Estimation and Forecasting Methods

Appendix to the User Manual of the Multi-Year Budgeting Model of the Russian Federation

This document was prepared as part of the ***Russian Federation Multi-Year Budgeting Project***, upon request of the Budget Department of the Ministry of Finance of the Russian Federation under the sponsorship of the World Bank. The research team consists of Jorge Martinez-Vazquez, Jameson Boex and Robert McNab

Table of Contents

1. Introduction 1

2. Overview of Statistical Methods and Forecasting Techniques 2
   - 2.1 Expert Analysis 2
   - 2.2 Statistical Analysis 2
   - 2.3 Ordinary Least Squares Regression Analysis 3
   - 2.4 Forecasting Technique: Time-Trend Models 4
   - 2.5 Forecasting Technique: Exponential Smoothing 5
   - 2.6 Forecasting Technique: Bi- and Multi-Variate Regression Models 5
   - 2.7 Time-Series Analysis and Other Advanced Techniques 5

3. Theoretical Review: Descriptive Statistics 7
   - 3.1 Descriptive Statistics - General 7
   - 3.2 Descriptive Statistics - Measures of Central Tendency 7
   - 3.3 Descriptive Statistics - Measures of Variability 9
   - 3.4 Estimating Confidence Intervals 10
   - 3.5 Hypothesis Testing for the Mean, Median, and Midrange 12

4. Theoretical Review: Regression Analysis 14
   - 4.1 Introduction 14
   - 4.2 Assumptions of the Multi-Variate Model 14
   - 4.3 Deriving the Estimator 16
   - 4.4 Determining the "Fit" of a Model and Parameters 19
   - 4.5 Exponential Smoothing 20

5. Statistical Procedures in Microsoft Excel 22
   - 5.1 Introduction 22
   - 5.2 Calculating Measures of Central Tendency 23
   - 5.3 Calculating Measures of Variability 26
   - 5.4 Calculating Confidence Intervals 26
   - 5.5 Conducting Hypothesis Tests in Microsoft Excel 30
   - 5.6 Using the Data Analysis Tool to Produce Summary Statistics 32

6. Regression Analysis in Microsoft Excel 36
   - 6.1 Forecasting Using a Time-Trend Model: The TREND Function 36
   - 6.2 Exponential Smoothing: Forecasting Using the GROWTH Function . 38
   - 6.3 Estimating a Linear Regression Using the LINEST Function 39
   - 6.4 Regression Analysis Using the Data Analysis Tool 42
   - 6.5 Concluding Remarks 45

1. Introduction

The purpose of this note is to provide an overview of the most common techniques used for obtaining summary statistics and producing forecasts of economic data. Also presented in this note is the underlying theory for the statistical analyses and regression approaches. In addition, step-by-step instructions are presented on how these procedures can be performed in Microsoft Excel. As a result, this document can serve both as the theoretical and practical foundation for forecasting economic variables as part of the *Multi-Year Budgeting Model of the Russian Federation*.

Two types of analyses can be performed to develop forecasts of economic data. Expert analysis (formally known as qualitative analysis) relies on the knowledge of experienced analysts to produce forecasts based upon the analysts' familiarity with historical trends in the data. Quantitative analysis, on the other hand, relies on economic and statistical theory to develop models and forecasts for economic variables. Neither method is mutually exclusive, and are often used together to produce more reliable forecasts than could be produced exclusively by either method.

Qualitative analysis methods, or expert analysis, will not be covered in depth in this note. However, it is often necessary to rely on expert analysis because frequently historical data is absent or incomplete. While qualitative methods can produce reasonably accurate forecasts, whenever possible, quantitative methods should be used to supplement the qualitative analysis process. Rather than viewing quantitative analysis as a replacement of the current use of expert analysis, one should consider quantitative analysis techniques as a set of tools that can readily generate forecasts based on objective statistical methods whose principles are widely known and transparent. In this manner, the ability to produce accurate forecasts for variables of interest is substantially enhanced.

The remainder of this note is organized into five sections. Section 2 provides a general overview of the statistics and forecasting techniques presented in this note. The theoretical background for these techniques is presented in Section 3 (Statistics) and Section 4 (Regression Analysis). The remaining sections (Sections 5 and 6) provide instructions on how to perform the reviewed analyses in Microsoft Excel. Thus, this note presents an in-depth review of the relevant theory, while also allowing readers who are already familiar with the theoretical background to immediately advance to the application in Sections 5 and 6.

2. Overview of Statistical Methods and Forecasting Techniques

A wide variety of techniques can be used to develop forecasts for the economic variables or "driving variables" for the *Multi-Year Budget Model of the Russian Federation*. Invariably, the best approach is to develop forecasts for the relevant variable using several different methods and then to evaluate the results provided by the different forecasting methods. In this manner, one can identify the most suitable way to estimate each variable of interest. This section provides a non-technical overview of the estimation techniques covered in this note. The theoretical background of each approach is considered in more detail in Sections 3 and 4, while the use of these techniques in Microsoft Excel is discussed in Sections 5 and 6.

2.1 Expert Analysis

Expert analysis, sometimes referred to as qualitative analysis, can be a valuable analytical tool in a variety of situations; the value of an individual with long-standing experience with the economy should not be discounted. Expert analysis is often relied on to produce forecasts when the underlying data is incomplete, suspect, or not of sufficient length to develop a forecast for the forecast period of interest. For example, few economic models produce accurate estimates of the economic growth rate for a multi-year framework. As a result, estimates for this variable are often produced an expert economic forecaster or by reaching consensus among a group of experts.

While expert analysis can produce reasonably accurate results, often more accurate estimates can be produced when expert analysis is used in conjunction with quantitative forms of analysis.

2.2 Statistical Analysis

Basic statistical methods can be applied in a variety of situations and often produce results that are as reliable as more sophisticated methods of analysis.

A straightforward example of the use of statistical techniques could be the projection of the rate of economic growth. Historical data could be used to estimate the average growth rate for the Russian economy for the past several years, and it could be assumed that the economy will grow at this rate in the medium term. On one hand, this approach is very intuitive and does not require significant effort by the analyst. On the other hand, the accuracy of this estimation technique is highly dependent on the stability of the economy, as the estimate does not capture the trends over time in the variable under investigation. In addition, the accuracy of this estimate is compromised by the fact that it is not based on the underlying forces that drive economic growth.

Despite these weaknesses of basic statistical estimates, the use of descriptive statistics allows the analyst to pose formal hypotheses and validate them using

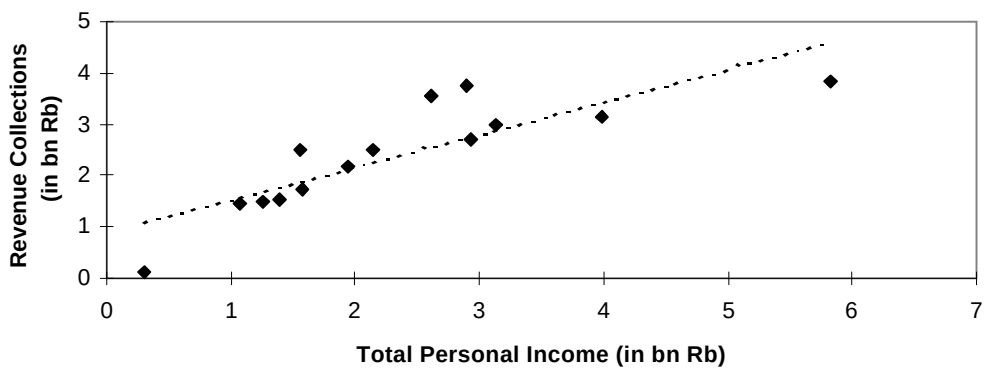
scientific and objective methods. As a result, the use of elementary statistical analysis is an important complement to expert analysis in the process of making multi-year economic and fiscal estimates.

2.3 Ordinary Least Squares Regression Analysis

Regression analysis is an extremely useful procedure that quantifies the relationship between one variable (the "dependent" variable) and one or more other "independent" variables. For example, we could employ regression analysis in the context of the revenue estimation to quantify the relationship between the amount of revenue collections for a certain tax and one or more variables that -- directly or indirectly -- influence the level of revenue collections.

A simple example may be illustrative. Figure 1 shows the relationship between total personal income and revenue collections for a group of regions. Each dot in the figure represents the amount of revenue collections and the level of personal income for a number of years. Just looking at the pattern of the dots, it is obvious that there exists a positive relationship between the level of revenue collections and the level of personal income. In order to quantify this relationship we could, for example, draw an straight line through the group of dots, and measure the slope and intercept of this line. This is in essence what a regression achieves.

Figure 1: Example of Regression Analysis



In this simple example, the relationship between revenue collections and personal income can be expressed by the equation:

$$\text{Tax Revenue} = B_0 + B_1 \cdot (\text{Personal Income}) + \text{errors}$$

The ordinary least squares (OLS) regression procedure will compute the values of the parameters B_0 and B_1 (the intercept and slope, respectively) that best "fit" the observations. In the given example, the parameters were calculated as $B_0 = 0.9$ and $B_1 = 0.6$.

Obviously, no straight line can exactly run through all of the points. The vertical distance between each observation and the line that fits "best" (the regression line) is referred to as error. The OLS regression procedure calculates the parameter values by minimizing the sum of the squared errors for all observations. Hence the name of the regression method: ordinary least squares.

OLS analysis has many desirable properties; it is the most reliable method of estimating linear relationships between economic variables. OLS is quite robust and can be utilized in a variety of environments, but unlike expert analysis or descriptive statistics, is not an "intuitive" approach to the data. OLS regressions do have limitations; for reasons discussed later, OLS should not be used if it appears that the data are correlated across time, that is, if the current period's value is a function of the value in the previous period.

When only two variables are included in a regression (one dependent and one independent variable) this is referred to a bi-variate regression. Exactly the same approach can be used when more than one independent variable is used; this is referred to as a multi-variate regression. The statistical properties of regression analysis are theoretically derived in Section 4, while Section 6 describes how to perform an OLS regression in Microsoft Excel. The remainder of this section discusses different forecasting techniques that rely on OLS regressions.

2.4 Forecasting Technique: Time-Trend Models

One OLS technique that is discussed in this note is time-trend modeling. Time-trend modeling attempts to explain the movement in a variable as a function of time. The trend variable is merely a counter, increasing by one for each time period (monthly, quarter, year). For example, one could imagine that the level of tax compliance (the degree to which taxpayers comply with the tax laws) is steadily increasing (or decreasing over time). Therefore, in this example,

$$\text{Tax Compliance} = B_0 + B_1 \text{ ?Year.}$$

Time trends can be quite effective in modeling variables that appear to be steadily increasing or decreasing over time. For example, later in this note we will discuss how to model the Consumer Price Index (CPI) with an increasing time- trend model since it appears that the CPI for the Russian Federation is steadily increasing over time. However, using a time-trend model will not provide the analyst with the same level of accuracy and understanding of the underlying relationships of the dependent variable with the economy.

