NOTES FOR DATA ANALYSIS

[Eighth Edition]

As stated in previous editions, the topics presented in this publication, which we have produced to assist our students, have been heavily influenced by the *Making Statistics More Effective in Schools and Business* Conferences held throughout the United States. The first conference was held at the University of Chicago in 1986. The College of Business Administration at California State University, Sacramento, hosted the tenth annual conference June 15-17, 1995. Most recent conferences were held at Babson College, (June 1999) and Syracuse University (June 2000).

As with any publication in its developmental stages, there will be errors. If you find any errors, we ask for your feedback since this is a dynamic publication we continually revise. Throughout the semester you will be provided additional handouts to supplement the material in this book.

StatGraphics Plus for Windows (ver 4.0), the statistical software used in MIS 101 and MIS 206, will work only on a Pentium chip computer. For the chapter discussions, the term StatGraphics is generic for StatGraphics® Plus for Windows (ver 4.0)

Manfred W. Hopfe, Ph.D.   Stanley A. Taylor, Ph.D.

Carmichael, California
June 2000

CALIFORNIA STATE UNIVERSITY, SACRAMENTO
College of Business Administration

NOTES FOR DATA ANALYSIS

[Eighth Edition]

TABLE OF CONTENTS

Part I

Part II

Part III

INTRODUCTION

The objective of this section is to ensure that you have the necessary foundation in statistics so that you can maximize your learning in data analysis. Hopefully, much of this material will be review. Instead of repeating Statistics 1, *the pre-requisite for this course,* we discuss some major topics with the intention that you will focus on concepts and not be overly concerned with details. In other words, as we "review" try to think of the overall picture!

Statistic vs. Parameter

In order for managers to make good decisions, they frequently need a fair amount of data that they obtain via a sample(s). Since the data is hard to interpret, in its original form, it is necessary to summarize the data. This is where statistics come into play -- a statistic is nothing more than a quantitative value calculated from a sample.

Read the last sentence in the preceding paragraph again. **A statistic is *nothing more than* a quantitative value *calculated from a sample***. Hence, for a given sample there are many different statistics that can be calculated from a sample. Since we are interested in using statistics to make decisions there usually are only a few statistics we are interested in using. These useful statistics estimate characteristics of the population, which when quantified are called *parameters*.[1]

The key point here is that managers must make decisions based upon their perceived values of parameters. Usually the values of the parameters are unknown. Thus, managers must rely on data from the population (sample), which is summarized (statistics), in order to estimate the parameters.

Mean and Variance

Two very important parameters which managers focus on frequently are the ***mean*** and ***variance***[2]. The mean, which is frequently referred to as "the average," provides a measure of the central

[1] Greek letters usually denotes parameters.

[2] The square root of the variance is called a standard deviation.

tendency while the variance describes the amount of dispersion within the population. For example, consider a portfolio of stocks. When discussing the rate of return from such a portfolio, and knowing that the rate of return will vary from time period to time period[3] one may wish to know the average rate of return (mean) and how much the variation there is in the returns [explain why they might be interested in the mean and variance].

Sampling Distribution

In order to understand statistics and not just “plug” numbers into formulas, one needs to understand the concept of a sampling distribution. In particular, one needs to know that ***every statistic has a sampling distribution, which shows every possible value the statistic can take on and the corresponding probability of occurrence***.

What does this mean in simple terms? Consider a situation where you wish to calculate the mean age of all students at CSUS. If you take a random sample of size 25, you will get one value for the sample mean (average)[4] which may or may not be the same as the sample mean from the first sample mean. Suppose you get another random sample of size 25, will you get the same sample mean? What if you take many samples, each of size 25, and you graph the distribution of sample means. What would such a graph show? The answer is that it will show the distribution of sample means, from which probabilistic statements about the **population** mean can be made.

Normal Distribution

For the situation described above, the distribution of the sample mean will follow a normal distribution. What is a **normal distribution**? The normal distribution has the following attributes:

- It depends on two parameters - the **mean** and **variance**
- It is bell-shaped
- It is symmetrical about the mean

[3] What is the random variable?

[4] The sum of all 25 values divided by 25.

[You are encouraged to use StatGraphics Plus and plot different combinations of means and variances for normal distributions.]

From a manager's perspective it is very important to know that with normal distributions approximately:

- 95% of all observations fall within 2 standard deviations of the mean
- 99% of all observations fall within 3 standard deviations of the mean.

Confidence Intervals

Suppose you wish to make an inference about the average income for a group of people. From a sample, one can come up with a **point estimate**, such as $24,000. But what does this mean? In order to provide additional information, one needs to provide a confidence interval. What is the difference between the following 95% confidence intervals for the population mean?

[23000 , 24500] and [12000 , 36000]

Hypothesis Testing

When thinking about hypothesis testing, you are probably used to going through the formal steps in a very mechanical process without thinking very much about what you are doing. Yet you go through the same steps every day.

Consider the following scenario:

> I invite you to play a game where I pull a coin out and will toss it. If it comes up heads you pay me $1. Would you be willing to play? To decide whether to play or not, many people would like to know if the coin is fair. To determine if you think the coin is fair (a hypothesis) or not (alternative hypothesis) you might take the coin and toss it a number of times, recording the outcomes (data collection). Suppose you observe the following sequence of outcomes, here H represents a head and T represents a tail -
>
> **H H H H H H H T H H H H H H T H H H H H H**
>
> What would be your conclusion? Why?

Most people look at the observations and notice the large number of heads (statistic) and conclude that they think the coin is not fair because the probability of getting 20 heads out of 22 tosses is very small, if the coin is fair (sampling distribution). It did happen; hence one rejects the idea of a fair coin and consequently does not wish to participate in the game.

Notice the steps in the above scenario

1. State hypothesis
2. Collect data
3. Calculate statistic
4. Determine likelihood of outcome, if null hypothesis is true
5. If the likelihood is small, then reject the null hypothesis
   If the likelihood is not small, then do not reject the null hypothesis

The one question that needs to be answered is “what is small?” To quantify what *small* is one needs to understand the concept of a Type I error. (We will discuss this more in class.)

