression in which the dependent variable and one or more independent variables are spuriously correlated, the result is a *spurious regression*, and the t-scores and overall fit of such spurious regressions are likely to be overstated and untrustworthy.

Stationary and nonstationary time series

Stationary time series	A series whose basis properties (for example its mean and its variance) do not change over time
Nonstationary time series	Has one or more basic properties that *do* change over time

A time-series variable X_t is stationary if:

1. The mean of X_t is constant over time
2. The variance of X_t is constant over time
3. The simple correlation coefficient between X_t and X_{t-k} depends on the length of the lag (k) but on no other variable (for all k).

If one or more of these properties is not met, then X_t is **non-stationary**. If a series is nonstationary, that problem is often referred to as **non-stationarity**. Besides *variables*, *error terms* can also be nonstationary. Many cases of heteroskedasticity in time-series data involve an error term with a variance that tends to increase with time. That kind of heteroskedastic error term is also nonstationary.

The major consequence of nonstationary regression analysis is the spurious correlation that inflates R^2 and the t-score of the nonstationary independent variables, which in turn leads to incorrect model specifications. Unfortunately, many economic time-series variables are nonstationary even after the removal of a time trend. This non-stationarity typically takes the form of the variable behaving as though it were a *random walk*.

Random walk	A time series variable where next period's value equals this periods value plus a stochastic error term.

A random-walk variable is nonstationary because it can wander up and down without an inherent equilibrium and without approaching a long-term mean of any sort.

Let's suppose that Y_t is generated by an equation that includes only past values of itself (an *autoregressive* equation):

$Y_t = \gamma Y_{t-1} + v_t$

Where;

- V_t = classical error term
- If $|\gamma| < 1$, then the expected value of Y_t will eventually approach 0 (and therefore be stationary) as the sample size gets bigger and bigger
- If $|\gamma| > 1$, then the expected value of Y_t will continuously increase, making Y_t nonstationary

Most importantly, what about If $|\gamma| = 1$?

- $Y_t = Y_{t-1} + v_t$

Than it is a random walk. The expected value of Y_t does not converge on any value, meaning that it is nonstationary.

Unit root	The circumstance where $\gamma = 1$. If a variable has a unit root, then the equation above holds, and the variable follows a random walk and is nonstationary.

The relationship between unit roots and non-stationarity is so strong that most econometricians use the words interchangeably, even though they recognize that both trends and unit-roots can cause non-stationarity.

Spurious regression

If the dependent variable and at least one independent variable in an equation are nonstationary, it is possible for the results of an OLS regression to be spurious. Consider the linear regression model:

$Y_t = \alpha_0 + ß_0 X_t + u_t$

If both X and Y are nonstationary, then they can be highly correlated for noncausal reasons, and our standard regression inference measures will almost surely be very misleading in that they will overstate $\bar{R}^2$ and the t-score for $\hat{\beta}_0$. To avoid spurious regression results, it is crucial to be sure that time-series variables are stationary before running regressions.

The Dickey-Fuller test

To ensure that the equations we estimate are not spurious, it is important to test for non-stationarity. If we can be reasonably sure that all variables are stationary, then we need not worry about spurious regressions.

Dickey-Fuller test	Standard method of testing for non-stationarity. Examines the hypothesis that the variable in question has a unit root and, as a result, is likely to benefit from being expressed in first-difference form.

How can you tell if a time series is nonstationary?

1. Visually examine the data
2. After this trend has been removed, the standard method of testing for non-stationarity is the **Dickey-Fuller test**

We looked at the value of γ to help us determine if Y was stationary or nonstationary

- $|\gamma| < 1$ then Y is stationary
- $|\gamma| > 1$ then Y is nonstationary
- $|\gamma| = 1$ then Y is stationary due to a unit root

Thus, we conclude that the autoregressive model is stationary if $|\gamma| < 1$ and nonstationary otherwise. So, it makes sense to estimate $Y_t = \gamma Y_{t-1} + v_t$ and determine if $|\gamma| < 1$ to see if Y is stationary, and that is almost exactly how the Dickey-Fuller test works:

1. Subtract Y_{t-1} from both sides of the equation
   - $(Y_t - Y_{t-1}) = (\gamma - 1)Y_{t-1} + v_t$
2. if we define $\Delta Y_t = Y_t - Y_{t-1}$ then we have the simplest form of the Dickey-Fuller test
   - $\Delta Y_t = ß_1 Y_{t-1} + v_t$
   - $ß_1 = (\gamma - 1)$
3. summarize the hypotheses
   - H_0: Y_t contains a unit root - $|\gamma| = 1$ and $ß_1 = 0$
   - H_A: Y_t is stationary - $|\gamma| < 1$ and $ß_1 < 0$
4. Hence, we construct a one-sided t-test on the hypothesis that $ß_1 = 0$
   - H_0: $ß_1 = 0$
   - H_A: $ß_1 < 0$

See page 406 for different forms of the Dickey-Fuller test. No matter what form of the Dickey-Fuller test we use, the decision rule is based on the estimate of $ß_1$.

- If $\hat{\beta}_1$ is significantly less than 0, then we can reject the null hypothesis of nonstationarity.

- If $\hat{\beta}_1$ is not significantly less than 0, then we cannot reject the null hypothesis of non-stationarity

The standard t-table does not apply to the Dickey-Fuller test. The critical values depend on the version of the Dickey-Fuller test that is applicable. Note that the equations for the Dickey-Fuller test and the critical values for each of the specifications are derived under the assumption that the error term is serially uncorrelated. If the error term is serially correlated, then the regression specification must be modified to take this serial correlation into account.

Cointegration

If the Dickey-Fuller test reveals non-stationarity, what should we do? The traditional approach has been to take the first differences ($\Delta Y = Y_t - Y_{t-1}$ and $\Delta X = X_t - X_{T-1}$) and use them in place of Y_t and X_t in the equation. Unfortunately, using first differences to correct for non-stationarity throws away information that economic theory can provide in the form of equilibrium relationships between the variables when they are expressed in their original units (X_t and Y_t). As a result, first differences should not be used without carefully weighing the costs and benefits of that shift, and in particular first differences should not be used until the residuals have been tested for **cointegration**.

Cointegration	Consist of matching the degree of nonstationarity of the variables in an equation in a way that makes the error terms (and residuals) of the equation stationary and rids the equation of any spurious regression results.

Even though individual variables might be nonstationary, it is possible for linear combinations of nonstationary variables to be stationary, or *cointegrated.* If a long-run equilibrium exists between a set of variables, those variables are said to be cointegrated. If the variables are cointegrated, then you can avoid spurious regressions even though the dependent variable and at least one independent variable are nonstationary. **Read page 408 – 409 (important).**

To sum, if the Dickey-Fuller test reveals that our variables have unit roots, the first step is to test for Cointegration in the residuals. If the nonstationary variables are not cointegrated, then the equation should be estimated using the first differences (ΔY and ΔX). however, if the nonstationary variables are cointegrated, then the equation can be estimated in its original units.

A standard sequence of steps for dealing with nonstationary time series

1. Specify the model. This model might be a time-series equation with no lagged variables, it might be a dynamic model in its simplest form, or it might be a dynamic model that includes lags in both the dependent and independent variables.
2. Test all variables for non-stationarity (technically unit roots) using the appropriate version of the Dickey-Fuller test
3. If the variables do not have unit roots, estimate the equation in its original units (Y and X)
4. If the variables have unit roots, test the residuals of the equation for cointegration using the Dickey-Fuller test.
5. If the variables have unit roots but are not cointegrated, then change the functional form of the model to first differences (ΔY and ΔX) and estimate the equation
6. If the variables have unit roots and are cointegrated, then estimate the equation in its original units

Online reading chapter 16: experimental and panel data

The experimental approach is important because it provides a possible way for regression analysis to provide evidence of causality. Panel data are formed when cross-sectional and time-series data sets are combined to create a single data set. Although some researchers use panel data to increase their sample size, the main reason for working with panel data is to provide an insight into analytical questions that can't be answered by using time-series or cross-sectional data alone.