2.5 Forecasting Technique: Exponential Smoothing

Another OLS technique that we will discuss in this note is exponential smoothing. Exponential smoothing tries to predict the change in a variable by looking at how this variable has changed in the past. In general,

$$\hat{y}_{t+1} = \alpha\, y_t + \alpha\,(1-\alpha)\, y_{t-1} + \alpha\,(1-\alpha)^2\, y_{t-2} + \ldots + \alpha\,(1-\alpha)^{t-1}\, y_1 + (1-\alpha)^t\, y_0$$

In essence, the dependent variable is defined as a weighted average of its past values, where the past values are weighted as a function of a parameter, ? . The value of ? is determined by the regression. Exponential smoothing is quite useful in updating forecasts generated by other processes but should not be used to generate estimates for multi-year periods. However, one could use this technique to update multi-year forecasts for one additional year that were generated using some other approach.

2.6 Forecasting Technique: Bi- and Multi-Variate Regression Models

Certain economic models attempt to examine the relationship between one variable of interest (the dependent variable) and one or more other variables (independent variables) using an OLS regression. A common example is that a person's monthly income is expected to be influenced by his or her age, years of education, years of work experience, geographical factors and gender. Given a sufficient amount of data, a multi-variate regression model would be able to compute the relative influence of all these factors on a person's monthly income.

We can use the bi-variate models (a regression model with only one independent variable) when we wish to investigate a specific relationship between variables, or when data is simply unavailable for other relevant variables. Multi-variate models, on the other hand, require more complex data sets but are often the most powerful models in assisting the analyst in understanding the underlying relationships between economic variables. One disadvantage to using regression analysis is that it is the least intuitive of the methodologies discussed in this note. It presents the analyst with many challenges and policy makers are likely to be less receptive to relying on the regression analysis approach for policy purposes.

2.7 Time-Series Analysis and Other Advanced Techniques

A variety of other, more complex regression techniques are available to forecast economic and fiscal variables, such as time-series analysis. These more complex techniques are not discussed in this note. In many cases, the data series are not of sufficient length to support these more advanced forms of statistical analysis. Second, the data series that are available for the Russian Federation are often subject to a high degree of volatility, invalidating the use of more sophisticated techniques. Finally, Microsoft Excel does not support these techniques, thus it would be impossible to demonstrate these methods without employing another software package.

3. Theoretical Review: Descriptive Statistics

The purpose of this section is to review some basic statistical concepts and methods that are used in the following sections. As mentioned before, the theoretical sections are included in this document for completeness and may serve to refresh the reader's memory of the basics of statistics. Readers who are familiar with these concepts may choose to skip this chapter and continue with Chapter 4 which presents the theoretical underpinnings of regression analysis, or straight to Chapters 5 and 6, which discuss the application of the estimation techniques in Microsoft Excel.

The remainder of this section will present a general overview of descriptive statistics such as measures of central tendency and dispersion; calculating confidence intervals; and hypothesis testing.

3.1 Descriptive Statistics - General

The primary objective of any statistical analysis is to make statistical inferences, that is, to draw testable conclusions about a population based upon information contained in a sample drawn from the population of interest. A **population** is any set of items of interest, from the population of taxpayers to the population of factories producing heavy machinery. A **sample** is subset of items drawn from the population of interest. Any characteristic of the population is called a **parameter** while any characteristic of the sample is denoted as a **statistic**. A statistic is merely an estimate, good or bad, of a population parameter.

Upon drawing a sample from the target population, the data contained in the sample can be examined using **descriptive statistics**. Descriptive statistics, to include such statistics as the sample mean, sample standard deviation, and sample variance, are useful in representing the data in recognizable terms. Statistical inferences are then made through two types of methods: **estimation** and **hypothesis testing**. Estimation is concerned with calculating the specific value of an unknown population parameter. Hypothesis testing is concerned with making a decision about a hypothesized value of an unknown population parameter.

3.2 Descriptive Statistics - Measures of Central Tendency

A descriptive statistic may be defined as any single statistical measure, derived from a sample of data, that is designed to illustrate a specific feature of the sample data set. The most common types of descriptive statistics are measures of **central tendency** and **variability**.

We first examine the measures of central tendency as these are necessary to calculate measures of variability. Measures of central tendency are used to calculate the 'average' value of a variable in the sample in easily recognizable units. Several measures of central tendency are commonly used, to include the sample mean, sample median, and sample midrange. In this subsection, we examine how each of these

measures of central tendency is calculated.

Sample Mean: The sample mean or 'average' of a variable X is noted as $\bar{X}$ and can be expressed as:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.1}$$

where n is the measure of sample size, X_1, X_2, ..., X_n are the n independent observations of X, and ? denotes summation. For example, if we had a sample of 50 observations on the federal share of Value Added Tax collections on a monthly basis, then $n = 50$, X_1, X_2, ..., X_n would each be equal to one specific month in the sample, and the sample mean would represent average federal Value Added Tax collections for the time period covered by the sample.

Sample Median: The sample median or 'middle' is equal to the middle ranked observation of the variable X and is denoted as **M**. For example, if we had a list that ranked all subjects of the Federation by their wealth, the sample would contain 89 observations. The median observation, **M**, would be equal to the 45th observation. If the data sample has an even number of observations, the median would be calculated by averaging the two middle observations.

Sample Midrange: The sample midrange measures the distance between the highest and lowest observations for the variable X and is denoted as **MR**. The sample midrange can be expressed as:

$$MR = \frac{X_{max} - X_{min}}{2} \tag{3.2}$$

The reader should note that the sample mean is often preferred to the sample median or sample midrange since the sample mean utilizes all the observations in the sample while the sample median only presents information on one observation and the sample midrange is the average of the difference between the two outlying observations.

This being said, the sample mean is not always the preferred measure of central tendency. Sometimes, the sample mean is subject to the influence of outliers. Outliers are observations that deviate significantly from the common pool of observations. In the event that a variable contains widely dispersed values, the other measures of central tendency (especially the median) may convey more pertinent

information than the sample mean. In the next section, we discuss measures of variability that the reader can use to determine what measure of central tendency best represents the sample data in question.

3.3 Descriptive Statistics - Measures of Variability

As just discussed, measures of central tendency are used as a measure of the "average" level of a variable. Measures of variability are used to determine how "spread out" the observations are around that "average." As more observations deviate from the common pool of observations, the dispersion or **variance** of the variable in question increases. In this subsection, we will examine two common measures of variability, the sample variance and sample standard deviation.

Sample Variance: The sample variance, S_x^2 , measures the variability of a variable X in terms of the average of the squared deviations of the sample observations from the sample mean. The sample variance can be expressed as:

$$\textit{Sample Variance} = S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{3.3}$$

where n is the number of observations in the sample; X_1, X_2, ..., X_n are the independent observations for the variable X; and $\bar{X}$ is the sample mean.

One problem with the sample variance is that it is expressed in squared units of the sample variable. To eliminate this problem so as to have a measure of the variability expressed in the same units as the variable of interest, we can use the sample standard deviation.

Sample Standard Deviation: The sample standard deviation, S_x, is the square root of the average of squared deviations of the sample observations from the sample mean. Simply put, the sample standard deviation is the square root of the sample variance or:

$$\textit{Sample Standard Deviation} = S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{3.4}$$

As before, n is the number of observations in the sample; X_1, X_2, ..., X_n are the independent observations for the variable X; and $\bar{X}$ is the sample mean.

Coupled with the measures of central tendency, the measures of variability

provide us with an assessment of the average value and dispersion of the variable of interest. However, calculating these measures by hand can prove quite tedious, especially in large samples. As illustrated in Section 5, we can use Microsoft Excel to circumvent this manual calculation process and also to reduce the probability of human error that arises when calculations are done by hand.

3.4 Estimating Confidence Intervals

In the previous two subsections, we defined the sample mean and sample variance, among others. The process of developing specific estimates for a unknown population parameter of interest is called **estimation** and the specific estimates are called **point-estimates**. Our line of inquiry now focuses on the probability that our point-estimates lie within a range of values that are likely to contain the true value of the population parameter. Simply put, we want to know whether our point-estimate is a reasonable estimate of the population parameter.

In general, to estimate a parameter of interest, the standard procedure is to draw a random sample of observations from the population of interest, develop the point-estimate for the population parameter, and then calculate a **confidence interval** for the population parameter. A confidence interval is a range of values within which we have a predetermined level of confidence that the value of the population parameter lies within the range covered by the confidence interval.

More technically, for any population parameter ?, we can develop a random variable $\hat{\theta}$, where $\hat{\theta}$ is the point-estimate of the population parameter ?. We utilize the '^' sign to differentiate between the point-estimate and the population parameter. The size of the confidence interval for the point-estimate is dependent upon what level of confidence with which we wish to be certain. For example, we could set the lower and upper limits of the confidence interval in a manner so that we know with 90 percent confidence that ? falls inside the confidence interval. This means that if we were to take repeated random samples of similar size from the same population, 90 percent of such intervals would statistically be expected to contain the parameter ?.

To calculate a confidence interval, three variables must be known. First, what is the level of confidence? Second, what is the point-estimate of the parameter? Third, what is the estimated standard error of the point estimate?

Confidence Intervals

When developing a confidence interval, note that:

- ? Parameter ? is a fixed and unknown constant, based on the full population;
- ? The lower and upper limits of the confidence interval are random quantities; so
- ? The boundaries of the confidence interval vary from sample to sample.

First, the desired level of confidence needs to be determined. The higher the level of confidence, the more likely that the confidence interval will contain ?. However, a trade-off exists, in that choosing a higher level of confidence will increase the distance between the boundaries of the confidence interval. For any given level of confidence, let ? = (1-Level of Confidence), where ? is the significance level of the confidence interval.

Second, determine the point-estimate ($\hat{\theta}$) of the parameter. As discussed previously, the point-estimate is a specific value calculated using a formula and a random sample of data drawn from the population of interest. A point-estimate may be an estimate of the population mean, population variance, population median, or a variety of other descriptive statistics.

Third, what is the estimated standard error the point-estimate? Slightly different than the sample standard error, the estimated standard error for a point-estimate can be expressed as:

$$\frac{S_x}{\sqrt{n}} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} \tag{3.5}$$

where S_X is the sample standard deviation of the variable X, n is the number of observations in the random sample, and $\bar{X}$ is the sample mean of X.