P-Values

In order to simplify the decision-making process for hypothesis testing, ***p-values*** are frequently reported when the analysis is performed on the computer. In particular a p-value[5] refers to where in the sampling distribution the test statistic resides. Hence the decision rules managers can use are:

- If the p-value is ≤ alpha, then reject Ho
- If the p-value is > alpha, then do not reject Ho.

The p-value may be defined as *the probability of obtaining a test statistic equal to or more extreme than the result obtained from the sample data, given the null hypothesis H_0 is really* true.

QUALITY -- COMMON VS SPECIFIC VARIATION

During the past decade, the business community of the United States has been placing a great deal of emphasis on quality improvement. One of the key players in this quality movement was the late W. Edwards Deming, a statistician, whose philosophy has been credited with helping the Japanese turn their economy around.

One of Deming's major contributions was to direct attention away from inspection of the final product or service towards monitoring the process that produces the final product or service with emphasis of statistical quality control techniques. In particular, Deming stressed that in order to improve a process one needs to reduce the variation in the process.

Common Causes and Specific Causes

In order to reduce the variation of a process, one needs to recognize that the total variation is comprised of **common causes** and **specific causes**. At any time there are numerous factors which individually and in interaction with each other cause detectable variability in a process and its output. Those factors that are not readily identifiable and occur randomly are refereed to as the **common causes**, while those that have large impact and can be associated with special circumstances or factors are referred to as **specific causes**.

To illustrate *common causes versus specific causes*, consider a manufacturing situation where a hole needs to be drilled into a piece of steel. We are concerned with the size of the hole, in particular the diameter, since the performance of the final product is a function of the precision of the hole. As we measure consecutively drilled holes, with very fine instruments, we will notice that there is variation from one hole to the next. Some of the possible common sources can be associated with the density of the steel, air temperature, and machine operator. As long as these sources do not

[5] Referred to frequently in statistical software as a Prob. Level or Sig. Value.

produce significant swings in the variation they can be considered common sources. On the other hand, the changing of a drill bit could be a specific source provided it produces a significant change in the variation, especially if a wrong sized bit is use!

In the above example what the authors choose to list as examples of common and specific causes is not critical, since what is a common source in one situation may be a specific source in another and vice versa. What is important is that one gets a feeling of a specific source, something that can produce a significant change and that there can be numerous common sources that individually have insignificant impact on the process variation.

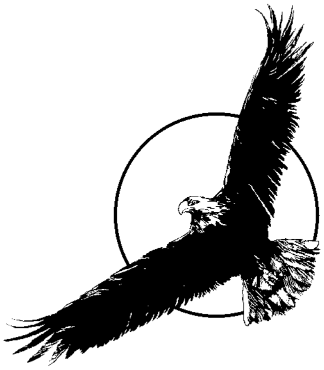
Stable and Unstable Processes

When a process has variation made up of only common causes then the process is said to be a stable process, which means that the process is in statistical control and remains relatively the same over time. This implies that the process is predictably, but does not necessarily suggest that the process is producing outputs that are acceptable as the amount of common variation may exceed the amount of acceptable variation. If a process has variation that is comprised of both common causes and specific causes then it is said to be an unstable process, which means that the process is not in statistical control. An unstable process does not necessarily mean that the process is producing unacceptable products since the total variation (common variation + specific variation) may still be less than the acceptable level of variation.

In practice one wants to produce a quality product. Since quality and total variation have an inverse relation (i.e. less {more} variation means greater {less} quality), one can see that a goal towards achieving a quality product is to identify the specific causes and eliminate the specific sources.[1] What is left then is the common sources or in other words a stable process. Tampering with a stable process will usually result in an increase in the variation that will decrease the quality. Improving the quality of a stable process (i.e. decreasing common variation) is usually only accomplished by a structural change, which will identify some of the common causes, and eliminating them from the process.

For a complete discussion of identification tools, such as times series plots to determine whether a process is stable (is the mean constant?, is the variance constant?, and is the series random -- i.e. no pattern?). The runs test is an identification tool that is used to identify nonrandom data.

[intentionally left blank]



CONTROL CHARTS

In this section we first provide a general discussion of control charts, then follow up with a description of specific control charts used in practice. Although there are many different types of control charts, our objective is to provide the reader with a solid background with regards to the fundamentals of a few control charts that can be easily extended to other control charts.

Control charts are statistical tools used to distinguish common and specific sources of variation. The format of the control chart, as shown in Figure 1 below, is a group made up of three lines where the center line = process average, upper control limit = process average + 3 standard deviations and lower control limit = process average - 3 standard deviations.

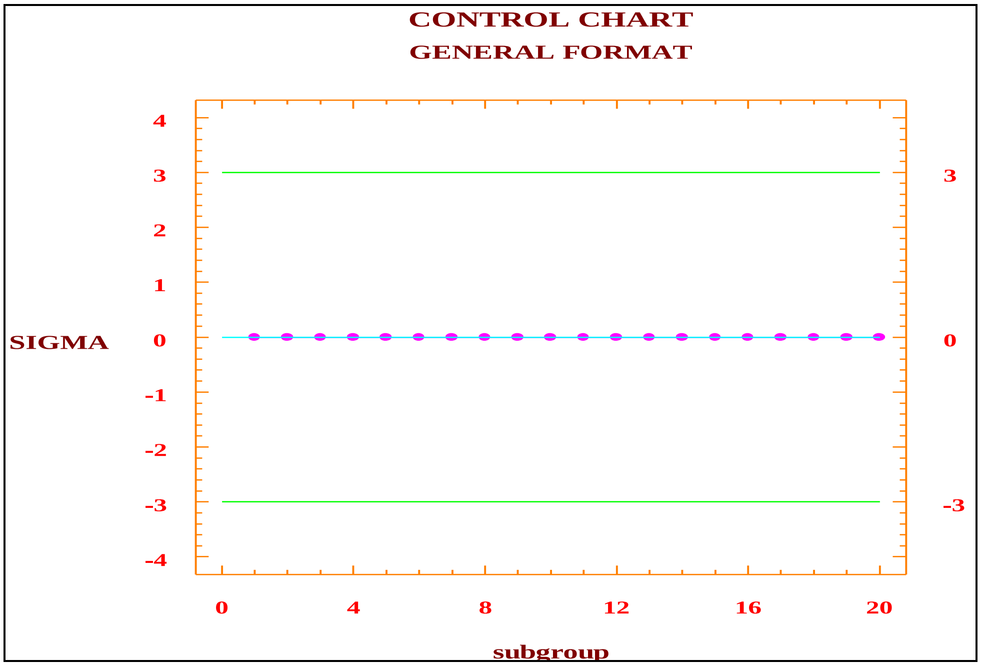


Figure 1. Control Chart (General Format)

