Notes based on Using Econometrics 5th Edition by A H Studenmund

by Bry42

Contents

1 An Overview of Regression Analysis

1.1 Introduction

Econometrics literally means ”Economic Measurement”. It is the quantitative measurement and analysis of actual economic and business phenomena. It attempts to quantify economic theory and the real world of human activity.
There are 3 major uses of econometrics, namely

- describing economic reality
- testing hypotheses about economic theory
- forecasting future economic activity.

The simplest use of econometrics is **description**. We can use econometrics to quantify economic activity and put them in equations that previously contained only abstract symbols. Consider the general relationship

$$Q = f(P, P_s, Y_d) \tag{1}$$

where Q is quantity demanded, P is the commodity's price, P_s is the price of a substitute and Y_d is disposable income. Econometrics allows this general and purely theoretical relationship to become explicit, like

$$Q = 27.6 - 0.61P + 0.09P_s + 0.24Y_d \tag{2}$$

The next use of econometrics is **hypothesis testing**, the evaluation of alternative theories with quantitative evidence. For example, you could test the hypothesis that the product in (**??**) is a normal good by applying various statistical tests to the **estimated regression coefficient** 0.24 of disposable income in (**??**). The evidence would seem to support this hypothesis because the coefficient's sign is positive, but the ”statistical significance” of that estimate would have to be investigated before such a conclusion could be justified.

The third and most difficult use of econometrics is to **forecast** what is likely to happen in the future, based on what has happened in the past. Eg. In (**??**) the president of a company who sold the product would want to decide whether to increase price of product based on collected past data.

1.2 What is Regression Analysis

Economic theory can give us the direction of a change but if we want to know "how much?" instead of just "how", then we need 2 other items, namely a sample of data and a way to estimate such a relationship (one of the most frequently used methods is **regression analysis**). Formally, regression analysis is 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, the **independent variables**, through the quantification of a single equation.

In (**??**), Q is the **dependent variable** and P, P_s and Y_d are the **independent variables**. Of course, a statistically significant regression result does NOT necessarily imply causality. We also need to rely on economic theory and, sometimes, common sense.

1.3 Single-Equation Linear Models

The simplest example of a Single-Equation Linear Model is

$$Y = \beta_0 + \beta_1 X \tag{3}$$

In (**??**), Y, the dependent variable, is a single-equation linear function of X, the independent variable. β_0 and β_1 are called **coefficients**. β_0 is the **constant** or **intercept** term. β_1 is the **slope coefficient**, the amount that Y will change when X changes by 1 unit. For a linear model, β_1 is constant over the entire function. We also know that, by definition,

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

We can only apply linear regression techniques *if and only if* the equation is linear, like (**??**). In contrast, the equation

$$Y = \beta_0 + \beta_1 X^2 \tag{4}$$

is **not linear** (but quadratic). However, we can easily make (**??**) linear by redefining it. First define

$$Z = X^2 \tag{5}$$

and, by substituting (**??**) into (**??**) we get

$$Y = \beta_0 + \beta_1 Z \tag{6}$$

and now (**??**) is now linear with the coefficients being β_0 and β_1 and the variables being Y and Z.

1.4 The Stochastic Error Term

The variation of the dependent variable, Y is understood to be caused by the independent variable, X. However, there is almost always variation caused by other variables too. It could be in the form of other potentially important explanatory variables that are missing (for example X_2 and X_3). However, even if they are added, there is going to be some variation of Y that simply could not be explained by the model. This variation could come from measurement errors, incorrect functional forms or purely random and totally unpredictable occurrences (noise or disturbance).

Econometricians include a **stochastic error term** or random error term, ϵ in their models. A stochastic error term is a term added to the regression model to introduce all of the variation in Y what cannot be explained by the included independent terms, or Xs. Now, with the stochastic error term, ϵ, (**??**) becomes

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{7}$$

In (**??**), there are two components, namely the **deterministic component** $[\beta_0 + \beta_1 X]$ and the **stochastic/random component** $[\epsilon]$.
The **deterministic component** is so called because it indicates the value of Y that is determined by a given value of X (which is assumed to be non-stochastic). It can be thought of as the expected value of Y given X, or the mean value of all the Ys associated with a particular value of X. Mathematically, the deterministic part of the equation may be expressed as $E(Y|X)$ and is written as

$$E(Y|X) = \beta_0 + \beta_1 X \tag{8}$$

This mathematical notation in (**??**) also means that this value is the **conditional expectation** of Y on X.

Of course, the value of Y is unlikely to be exactly equal to the deterministic, expected value $E(Y|X)$ and so the stochastic element has to be added to the equation, making

$$Y = E(Y|X) + \epsilon = \beta_0 + \beta_1 X + \epsilon \tag{9}$$

Example: Aggregate Consumption Function
Aggregate consumption, as a function of aggregate income, may be higher or lower than it would otherwise because of consumer uncertainty but it is hard to measure and hence, usually omitted. Next, there could be a measurement error or sampling error so the observed consumption and actual consumption could differ. Third, human behaviour always contains elements of pure chance which are unpredictable so random events may change aggregate consumption at any time. Fourth, the underlying consumption function might be estimated to be linear when it is, in reality, nonlinear.

Whenever one or more of these factors are at play, the observed Y will differ from the Y predicted from the deterministic part, $[\beta_0 + \beta_1 X]$

1.5 Extending the Notation

The single-equation linear case is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2, \cdots, N \tag{10}$$

where Y_i is the i-th observation of the dependent variable, X_i is the i-th observation of the independent variable, ϵ_i is the i-th observation of the stochastic error term, β_0, β_1 are the regression coefficients and N is the number of observations. So, in reality, like in (**??**), there are N equations, one for each observation. The coefficients β_0 and β_1 are **the same**, while the values of Y, X and ϵ **differ** across the observations.

In the general case, **multivariate** regression is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \qquad i = 1, 2, \cdots, N \tag{11}$$

where, X_{1i} is the i-th observation of the first independent variable, X_{2i} is the i-th observation of the second independent variable, X_{3i} is the i-th observation of the third independent variable. Now, the meaning of the regression coefficient β_1 in (**??**) is the impact of a one unit increase of X_1, *holding constant* the other included independent variables, X_2 and X_3. The same can be said for β_2 and β_3. These **multivariate regression coefficients** serve to isolate the impact on Y when a particular variable of the Xs change.

1.6 Example: Wage Regression

Redefining (**??**), let wages ($WAGE$) depend on the years of work experience (EXP), years of eduction (EDU) and gender of the worker ($GEND$, where

$GEND = 1$ if male and $GEND = 0$ if female). Substituting into (**??**), we get

$$WAGE = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \epsilon_i \tag{12}$$

So β_1 would mean the increase in a person's wage for and additional year of experience, holding education and gender constant.

Concluding, the general equation for K independent variables is

$$\begin{aligned} &Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \\ &\text{where } i = 1, 2, \cdots, N \end{aligned} \tag{13}$$

If the sample consists of **time series** data like data stretching years, months or days (eg. the daily exchange rate), use the subscript $_t$. If it contains **cross section** data or data on individuals , use the subscript "$_i$". Finally, use the subscript "$_{it}$" for **panel data**, when you look at multi-dimensional data frequently involving measurements over time.

1.7 The Estimated Regression Equation

Once a specific equation has been decided upon it must be quantified. This quantified version of the theoretical regression equation is called the **estimated regression equation** and is obtained from a sample of data for actual Ys and Xs. Although the theoratical equation is purely abstract in nature like

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{14}$$

the estimated regression equation has numbers in it, like

$$\hat{Y}_i = 103.40 + 6.38 X_i \tag{15}$$

The observed, real-world values of X and Y are used to calculate the coefficient estimates 103.40 and 6.38. These estimates are used to determine $\hat{Y}$, the *estimated* or *fitted* value of Y. The theoretical regression coefficients, β_0 and β_1, of the theoretical regression equation has been replaced by*estimates* of those coefficients, 103.40 and 6.38. We cannot actually observe the true regression coefficients so we calculate estimates of those coefficients from the data. the **estimated regression coefficients**, generally denoted as $\hat{\beta_0}$ and $\hat{\beta_1}$ are emperical, best guesses of the true regression coefficients and are obtained from a sample of Ys and Xs. The expression

$$\hat{Y}_i = \hat{\beta_0} + \hat{\beta_1} X_i \tag{16}$$

is the empirical counterpart of the theoretical regression equation of (**??**). The calculated estimates in (**??**) are examples of estimated regression coefficients $\hat{\beta_0}$ and $\hat{\beta_1}$ For each sample we calculate different set of estimated regression coefficients.
The difference between the estimated value of the dependent variable, $\hat{Y_i}$, and the actual value of the dependent variable, Y_i is defined as the residual. Mathematically,

$$e_i = Y_i - \hat{Y_i} \tag{17}$$

and this is **different** from the error term, ϵ_i, given as

$$\epsilon_i = Y_i - E(Y_i|X_i) \tag{18}$$

The residual is the difference between the observed Y and the estimated regression line, $\hat{Y}$ while the error term is the difference between the observed Y and the true regression equation. The error term is a theoretical concept and can never be observed, but the residual is a real world value which is calculated for each observation. The residual could be thought of as an estimate of the error term or $\hat{\epsilon}$. The smaller the residual, the better the fit, the closer the $\hat{Y}$s to the actual Ys.

Of course, the estimated regression model could be extended to more than one independent variable. Extending (**??**)

$$\begin{aligned}&\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1}X_{1i} + \hat{\beta_2}X_{2i} + \cdots + \hat{\beta_K}X_{Ki} \\ &\text{where } i = 1, 2, \cdots, N\end{aligned} \tag{19}$$

1.8 Example: Using Regression to Explain Housing Prices

Houses are not homogeneous products like corn or gold, which have generally known prices. So how do we appraise a house? Consider a specific case where the area is 1600 sq feet and its asking price was \$230,000. Let the price of a house depend on its size. We first construct the theoretical model of the price of a house by saying

$$\begin{aligned}PRICE_i &= \overbrace{f(SIZE_i)}^{+} + \epsilon_i \\ &= \beta_0 + \beta_1 SIZE_i + \epsilon_i\end{aligned} \tag{20}$$

where $PRICE_i$ is price of the i-th house in \$000, $SIZE_i$ is the size of that house in sq feet and ϵ_i is the stochastic error term for that house.
We then collect *cross-sectional* data on prices and sizes, respectively for, say

43 houses and this yields the estimated regression line

$$\hat{PRICE}_i = 40.0 + 0.138 SIZE_i \tag{21}$$

In (**??**), $\hat{\beta}_0$ means that every 1 sq feet increase in the size of the house commands an increase of $138. Now, we substitute $SIZE = 1600$ into (**??**) and yield $PRICE = 260.8$. Hence, it is a good deal because the appraised price, $260,800 is higher than the asking price of $230,000.

Notes

1. the interpretation of the intercept term, 40.0 in this case, is problematic because it represents the price of house with 0.00 sq feet. In this case, it is better to not interpret the value of $\hat{\beta}_0$ at all.

2. the example is very simplified as it assumes that the price of a house only depends on its size.

2 Ordinary Least Squares

2.1 Estimating Single-Independent Variable Models with OLS

The objective of regression analysis is to start from the purely theoretical equation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{22}$$

and, through the use of data, arrive at the estimated equation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{23}$$

where each of the "hats" is a sample estimate of the population value. How do we move from (**??**) to (**??**)? One of the most widely used methods is **Ordinary Least Squares (OLS)**. OLS is a regression estimation technique that calculates the values of β_0 and β_1 to minimize the sum of the squared residuals, or

$$\text{minimize} \quad \sum_{i=1}^{N} e_i^2 \tag{24}$$

or the sum of the squared deviations of the **vertical distance** between the observations. (**??**) is also expressed as $\sum_i^N (Y_i - \hat{Y}_i)^2$.

2.1.1 Why Use Ordinary Least Squares?

There are at least three reasons for using OLS to estimate regression models. First, it is relatively easy to use. Second, the goal of minimizing $\sum e_i^2$ is appropriate from a theoretical point of view or it can take into account positive or negative deviations of $(Y_i - \hat{Y}_i)$. Finally, and OLS estimates have a number of useful characteristics. The reasons are

- the sum of the residuals is *exactly* zero
- OLS can be shown to be the "best" estimator under a set of specific assumptions

An **estimator** is a mathematical technique that is applied to a sample of data to produce real-world **estimates** of the true population regression coefficients (or other parameters). Thus, OLS is an **estimator** and βs produced by OLS are called **estimates**.

2.1.2 How does OLS work?

From (**??**), OLS selects those estimates of β_0 and β_1 that minimizes the squared residuals, summed over all the sample data points. For an equation with one variable, the coefficients are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left[\left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right) \right]}{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2} \tag{25}$$

and, given the definition of β_1 by (**??**),

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{26}$$

where $\bar{X}$ is the mean of X, or $\frac{\sum X}{N}$ and $\bar{Y}$ is the mean of Y, or $\frac{\sum Y}{N}$. For each data set, we obtain different estimates of β_0 and β_1. We move now to a proof of (**??**) and (**??**).

Since

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \qquad \cdots\cdots(\clubsuit)$$

and

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \cdots\cdots(\spadesuit)$$

Substituting ($\spadesuit$) into ($\clubsuit$) yields

$$\sum_i e_i^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Let $S = \sum_i e_i^2$. Differentiating S w.r.t. β_0,

$$\begin{aligned} \frac{\partial S}{\partial \hat{\beta}_0} &= \frac{\partial}{\partial \hat{\beta}_0} \left[\textstyle\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] \\ &= -2 \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \qquad \cdots\cdots \heartsuit \end{aligned}$$

and, similarly, differentiating S w.r.t. β_1,

$$\begin{aligned} \frac{\partial S}{\partial \hat{\beta}_1} &= \frac{\partial}{\partial \hat{\beta}_1} \left[\textstyle\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] \\ &= 2 \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i) \qquad \cdots\cdots \diamondsuit \end{aligned}$$

Setting ($\heartsuit$) and ($\diamondsuit$) to 0,

$$\begin{aligned}
0 &= -2\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
&= \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
0 &= \sum_i Y_i - N\hat{\beta}_0 - \beta_1 \sum_i X_i \\
\sum_i Y_i &= N\hat{\beta}_0 + \beta_1 \sum_i X_i \qquad \cdots\cdots \mathcal{A}
\end{aligned}$$

and

$$\begin{aligned}
0 &= 2\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i) \\
&= \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i) \\
0 &= \sum_i X_i Y_i - \hat{\beta}_0 \sum_i X_i - \hat{\beta}_1 \sum_i X_i^2 \\
\sum_i X_i Y_i &= \hat{\beta}_0 \sum_i X_i + \hat{\beta}_1 \sum_i X_i^2 \qquad \cdots\cdots \mathcal{B}
\end{aligned}$$

From A,

$$\hat{\beta}_0 = \frac{\sum_i Y_i - \hat{\beta}_1 \sum_i X_i}{N} \qquad \cdots\cdots \mathcal{C}$$

and substituting $\mathcal{C}$ into $\mathcal{B}$,

$$\begin{aligned}
\sum_i X_i Y_i &= \frac{(\sum_i X_i)(\sum_i Y_i) - (\sum_i X_i)(\hat{\beta}_1 \sum_i X_i)}{N} + \hat{\beta}_1 \sum_i X_i^2 \\
N\sum_i X_i Y_i &= \sum_i X_i \sum_i Y_i - \hat{\beta}_1 (\sum_i X_i)^2 + N\hat{\beta}_1 \sum_i X_i^2 \\
N\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i &= \hat{\beta}_1 [N\sum_i X_i^2 - (\sum_i X_i)^2] \\
\hat{\beta}_1 &= \frac{N\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{N\sum_i X_i^2 - (\sum_i X_i)^2} \qquad \cdots\cdots \mathcal{D}
\end{aligned}$$

Finally, from $\mathcal{D}$,

$$\begin{aligned}
N \cdot E(Y) &= N\hat{\beta}_0 + N\hat{\beta}_1 \cdot E(X) \\
\beta_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \cdots\cdots \mathcal{E}
\end{aligned}$$

Equations $\mathcal{D}$ and $\mathcal{E}$ are actually (**??**) and (**??**), respectively.

2.1.3 An Illustration of OLS Estimation

The exercise has been completed and the results are in the Excel File **Topic2Data_Bry [Table 2.1]**. After the OLS estimator has been run on the data, $\beta_0 = 103.397$ and $\beta_1 = 6.377$ and so the estimated regression equation is

$$\hat{Y_i} = 103.397 + 6.377X_i \tag{27}$$

The actual values from the textbook is $\beta_0 = 103.4$ and $\beta_1 = 6.38$. They are verified as rounded off values of the OLS that was run in the Excel File.

2.2 Estimating Multivariate Regression Models with OLS

2.2.1 The Meaning of Multivariate Regression Coefficients

Recall that the general multivariate model with K independent variables is

$$\begin{aligned}&Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i\\ &\text{where } i = 1, 2, \cdots, N \end{aligned} \tag{$\Diamond$}$$

The slope coefficients of ($\Diamond$), often called **partial regression coefficients** and are defined to allow a researcher to distinguish the impact of one variable on the dependent variable from that of the other independent variables. A **multivariate regression coefficient** indicates the change in the independent variable associated with a one-unit increase of the independent variable *holding all the other independent variables constant.*

2.2.2 Example: Beef Consumption in USA

Consider the model for the annual model of per capita demand for beef in the US

$$\widehat{CB}_t = 37.54 - 0.88P_t + 11.9Yd_t \tag{28}$$

where, in (**??**), $\widehat{CB}$ is the per capita consumption of beef in time t in lb/person, P_t is price of beef in year t in cents/lb and Yd_t is per capita disposable income in year t in \$000.

The estimated coefficient of income, 11.9 says that beef consumption increases by 11.9 lb/person if the disposable income per person increases by \$1000, holding the price of beef constant.

2.2.3 OLS Estimation of Multivariate Regression Models

Consider the application of OLS to an equation with two independent variables

$$Y_i = \beta_0 + \beta_1 X_{1i} = \beta_2 X_{2i} + \epsilon_i \tag{29}$$

The goal of OLS is to choose $\hat{\beta}$s that minimize the summed square residuals. Now, they form a *multivariate* model. But, they can be minimized using the similar approach. The equations are cumbersome but the principles are the same.

2.2.4 An Example of a Multivariate Regression Model

Consider a model of financial aid at a liberal arts college, where

$$FINAID_i = f(\overbrace{PARENT_i}^{-}, \overbrace{HSRANK_i}^{+}) \tag{30}$$

and, making it explicit

$$FINAID_i = \beta_0 + \beta_1 PARENT_i + \beta_2 HSRANK_i + \epsilon_i \tag{31}$$

where $FINAID_i$ is the financial aid, in dollars, awarded to the i-th applicant, $PARENT_i$ is the amount in dollars the i-th applicant's parents can contribute to college and $HSRANK_i$ is the GPA rank of the i-th applicant in high school, as measured as a percentage. OLS has been run on the data and the results are in results are in the Excel File **Topic2Data_Bry [Table 2.2]**

If we estimate (**??**) using OLS and the data in Table 2.2, we get the equation

$$\widehat{FINAID}_i = 8927 - 0.36 PARENT_i + 87.4 HSRANK_i \tag{32}$$

(Note that OLS has been attempted on the data but it has not been successful yet.)

The coefficient -0.36 means that the model implies that the i-th student's financial aid grant will fall by \$0.36 for every dollar increase in the parent's ability to pay, holding high school rank constant. This coefficient makes sense and meets our expectations. The coefficient 87.4 means that the i-th's student financial aid grant will increase by \$87.4 for every one percentage point increase in the high school ranking, holding the family's ability to pay constant.

2.3 Total, Explained and Residual Sum of Squares

Total sum of squares or **TSS** is the measure of the amount of variation to be explained by the the regression. Mathematically, TSS is expressed as

$$\text{TSS} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \tag{33}$$

For OLS, the total sum of squares has two components – the variation which could be explained by the regression and the variation which cannot

$$\begin{aligned}
\overbrace{\sum_i (Y_i - \bar{Y})^2}^{\text{Total Sum of Squares, TSS}} \quad &= \quad \overbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}^{\text{Explained Sum of Squares, ESS}} \quad + \quad \overbrace{\sum_i e_i^2}^{\text{Residual Sum of Squares, RSS}} \\
&= \sum_i (\hat{Y}_i - \bar{Y})^2 - \sum_i (Y_i - \hat{Y}_i)
\end{aligned} \tag{34}$$

Equation (**??**) is usually called the "decomposition of variance". The first component of (**??**) measures the amount of the squared deviation of Y_i from its mean that is explained by the regression line. The component of the total sum of the squared deviations, the **explained sum of squares** or **ESS** is attributable to the fitted regression line. The unexplained portion of TSS is called the **residual sum of squares** or **RSS**.

From (**??**), the smaller the RSS relative to the ESS, the better the estimated regression line fits the data. OLS is the estimation technique which minimizes the RSS and, therefore maximizes the ESS.

2.4 Describing the Overall Fit of the Estimated Model

2.4.1 R^2, the Coefficient of Determination

The most common measure of fit is the **coefficient of determination**, R^2. The coefficient of determination is the ration of the ESS relative to the TSS

$$\begin{aligned}
R^2 &= \frac{\text{ESS}}{\text{TSS}} \\
&= 1 - \frac{\text{RSS}}{\text{TSS}} \\
&= 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}
\end{aligned} \tag{35}$$

The higher the R^2, the closer the estimated regression equation fits the sample data. Since TSS, ESS and RSS are all nonnegative, and ESS $\leq$ TSS, R^2 must lie in the interval $0 \leq R^2 \leq 1$. A high value of R^2 demonstrates a relationship between X and Y that can be explained quite well by a linear regression equation. Most of the variation has been explained but there still remains a portion of the variation that is essentially random or unexplained by the model.

2.4.2 The Simple Correlation Coefficient, r

The **simple correlation coefficient**, r is a measure of the strength and direction of the linear relationship between the two variables. The interval of r is $-1 \leq r \leq 1$, and the sign of r indicates the direction of the correlation between the two variables. The stronger the correlation between the two variables, the closer the *absolute value* of r to 1.
Thus, if two variables are perfectly positively correlated, $r = 1$. If they are perfectly negatively correlated, $r = -1$ and if there totally no correlation between the variables, then $r = 0$.

2.4.3 The Adjusted R^2, $\bar{R}^2$

The problem with R^2 is that adding another independent variable to a particular equation *will never decrease* R^2. Hence, if you compare two models, say P and Q and model Q has one more independent variable, then Q will always have a better (or equal) fit as measured by R^2. Recall from (**??**) that

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

If RSS falls and TSS remains constant, then R^2 increases. The dependent variable has not changed so TSS stays the same. Since OLS ensures that adding a variable will not increase the summed squared residuals, RSS will only stay the same or fall.

The inclusion of another variable decreases the **degrees of freedom** because it requires the estimation of another coefficient. Degrees of freedom is described as the excess number of observations (N) over the number of coefficients (including the intercept, commonly known as β_0) ($K+1$). Mathematically, $df = N - K - 1$. In the weight/height example, adding another independent variable still keeps $N = 20$ but changes $K = 2$ to $K = 3$. Hence, the decreases the degrees of freedom from $(20-2=)18$ to $(20-3=)17$. This decrease has a cost as the lower the degrees of freedom, the less reliable the estimates are likely to be. Hence, the increase in the quality of the fit caused by the addition of a variable needs to be compared to the decrease in the degrees of freedom before a decision can be made with respect to the statistical impact of the added variable.

In essence, R^2 is of little help if we are trying to decide whether adding a variable to an equation improves our ability to meaningfully explain the dependent variable. We use $\bar{R}^2$ which is defined as the percentage of the variation of Y around its mean, $\bar{Y}$ that is explained by the regression equation, *adjusted for degrees of freedom.*

$$\bar{R}^2 = 1 - \frac{\frac{\sum e_i^2}{N-K-1}}{\frac{\sum (Y_i - \bar{Y})^2}{N-1}} \tag{36}$$

3 Learning to Use Regression Analysis

3.1 Estimating Single-Independent Variable Models with OLS

The respective steps in Applied Regression Analysis are:

3.1.1 Review the Literature; Develop the Theoratical Model

Start reviewing scholarly literature about the topic you are interested in doing. If the topic is new, figure out ways to transfer existing theories to current ones, or ask experts or professionals in the subject matter.

3.1.2 Specify the Model: Select the Independent Variables and Functional Form

Select the dependent variable, then based on that, **specify** the model with the independent variables, the functional form and the nature of the stochastic error term. A mistake in any of these elements is *specification error.* Researchers deciding on their independent variables are imposing their *priors* or working hypothesis on the regression equation.

Some variables are called **dummy** variables which take the value one or zero, depending on whether the condition is met.

3.1.3 Hypothesize the Expected Signs of the Coefficients

Putting signs above the variables indicate the hypothesized sign of the respective regression coefficient in a linear model.

3.1.4 Collect, Inspect & Clean the Data

A general rule in this stage is the more the observations, the better. In regression analysis, *all* the variables have the same number of observations and, in the case of a time series, the same frequency.

There should be as many observations as possible due to the statistical concept of *degrees of freedom.* The more degrees of freedom there are, the less likely that the stochastic component of the equation will affect inferences about the deterministic portion. While inspecting data, look for outliers. They could be due to data entry errors. However, the mere existence of an outlier does not justify dropping an observation from the sample. A regression neesd to be able to explain all the observations in the sample.

3.1.5 Estimate & Evaluate the Equation

Use computer software to easily perform this step. EViews, SPSS or Stata are computer packages used to do this.

3.1.6 Document the Results

State the functional form, the t-value, the value used to test the hypothesis that the true value of the coefficient is different from zero. The number of observations and the value of $\bar{R}^2$.

4 The Classical Model

4.1 The Classical Assumptions

The **Classical Assumptions** must be met in order for OLS estimators to be the best available. They are

1. The regression model is linear, correctly specified and has an additive error term
2. The error term has a zero population mean
3. All explanatory variables are uncorrelated with each other (no serial correlation)
4. Observations of the error term are not correlated 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 invoked)

An error term satisfying Assumptions I through V is called a **classical error term** and if Assumption VII is added, the error term is called a **classical normal error term**.

4.1.1 The regression model is linear, correctly specified and has an additive error term

. The model is assumed to be linear

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \tag{37}$$

The underlining theory does not need to be linear, though. An exponential function, for example

$$Y_i = e^{\beta_0} X_i^{\beta_1} e^{\epsilon_i} \tag{38}$$

can be transformed to

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \epsilon_i \tag{39}$$

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

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_i \tag{40}$$

In equation (**??**), the properties of the OLS estimator of the βs hold because the equation is still linear. Two additional properties must hold. First, we assume that the equation is properly specified. Second, we assume that a stochastic error term has been added to the equation. This error term *must* be additive and cannot be multiplied by or divided into any of the variables in the equation.

4.1.2 The error term has a zero-population mean

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. Classical Assumption II says that the mean of this distribution is zero. For a small sample, the mean is not likely to be zero but as the sample size approaches ∞, the mean approaches zero.