For any point-estimate $\hat{\theta}$, the boundaries of the confidence interval are defined as:

$$\hat{\theta} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{S_X}{\sqrt{n}} \tag{3.6}$$

where t represents the t distribution with n - 1 degrees or freedom at ? significance level. The respective values of the t distribution are contained in many standard statistical books or can be computed using Microsoft Excel.

3.5 Hypothesis Testing for the Mean, Median, and Midrange

In the previous section, we discussed how to estimate a confidence interval for the point-estimator of a population parameter. Given an estimated confidence interval, we can determine whether or not the point-estimator is 'reasonable' estimate

of the population parameter. For example, if we developed a confidence interval for an estimate for the sample mean, we would be able to determine whether the sample mean was within the plausible range of values for the population mean.

However, the range of plausible values may be relatively large if one wishes to make inferences about a specific value. If we wished to test whether the population parameter was equal to X_1 instead of X_2, where both X_1 and X_2 were within the confidence interval for the population parameter; we would not be able to conduct this test using the confidence interval approach.

We can, however, use hypothesis testing to test whether the estimated value of a parameter of interest varies significantly from the hypothesized value. Simply put, we can use hypothesis testing to conclude whether or not the population parameter is equal to X_1 instead of X_2.

As with the estimation of a confidence interval, we must first answer a series of question in order to test a hypothesis. First, what are the null and alternative hypotheses? Second, what is the significance level of the hypothesis test? Third, what test statistic will be used to test the null hypothesis? Fourth, what is the critical value of the test statistic? From these questions, we can compute the value of the test statistic and determine whether or not to reject the null hypothesis.

<u>Null and Alternative Hypotheses</u>: In general, hypothesis testing involves the statement of a rule, expressed in terms of the data, that dictates whether or not the null hypothesis should be rejected or not. Traditionally, the null hypothesis (H_o) is the statement that we wish to examine and the alternative hypothesis (H_a) is the conclusion that we will draw if we reject the null hypothesis. Let us assume that we wish to test whether the population mean income is equal to $54,000. Our null and alternative hypotheses would then be:

$$\begin{aligned} H_O &: \textit{Mean Income} = \$54{,}000 \quad \textit{or} \quad \mu = \$54{,}000 \\ H_A &: \textit{Mean Income} \neq \$54{,}000 \quad \textit{or} \quad \mu \neq \$54{,}000 \end{aligned} \tag{3.7}$$

<u>Significance Level of a Hypothesis Test</u>: Ideally, we would like to eliminate the possibility of incorrectly rejecting the null hypothesis when it is true. However, a tradeoff exists in that as we decrease the likelihood of incorrectly rejecting the null hypothesis, we increase the probability that we will fail to reject the null hypothesis when it is false. Thus, the best possible course of action is that for a given probability of rejecting the null hypothesis when it is true (?), the testing procedure used has the smallest probability of failing to reject the null hypothesis when it is false (?). We denote ? as the size of the test and ? is also known the significance level of the test. The power of the test is expressed as (1 - ?).

In our current example, we wish to test whether mean population income is equal to $54,000. As previously discussed with respect to developing a confidence

interval, we wish to have a predetermined level in confidence in the test. For this example, we will test the null hypothesis at a 90% level of confidence or 10% significance level.

<u>The Test Statistic</u>: Recall that the general form of the confidence interval is:

$$\hat{\theta} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{S_X}{\sqrt{n}} \tag{3.8}$$

and that the range of plausible values for the population parameter of interest is given by Equation 3.8. We can easily manipulate Equation 3.8 to construct the following test statistic:

$$t_{n-1,1-\frac{\alpha}{2}} = \left| \frac{\hat{\theta} - \theta}{\frac{S_X}{\sqrt{n}}} \right| \tag{3.9}$$

where $\hat{\theta}$ is the point estimator of ? ; S_X is the estimated standard deviation; and *n* the number of observations in the sample. As with the construction of the confidence interval for the sample mean, the T-statistic is distributed with *(n-1)* and (1-? /2) degrees of freedom.

Using the point-estimate for the population parameter of interest, we will be able to test the null hypothesis. If the calculated T-statistic in Equation 3.9 is less than the critical T-statistic, then we "fail to reject" the null hypothesis. Failing to reject the null hypothesis leads one to conclude, based upon the sample information, that it is likely that the population parameter is equal to the hypothesized value in the null hypothesis. If the calculated T-statistic in Equation 3.9 is greater than the critical T-statistic, we would then "reject" the null hypothesis, and we would conclude, based on the sample information, that the population parameter is not equal to the value contained in the null hypothesis.

4. Theoretical Review: Regression Analysis

4.1 Introduction

As discussed in Section 2, linear regression models are used to study the relationship between a **dependent** variable and a set of **independent** variables. Multi-variate regression analysis attempts to explain variations in the dependent variable **y** through movements in the *k* explanatory (independent) variables $\mathbf{x_1, x_2, ..., x_k}$. The general form of the multi-variate linear regression model is:

$$\begin{aligned} y_i &= f\,(\,x_{i1}\,,\,x_{i2}\,,\,\ldots\,,\,x_{ik}\,) + \epsilon_i \\ y_i &= \beta_0 + \beta_1\,x_{i1} + \beta_2\,x_{i2} + \ldots + \beta_k\,x_{ik} + \epsilon_i \end{aligned} \tag{4.1}$$

The model incorporates the randomness that is prevalent in the real-world, hence **y** is dependent not only on $\mathbf{x_1, x_2, ..., x_k}$ but also on the stochastic element **?**, hence y is a function of **x** and **?** or **y** = f(**x,?**).

The remainder of this chapter will discuss the assumptions of the multi-variate model, the derivation of the model's estimators, and a measure of the model's "fit.". Since a time trend model is simply a regression model that uses a measure of time as its only independent variable, this approach requires little further theoretical explanation. Exponential smoothing is discussed in greater detail in subsection 4.5. Instructions on how to estimate regressions with Microsoft Excel are discussed in Section 6.

4.2 Assumptions of the Multi-Variate Model

The purpose of this section is to briefly review the underlying assumptions of the classical linear regression model. In general, these assumptions address the following issues:

1. Linear functional form;

2. Identifyability of the model parameters;

3. Expected value of the disturbance given observed information;

4. Variances and covariances of the disturbances give observed information;

5. Nature of the sample data with respect to the independent variables, and;

6. Probability distribution of the stochastic part of the model.

The purpose of these assumptions is to describe the form of the multi-variate linear regression model and the relationships among its components. Furthermore, these assumptions dictate the appropriate estimation and inference procedures and are therefore critical to the model.

Linear Functional Form: In the context of the multi-variate linear regression model, the assumption of *linearity* refers to the manner in which the regression parameters (? 's) and the disturbance term (?) enter the regression equation. In general, the assumption of linearity means that the function $y = f(x_1, x_2, ..., x_k)$ has to be expressed as a linear equation, or transformable into a linear equation. However, by transforming the independent variables, one can estimate non-linear functions. For example, the equations:

$$y = ?_0 + ?_1 x_1 + ?$$

$$y = ?_0 + ?_1 \log(x_1) + ?$$

$$y = ?_0 + ?_1 x_1 + ?_2 x_1^2 + ?$$

$$y = e^{?0} x_1^{?1} x_2^{?2} \quad (\text{equivalent to } \log(y) = ?_0 + ?_1 \log(x_1) + ?_2 \log(x_2) + ?)$$

are all linear in some function of **x** and meet the linearity assumption. More formally, the linearity of the regression model assumption is stated as:

$$y = X\beta + \epsilon \tag{4.2}$$

where y is a T x 1 vector containing the independent variable, X is a T x k matrix of independent variables and ? is the k x 1 vector of parameters; ? is an T x 1 vector of the stochastic error.

Identifyability of the model parameters: The second assumption is that no exact linear relationship exists among the variables in the model; the columns of the matrix **X** are linearly independent. This means that none of the independent variables in **X** are equal to a linear combination of one or more of the other independent variables in **X**. More formally, **X** is a *n* x *k* matrix with rank *k*.

Expected value of the disturbance given observed information: Given the matrix of independent variables **X**, we further assume that the disturbance term ? has an expected value of zero at every observation. In principle, the mean of each disturbance term $?_i$ conditioned on all observations x_i is zero, that is, no observations in **X** convey information about the expected value of the disturbance term. This implies that the disturbance terms are 'white noise' or purely random draws from

some population of disturbance terms. This assumption can be formally expressed as:

$$E[\epsilon \mid X] = E[\epsilon_1 \mid X] = E[\epsilon_2 \mid X] = \ldots = E[\epsilon_n \mid X] = 0 \tag{4.3}$$

Variances and covariances of the disturbances give observed information: The fourth assumption concerns the variances and covariances of the disturbance terms:

$$Var[\epsilon_i \mid X] = \sigma^2 \qquad for\ all\ i = 1, \ldots, n \tag{4.4}$$

$$Cov[\epsilon_i, \epsilon_j \mid X] = 0 \qquad for\ all\ \ i \neq j \tag{4.5}$$

We assume that the data has constant variance or is **homoscedastic** and displays no **autocorelation**. Note that these assumptions pertain to the variances and covariances conditional upon the independent variables. These assumptions, which are relaxed in more refined models, imply that (1) the variation in the disturbance terms is constant, and (2) disturbances are truly independent, that is, a preceding or succeeding disturbance has no influence on another disturbance term.

Nature of the sample data with respect to the independent variables: We further assume that the independent variables in the observed sample are non-stochastic, that is, the explanatory variables are a known series of constants. This assumption, which is equivalent to stating that the explanatory variables are fixed in repeated samples, allows us to disregard the sources of variation in the explanatory variables and focus on the relationship between *y* and **X**.

Probability distribution of the stochastic part of the model: The final assumption of the classical linear regression model is that the disturbances are normally distributed with zero mean and constant variance. This adds the assumption of normality to the assumptions about the mean, variance, and covariance of the disturbance terms conditional on the matrix of explanatory variables.

4.3 Deriving the Estimator

As previously discussed, the objective of the least squares method is to minimize the sum of the squared errors, that is, to minimize the sum of the distances between the observed and predicted values of **y**. This objective function can be written as:

$$e'e = (y - \hat{y})'(y - \hat{y}) = (y - X\beta)'(y - X\beta) \tag{4.6}$$

where e is a T x 1 vector of residuals, and e' is the transpose of e. Thus, e'e results is a scalar result, that is, the sum of squared errors.