The control charts are completed by graphing the descriptive statistic of concern, which is calculated for each subgroup. There are usually 20 to 30 subgroups used per each graph. The concept of how to form subgroups is very important and will be discussed later. For now it is important to state that *the horizontal axis is time*, so that we can view the graphed points from earliest to latest as we read the graph.

Recall that our goal in constructing control charts is to detect sources of specific variation, which, if they exist can be eliminated, thereby decreasing the variation of the process and hence increasing quality. Furthermore, recall that the existence of specific variation is the difference between an unstable process and a stable process. Therefore the detection of specific variation will be equivalent to being able to differentiate between unstable and stable processes.

Since stable processes are made up of only common causes of variation, the control charts of stable processes will exhibit no pattern in the time series plot of the observations. Departures, i.e. a pattern in the time series plot, indicate an unstable process that means that specific sources of variation exist, which need to be exposed of and eliminated in order to reduce variation and hence improve quality. As we consider each control chart, we will focus on whether there is any information in the series of observations that would be evident by the existence of a pattern in the time series plot of the observations.

Rather than showing what the control chart of a stable process looks like, it is helpful to first consider charts of unstable processes that occur frequently on practice.

We present seven graphs on the following pages for consideration. The following will summarize the seven examples displayed:

> Note that in Figure 2. Chart A the process appears to be fairly stable with the exception of an outliner (see subgroup 7). If this were the case then one would want to determine what caused that specific observation to be outside the control limits and based upon the source take appropriate action.
>
> In Figure 3, Chart B, note that there are two observations, close to each other that are outside the control limits. When this occurs there is much stronger evidence that the process is out of control than in Figure 2. Chart A. Again one would need to investigate the reason for these outliers and take appropriate action.
>
> Illustrated in Figure 3. Charts C and D are the concept of a trend. Notice in Chart C there is a subset of observations that constitute a downward trend is a subset of

observations that constitute a downward trend, while in Figure 3. Chart D there is a subset that makes up an upward trend.

In Figure 3, Chart E, a cyclical pattern is depicted. These types of patterns occur frequently when the process is subject to a seasonal influence. If this is the case, then one needs to account for the seasonally and make the necessary adjustments.

Presented in Figure 3. Chart F is a situation where there is a change in the level of a process. Notice how the level slides upward, thereby indicating a change in the level. In this situation, one would need to ascertain why the slide took place and then take appropriate action.

The final case illustrated, Figure 3. Chart G, is one where there is a change in the variance (dispersion). Notice that the first part of the sequence has a much smaller variance than the latter part. Clearly an event occurred which altered the variance and needs to be dealt with appropriately.

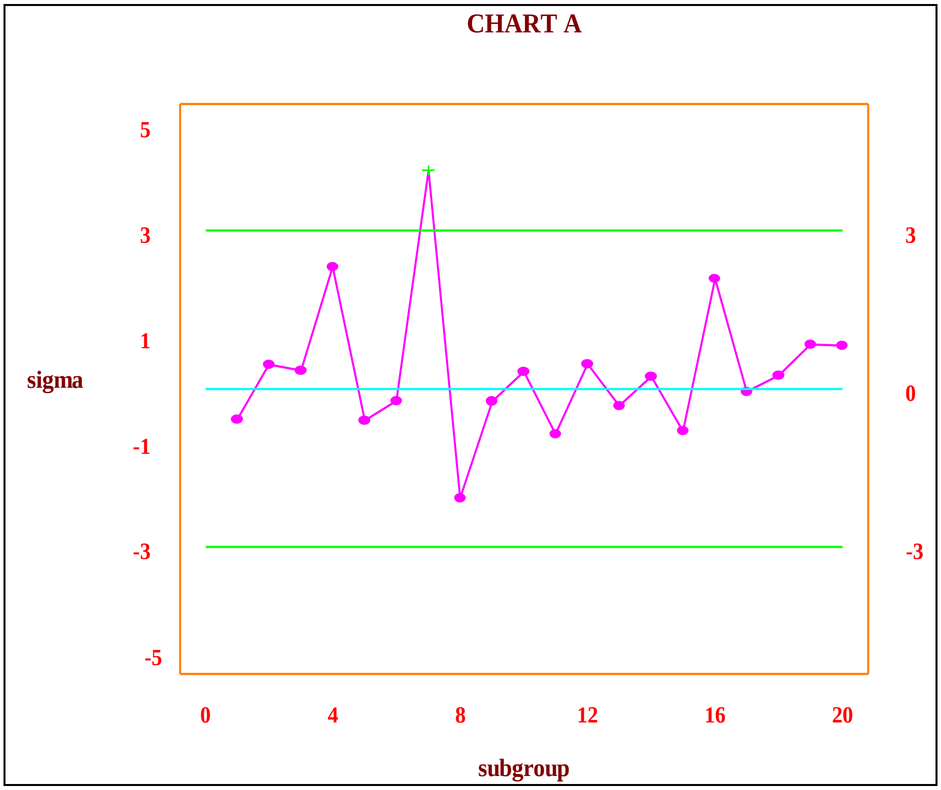


Figure 2. Chart A

Charts B through F appear in Figure 3 on the next page.