To compensate for the change that the mean of the population ϵ might not equal zero, the mean of ϵ is forced to be zero by the existence of the constant term in the equation. An example equation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{41}$$

might have the stochastic term, ϵ_i to have a mean of 3. Then, consider $\mathbb{E}(\epsilon - 3) = 0$. Add 3 to the constant term and subtract 3 from the error term to get

$$Y_i = (\beta_0 + 3) + \beta_1 X_i + (\epsilon_i - 3) \tag{42}$$

From (**??**), substitute $\beta_0^* = \beta_0 + 3$ and $\epsilon_i^* = \epsilon_i - 3$ to get

$$Y_i = (\beta_0^*) + \beta_1 X_i + (\epsilon_i^*) \tag{43}$$

and now (**??**) conforms to Classical Assumption II.

4.1.3 All explanatory variables are uncorrelated with the error term

It is assumed that the observed values of the explanatory variables are determined independently of the values of the error term. Explanatory variables

Xs are considered to be determined outside the context of the regression equation.

4.1.4 Observations of the error term are uncorrelated with each other

The observations of the error term are independent of each other. In some cases, though this assumption is unrealistic. For example, in time series models, this classical assumption states that the increase in the error term in one period due to a shock does not show up or affect, in any way, the error term in another time period. However it is unrealistic as the shock lasts a number of time periods. If, over *all* the observations, ϵ_{t+1} is correlated with ϵ_t, then this assumption is violated.

4.1.5 The error term has a constant variance

The variance of the distribution from which the observations of the error term are drawn is constant. In other words, 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. The violation of Classical Assumption V is called **heteroskedasticity**.

For example, suppose I am studying about public expenditure on education of 50 states. Because New York is larger than smaller states like Nevada, it is probable that the value of the stochastic error term is bigger than that of smaller states. The amount of unexplained variation in big states like New York is likely to be larger than that in small states like Nevada.

4.1.6 No explanatory variable is a perfect linear function of any other explanatory variable(s)

Perfect **collinearity** between two independent variables imply that they are really the same variable, or one is a multiple of the other, or a constant has been added to both variables. Because of this, the OLS estimator cannot distinguish one variable from another.

Perfect multicollinearity also can occur when two independent variables always sum to a third; or when one of the explanatory variables has a variance of zero. With perfect multicollinearity, the OLS computer program will be unable to estimate the coefficients of the collinear variables.

4.1.7 The error term is normally distributed

Assumption VII states that the observations of the error term are drawn from a normal distribution. This assumption is not necessary for OLS estimation but critical for **hypothesis testing**, which uses the estimated regression coefficient to investigate hypotheses about economic behaviour. Although Assumption VII is optional, it is usually advisible to add to the assumption for two reasons:

- The error term ϵ_i can be thought of as a sum of several minor errors. As the number of these minor errors get larger, the distribution of the error term ends to approach the normal distribution (by Central Limit Theorem).
- The t-statistic and F-statistic, are not truly applicable unless the error term is normall distributed.

4.2 The Sampling Distribution of $\hat{\beta}$

Just as the error term follows a probability distribution, so do the estimates of β. The probability distribution of these $\hat{\beta}$ values across different samples is called the **sampling distribution of** $\hat{\beta}$.

Recall that an *estimator* is a formula like the OLS formula but an *estimate* is the value of $\hat{\beta}$ for a given sample. Although many beginner econometricians assume that regression analysis can produce only one estimate of β for a given population, in reality each different sample from the same population will produce a different estimate of β. The collection of all the possible samples has a distribution with a *mean* and a *variance.*

Consider an example of a sampling distribution of $\hat{\beta}$. A researcher intends to build a regression model to explain the starting salaries of last year's graduates of a school as a function of GPAs

$$\begin{aligned} SALARY_i &= f(\overbrace{GPA_i}^{+}) \\ &= \beta_0 + \beta_1 GPA_i + \epsilon_i \end{aligned} \tag{44}$$

The researcher selects a sample of 25 students and estimates (**??**). Now he selects a second sample of students and does the same thing. He will definitely get a different value of $\hat{\beta}_1$ for both samples because the students drawn are different. Sometimes the value of β_1 is higher and sometimes it is lower. After collecting 5 samples of $\hat{\beta}_1$, he obtains the vector

$$\boldsymbol{\hat{\beta}_1} = \begin{pmatrix} 8612 & 8101 & 11355 & 6934 & 7994 \end{pmatrix}^I$$

The average of $\boldsymbol{\hat{\beta}_1}$ is 8599. For a "good" estimation technique, we want the mean of the sampling distribution of $\hat{\beta}$ to be equal to our true population of β, or *unbiasedness.* In this case, let the true value of β_1 be 8400.

4.2.1 Properties of the Mean

A desirable property of the distribution of the estimates is that the mean is equal to the true mean of the variable being estimated. An estimator which yields such an estimate is an **unbiased estimator**. In other words, an estimator, $\hat{\beta}$ is **unbiased** if the mean of its sampling distribution and its true value of β have the same value

$$\mathbb{E}(\hat{\beta}) = \beta \tag{45}$$

Only one value of $\hat{\beta}$ is obtained in practice but this property is useful because a single estimate drawn from an unbiased distribution is more likely to be near the true value than one not centred around the true value. If an estimator produces $\hat{\beta}$ which does not center around the true value of β then the estimator is referred to be a **biased estimator**. Usually, without any other information about the distribution of the estimates, we would always rather have an unbiased estimate than a biased one.

4.2.2 Properties of the Variance

We would like the distribution to be narrow too. For a β distribution with a small variance, the estimates are likely to be close to the mean of the sampling distribution. The variance of the distribution of β can be decreased by increasing the size of the sample (this also increases the degrees of freedom).

The element of chance, a random occurence is always present in estimating regression coefficients, and sometimes estimates may be far from the true value no matter how good the estimating technique. However, if the distribution is centered around the true value and has as small a variance as possible, the element of chance is less likely to induce a poor estimate. Of course, if the sampling distribution is centered around the wrong value then a lower variance implies that most of the sampling distribution of $\hat{\beta}$ is concentrated on the wrong value.

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** or MSE. The Mean Square Error is equal to the variance plus the square of the bias. The lower the MSE, the better.

Another item is that as the variance of the error term increases, so does the variance ofthe distribution of $\hat{\beta}$. The reason for the increased variance of $\hat{\beta}$ is that 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 value of Y_i.

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 we have mentioned.

4.3 The Gauss-Markov Theorem and the Properties of OLS Estimators

The **Gauss-Markov Theorem** states that given Classical Assumptions I through VI, the OLS estimator of β_k is the minimum variance estimator from all the linear unbiased estimators of $\beta_k \forall k = 0, 1, 2, \cdots, K$.

It is most easily remembered that OLS is "BLUE", where BLUE stands for Best Linear Unbiased Estimator, where 'best' here means minimum variance. If an equation's coefficient estimation is unbiased,then

$$\mathbb{E}(\hat{\beta_k}) = \beta_k \quad \forall k = 0, 1, 2, \cdots, K$$

An unbiased estimator with the smallest variance is said to be **efficient**, and that estimator is said to have the property of efficiency.

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

1. ***They are unbiased.*** In other words, $\mathbb{E}(\hat{\beta}) = \beta$. The OLS estimates of the coefficients are centered around the true population values of the parameters being estimated.

2. ***They are minimum variance.*** The distribution of the coefficient estimates around the true parameter is as narrowly distributed as possible for an unbiased distribution.

3. ***They are consistent.*** As the sample size approaches to ∞, the estimates converge to the true population parameters.

4. ***They are normally distributed.*** The $\hat{\beta}$s fulfil the property $\hat{\beta} \sim \text{N}(\beta, VAR[\hat{\beta}])$.

If the seven Classical Assumptions are met and if OLS is used to calculate all the $\hat{\beta}$s, then it can be stated that the estimated regression coefficient is an unbiased, minimum variance estimate of the impact on the dependent variable of a one-unit increase in a given independent variable, holding constant all other independent variables. This estimate is drawn from a distribution of estimates that is centered around the true population coefficient and has the smallest possible variance for such unbiased distributions.

4.4 Standard Econometric Notation

Population Parameter (True, unobserved values)

Regression Coefficient	β_k
Expected value of the estimated coefficient	$\mathbb{E}(\hat{\beta}_k)$
Variance of the error term	σ^2 or VAR(ϵ_i)
Standard deviation of the error term	σ
Variance of the estimated coefficient	$\sigma^2(\hat{\beta}_k)$ or VAR($\hat{\beta}_k$)
Error or Disturbance Term	ϵ

Estimate (Observed from sample)

Estimated Regression Coefficient	$\hat{\beta}_k$
Estimated Variance of the error term	s^2 or $\hat{\sigma}^2$
Standard error	s or SE
Estimated Variance of the estimated coefficient	$s^2(\hat{\beta}_k)$ or $\widehat{\text{VAR}}(\hat{\beta}_k)$
Residual	e_i

5 Hypothesis Testing

5.1 What is Hypothesis Testing?

The three topics that are central to hypothesis testing are

1. the specification of the hypothesis to be tested
2. the decision rule to use in deciding whether to rejected the hypothesis
3. the kinds of errors that might be encountered if the application of the decision rule to the appropriate statistics yields an incorrect reference

5.1.1 Classical Null and Alternative Hypothesis

The first step in hypothesis testing is to state the hypotheses to be tested. The **null hypothesis** typically is a statement of the values that the researcher does *not* expect. The notation used to specify the null hypothesis is "H_0:" followed by a statement of the range of values you *do not expect.* For example if you expect a positive coefficient, then you do not expect a zero or negative coefficient and so the null hypothesis is H_0: $\beta \leq 0$
The **alternative hypothesis** is a statement of the values that the researcher expects. The notation used to specify the alternative hypothesis is "H_A:" followed by a statement of the range of values you expect. To continue, the alternative hypothesis is H_A: $\beta > 0$.
Alternatively if you expect a negative coefficient then

$$\mathrm{H}_0\text{: } \beta \geq 0 \quad \mathrm{H}_\mathrm{A}\text{: } \beta < 0$$

This is for a **one-sided test** because the alternative hypotheses have values only on one side of the null hypothesis. Another approach is the **two-sided test** (or *two-tailed test*) where the alternative hypothesis has values on both sides of the null hypothesis. This is useful if the researcher does not expect β to be equal to zero but is unsure to expect a positive or negative value. For a two-sided test around zero, the null and alternative hypotheses are:

$$\mathrm{H}_0\text{: } \beta = 0 \quad \mathrm{H}_\mathrm{A}\text{: } \beta \neq 0$$

5.1.2 Type I and Type II Errors

Recall that the regression coefficients are only estimates of the true population parameters. Hence, it would be unrealistic to think that conclusions drawn from regression analysis will always be right. There are two kinds of

errors that we make in such hypothesis testing, namely **Type I** errors where we reject a true null hypothesis and **Type II** errors where we do not reject a false null hypothesis.
Suppose we have the null and alternative hypothesis as H_0: $\beta \leq 0$ and H_A: $\beta > 0$. There are two distinct possibilities.

- The first is that the true β in the population is less than 0, as specified by the null hypothesis H_0. When the true β is not positive, unbiased estimates of β will be distributed around 0 or some negative number, but any given estimate is very unlikely to be exactly equal to that number. Any single sample (and therefore any estimate of β calculated from that sample) might be quite different from the mean of the distribution. As a result, even if the true parameter β is not positive, the particular estimate obtained by a researcher may be sufficiently positive to lead to the rejection of the null hypothesis that $\beta \leq 0$. This is a Type I Error – *We have rejected the truth!*

- The second is that that true β is greater than 0, as specified in the alternative hypothesis H_A. Depending on the specific value of the population β, it is possible to obtain an estimate of β that is close enough to zero (or negative) to be considered "not significantly positive". This occurs because the sampling distribution of β, even if unbiased, has a portion of its area in the region $\beta \leq 0$. Such a result may lead the researcher to not reject the null hypothesis $\beta \leq 0$ when, *in truth*, $\beta > 0$. This is a Type II Error – *We have failed to reject a false null hypothesis!*

As an example of Type I and Type II Errors, consider a jury in a murder case. In such a situation, the presumption "*innocent until proven guilty*" implies that:

- H_0: The defendent is innocent and
- H_A: The defendent is guilty

The associated Type I and Type II errors would mean

- **Type I Error**: Falsely rejecting H_0 or sending the innocent defendant to jail
- **Type II Error**: Falsely not rejecting H_A or setting free the guilty (specifically, not being able to prove the defendant's wrongdoing)

In the real world, decreasing the probability of Type I Error would mean to increase the probability of a Type II Error.

5.1.3 Decision Rules of Hypothesis Testing

In testing a hypothesis, a sample statistic must be calculated that determines when the null hypothesis can be rejected depending on the magnitude of the sample statistic compared with a preselected *critical value* . This procedure is referred to as a **decision rule**.

A decision rule is formulated before regression estimates are obtained. The range of possible values of $\hat{\beta}$ is divided into two regions, a *rejection* region and an "*acceptance*" region. The terms are expressed *relative to the null hypothesis.* The **critical value** is a value that divides the "acceptance" region from the rejection region when testing a null hypothesis.

To use a decision rule, we need to select a critical value. Consider the one-sided test H_0: $\beta \leq 0$ and H_{A}: $\beta > 0$ and let the critical value be 1.8. If the observed $\hat{\beta}$ is greater than 1.8, then we reject the null hypothesis because the value falls into the rejection region. Similarly, any observed value, let's say $\hat{\beta}\prime$ where $\hat{\beta}\prime < 1.8$ can be seen to fall in the "acceptance" region. We can also see that the rejection region is the probability that the null hypothesis is rejected if it is, in fact true. This also means that the Type I Error takes this definition. Formally, the Type I Error is the probability of rejecting the null hypothesis, when it is in fact true.

5.2 The *t*-Test

The t-test is usually used to test hypotheses about individual regression slope coefficients. It is easy to use and it is appropriate when the stochastic error term is normally distributed and when the variance of that distribution must be estimated.

5.2.1 The *t*-Statistic

Fore a typical multiple regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{46}$$

we can calculate t-values for each of the estimated coefficients in the equation. t-tests are usually done only on slope coefficients and for these, the relevant form of the ***t*-statistic** for the $k-th$ coefficient is

$$t_k = \frac{\hat{\beta}_k - \beta_{H_0}}{\mathrm{SE}(\hat{\beta}_k)} \qquad \forall \quad k = 1, 2, \cdots, K \tag{47}$$

where
$\hat{\beta_k}$ is the estimated value of the regression coefficient of the k-th variable
β_{H_0} is the critical value implied by the null hypothesis of the k-th variable
SE($\hat{\beta_k}$) is the estimated standard error of $\hat{\beta_k}$ Since most regression hypotheses test whether a particular regression coefficient is significantly from 0, β_{H_0} is typically 0 and the most used form of the t-statistic becomes

$$t_k = \frac{\hat{\beta_k} - 0}{\text{SE}(\hat{\beta_k})} \qquad \forall \quad k = 1, 2, \cdots, K$$

which simplifies to

$$t_k = \frac{\hat{\beta_k}}{\text{SE}(\hat{\beta_k})} \qquad \forall \quad k = 1, 2, \cdots, K \tag{48}$$

or the estimated coefficient divided by the standard error.

Consider the example at Woody's restaurants. Woody's wants to open a new chain and is evaluating the suitability of a location. They have hypothesized that sales depend on 3 factors

- N = Competition: the number of competitors in a 2-mile radius of the Woody's location
- P = Population: the number of people living in a 3-mile radius of the Woody's location
- I = Income: the average household income of the population measured in P

The equation for the check volume is

$$\begin{aligned} \hat{Y_i} &= 102192 - 9075N_i + 0.3547P_i + 1.288I_i \\ & \qquad\qquad\quad (2053) \quad (0.0727) \quad (0.543) \\ t &= \qquad\qquad -4.42 \qquad 4.88 \qquad 2.37 \\ N &= 33 \qquad \bar{R}^2 = 0.579 \end{aligned} \tag{49}$$

The numbers in the parentheses are the estimated standard errors of the estimated $\hat{\beta}$, and the numbers below them are t-values calculated according to (**??**). Note that the sign of the t-value is always the same as that of the estimated regression coefficient and the standard error is always positive.

Using the regression results in (**??**), the t-value for the estimated coefficient for P Given that the estimated regression coefficient is 0.3547 and the standard error is 0.0727, and given that H_0: $\beta_P \leq 0$, then the relevant t-value is indeed 4.88 as specified in (**??**)

$$t_P = \frac{\hat{\beta_P}}{\text{SE}(\hat{\beta_P})} = \frac{0.3547}{0.0727} = 4.88$$

The larger the absolute value of the t-value, the greater the likelihood that the estimated regression coefficient is significantly different from zero.

5.2.2 The Critical t-value and the t-Test Decision Rule

The decide whether to reject the null hypothesis based on a calculated t-value, we use a critical t-value. A **critical t-value** is the value that distinguishes the "acceptance" region from the "rejection" region. The critical t-value, t_C, is selected from a t-table depending on whether it is one-sided or two-sided, the level of Type I Error specified and the degrees of freedom $N-K-1$. The level of Type I Error is also known as the *level of significance* of that test.

Once a t-value, t_k and a critical t-value, t_C has been calculated, we reject the null hypothesis if $|t_k| > t_C$ AND the t_k has the sign implied by H_A. In short,

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 works for calculated t-values and critical t-values for one-sided hypotheses around zero:

$$\text{H}_0\text{: } \beta_k \leq 0 \text{ and } \text{H}_\text{A}\text{: } \beta_k > 0$$
$$\text{H}_0\text{: } \beta_k \geq 0 \text{ and } \text{H}_\text{A}\text{: } \beta_k < 0$$

or for two-sided hypotheses around zero:

$$\text{H}_0\text{: } \beta_k = 0 \text{ and } \text{H}_\text{A}\text{: } \beta_k \neq 0$$

or the same idea, around values other than zero. The one-sided test is

$$\mathrm{H}_0\text{: } \beta_k \leq S \text{ and } \mathrm{H}_\mathrm{A}\text{: } \beta_k > S$$
$$\mathrm{H}_0\text{: } \beta_k \geq S \text{ and } \mathrm{H}_\mathrm{A}\text{: } \beta_k < S$$

or for two-sided hypotheses around zero:

$$\mathrm{H}_0\text{: } \beta_k = S \text{ and } \mathrm{H}_\mathrm{A}\text{: } \beta_k \neq S$$

The decision rule fo all the above is the same. Reject the null hypothesis if $|t_k| > t_C$ and has the same sign as the coefficient implied by H_A.

Going back to the Woody's Restaurant example in Section 5.2.1, recall that we hypothesized that the coefficient would be positive, so it would be a one-sided test H_0: $\beta_P \leq 0$ and H_A: $\beta_P > 0$. This is a one-sided test and there are 29 degrees of freedom. Also, the level of significance is 5%. Hence, the critical value is 1.699. The decision rule becomes to reject H_0 if the t-value is greater than 1.699. Since, earlier we calculated that t_P is +4.88, we can reject the null hypothesis and conclude that the population does have a positive relationship with Woody's sales, holding all other variables in the equation constant.

5.2.3 Choosing a Level of Significance

It was necessary to pick a statistic level of significance before a critical t-value could be found. The phrase "statistically positive" would mean 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.

Beginning econometricians assume that the lower the level of significance the better. However an extremely low level of significance also dramatically increases the probability of making a Type II Error. Instead, we use a 5% level of significance except in circumstances where you know something unusual about the relative costs of making Type I or Type II errors. For example, if making a Type II Error is extremely costly then it would be reasonable to revise the level of significance up to 10%.

If we can reject a null hypothesis at the 5% level of significance we can say that the coefficient is "statistically significant" at the 5% level. Since the 5% level is arbitrary, we should not jump to conclusions about the value of a variable simply because its coefficients misses being significant by a small amount; if a different level of significance had been chosen the result might be different.

5.2.4 Confidence Intervals

A **confidence interval** is a range which contains the true value of an item a specific percentage of the time. 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 coefficent:

$$\text{Confidence Interval} = \hat{\beta} \pm t_C \cdot \text{SE}(\hat{\beta}) \tag{50}$$

Going back to Equation (**??**)

$$\begin{aligned}
\hat{Y_i} &= 102192 - 9075N_i + 0.3547P_i + 1.288I_i \\
& \qquad\qquad\quad (2053) \quad (0.0727) \quad (0.543) \\
t &= \qquad\qquad -4.42 \qquad 4.88 \qquad 2.37 \\
N &= 33 \qquad \bar{R^2} = 0.579
\end{aligned}$$

A 90% confidence interval for $\hat{\beta_P}$ would be $0.3547 \pm 1.699 \cdot 0.0727$ or 0.3547 ± 0.1235. In other words, we are confident 90% of the time that the true coefficient will fall between 0.2312 and 0.4782 or, Mathematically,
$0.2312 \leq \hat{\beta_P} \leq 0.4782$.

If a hypothesized border value, β_{H_0}, falls within the 90% confidence interval for an estimated coefficient then we are not able to reject the null hypothesis at the 10% level of significance in a two-sided test. On the other hand, if β_{H_0} falls outside the 90% confidence interval the we can reject the null hypothesis.

5.2.5 *p*-Values

An alternative approach to the t-test is the p-test based on a measure called the p-value, or *marginal significance level.* a ***p*-value** for a t-score is the probability of observing a t-score that is big or bigger if the null hypothesis were true.

A p-value is a probability, so $0 \leq p\text{-value} \leq 1$. It tells us the lowest level of significance at which we could reject the null hypothesis. A small p-value casts doubt on the null hypothesis, so to reject the null hypothesis, we need a small p-value.

To use a p-value to run a t-test, let your level of significance be 0.05. Reject the null hypothesis as long as the p-value is lower than 0.05 and the sign of $\hat{\beta}_k$ is in the direction as $\mathrm{H_A}$. Thus, the p-value decision rule is

> Reject $\mathrm{H_0}$ if p-value$_k$ < level of significance **AND** if $\hat{\beta}_k$ also has the sign implied by $\mathrm{H_A}$. Do not reject $\mathrm{H_0}$ otherwise.

Back to the Woody's restaurant example, if we run a one-sided test on the coefficient of I, the null and alternative hypotheses are $\mathrm{H_0}$: $\beta_I \leq 0$ and $\mathrm{H_A}$: $\beta_I > 0$ The p-value for $\hat{\beta}_I$ is 0.0246. This value is the two-sided p-value so we need to divide by 2 to get the actual p-vlue, 0.0123. Since 0.0123 is lower than our chosen level of 0.05 and the sign of $\hat{\beta}_I$ agrees with $\mathrm{H_A}$, we can reject $\mathrm{H_0}$ in favour of the alternative hypothesis.

5.3 Examples of *t*-Tests

Consider a simple model of aggregate retail sales of new cars that hypothesizes the sales of new cars (Y) are a function of real disposable income (X_1) and the average retail price of a new car adjusted by the consumer price index (X_2). After reviewing literature you have decided to add a third independent variable, the number of sports utility vehicles sold (X_3). The following model is hypothesized:

$$Y = f(\overbrace{X_1}^{+}, \overbrace{X_2}^{-}, \overbrace{X_3}^{-}) + \epsilon \tag{51}$$

With (**??**), β_1 is expected to be positive while β_2 and β_3 are expected to be negative. Based on these values, the respective null and alternative hypotheses are:

$$\begin{aligned} \mathrm{H_0}&: \beta_{X_1} \leq 0 \\ \mathrm{H_A}&: \beta_{X_1} > 0 \\ \\ \mathrm{H_0}&: \beta_{X_2} \geq 0 \\ \mathrm{H_A}&: \beta_{X_2} < 0 \\ \\ \mathrm{H_0}&: \beta_{X_3} \geq 0 \\ \mathrm{H_A} \ &: \beta_{X_3} < 0 \end{aligned}$$

Consider that there are 10 observations and you are picking the 5% level of significance. The number of degrees of freedom is $10 - 3 - 1 = 6$. For all 3 variables, $t_C = 1.943$.

After running the regression, the OLS computer package produces:

$$\begin{array}{rlll} \hat{Y}_t = 1.30 + & 4.91X_{1t} & + 0.00123X_{2t} & - 7.14X_{3t} \\ & (2.38) & (0.00022) & (71.38) \\ t = & -2.1 & 5.6 & -0.1 \end{array} \tag{52}$$

where
Y is new car sales in ('00,000) units in year t
X_1 is real U.S. disposable income in \$('000,000,000) in year t
X_2 is average retail price of a new car in in year t
X_3 is the number of sports utility vehicles sold in year t in ('000,000)

By applying the decision rule based on the hypotheses that was set up,

1. reject H_0 for β_1 if the t-value is greater than 1.943. Hence, reject H_0 in favour of $\text{H}_\text{A}$for X_1.

2. reject H_0 for β_2 if the t-value is smaller than -1.943. Hence, do not reject H_0 for X_2. In this case, the absolute value exceeds the critical value but the sign does not agree.

3. reject H_0 for β_3 if the t-value is smaller than -1.943. Hence, do not reject H_0 for X_3.

5.3.1 Examples of Two-Sided t-Tests

The two-sided tests are used to test if

- if an estimated coefficient is significantly different from zero
- if an estimated coefficient is significantly different from a specific nonzero value

Consider back the Woody's Restaurant example. We suspect that the impact of the average income on an area on the total customers going out to dinner is ambiguous. High incomes might mean Woody's is an inferior good and they dine at another restaurant. To do so, run a two-sided test around zero to determine whether or not the estimated coefficient of income is significantly different from zero in *either* direction. The hypothesis set up is H_0: $\beta_I = 0$ and H_A: $\beta_I \neq 0$

Keeping the level of significance to be at 5%, the critical values are now 2.045 and -2.045. That means, reject the null hypothesis if the t-value is outside this range. The calculated t-value is 2.37. Since it is outside the "acceptance" region, we reject the null hypothesis in favour of the alternative hypothesis. Because of the *positive* sign, we say that the higher the income of the population, the more check volume Woody's will get.

5.4 The *F*-Test

Although the t-test is invaluable for hypotheses about individual regression coefficients, it cannot be used to test multiple hypotheses *simultaneously.* For example, say you want to test the null hypothesis that there is no seasonal variation in a quarterly regression equation that has dummy variables for the seasons. For such hypotheses, most researchers will use the F-test.

5.4.1 What is the *F*-Test?

The F-test is a formal hypothesis test that is designed to deal with a null hypothesis which contains multiple hypotheses or a single hypothesis about a group of coefficients. Such "joint" or "compound" null hypotheses are appropriate whenever the underline economic theory specifies values for multiple coefficients simultaneously.

First translate the particular null hypothesis to constraints that will be placed in the equation. The resulting constrained equation can be though of as what the equation would look like if the null hypothesis were correct – you substitute the hypothesized values into the regression equation in order to see what would happen if the equation was constrained to agree with the null hypothesis. As a result, in the F-test the null hypothesis always leads to a constrained equation even if this violates our standard practice that the alternative hypothesis contains that we expect is true.