To minimize (4.6), we must first obtain the first order conditions. We differentiate (4.6) with respect to ? and setting the resulting expression equal to 0 results in:

$$\frac{\delta(e'e)}{\delta\beta} = 2X'(y - X\beta) = 0 \tag{4.7}$$

We can proceed to solve (4.7) by dividing through by 2 and multiplying through by the transpose of the **X** matrix (X'):

$$\begin{aligned} 2X'(y - X\beta) &= 0 \\ X'y - X'X\beta &= 0 \\ X'y &= X'X\beta \end{aligned} \tag{4.8}$$

In order to solve (4.8) for ?, we must multiply both sides by the inverse of X'X, which results in the least squares estimator:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4.9}$$

<u>Statistical Properties of the Estimator</u>: We can now examine the properties of the least squares estimator under the standard set of assumptions. We are concerned about the least squares estimator's properties in that the estimator measures the true population parameter and we wish to discern how accurately this measurement occurs. Note that in order to examine the properties of the least squares estimator, we must assume that the linear model is, in itself, the correct model to estimate the population parameters.

We can use the following four measures to evaluate the statistical properties of the least squares estimator:

1. Unbiasedness

2. Consistency

3. Variance

4. Mean Square Error

Unbiasedness: An estimator is unbiased if its expected value is equal to the population parameter, that is, $\hat{\beta}$ is unbiased if $E[\hat{\beta}] = E[\beta]$. On average, an unbiased estimator will be equal to the true value of the population parameter.

Consistency: An estimator is consistent if the estimator converges in probability to the population parameter. In essence, as we increase the sample size to infinity, the probability that the estimator deviates from the true population parameter approaches zero.

Variance: As the variance of an unbiased estimator declines, the precision of the estimator increases. Intuitively, if we have two unbiased estimators, the estimator with the smaller variance is the more accurate estimator.

Mean Square Error: Mean square error is equal to the variance of an estimator plus the square of its bias or:

$$MSE = var(\hat{\beta}) + bias(\hat{\beta})^2 = var(\hat{\beta}) + (E[\hat{\beta}] - \beta)^2 \qquad (4.10)$$

As with the variance measure, as the Mean Square Error of an estimator declines, the precision of the estimator increases. Thus, it is possible for an unbiased estimator with a relatively large variance to be less precise than a biased estimator with a relatively small variance and bias.

Thus, coupled with the assumptions of subsection (4.2) and the assumption that the columns of the **X** matrix are linearly independent, we know:

1. The model is well-defined.

2. On average, the omitted variables, sampling error, and other forms of error have no impact on the dependent variable.

3. The least squares estimator is consistent.

4. If the columns of **X** are independent of the error terms, then the least squares estimator is also unbiased.

5. The variance of the least squares estimator is $var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$.

These assumptions and results are sufficient to show that, within the class of linear, unbiased estimators for ? , the least squares estimator has minimum variance, and is thus the most precise estimator for ? .

4.4 Determining the "Fit" of a Model and Parameters

Several measures exist to determine how well a model 'fits' the observed data. We could run several models and compare the Residual Sum of Squares and choose the model that minimizes the Residual Sum of Squares. However, it can be shown that the Residual Sum of Squares can be scaled arbitrarily just by multiplying the values of **y** by the desired scaled factor. Another measure that may be used is the **coefficient of determination** or $\mathbf{R^2}$.

To calculate R^2 , we must first determine how well a specific model 'captures' the variation in the dependent variable. Let the Total Sum of Squares be equal to:

$$SST = \Sigma\ (y_i - \bar{y})^2 = SSR + SSE \tag{4.11}$$

that is, the total variation in **y** is defined as the sum of the squared deviations of **y** from its mean. The Total Sum of Squares is also equal to the Regression Sum of Squares (SSR) plus the Residual Sum of Squares (SSE).

Proceeding from the definition of the Total Sum of Squares, we can define the Residual Sum of Squares as:

$$SSE = \Sigma\ e_i^2 = \Sigma\ (\ y_i - \hat{y}_i\ )^2 \tag{4.12}$$

and the Regression Sum of Squares as:

$$SSR = \hat{\beta}\ \Sigma\ (\ x_i - \bar{x}\ )^2 = \Sigma\ (\ \hat{y}_i - \bar{y}\ )^2 \tag{4.13}$$

A regression will "fit" well if the deviations of **y** from its mean are more largely explained by deviations of **x** from its mean than by the residuals. R^2 is a measure that captures this concept and is equal to:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{4.14}$$

The coefficient of determination (R^2) will vary between zero, in which case the

independent variables have no explanatory power whatsoever, and unity, in which case the independent variables explain all variation in the dependent variable.

An other important measure how well the regression is performing is whether each parameter has a statistically significant influence on the dependent variable. We can use Equation 3.6 from Section 3 to develop test statistics for the estimated intercept and slope coefficients. To test the null hypothesis whether an estimated regression coefficient is independently equal to zero, we can use the following equations:

$$t(\hat{\beta}) = \frac{(\hat{\beta} - 0)}{se(\hat{\beta})} = \frac{\hat{\beta}}{se(\hat{\beta})} \tag{4.15}$$

Note that for a two-tailed test where the null hypothesis is that the estimated coefficient is equal to zero, the calculated T-statistic is equal to the estimated regression coefficient divided by its sample standard error.

4.5 Exponential Smoothing

Exponential smoothing is used for continually updating a forecast to account for recent changes in the data series of interest. Exponential smoothing applies to data series without time-trends and to series with both linear and nonlinear time-trends. This method of forecasting is primarily used to update or revise forecasts generated by other processes and is best used for updating forecasts for one period ahead. The caveat is that exponential smoothing should not be used for long-term forecasts.

Exponential smoothing is most powerful when the data series has been subject to large, unexpected shocks or innovations. The reader should be aware that even a series governed by a strong linear time-trend may be subject to wide variations around the time-trend. While a linear time-trend model may depict the general, long-term movement of the data series of interest, point forecasts based on the linear model may, in fact, miss the unexpected fluctuations in the series. An updated forecast generated through the use of exponential smoothing may provide more accurate short-term forecasts.

With exponential smoothing, the revised estimate for the next forecast period is simply the sum of the estimate for the current period and a portion of the forecast error in the current period or:

$$\hat{y}_{t+1} = \hat{y}_t + \alpha\,(y_t - \hat{y}_t) \tag{4.16}$$

where ? is the weight given by the exponential smoothing function to the forecast error. Sensible values for ? range from 0 to 1, although ? can vary outside this range. We can rewrite Equation 4.16 so that the updated forecast is the weighted sum of all past values of the series with the weights equal to powers of the ? parameter or:

$$\begin{aligned} \hat{y}_{t+1} &= \alpha\, y_t + \alpha\,(1-\alpha)\, y_{t-1} + \alpha\,(1-\alpha)^2\, y_{t-2} \\ &+ \ldots + \alpha\,(1-\alpha)^{t-1}\, y_1 + (1-\alpha)^t\, y_0 \end{aligned} \tag{4.17}$$

Consider the impact of a large ? . A large ? will result in more recent observations having more influence on the new forecast relative to older observations. Thus, forecasts developed using large values of ? will respond more quickly to changes in the data series than forecasts developed with a small ? . The tradeoff is that forecasts developed with large ? values will be subject to the influence of large shocks or innovations in the data series. However, in most cases, it is not necessary to estimate the ? value in that the best choice for ? is the value that minimizes the sum of squared prediction errors.

5. Statistical Procedures in Microsoft Excel

5.1 Introduction

The User Manual for the *Multi-Year Budget Model for the Russian Federation* contains a section with general directions for the use of Microsoft Excel. This section complements and builds on the general instructions contained in the User Manual. Readers who are familiar with Microsoft Excel and the use of basic statistical operators may choose at move to the next section at their discretion.

Accompanying this note is a floppy disk containing this note and a Microsoft Excel data file. We will use the Microsoft Excel data file throughout this note to illustrate how to calculate various functions and to estimate relationships between variables of interest. Our first step is to open this data file. At this point, start the Microsoft Excel program.

Opening a Workbook: To open the data file on the floppy disk, you must use the **OPEN** command. Look at the top of Microsoft Excel program for a list of commands, to include **File, Edit, View, Insert, Format**, among others. Taking the mouse, point at the **File** command and click on **File**. When the **File** menu opens, look for the **Open** command and point and click on the **Open** command. A rectangular window will appear and at the top of the window is a small box with **Look In** to the left of the small box and a arrow to the right of the small box. Click on the small arrow to see a list of drives available to you at this period in time. Look for and click on the 'A' or floppy drive. When the 'A' drive opens in the rectangular window, look for the file called **Examples**. Click on this file and then click on **Open** so that Microsoft Excel will open the appropriate file. Figure 5.1 illustrates the **Example** data file open in Microsoft Excel.

Workbooks and Worksheets: Note that in Figure 5.1 and in the open data file in Microsoft Excel that there are a series of tabs starting in the bottom left hand corner of the data file. In Microsoft Excel, the entire data file is called a **Workbook**. A workbook consists of several individual **Worksheets** which contain data, charts, and other items of interest. To move from one worksheet to another, simply click on the tab of the worksheet that you wish to move to and this worksheet will appear. The active worksheet is highlighted in white as illustrated in Figure 5.1.

Worksheet Organization: Each worksheet consists of a spreadsheet of rows and columns which resembles a chessboard. As illustrated in Figure 5.1, each row is numbered and each column is represented by a letter. A specific **cell** is defined by its location in the spreadsheet, so in this example, cell B2 contains the number 4210.40.

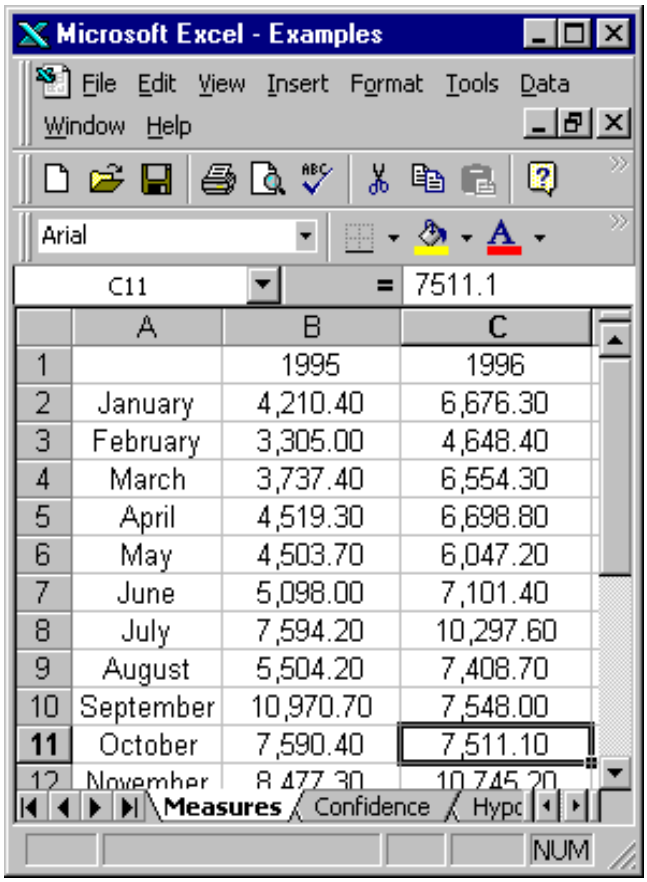
Entering and Deleting Data: To enter data in Microsoft Excel, the reader merely has to select the cell into which the data is to be entered. Selecting a specific cell is done by pointing the mouse arrow at the cell and then clicking on the specific

cell of interest. Once the cell is highlighted (see cell C11 in Figure 5.1), type in the data and press ENTER. The data is now a part of the worksheet. To delete a specific data cell, click on the cell containing the data you wish to delete and press the DELETE button.