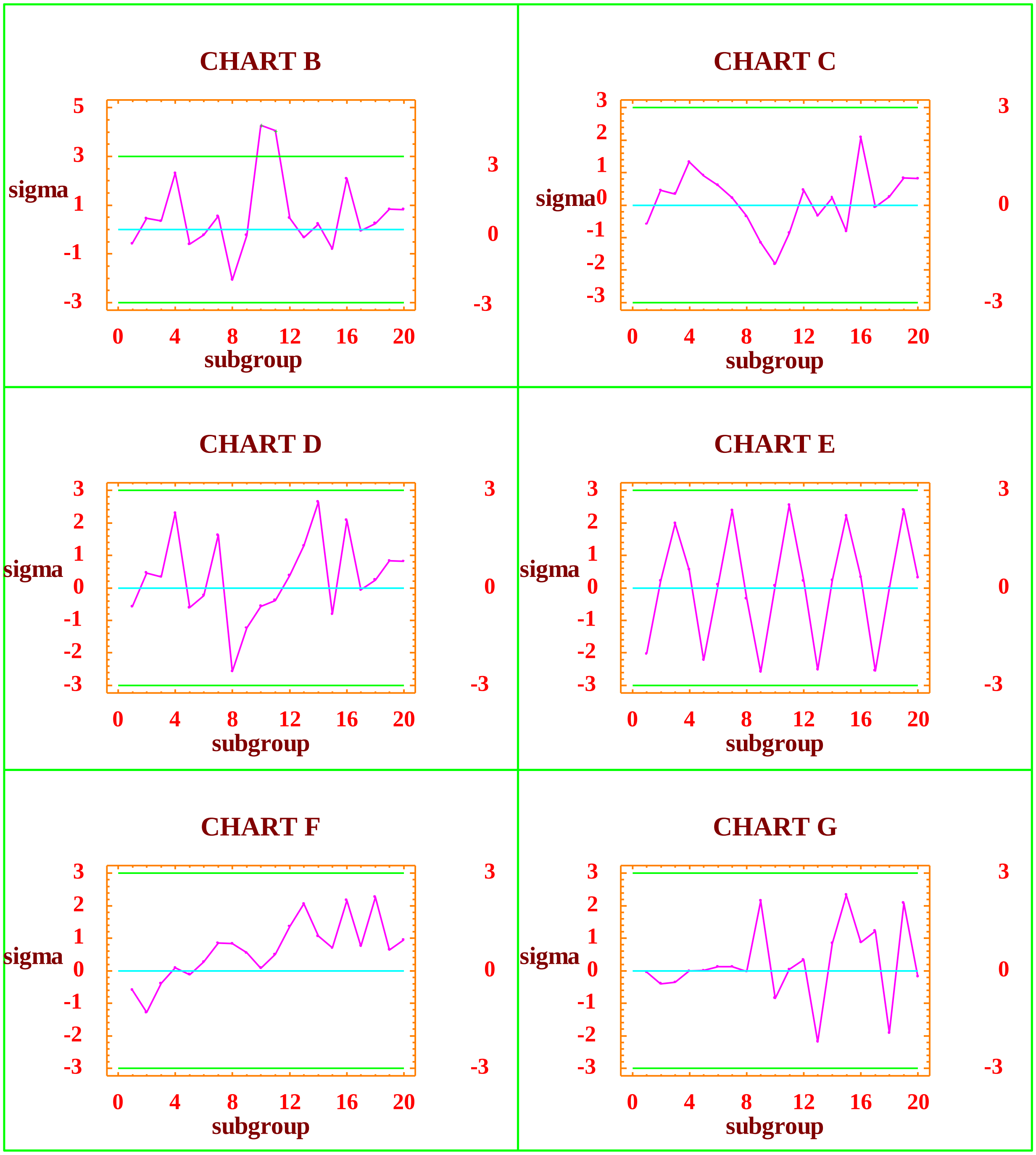


Figure 3. Charts B through G

Types Of Control Charts

As we mentioned previously, there are a large number of different control charts that are used in practice and that for our purposes we will consider just a few. For a given application the type of control chart that should be employed depends upon the type of data being collected. There are three general classes of data:

- continuous data
- classification data
- count data

Continuous data meanwhile is measurable data such as thickness, height, cost, sales units, revenues, etc. The latter two classes (classification and count) are examples of attribute data. For classification, data is bi-polar, for example, success/failure, good/bad, yes/no or conforming/non-conforming. Count data is rather straightforward -- number of customers served during the lunch hour, number of blemishes per sheet (8' by 4') of particleboard, number of failed parts per case, and so forth.

For many applications the data to be collected can be either ***continuous*** or ***attribute***. For example, when considering the size of holes discussed earlier one can record the diameter in millimeters (continuous) or as simply acceptable or unacceptable (attribute). Whenever possible, one should elect to record continuous data since fewer measurements are required per subgroup for continuous charts, 1 to 10, than for attribute charts which typically require 30 to 1000. The fewer the number of observations needed, the quicker the possible response time when problems surface.

We now consider examples for each of the control charts stated previously. First we will consider continuous data, in particular the X-bar and R charts. Then we will consider the P chart (classification data). Lastly we present the C chart (count data).

Continuous Data

X-bar and R Charts

To demonstrate the X-bar and R charts, we utilize data generated over a twenty-week period of time from the SR Mattress Co. The daily output of usable mattress frames for both shifts are shown below:

SR Mattress Company					
Week	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**
1	53	56	44	57	51
2	46	58	53	59	46
3	47	56	55	44	57
4	58	53	46	44	51
5	50	55	55	46	46
6	54	55	44	51	53
7	54	54	54	49	55
8	46	58	52	51	58
9	46	49	46	45	52
10	54	47	55	45	47
11	48	51	46	54	49
12	58	45	55	44	45
13	56	44	54	56	52
14	49	48	55	53	57
15	59	45	54	58	50
16	53	50	44	55	53
17	54	50	59	45	52
18	58	51	55	47	55
19	56	44	46	52	53
20	54	47	51	54	59

Table 1. SR Mattress Company Data

The first question one needs to answer before analyzing the data is “How will the subgroups be formed?” We will address this issue later, but to keep things simple, we will define the subgroups as being made up of the 5 daily outputs for each shift ***per week***. The respective time series plots, x-bar equals 51.42 with the lower and upper control limits of 44.931 and 59.909, respectively. [When using StatGraphics, grid lines appear in the graphs and the control limits are not initially shown. One can insert the control limits by left clicking on the graph (pane) and then right clicking in order to "pull up" the options selection. We eliminated the background grids in order to highlight the other features.