Second, estimate this constrained equation with OLS and compare the fit of this constrained equation with the fit of the un-constrained equation. If 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 unconstrained equation, 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-statistic where

$$F = \frac{(\mathrm{RSS}_M - \mathrm{RSS})/M}{\mathrm{RSS}/N-K-1} \tag{53}$$

where
RSS is the residual sum of squares from the unconstrained equation
RSS_M is the residual sum of squares from the constrained equation
M is the number of constraints placed on the equation
$(N - K - 1)$ is the 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. If the unconstrained regression yields exactly the same estimated coefficients as does the constrained regression, the RSS are equal and the F-statistic is zero. In this case, H_0 is not rejected because the data indicate that the constraints appear to be correct. As the difference between the constrained coefficients and the unconstrained coefficients increases, the data indicate that the null hypothesis is less likely to be true. Thus, when F gets larger than the critical F-value, the hypothesized restrictions specified in the null hypothesis are rejected by the test.

The decision rule to use in the F-test is to reject the null hypothesis if the calculated F-value is greater than the appropriate critical F-value, F_C.

Reject $\mathbf{H}_0$ if $F \geq F_C$.

Do not reject $\mathbf{H}_0$ if $F < F_C$

The critical F-value, F_C is determined from statistical tables depending 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: the degrees of freedom for the numerator of (**??**) due to M and the degrees of freedome for the denominator due to $(N-K-1)$. The underlying principle here is that the if the calculated F-value is greater than the critical value then the estimated equation's fit is significantly better than the constrained equation's fit, and we can reject the null hypotheiss of no effect.

5.4.2 The F-Test of Overall Significance

Although R^2 and $\bar{R}^2$ measure the overall degree of fit of an equation, they don't 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. For an equation with K independent variables, this means that the null and alternative hypothesis would be

$$\begin{aligned}&\mathrm{H_0}\colon\ \beta_1 = \beta_2 = \beta_3 = \cdots = 0\\ &\mathrm{H_A}\colon\ \mathrm{H_0} \text{ is not true}\end{aligned}$$

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, (**??**) simplifies to

$$F = \frac{\mathrm{ESS}/K}{\mathrm{RSS}/N-K-1} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i e_i^2/(N-K-1)} \tag{54}$$

This is the ratio of the explained sum of squares (ESS) to the residual sum of squares (RSS), adjusted for the number of independent variables, K, and the number of observations in the sample, N. In this case, the "constrained" equation to which we are comparing the overall fit is :

$$Y_i = \beta_0 + \epsilon_i \tag{55}$$

which is nothing more than saying that $\hat{\beta}_i = \bar{Y}$. Thus the F-test of overall significance is really testing the null hypothesis that the fit of the equation isn't significantly better than that provided by using the mean alone.

Let us continue to test the overall significance of the Woody's restaurant model. Since there are three independent variables, the null and alternative hypotheses are

$$\begin{aligned}&\mathrm{H_0}\colon\ \beta_N = \beta_P = \beta_I = 0\\ &\mathrm{H_A}\colon\ \mathrm{H_0} \text{ is not true}\end{aligned}$$

To decide whether to reject or not reject this null hypothesis, we need to calculate (**??**) for the Woody's example. There are three constraints in the null hypothesis so $K = 3$. Given, from the EViews example that $N = 33$ and RSS$= 6,130,000,000$. It can be also calculated that ESS$= 9,929,450,000$. Thus, the appropriate F-ratio is

$$F = \frac{\text{ESS}/K}{\text{RSS}/N-K-1} = \frac{9{,}929{,}450{,}000/3}{6{,}130{,}000{,}000/29} = 15.65 \tag{56}$$

Our decision rule tells us to reject the null hypothesis if the calculated F-value is greater than the critical F-value. To determine that critical F-value, we need to know the level of significance and the degrees of freedom. Taking 5% level of significance and, taking degrees of freedom defined with $K = 3$ and $N = 29$, then the critical F-value, F_C, is 2.93 (verified by the F-table). Since the calculated F-ratio is higher than F_C, reject the null hypothesis and conclude that the Woody's equation does indeed have a significant overall fit.

Two final comments of the F-test. First, if there is only one independent variable, then an F-test and a t-test of whether the slope coefficient equals zero will always produce the same answer. This property does not hold if there ar two or more independent variables. In such a situation, an F-test could determine that the coefficients *jointly* are not significantly different from zero even though a t-test on one of the coefficients might show that *individually* it is significantly different from zero. The vice versa holds too.

Second, 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. Most standard regression estimation programs report both the F-value and the p-value associated with this test.

5.4.3 Other Uses of the *F*-Test

Besides the test of overall significance, there are many other uses of the F-test. Consider the Cobb-Douglas production function. Recall that the general Cobb-Douglas production function is $Q = AL^{\lambda}K^{\mu}$. Now, consider the production function

$$Q_t = \beta_0 + \beta_1 L_t + \beta_2 K_t + \epsilon_t \tag{57}$$

where
Q_t is the natural log of total output in the USA in year t
L_t is the natural log of labour input in the USA in year t
K_t is the natural log of capital input in the USA in year t
$/epsilon_t$ is a well-behaved stochastic term

Since that it can be shown that a Cobb-Douglas production function with constant returns to scale is one where β_1 and β_2 add up to exactly one, so the null hypothesis to be tested is

$$\begin{aligned} \mathrm{H_0}&: \beta_1 + \beta_2 = 1 \\ \mathrm{H_A}&: \text{otherwise} \end{aligned}$$

To test this null hypothesis with the F-test, we run regressions on the unconstrained (**??**) and and equation that is constrained to conform to the null hypothesis. To create such a constrained equation, we solve the null hypothesis for β_2 and substitute into (**??**) to form

$$\begin{aligned} Q_t &= \beta_0 + \beta_1 L_t + (1 - \beta_1) K_t + \epsilon_t \\ Q_t &= \beta_0 + \beta_1 (L_t - K_t) + K_t + \epsilon_t \end{aligned} \tag{58}$$

and rearranging K_t to the left side of the equation yields the constrained equation

$$(Q_t - K_t) = \beta_0 + \beta_1 (L_t - K_t) + \epsilon_t \tag{59}$$

Equation (**??**) will be the equation that will hold if our null hypothesis were correct.

To run an F-test on our null hypothesis of constant returns to scale, we need to run regressions on the constrained (**??**) and the unconstrained (**??**) and compare the fits of the two equations with the F-ratio from (**??**). If we use annual US data, we obtain an unconstrained equation of

$$\begin{aligned} \hat{Q}_t &= -38.08 - 1.28 L_t + 0.72 K_t \\ &\qquad\qquad\quad (0.30) \qquad (0.05) \\ t &= \qquad\qquad\ \ 4.24 \qquad 13.29 \\ N &= 24 \qquad \bar{R}^2 = 0.997 \qquad F = 4118.9 \end{aligned} \tag{60}$$

If we run the constrained equation and substitute the appropriate RSS into (**??**), we obtain $F = 16.26$. When this F is compared to a 5% critical F-value of (with $N = 21$ and $M = 1$) only F_C =4.32. The degrees of freedom in the numerator is only 1 because one coefficient has been eliminated from the equation by the constraint, the value is much higher so we reject the null hypothesis that constant returns to scale characterize the US economy. Interestingly, the sum of the estimated regression coefficients $\beta_1 + \beta_2 = 2.00$

so there is a drastic increase returns to scale. However since $\beta_1 > 1$ and the slope coefficient must be between 0 and 1, caution should be taken before we comfortably reach this conclusion.

A different use of the F-test involves testing null hypotheses that apply to various subsets of the coefficients in the equation. Consider a problem of testing the significance of seasonal dummies. **Seasonal dummies** are dummy variables that are used to account for seasonal variation in the data in time-series models. In a quarterly model, if

$$X_{1t} = \begin{cases} 1 \text{ in quarter } 1 \\ 0 \text{ otherwise} \end{cases}$$

$$X_{2t} = \begin{cases} 1 \text{ in quarter } 2 \\ 0 \text{ otherwise} \end{cases}$$

$$X_{3t} = \begin{cases} 1 \text{ in quarter } 3 \\ 0 \text{ otherwise} \end{cases}$$

then

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \epsilon_t \tag{61}$$

where X_4 is a non-dummy independent variable and t is quarterly. 3 dummy variables are required to represent four seasons (simply because the last season is in play if $X_{1t} = X_{2t} = X_{3t} = 0$. In this formulation β_1 shows the extent to which the expected value of Y in the first quarter differ from its expected value in the fourth quarter, the omitted condition. The other variables β_2 and β_3 can be interpreted similarly.

Inclusion of a set of seasonal dummies "deseaonalizes" Y and andy independent variables that are not seasonally adjusted. This procedure may be used as Y and X_4 are not "seasonally adjusted" prior to estimation. In this case, the null hypothesis is that there is *no* seasonality

$$\text{H}_0\text{: } \beta_1 = \beta_2 = \beta_3 = 0$$
$$\text{H}_\text{A}\text{: H}_1 \text{ is not true}$$

The constrained equation would then be $Y = \beta_0 + \beta_4 X_4 + \epsilon$. To determine whether the whole set of seasonal dummies should be included, the fit of the estimated constrained equation would be compared to the fit of the estimated unconstrained equation by using the F-test.

6 Specification: Choosing the Independent Variables

6.1 Omitted Variables

Say you forget to include one of the relevant independent variables when you first specify an equation or you cannot get data for one of the variables that you *do* think of. These are examples of situations with an **omitted variable**, defined as an important explanatory variable that has been left out of a regression equation.

The bias caused by leaving a variable out of an equation is called **omitted variable bias** (or, more generally, **specification bias**). In an equation, the coefficient β_k is the impact of change in the dependent variable Y given a one unit increase in the independent variable X_k, holding constant all the other independent variables. If a variable is omitted, then it is not included as an independent variable, and it is not held constant for the calculation and interpretation of $\hat{\beta_k}$. The omission can cause bias: it can force the expected value of the coefficient away from the true value of the population coefficient.

6.1.1 The Consequences of an Omitted Variable

The major consequence of omitting a relevant independent variable from an equation is to cause bias in the regression coefficients that remain in the equation. Say the true regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{62}$$

where ϵ_i is a classical error term. If you omit X_2 from the equation then it becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i^* \tag{63}$$

where

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i} \tag{64}$$

because the stochastic term includes the effects of any omitted variables. From (**??**) and (**??**), it might seem as though we could get unbiased estimates of β_0 and β_1 even if we left X_2 out of the equation. Unfortunately, this is not the case because the included coefficients almost surely pick up some of the effect of the omitted variable and therefore will change, causing bias.

If we leave an important variable out of an equation, we violate Classical Assumption III which says that the explanatory variables are independent of the error term. Unless the omitted variable is uncorrelated with **all** the included independent variables. In general, when there is a violation of one of the Classical Assumptions, the Gauss-Markov Theorem does not hold, and the OLS estimates are not BLUE. They are not unbiased, or minimum variance, or both.

An omitted variable causes Classical Assumption III to be violated in a way that causes bias. Estimating (**??**) when (**??**) is the truth will cause bias. This means that

$$\mathbb{E}(\hat{\beta_1}) \neq \beta_1 \tag{65}$$

Instead of having an expected value equal to the true β_1, the estimate will compensate for the fact that X_2 is omitted from the equation, then the OLS estimation procedure will attribute to X_1 variations in Y actually caused by X_2, and a biased estimate will result.

To see this, consider a production function that states that output, Y depends on the amount of labour, X_1 and capital, X_2. What would happen if data on capital were unavailable for some reason and X_2 was omitted from the equation? We leave the impact of capital on the model. This omission would almost surely biases the estimate of the coefficient of labour because it is likely that capital and labour are positively correlated. The OLS program would attribute to labour the increase in output actually caused by capital to the extent that labour and capital were correlated. Thus the bias would be a function of the *impact of capital on output and the correlation between capital and labour.*

To generalize a model with two independent variables, the expected value of the coefficient of an included variable, X_1, when a *relevant* variable, X_2 is omitted from the equation equals

$$\mathbb{E}(\beta_1) = \beta_1 + \beta_2 \cdot \alpha_1 \tag{66}$$

where α_1 is the slope coefficient of the secondary regression that relates X_2 to X_1:

$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + u_i \tag{67}$$

where u_i is a classical error term. α_1 can be expressed as a function of the correlation between X_1 and X_2, the included and excluded variables, or $f(r_{12})$. Equation (**??**) states that the expected value of the included variables coefficient is equal to its true value plus the omitted variable's true coefficient times a function of the correlation between the included (in) and the omitted (om) variables. Since the expected value of an unbiased estimate is the true value, the right hand term measures the omitted variable bias in the equation:

$$\text{Bias} = \beta_2\alpha_1 \quad \text{or} \quad \text{Bias} = \beta_{\text{om}} \cdot f(r_{\text{in, om}}) \tag{68}$$

In general terms, the bias thus equals β_{om}, the coefficient of the omitted variable, times $f(r_{\text{in, om}})$, a function of the correlation between the included and omitted variables. The bias exists unless:

- the true coefficient equals zero
- the included and omitted variables are not correlated.

The term $\beta_{\text{om}} f(r_{\text{in, om}})$ is the amount of specification biased introduced to the estimate of the coefficient of the included variable by leaving out the omitted variable.

6.1.2 An Example of Specification Bias

Consider the following equation for the annual consumption of chicken in the US.

$$\begin{aligned}
\hat{Y}_t &= 27.6 - 0.61PC_t + 0.09PB_t + 0.24YD_t \\
& \qquad\qquad (0.16) \qquad (0.04) \qquad (0.011) \\
t &= \qquad\quad -3.86 \qquad +2.31 \qquad +22.07 \\
\bar{R^2} &= 0.990 \qquad N = 40 \text{ (Annual 1960 -1999)}
\end{aligned} \tag{69}$$

where
Y_t is the per capita consumption (in pounds) in year t
PC_t is the price of chicken (in cents per pound) in year t
PB_t is the price of beef (in cents per pound) in year t
YD_t is the per capita disposable income (in hundreds of dollars) in year t

This is a simple demand for chicken model that includes the price of the good, a close substitute and income variable. If we estimate the equation without the price of the substitute, we get

$$\begin{aligned}
\hat{Y}_t &= 27.5 - 0.42PC_t + 0.27YD_t \\
& \qquad\qquad (0.14) \qquad (0.005) \\
t &= \qquad\quad -2.95 \qquad +55.00 \\
\bar{R^2} &= 0.988 \qquad N = 40 \text{ (Annual 1960 -1999)}
\end{aligned} \tag{70}$$

Compare equations (**??**) and (**??**) to see if dropping the beef price variable had an impact on the estimated equations. Comparing the overall fit, $\bar{R^2}$ fell from 0.990 to 0.988 (0.002 points) when PB was dropped, exactly what we'd expect when a relevant variable was omitted.

Dropping PB caused $\hat{\beta}_{PC}$ to become more positive, from -0.61 to -0.42. Similarly, a shift in the same direction was observed for $\hat{\beta}_{YD}$ from 0.24 to 0.27. The direction of this bias, by this way, is considered *positive* because the biased coefficient of PC of -0.42 is more positive than the suspected unbiased one with -0.61 and biased coefficient of YD is more positive than the suspected unbiased one of 0.24.

The fact that the bias is positive could have been guessed before any regressions were run if (**??**) were used. The specification bias by omitting PB is expected to be positive because the expected sign of the coefficient of

PB is positive and the expected correlation between the price of beef and chicken are positive.

$$\text{Expected bias in } \hat{\beta}_{PC} = \beta_{PB} \cdot f(r_{PC,PB}) = (+) \cdot (+) = (+) \quad \text{and}$$
$$\text{Expected bias in } \hat{\beta}_{PB} = \beta_{YD} \cdot f(r_{PB,YD}) = (+) \cdot (+) = (+)$$

Both correlation coefficients are anticipated to be positive. Using economic theory, an increase in the price of chicken will result in an increase in the price of beef. An increase in income also increases the price of beef.

To sum, if a relevant variable is left out of a regression equation,

- there is no longer an estimate of the coefficient of that variable in the equation
- the coefficients of the remaining variables are likely to be biased

6.1.3 Correcting for an Omitted Variable

The solution to a specification error seems easy – add the omitted variable to the equation. However, it is easier said than done.

First, omitted variable bias is hard to detect. The amount of bias might be small and hard to detect. This is especially true when there is no reason to believe that you have mis-specified the model. Some indications of specification bias are obvious but some are not so. The best indicators for an omitted relevant variable are the theoretical underpinnings of the model itself. The best way to avoid omitting an important variable is to think through the equation and model before entering anything into the computer.

Second, there is the problem of choosing which variable to add to an equation once you have decided that it is suffering from omitted variable bias. Some beginning researchers will add all the possible relevant variables to the equation at once but this leads to less precise estimates. Others will test a number of different variables and keep the one in the equation that does the best statistical job of appearing to reduce the bias. This technique is invalid because the variable that best corrects a case of specification bias might only do so only by chance rather than by being the true solution to the problem. It might give superb statistical results but does not describe the characteristics of the true population.

If the estimated coefficient is significantly different from our expectations in sign or magnitude, then it is extremely likely that some sort of specification bias exists in our model. If an unexpected result leads you to believe that you have an omitted variable, one way to decide which variable to add to the equation is to use expected bias analysis. **Expected bias** is the likely bias that omitting a particular variable would have caused in the estimated coefficient of one of the included variables. It can be estimated with (**??**)

$$\text{Bias} = \beta_{\text{om}} \cdot f(r_{\text{in, om}})$$

If the sign of the expected bias is the same as the sign of your unexpected result then the variable might be the source of the apparent bias. If the sign of the expected bias is *not* the same as the sign of your unexpected result then the variable is extremely unlikely to have caused your unexpected result. Expected bias analysis should only be used when you are choosing between theoretically sound potential variables.

As an example, return to (**??**). Assume you expect the coefficient of β_{PC} to be in the range of -1.0 and that you were surprised by the unexpected positive coefficient of PC in (**??**). This unexpectedly positive result could have been caused by an omitted variable with positive expected bias. One such variable is the price of beef. The expected bias in $\hat{\beta}_{PC}$ due to leaving out PB is positive since both the expected coefficient of PB and the expected correlation between PC and PB are positive. Hence, the price of beef is a reasonable candidate to be an omitted variable in (**??**).

6.2 Irrelevant Variables

When you include a variable in an equation that does not belong there, you are adding **irrelevant variables**. This is the converse of omitted variables and can be analysed using the model in Section 1. The addition of a variable to an equation where it does not belong does not cause bias but it increases the variances of the estimated coefficients of the included variables.

6.2.1 Impact of Irrelevant Variables

If the true regression coefficient is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \tag{71}$$

but the researcher for some reason includes and extra variable,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**} \tag{72}$$

where the misspecified error term is

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i} \tag{73}$$

Such a mistake would not cause bias if the true coefficient of the extra variable is zero. That is, $\hat{\beta}_1$ in (**??**) is unbiased if $\beta_2 = 0$. However, the inclusion of an irrelevant variable will increase the variance of the estimated coefficients, and this increased variance will tend to decrease the absolute value of the t-scores. An irrelevant variable will also decrease the $\bar{R}^2$.

6.2.2 An Example of an Irrelevant Variable

Return to the annual consumption of chicken example. Recall that

$$\begin{aligned} \hat{Y}_t &= 27.6 - 0.61PC_t + 0.09PB_t + 0.24YD_t \\ &\qquad\quad (0.16) \qquad (0.04) \qquad (0.01) \\ t &= \qquad -3.86 \qquad +2.31 \qquad +22.07 \\ \bar{R}^2 &= 0.990 \qquad N = 40 \text{ (Annual 1960 -1999)} \end{aligned}$$

Consider that you hypothesize that the demand for chicken also depends on IR, the interest rate. With the new, but irrelevant variable, the equation becomes

$$\begin{aligned} \hat{Y}_t &= 27.6 - 0.58PC_t + 0.12PB_t + 0.24YD_t - 0.14IR_t \\ &\qquad\quad (0.16) \qquad (0.05) \qquad (0.01) \qquad (0.18) \\ t &= \qquad -3.64 \qquad +2.33 \qquad +18.72 \qquad -0.81 \\ \bar{R}^2 &= 0.989 \qquad N = 40 \text{ (Annual 1960 -1999)} \end{aligned} \tag{74}$$

A comparison between (**??**) and (**??**) will explain the theory in Section 2.1. $\bar{R}^2$ has fallen indicating the reduction in fit adjusted for degrees of freedom. None of the estimated regression coefficients changed significantly, as compared between (**??**) and (**??**). The standard errors also increased or remained the same. The t-score for the interest rate is very small, indicating that it is not significantly different from zero. Given the theoretical shakiness (not entirely theoretically sound) of the new variable, these results indicate that it is irrelevant and never should have been included in the regression.

6.2.3 Four Important Specification Criteria

The four criteria used to decide whether a given variable belongs in the equation are:

1. ***Theory*** – is the variable's place in the equation theoretically sound and unambiguous?

2. ***t-test*** – is the variable's estimated coefficient significant in the expected direction?

3. $\boldsymbol{R^2}$ – Does the overall fit of the equation improve when the variable is added to the equation?

4. ***Bias*** – Do other variables' coefficients change significantly when the variable is added into the equation?

If all of these conditions hold then the variable belongs in the equation. If none of them hold then it does not and can be safely excluded from the equation.

6.3 An Illustration of the Misuse of Specification Criteria

Since economic theory is the most important test for including a variable, consider an example of why a variable should *not* be dropped from an equation simply because it has an insignificant t-score.

Suppose you believe that the demand for Brazilian coffee in the US is a negative function of the real price of Brazilian Coffee, P_{bc}, and a positive function of both the real price of tea, P_t and the real disposable income of the US, Y_d. Suppose further that you obtain data, run the implied regression and observe the following results

$$\begin{aligned} \widehat{COFFEE} &= 9.1 + 7.8P_{bc} + 2.4P_t + 0.0035Y_d \\ &\qquad\quad (15.6) \quad (1.2) \quad\; (0.0010) \\ t &= \qquad\; +0.5 \quad +2.0 \quad\; +3.5 \\ \bar{R}^2 &= 0.60 \qquad N = 25 \end{aligned} \tag{75}$$

The coefficients of P_t and Y_d appear to be fairly significant in the direction you hypothesized. But P_{bc} appears to have an insignificant coefficient with an unexpected sign. If you think that there is a possibility that the demand for Brazilian coffee is perfectly price inelastic (coefficient is zero) you run the same equation without the price variable, obtaining

$$\begin{aligned} \widehat{COFFEE} &= 9.3 + 2.6P_t + 0.0036Y_d \\ &\qquad\quad (1.0) \quad\ (0.0009) \\ t &= \qquad +2.6 \quad\ +4.0 \\ \bar{R}^2 &= 0.61 \qquad N = 25 \end{aligned} \tag{76}$$

By comparing (**??**) and (**??**), we can apply four specification criteria for the inclusion of a variable in an equation learnt earlier:

1. Since the demand for coffee could be perfectly price inelastic, the theory behind dropping the variable might be plausible

2. The t-score of the possibly irrelevant variable is 0.5, insignificant at any level

3. The $\bar{R}^2$ increased, meaning the fit increased, indicating that the variable is irrelevant

4. The coefficients changed little when P_{bc} was dropped,suggesting that there was little, if any biased caused by dropping the variable

Based on this analysis you could conclude that the demand for Brazilian coffee is indeed price inelastic and that the variable is therefore irrelevant. However, this is not the end, indeed.

Although the elasticity of demand of coffee is generally low, it is hard to believe that Brazilian coffee is immune to price competition from other kinds of coffee. Indeed, one would expect a bit of sensitivity in the demand for Brazilian coffee with respect to the price of Colombian coffee, a substitute. To test this hypothesis the price of Colombian coffee, P_{cc} was added to (**??**)

$$\begin{aligned} \widehat{COFFEE} &= 10.0 + 8.0P_{cc} - 5.6P_{bc} + 2.6P_t + 0.0030Y_d \\ &\qquad\qquad (4.0) \qquad (2.0) \qquad (1.3) \qquad (0.0010) \\ t &= \qquad\ +2.0 \quad\ -2.8 \quad\ +2.0 \qquad +3.0 \\ \bar{R}^2 &= 0.65 \qquad N = 25 \end{aligned} \tag{77}$$

By comparing (**??**) and (**??**) with, once again the four specification criteria

1. Both prices should have been included in the model. The logical justification is strong

2. The t-scores of the new variable is 2.0, significant on most levels
3. The value of $\bar{R}^2$ increased with the addition of th new variable, indicating the variable was an omitted variable
4. Two of the coefficients did not change significantly indicating that the correlations between these variables and the price of Colombian coffee are low. However the coefficient for the price of Brazilian coffee changed significantly, indicating bias in the original result.

Since the expected sign of the coefficient of the omitted variable, P_{cc} is positive and the simple correlation coefficient between the two competitive prices, $(r_{P_{bc},P_{cc}})$ is also positive, then the direction of the expected bias in $\hat{\beta}_{P_{bc}}$ is positive. The positive bias could be seen as the coefficient moved a lot from -5.6 to +7.8.

6.4 Specification Searches

One of the weaknesses of econometrics is that a researcher could potentially manipulate a data set to produce almost *any* result by specifying different regressions until estimates with the desired properties are obtained. The subject of how to search for the best specification is quite controversial among econometricians. This section does not aim to solve the controversy but provide some guidance and insight for beginning researchers.

6.4.1 Best Practices in Specification Searches

The recommendations for specification searches include

- Rely on theory rather than statistical fit as much as possible when choosing variables, functional forms and the like
- Minimize the number of equations estimated
- Reveal, in a foot note or appendix, all alternative specifications estimated

If theory, not $\bar{R}^2$ or t-scores, is the most important criterion for the inclusion of avariable in a regression equation, then it follows that most of the work of specifying a model should be done before you attempt to estimate the equation.

6.4.2 Sequential Specification Searches

The **sequential specification search** technique allows a researcher to estimate an undisclosed number of regressions and then present a final choice as if it were the only specification estimated. Such a method misstates the statistical validity of the regression results for two reasons:

- The statistical significance of the results is overestimated because the estimations of the previous regressions are ignored

- The expectations used by the researcher to choose between various regression results rarely, if ever, are disclosed. Thus the reader has no way of knowing whether or not all the other regression results had opposite signs or insignificant coefficients for the important variables

Unfortunately, there is no universally accepted way of conducting sequential searches, primarily because the appropriate test at one stage in the process depends on which tests previously were conducted, and also because the tests have been very difficult to invent. One possibility is to reduce the degrees of freedom in the "final" equation by one for each alternative specification attempted. This procedure is far from exact, but it does impose an explicit penalty for specification searches.

Instead, we recommend trying to keep the number of regressions estimated to be as low as possible; to focus on theoretical considerations when choosing variables or functional forms; and to document all the various specifications investigated. That means, we recommend parsimony (using theory and analysis to limit the number of specifications estimated) with disclosure (reporting all the equations estimated).

6.4.3 Bias Caused by Relying on the t-Test to Choose Variables

We stated in the previous section that sequential specification searches are likely to mislead researchers bout the statistical properties of their results. In particular, the practice of dropping a potential independent variable simply because its t-score indicates that its estimated coefficient is insignificantly different from zero will cause systematic bias in the estimated coefficients (and their t-scores) of the remaining variables.

Consider the hypothesized model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{78}$$

Assume further that, on the basis of theory, we are certain that X_1 belongs to the equation but we are not certain that X_2 belongs. Even though we have stressed our four criteria to determine whether X_2 should be included, many researchers just use the t-test on $\hat{\beta}_2$ to determine whether X_2 should be included. If this preliminary t-test is significantly different from zero, then researchers drop X_2 from the equation and consider Y to be a function of X_1.