Now that we have covered the basic operations of opening an Excel Workbook, finding and selecting a Worksheet, and Entering and Deleting Data, we turn to calculating the measures of central tendency and variability using Microsoft Excel.

Figure 5.1

Illustration of Open Excel Workbook



	A	B	C
1		1995	1996
2	January	4,210.40	6,676.30
3	February	3,305.00	4,648.40
4	March	3,737.40	6,554.30
5	April	4,519.30	6,698.80
6	May	4,503.70	6,047.20
7	June	5,098.00	7,101.40
8	July	7,594.20	10,297.60
9	August	5,504.20	7,408.70
10	September	10,970.70	7,548.00
11	October	7,590.40	7,511.10
12	November	8,477.30	10,745.20

5.2 Calculating Measures of Central Tendency

In this subsection, we will examine how to calculate the measures of central tendency and variability using Microsoft Excel. In many cases, Microsoft Excel has "built-in" functions that we will be able to use to calculate the statistics of interest; in other circumstances, we will have to calculate these statistics without relying on the

existing functions in Microsoft Excel. For the remainder of this section, please refer to the **Measures** worksheet of the **Examples** data file.

The first task of this subsection is to calculate the measures of central tendency, to include the sample mean, sample median, and sample midrange. The **Measures** worksheet already contains the necessary data. 36 observations on the federal share of the Value Added Tax collections (in millions or Rubles) from January 1995 to December 1997 are contained in the **Measures** worksheet.

Note that the first data cell containing observations for Value Added Tax collections is cell B2 and that the data is separated into columns, where each column contains observations for one year. Thus, the observations are in cells B2 to B13, C2 to C13 and D2 to D13 for 1995, 1996, and 1997, respectively.

Calculating the Sample Mean: We can use the **average** function in Microsoft Excel to calculate the sample mean for a variable of interest. The general form of the **average** function is:

=average(range of cells)

The range of cells should refer to those cells that contain the observations for which we wish to calculate the sample mean. A range of cells is specified by typing the cell reference of the top-left corner cell of the range, followed by a colon, followed by the cell reference of the bottom-right corner cell of the range. Note that Microsoft Excel will not discern between two different variables, so caution must be exercised to correctly input the range of cells.

Let us first calculate the sample mean for the observations for 1995. As previously discussed, the observations for 1995 are contained in cells B2 to B13. To calculate the sample mean for 1995, we merely have to type the following formula into a blank cell and press ENTER:

=average(b2:b13)

If the sample mean for 1995 is calculated correctly, Microsoft Excel should return a value of 5,891.99. Since our data is denominated in millions of Rubles, we can state that average federal share of Value Added Tax collections for 1995 was 5,891.99 million Rubles. In the **Measures** worksheet, we have output this result to cell B15.

The reader should note that we have added a "=" sign to the **average** function. The "=" sign is an indicator that the text following the "=" sign is a function and not merely plain text. If we omit the "=" sign and press ENTER, we would observe that what we had typed appears in the cell instead of the desired descriptive statistic. All functions in Excel must be entered with a "=" sign.

How does this value compare to the sample mean for the entire sample of observations from 1995 to 1997? In order to calculate average Value Added Tax collections (Federal Share) from 1995 to 1997, we can modify the **average** formula to state:

=average(b2:d13)

which is entered to cell F15; Microsoft Excel should return of 8044.33. It does appear that average Value Added Tax collections (Federal Share) increased over time.

Calculating the Sample Median and Midrange: We can use the **median** function in Microsoft Excel to calculate the median observation for a range of values. Microsoft Excel will return the median value if there are a odd number of observations in the range of cells for which Microsoft Excel is to calculate the median value. If there are an even number of observations in the range of cells, Microsoft Excel will calculate the average of the two median observations instead of returning the values for the two median observations.

The general form of the **median** function is similar to the **average** function and is expressed as:

=median(range of cells)

or, for example,

=median(b2:b13)

returns the median value for monthly VAT collections for the year 1995, which equals 5145.65.

Calculating the sample midrange proves slightly more difficult, since Microsoft Excel does not have a built-in function for the calculation of the sample midrange. Note that computation of the sample midrange requires two values, the minimum and maximum observations for the variable of interest. Microsoft Excel does have two functions, **Min** and **Max**, that we can use to obtain the minimum and maximum observations for the Value Added Tax collections (Federal Share) series. Similar the average and median, the general form of the **Min** and **Max** functions are:

=min(range of cells)

=max(range of cells)

Recall that the sample midrange is merely the maximum observation minus the

minimum observation divided by two. Using 1995 as an example, we can calculate the sample midrange by typing:

=((max(b2:b13)-min(b2:b13))/2)

5.3 Calculating Measures of Variability

We now turn to the calculation of the measures of variability that were previously discussed in Section 2.3. Please continue to use the **Measures** worksheet in the **Examples** data file. The general form of the sample variance function and the sample standard deviation function are similar to that for the functions for the sample mean and sample median in that the general form is:

=var(range of cells)

and

=stdev(range of cells)

Continuing with our example of calculating the descriptive statistics for 1995 and the entire sample of observations, we can calculate the sample variance and sample standard deviation for 1995 by typing:

=var(b2:b13)

and

=stdev(b2:b13)

If the formulas were correctly entered, cells B18 and B19 should return the value of 5,227,486.78 for 1995 and 2,286.37, respectively. The sample variance and sample standard deviation for all observations are calculated in cells F18 and F19.

5.4 Calculating Confidence Intervals

In this subsection, we examine how to calculate confidence intervals for point-estimates using Microsoft Excel. The reader should now open the **Examples** data file that accompanied this note and go to the **Confidence** worksheet in the **Examples** file.

Introduction and Assumptions: The **Confidence** worksheet contains two

hypothetical samples that we will use to develop confidence intervals for several point-estimates of interest. Let us assume that each of the hypothetical samples are drawn from a population of 5,000 individuals and that the samples are drawn at random. The sample size is set at 100 individuals and we observe gross income for each individual in the sample. Due to time and resource constraints, we are unable to sample the entire population. For convenience, the data samples have been highlighted in red.

We are able, however, to draw inferences about the entire population from the observed samples. We can calculate descriptive statistics and develop confidence intervals for those statistics so that we know how confident we can be about their accuracy. In this fashion, we are able to make inferences about the population at large based upon a relatively small sample of observations.

Using our previous discussions on the measures of central tendency and variability, we can calculate the sample mean, sample median, sample midrange, sample variance and sample variance using Microsoft Excel. The results are presented in Table 5.1 and are also contained in cells D3 through D8 of the **Confidence** worksheet.

Table 5.1

Descriptive Statistics for Sample 1 in the Confidence Worksheet

Point-Estimator	Point-Estimate
Sample Mean	$50,928
Sample Median	$52,996
Sample Midrange	$48,523
Sample Variance	877,471,499
Sample Standard Deviation	$29,622
n	100

The reader should note that instead of counting the number of observations by hand, we have used the Count function in Microsoft Excel to confirm there are actually 100 observations in Sample 1. The general form of the Count function is:

=count(range of cells)

The Count function will return the number of non-blank, numerical cells only. Thus

it will not include cells that include text or invalid arguments in the numerical count. In order to determine the number of observations in Sample 1, we typed:

=count(b3:b102)

Recall that we previously stipulated that the sample standard deviation was equal to the square root of the sample variance. To illustrate this fact, we calculate the square root of the sample variance in cell D10 by using the SQRT function or:

=sqrt(D8)

Calculating the Confidence Interval for the Sample Mean: Recall from Equation 2.6 that in order to calculate a confidence interval, we must first determine the following four values:

- ? The Level of Confidence;
- ? The Estimated Standard Error for the Point-Estimate;
- ? The Square Root of the Number of Observations in the Sample;
- ? The T-Statistic at *n-1* and *1-? /2* degrees of freedom.

Let us assume that we wish to calculate a confidence interval for the sample mean at a 95% level of confidence. Recall that ? , which is the level of significance, is equal to (1- level of confidence), so ? is equal to 5%.

The estimated standard error for the sample mean, as defined in Equation 3.5, is equal to the estimated sample standard deviation (Equation 3.4) divided by the square root of the number of observations in the sample. As illustrated in the **Confidence** worksheet, the sample standard deviation is equal to $29,622 (Cell D7) and the square root of the number of observations in the sample is equal to 10 (Cell D10). Thus, the estimated sample standard error is equal to $2,962, which is the result in Cell D11 and is also illustrated in Table 5.2.

Table 5.2

Calculating the Confidence Interval for the Sample Mean

Point-Estimator	Point-Estimate
Sample Mean	$50,928
Estimated Standard Error	$2,962
Critical T-Statistic	1.984
Upper Bound for the Sample Mean	$45,050
Lower Bound for the Sample Mean	$56,805
n, n-1, ?	100, 99, 0.95

Recall from Equation 3.6 that the upper and lower bounds of the confidence interval are equal to the point estimate plus/minus the critical t-statistic multiplied by the estimated sample standard error. We now need to determine the value of the critical t-statistic in order to calculate the upper and lower boundary points. As illustrated in Equation 3.6, the t-statistic is distributed with *n-1* degrees of freedom and *1-(? /2)* level of confidence.

<u>The TINV Function</u>: We can use the Tinv function in Microsoft Excel to calculate the critical t-statistic. The general form of the Tinv function is:

=TINV(level of significance, degrees of freedom)

Note that instead of entering 1-(? /2) as the operator of the t distribution, we only have to enter ? in the Tinv function; by default, Microsoft Excel calculates the correct (two-tailed) t-statistic.