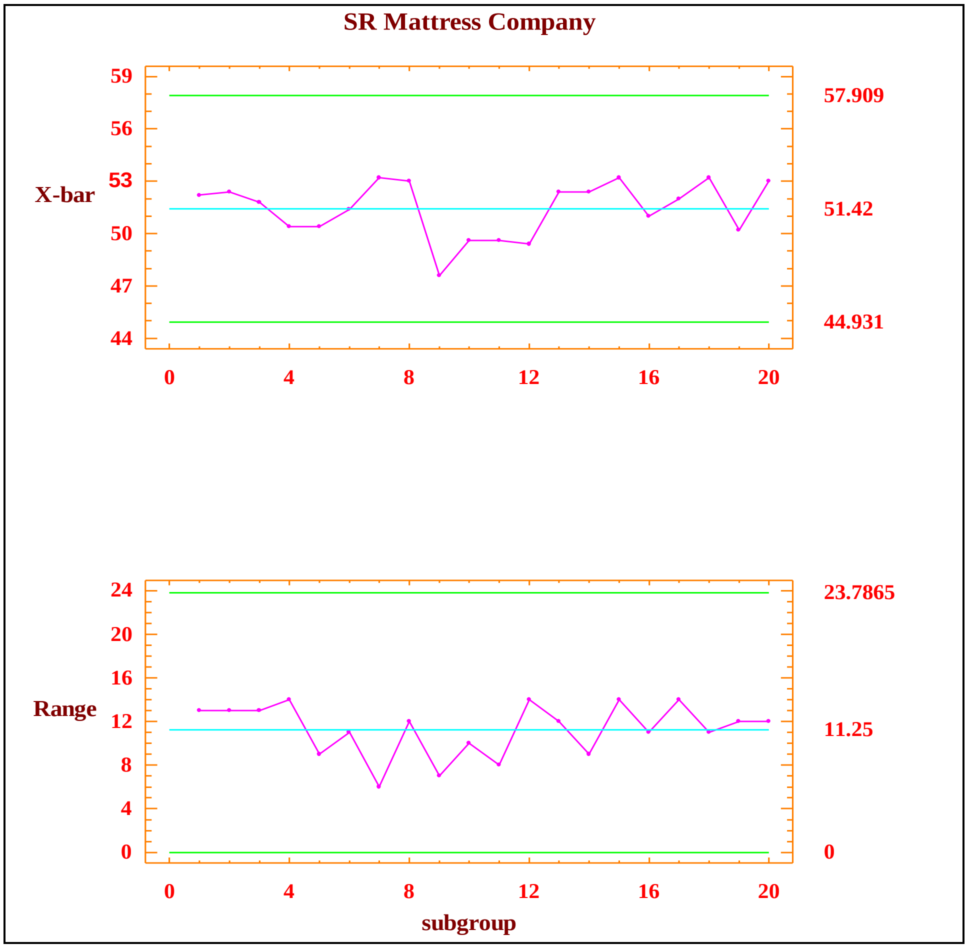


Figure 4. SR Mattress Company X-Bar and Range Chart

From these charts, the X-bar chart and range chart, we can see that none of the values are outside the control limits, thereby suggesting a possible stable process. On closer examination one may see some possible patterns that should be investigated for possible sources of specific variation. Do you see any such patterns? If so, what might be a possible scenario to describe the pattern and what type of action might management take if your scenario is true.

Given the previous example, hopefully the reader has an intuitive feel for what X-bar charts and R charts represent. We will leave it to the computer to calculate the upper and lower control limits.

Before moving on, we need to take another look at the question about how the subgroups were defined. The division described above will highlight differences between different weeks. However, what if there was a difference between the days of the workweek? For example, what if a piece of required machinery is serviced after closing every Wednesday, resulting in higher outputs every Thursday, would our sub-grouping detect such an impact? In this case one might choose to subgroup by day of the week. Hopefully, one can see how the successful implementation of control charts may depend upon the design of the control chart itself that is a function of knowing as much as possible about possible sources of specific variation.

Two final points about continuous variable control charts. The first is that when the subgroups are of size one, the X-bar chart is the same as a chart for the original series. In this case the R chart may be replaced by a moving average chart based upon past observations. The second point is that in our scenario we required each subgroup to be of the same size (equal number of observations). For example, what if there were holidays in our sample? In this case an R chart, where the statistic of concern is the range, could be replaced by an S chart, which relies on the sample standard deviation as the statistic of concern. In practice, the R charts are used more frequently with exceptions such as the holiday situation just noted.

P Charts

The P chart is very similar to the X-bar chart except that the statistic being plotted is the sample proportion rather than the sample mean. Since the proportion deals with the percentage of successes[6], clearly the appropriate data for P charts needs to attribute data where the outcomes for each trial can be classified as either a success or a failure (conform or non-conform, yes or no, etc.). The subgroup size must be equal so the proportion can be determined by dividing the outcome by the subgroup size.

To illustrate the P chart, a situation is considered where we are concerned about the accuracy of our data entry departments work. In auditing their work over the last 30 days, we randomly selected a sample of 100 entries for each day and classify each entry as correct or incorrect. The results of this audit are as follows:

Day	**# Incorrect**	**Day**	**# Incorrect**
1	2	16	1
2	7	17	5
3	6	18	9
4	2	19	6
5	4	20	4
6	3	21	3
7	2	22	3
8	6	23	5
9	6	24	3
10	2	25	6
11	4	26	6
12	3	27	5
13	6	28	2
14	2	29	3
15	4	30	4

Table 2. Number Incorrect Entries in Sample Size of 100

Given the data above, one can easily calculate the proportions of incorrect entries per day by taking the number of incorrect entries and dividing by the total number of entries for that day, which in our example were 100 each day. This may seem to be an unnecessary task at this time, since we are

[6] Recall the binomial distribution where one of the parameters is the probability of success.

essentially just scaling the data. This scaling, however, does allow us to work with the statistic, P rather than the total number of occurrences that would produce another type of chart called the NP chart. We have chosen not to discuss the NP chart since it provides the same information as the P chart for subgroups of the same size, while the P chart allows us more flexibility, so that we can consider cases when the subgroups are not all of the same sample size.[7] The P chart for the data entry example is shown below.