16.1 Experimental methods in economics

correlation does not prove causality. Can experimental methods provide evidence of causality in economics?

Random assignment experiments

Random assignment is an experimental design in which the following steps are followed:

1. A sample of subjects is chosen or recruited, and then they are randomly assigned to one of two groups – the *control group* and the *treatment group*.

Treatment group	Receives the medicine that is being tested or the group who receives training
Control group	Receives a harmless ineffective placebo or the group who does not receive training
Random assignment experiments	If the treatment and control groups are chosen randomly, then such experiments are called random assignment experiments

Randomization helps ensure that any difference in outcome was caused by the treatment and not merely correlated with the treatment. The subjects' random assignment to the group should be enough to guarantee that the only *systematic* reason for observed differences are the chance consequence of the random assignment. The larger the sample, the more likely it is that random fluctuations will balance out. Factors other than the treatment that may affect the outcome are put in the error term, and the resulting equation is:

$$OUTCOME_i = \beta_0 + \beta_1 TREATMENT_i + \epsilon_i$$

Where;

- $OUTCOME_i$ = a measure of the desired outcome in the *i*th individual
- $TREATMENT_i$ = a dummy variable equal to 1 for individuals in the treatment group and 0 for individuals in the control group

$ß_1$ is often called the **difference estimator** because it measures the difference between the average outcome for the treatment group and the average outcome for the control group. If the estimated value of $ß_1$ is substantially different from zero in the direction predicted by theory, then we have evidence that the treatment did indeed cause the outcome to move in the expected direction. However, random assignment can't always control for all possible other factors, and we may be able to identify some of these factors and add them to our equation.

$$OUTCOME_i = \beta_0 + \beta_1 TREATMENT_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \epsilon_i$$

Where;

- X_1 = a dummy variable for the individual's gender
- X_2 = the individual's age

Unfortunately, random assignment experiments are not common in economics because they are subject to problems that typically do not plague medical experiments

1. Non-random examples
   - Most subjects in economic experiments are volunteers, and samples of volunteers often are not random.
   - The characteristics of the volunteer sample are not necessarily representative of the population.
2. Unobservable heterogeneity
   - In the second equation, we added observable factors to avoid omitted variable bias, but not all omitted factors in economics are observable.
3. The Hawthorne effects
   - The fact that human subjects know that they are being observed sometimes can change their behaviour, and this change in behaviour could clearly change the results of the experiment.
   - **Hawthorne effect**: the fact that people behave differently when they know they are being watched.
4. Impossible experiments

Natural Experiments

If random assignment experiments are not always feasible in economics, one alternative approach is to use data from natural experiments to try to get at issues of causality.

Natural experiments *(quasi experiments)*	are similar to random assignment except those observations fall into treatment and control groups *naturally* (because of an exogenous event) instead of being randomly assigned by the researcher.

This approach requires finding natural events or policy changes that can be analysed as if they were treatments in a random assignment experiment. If the natural exogenous (not under control of either of the groups), it turns out that a natural experiment can come very close to mimicking a random assignment experiment. The key is thus to find naturally occurring events that mimic a random assignment experiment.

A strict approach to natural experiments would seem to require that one find equivalents of *treatment* and *control* groups that have no systematic differences except for the treatment variable and other factors that can be observed and added to the equation. However, in economics, the treatment and control groups seem quite likely to have started off with different levels of the outcome measure. In addition, unobserved heterogeneity or non-random samples could result in the groups having different outcome measures. If the outcomes do not start of equal, then comparing outcomes after the treatment won't give us a true measure of the impact of the treatment.

To get around this problem, economists who run natural experiments do not compare outcomes between the treatment and control groups. Instead, they compare the change in outcomes. The resulting *difference in differences* measures the impact of the treatment on the outcome of the natural experiment. In a regression equation, the appropriate dependent variable in such a natural experiment thus is the difference in the outcome measure, not the outcome level:

$$\Delta OUTCOME_i = \beta_0 + \beta_1 TREATMENT_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \epsilon_i$$

Where;

- ΔOUTCOME = the outcome after the treatment minus the outcome before the treatment for the *i*th observation

$ß_1$	The **difference-in-differences estimator**. It measures the difference between the change in the treatment group and the change in the control group, holding constant X_1 and X_2.