Two kinds of mistakes could be made with such a system. First, X_2 can sometimes be left in the equation when it does not belong there, but this does not change the expected value of $\hat{\beta}_1$. This is a Type I Error.

Second, X_2 sometimes can be dropped from the equation when it belongs. This is a Type II Error. In this second case, the estimated coefficient of X_1 will be biased by the value of the true β_2 to the extent that X_1 and X_2 are correlated.

6.4.4 Sensitivity Analysis

Sensitivity Analysis consists of purposely running a number of alternative specifications to determine whether particular results are *robust* (not statistical flukes). In essence, we are trying to determine how sensitive a potential "best" equation is to a change in specification because the true specification is **not** known. Researchers who use sensitivity analysis run (and report on) a number of different reasonable specifications and tend to discount a result that appears significant in some specifications and insignificant in others. Indeed, the whole purpose of sensitivity analysis is to gain confidence that a particular result is significant in a variety of alternative specifications, functional forms,variable definitions and/or subsets of the data.

6.4.5 Data Mining

Data Mining involves exploring a data set not for the purpose of testing hypotheses or finding a specification, but for the purpose of uncovering empirical regularities that can inform economic theory. **However**, if you develop a hypothesis using data mining techniques, you **must** test that hypothesis on a ***different*** data set. A new data set must be used because our typical statistical tests have little meaning if the new hypothesis is tested on the data set that was used to generate it. After all, the researcher already knows, ahead of time, what the results will be! The use of dual data sets is easiest when there is a plethora of data.

In essence, improper data mining to obtain desired statistics for the final regression equation is a potentially unethical empirical research method. Whether the improper data mining is accomplished by estimating one equation at a time or by estimating batches of equations or by techniques like stepwise regression procedures, the conclusion is the same. Hypotheses developed by data mining should always be tested on a data set different from the one that was used to develop the hypothesis. Otherwise the researcher has not found any scientific evidence to support the hypothesis; rather a specification has been chosen in a way that is essentially misleading.

6.5 An Example of Choosing Independent Variables

Suppose a friend surveys all 25 members of her econometrics class and asks for your help in choosing a specification:

- GPA_i is the cumulative college GPA on the i-th student on a four point scale
- $HGPA_i$ is the cumulative high school GPA on the i-th student on a four point scale
- $MSAT_i$ is the highest score obtained by the i-th student on the mathematics section of the SAT test, maximum 800
- $VSAT_i$ is the highest score obtained by the i-th student on the verbal section of the SAT test, maximum 800
- $SAT_i = MSAT_i + VSAT_i$
- $GREK_i$ is a dummy variable equal to 1 if the i-th student is a member of a fraternity or sorority, 0 otherwise
- HRS_i is the i-th student's estimate of the average number of hours spent studying per course per week in college
- $PRIV_i$ is a dummy variable equal to 1 if the i-th student graduated from a private high school, 0 otherwise
- $JOCK_i$ is a dummy variable equal to 1 if the i-th student is or was a member of a varsity intercollegiate athletic team for at least one season, 0 otherwise

- $\ln EX_i$ is the natural log of the number of full courses that the i-th student has completed in college

Letting GPA_i be the dependent variable, which independent variables would you choose? After the author's arguments, and running the regressions, he derived the following OLS model:

$$\begin{array}{rllll} \widehat{GPA_i} = & -0.26 + 0.49 HGPA_i & + 0.06 HRS_i & + 0.42 \ln EX_i \\ & (0.21) & (0.02) & (0.14) \\ t = & +2.33 & +3.00 & +3.00 \\ \bar{R}^2 = 0.585 & N = 25 & & \end{array} \tag{79}$$

Since the overall fit seems reasonable and since each each coefficient meets our expectations in terms of sign, size and significance, we consider this an acceptable equation. If we believe that we might have omitted SAT scores, we consider

$$\begin{array}{rlllll} \widehat{GPA_i} = & -0.92 + 0.47 HGPA_i & + 0.05 HRS_i & + 0.44 \ln EX_i & + 0.00060 SAT_i \\ & (0.22) & (0.02) & (0.14) & (0.00064) \\ t = & +2.12 & +2.50 & +3.12 & +0.93 \\ \bar{R}^2 = 0.583 & N = 25 & & & \end{array} \tag{80}$$

Using the four specification criteria to compare (**??**) and (**??**):

- The theoretical validity of SAT scores is controversial but it is the most widely accepted way of testing academic potential
- t-test The coefficient of SAT is positive but not significantly different from 0
- $\bar{R}^2$ decreases when SAT scores were added
- None of the estimated slopes changed significantly when SAT was added, though some of the t-scores did change because of the increase in the $\mathbb{SE}(\hat{\beta})$s caused by the addition of the SAT

Thus, the statistical criteria support our theoretical contention that SAT is irrelevant.

6.6 Additional Specification Criteria

We shall describe three of the most popular specification criteria.

6.6.1 Ramsey's Regression Specification Error Test (RESET)

The **Ramsey RESET test** is a general test that determines the likelihood of an omitted variable or some other specification error by measuring whether the fit of a given equation can be significantly improved by the addition of $\hat{Y_2}$, $\hat{Y_3}$ or $\hat{Y_4}$ terms. The additional terms act as proxies for any possible unknown omitted variables or incorrect functional forms. If the proxies can be shown by the F-test to have improved the overall fit of the original equation, then we have evidence that there is some sort of specification error in our equation.

The Ramsey RESET test has three steps:
Estimate the equation to be tested using OLS:

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_{1i} + \hat{\beta_2} X_{2i} \tag{81}$$

Take the $\hat{Y_i}$ values from (**??**) and create $\hat{Y_2}$, $\hat{Y_3}$ and $\hat{Y_4}$ terms. Add these terms to equation (**??**) as additional explanatory variables and estimate the new equation with OLS:

$$\hat{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \hat{Y_i}^2 + \beta_4 \hat{Y_i}^3 + \beta_5 \hat{Y_i}^4 + \epsilon_i \tag{82}$$

Compare the fits of (**??**) and (**??**) using the F-test. If the two equations are significantly different in overall fit, then we can conclude that (**??**) is misspecified.

While the Ramsey RESET test is fairly easy to use, it does little more than signal *when* a major specification error might exist. If you encounter a significant Ramsey RESET test result then you face the daunting task of figuring out what the error is.

As an example of the Ramsey RESET test, consider the chicken demand model to see if RESET could detect that the known specification error. In step 1, run the equation with the omission of PB.

$$\begin{aligned}
\hat{Y_t} &= 27.5 - 0.42PC_t + 0.27YD_t \\
& \qquad\qquad (0.14) \qquad (0.005) \\
t &= \qquad\quad -2.95 \qquad +55.00 \\
\bar{R^2} &= 0.988 \qquad N = 40 \text{ (Annual 1960 -1999)}
\end{aligned}$$

Take $\hat{Y}_t$ from (**??**), calculate $\hat{Y}_t^2$, $\hat{Y}_t^3$ and $\hat{Y}_t^4$ and reestimate (**??**) with the new terms added in:

$$\begin{aligned}
\hat{Y}_t &= 243.8 - 6.30PC_t + 4.20YD_t - 0.41\hat{Y}_t^2 + 0.005\hat{Y}_t^3 + 0.00002\hat{Y}_t^4 + e_t \\
&\qquad\qquad (0.98) \qquad (0.66) \qquad (0.07) \qquad (0.0009) \qquad (0.000004) \\
t &= \qquad -6.39 \qquad +6.35 \qquad -5.84 \qquad +5.73 \qquad -5.61 \\
\bar{R}^2 &= 0.994 \qquad N = 40 \text{ (Annual 1960 -1999)} \\
\text{RSS} &= 79.27
\end{aligned} \tag{83}$$

In step 3, we compare the fits of the two equations by using the F-test. Specifically, we test the hypothesis that the coefficients of all three added terms are equal to zero:

$$\text{H}_0\text{: } \beta_3 = \beta_4 = \beta_5 = 0$$
$$\text{H}_\text{A}\text{: otherwise}$$

The appropriate F-statistic to use is

$$F = \frac{(\text{RSS}_M - \text{RSS})/M}{\text{RSS}/N-K-1} = 12.16 \tag{84}$$

The critical value to use is $M = 3$ and degrees of freedom=34 so the value is 2.89. Since $12.16 > 2.89$ then we can reject the null hypothesis that the coefficients of the added variables are jointly zero, concluding that there is a specification error in (**??**). True enough, the price of beef was not included in the equation. Remember that the RESET test only tells us that there is a specification error but does not tell us the details of the error.

6.6.2 Akaike's Information Criterion and Schwarz Criterion

Akaike's Information Criterion (AIC) and **Schwarz Criterion (SC)** are methods of comparing alternative specifications by adjusting RSS for the sample size, N and the number of independent variables, K. These criteria can be uses to augment our four basic specification criteria when we try to decide if the improved fit caused by an additional variable is worth the decreased degrees of freedom and increased complexity caused by addition. Their equations are:

$$\text{AIC} = \log\frac{\text{RSS}}{N} + \frac{2(K+1)}{N} \tag{85}$$

$$\mathrm{SC} = \log \frac{\mathrm{RSS}}{N} + \frac{(K+1)\log N}{N} \tag{86}$$

To use AIC and SC, estimate two alternative specifications and calculate AIC and SC for each equation. The lower the AIC or SC are, the better the specification. Both criteria penalize the addition of another explanatory variable more than $\bar{R}^2$ does. Applying AIC and SC to the chicken demand example, we need to calculate AIC and SC for equations with and without the price of beef. The equation with the lower AIC and SC values, all else being equal, will be our preferred specification. Plugging the numbers from (**??**) into (**??**) and (**??**), AIC and SC can be

$$\mathrm{AIC} = \log \frac{143.07}{40} + \frac{2(4)}{40} = 1.47$$

$$\mathrm{SC} = \log \frac{143.07}{40} + \frac{(4)\log 40}{40} = 1.64$$

against (**??**) which omits the price of beef,

$$\mathrm{AIC} = \log \frac{164.31}{40} + \frac{2(3)}{40} = 1.56$$

$$\mathrm{SC} = \log \frac{164.31}{40} + \frac{(3)\log 40}{40} = 1.69$$

Since AIC and SC are better with the inclusion of the BC variable, we say that AIC and SC provide evidence that (**??**) is preferable to (**??**). As it turns out, all three new specification criteria indicate that the presence of a specification error when we leave the price of beef out of the equation.

RESET is most useful as a general test of the existence of a specification error, while AIC and SC are more usefulas a measn of comparing two or more alternative specifications.

7 Specification: Choosing a Functional Form

7.1 The Use and Interpretation of the Constant Term

In the linear regression model, β_0 is the intercept or constant term, the expected value of Y when all the explanatory variables (and the error term) equal zero. An estimate of β_0 has at least three components:

1. the true β_0
2. the constant impact of any specification errors (an omitted variable, for example)
3. the mean of ϵ for the correctly specified equation (if not equal to zero)

Unfortunately, these components cannot be distinguished from one another because we can only observe β_0, the sum of the three components. The result is that we have to analyze β_0 differently from the way we analyze the other coefficients in the equation.

At times, β_0 is of theoretical importance. At others, it is not. Those are times when the researcher suppresses the constant term. Neither suppressing the constant term nor relying on it for inference is advisable, however, and reasons for these conclusions are as follows.

7.1.1 Do not Suppress the Constant Term

Suppressing the constant term leads to a violation of the Classical Assumptions. This is because Classical Assumption II can be met only if the constant term absorbs any nonzero mean that the observations of the error might have in a given sample. If you omit the constant term, then the constant effect of omitted variables, non-linearities etc is forced into the estimates of the other coefficients, causing potential bias.

The consequence of suppressing the constant term is that the slope coefficient estimates are potentially biased and their t-scores are potentially inflated.

7.1.2 Do not Rely on Estimates of the Constant Term

There are reasons that suggest that the intercept should *not* be relied on for purposes of analysis or inference. First, the error term is generated, in part, by the omission of a number of marginal independent variables, the mean effect of which is placed in the constant term. Second, the constant term is the value of the dependent variable when all the dependent variables and the error term are zero, but these values almost never equal zero because the variables used for economic analysis are usually positive. Hence, the origin often lies *outside* the range of sample ovservations. Since the constant term is an estimate of Y when the Xs are outside the range of the sample observations, estimates of it are tenuous. Estimating the constant term is like forecasting beyond the range of the sample data, a procedure that inherently contains greater error than within-sample forecasts.

7.2 Alternative Functional Forms

Before talking about functional forms, we need to differentiate between ean equation that is liner in the coefficients and one that is linear in the variables.

An equation that is **linear in the variables** if plotting the function in terms of X and Y generates a straight line. In the following

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{87}$$

$$Y = \beta_0 + \beta_1 X^2 + \epsilon \tag{88}$$

(**??**) is linear in the variables but (**??**) is not linear in the variables because a plot of (**??**) results in a quadratic line, not a straight line.

An equation is **linear in the coefficients** only if the coefficients (the βs) appear in their simplest form – they are not raised to any powers (other than one), are not multiplied or divided by other coefficients, and do not themselves include some sort of function (like logs or exponents). In the following

$$Y = \beta_0 + X^{\beta_1} \tag{89}$$

(**??**) is not linear in the coefficients β_0 and β_1. (**??**) is not linear because there is no rearrangement of the equation that will make it linear in the βs of original interest. Of all possible equations for a single explanatory variable, *only* functions of the general form

$$f(Y) = \beta_0 + \beta_1 f(X) \tag{90}$$

are linear in the coefficients β_0 and β_1. In essence, any sort of configuration of the Xs and Ys can be used and the equation will continue to be linear in the coefficients.

Although linear regressions, as defined by the Classical Assumptions, need to be linear in the coefficients, they do not necessarily need to be linear in the variables. Linear regression analysis can be formulated in a way that is linear in the coefficients.

The use of OLS requires that the equation be linear in the coefficients, but there is a wide variety of function forms that are linear in the coefficients while being nonlinear in the variables. The choice of a functional form almost always should be based on the underlying economic or business theory and only rarely on which form provides the best fit. The next paragraphs will characterize the most frequently used forms.

7.2.1 Linear Form

This is based on the assumption that the slope of the relationship between the independent variable and the dependent variable is constant:

$$\frac{\Delta Y}{\Delta X_k} = \beta_k \quad \forall\, k = 1, 2, \cdots, K$$

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 functional form should be used.

Since the slope is constant, the **elasticity** of Y with respect to X, defined as the percentage change in Y with a one percent change in the independent variable holding all other variables constant, can be calculated fairly easily

$$\text{Elasticity}_{Y,X_k} = \frac{{}^{\Delta Y}\!/_{Y}}{{}^{\Delta X_k}\!/_{X_k}} = \frac{\Delta Y}{\Delta X_k} \cdot \frac{X_k}{Y} = \beta_k \frac{X_k}{Y}$$

Theory frequently predicts only the sign of a relationship and not functional form. Unless theory or other reasons justifies using some other functional form, you should use the linear model.

7.2.2 Double-Log Form

The double-log form is the most common functional form that is non-linear 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:

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \epsilon \tag{91}$$

where, in (**??**), the double-log form, sometimes called the log-log form, often is used because the researcher has specified that the elasticities of the model are constant and the slopes are not. In a double-log equation, an individual regression coefficient can be interpreted as an elasticity because

$$\beta_k = \frac{\Delta(\ln Y)}{\Delta(\ln X_k)} = \frac{\Delta Y / Y}{\Delta X_k / X_k} = \text{Elasticity}_{Y,X_k} \tag{92}$$

Since regression coefficients are constant, the condition that the model have a constant elasticity is met by the double-log equation.The way to interpret β_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 β_k percent.

7.2.3 Semilog Form

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

$$Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{93}$$

In this case the economic meanings of the two slope coefficients are different, since X_2 is linearly related to Y while X_1 is nonlinearly related to Y.

Applications of the semilog form are quite frequent in economics and business. Most consumption functions tend to increase at a decreasing rate past some level of income. These *Engel curves* tend to flatten out because as incomes get higher, a smaller percentage of income goes to consumption and a greater percentage goes to saving. Consumption thus increases at a decreasing rate. For example, consider the beef consumption example

$$\begin{aligned}
\widehat{BC_t} &= 37.54 - 0.88P_t + 11.9Yd_t \\
& \qquad\qquad\;\; (0.16) \qquad (1.76) \\
t &= \qquad\quad -5.36 \qquad +6.75 \\
\bar{R^2} &= 0.631 \qquad N = 28
\end{aligned}$$

where BC is per capita consumption of beef, P is price of beef in cents per pound and Yd is US disposable income in \$'000

If we substitute the log of disposable income for disposable income, we get

$$\begin{aligned}\widehat{BC_t} &= -71.75 - 0.87P_t + 98.87\ln Yd_t \\ & \qquad\qquad\quad (0.13) \qquad\quad (11.11) \\ t &= \qquad\quad -6.93 \qquad\ + 8.90 \\ \bar{R}^2 &= 0.750 \qquad N = 28\end{aligned} \tag{94}$$

In (**??**) the independent variables include the price of beef and the *log* of disposable income. (**??**) would be appropriate if we hypothesize that as income rises, consumption will increase at a decreasing rate.

Some semilog functions have the log of the left-hand side of the equation

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{95}$$

This model neigher has a constant slope or a constant elasticity. If X_1 increases by one *unit*, then Y will change in *percentage* terms. Specifically Y will change by $\beta_1 \cdot 100$ percent, holding X_2 constant, for every 1 unit increase in X_1.

7.2.4 Polynomial Form

Polynomial functional forms express Y as a function of independent variables, some raised to powers other than one. For example, in a second-degree polynomial equation, at least one independent variable is squared

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \epsilon_i \tag{96}$$

Such a model can produce slopes that change as the independent variables change, for example differentiating w.r.t X_1

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1 \tag{97}$$

The slope depends on the value of X_1.

Consider a model of annual employee earnings as a function of age of each employee and a number of other measures of productivity.Such a theoretical relationship could be modeled with a quadratic equation

$$EARNINGS_i = \beta_0 + \beta_1 AGE_i + \beta_2 (AGE_i)^2 + \cdots + \epsilon_i \tag{98}$$

Since you expect earnings to increase up to a certain age, and due to retirement earnings will drop after a certain age, then the sign of $\hat{\beta}_2$ should be negative.

7.2.5 Inverse Form

The **inverse functional form** expresses Y as a function of the *reciprocal* of one or more of the independent variables

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_{1i}} + \beta_2 X_{2i} + \epsilon_i \tag{99}$$

The inverse (or reciprocal) 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 (**??**), X_1 cannot equal zero as dividing by zero will produce and undefined result. The slope with respect to X_1 is

$$\frac{\Delta Y}{\Delta X_1} = \frac{-\beta_1}{X_1^2} \tag{100}$$

An example of the application of the inverse functional form is the Philips curve which explores the rate of unemployment and the percentage change in wages.

$$W_t = \beta_0 + \beta_1 \frac{1}{U_t} + \epsilon_t \tag{101}$$

and estimating the equation with OLS yields

$$\begin{aligned} \hat{\beta_t} &= 0.00679 - 0.1842 \frac{1}{U_t} \\ &\qquad\qquad (0.0590) \\ t &= \qquad\qquad +3.20 \\ \bar{R^2} &= 0.397 \end{aligned} \tag{102}$$

This indicates that W and U are related in a way similar to that hypothesized, but it doesn't provide any evidence that the inverse functional form is the best way to depict this particular theory.

7.3 Lagged Independent Variables

So far, the equations we have studied involve equations that include independent and dependent variables from the same time period, as in

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t \tag{103}$$

where the subscript t is used to refer to a particular point in time. If all variables have the same subscript, then the equation is instantaneous.

In many cases, in contrast, we must allow for the possibility that time might elapse between the change in the independent variable and the resulting change in the dependent variable. The length of time between cause and effect is called a **lag**. Many econometric equations include one or more *lagged independent variables* like X_{1t-1} where the subscript $t-1$ indicates that the observation of X_1 is from the time period previous to time period t, as in the following equation

$$Y_t = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{2t} + \epsilon_t \tag{104}$$

In this equation, X_1 is lagged by one time period, but the relationship between Y and X_2 is still instantaneous.

For example, consider the process by which the supply of an agricultural product is determined. Since agricultural goods take time to growth, decisions on how many acres to plant must be made ahead of time. Any change in an agricultural market, such as increase in the price that the farmer can earn for providing cotton, has a lagged effect on the supply of that product

$$C_t = f(\overbrace{PC_{t-1}}^{+}, \overbrace{PF_t}^{-}) + \epsilon_t = \beta_0 + \beta_1 PC_{t-1} + \beta_2 PF_t + \epsilon_t \tag{105}$$

where C_t is the quantity of cotton supplied in year t, PC_{t-1} is the price of cotton in year $t-1$ and PF_t is price of farm labor in year t.
This equation hypothesizes a lag between the price of cotton and the production of cotton. The meaning of the regression coefficient of a lagged variable is the estimated change of *this year's* cotton production given a one-unit change in *last year's* price of cotton, holding this year's price of farm labor constant. If the lag structure is hypothesized to take place over *more than one* time or a lagged dependent variable is included on the right hand side of the equation, then the question becomes significantly more complex. They are called *distributed lags*.

7.4 Using Dummy Variables

This section focuses on the use of a dummy variable 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

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i \tag{106}$$

textwhere

$$D_i = \begin{cases} 1 \text{ if the } i\text{-th observation meets a particular condition} \\ 0 \text{ otherwise} \end{cases}$$

The intercept dummy does indeed change the intercept depending on the value of D, but the slopes remain constant, independent on the value of D. The event not explicitly represented by a dummy variable, called an **omitted condition**, forms the basis against which the included conditions are compared. Thus, for dual situations only one dummy is entered.

Consider the study of the relationship between fraternity/sorority membership and GPA. We want to build a regression model that explains college GPA. Independent variables would not only include Greek membership but also other predictors of academic performance like the SAT and high school school grades. Being a member of a social organization is a qualitative variable so we assign it a dummy variable

$$G_i = \begin{cases} 1 \text{ if the } i\text{-th student is an active member of a fraternity or sorority} \\ 0 \text{ otherwise} \end{cases}$$

By collecting data and estimating the equation implied above

$$\begin{aligned} \widehat{CG_i} &= 0.37 - 0.81HG_i + 0.00001S_i - 0.38G_i \\ \bar{R}^2 &= 0.45 \qquad N = 25 \end{aligned} \tag{107}$$

where CG_i is the cumulative college GPA out of 4 for the i-th student, HG_i is the cumulative high school GPA out of 4 for the i-th student and S_i is the sum of the highest verbal and mathematics SAT scores earned by the i-th student (1600 maximum)

The meaning of the estimated coefficient G_i is the GPA of fraternity/sorority members is 0.38 lower than that for non-members, holding GPAs and SAT scores constant. Before taking these results and generating a conclusion, be careful of the interpretation of them.

7.5 Slope Dummy Variables

Until now, every independent variable in the text has been multiplied by exactly one other term, the slope coefficient. Consider again

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i$$

Here, X is only multiplied by β_1 and D is only multiplied by β_2. There are no other factors involved.

This restriction does not apply to a new kind of variable called an interaction term. An **interaction term** is 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 end result is that the interaction term has three or more components, as in $\beta_3 X_i D_i$. Such interaction terms are used when the change in Y with respect to one independent variable depends on the level of another independent variable.

The most frequent application of interaction terms is to create slope dummies. **Slope dummy variables** allow the slope of the relationship between the dependent variable and an independent variable to be different depending on whether the condition specified by a dummy variable is met. This is in contrast to an intercept dummy variable which changes the intercept but not the slope, when a particular condition is met.

In practice, slope dummy variables have may realistic uses. It could be used whenever the impact of an independent variable on the dependent variable is hypothesized to a change if some qualitative condition is met.
For example consider a consumption function that is estimated over a time period that includes a major war. Being in a war would surely reduce the marginal propensity to consume and such a change can be modeled with a slope dummy that takes on one value during war years and other during non-war years.

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

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \epsilon_i \tag{108}$$

In the case of the consumption function, Y would be consumption, X would be disposable income, D would measure if the i-th year was a war year. To see whether the slope of Y w.r.t X changes if D changes:

$$\text{When } D = 0, \quad \frac{\Delta Y}{\Delta X} = \beta_1$$

$$\text{When } D = 1, \quad \frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3$$

The coefficient of X changes when the condition specified by D is met. Note that (**??**) includes *both* a slope dummy and an intercept dummy. It turns out, whenever a slope dummy is used, it is vital to also have $\beta_1 X_1$ and $\beta_2 D$ in the equation to avoid bias in the estimate of the coefficient of the slope dummy term. Such a specification should be used in all highly unusual and forced situations.

Consider the question of earning differentials between men and women. Suppose you decide to build a model of earnings to understand whether sexual discrimination exists and hypothesize that men earn more than women on average. Ignoring the other relevant factors eg. experience, education etc., you come up with the following

$$\ln(EARNINGS_i) = \beta_0 + \beta_1 D_i + \beta_2 EXP_i + \cdots + \epsilon_i \tag{109}$$

where D_i is 1 if the i-th worker is male and 0 otherwise, EXP_i is the years of experience of the i-th worker and ϵ_i is a classical error term.
In (**??**), $\hat{\beta_1}$ would be an estimate of the average difference between males and females holding constant their experience and the other factors in the equation. (**??**) also forces the impact of increases in experience to have the same effect for females and males because the slopes are the same for both genders.

If you hypothesize that men also increase their earnings more per year of experience than women then you should add a slope dummy and an intercept dummy:

$$\ln(EARNINGS_i) = \beta_0 + \beta_1 D_i + \beta_2 EXP_i + \beta_3 EXP_i D_i + \cdots + \epsilon_i \tag{110}$$

In (**??**), $\hat{\beta_3}$ would be an estimate of the differential impact of an extra year of experience on earnings between men and women. We could test the possibility of a positive true β_3 by running a one-tailed test on $\hat{\beta_3}$. If $\hat{\beta_3}$ were significantly different from zero in a positive direction, then we could reject the null hypothesis and claim that there is a difference due to gender in the impact of experience on earnings, holding the other variables constant.
Chapters 8, 9 and 10 deal with violations of the Classical Assumptions and remedies for those violations. This chapter introduces Multicollinearity. The next two will introduce serial correlation and heteroskedasticity.