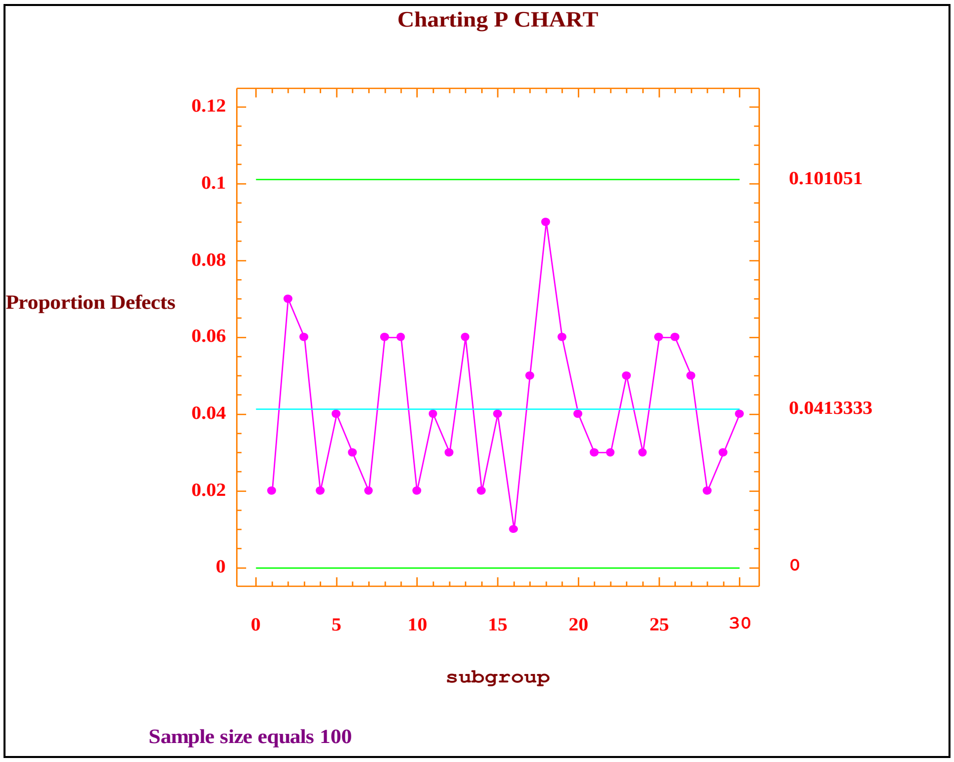


Figure 5. Proportion Control Chart

From the P chart displayed above, one can see that all of the observed values fall within the control limits and that there does not appear to be any significant pattern. One might be concerned with the

[7] When the sample sizes are different the calculations become more complicated. For our purposes we will just note this and leave the details for the software programmers.

value for the 18th observation that is .09 and look to see if a particular event triggered this *larger* value. Keep in mind, however, that common variation may very well cause this *larger variation*.

C Charts

The C chart is based upon the statistic that counts the number of occurrences in a unit, where the unit may be related to time or space. Whereas the P chart was related to the binomial distribution, the C chart is related to the Poisson distribution. To demonstrate the C chart we consider a situation where we are interested in the number of defective parts produced daily at the AKA Machine Shop. Over the past 25 days the number of defective parts per day are as shown below:

Day	**# Defective Parts**	**Day**	**# Defective Parts**
1	5	14	7
2	10	15	3
3	7	16	4
4	5	17	8
5	8	18	5
6	8	19	3
7	8	20	6
8	5	21	10
9	7	22	1
10	7	23	6
11	10	24	5
12	6	25	4
13	6		

Table 3. Number of Defective Parts per Day

The C chart[8], which appears on the next page, shows that the process appears to be stable. In particular, there are no values outside the control limits, nor does there appear to be any systematic pattern in the data. (Note: no reference made to sample size.)

[8] The notation in the StatGraphics software may confuse you some as it relates the C chart option with “count of defects” and the U chart option with “defects per unit”. We are not discussing the U chart in class or this write up. The U chart allows for the “units” to change from subgroup to subgroup.

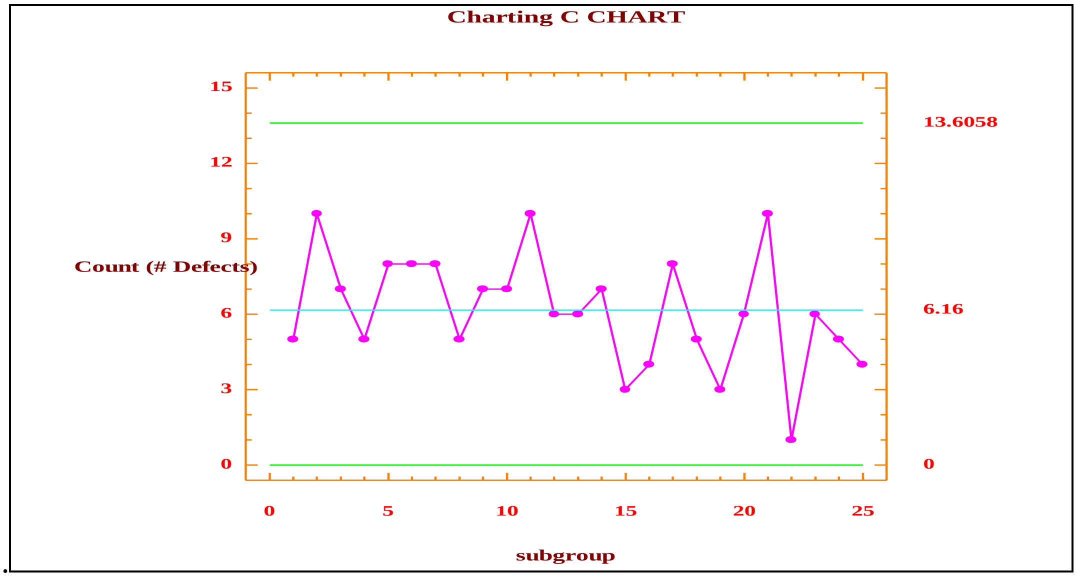


Figure 6. Count Control Chart

Conclusion

In our discussion of control charts we first discussed the common attributes of different control charts available (center line, upper control limit and lower control limit) and focused on what one looks for in trying to detect sources of specific variation (outliers, trends, oscillating, seasonality, etc.). We then looked at some of the most commonly used control charts in practice, namely the X-bar and R, P, and C control charts.