In essence, the difference-in-differences estimator uses the change in the control group as a measure of what would have happened to the treatment group if there hadn't been a treatment. The validity of this approach thus depends on the assumption that the changes in the outcome would have been the same in both the treatment and control group had there been no treatment.

In the equation above

- $ß_2$ = measures the impact of one-unit increase in X_1 on the *change* in the outcome (not the level of the outcome as before)

one final note. It is important to think through the appropriate *before* and *after* time frames when you are collecting data for a natural experiment. Data on the control and treatment groups should come from a time period far enough in advance of the policy change (treatment) that you are not picking up any anticipatory effects of the intended policy change.

16.2 Panel Data

when a case has both time-series and cross-sectional dimensions, it is neither time-series nor cross-sectional. It is a **panel data set**.

What are panel data

Panel data	Combine time-series and cross-sectional data in a very specific way. Panel data include observations on the same variables from the same cross-sectional sample from two or more *different* time periods

Not every data set that combines time-series and cross-sectional data meet this definition. If different variables are observed in the different time periods or if the data are drawn from different samples in the different time periods, then the data are not considered to be panel data.

Why use panel data?

1. Panel data certainty will increase the sample sizes
2. Panel data provide insight into analytical questions that cannot be answered by using time-series or cross-sectional data.
3. Panel data often allow researchers to avoid omitted variable problems that otherwise would cause bias in cross-sectional studies.

There are four different kinds of variables that we encounter when we use panel data

1. Variables that can differ between individuals but do not change over time (gender, ethnicity)
2. Variables that change over time but are the same for all individuals in a given time period (retail price index, national unemployment rate)
3. Variables that vary both over time and between individuals (income, marital status)
4. Trend variables that vary in predictable ways (individual's age)

To estimate an equation using panel data, it is crucial to ensure that the data are in the right order. Typically, panel data are grouped by starting with the first cross-section for all time periods, followed by the second cross-section for all time periods, and so on. This format is usually called **long form** because it results in a narrow but long data file.

Finally, the use of panel data requires a slight expansion of our notation. In the past we have used the subscript *i* to indicate the observation number in a cross-sectional data set, so Y_i indicated Y for

the *i*th cross-sectional observation. Similarly, we have used the subscript *t* to indicate the observation number in a time series, so Y_t indicated Y for the *t*th time-series observation. In a panel data set, however, variables will have both cross-sectional and time-series components, so we will use *both* subscripts. As a result, Y_{it} indicates Y for the *i*th cross-sectional and *t*th time-series observation. This notation expansion also applies to independent variables and error terms.

The Fixed effects model

There are several alternative estimation procedures for estimating panel data equations, but most researchers use the *fixed effects model.* The **fixed effect model** is a method of estimating panel data equations that work by allowing each cross-sectional unit to have a different intercept:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 D2_i + \cdots + \beta_N DN_i + v_{it}$$

Where,

- D2 = intercept dummy equal to 1 for the second cross-sectional entity and 0 otherwise
- DN = intercept dummy equal to 1 for the *N*th cross-sectional entity and 0 otherwise

What we are doing is allowing each cross-section's intercept to differ. Because the ßs are constant across units, in essence, we have N parallel regression lines. Observations across time in each unit vary around a baseline level specific to that unit. One major advantage of the fixed effects model is that it avoids bias due to omitted variables that do not change over time (race or gender). Such time-invariant omitted variables often are referred to as unobserved heterogeneity or a fixed effect. To understand how the fixed effects model does this, look how the equation would look like with only two years' worth of data:

Y_{it} $= \beta_0 + \beta_1 X_{it} + \beta_2 D2_i + v_{it}$

v_{it} = classical error term + unobserved impact of the time-invariant omitted variables