Perfect multicollinearity is the violation of Classical Assumption VI that no independent variable is a perfect linear function of one or more independent variables. 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 Multicollinearity

8.1 Perfect vs Imperfect Multicollinearity

8.1.1 Perfect Multicollinearity

Perfect multicollinearity violates Classical Assumption VI, which specifies that no explanatory variable is a perfect linear function of any other explanatory variables. Perfect means that the variable can be *completely* explained by movements in another explanatory variable. Such a perfect linear function would be:

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} \tag{111}$$

where the αs are constants and the Xs are independent variables in"

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{112}$$

Notice that there is no error term in (**??**), implying that X_1 can be exactly calculated given X_2 and the equation.Examples of such perfect linear relationships would be

$$X_{1i} = 3X_{2i} \qquad X_{1i} = 6 + X_{2i} \qquad X_{1i} = 2 + 4X_{2i} \tag{113}$$

An example of perfect multicollinearity is interest rates. When nominal and real interest rates are included as explanatory variables in an equation, the relationship between the two continually change because the difference between the two, the rate of inflation, is changing. If the interest rate were constant, the difference between the two would be constant. They would be perfectly linearly related and perfect multicollinearity would result:

$$IN_t = IR_t + INF_t = IR_t + \alpha \tag{114}$$

where
IN_t is the nominal interest rate in time t
IR_t is the real interest rate in time t
INF_t is the rate of inflation in time t
α is the constant rate of inflation

When estimation of an econometric equation where there exists perfect multicollinearity, OLS is incapable of generating estimates of the regression coefficients and most OLS programs will print out an error message. Using

(**??**), we theoretically would obtain the following estimated coefficients and standard errors:

$$\hat{\beta}_1 = \text{indeterminate} \qquad \mathbb{SE}(\hat{\beta}_1) = \infty \tag{115}$$

$$\hat{\beta}_2 = \text{indeterminate} \qquad \mathbb{SE}(\hat{\beta}_2) = \infty \tag{116}$$

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 by definition so highly correlated with the dependent variable that it completely masks the effects of all other independent variables in the equation. In a sense, this is a case of perfect collinearity between the dependent and an independent variable.

8.1.2 Imperfect Multicollinearity

When econometricians talk about multicollinearity, they usually refer to severe imperfect multicollinearity. **Imperfect multicollinearity** can be defined as 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. In other words, imperfect multicollinearity occurs when two or more explanatory variables are imperfectly linearly related as in:

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + u_i \tag{117}$$

Comparing (**??**) to (**??**), notice that (**??**) includes a stochastic error term. This implies that although the relationship between X_1 and X_2 might be fairly strong, it is not strong enough to allow X_1 to be completely explained by X_2; some unexplained variation still remains.

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 will be considered significantly multi-collinear. It is fair to say that multi-collinearity is a sample phenomenon as well as a theoretical phenomenon.

8.2 The Consequences of Multicollinearity

If the multi-collinearity in a particular sample is severe, what will happen to estimates calculated from that sample? What consequences does significant imperfect multicollinearity imply? Recall that the OLS estimators are BLUE if the Classical Assumptions hold. This means that OLS estimates can be thought of as being unbiased and having the minimum variance possible for unbiased linear estimators.

8.2.1 What are the Consequences of Multicollinearity?

The three major consequences of Multicollinearity are:

1. ***Estimates will remain unbiased***. Even if an equation has significant multicollinearity, the estimates of the βs will still be centered around the true population βs if all the Classical Assumptions are met for a correctly specified equation.

2. ***The variances and standard errors of the estimates will increase***. This is the principal consequence of multicollinearity. Since two or more of the explanatory variables are significantly related, it becomes difficult to precisely identify the separate effects of the multicollinear variables. When it becomes hard to distinguish the effect of one variable from the effect of another, then we are much more likely to make large errors in estimating the βs than we were before we encountered multicollinearity. As a result the estimated coefficients, although still unbiased come from distributions with much larger variances and, therefore larger standard errors.

   In particular, a consequence of having a large variance due to multicollinearity is a higher probability of obtaining a $\hat{\beta}$ that is dramatically different from the true β. For example, it turns out that multicollinearity increases the likelihood of obtaining an unexpected sign for a coefficient even though, as mentioned above, multicollinearity causes no bias.

3. ***The computed t-scores will fall***. Multicollinearity tends to decrease the t-scores of the estimated coefficients mainly because of the formula for the t-statistic:

$$t_k = \frac{(\hat{\beta}_k - \hat{\beta}_{H_0})}{\mathbb{SE}(\hat{\beta}_k)} \tag{118}$$

Notice that this equation is divided by the standard error of the estimated coefficient. Multicollinearity increases the standard error of the estimated coefficient, and if the standard error increases, then the t-score must fall. Not surprisingly it is common to observe low t-scores in equations with severe multicollinearity.

4. ***Estimates will become very sensitive to changes in specification.*** The addition or deletion of an explanatory variable or of a few observations will often cause major changes in the values of the $\hat{\beta}$s when significant multicollinearity exist. If you drop a variable, even one that appears to be statistically insignificant, the coefficients of the remaining variables in the equation will sometimes change dramatically.

5. ***The overall fit of the equation and the estimation of the coefficients of non-multicollinear variables will be largely unaffected.*** Even if the t-scores are low in a multicollinear equation, the overall fit of the equation, as measured by $\bar{R}^2$ will not fall much, if at all, in the face of significant multicollinearity.

8.2.2 Two Examples of the Consequences of Multicollinearity

Consider that you decide to estimate a "student consumption function" and come up with the hypothesized equation

$$CO_i = f(\overbrace{Yd_i}^{+}, \overbrace{LA_i}^{+}) + \epsilon_i = \beta_0 + \beta_1 Yd_i + \beta_2 LA_i + \epsilon_i \tag{119}$$

where
CO_i is the annual consumption expenditures of the i-th student on items other than tuition and orom and board
Yd_i is the annual disposable income (including gifts) of that student
LA_i is the liquid assets (savings, etc.) of the i-th student
ϵ_i is a stochastic error term After running the OLS regression on the data set for (**??**) you get

$$\begin{aligned} \widehat{CO_i} &= -367.83 + 0.5113 Yd_i + 0.0427 LA_i \\ &\qquad\qquad\quad (1.0307) \qquad (0.0942) \\ t &= \qquad\qquad\quad +0.496 \qquad +0.453 \\ \bar{R}^2 &= 0.835 \qquad N = 7 \end{aligned} \tag{120}$$

on the other hand, if you had consumption as a function of disposable income alone, you get

$$\begin{aligned} \widehat{CO_i} &= -471.43 + 0.9714Yd_i \\ & \qquad\qquad\quad (0.157) \\ t &= \qquad\qquad\quad +6.187 \\ \bar{R^2} &= 0.861 \qquad N = 7 \end{aligned} \tag{121}$$

Notice from (**??**) to (**??**), the t-score for disposible income increased drastically (more than 10-fold) when the liquid assets variable is dropped from the equation. To explain, first the simple correlation coefficient between Yd and LA is quite high: $r_{Yd,LA} = 0.986$. This high degree of correlation causes the standard errors of the estimated coefficients to be very high when both variables are included. In the case of $\hat{\beta_Y d}$, the standard error increased form 0.157 to 1.03 with the inclusion of LA! The coefficient estimate itself changes somewhat. Further, note that the $\bar{R^2}$s of the two equations are quite similar despite the large differences in the significance of the explanatory variables in the two equations. It's quite common for $\bar{R^2}$ to stay virtually unchanged when multicollinear variables are dropped. All of these results are typical equations with multicollinearity.
Which equation is better? Adding the liquid assets variable means certain multicollinearity. Dropping it will result in omitted variable bias. There is no automatic answer.

A second example of the consequences of multicollinearity is based on actual, rather than hypothetical data. Say we build a cross-sectional model of the demand for gasoline by state:

$$PCON_i = f(\overbrace{UHM_i}^{+}, \overbrace{TAX_i}^{-}, \overbrace{REG_i}^{+}) + \epsilon_i \tag{122}$$

where
$PCON_i$ is petroleum consumption in the i-th state (trillions of British Thermal Units or BTU)
UHM_i is the urban highway miles within the i-th state
TAX_i is the gasoline tax rate in the i-th state
REG_i is the motor vehicle registrations in the i-th state

Using a classical error term, the OLS estimation gives us

$$\begin{aligned}\widehat{PCON}_i &= 389.6 - 60.8UHM_i - 36.5TAX_i - 0.0061REG_i \\ & \qquad\qquad (10.3) \qquad\quad (13.2) \qquad\quad (0.043) \\ t &= \qquad +5.92 \qquad\quad -2.77 \qquad\quad -1.43 \\ \bar{R}^2 &= 0.919 \qquad N = 50\end{aligned} \tag{123}$$

The motor vehicle registrations variable has an insignificant coefficient with an unexpected sign, but it is hard to believe the variable is irrelevant. Knowing that there is a strong correlation between REG and UHM with a correlation coefficient of 0.98, we are dealing with a case of multicollinearity. It seems fair to say that one of the two variables is redundant.

Note the impact of the multicollinearity on the equation. The coefficient of a variable such as motor vehicle registrations, which has a very strong theoretical relationship to petroleum consumption, is insignificant and has a sign contrary to our expectations. If we drop one of the multicollinear variables, we get

$$\begin{aligned}\widehat{PCON}_i &= 551.7 - 53.6TAX_i + 0.186REG_i \\ & \qquad\qquad (16.9) \qquad\quad (0.012) \\ t &= \qquad -3.18 \qquad\quad +15.88 \\ \bar{R}^2 &= 0.861 \qquad N = 50\end{aligned} \tag{124}$$

Dropping UHM made REG extremely significant. The standard error of the coefficient of REG has fallen substantially now that the multicollinearity has been removed from the equation. Also note that the sign of the estimated coefficient has now become positive as hypothesized. The reason is that REG and UHM are virtually indistinguishable from an empirical point of view and so the OLS program latched onto minor differences between the variables to explain the movements of $PCON$. Once the multicollinearity was removed, the direct positive relationship between REG and $PCON$ was obvious. Note, however, that the coefficient of the REG variable now measures the effect of both REG and UHM on $PCON$. Since we dropped a variable, the remaining coefficient helps explain the effect of the omitted variable.

Either UHM or REG could have been dropped with similar results because they are quantitatively similar. Even though $\bar{R}^2$ fell when UHM was dropped, (**??**) should be considered superior to (**??**). This is an example that the fit of the equation is *not* the most important criterion to be used in determining its overall quality.

8.3 The Detection of Multicollinearity

To first begin this section, we should recognize that some multicollinearity exists in every equation. It is hard to find a real-world example to find a set of explanatory variables in which the explanatory variables are totally uncorrelated with each other. Hence, the main purpose of this section is to learn to determine *how much* multicollinearity exists in an equation, not *whether* any multicollinearity exists.

Second, multicollinearity is a sample phenomenon as well as a theoretical one. Hence, 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. The trick is to find variables that are theoretically relevant and are also statistically non-multicollinear.

Let us examine two of the most used characteristics that determine the severity of multicollinearity in an estimated equation.

8.3.1 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 as high in absolute value, then we know that the two particular Xs are quite correlated and that multi-collinearity is a potential problem. For example in (**??**) the simple coefficient between Yd and LA is 0.986. A simple correlation coefficient this high is a ertain indication of severe multicollinearity.

How high is high? Some researchers pick an arbitrary number, like 0.80, and become concerned if the value of a simple correlation coefficient exceeds 0.80. A better answer might be that r is high if it causes unacceptably large variances in the coefficient estimates in which we are interested.

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. It is quite possible for groups of independent variables, acting together, to cause multicollinearity without any single simple correlation coefficient being high enough to indicate that multicollinearity is in fact severe.

8.3.2 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 the severity of multicollinearity that is easy to use and that is gaining in popularity is the variance inflation factor. The **variance inflation factor (VIF)** is 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 variable in an equation. There is a VIF for each explanatory variable in an 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 estimated variance of the estimated coefficient by quite a bit, yielding a decreased t-score.

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_K X_K + \epsilon$$

Doing so requires calculating K different VIFs, one for each X_i. Calculating the VIF for each X_i involves three steps:

1. ***Run an OLS regression that has X_i as a function of all the other explanatory variables in the equation.*** For example, if $i = 1$, then this equation would be

   $$X_1 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_K X_K + v \tag{125}$$

   where v is a classical stochastic error term. Note that X_1 is not included on the right hand side of (**??**), which is referred to as an auxiliary regression. Thus, there are K auxiliary regressions, one for each independent variable in the equation.

2. ***Calculate the variance inflation factor for $\hat{\beta}_i$***

   $$\text{VIF}(\hat{\beta}_i) = \frac{1}{(1 - R_i^2)} \tag{126}$$

   where R_i^2 is the coefficient of determination (the unadjusted R^2) of the auxiliary regression in step one. Since there is a separate auxiliary regression for each independent variable in the original equation, there also is an R_i^2 and a VIf($\hat{\beta}_i$) for each X_i.

3. ***Analyze the degree of multicollinearity by evaluating the size of the VIF($\hat{\beta}_i$).*** The higher a given variable's VIF, the higher the variance of that variable's estimated coefficient (holding constant the variance of the error term). Hence, the higher the VIF, the more the severe the effects of multicollinearity.

How high is high? An R_i^2 of one, indicating perfect multicollinearity, produces a VIF of ∞, where an R_i^2 of zero, indicating no multicollinearity at all produces a VIF of one. The rule of thumb is that if VIF(β_i)¿5 the multicollinearity is severe. As the number of independent variables increase, it makes sense to increase this number too.

Consider (**??**) again and calculate the VIFs for both independent variables. Both VIFs equal 36, confirming the severe multicollinearity which we already know exists. In an equation with two independent variables, the two auxiliary equations will have identical R_i^2s leading to equal VIFs.

Thus the VIF is a method of detecting multicollinearity that takes into account all the explanatory variables at once. Some statistical software programs replace the VIF with its reciprocal, the **tolerance** which is $(1 - R_i^2)$ or TOL.

There are problems with using VIFs too. It is possible to have multicollinear effects in an equation that has no large VIFs. Foe example if the simple correlation coefficient between X_1 and X_2 is 0.88, multicollinear effects are likely but the VIF for the equation (with no other Xs) could be as low as 4.4. Also, there is no hard-and-fast decision rule for VIF. We mentioned a "critical value" of 5. In essence, VIF is sufficient but not necessary when testing for multicollinearity, just like the simple correlation coefficient.

8.4 Remedies for Multicollinearity

What can be done to minimize the consequences of severe multicollinearity? There is no one answer because multicollinearity is a phenomenon that could change from sample to sample even for the same specification of a regression equation. This section aims to outline a number of alternative remedies for multicollinearity that might be appropriate under certain circumstances.

8.4.1 Do Nothing

One reason for doing nothing is that multicollinearity in an equation will not always reduce the t-scores enough to make them insignificant or change the $\hat{\beta}$ enough to make them differ from expectations. In other words the mere existence of multicollinearity does not necessarily mean anything. A remedy for multicollinearity should only be considered if the consequences cause insignificant t-scores or unreliable estimated coefficients.

A second reason for doing nothing is tat the deletion of a multicollinear variable that belongs in an equation is fairly dangerous because it will cause specification bias. Dropping a variable will *purposely* create bias.

Finally, every time a regression is rerun, we risk encountering a specification that fits because it accidentally works for the particular data set involved, not because it is the truth. Only with larger experiments, we have a greater chance of finding the accidental result. In addition, when there is significant multicollinearity in the sample, the odds of strange results increase rapidly because of the sensitivity of the coefficient estimates to slight specification changes.

8.4.2 Drop a Redundant Variable

On occasions, the simple solution of dropping one of the multicollinear variables is a good one. For example,some researchers, not wanting to face omitted variable bias, include too many variables in their equation. As a result, they often have two or more variables in their equations that are essentially measuring the same thing. The variables might be called **redundant** where only one of them is needed to represent the effect on the dependent variable that all of them currently represent. Dropping such redundant multicollinear variables is doing nothing more than making up for a specification error; the variables should never have been included in the first place. Return to the student consumption example earlier:

$$\begin{aligned}
\widehat{CO_i} &= -367.83 + 0.5113Yd_i + 0.0427LA_i \\
& \qquad\qquad\qquad (1.0307) \qquad (0.0942) \\
t &= \qquad\qquad\quad +0.496 \qquad + 0.453 \\
\bar{R}^2 &= 0.835 \qquad N = 7
\end{aligned}$$

When we first discussed this example we compared this result with the same equation without the liquid assets variable:

$$\begin{aligned}\widehat{CO}_i &= -471.43 + 0.9714Yd_i \\ & \qquad\qquad\quad (0.157) \\ t &= \qquad\qquad\quad +6.187 \\ \bar{R}^2 &= 0.861 \qquad N = 7\end{aligned}$$

If we had instead dropped the disposable income variable, we would have obtained:

$$\begin{aligned}\widehat{CO}_i &= -199.44 + 0.08876LA_i \\ & \qquad\qquad\quad (0.01443) \\ t &= \qquad\qquad\quad +6.153 \\ \bar{R}^2 &= 0.860 \qquad N = 7\end{aligned} \tag{127}$$

Since dropping a variable changes the meaning of the remaining coefficient (because the dropped variable is not being held constant), such dramatic changes are not unusual. The coefficient of the remaining included variable measures almost all of the *joint* impact on the dependent variable of the multicollinear explanatory variables.

Assuming a multicollinear variable is dropped, which one do you choose? In the cases of severe multicollinearity, there is no statistical difference which variable is dropped. Hence, it does not make sense to pick the variable to be dropped on the basis of which one gives superior fit or which one is more significant (or has the expected sign) in the original equation. Instead the theoretical underpinnings of the model should be the basis for such a decision. In the student consumption example there is more theoretical support that disposable income determines consumption so (**??**) is a better option than (**??**).

8.4.3 Transform the Multicollinear Variables

On rare occasions the consequences of multicollinearity are serious enough to warrant the consideration of remedial action when the variables are extremely important on theoretical grounds. In these cases neither inaction nor dropping a variable is helpful. However it is sometimes possible to transform the variables in the equation to get rid of at least some multicollinearity. Some of the most common transformations include:

- form a combination of the multicollinear variables
- transform the equation into first differences

The technique of forming a **combination** of two or more variables consists of creating a new variable that is a function of the multicollinear variables and using the new variable to replace the old ones in the regression equation.

For example, if X_1 and X_2 are highly multicollinear, consider a new variable, X_3 where $X_3 = X_1 + X_2$ (or more generally, any linear combination such that $X_3' = k_1X_1 + k_2X_2$) might be substituted for a reestimation of the model. This technique is especially useful for forecasting since the multicollinearity outside the sample might not exist or might not follow the pattern that it did inside the sample.

A major disadvantage of the technique is that both portions of the combination variable are forced to have the same coefficient in the reestimated equation. For example if $X_{3i} = X_{1i} + X_{2i}$ then

$$Y_i = \beta_0 + \beta_3 X_{3i} + \epsilon_i = \beta_0 + \beta_3(X_{1i} + X_{2i}) + \epsilon_i \tag{128}$$

Care must be taken not to include (in a combination) variables with different expected coefficients (like signs) or dramatically different mean values (like in orders of magnitude) without adjusting for these differences by using appropriate constants in the more general equations.

The second kind of transformation is to consider to switch the functional form of the equation to first differences. A **first difference** is nothing more than the change in a variable from the previous time period to the current time period, otherwise referred to earlier as Δ. The first difference, mathematically, is defined as

$$\Delta X_t = X_t - X_{t-1}$$

If an equation is switched from its normal specification to a first difference specification, it is quite likely that the degree of multicollinearity will be significantly reduced for two reasons. First, any change in the definitions of the variables (except a linear change) will change the degree of multicollinearity. Second, multicollinearity takes place most frequently in time-series data, in which first differences are far less likely to move steadily upward than are the aggregates from which they are calculated. For example, although GDP might grow 3% to 4% year to year, the *change* in GDP could fluctuate severely.

Using first differences has one unexpected advantage, however. This involves the concept of a *non-stationary* time series, or a time series that has a significant trend of some sort. Evidence of multicollinearity in a time series is often evidence that a number of the the independent variables are nonstationary. By coincidence, one possible remedy for nonstationary variables is to convert to first differences, so first differences are possible remedies for *both* multicollinearity and nonstationarity.

8.4.4 Increase the Sample Size

The idea behind increasing the size of the sample is that a larger data set will allow more accurate estimates than a small one, since the large sample normal will reduce somewhat the variance of the estimated coefficients, diminishing the impact of the multicollinearity.

8.5 Choosing the Proper Remedy

The following are examples to illustrate general guidelines to follow when attempting to rid an equation of severe multicollinearity.

8.5.1 Why Multicollinearity Often Should Be Left Unadjusted

Consider that you work in the marketing department of a soft drink company and build a model of the impact of sales on advertising:

$$\begin{aligned} \hat{S}_t &= 3080 - 75000P_t + 4.23A_t - 1.04B_t \\ & \qquad\qquad (25000) \quad (1.06) \quad (0.51) \\ t &= \qquad\quad -3.00 \quad +3.99 \quad -2.04 \\ \bar{R}^2 &= 0.825 \qquad N = 28 \end{aligned} \tag{129}$$

where
S_t is the sales of the soft drink in year t
P_t is the average relative price of the drink in year t
A_t is the advertising expenditures for the company in year t
B_t is the advertising expenditures for the company's main competitor in year t

Assume that there are no omitted variables and variables are measured in real dollars.

Suppose, now that you were told that advertising in this industry is cutthroat and firms tend to match their main competitor's advertising expenditures.This would lead you to suspect that there *could* be significant multicollinearity between A_t and B_t. True enough, the simple correlation coefficient between the two variables is 0.974.

We can conclude that there is evidence that severe multicollinearity exists in the equation, but there is no reason to do anything about it. The coefficients are so powerful that their t-scores remain significant, even when severe multicollinearity exists. Unless multicollinearity in the equation causes problems in the equation, it should be left unadjusted.

When a variable is dropped from the equation, its effect will be absorbed by other explanatory variables to the extent they are correlated with the newly omitted variable. It is likely that the remaining multicollinear variable(s) will absorb virtually all the bias, since the variables are highly correlated. This bias may destroy whatever usefulness the estimates had before the variable was dropped.

For example, consider dropping B from the equation to fix the multicollinearity and the following might occur:

$$\begin{aligned} \hat{S}_t &= 2586 - 78000P_t + 0.52A_t \\ & \qquad\qquad (24000) \quad (4.32) \\ t &= \qquad\quad -3.25 \quad\ +0.12 \\ \bar{R^2} &= 0.531 \qquad N = 28 \end{aligned} \tag{130}$$

The company's advertising coefficient becomes less instead of more significant than when one of the multicollinear variables is dropped. To understand why, note that the expected bias on $\hat{\beta_A}$ is negative because the product of the expected sign of the coefficient of B and the correlation of A and B is negative:

$$\text{Bias} = \beta_B \cdot f(r_{A,B}) = (-) \cdot (+) = (-) \tag{131}$$

Second, this negative bias is strong enough to decrease the estimated coefficient of A until being insignificant. Although the problem could have been avoided by using a relative advertising variable, this formulation would have forced both A and $^1/_B$ to be identical. Most of the times, these kinds of constraints will force bias into and equation which previously does not have none.

8.6 A More Concrete Example of Dealing with Multicollinearity

Consider the model of annual demand for fish in the US from 1946 to 1970. Suppose that you decide to try to confirm the idea that the Pope's 1966 decision to allow Catholics to eat meat on non-Lent fridays caused a shift in the demand function for fish. Consider the hypothesized equation:

$$F_t = f(\overset{-}{PF_t}, \overset{+}{PB_t}, \overset{+}{Yd_t}, \overset{+}{N_t}, \overset{-}{P_t}) + \epsilon_t \tag{132}$$

where
F_t is the average pounds of fish consumed per capita in year t
PF_t the price index for fish in year t
PB_t the price index for beef in year t
Yd_t real per capita disposable income in year t (\$ 000,000,000)
N_t the number of Catholics in the United States in year t ('0000)
P_t a dummy variable equal to 1 after the Pope's 1966 decision and 0 otherwise

and you choose the following functional form

$$F_t = \beta_0 + \beta_1 PF_t + \beta_2 PB_t + \beta_3 \ln Yd_t + \beta_4 N_t + \beta_5 P_t + \epsilon_t \tag{133}$$

Note that the method chosen is an intercept dummy. Since you have stated that you expect this coefficient to be negative, the null hypothesis should be H$_0$: $\beta_5 \geq 0$ Second you have chosen a semilog function to relate the disposable income and the quantity of fish consumed; this is consistant with the theory that as income rises, the portion of that extra income devoted to the consumptin of fish to decrease. After collecting the data you obtain the following OLS estimates:

$$\begin{array}{llllll}
\hat{F_t} = -1.99 + 0.039PF_t & - 0.00077PB_t & + 1.77\ln Yd_t & - 0.0031N_t & - 0.355P_t \\
\qquad\qquad (0.031) & (0.02020) & (1.87) & (0.0033) & (0.353) \\
t = \qquad\quad +1.27 & -0.0384 & +0.945 & -0.958 & -1.01 \\
\bar{R^2} = 0.666 \quad N = 25 & & & &
\end{array} \tag{134}$$

This result is not encouraging. None of the estimated coefficients are significant. Also, three of the coefficients have unexpected signs. The problems could have been caused by omitted variables, irrelevant variables or multicollinearity. Supporting this decision is the value of $\bar{R^2}$ to be 0.666. This is high for the insignificant t-scores. One measure of severe multicollinearity is the size of the simple correlation coefficients. As expected the annual per capita disposable income Yd_t and number of Catholics are likely to be highly correlated and sure enough, the value of $r_{(\ln Yd,N)} = 0.946$.

In addition, it is not unreasonable to think that food prices might move together. Price of substitutes tend to move together. True enough, the simple correlation between the two price variables, or $r(PB, PF) = 0.958$. With multicollinearity of this severity between two variables with opposite signs, it is no surprise that the coefficients of both variables "switched signs". As multicollinearity increases, the distribution of $\hat{\beta}$ widens, and the probability of observing an unexpected sign increases.

The second sign of detecting multicollinearity, the size of the variance inflation factors or VIFs also indicate severe problems. All the VIFs for (**??**) except VIF$_P$ are above 5, an indicator of severe multicollinearity.

$$\begin{aligned} \text{VIF}_{PF} &= 43.4 \qquad \text{VIF}_{\ln Yd} = 23.3 \qquad \text{VIF}_{PB} = 18.5 \\ \text{VIF}_{N} &= 18.5 \qquad \text{VIF}_{P}U = r.4 \end{aligned}$$

So there appears to be significant multicollinearity in the model. Dropping N when choosing between Yd and N, the model becomes:

$$\begin{aligned} \hat{F_t} &= 7.96 + 0.03PF_t + 0.0047PB_t + 0.36\ln Yd_t - 0.12P_t \\ & \qquad\quad (0.03) \qquad (0.019) \qquad\quad (1.15) \qquad (0.26) \\ t &= \qquad\; +0.98 \qquad +0.24 \qquad\quad +0.31 \qquad -0.48 \\ \bar{R}^2 &= 0.667 \qquad N = 25 \end{aligned} \tag{135}$$

Dropping N clearly eliminated a redundant variable from (**??**) but (**??**) still has multicollinearity as measured by both detection techniques. The remaining multicollinearity lies in the price variables.