What differentiates the various control charts is the statistic that is being plotted. Since different types of data can produce different types of statistics it is clear that the type of data available will suggest the type of statistic that can be calculated and hence the appropriate control chart.

One final but important point is that the control charts generated, including those in this write up, frequently use the data set being examined to construct centerline and control limits (upper and lower). The problem this may cause is that if the process is unstable then the data it generates may alter the components of the control chart (different centerlines and different control limits) and hence be unable to detect problems that may exist. For this reason, in practice, when a process is believed to be stable the resulting statistics are frequently used to establish the control limits (center, upper, and lower) for future windows. What we mean by *window* is that if we decide to monitor say 30 subgroups at a time, as time evolves subgroups are added and consequently the same number are dropped from the other end, hence a revolving window. Useful software, such as StatGraphics, will allow one to specify the limits as an option.

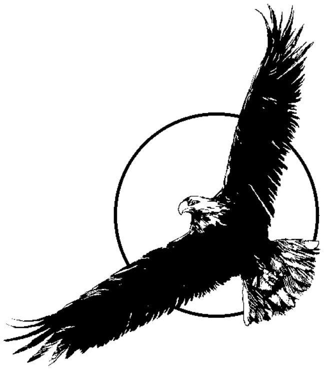
In summary:

- X-bar and Range charts are used when sample subgroups are of equal size, sample subgroups are taken at equal time intervals, and the subgroup means and range of highest and lowest values are of interest.

- Proportion charts are used when samples are of equal size and the defect proportions are of interest.

- Count charts are used when either the sample size is unknown or the sample sizes are not uniform.

[intentionally left blank]



TRANSFORMATIONS & RANDOM WALK

In the previous chapter we focused our attention on viewing variability as being comprised of two parts, common variation and specific variation. With the exception of manufacturing systems, most economic variables when viewed in their measured formats demonstrate sources of specific variation. In data analysis, whether we are trying to forecast or explain economic relationships, our goal is to model those sources of specific variation with the result being that only common variation is "left over." This can be depicted by the expression:

ACTUAL = FITTED + ERROR.

Where the FITTED values are generated from the model (specific variation), the ACTUAL values are the observed values and the ERROR values represent the differences and are a function of common sources of variation. If the common sources of variation of the model appear to be random, the model may better predict future outcomes as well as providing a more thorough understanding of how the process works.

Random Walk

One of the simplest, yet widely used models in the area of finance is the random walk model. A common and serious departure from random behavior is called a *random walk*. By definition, a series is said to follow a random walk if the first differences are random. What is meant by *first differences* is the difference from one observation to the next, which if you think about as the steps of a process and the sequence of steps as a walk, suggest the name random walk. (Do not be mislead by the term "random" in "random walk." A random walk is not random.) Relating this back to the equation we see that the ACTUAL values are the observed values for the current time period, while the FITTED values are the last periods observed values.

Hence we can write the equation as:

$$X_t = X_{t-1} + e_t$$

where: X_t is the value in time period t,
X_{t-1} is the value in time period t-1 (1 time period before)
e_t is the value of the error term in time period t.

Since the random walk was defined in terms of first differences, it may be easier to see the model expressed as:

$$X_t - X_{t-1} = e_t$$

Therefore, as one can see from the resulting equation, the series itself is not random. However, when we take the first differences the result is a transformed series $X_t - X_{t-1}$, which is random.

To illustrate the random walk model, we consider the series of stock prices for Nike as it was posted on the New York Stock Exchange at the end of each month, from May 1995 to May 2000. The time sequence plot of the series Nike (see data file) is shown in the figure below.

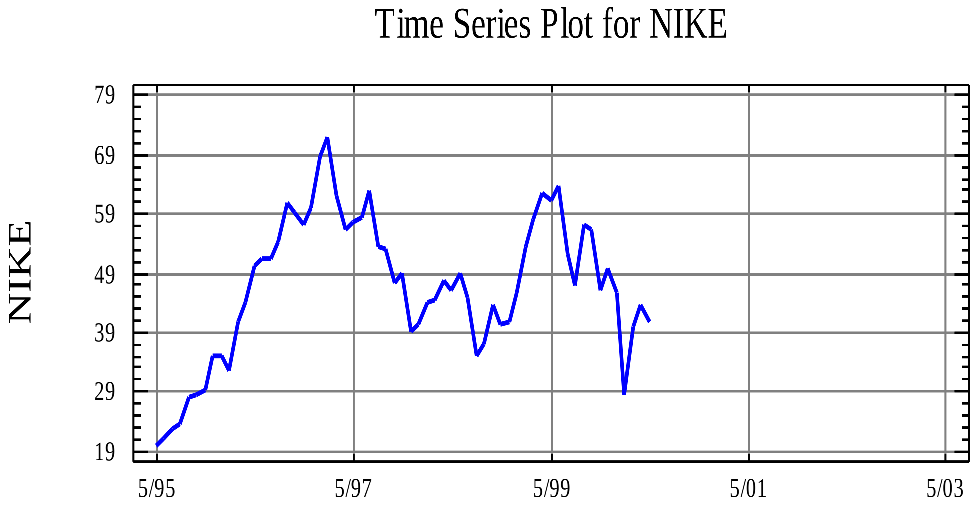


Figure 1. Times Sequence Plot of Nike

As one can see the original series for Citicorp does not appear to be random. In fact, when the nonparametric runs test is performed on the original series, the p-value is 0.000020, which indicates compelling evidence to reject the null hypothesis. Hence, the *original* series of Nike is not random.