$= \epsilon_{it} + a_i$

Y_{it} $= \beta_0 + \beta_1 X_{it} + \beta_2 D2_i + \epsilon_{it} + a_i$

the unobserved impact *a* has only one subscript because it is a function of omitted variables that do not change over time. The key is to think about how much each observation of a variable differs from the *average* for that variable:

$$\overline{Y}_i = \beta_0 + \beta_1 \overline{X}_i + \beta_2 D2_i + \overline{\epsilon}_i + a_i$$

where the bar over a variable indicates the mean of that variable with respect to time.

$$Y_{it} - \overline{Y}_i = \beta_1 (X_{it} - \overline{X}_i) + \epsilon_{it} - \overline{\epsilon}_i$$

If we use the symbol θ to indicate a *demand variable* (a variable that has had its mean subtracted from it) and if we add $ß_0$ to the equation above to avoid violating Classical Assumption II, we obtain

$$\theta Y_{it} = \beta_0 + \beta_1 \theta X_{it} + \theta \epsilon_{it}$$

where ;

- θY_{it} = the demeaned Y = $Y_{it} - \overline{Y}_i$
- θX_{it} = the demeaned X = $X_{it} - \overline{X}_i$
- $\theta \epsilon_{it}$ = the demeaned $\epsilon = \epsilon_{it} - \overline{\epsilon}_i$

in actual practice many researchers use $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 D2_i + \cdots + \beta_N DN_i + v_{it}$.

The fixed effects model has some drawbacks

1. degrees of freedom for the fixed-effects model tend to be low because we lose one degree of freedom for cross-sectional observations because of the time-demeaning.
2. Any substantive explanatory variables that do not vary across time in each unit will be perfectly collinear with the fixed effects, so we cannot include them in the model or estimate their coefficients.

16.3 Fixed versus random effects

The fixed-effects model does a good job estimating panel data equations, and it also helps avoid omitted variable bias due to unobservable heterogeneity. As a result, the fixed effects model is the panel data estimation procedure that we recommend. However, many researchers use an advanced panel data method called the random-effects *model*.

The Random-effects model

An alternative to the fixed effects model is called the random-effects model. While the fixed effects model assumes that each cross-sectional unit has its own intercept, the random-effects **model** assumes that the intercept for each cross-sectional unit is drawn from a distribution that is centred around a mean intercept. Thus, each intercept is a random draw from an *intercept distribution* and therefore is independent of the error term for any observation.

The random-effects model has several clear advantages over the fixed effects model

1. A random-effects model will have quite a few more degrees of freedom than a fixed model because rather than estimating an intercept for virtually every cross-sectional unit, all we need to do is to estimate the parameters that describe the distribution of the intercepts.
2. You can estimate coefficients for explanatory variables that are constant over time

However, the random effects estimator has a major disadvantage

1. It requires us to assume that a_i is uncorrelated with the independent variables, the Xs, If we are going to avoid omitted variable bias.

Choosing between fixed and random effects

1. One key is the nature of the relationship between a_i and the Xs
   - If they are likely to be correlated, then it makes sense to use the fixed effects model, as that sweeps away the a_i and the potential omitted variable bias.

Many researchers use the **Hausman test** to see whether the regression coefficients under the fixed effects and random effects model are statistically different from each other. If they are different, then the fixed effects model is preferred even though it uses up many more degrees of freedom. If the coefficients are not different, then researchers either use the random effects model (in order to conserve more degrees of freedom) or provide estimates of both the fixed effects and random effects model.

Learning objectives

The aim of the course is to teach bachelor students to apply econometric models for cross-sectional data (measured at one point in time) and longitudinal data (repeated measures over time) to answer substantial research questions using the general-purpose statistical software package Stata. Techniques discussed in the course are simple and multiple regression analysis, time series analysis, cointegration, and analysis of panel data. The focus will be on determining economic associations among variables, performing statistical tests of the associations, economic interpretation of the results and presenting the results in a scientific paper. Students also learn to work with Stata syntax (do) file

Disclaimer

ESV Nijmegen tries to keep the content of this summary up to date and where needed complements it. Despite these efforts, it is still possible that the content is incomplete or incorrect. The offered material is a supplement for studying next to the appointed literature. The material is offered without any guarantee or claim for correctness.

Tips and remarks about the summary can be sent to secretaris@esvnijmegen.nl.