Given that we know that ? =0.5 (cell D13), and *n-1* = 99 (cell D15), we can calculate the critical T-statistic using either two of the following examples:

=TINV(d13,d15)

The result of this calculation is contained in cell D15 and the critical T-statistic is equal to 1.984.

Calculating the upper and lower bounds of the confidence interval for the sample mean is illustrating in cells D17 and D18 and the results are contained in Table 2. For example, the upper bound of the confidence interval is equal to:

=(D3 + (D11*D15))

or

$$\$50{,}928 + (1.984 * \$2{,}962) = \$56{,}805$$

We can now conclude that the sample mean of $50,928 is a reasonable estimate for the population mean and that in repeated samples, the population mean is likely to lie within the range between $45,050 and $ 56,805 with a level of confidence of 95 percent. Sample 2 in the **Confidence** worksheet presents one more example of how to calculate a confidence interval for the sample mean.

5.5 Conducting Hypothesis Tests in Microsoft Excel

In the previous subsection, we developed estimates for a range of statistically possible values for the sample mean. However, as evidenced by Sample 1 in the **Confidence** worksheet, the confidence interval (the range of values that are plausible) can be relatively large if one wishes to make inferences about a specific value. In the previous example, if we stated the hypothesis that the population mean was equal to $54,000 rather than the estimated value of $50,928, we would be unable to discern whether or not this hypothesis was indeed true using the confidence interval approach.

We can, however, specify a hypothesis test to determine whether the estimated value of a parameter of interest varies significantly from the hypothesized value. Simply put, we can use a hypothesis test to conclude whether or not the population sample mean is indeed equal to $54,000. For this section, please refer to the **Hypothesis** worksheet of the **Examples** data file.

Recall that as with the estimation of a confidence interval, we must first answer a series of question in order to test a hypothesis. First, what are the null and alternative hypotheses? Second, what is the significance level of the hypothesis test? Third, what test statistic will be used to test the null hypothesis? Fourth, what is the critical value of the test statistic? From these questions, we can compute the value of the test statistic and determine whether or not to reject the null hypothesis.

<u>Null and Alternative Hypotheses</u>: Given that we wish to test whether the population mean is actually equal to $54,000, the null and alternative hypotheses are:

$$H_O : \textit{Mean Income} = \$54{,}000 \quad or \quad \mu = \$54{,}000$$
$$H_A : \textit{Mean Income} \neq \$54{,}000 \quad or \quad \mu \neq \$54{,}000$$

<u>Significance Level of a Hypothesis Test</u>: In our current example, we wish to test whether mean population income is equal to $54,000. As previously discussed

with respect to developing a confidence interval, we wish to have a predetermined level in confidence in the test. For this example, we will test the null hypothesis at a 90% level of confidence or 10% significance level.

The Test Statistic: As with the estimation of a confidence interval for the sample mean, we can use the T-statistic to test the hypothesis of whether mean population income is equal to $54,000. If the calculated T-statistic is larger than the critical T-statistic, then we reject the null hypothesis and conclude that mean population income is unlikely to be equal to $54,000. On the other hand, if the calculated T-statistic does not exceed the critical T-statistic, then we fail to reject the null hypothesis, and thus we can conclude that, based on the sample information, that mean population income is indeed equal to $54,000.

Critical Value of the Test Statistic: As discussed earlier in this section, the critical value of the test statistic is determined by the number of observations in the sample (*n*) and the significance level of the test (?). Given that *n=100* and (?)=.10, the t-statistic can be calculated using the following function:

=tinv(.10, 99)

which is equal to 1.66. If the calculated test statistic exceeds 1.66, then it falls within the rejection area of the test and we may reject the null hypothesis at a 10% level of significance. This means that, in repeated samples, that we are likely to reject the null hypothesis 90% of the time. Likewise, if the test statistic does not exceed 1.66, then we fail to reject the null hypothesis and it is possible that mean population income is equal to $54,000.

Calculating the Test Statistic for the Sample Mean: As illustrated in cells D22-D31 of the **Hypothesis** worksheet and in Table 5.3, we have calculated the critical T-statistic for the null hypothesis that mean population income is equal to $54,000 using the expression in Equation (3.9):

$$T = \left|\frac{\hat{\theta}-\theta}{\frac{S_X}{\sqrt{n}}}\right| = \left|\frac{50{,}928 - 54{,}000}{\frac{29{,}622.15}{\sqrt{100}}}\right| = 1.037$$

Note that in the **Hypothesis** worksheet we have calculated the values for the sample mean, the sample standard error, and the number of observations using functions previously discussed. We then referred to these results in the calculation of the T-statistic. The following expressions was used in Microsoft Excel for the calculation of the critical T-statistic:

=abs((D3-D4)/D9)

where the **abs** function produces the absolute value of term inside the inside the parentheses.

Interpreting the Result: As illustrated in Table 5.3, the calculated value for the T-statistic (1.037) does not exceed the critical T-statistic of 1.66. We thus "fail to reject" the null hypothesis that mean population income is equal to $54,000 and conclude that mean population income is likely to be equal to $54,000. Note that we can never conclude with certainty that the population's mean income is equal to $54,000 because we do not observe the entire population and only may draw inferences about mean population income (or some other population parameter).

Table 5.3

Hypothesis Testing

Point-Estimator	Point-Estimate
Sample Mean	$50,928
Hypothesized Population Mean	$54,000
Estimated Standard Error	2,962
Critical T-Statistic	1.66
Calculated T-Statistic	1.04
n, n-1, ?	100, 99, 10%

5.6 Using the Data Analysis Tool to Produce Summary Statistics

Before proceeding to the next chapter, it will be useful to cover one more tool in Microsoft Excel, the **Data Analysis Tool (DAT)**. The Data Analysis tool comes with the Excel program, but is not installed as a standard component. Therefore if the Data Analysis tool appears to be missing from your version of Excel, you would have to install this component before you can use the procedures int this subsection. It should be pointed out that all the functions that the Data Analysis tool performs can be produced without installing the Data Analysis tool.

The Data Analysis Tool (DAT) automates many statistical functions in Microsoft Excel and, for simple statistical analysis, is more efficient than entering the statistical functions by hand. A note of caution, however, in that the DAT does not automate every function, so the DAT should be used to support statistical analysis and not replace the analyst's role in determining what are those appropriate statistics of interest. Please refer to the **DAT-STAT** worksheet in the **Examples** data file for

the remainder of this subsection.

Figure 5.2

Selection of Analysis Tools in the Data Analysis Tool Window

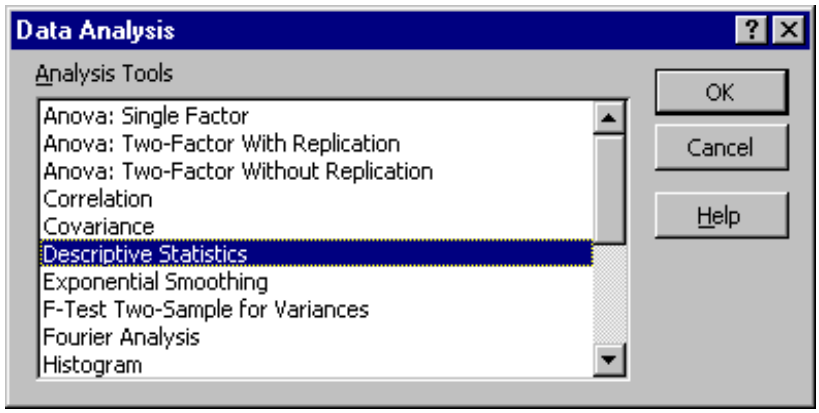
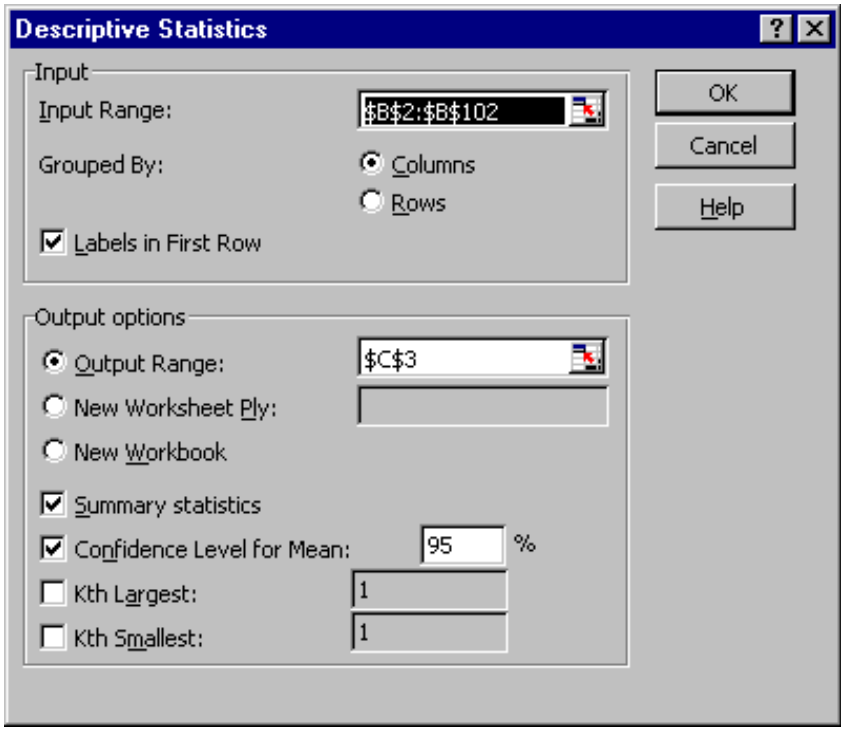


Figure 5.3

Calculating Summary Statistics Using the Data Analysis Tool



Operating the Data Analysis Tool: The DAT is quite simple to operate. Go to **Tools** and then click on **Data Analysis**. The Data Analysis tool will open as illustrated in Figure 5.2.

Note that there are a variety of statistical functions that we may use with the Data Analysis tool. Click on **Descriptive Statistics** so that is highlighted as in Figure 5.2 and click on OK. The Descriptive Statistics function window will open as illustrated in Figure 5.3.

Input Range: The first 'box' in the Descriptive Statistics function window is labeled *Input Range*. In this 'box' we will enter the range of data for which the descriptive statistics will be produced. Note that in the **DAT-STAT** spreadsheet the data range is B2:B102 and the first data cell (B2) contains the label for the data series on income.

There are two methods of entering the range of data. First, we can enter the range B2:B102 in the *Input Range* box. Second, note that at the right hand edge of the box is a small rectangle with an upward pointing arrow. If we click on this small box, another entry box will appear. We can now simply highlight the cells containing our data and then click on the small box with the downward pointing arrow to return to the Data Analysis tool.