In the case of prices, we do not have the option of dropping one of the multicollinear variables because both PB and PF are too theoratically important in the model. What we could do, in this case is to transform the two price variables into a new relative price variable:

$$RP_t = \frac{PF_t}{PB_t}$$

Such a variable would make sense if theory called for keeping both variables in the equation and of the two coefficients could be expected to be close in absolute but of opposite signs. Suppose you hypothesize

$$F_t = f(\overset{-}{RP_t}, \overset{+}{Yd_t}, \overset{-}{P_t}) + \epsilon_t$$

and obtain

$$\begin{array}{llll} \hat{F_t} = -5.17 - 1.93RP_t & + 2.71 \ln Yd_t & + 0.0052P_t & \\ \qquad\qquad\quad (1.43) & (0.66) & (0.2801) & \\ t = \qquad\qquad -1.35 & + 4.13 & + 0.019 & \\ \bar{R^2} = 0.588 \qquad N = 25 & & & \end{array} \qquad (136)$$

Although these are all questions of judgement, the two changes appear to work reasonably in ridding the equation from much of its multicollinearity. The VIFs now are below 3. We can now test the hypothesis that the Pope's decision caused a significant change in the demand for fish. We cannot reject the null hypothesis and conclude that the Pope's decision did not cause a cut down in the consumption of fish.

Finally, notice that someone else might take a completely different form. There is no correct remedy. Indeed, if you want to be sure that your choice of specification did not influence your ability to reject the null hypothesis,you might see how sensitive that conclusion is to an alternative approach to fixing the multicollinearity.

9 Serial Correlation

9.1 Pure versus Impure Serial Correlation

9.1.1 Pure 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:

$$\mathbb{E}(r_{\epsilon_i \epsilon_j}) = 0 \qquad (i \neq j)$$

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.

The most commonly 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:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \tag{137}$$

where:
ϵ is the error term of the equation in question
ρ is the parameter depicting the functional relationship between observations of the error term
u is a classical (not serially correlated) error term

The function form in (**??**) is called a first-order Markov scheme, and the new symbol, ρ is called the **first-order auto-correlation coefficient**. For this kind of serial correlation, all that is needed is for the value of one observation of the error term to directly affect the value of the next observation of the error term.

The magnitude of ρ indicates the strength of the serial correlation in an equation. If $\rho = 0$, then there is no serial correlation. As $|\rho| \to 1$ the value of the previous observation of the error term becomes more important in determining the current value of ϵ_t, and a high degree of serial correlation exists. ρ cannot be greater than one because it means the error term will continually increase in absolute value over time. Hence, we can state that

$$-1 < \rho < 1 \tag{138}$$

The sign of ρ indicates the nature of the serial correlation in an equation. A positive value for ρ implies that the error term tends to have the same sign from one time period to the next; this is called **positive serial correlation**. Such a tendency means if ϵ_t happens by chance to take on a large value in one time period, subsequent observations would tend to retain a portion of this original, large value and would have the same sign as the original. For example in time-series models, a large external shock to an economy in one period may linger on for several time periods.

A negative value for ρ implies that the error term has a tendency to switch signs from negative to positive and back again in consecutive observations; this is called **negative serial correlation**. It implies that there is some sort of cycle behind the drawing of stochastic differences. Negative serial correlation might exist in the error term of an equation that is in first differences because *changes* happen in a cyclical pattern.

Serial correlation can take on many forms other than first-order serial correlation. For example, in a quarterly model, the current quarter's error term observation many be functionally related to the observation of the error term from the same quarter in the previous year. This is called seasonally based serial correlation:

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

Similarly, it is possible that the error term in an equation might be a function of more than one previous observation of the error term:

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

Such a formulation is called *second-order* serial correlation. Higher-order expressions are similarly formed, but the justifications for assuming these higher-order forms are usually weaker than the justification for the second-order form, which itself is not always all that strong.

9.1.2 Impure Serial Correlation

By **impure serial correlation** we mean serial 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.

Recall that the error term might be the effect of omitted variables, nonlinearities and pure stochastic disturbances on the dependent variable. This means, for example that if we omit a relevant variable or use the wrong functional form,the the portion of that omitted effect that cannot be explained by the explanatory variables must be absorbed by the error term. This new error term might be serial correlated even if the true one is not. Of course, in this case the serial correlation is due to the researcher's choice of specification and not by the pure error term associated with the correct specification.

Let us see how an omitted variable can cause the error term to be serially correlated. Suppose that the true equation is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t$$

where ϵ_t is a classical error term. As seen in Section 6 if X_2 is accidentally omitted from the equation then

$$Y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t^* \qquad \text{where} \qquad \epsilon_t^* = \beta_2 X_{2t} + \epsilon_t \tag{139}$$

Thus the error term being used in the omitted variable case is not the classical error term ϵ but a function of the independent variable X_2. As a result, the new error term ϵ^* can be serially correlated even if the true error term ϵ is not. In particular, the new error term will tend to be serially correlated when:

1. X_2 itself is serially correlated ***and***
2. the size of ϵ is small compared to the size of $\beta_2 \bar{X}_2$

These tendencies hold even if there are a number of included and/or omitted variable.

Note that while the error term ϵ^* appears to have a non-zero mean, this will not actually occur since the OLS estimate of the constant term, $\hat{\beta}_0^*$, will adjust to offset this problem. Second, since the impure serial correlation implies a specification error such as an omitted variable, impure serial correlation is likely to be associated with biased coefficient estimates. Both the bias and the impure serial correlation will disappear if the specification error is corrected.

An example of how an omitted variable might cause serial correlation in the error term in an incorrectly-specified equation is as follows. Consider the fish-demand equation:

$$F_t = \beta_0 - \beta_1 RP_t + \beta_2 \ln Yd_t + \beta_3 D_t + \epsilon_t \tag{140}$$

where F_t is the average pounds of fish consumed per capita in year t
PF_t the price index for fish in year t
Yd_t real per capita disposable income in year t ($ 000,000,000)
D_t a dummy variable equal to 1 after the Pope's 1966 decision and 0 otherwise
ϵ_t is a classical (not serially correlated) error term

Assume that (**??**) is the "correct" specification. What would happen to this equation if Yd were omitted?

$$F_t = \beta_0 + \beta_1 RP_t + \beta_3 D_t + \epsilon_t^* \tag{141}$$

The first effect is that the estimated coefficients of RP and D would be biased, depending on the correlation of RP and D with Yd. The second effect is that the error term would now include a large portion of the left out effect of disposable income on the consumption of fish. That is, ϵ_t^* would be a function of $\epsilon_t + \beta_2 \ln Yd_t$. It is reasonable to expect that disposable income might follow a fairly serially correlated pattern:

$$\ln Yd_t = f(\ln Yd_{t-1}) + u_t \tag{142}$$

Because of the continual rise of disposable income over time makes it (and its log) act in a serially correlated or autoregressive manner. Since disposable income is serially correlated, then ϵ^* is likely to also be serially correlated, which can be expressed as:

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

where ρ is the coefficient of serial correlation and u is a classical error term. Hence, it is indeed possible for an omitted variable to introduce "impure" serial correlation into an equation.

Another common kind of impure serial correlation is caused by an incorrect functional form. Here, the choice of the wrong functional form can cause the error term to be serially correlated. Say the true equation is polynomial in nature:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{1t}^2 + \epsilon_t \tag{143}$$

but a linear regression was run

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \epsilon_t^* \tag{144}$$

The new error term ϵ^* is now a function of the true error term ϵ and of the differences between the linear and the polynomial functional forms. These differences often follow fairly autoregressive patterns, that is positive differences tend to be followed by positive differences and negative differences tend to be followed by negative differences. As a result, using a linear functional form when a nonlinear one is appropriate will usually result in positive impure serial correlation.

9.2 The Consequences of Serial Correlation

The three major consequences of serial correlation are:

1. Pure serial correlation does not cause bias in the coefficient estimates
2. Serial correlation causes OLS to no longer be the minimum variance unbiased estimator
3. Serial correlation causes toe OLS estimates of the $\mathbb{SE}(\hat{\beta})$s to be biased, leading to unreliable hypothesis testing.

Explaining the consequences, we shall focus purely on first-order serial correlation.

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.*** Recall that the most important property of the OLS estimation technique is that it is minimum variance for the class of linear unbiased estimators. If the errors are serially correlated, onf of the assumptions of the Gauss-Markov Theorem is violated but this violation does not cause the coefficient estimates to be biased. This conclusion does not depend on whether the serial correlation is positive or negative or first order. If the serial correlation is impure, however, bias may be introduced by the use of an incorrect specification.
   This lack of bias does not necessarily mean that the OLS estimates of the coefficients of a serially correlated equation will be close to the true

coefficient values; the single estimate observed in practice can come from a wide range of possible values. In addition, the standard errors of these estimates will typically be increased by the serial correlation This increase will raise the probability that $\hat{\beta}$ will differ significantly from the true value of β.

2. ***Serial correlation causes OLS to no longer be the minimum variance estimator (of all unbiased estimators.)*** Although the violation of Classical Assumption IV causes no bias, it does affect the other main conclusion of the Gauss-Markov Theorem, that is of minimum variance. In particular, we cannot prove that the distribution of the OLS $\hat{\beta}$s is minimum variance of all linear unbiased estimators when Assumption IV is violated. As a result, if the error term is serially correlated, then OLS no longer provides minimum variance linear unbiased estimates of the coefficients.
   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. Thus, OLS is more likely to mis-estimate the true value of β in the face of serial correlation. On balance, the $\hat{\beta}$s are still unbiased because overestimates are just likely as underestimates, but these errors increase the variance of the distribution of the estimates, increasing the amount that any given estimate is likely to differ from the true β.

3. ***Serial correlation causes the OLS estimates of the*** $\mathbb{SE}(\hat{\beta})$***s to be biased, leading ot unreliable hypothesis testing.*** With serial correlation, the OLS formula for the standard error produces biased estimates of the $\mathbb{SE}(\hat{\beta})$s. Because the $\mathbb{SE}(\hat{\beta})$ is a prime component of the t-statistic, these biased $\mathbb{SE}(\hat{\beta})$s cause biased t-scores and unreliable hypothesis testing in general. In essence, serial correlation causes OLS to produce incorrect $\mathbb{SE}(\hat{\beta})$s and t-scores!
   Typically, the bias in the estimate of$\mathbb{SE}(\hat{\beta})$ is negative, meaning that OLS underestimates the size of the standard errors of the coefficients. This is because serial correlation usually results in a pattern of observations that allows a better fit than the actual not serially correlated observations would other justify. The tendency of OLS to underestimate the $\mathbb{SE}(\hat{\beta})$ means that OLS typically overestimates the$t$-scores of the estimated coefficients since

$$t = \frac{(\hat{\beta} - \beta_{H_0})}{\mathbb{SE}(\hat{\beta})} \tag{145}$$

Thus the t-scores b a software regression package in the face of serial correlation is likely to be too high.
If OLS underestimates the $\mathbb{SE}(\hat{\beta})$s and therefore overestimates the t-scores, it will cause a t-score to be "too high" and it is more likely we will reject the null hypothesis (usually H_0: $\beta \leq 0$) when it is in fact, true. This means we are more likely to make a Type I Error, and more likely to keep an irrelevant variable in an equation. This, in conclusion, makes hypothesis testing biased and unreliable, with the presence of serial correlation.

9.3 The Durbin-Watson *d*-Test

Although the first step in detecting serial correlation often is observing a pattern, the test for serial correlation that is widely used today is the Durban-Watson d-test.

9.3.1 The Durbin-Watson *d* Statistic

The **Durbin-Watson *d* statistic** is 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 Durban-Watson d statistic only when the assumptions that underlie its deviation are met:

1. The regression model includes an intercept term
2. The serial correlation is first-order in nature:

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

   where ρ is the coefficient of serial correlation and u is a classical, normally distributed error term

3. The regression model does not include a lagged dependent variable as an independent variable

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

$$d = \frac{\sum_2^T (e_t - e_{t-1})^2}{\sum_1^T e_t^2} \tag{146}$$

where the e_ts are the OLS residuals. Not that the numerator has one fewer observation than the denominator because an observation is needed to calculate e_{t-1}. The Durbin-Watson d statistic equal zero if there is extreme serial correlation, two if there is no serial correlation and four if there is negative correlation. To see this, consider back (**??**):

- If there is extreme serial positive correlation, $e_t = e_{t-1}$ and so $e_t - e_{t-1} = 0$ and so $d = 0$
- If there is extreme negative serial correlation, $e_t = -e_{t-1}$ and so $e_t - e_{t-1} = 2e_t$. Substituting to (**??**), $^{d=\sum(2e_t)^2}/_{\sum(e_t)^2} \approx 4$
- If there is no serial correlation, the mean of the distribution of d is 2.

9.3.2 Using the Durbin-Watson *d* Test

The Durbin-Watson d test is unusual in two respects. First, econometricians almost never test the one-sided null hypothesis that there is negative serial correlation in the residuals because negative serial correlation, as mentioned is quite difficult to explain theoretically in economic or business analysis. Its existence often means that impure serial correlation has been caused by some error of specification.

Second, the Durbin-Watson test is sometimes inconclusive. This test has an "inconclusive region". Hence, for some cases, there is no conclusion to this case.

The use of the d test is quite similar to the t-test and F-test. To test for positive serial correlation, perform the following steps:

1. Obtain the OLS residuals from the equation to be tested and calculate the d statistic.
2. Determine the sample size and the nuber of explanatory variables, find the upper critical d value, d_U and the lower critical d value, d_L, respectively.
3. Given the null hypothesis of no positive serial correlation and a one-sided alternative hypothesis:

$$\begin{aligned} &\mathrm{H_0}\colon\ \rho \leq 0 \quad \text{(no positive serial correlation)} \\ &\mathrm{H_A}\colon\ \rho > 0 \quad \text{(positive serial correlation)} \end{aligned} \tag{147}$$

The appropriate decision rule is

$$\begin{array}{rl} \text{if } d < d_L \text{ then} & \text{Reject H}_0 \\ \text{if } d > d_U \text{ then} & \text{Do Not Reject H}_0 \\ \text{if } d_L \leq d \leq D_U \text{ then} & \text{Test is Inconclusive} \end{array}$$

For a two-sided d test, the third step becomes:

Given the null hypothesis of no serial correlation and two-sided alternative hypothesis:

$$\begin{array}{rl} \text{H}_0\text{: } \rho = 0 & \text{(no serial correlation)} \\ \text{H}_\text{A}\text{: } \rho \neq 0 & \text{(serial correlation)} \end{array} \tag{148}$$

The appropriate decision rule becomes

$$\begin{array}{rl} \text{if } d < d_L \text{ or if } (4 - d_L) < d \text{ then} & \text{Reject H}_0 \\ \text{if } d_U < d < (4 - d_U) \text{ then} & \text{Do Not Reject H}_0 \\ \text{otherwise then} & \text{Test is Inconclusive} \end{array}$$

9.3.3 Examples of the Use of the Durbin-Watson *d* Statistic

Consider setting up a one-sided 5% test for a regression with three explanatory variables and sample size 25. The critical values are, with $k' = 3$ and $N = 25$, $d_U = 1.654$ and $d_L = 1.123$. So if the hypotheses are:

$$\begin{array}{rl} \text{H}_0\text{: } \rho \leq 0 & \text{(no positive serial correlation)} \\ \text{H}_\text{A}\text{: } \rho > 0 & \text{(positive serial correlation)} \end{array} \tag{149}$$

then the appropriate decision rule becomes

$$\begin{array}{rl} \text{if } d < 1.123 \text{ then} & \text{Reject H}_0 \\ \text{if } d > 1.654 \text{ then} & \text{Do Not Reject H}_0 \\ \text{if } 1.123 \leq d \leq 1.654 \text{ then} & \text{Test is Inconclusive} \end{array}$$

Given a computed d statistic of

- 1.78, we do not reject H_0. There is insufficient evidence of positive serial correlation

- 1.28, the test is inconclusive
- 0.60, we reject H_0. There is sufficient evidence of positive serial correlation

Consider the chicken demand model of Chapter 6. The Durbin-Watson d statistic is 0.90. We want to see if there is any serial correlation. Given that $k' = 2$ and $N = 40$, the appropriate critical values are $d_L = 1.338$ and $d_U = 1.659$. Hence, the decision criteria is, using a one-tailed test:

$$\begin{aligned} \text{if } d < 1.338 \text{ then} \quad & \text{Reject H}_0 \\ \text{if } d > 1.659 \text{ then} \quad & \text{Do Not Reject H}_0 \\ \text{if } 1.338 \leq d \leq 1.659 \text{ then} \quad & \text{Test is Inconclusive} \end{aligned}$$

So we reject H_0. There sufficient evidence to show that there is positive serial correlation, and now have to decide how to deal with the serial correlation. Of course, it has been dealt with in the example.

9.4 Remedies for Serial Correlation

Say the Durbin-Watson d statistic detects serial correlation in the residuals of the equation. The place to start in correcting a serial correlation problem is to look carefully at the specification of the equation for possible errors that might cause the impure serial correlation – functional form, omission of errors or other specification errors. Are there specification errors that might have some pattern over time that could have introduced impure serial correlation into the residuals? Only after the specification of the equation has been reviewed carefully should the possibility of an adjustment for pure serial correlation be considered.

A significant Durbin-Watson statistic can easily be caused by an omitted variable or an incorrect functional form. In such circumstances, the Durbin-Watson test does not distinguish between pure and impure serial correlation, but the detection of negative serial correlation is often a strong hint that the serial correlation is impure.

If you conclude that you have pure serial correlation, then the appropriate response is to consider applying the Generalized Least Squares or Newey-West standard errors.

9.4.1 Generalized Least Squares

Generalized least squares (GLS) is a method of ridding an equation of pure first-order serial correlation and in the process restoring the minimum variance property to its estimation. GLS starts with an equation that does not meet the Classical Assumptions and transforms it into one that does meet those assumptions. Start with an equation that has first-order serial correlation:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t \tag{150}$$

where $\epsilon_t = \rho(\epsilon_{t-1}) + u_t$ so (**??**) can be rewritten as

$$Y_t = \beta_0 + \beta_1 X_{1t} + \rho\epsilon_{t-1} + u_t \tag{151}$$

where ϵ is the serially correlated error term, ρ is the coefficient of serial correlation and u is a classical (not serially correlated) error term.
If we could get the $\rho\epsilon_{t-1}$ term out of (**??**) then the serial correlation would be gone. To do so, multiply (**??**) by ρ and lag the new equation by one time period, obtaining:

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{1t-1} + \rho\epsilon_{t-1} \tag{152}$$

Subtract (**??**) from (**??**) to remove the serially correlated component of the error term

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + u_t \tag{153}$$

and (**??**) can be re-written as

$$Y_t^* = \beta_0^* + \beta_1 X_{1t}^* + u_t \tag{154}$$

where

$$\begin{aligned} Y_t^* &= Y_t - \rho Y_{t-1} \\ \beta_0^* &= \beta_0(1-\rho) \\ X_{1t}^* &= (X_{1t} - \rho X_{1t-1}) \end{aligned} \tag{155}$$

(**??**) is called a Generalized Least Squares (or "quasi-differenced") version of (**??**). Notice that

- The error term is not serially correlated. Hence, OLS estimation of (**??**) will be minimum variance.

- The slope coefficient β_1 is the same as the slope coefficient of the original serially correlated equation, (**??**). Hence, GLS estimates have the same economic meaning as OLS estimates.

- The dependent variable has changed compared to that in equation (**??**). This means that the GLS $\bar{R}^2$ is not directly comparable to OLS $\bar{R}^2$.

- To forecast with GLS, adjustments need to be done.

Unfortunately, we cannot use OLS to estimate a Generalized Least Squares model because GLS equations are inherently nonlinear in the coefficients. Looking at (**??**), we need to estimate values for not only β_0 and β_1, but also ρ, the coefficient of serial correlation, and ρ is multiplied with β_0 and β_1. Since OLS requires that the equation be liner in the coefficients, we need a different estimation procedure.

The best known way to estimate GLS equations is the **Cochrane-Orcutt method**, a two-step iterate technique. The first produces an estimate of ρ and then estimates the GLS equation using that $\bar{\rho}$. The two steps are:

- Estimate ρ by running a regression based on the residuals of the equation suspected of having serial correlation:

  $$e_t = \rho\epsilon_{t-1} + u_t \tag{156}$$

  where the e_ts are the OLS residuals from the equation suspected of having pure serial correlation and u_t is a classical error term

- Use this $\hat{\rho}$ to estimate the GLS equation by substituting $\hat{\rho}$ into (**??**) and using OLS to estimate (**??**) with the adjusted data.

These two steps are repeated (iterated) until further iteration results in little change in $\hat{\rho}$. Once $\hat{\rho}$ has converged, the last estimate of Step 2 is used as a final estimate of (**??**).

As popular as Cochrane-Orcutt is, there is a different, suggested method for GLS models. The **AR(1) method** estimates a GLS equation like (**??**) by estimating β_0, β_1 and ρ simultaneously with iterative nonlinear regression techniques. The AR(1) method tends to produce the same coefficient estimates as Cochrane-Orcutt but with superior estimates of the standard errors.

Consider applying GLS using AR(1) estimation method to the chicken demand example that was found to have positive serial correlation. Recall that estimated the per capita demand for chicken as a function of the price of chicken, price of beef and disposable income:

$$\begin{aligned} \hat{Y}_t &= 27.6 - 0.61PC_t + 0.09PB_t + 0.24YD_t \\ & \qquad\qquad (0.16) \qquad (0.04) \qquad (0.011) \\ t &= \qquad\quad -3.86 \qquad +2.31 \qquad +22.07 \\ \bar{R^2} &= 0.990 \qquad N = 40 \text{ (Annual 1960 -1999)} \\ \text{DW } d &= 0.90 \end{aligned} \tag{157}$$

If we reestimate (**??**) with the AR(1) approach to GLS we obtain

$$\begin{aligned} \hat{Y}_t &= 23.5 - 0.09PC_t + 0.09PB_t + 0.24YD_t \\ & \qquad\qquad (0.10) \qquad (0.05) \qquad (0.02) \\ t &= \qquad\quad -0.89 \qquad +1.95 \qquad +15.33 \\ \bar{R^2} &= 0.995 \qquad N = 39 \text{ (Annual 1960 -1999)} \\ \hat{\rho} &= 0.80 \end{aligned} \tag{158}$$

Comparing (**??**) and (**??**), note that $\hat{\rho}$ used in equation (**??**) is 0.80. This means Y was run as $Y^{=}Y_t - 0.80Y_{t-1}$, PC becomes $PC* = (1 - 0.80)PC$ etc. Also, $\hat{\rho}$ replaces DW in the documentation of the GLS estimates in part because of the DW of equation (**??**) is not strictly comparable to non-GLS DW. Finally, the sample size dropped to 39 because the first observation has to be used to create the lagged values for the calculation of the quasi-differenced variables in (**??**).

Generalized Least Squares Estimates, no matter how produced, have at least two problems. First, even though serial correlation causes no bias in the estimates of $\hat{\beta}$s, the GLS estimates usualy are different from the OLS ones. For example, $\hat{\beta_0}$ and β_{PC} changed as we moved from OLS in (**??**) to GLS in (**??**). Second, it turns out that GLS works well if $\hat{\rho}$ is close to the actual ρ, but the GLS $\hat{\rho}$ is biased in small samples. If $\hat{\rho}$ is biased, then the biased $\hat{\rho}$ introduces bias into the GLS estimates of the $\hat{\beta}$s. Luckily, there is a remedy for serial correlation that avoids both of these problems: Newey-West standard errors.

9.4.2 Newey-West Standard Errors

Not all corrections for pure serial correlation involve Generalized Least Squares. **Newey-West standard errors** are $\mathbb{SE}(\hat{\beta})$s that take account of serial correlation without changing the $\hat{\beta}$s themselves in any way. The logic behind the Newey-West standard errors is powerful. If serial correlation does not cause bias in the $\hat{\beta}$ but does impact the standard errors, then it makes sense to adjust the estimated equation in a way that changes the $\mathbb{SE}(\hat{\beta})$s but not the $\hat{\beta}$.

The Newey-West standard errors have been calculated specifically to avoid the consequences of pure first-order serial correlation. The Newey-West procedure yields an estimator of the standard errors that, while they are biased, is generally more accurate than uncorrected standard errors for large samples with a possibility of serial correlation. As a result, Newey-West can be used for t-tests and other hypothesis tests in most samples without the errors of inference potentially caused by serial correlation. Typically, Newey-West $\mathbb{SE}(\hat{\beta})$s are larger than OLS $\mathbb{SE}(\hat{\beta})$s, thus producing lower t-scores and decreasing the probability that a given estimated coefficient is significantly different from zero. Applying the Newey-West standard errors to the chicken demand equation in (**??**), we get

$$\begin{aligned} \hat{Y}_t &= 27.6 - 0.61PC_t + 0.09PB_t + 0.24YD_t \\ & \qquad\qquad (0.167) \qquad (0.042) \qquad (0.010) \\ t &= \qquad\qquad -3.64 \qquad +2.20 \qquad +24.65 \\ \bar{R}^2 &= 0.990 \qquad N = 40 \text{ (Annual 1960 -1999)} \\ \text{DW } d &= 0.90 \end{aligned} \tag{159}$$

Comparing (**??**) and (**??**), the $\hat{\beta}$s are identical as Newey-West standard errors do not change the OLS $\hat{\beta}$s. As expected, also the Newey-West standard errors are larger than the OLS $\mathbb{SE}(\hat{\beta})$s for two of three $\hat{\beta}$s, producing lower t-scores for those coefficients. Heteroskedasticity is the violation of Classical Assumption V, which states that the observations of the error term are drawn from a distribution with a constant variance. The assumption of constant variances for different observations of the error term is not always realistic. As we will see, the distinction between heteroskedasticity and homoskedasticity is important because OLS, when applied to heteroskedastic models, is no longer the minimum variance estimator.

10 Heteroskedasticity

10.1 Pure versus Impure Heteroskedasticity

Similar to serial correlation, there is pure and impure heteroskedasticity. Pure heteroskedasticity is caused by the error term of the correctly specified equation while impure heteroskedasticity is caused by a specification error like an omitted variable.

10.1.1 Pure Heteroskedasticity

Pure heteroskedasticity refers to heteroskedasticity that is a function of the error term of a correctly specified equation. Such **pure heteroskedasticity** occurs when Classical Assumption V, which assumes that the error term is constant, is violated in a correctly specified equation. Recall that Assumption V assumes that:

$$\mathbb{VAR}(\epsilon_i) = \sigma^2 = k \qquad \text{where } k \text{ is a constant} \quad (i = 1, 2, \cdots, N) \tag{160}$$

If this assumption is met, all the observations of the error term can be thought as being drawn from the same distribution: a distribution with mean zero and variance σ^2. This σ^2 does not change for different observations of the error term and this is called homoskedasticity.

With heteroskedasticity, this error term variance is not constant. Instead the variance of the distribution of the error term depends on exactly which observation is being discussed:

$$\mathbb{VAR}(\epsilon_i) = \sigma_i^2 \qquad (i = 1, 2, \cdots, N) \tag{161}$$

Heteroskedasticity often occurs in data sets in which there is a wide disparity between the largest and smallest observed value of the dependent variable. The larger the disparity between the size of observations of the dependant variable in a sample, the larger the likelihood that the error term observations associated with them have different variances and therefore be heteroskedastic.

In cross sectional data sets, it is easy to get such a large range between the largest and smallest values of the variables. Recall that in cross-sectional models, the observations are from the same time period but of different entities. Since cross-sectional models often include observations of widely different sizes of the same sample, heteroskedasticity is hard to avoid if economic topics are to be studied cross sectionally.