Grouped By: The *Grouped By* consists of two inputs, whether the data is oriented by columns or by rows. The data in the **DAT-STAT** worksheet is column oriented, that is, the observations for one series are in one column. As illustrated in Figure 5.3, we select <u>Columns</u> since our data is oriented by column. If the observations for one series were contained in one row, then we would select <u>Rows</u> instead of <u>Columns</u>.

Labels in First Row: If the first row (or cell) of the data series is a label, then make sure that this box is checked else the DAT function will return an error message. Since cell B2 contains the label 'Income', we have checked the cell in the example illustrated in Figure 5.3.

Output: There are three *Output Options*. If *Output Range* is selected, the analyst should enter the reference for the upper-left cell where the output should be presented by Microsoft Excel. Be sure to allow for at least two columns and several rows for the output of the DAT. If the *New Worksheet* function is chosen, the output is directed to a new spreadsheet within the existing file and if the *New Workbook* is chosen, the output is directed to an entirely new file. As illustrated in Figure 5.3, we have chosen to send the output to cell C3.

Summary Statistics: If this box is checked, Microsoft Excel will output the sample mean, sample standard error, sample median, sample mode, sample standard deviation, sample variance, the minimum and maximum observations of the sample, the number of observations and several other statistics to the cell designated in the *Output* box. If this box is left unchecked, no summary statistics will be produced by the DAT.

Confidence Level for Mean: If this box is checked, the confidence level for the

sample mean is produced by Microsoft Excel. The default confidence level is 95%, although this change be changed by entering a new confidence level in the box by the percentage sign. Note that the confidence level must be added to/subtracted from the sample mean to produce the confidence interval for the sample mean.

K-th Largest and *K-th Smallest*: These functions produce the maximum and minimum observations if there are separate ranges of data in the *Input Range* of the DAT. If only one range of data is used, these two functions produce the maximum and minimum observations in the sample. If more than one range of data is used in the DAT, make sure to enter the number of separate data ranges in the boxes adjoining the *K-Largest* and *K-th Smallest* check boxes.

We have presented one example of the operation of the DAT in the **DAT-STAT** worksheet in the **Examples** data file. Recall that the results produced by the DAT are identical to the descriptive statistics computed using formulas.

6. Regression Analysis in Microsoft Excel

As discussed earlier in Section 2 and Section 4, a number of estimation techniques based on OLS regressions can be used to forecast economic and fiscal variables. In particular, we discussed time-trend modeling, exponential smoothing and bi-variate and multi-variate regression models. Since all of these techniques rely on OLS regressions, they can all be estimated using the standard linear regression function in Excel (LINEST). However, Excel allows the user to estimate time trends using the TREND function, while exponential smoothing can be performed with the GROWTH function.

6.1 Forecasting Using a Time-Trend Model: The TREND Function

The TREND function fits a straight line using the method of least squares to an array containing the dependent and independent variables. If the set of explanatory variables is expanded beyond the range of observations containing the known dependent and independent variables, the TREND function can produce forecasts of the dependent variable. In the time-trend model, expanding the range of observations requires only creating the values of the time-trend variable as far into the future as desired. For this subsection, please refer to the **Trend** worksheet in the **Examples** data file.

Array Producing Functions

Unlike the functions for the sample mean and sample variance, which return a scalar result (a single value), the TREND function returns a series of results in array format. An array formula performs multiple calculations and then returns multiple results.

Array functions are created in the same way that basic, single-value functions are created. However, before typing the function, the user should select the group of cells that will contain the function. Next, the user should type the function and then press CTRL+SHIFT+ENTER at the same time to apply the formula to the entire array.

As illustrated in the **Trend** worksheet, we have included time series data on a number of economic variables from January 1994 to February 1998. The TREND function is used to estimate these variables from March 1998 through December 1998. For example, we can use this data to create an “out-of-sample” forecast for the Consumer Price Index of the Russian Federation. Such forecast is referred to as “out-of-sample” because we are estimating a value for the dependent variable outside the

range for which we have data (in this case January 1994-February 1998).

The general form of the TREND function is:

=trend(**y** (in-sample), **x** (in-sample), **x** (out-of-sample), *constant*)

where **y** is the dependent variable that we are trying to estimate and **x** is the explanatory variable. Often **x** is a variable measuring time, but could conceivably also be another relevant economic or fiscal variable. In the example presented in the **Trend** worksheet, the known (in-sample) values for **x** (the time trend) and **y** (CPI) are contained in cells C4 through C53 (**y**) and B4 through B53 (**x**). The out-of-sample observations for **x** (B54:B63) correspond to those observations of the time-trend variable from March 1998 to December 1998.

The *constant* option in the TREND function is set to TRUE or omitted if the intercept coefficient is to be calculated normally or set to FALSE is the intercept is to be forced equal to zero.

It is important to note that unlike the LINEST function (as we will see later), the TREND function does not return the estimated regression coefficients nor does it provide the analyst with statistical information to test the validity of the model. The TREND function ***only*** returns values for the dependent variable based on the out-of-sample independent variable(s).

Forecasting with the Trend Function: The TREND function returns an array of values for the dependent variable equal to the number of out-of-sample observations for the explanatory variable. In our example, the number of out-of-sample observations for the time-trend variable that we seek to forecast is equal to 10, so before entering the TREND function, it is first necessary to highlight a 10 x 1 block of cells. In the Trend worksheet, we selected the cell range from cell H4 to cell H13. Next, we entered:

=trend(c4:c53,b4:b53,b54:b63,true)

and pressed CTRL+SHIFT+ENTER. As illustrated in cells H4 to H13 of the **Trend** worksheet, the forecasted values for the Consumer Price Index of the Russian Federation continue to increase through December 1998. Thus, it appears, using the time trend model and based on the sample of observations from January 1994 to February 1998, that the Consumer Price Index will continue to increase steadily throughout 1998.

Forecasts or predicted values are, of course, uncertain. In all forecasts based on regression models, there are two sources of uncertainty. First, uncertainty arises because the model cannot include all the factors that may influence the series of interest. This uncertainty appears in the error terms that are included in each regression equation. The second source of uncertainty is the estimators used for the

model's parameters. In the time-trend model, the slope and the intercept which underlie the time-trend estimates are computed using the least squares method. As illustrated through the use of the LINEST function in subsection 6.3, each estimate is accompanied by an estimated standard error that measures the precision of the estimate. The analyst should be aware of these sources of uncertainty and that the forecasts should be updated when new data becomes available.

6.2 Exponential Smoothing: Forecasting Using the GROWTH Function

The GROWTH function uses the exponential smoothing method to develop estimates of the dependent variable of interest. As with the TREND function and other regression techniques, the GROWTH function fits a line through an array containing the dependent and independent variables using the method of least squares. For the remained of this subsection, please refer to the **Growth** spreadsheet in the **Examples** data file.

The general form of the GROWTH function is similar to that of the **TREND** function:

=growth(**y** (in-sample), **x** (in-sample), **x** (out-of-sample), *constant*)

where **y** is the dependent variable, and **x** is the explanatory variable. The previous discussion on in-sample and out-of-sample observations with respect to the TREND function applies equally to the operation of the GROWTH function. The *constant* option is set to TRUE or omitted if the intercept coefficient is to be calculated normally or set to FALSE is the intercept is to be forced equal to zero.

The reader should note that when the GROWTH function is used Microsoft Excel estimates the value of the weighting parameter ? to minimize the sum of the squared prediction errors. If the reader wishes to chose a value for ?, it is recommended that the reader use the Data Analysis Tool that is covered later on in this section.

<u>Forecasting with the GROWTH function</u>: The GROWTH function returns an array of values for the dependent variable equal to the number of out-of sample observations for the explanatory variable. In the first example in the **Growth** worksheet, our objective is to develop a forecast of the Consumer Price Index for the Russian Federation using the GROWTH function. The in-sample observations for **y** (Consumer Price Index) and **x** (time-trend) are in cells C4 through C53 and B4 through B53, respectively. The out-of-sample observations for the time-trend variable are in cells B54 through B63.

Given that we have ten out-of-sample observations for the explanatory variable, we must first highlight a 10 x 1 block of cells (in this case, from cell H4 to H13) and type the following formula:

=growth(c4:c53,b4:b53,b54:b63,true)

and pressing CTRL+SHIFT+ENTER to apply the formula to the entire array. The results are contained in cells H4 through H13 of the **Growth** worksheet.

6.3 Estimating a Linear Regression Using the LINEST Function

The basic idea behind a multi-variate model is that the variation in a variable may be caused by the variation in several other observable series. Suppose, for example, that we would like to attempt to explain the level of Federal Enterprise Profits Tax collections. A multi-variate regression model would attempt to explain movements in the Federal Enterprise Profits Tax collections data series with a set of explanatory variables such as production, the rate of inflation, the unemployment rate, among others. For this subsection, please refer to the **Linest** worksheet in the **Examples** data file.

Multi-variate models are interpreted as **causal** models. This means that the independent variables (i.e., the explanatory variables, such as production, inflation, etc) should cause changes in the dependent variable (in this case, Federal Enterprise Profits Tax collections). The direction of causality is assumed to be from the explanatory variables to the dependent variable, that is, a uni-directional relationship.

Multi-variate or structural models are quite useful in estimation for several reasons. First, models based on economic variables, as opposed to statistical processes, may provide a good fit for series that are quite erratic. Second, by explaining the variation in an economic variable with a set of other economic variables, the analyst is more able to discern how policy may impact variables of interest. Finally, multi-variate models are the most efficient, reliable tool available for the estimation of economic relationships available in Excel.

<u>Linear Regression Using the LINEST function</u>: The LINEST function in Microsoft Excel is used to obtain estimated regression coefficients for a linear regression model. The LINEST function also reports a variety of statistical tests that are useful in determining the fit of the model.

The general form of the LINEST function is somewhat different from the TREND and GROWTH functions:

=LINEST(**y**,**X**, *constant*, *statistics*)

where **y** is the range of cells containing dependent variable and **X** is the range of cells (like a matrix) of independent variables. The *constant* option is set to TRUE or omitted if the intercept coefficient is to be calculated normally or set to FALSE is the intercept is to be forced equal to zero. The *statistics* option is set to TRUE if a series of additional regression statistics are desired or set to FALSE or omitted if the regression statistics are not needed.

Given that the LINEST function returns its results in array format, we must create a block of cells for the array output, similar to the TREND and GROWTH function. Before entering in the LINEST function, highlight a group of cells by clicking on the first cell of the output range and dragging the mouse until a sufficient block of cells is highlighted.