The simplest way to visualize pure heteroskedasticity is to picture a world in which the observations of the error term could be grouped into two different distributions – the ‘wide’ and ‘narrow’ distribution. This simple version is called *discrete heteroskedasticity.* In this model both distributions are centred around zero but one would have a larger variance than the other.

Heteroskedasticity takes on many more forms but we shall focus on the most frequently specified model of pure heteroskedasticity. In this model, the variance of the error term is related to an exogenous variable Z_i. For a typical regression equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{162}$$

the variance of the otherwise classical error term, ϵ might be equal to

$$\mathbb{VAR}(\epsilon_i) = \sigma^2 Z_i^2 \tag{163}$$

where Z might or might not be one of the independent variables in the equation. The variable Z is called a **proportionality factor** because the variance of the error term changes proportionally to the square of Z_i. The higher the value of Z_i, the higher the variance of the distribution of the i-th observation of the error term. There would be N different distributions, one for each observation, from which the observations of the error term could be drawn depending on the number of different values that Z takes.

This model could occur in

- the study of population and household consumption (variances vary greatly across states due to the different size of the population in each state);
- sales of DVD players from 1980 to 2005 (as the years pass, the more DVD players buy and the variance term also increased greatly)

10.1.2 Impure Heteroskedasticity

Heteroskedasticity caused by an error in specification, such as an omitted variable is called **impure heteroskedasticity**. Although improper functional form is less likely to cause impure heteroskedasticity, than it is to cause serial correlation, the two concepts are similar in most other ways.

An omitted variable can cause a heteroskedastic error term because the portion of the omitted effect not represented by one of the included explanatory variables must be absorbed by the error term. If this effect has a heteroskedastic component, the error term of the misspecified equation might be heteroskedastic even if the error term of the true equation is not. This distinction is important because with impure heteroskedasticity the correct remedy is to find and omitted variable and include it into the regression. It is therefore important to be sure that the specification is correct before attempting to detect or remedy pure heteroskedasticity.

10.2 The Consequences of Heteroskedasticity

If the error term of your equation is known to be heteroskedastic, there are three major consequences:

1. ***Pure heteroskedasticity does not cause bias in the coefficient estimates.*** If the error term of an equation is known to be purely heteroskedastic, that heteroskedasticity will not cause bias in the OLS estimates of the coefficients. Even though large positive errors are more likely, so to are large negative errors. The two tend to average out and leave the OLS estimator to be unbiased.
   As a result, we can say that an otherwise correctly specified equation that has pure heteroskedasticity still has the property that
   $\mathbb{E}(\hat{\beta}) = \beta \, \forall \, \beta$s.
   Lack of bias does not guarantee "accurate" coefficient estimates, especially since heteroskedasticity increases the variance of the estimates, but the distribution of the estimates is still centered around the true β. Equations with impure heteroskedasticity caused by an omitted variable will have possible specification bias.

2. ***Heteroskedasticity typically causes OLS to no longer be the minimum variance estimator.*** Pure heteroskedasticity causes no bias in the estimates of the OLS coefficients, but it does affect the minimum variance property. If the error term of an equation is heteroskedastic with respect to a proportionality factor Z:

   $$\mathbb{VAR}(\epsilon_i) = \sigma^2 Z_i^2 \tag{164}$$

   then the minimum variance portion of the Gauss-Markov Theorem cannot be proven because there are other linear unbiased estimators that have smaller variances. This is because the heteroskedastic error term causes the dependent variable to fluctuate, and the OLS estimation procedure attributes this fluctuation to the independent variables. Thus, OLS is more likely to misestimate the true β in the face of heteroskedasticity. On balance, the $\hat{\beta}$s are still unbiased because overestimates are just as likely as underestimates.

3. ***Heteroskedasticity causes the OLS estimates of the*** $\mathbb{SE}(\hat{\beta})s$ ***to be biased, leading to unreliable hypothesis testing.*** With heteroskedasticity, the OLS formula for the standard error produces biased estimates of the $\mathbb{SE}(\hat{\beta})s$. Because the $\mathbb{SE}(\hat{\beta})$ is a prime component of the t-statistic, these biased $\mathbb{SE}(\hat{\beta})s$ caused biased t-scores and unreliable hypothesis testing in general.
   What sort of bias does heteroskedasticity tend to cause? Typically, the bias in the estimate of $\mathbb{SE}(\hat{\beta})$ is negative, meaning that OLS underestimates the size of the standard errors of the coefficients. This comes about because heteroskedasticity usually results in a pattern of observations that allows a better fit than the actual homoskedastic observations would otherwise justify. This tendency of OLS to underestimate the $\mathbb{SE}(\hat{\beta})$ means that OLS typically overestimates the t-scores of the estimated coefficients.
   Since OLS underestimates the $\mathbb{SE}(\hat{\beta})$ and hence, overestimates the t-score, we are more likely to reject the null hypothesis when it is in fact, true. We are more likely to commit a Type I Error, and we are more likely to make the mistake of keeping an irrelevant variable in the equation because its coefficient's t-score has been overestimated. Hypothesis testing becomes both biased and unreliable.

10.3 Testing for Heteroskedasticity

Econometricians do not all use the same test for heteroskedasticity because heteroskedasticity takes a number of different forms, and its precise form in a given equation is almost never known. the "Z_i proportionality factor" approach of this chapter is merely one of the many specifications of the form of heteroskedasticity. Hence, there is no universally agreed upon method of testing for heteroskedasticity – basic textbooks list up to eight different methods for testing.

We shall describe the two different tests for heteroskedasticity. One is the *Park test* because it tests the proportionality factor form and the *White test* which is more generally used than the Park test. No test for heteroskedasticity can "prove" that heteroskedasticity exists in an equation, though so the best we can do is to get a general indication of this likelihood.

There is no need to run a heteroskedasticity test for every specification estimated, however, so before using any test, it is a good idea to ask the following preliminary questions

1. Are there any obvious specification errors? If it is suspected to have an omitted variable then a test for heteroskedasticity should be delayed until the specification is as good as possible
2. Is the subject of the research often afflicted with heteroskedasticity? Cross-sectional studies with large variances are very susceptible to heteroskedasticity, for example.
3. Does a graph of the residuals show any evidence of heteroskedasticity? By plotting the residuals the graphs can show without a test that heteroskedasticity is or is not likely.

10.3.1 The Park Test

The **Park test** is a formal procedure that attempts to test the residuals for this heteroskedasticity in a manner similar to the way that the Durban-Watson d statistic tests are used for serial correlation. The Park test has three basic steps. First, the regression equation is estimated by OLS and the residuals are calculated. Second, the log of the squared residuals is used as the dependent variable of an equation whose sole explanatory variable is the log of the proportionality factor Z. Finally, the results of the second regression are tested to see if there is any evidence of heteroskedasticity.

If there is reason to suspect heteroskedasticity, it is appropriate to run a Park test. The steps are:

1. **Obtain the residuals of the estimated regression equation.** The first step is to estimate the equation with OLS and then find the residuals from their estimation:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}) \tag{165}$$

   These residuals are the same ones used to calculate the Durban-Watson d statistic.

2. **Use these residuals to form the dependent variable in a second regression.** In particular the Park test suggests that you run the following the double-log regression:

$$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Z_i + u_i \tag{166}$$

   where
   e_i is the residual form fro the i-th observation from (**??**)
   Z_i is your best choice for the proportionality factor
   u_i is a classical homoskedastic error term

3. **Test the significance of the coefficient of Z in (??) with a t-test.** Use the t-statistic to test the significance of $\ln Z$ in explaining $\ln e^2$ in (**??**). If the coefficient of Z is significantly from zero, then there is evidence of heteroskedastic patterns in the residuals with respect to Z; otherwise, heteroskedasticity related to this particular value of Z is not support by the evidence in these residuals.

The Park test is not always easy to use. Its major problem is the identification of the proportionality factor Z. Although Z is often an explanatory variable in the original regression equation, there is no guarantee of that. A particular Z should be chosen for your Park test only after investigating the type of potential heteroskedasticity in your equation. A good Z is a variable that seems likely to vary with the variance of the error term.

In a cross-sectional model of countries and states, a good Z would be one that measured the size of the observation relative to the dependent variable in the equation. For a dependent variable like gallons of gasoline consumed, the number of drivers is better than the whole population.

10.3.2 An Example of the use of the Park Test

Consider the Woody's Restaurants example again. The regression equation explained the number of customers, as measured by check volume Y at a cross section of 33 different Woody's restaurants as a function of the number of nearby competitors N, the nearby population Y and the average household income of the local area I.

$$\begin{array}{llll} \hat{Y}_i = 102192 - 9075N_i & + 0.3547P_i & + 1.288I_i \\ \qquad\qquad (2053) & (0.0727) & (0.543) \\ t = \qquad\quad -4.42 & 4.88 & 2.37 \\ N = 33 \qquad \bar{R}^2 = 0.579 & & \end{array} \tag{167}$$

The equation is cross sectional so the heteroskedasticity is a theoretical possibility. However, the dependent variable does not change much in size from restaurant to restaurant, so heteroskedasticity is not likely to be a major problem. As a result, the assumption of a constant variance of the error term is reasonable. Consider the Park test to see if (**??**) gives any indication of heteroskedasticity.

1. *Calculate the residuals.* First, obtain the residuals from the equation you want to test.

2. *Use these residuals as the dependent variable in a second regression* Run a regression of the log of the squared residual as the dependent variable as a function of the log of the suspected proportionality factor Z as first outlined in (**??**).

   $$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Z_i + u_i$$

   It is possible that there exists no Z but if one does, it seems likely to be related to the size of the market. Larger error term variances might exist in more heavily populated areas, P is a reasonable choice to use to conduct the Park test.
   If the logged and squared residuals from (**??**) were regressed as a function of $\log P$ then we get

   $$\begin{aligned} \widehat{\ln(e_i^2)} &= 21.05 - 0.2865 \ln P_i \\ & \qquad\qquad (0.6263) \\ t &= \qquad\qquad -0.457 \\ \bar{R}^2 &= 0.0067 \qquad N = 33 \end{aligned} \tag{168}$$

3. *Test the significance of $\hat{\alpha_1}$ in (**??**).* As can be seen from the calculated t-score there is virtually no measurable relationship between the squared residuals of (**??**) and the population. The calculated t-score of -0.457 is smaller in absolute value than the critical t-test for a two-tailed, 1% test. We cannot reject the null hypothesis of homoskedasticity:

   $$\begin{aligned} &\mathrm{H_0}\colon\ \alpha_1 = 0 \quad \text{and} \\ &\mathrm{H_A}\colon\ \alpha_1 \neq 0 \end{aligned}$$

10.3.3 The White Test

To use the Park test, we need to know the particular value of Z_i, the variable suspected of being proportional to the possible heteroskedasticity. Quite often, however, we may want to test the possibility that more than one proportionally factor is involved simultaneously. Less frequently, we might not be able to decide which of a number of possible Z factors to test. In this case, it is more appropriate to use the White test

The **White test** approaches the detection of heteroskedasticity by running a regression with the squared residuals as the dependent variable. This time, the right-hand side of the secondary equation includes all the original independent variables, the squares of all the independent variables and the cross products of all the original independent variables with each other. The White test, hence, does not assume any particular form of heteroskedasticity. To run a White test,

1. Obtain the residuals of the estimated regression equation. This is identical to the first step of the Park test.

2. Square these residuals. Use this result as the dependent variable in a second equation that includes, as explanatory variables, each X, each X^2 and the product of each X multiplied by all other Xs. For example if the equation has 3 independent variables X_1, X_2 and X_3, the appropriate White test equation is:

$$\begin{aligned}(e_i^2) =&\alpha_0\\ &+\alpha_1 X_{1i}+\alpha_2 X_{2i}+\alpha_3 X_{3i}\\ &+\alpha_4 X_{1i}^2+\alpha_5 X_{2i}^2+\alpha_6 X_{3i}^2\\ &+\alpha_7 X_{1i}X_{2i}+\alpha_8 X_{1i}X_{3i}+\alpha_9 X_{2i}X_{3i}+u_i\end{aligned} \tag{169}$$

3. Test the overall significance of (**??**) with the chi-square test. The appropriate test statistic is NR^2 or sample size N multiplied by coefficient of determination (unadjusted R^2) of equation (**??**). If NR^2 is larger than the critical chi-square value, we cannot reject the null hypothesis of homoskedasticity.

One problems with the White test is that, in some situations, the secondary equation cannot be estimate because it has negative degrees of freedom. This could happen if the original equation has a small sample size and/or so many variables that the secondary equation has more independent variables than observations.

10.4 Remedies for Heteroskedasticity

The first thing to do if you suspect heteroskedasticity is to examine the equation carefully for specification errors. If there are no obvious specification errors, the heteroskedasticity is probably pure in nature and one of the remedies should be considered:

- weighted least squares
- heteroskedasticity-corrected standard errors
- redefining the variables

Consider an equation with pure heteroskedasticity caused by a proportionality factor Z:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{170}$$

where the variance of the error term, instead of being constant, is

$$\mathbb{VAR}(\epsilon_i) = \sigma_i^2 = \sigma^2 Z_i^2 \tag{171}$$

where σ^2 is the constant variance of a classical (homoskedastic) error term u_i and Z_i is the proportionality factor. Given that pure heteroskedasticity exists, then (**??**) can be shown to equal

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i \tag{172}$$

The error term in (**??**) is heteroskedastic because the variance $\sigma^2 Z_i^2$ is not a constant.

To make $Z_i u_i$ into just u_i, the easiest method is to divide the entire equation through by the proportionality factor Z_i, resulting in just the error term u_i with a constant variance of σ^2. This new equation satisfies the Classical Assumptions and a regression run on this new equation will no longer be expected to have heteroskedasticity.

The general remedy to heteroskedasticity is called Weighted Least Squares, which is a version of GLS. **Weighted least squares** involves dividing (**??**) through by whatever will make the term homoskedastic and then re-running the regression on the transformed variables. This technique has three steps, from (**??**)

1. Divide (**??**) through the proportionality factor Z:

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + u_i \tag{173}$$

   The error term of (**??**) is now just u_i, which is homoskedastic.

2. Recalculate the data for the variables to conform to (**??**)

3. Estimate (**??**) with OLS

The third step in Weighted Least Squares, the estimation of the transformed equation is tricky because the exact detail of how to complete this regression depend on *whether the proportionality factor Z is also an explanatory variable in (**??**).* If Z is not an explanatory variable in (**??**), then the regression to be run might seem to be

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + u_i \tag{174}$$

Note, however, that this equation has *no constant term* (simply because the value of ${}^{\beta_0}/_{Z_i}$ changes for every observation). Most OLS computer packages can run such a regression only if the equation is forced through the origin by specifically suppressing the intercept with an instruction to the computer.

A superior alternative to (**??**) is to add a constant term before the transformed equation is estimated. This is, if Z is not identical to one of the Xs in the original equation, we suggest that the following specification be run as step 3 in Weighted Least Squares:

$$\frac{Y_i}{Z_i} = \alpha_0 + \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + u_i \tag{175}$$

If Z is an explanatory variable in (**??**) then no constant term needs to be added because it already exists. Consider (**??**) again. If $Z = X_1$ or $Z = X_2$ then one of the slope coefficients becomes the constant term in the transformed equation because ${}^{X_1}/_{Z} = 1$:

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \beta_1 + \frac{\beta_2 X_{2i}}{Z_i} + u_i \tag{176}$$

If this form of Weighted Least Squares is used, however, coefficients obtained from an estimation of (**??**) must be interpreted very carefully. Note that β_1 is now an intercept in (**??**) term even though it is a slope coefficient in (**??**)

and that β_0 is a slope coefficient in (**??**) even though it is an intercept in (**??**). As a result a researcher interested in an estimate of the coefficient of X_1 in (**??**) will have to examine the intercept of (**??**). A researcher interested in an estimate of the intercept term of (**??**) would have to examine the coefficient of ${}^1/{Z_i}$ in (**??**).

There are two other major problems with using Weighted Least Squares:

1. The job of identifying the proportionality factor Z is, and hs been pointed out, to be difficult

2. The functional form that relates the Z factor to the variance of the error term of the original equation may not be our assumed squared function of (**??**). When some other functional relationship is involved, a different transformation is required.

10.4.1 Heteroskedasticity-Corrected Standard Errors

The most popular remedy for heteroskedasticity is heteroskedasticity-corrected standard errors, which take a completely different approach to the problem. It focuses on $\mathbb{SE}(\hat{\beta})$s while still using the OLS estimates of the slope coefficients. Since heteroskedasticity causes problems with the $\mathbb{SE}(\hat{\beta})$s but not with the estimated regression coefficients $\hat{\beta}$s, it makes sense to improve the estimation of the $\mathbb{SE}(\hat{\beta})$s in a way that does <u>not</u> alter the estimates of the slope coefficients. This differs from our other two remedies because both WLS and reformulating the equation affect both the $\hat{\beta}$s and the $\mathbb{SE}(\hat{\beta})$s.

Thus, **heteroskedasticity-corrected (HC) correct errors** are $\mathbb{SE}(\hat{\beta})$ that have been calculated specifically to avoid the consequences of heteroskedasticity. The HC procedure yields an estimator of the standard errors that, while being biased, are generally more accurate than uncorrected standard errors for large samples in the face of heteroskedasticity. As a result,the HC $\mathbb{SE}(\hat{\beta})$s can be used in t-tests and other hypothesis tests in most samples without the errors of inference potentially caused by heteroskedasticity. Typically, the HC $\mathbb{SE}(\hat{\beta})$s are greater than the OLS $\mathbb{SE}(\hat{\beta})$, thus producing lower t-scores and decreasing the probability that a given estimated coefficient will be significantly different from zero.

The problems associated with this is that the technique works best in large samples.

10.4.2 Redefining the Variables

Another approach to ridding an equation of heteroskedasticity si to go back to the basic underlying theory of the equation and redefine the variables in a way that avoids it.

Consider a cross-sectional model of the total expenditures by the governments of different cities. For example, the larger the total income of a city's residents and businesses for the city, the larger the city's government expenses. It is not very enlightening to know that larger cities spend more than smaller ones.

Fitting a regression line to such data also gives undue weight to the larger cities because they would otherwise give rise to large squared residuals. That means since OLS minimizes the summed squared residuals, and since the residuals from the large cities are likely to be larger due to size of the city, the regression estimation will be especially sensitive to the residuals from the larger cities. This is often called "spurious correlation" due to size.

In addition, the residuals may indicate heteroskedasticity. The remedy for this kind of heteroskedasticity is not to automatically use Weighted Least Squares or throw out observations from large cities. It makes sense to consider reformulating the model to discount the scale factor, in this case shrinking to discount the scale factor. In this case, per capita expenditure would be more logical.

This transformation is in some ways similar to Weighted Least Squares. The differences are that there is no term equal to the reciprocal of population and not all explanatory variables divided by population. For the original equation

$$EXP_i = \beta_0 + \beta_1 POP_i + \beta_2 INC_i + \beta_3 WAGE_i + \epsilon_i \tag{177}$$

The weighted least squares version is

$$\frac{EXP_i}{POP_i} = \beta_1 + \frac{\beta_0}{POP_i} + \beta_2 \frac{INC_i}{POP_i} + \beta_3 \frac{WAGE_i}{POP_i} + u_i \tag{178}$$

and the directly transformed equation would be

$$\frac{EXP_i}{POP_i} = \alpha_0 + \alpha_1 \frac{INC_i}{POP_i} + \alpha_2 WAGE_i + u_i \tag{179}$$

where
EXP_i refers to expenditures of the i-th city
INC_i refers to income of the i-th city
$WAGE_i$ refers to average wage of the i-th city
POP_i refers to the population of the i-th city

The weighted least squares equation of (**??**) divides through the entire equation by population but the theoretically transformed one divides on expenditures and income by population.
While the directly transformed equation equation (**??**) does indeed solve any potential heteroskedasticity in the model, such a solution should be considered incidental to the benefits of rethinking th equation in a way that focus on the basic behaviour being examined.

Note that it is possible for the *reformulated* (**??**) to have heteroskedasticity; the error variances might be larger for the observations having larger per capita values. Hence, it is legitimate to suspect and test for heteroskedasticity even in the transformed model.

10.5 A More Complete Example

Consider a more complete example involving a cross-sectional data set and heteroskedasticity. In the mid-1970s, the US Department attempted to allocate gasoline to regions, states and even individual retailers based on past usage, changing demographics and other factors. Underlying those allocations must have been some sort of model of the usage of petroleum by state as a function of a number of factors. It seems likely that such across-sectional model, if ever estimated, would have had to cope with the problem of heteroskedasticity.

In a model where the dependent variable is petroleum consumption by state, possible explanatory variables include functions of the size of the state and variables that are not functions of the size of the state. Since there is little to be gained to include more than one variable that measures the size of the state (because this might introduce collinearity) a reasonable model might be:

$$PCON_i = f(\overbrace{REG}^{+}, \overbrace{TAX}^{-}) + \epsilon_i = \beta_0 + \beta_1 REG_i + \beta_2 TAX_i + \epsilon_i \tag{180}$$

where:
$PCON_i$ is the petroleum consumption in the i-th state
REG_i is the motor vehicle registrations in the i-th state
TAX_i is the gasoline tax rate in the i-th state
ϵ_i is a classical error term

The more cars registered in a state, we would think the more petroleum consumed while a high tax rate would decrease aggregate gasoline purchases in the same state. If we collect the data and run OLS estimator we would yield:

$$\begin{aligned}\widehat{PCON}_i &= 551.7 + 0.1861 REG_i - 53.59 TAX_i \\ & \qquad\qquad (0.0117) \qquad\quad (16.86) \\ t &= \qquad\quad +15.88 \qquad\quad -3.18 \\ \bar{R^2} &= 0.861 \qquad N = 50\end{aligned} \tag{181}$$

The equation seems to have no problems. Given the discussion in the previous sections, let us investigate the possibility of heteroskedasticity caused by variation in the size of the states.

We obtain the residuals from (**??**) and run a Park test on them. As explained, before we run a Park test, we need to find an appropriate value of Z, the proportionality factor.

REG is a reasonable choice. The residuals potentially look heteroskedastic when looking at the plot of residual against size of the state, or in this case REG. So we run the Park test:

$$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln REG_i + u_i \tag{182}$$

where e_i is the residual for the i-th state from (**??**) and u_i is a classical (homoskedastic) error term. Running the Park test regression, we get:

$$\begin{aligned}\widehat{\ln(e_i^2)} &= 1.650 - 0.952 \ln REG_i \\ & \qquad\qquad (0.308) \\ t &= \qquad\quad +3.09 \\ \bar{R^2} &= 0.148 \qquad N = 50\end{aligned} \tag{183}$$

The critical t-value for a one-percent two-tailed t-test is about 2.7 so the appropriate decision rule is:

$$\begin{gathered}\text{Reject H}_0\text{: } \alpha_1 = 0 \text{ if } |t_{\text{Park}}| > 2.7 \text{ and} \\ \text{Do not Reject H}_0 \text{ if } \alpha_1 \leq 2.7\end{gathered}$$

Hence, we reject the null hypothesis and conclude that there is heteroskedasticity.

Since there appears to be heteroskedasticity, what do we do? First, we should think through the specification of the equation and maybe search for an omitted variable. While there are a number of possible ones for this equation, it turns out that the heteroskedasticity in the equation is very clearly pure heteroskedasticity.

As a result, we shall reestimate (**??**) with Weighted Least Squares, using REG as the proportionality factor Z.

$$\frac{PCON_i}{REG_i} = \frac{\beta_0}{REG_i} + \beta_1 + \beta_2 \frac{TAX_i}{REG_i} + u_i \tag{184}$$

resulting in the estimates

$$\begin{aligned} \frac{\widehat{PCON_i}}{REG_i} &= \frac{218.54}{REG_i} + 0.168 - 17.389 \frac{TAX_i}{REG_i} \\ & \qquad\qquad\quad (0.014) \qquad (4.682) \\ t &= \qquad\qquad +12.27 \qquad -3.71 \\ \bar{R}^2 &= 0.333 \qquad N = 50 \end{aligned} \tag{185}$$

Comparing this with (**??**),

1. The coefficient of the reciprocal of REG in (**??**) is really an estimate of the intercept of (**??**), and therefore no t-test is conducted even though the OLS regression program will indicate that it is a slope coefficient.

2. What appears to be the interception of (**??**) is actually an estimate of the coefficient of REG in (**??**). Note that this particular estimate is quite close in magnitude and significance to the original equation in (**??**) with 0.186 in (**??**) and 0.168 in (**??**)

3. The t-score of the coefficient of the proportionality factor REG is lower in the Weighted Least Squares estimate than it is in the potentially

heteroskedastic (**??**). The overall fit is also worse, but this has no particular importance because the dependent variables are different in the two equations.

However, as mentioned before an alternative is to rethink the purpose of the regression and reformulate the variables of the equation to try to avoid heteroskedasticity resulting from spurious correlation due to size. If we were to rethink (**??**) we could decide to attempt per capita petroleum consumption and get

$$\frac{PCON_i}{POP_i} = \beta_0 + \beta_1 \frac{REG_i}{POP_i} + \beta_2 TAX_i + \epsilon_i \tag{186}$$

where POP_i is the population of the i-th state in thousands of people.
We reformulated the equation in a similar way to the Weighted Least Squares, but now have an equation that can stand on its own from a theoretical point of view. Estimation of (**??**) yields:

$$\begin{aligned} \frac{\widehat{PCON_i}}{POP_i} &= 0.168 + 0.1082 \frac{REG_i}{POP_i} - 0.0103 \frac{TAX_i}{POP_i} \\ & \qquad\qquad\quad (0.0716) \qquad\qquad (0.0035) \\ t &= \qquad\quad\ +1.51 \qquad\qquad\quad -2.95 \\ \bar{R}^2 &= 0.165 \qquad N = 50 \end{aligned} \tag{187}$$

Comparing (**??**) and (**??**) and (**??**), we see this approach is not necessarily better but quite different. The statistical properties of (**??**), though not directly comparable to the other equations do not seem as strong but this is not necessarily the deciding factor.

Which is better? It depends on the purposes of the research. If your goal is to determine the impact of tax rates on gasoline consumption all three models give virtually the same results in terms of the signs and significance of the coefficient. However, the latter two models avoid the heteroskedasticity.
If your goal is to allocate petroleum in aggregate amounts to states then the first equation is fine.

Finally let us apply HC standard errors to this example. Staring with (**??**) and using White's suggested method for estimating $\mathbb{SE}(\hat{\beta})$s that are minimum variance while facing heteroskedasticity, we get

$$\begin{aligned}\widehat{PCON}_i &= 551.7 + 0.186REG_i - 53.59TAX_i \\ & \qquad\qquad (0.022) \qquad\quad (23.90) \\ t &= \qquad\quad +8.64 \qquad\quad -2.24 \\ \bar{R}^2 &= 0.86 \qquad N = 50\end{aligned} \tag{188}$$

Comparing (**??**) and (**??**), the slope coefficients are identical, as you'd expect the HC approach uses OLS to estimate the coefficients. Also note that the HC $\mathbb{SE}(\hat{\beta})$s are higher than the OLS $\mathbb{SE}(\hat{\beta})$s, as is usually but not necessarily the case. Although the resulting t-scores are lower they are still significantly different from zero in the direction we expected, making (**??**) very appealing, still.

Is the HC standard error approach the best for this example? Not necessary, because the sample size of 50 makes it unlikely that the large sample properties of HC estimation hold in this case. Finally, if t-scores are not used to test hypotheses or retain variables, as it is true in this example, it is not at all clear that any sort of remedy for heteroskedasticity is even necessary. The purpose of this chapter is to provide an introduction to a number of interesting models that have been designed to cope with and take advantage of special properties of time series data. Working with these often cause complications that simply can't happen with cross-sectional data.

The most important of the topics concerns a class of dynamic models in which a lagged value of the dependent variable appears in the right hand side of the equation. This implies that the impact of the independent variables could spread out over a number of time periods. We will also learn about models in which different numbers of lags appear and investigate ho the presence of these lags affects our estimators.

11 Time Series Models

11.1 Dynamic Models

11.1.1 Distributed Lag Models

Lagged independent variables can be used when you expect X to affect Y after a period of time. For example if the underlying theory suggests that X_1 affects Y with a one-time-period lag, we use equations like

$$Y_t = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{2t} + \epsilon_t \tag{189}$$

Such lags are called simple lags, and the estimation of β_1 with OLS is no more difficult than the estimation of coefficients of non-lagged equations, except for possible impure serial correlation if the lag is misspecified. Recall also in (**??**) that β_2 is the effect of a one-unit increase on this time's X_2 on this time's Y holding *last time's* X_1 constant.

A case that is more complicated than this simple lag model occurs when the impact of an independent variable is expected to spread out over a number of periods. An example is the effect of money supply on GDP. In such a case, the appropriate econometric model would be a distributed lag model:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \epsilon_t \tag{190}$$

A **distributed lag model** explains the current value of Y as a function of current and past values of X, thus 'distributing' the impact of X over a number of time periods. The coefficients β_0, β_1 all the way to β_p measure the effects of the various lagged values of X on the current value of Y. In most econometric applications, we expect the impact of X on Y to decrease as the length of the lag increases.

Unfortunately, the estimation of (**??**) with OLS causes a number of problems:

1. The various lagged values of X are likely to be severely multi-collinear, making coefficient estimates imprecise.

2. In large part because of this multi-collinearity, there is no guarantee that the estimates βs will follow the smoothly declining pattern that economic theory would suggest. Instead, it is quite typical for the estimated coefficients of (**??**) to follow a fairly irregular pattern.

3. The degrees of freedom tend to decrease, sometimes substantially for two reasons. First, we have to estimate a coefficient for each lagged X, thus increasing K and lowering the degrees of freedom $(N - K - 1)$. Second, unless data for lagged Xs outside the sample are available, we have to decrease the sample size by one for each lagged X we calculate. This lowers the number of observations N and therefore the degrees of freedom.

As a result of these problems with OLS estimation of functions like (**??**) called *ad hoc* distributed lag equations, it is standard practice to use a simplifying assumption in such situations. The most commonly used simplification is to replace all the lagged independent variables with a lagged value of the dependent variable, and this model is called the *dynamic model.*

11.1.2 What is a Dynamic Model?

The simplest dynamic model is:

$$Y_t = \alpha_0 + \beta_0 X_t + \lambda Y_{t-1} + u_t \tag{191}$$

Note that Y is on both sides of the equation. The one on the right hand side, however, is Y_{t-1}. It is this difference in time period that makes the equation dynamic. Thus, the simplest **dynamic model** is an equation in which the current variable of the dependent variable Y is a function of the current value of X and a lagged value of Y itself. Such a model with a lagged dependent variable is often called an *autoregressive* regression.

Consider(**??**) to see why it can be used to represent a distributed lag model. Suppose we lag (**??**) one time period:

$$Y_{t-1} = \alpha_0 + \beta_0 X_{t-1} + \lambda Y_{t-2} + u_{t-1} \tag{192}$$

Substitute (**??**) into (**??**) yields:

$$Y_t = \alpha_0 + \beta_0 X_t + \lambda(\alpha_0 + \beta_0 X_{t-1} + \lambda Y_{t-2} + u_{t-1}) + u_t \tag{193}$$

or

$$Y_t = (\alpha_0 + \lambda\alpha_0) + \beta_0 X_t + \lambda\beta_0 X_{t-1} + \lambda^2 Y_{t-2} + (\lambda u_{t-1} + u_t) \tag{194}$$

Do this one more time and we get

$$Y_t = \alpha_0^* + \beta_0 X_t + \lambda\beta_0 X_{t-1} + \lambda^2 Y_{t-2} + \lambda^3 Y_{t-3} + (u_t^*) \tag{195}$$

where α_0^* is the new combined intercept and u_t^* is the new combined error term. In other words, $Y_t = f(X_t, X_{t-1}, X_{t-2})$ and we have shown that a dynamic model can indeed be used to represent a distributed lag model.

In addition, note that the coefficients of the lagged Xs follow a clear pattern. Comparing the coefficients in (**??**) and (**??**), we get

$$\begin{aligned}
\beta_1 &= \lambda\beta_0 \\
\beta_2 &= \lambda^2\beta_0 \\
\beta_3 &= \lambda^3\beta_0 \\
&\vdots \\
\beta_p &= \lambda^p\beta_0
\end{aligned} \tag{196}$$

As long as $0 < \lambda < 1$, these coefficients will indeed smoothly decline as with a geometric progression.

Dynamic models like those in (**??**) avoid the three major problems associated with ad hoc distributed lag equations. The degrees of freedom have increased dramatically, and the multicollinearity problem has disappeared. If u_t is well behaved, OLS estimation of (**??**) can be shown to have desirable properties for extremely large samples. The author recommends a sample of at least 50 observations. The smaller the sample size, the more likely to encounter bias and hypothesis testing becomes untrustworthy.

In addition to this sample size issue, dynamic models are more likely to encounter serial correlation than are equations without a lagged independent variable as an independent variable. To make things worse, serial correlation almost surely will cause bias in the OLS estimates of dynamic models no matter how large the sample size is.

11.1.3 An Example of a Dynamic Model

Consider an aggregate consumption function from a macroeconomic equilibrium GDP model. Many economists argue that in such a model, consumption CO_t is not just an instantaneous function of disposable income YD_t but also influenced by past levels of disposable income (YD_{t-1}, YD_{t-2} etc.):

$$CO_t = f(\overbrace{YD_t}^{+}, \overbrace{YD_{t-1}}^{+}, \overbrace{YD_{t-2}}^{+}, \text{etc}) + \epsilon_t \tag{197}$$

Such an equation fits well with simple models of consumption, but makes sense only if the weights given past levels of income decrease as the length of the lag increases. Consequently we expect the coefficient to decrease as the lag time increases.

As a result, most econometricians would model (**??**) with a dynamic model:

$$CO_t = \alpha_0 + \beta_0 YD_t + \lambda CO_{t-1} + u_t \tag{198}$$

To estimate (**??**), we use data from section 14.3 and the OLS estimates of (**??**) for this data set are:

$$\begin{aligned} \widehat{CO}_t &= -279.21 + 0.49 YD_t + 0.54 CO_{t-1} \\ & \qquad\qquad\quad (0.12) \qquad\quad (0.12) \\ t &= \qquad\qquad +4.22 \qquad +4.54 \\ \bar{R}^2 &= 0.999 \qquad N = 28 \text{ (Annual 1976 - 2003)} \end{aligned} \tag{199}$$

If we substitute $\hat{\beta}_0 = 0.43$ and $\hat{\lambda}_0 = 0.59$ into (**??**) for $i = 1$, we obtain $\hat{\beta}_0 = \hat{\beta}_0 \hat{\lambda}_1 = (0.43)(0.59)^1 = 0.25$ and if we continue this process, it turns out (**??**) is equivalent to

$$\begin{aligned} \widehat{CO}_t &= -606.98 + 0.49 YD_t + 0.26 YD_{t-1} \\ &\quad + 0.14 YD_{t-2} + 0.08 YD_{t-3} + \cdots \end{aligned} \tag{200}$$

As can be seen the coefficients of YD in (**??**) do indeed decline as we'd expect in a dynamic model.

To compare this estimate with an OLS estimate of the same equation without the dynamic model format, we need to estimate an ad hoc distributed lag equation with the same number of lag variables:

$$CO_t = \alpha_0 + \beta_0 YD_t + \beta_1 YD_{t-1} + \beta_2 YD_{t-2} + \beta_3 YD_{t-3} + \epsilon_t \tag{201}$$

If we estimate (**??**) using the same data set we get

$$\widehat{CO}_t = -663.55 + 0.77 YD_t + 0.37 YD_{t-1} + -0.004 YD_{t-2} - 0.11 YD_{t-3} \tag{202}$$

In (**??**), as the lag increases the coefficients of YD decrease sharply, becoming negative in YD_{t-3}. Neither economic theory nor common sense leads us to expect this pattern. Such a poor result is due to the severe multicollinearity between the lagged Xs. Hence, the dynamic model is preferred to the ad hoc model.

An increasing interpretation of the results of (**??**) is the long-run multiplier implied by the model. The long-run multiplier measures the total impact of a change in income on consumption after all the lagged effects have been felt. Since it is a geometric progression, we should calculate $\frac{\beta_0}{1-\lambda}$. In this case it is $\frac{0.49}{1-0.54} = 0.93$.

11.2 Serial Correlation and Dynamic Models

The consequences of serial correlation depend crucially on the type of model we are talking about. For a distributed lag model like (**??**), serial correlation causes the OLS to no longer be the minimum variance unbiased estimator, causes the $\mathbb{SE}(\hat{\beta})$s to be biased and causes no biased in the OLS $\hat{\beta}$ themselves.

For dynamic models like (**??**), however, serial correlation does indeed cause bias in the $\hat{\beta}$s produced by OLS. Compounding this is the fact that the consequences, detection and remedies for serial correlation discussed earlier are all either incorrect or need to be modified in the presence of a lagged variable.

11.2.1 Serial Correlation Causes Bias in Dynamic Models

If an equation that contains a lagged dependent variable as an independent variable has a serially correlated error term, then OLS estimates of the coefficients of that equation will be biased, even in large samples. Consider a dynamic model like (**??**):

$$Y_t = \alpha_0 + \beta_0 X_t + \overset{\uparrow}{\lambda Y_{t-1}} + \overset{\uparrow}{u_t}$$

and assume that the error term u_t is serially correlated: $u_t = \rho u_{t-1} + \epsilon_t$ where ϵ_t is a classical error term. If we substitute this serially correlated error term into (**??**) we get

$$Y_t = \alpha_0 + \beta_0 X_t + \overset{\uparrow}{\lambda Y_{t-1}} + \overset{\uparrow}{\rho u_{t-1}} + \epsilon_t \tag{203}$$

Consider (**??**) lagged one time period:

$$\overset{\uparrow}{Y_{t-1}} = \alpha_0 + \beta_0 X_{t-1} + \lambda Y_{t-2} + \overset{\uparrow}{u_{t-1}} \tag{204}$$

In (**??**), the positive error term u_{t-1} causes Y_{t-1} to be larger than it would have been otherwise (these changes are marked by the upward arrows). In addition, the positive u_{t-1} is quite likely to cause u_t to be positive in (**??**) because $u_t = \rho u_{t-1} + \epsilon_t$ and ρ is usually positive.

Consider the errors in (**??**). Y_{t-1} and u_t are correlated! Such a correlation violates Classical Assumption III, which assumes that the error term is not correlated with any of the explanatory variables.

The consequences of this correlation include biased estimates, in particular of the coefficient λ because OLS attributes to Y_{t-1} some of the change in Y_t actually caused by u_t. In essence, the uncorrelated serial correlation acts like an omitted variable u_{t-1}. Since an omitted variable causes bias whenever it is correlated with one of the included independent variables, and since u_{t-1} is correlated with Y_{t-1} the combination of a lagged dependent variable and serial correlation causes bias in the coefficient estimates.

Serial correlation in a dynamic model also causes estimates of the standard errors of the estimated coefficients and the residuals to be biased. The former bias means that hypothesis testing is invalid even for large samples. The latter bias means that tests based on the residuals like the Durbin-Watson d test are potentially invalid.

11.2.2 Testing for Serial Correlation in Dynamic Models

We relied on the Durbin-Watson d test to test for serial correlation, but it is potentially invalid for an equation that contains a lagged dependent variable as an independent variable. This is because the biased residuals cause the DW d statistic to be biased toward 2. This bias toward 2 means that the Durbin-Watson test sometimes fails to detect the presence of serial correlation in a dynamic model.

The widely used alternative is to use a special case of a general testing procedure called the Lagrange Multiplier Test. The **Lagrange Multiplier Serial Correlation (LMSC) Test** is a method that can be used to test for serial correlation by analysing how well the lagged residuals explain the residuals of the original equation. If the lagged residuals are significant in explaining this time's residuals, then we can reject the null hypothesis of no serial correlation.

Using the Langrange Multiplier to test for serial correlation for a typical dynamic model involves three steps:

1. Obtain the residuals from the estimated equation

$$e_t = Y_t - \hat{Y}_t = Y_t - \hat{\alpha}_0 - \hat{\beta}_0 X_{1t} - \hat{\lambda}_0 Y_{t-1} \tag{205}$$

2. Use these residuals as the dependent variable in an auxiliary equation that includes as independent variables all those on the right-hand side of the original equation as well as the lagged residuals

$$e_t = a_0 + a_1 X_t + a_2 Y_{t-1} + a_3 e_{t-1} + u_t \tag{206}$$

3. Estimate (**??**) using OLS and then test the hypothesis that $a_3 = 0$ with the following test statistic:

$$LM = NR^2 \tag{207}$$

   where N is the sample size and R^2 is the unadjusted coefficient of determination, both from the auxiliary equation (**??**). For large samples, LM has a chi-square distribution with degrees of freedom equal to the number of restrictions in the null hypothesis (in this case, one). If LM is greater than the critical chi-square value then we reject the null hypothesis that $a_3 = 0$ and conclude that there is indeed serial correlation in the original equation.

To run an LMSC test for second-order or higher-order serial correlation, add lagged residuals (e_{t-2} for second order, e_{t-2} and e_{t-3} for third order and so on) to the auxiliary equation (**??**). This latter change makes the null hypothesis $a_3 = a_4 = a_5 = 0$. Such a null hypothesis raise the degrees of freedom in the chi-square test to three because we have imposed three restrictions on the equation (three coefficients are jointly set equal to zero). To run an LMSC test with more than one lagged independent variable, add the lagged variables (Y_{t-2}, Y_{t-3}, etc.) to the original equation.

11.2.3 Correcting for Serial Correlation in Dynamic Models

There are three strategies for attempting to rid a dynamic model of serial correlation: improve the specification, instrumental variables and modified GLS.

The first strategy is to consider the possibility that the serial correlation could be impure, caused by either omitting a relevant variable or by failing to capture the actual distributed lag pattern accurately. Unfortunately, finding an omitted variable or an improved lag structure is easier said than done. Because of the dangers of sequential specification searches, this option should be considered only if an alternative specification exists that has a theoretically sound justification.

The second strategy, called instrumental variables, consists of substituting an "instrument" (a variable that is highly correlated with Y_{t-1} but uncorrelated with u_t) for Y_{t-1} in the original equation, thus eliminating the correlation between Y_{t-1} and u_t. Although using an instrument is a reasonable option, it is not always easy to find a proxy that retains the distributed lag nature of the original equation.

The final solution to serial correlation in dynamic models is to use an iterative maximum likelihood technique to estimate the components of the serial correlation and then to transform the original equation so that the serial correlation has been eliminated. This technique, similar to the GLS procedure is not without its complications. In particular the sample needs to be large, the standard errors of the estimated coefficients potentially need to be adjusted, and the estimation techniques are flawed under some circumstances.

11.3 Granger Causality

One application of ad hoc distributed lag models is to provide evidence about the direction of causality in economic relationships. Such a test is useful when we know that two variables are related but we do not know which variable causes the other to move.

One approach to such a question of indeterminate causality is to theorize that the two variables are determined simultaneously. A second approach is to test for what is called "Granger Causality". **Granger Causality** or precedence is a circumstance in which one time-series variable consistently and predictably changes before another variable. Granger causality is important because it allows us to analyse which variable precedes or "leads" the other, and such leading variables are useful for forecasting purposes.

Despite the value of Granger causality, however, we should not let ourselves be lured into thinking that it allows us to prove economic causality in any rigorous way. If one variable precedes ("Granger causes") another we can't be sure that the first variable "causes" the other to change.

As a result, even if we are able to show that Event A always happens before Event B, we have not shown that Event A "causes" event B. All we have shown is that Event A preceded or "Granger caused" Event B.

Granger suggested that to see of A Granger-caused Y, we should run:

$$\begin{aligned} Y_t =&\beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} \\ &+ \alpha_1 A_{t-1} + \alpha_2 A_{t-2} \cdots + \alpha_p A_{t-p} + \epsilon_t \end{aligned} \tag{208}$$

and test the null hypothesis that the coefficients of the lagged As (specifically the αs) equal jointly equal zero. If we can reject this null hypothesis using the F-test, then we have evidence that A Granger-causes Y. Note that if $p = 1$, (**??**) is similar to the dynamic model, (**??**)

Applications of this test involve running two Granger tests, one in each direction. That is, run (**??**) and also run:

$$\begin{aligned} A_t =&\beta_0 + \beta_1 A_{t-1} + \cdots + \beta_p A_{t-p} \\ &+ \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} \cdots + \alpha_p Y_{t-p} + \epsilon_t \end{aligned} \tag{209}$$

testing for Granger causality in both directions by testing the null hypothesis that the coefficients of the lagged Ys (the αs) jointly equal zero. If the F-test is significant for (**??**) but not for (**??**) then we can conclude that A Granger-causes Y.

11.4 Spurious Correlation and Nonstationarity

One problem with time-series data is that independent variables can appear to be more significant than they actually are if they have the same underlying trend as the dependent variable.
For example in a country with rampant inflation, almost any nominal variable will appear to be highly correlated with all other nominal variables as they have not been adjusted for inflation. Hence, they have a powerful inflationary component. This inflationary component will usually outweigh any real causal relationship causing nominal variables to appear to be correlated even if they aren't.

This is an example of **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 ore 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.

11.4.1 Stationary and Nonstationary Time Series

A stationary series is one whose basic properties, for example its mean and its variance, do not change over time. In contrast, a nonstationary series has one or more basic properties that *do* change over time. In our inflation example the mean of a nominal variable changes over time, but a time series can be nonstationary even with a constant mean if another property like variance changes over time, so it is nonstationary. By contrast, the growth *rate* of real per capita output often does not increase over time, so this variable is stationary even though the variable it's based on, real per capita output, is nonstationary.

Formally, a time-series variable is **stationary** if:

1. the mean of X_t is constant over time
2. the variance of X_t is constant over time and
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 $\forall\ k$

If one or more of these properties are violated then X_t is **nonstationary**. If a series is nonstationary, that problem is often referred to as **nonstationarity**.

Although our definition of stationary series focuses on stationary and nonstationary *variables*, it is important to note that *error terms* (and therefore, *residuals*) can also be nonstationary. In fact, 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 nonstationarity for regression analysis is spurious correlation that inflates R^2 and the t-scores of the nonstationary independent variables which in turn leads to incorrect model specification. This occurs because the regression procedure attributes to the nonstationary X_t

changes in Y_t that were actually caused by some factor (trend, for example) that also affects X_t. Thus the variables move together because of the nonstationarity, increasing the R^2 and relevant t-scores.

Some variables are nonstationary mainly because they increase rapidly over time. Spurious regression results involving these kinds of variables often can be avoided by the addition of a simple time trend ($t = 1, 2, 3, \cdots, T$) to the equation as an independent variable.

Unfortunately, many economic time-series variables are nonstationary even after the removal of a time trend. This nonstationarity typically takes the form of the variable behaving as though it were a "random walk". A **random walk** is a time-series variable where next period's value equals this period's 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 any long term mean of any sort.

Suppose that Y_t is generated by an equation that includes only past values of itself (an *autogressive* equation):

$$Y_t = \gamma Y_{t-1} + v_t \tag{210}$$

where v_t is a classical error term. If $|\gamma| < 1$, then the expected value of Y_t will eventually approach 0 (and be stationary) as the sample gets bigger and bigger. v_t is a classical error term with expected value = 0.If $|\gamma| > 1$ then the expected value of Y_t will continuously increase making Y_t nonstationary. This is nonstationarity due to a trend. If $\gamma = 1$, then

$$Y_t = Y_{t-1} + v_t \tag{211}$$

and becomes a random walk. The expected value of Y_{t-1} does not converge on any value, meaning that it is nonstationary. This circumstance, where $\gamma = 1$ in (**??**) is called a **unit root**. If a variable has a unit root then (**??**) holds, and the variable follows a random walk and is nonstationary.

11.4.2 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:

$$Y_t = \alpha_0 + \beta_0 X_t + u_t \tag{212}$$

If both X and Y are nonstationary then they can be highly correlation for non-causal reasons, and our standard regression inference measures will almost be surely be very misleading in that they'll overstate $\bar{R}^2$ and the t-score for $\hat{\beta}_0$.

For example, consider the following estimated equation:

$$\begin{aligned} \widehat{PRICE}_t &= -27.8 - 0.070 TUITION_t \\ &\qquad\qquad\quad (0.006) \\ t &= \qquad\qquad\quad +11.4 \\ \bar{R}^2 &= 0.94 \qquad T = 10 \end{aligned} \tag{213}$$

The R^2 of this equation and the t-score for the coefficient of $TUITION$ are clearly significant, but $PRICE$ is price of a gallon of gasoline in Portland and $TUITION$ is tuition fees for a semester of study at a college in Los Angeles. What is going on? The 1970s was a decade of inflation and so any nominally measured variables are likely to result in an equation that fits as well as (**??**). Both variables are nonstationary and this particular regression is spurious. To avoid spurious regression results, it is crucial to be sure that time-series variables are stationary before running regressions.

11.4.3 The Dickey-Fuller Test

To ensure that the equations we estimate are not spurious, it is important to test for nonstationarity. The standard method for testing for nonstationarity is the **Dickey-Fuller test**, which 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.

Return to the discussion of the role that unit roots play in the distinction between stationarity and nonstationarity. Recall that we looked at the value of γ in (**??**) to help us determine if Y was stationary or nonstationary. We concluded taht the autoregressive model is stationary if $|\gamma| < 1$ and nonstationary otherwise.

From this discussion of stationarity and unit roots, it makes sense to estimate (**??**) and determine if $|\gamma| < 1$ to see if Y is stationary and that is almost exactly how the Dickey-Fuller test works. First, subtract Y_{t-1} from both sides of (**??**) yielding

$$Y_t - Y_{t-1} = (\gamma - 1)Y_{t-1} + v_t \tag{214}$$

If we define $\Delta Y_t = Y_t - Y_{t-1}$ then we have the simplest form of the Dickey-Fuller test:

$$\Delta Y_t = \beta_1 Y_{t-1} + v_t \tag{215}$$

where $\beta_1 = \gamma - 1$. The null hypothesis is that Y_t contains a unit root and the alternative hypothesis is that Y_t is stationary.
If Y_t contains a unit root, $\gamma = 1$ and $\beta_1 = 0$
If Y_t is stationary then $|\gamma| < 1$ and $\beta_1 < 0$
We construct a one-sided t-test on the hypothesis that $\beta_1 = 0$:

$$\begin{aligned} &\mathrm{H_0}\colon\ \beta_1 = 0 \text{ and} \\ &\mathrm{H_A}\colon\ \beta_1 \neq 0 \end{aligned}$$

The Dickey-Fuller test actually comes in three versions:

- Equation (**??**)
- Equation (**??**) with an added constant term
- Equation (**??**) with a constant term and a trend term

The form of the Dickey-Fuller test in (**??**) is correct if Y_t follows (**??**).

If we believe that (**??**) includes a constant then the appropriate Dickey-Fuller test equation is:

$$\Delta Y_t = \beta_0 + \beta_1 Y_{t-1} + v_t \tag{216}$$

Similarly, if we belive that Y_t contains a trend t where $t = 1, 2, 3, \cdots, T)$ then we add t to the equation as a *variable* with a coefficient and the appropriate Dickey-Fuller test becomes

$$\Delta Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + v_t \tag{217}$$

With the three versions, the appropriate decision rule is based on the estimate of β_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 nonstationarity. Note that they use different tables from the t-table.

The equation 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. This adjustment takes the form of adding in several lagged first differences as independent variables in the equation for the Dickey-Fuller test. There are several good methods for choosing the number of lags to add but there is no universal agreement as which of these methods is optimal.

11.4.4 Cointegration

If the Dickey-Fuller test reveals nonstationarity, 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. With economic data, taking a first difference usually is enough to convert a nonstationary series to a stationary one. Unfortunately, using first differences to correct for nonstationarity 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. In particular, first differences should not be used until the residuals have been tested for *cointegration.*

Cointegration consists of matching the degree of nonstationarity of the variables in an equation in a way that makes the error term (and residuals) of the equation stationary and rids the equation of any spurious regression results. Even though individual variables might not be stationary, linear combinations of such variables *could* be stationary or *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.

To see this return to (**??**):

$$Y_t = \alpha_0 + \beta_0 X_t + u_t$$

If X_t and Y_t are nonstationary, it is likely that we get spurious regression results. To understand how to get sensible results from (**??**) if the nonstationary variables are cointegrated, focus on the case in which both X_t and Y_t contain one unit root. They key to cointegration is the behaviour of u_t. Solving (**??**) for u_t, we get

$$u_t = Y_t - \alpha_0 - \beta_0 X_t \tag{218}$$

In (**??**), u_t is a function of two nonstationary variables, so you expect u_t to be nonstationary, but that is not necessarily the case. In particular, suppose X_t and Y_t are related. the error u_t may well be stationary even though X_t and Y_t are nonstationary. If u_t is stationary, then the unit roots in Y_t and X_t 'cancel out' and Y_t and X_t are said to be cointegrated.

If X_t and Y_t are cointegrated then OLS estimation of the coefficients of (**??**) can avoid spurious results. To determine of X_t and Y_t are cointegrated, we begin with OLS estimation of (**??**) and calculate the OLS residuals.

$$e_t = Y_t - \hat{\alpha}_0 - \hat{\beta}_0 X_t \tag{219}$$

We then perform a Dickey-Fuller test on the residuals. If we are able to reject the null hypothesis of a root unit in the residuals, we can conclude that Y_t and X_t are cointegrated and our OLS estimates are not spurious.

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 first differences (ΔY and ΔX). However, if the nonstationary variables are cointegrated then the estimation could be estimated in its original units.

11.4.5 A Standard Sequence of Steps for Dealing with Nonstationary Time Series

To deal with the possibility that nonstationary time series might be causing regression results to be spurious, most empirical work in time series follow a standard series of steps:

1. Specify the model. This might be a time-series equation with no lagged variables or a dynamic model in its simplest form, or a dynamic model that includes lags in both the independent and dependent variables.

2. Test all variables for nonstationarity (technically unit roots) using the appropriate version of the Dickey-Fuller test.

3. If the variables don't 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 version of the Dickey-Fuller test of 12.4.4

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 also are cointegrated then estimate the equation in its original units.

–END–