If, for example, we choose to have the LINEST function produce all the available statistics for a regression equation with *k* independent variables (excluding the intercept), we must highlight an array of *k+1* columns wide and 5 rows long. The general output format of the LINEST function is somewhat counter-intuitive; it is illustrated in Table 6.1.

As can be seen in Table 6.1, the top row of the output array, contains the parameter estimates produced by the regression, from right to left (*not from left to right*). The second row of the array presents the standard errors below each parameter estimate. Other regression statistics (including the R^2 are included in the output as indicated.

Table 6.1

Output Format of the LINEST Function

$\hat{\beta}_n$	$\hat{\beta}_{n-1}$	...	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\alpha}$
$se(\hat{\beta}_n)$	$se(\hat{\beta}_{n-1})$		$se(\hat{\beta}_2)$	$se(\hat{\beta}_1)$	$se(\hat{\alpha})$
R^2	$se(\hat{y})$				
F-statistic	Degrees of Freedom				
Regression Sum of Squares	Residual Sum of Squares				

In the current example, the dependent variable is Real Federal Enterprise Profits Tax Collections (EPT) and the explanatory variables are the percentage change in the inflation rate (Inflation), the index for Real GDP (GDP), and the index for retail sales of goods and services in constant prices (Sales). The model can be expressed as:

$$EPT_t = \beta_0 + \beta_1 \, Inflation_t + \beta_2 \, GDP_t + \beta_3 \, Sales_t + \epsilon_t$$

After highlighting the output array (which is highlighted in bright yellow for

clarity), the specific form for the LINEST function used in cells I5 through L9 of the **Linest** worksheet:

=LINEST(g4:g52,d4:f52, true, true)

Recall, the LINEST function creates an output array for the estimated regression coefficients, so it is necessary to press CNTL+SHIFT+ENTER to calculate the LINEST function. The results are illustrated in Table 6.2 and in the **Linest** worksheet.

Table 6.2

Estimation Results - LINEST example

	Sales	**GDP**	**Inflation**	**Intercept**
Estimated Coefficients	85.24	199.6234	4456.932	-25013.37
Estimated Standard Error	26.61	71.41242	5288.52	6084.944
R^2	0.62	1144.06		
F-Statistic/ degrees of freedom	24.27	45		
Regression and Residual Sum of Squares	95,307,449.15	58900139		

As illustrated in by the R^2, the model explains approximately 62 percent of the variation in Real Federal Enterprise Profits Tax collections. An F test is performed to see if the model's explanatory power is statistically significant. The critical value for the F-statistic, which is calculated in cell J20, equals 4.24 at the 1 percent significance level. Since the F-statistic computed by the model (F=24.27) far exceeds this critical value, thus we reject the null hypothesis that all the estimated coefficients are jointly equal to zero. Table 6.2 and cells J22 through J27 of the **Linest** worksheet contain the results of the hypothesis tests for the estimated regression coefficients.

A major advantage of the LINEST function is that it allows us to carefully analyze the influence of each independent variable on the dependent variable. We find that the estimated coefficients for real GDP and the Sales have a statistically significant influence on the amount of EPT collections, even at the 1 percent level of significance. Furthermore, each of the estimated coefficients for these variables is the "expected" sign. The results indicate a positive relationship between the levels of real

GDP and Federal Enterprise Profits Tax collections as well as for retail sales and EPT collections. On the other hand, while the estimated coefficient for the inflation rate is positive, it is not statistically significant at any meaningful significance level. Thus, this suggests that the monthly inflation rate does not appear to significantly influence Real Federal Enterprise Profits Tax collections.

6.4 Regression Analysis Using the Data Analysis Tool

In the preceding analyses, we developed and estimated linear regression models without the aid of the **Data Analysis Tool**. As mentioned earlier, this Tool is an option of Excel that often needs to be installed separately. As with the production of descriptive statistics, the Data Analysis Tool can assist in the estimation of linear regression models. In this section, we discuss the Regression Function (REG) in the Data Analysis tool (DAT).

Please refer to the **DAT-REG** worksheet in the **Examples** data file that accompanied this note for the examples contained in this section.

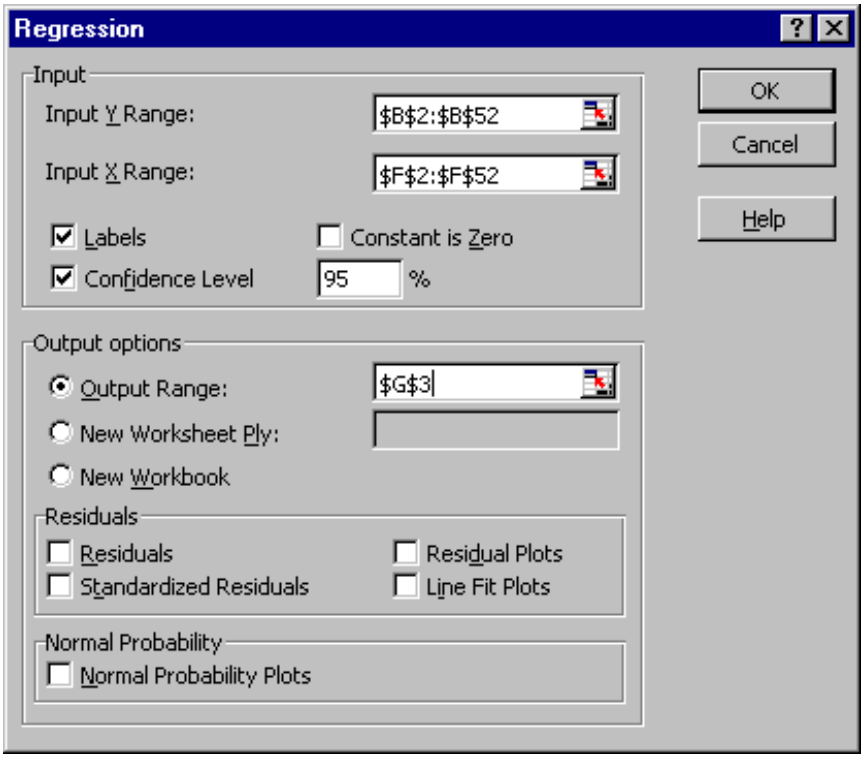
Operating the Data Analysis Tool: The DAT-REG option is quite simple to operate. Go to **Tools** and then click on **Data Analysis**. Note that there are a variety of statistical functions that we may use. Click on **Regression** so that is highlighted and click on OK. The Regression Analysis function window will open as illustrated in Figure 6.1.

In this example, we will replicate the estimation of a linear time-trend model using the DAT-REG tool. As illustrated in Figure 6.2, we have selected the input range for the dependent (**y**) variable to correspond to the range for the Consumer Price Index, including the data label (B2:B52). We have also selected the input range for the explanatory variable (**x**) to correspond to the range for the linear time-trend variable, to include the data label (F2:F52). As in earlier examples, the in-sample observations range from January 1994 to February 1998.

Since the first observation in each of the data series is the data series label, we have checked the *Labels* option. If this box remained unchecked, the DAT would return an error message. We have also chosen the *Confidence Level* option to generate confidence intervals for the regression coefficients. We have chosen to have Microsoft Excel generate a 95% confidence interval for each of the regression coefficients.

Figure 6.2

The Regression Analysis Function Window



The *constant is zero* option should only be checked if the regression is to be run without a constant term. In essence, choosing this option forces the regression line to pass through the origin.

There are three *Output Options*. If *Output Range* is selected, the analyst should enter the reference for the upper-left cell of the output. Be sure to allow for at least seven columns for the output of the regression function. If the *New Worksheet* function is chosen, the output is directed to a new spreadsheet within the existing file and if the *New Workbook* is chosen, the output is directed to an entirely new file. We have chosen to have the results sent to cell G3 of the **DAT-REG** worksheet.

The analyst can also choose to include the residuals and/or the standardized residuals in the output of the regression function by choosing the *Residuals* and/or *Standardized Residuals* options. The *Residual Plots* option should be selected if the analyst wishes to have Excel generate a chart for each independent variable versus the residuals.

Finally, if the *Line Fit Plots* option is selected, Microsoft Excel will generate a chart for the predicted values versus the observed values and if *Normal Probability*

Plots is selected, a chart that plots normal probability will generated.

Interpreting the Results of the Regression Analysis Tool: The output from the regression analysis tool is presented in three main categories: regression statistics, ANOVA statistics, and estimated regression coefficients.

Regression Statistics: Table 6.3 illustrates the regression statistics report by the DAT procedure.

Table 6.3

Output from the Data Analysis Tool -Regression Analysis

Regression Statistics	
Multiple R	0.965603848
R Square	0.932390791
Adjusted R Square	0.930982266
Standard Error	11.07274238
Observations	50

Analysis of Variance Table: The regression analysis tool also reports the degrees of freedom (df), the Sum of Squares (SS), and the Mean Square (MS) for the regression and residuals. The Analysis of Variance (ANOVA) table also contains the Total Sum of Squares and the F-test that the estimated coefficients are jointly equal to zero. The *Significance F* statistic is the probability that we accept the null hypothesis that the estimated coefficients are jointly equal to zero.

Estimated Coefficients: As illustrated in Table 6.4, the Regression Analysis Function reports the estimated coefficients, the estimated standard error, the T-statistic, the p-value, and the upper and lower bounds of the 95% confidence interval for each of the model's coefficients. The reader should note that, as with the *Significance F* statistic reported in the ANOVA table, the *P-value* statistic is the probability that we accept the null hypothesis that the estimated coefficient is equal to zero. Given the relatively high T-statistics, one can see why the *P-values* are close to zero.

Table 6.4

Estimated Coefficients - Data Analysis Tool -Regression Analysis

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	18.0600	3.1794	5.6802	7.68 E-07	11.6673	24.4526
Time	2.7918	0.1085	25.7286	9.85 E-30	2.5736	3.0100

6.5 Concluding Remarks

In this note, we have examined how to estimate linear regression models in Microsoft Excel and how to develop out-of-sample forecasts based upon the linear regression analysis. We discussed the theoretical underpinnings of the linear regression model and what were the strengths and weaknesses of the Microsoft Excel software package. This note provides the reader with the necessary materials to conduct regression analysis, interpret the results, and to construct statistically reliable out-of-sample forecasts.

In conclusion, we note that the process of quantitative analysis is a give and take process, where the theoretical knowledge and intuition of the analyst combines to form a statistical model that is estimated using the available data. As we have illustrated, the qualitative knowledge of the analysts is not surplanted by these methodologies, merely enhanced.