H_0: The [original] series is random[9]
H_1: The [original] series is **NOT** random

Now consider the first differences of Nike with the time series plot shown below:

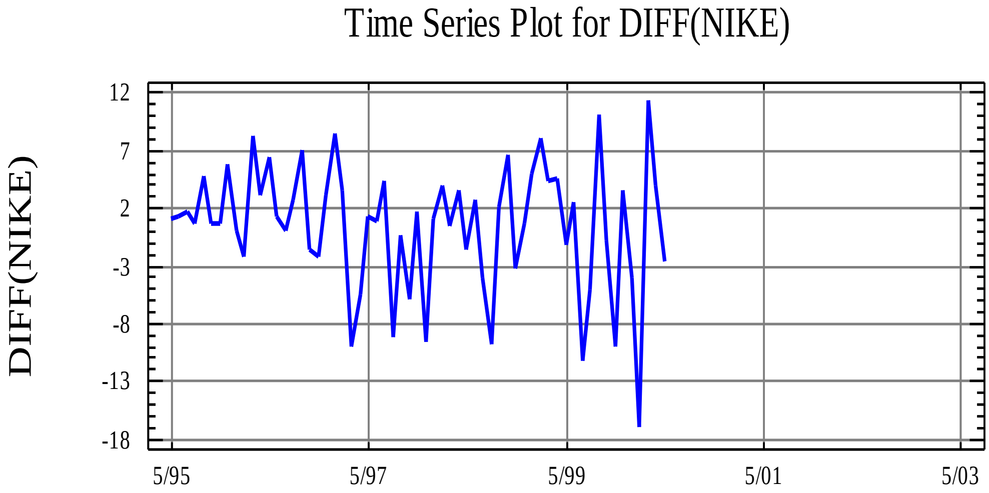


Figure 2. First Differences of Nike

As we can see from the time series plot, by taking first differences the transformed series appears to be random. (Note that we are only discussing whether the series is random, nothing is being said about it being stable since the variance increases with time.) To confirm our visual conclusion that the differenced series is random, we perform the runs test and find out that the p-value is 0.7191.

[9] The use of [original] is for emphasis only ... it is not normally used when stating the null hypothesis.

The p-value exceeds $\alpha = 0.05$ and thus provides supporting evidence to retain the null hypothesis, the differenced series is random, and thus the stock price of Nike tends to follow a random walk model.

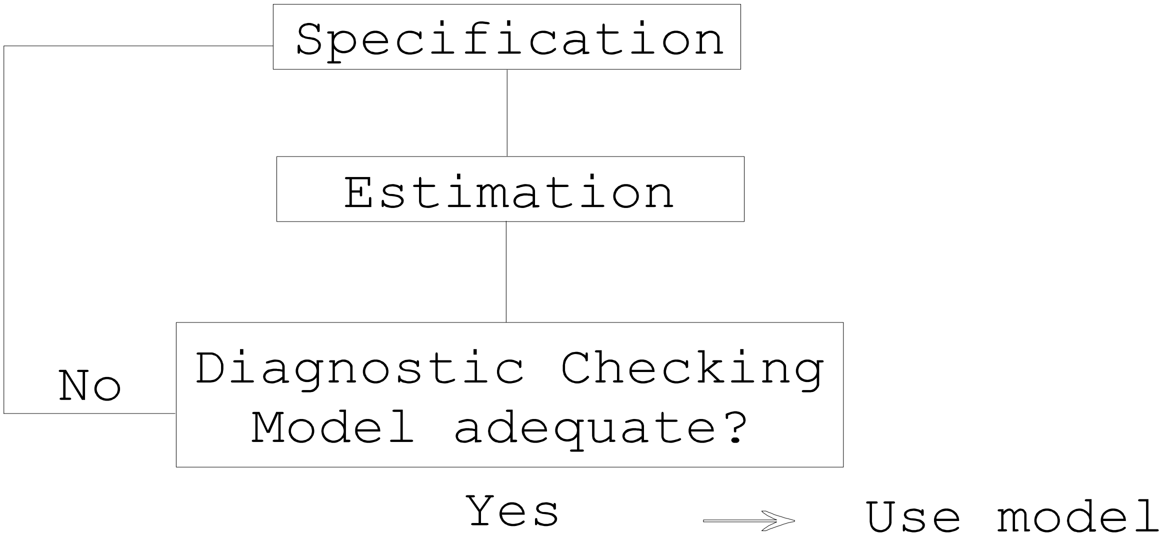
H_0: The (first differenced) series is random[10]
H_1: The (first differenced) series is **NOT** random

Information is not lost by differencing. In fact, use of differencing, or inspecting changes, is a very useful technique for examining the behavior of meandering time series. Stock market data generally follows a random walk and by differencing, we are able to get a simpler view of the process.

[10] Use of [first differenced] for emphasis only. (See footnote 10.)

MODEL BUILDING

Building a statistical model is an iterative process as depicted in the following flowchart:



As one can see, when constructing a statistical model for use there are three phases that must be followed. In fact most models used in practice require going through the three phases multiple times, as seldom is the model builder satisfied without refining the initial model at least once.

Each of these phases is discussed below in general terms, for all statistical models, and later will be described in detail for specific models (regression, time series, etc.)

Specification

The specification or identification phase involves answering two questions:

1. What variables are involved?

and

2. What is the mathematical relationship between variables?

When establishing a mathematical model there are parameters involved which are unknown to the practitioner. These parameters need to be estimated, hence, the need for the estimation phase which is discussed in the next section. When answering the questions above, it is essential that the model builder use economic theory to help establish a tentative model. A model that is based upon theory has a much better chance of being useful than one based upon guesswork.

Estimation

As mentioned previously, the models developed in the specification phase possess parameters that need to be estimated. To obtain these estimates, one gathers data and then determines the estimates that best fit the data. In order to obtain these estimates, one has to establish a criterion that can be used to ascertain whether one set of estimates is "better" than another set. The most commonly used criterion is referred to as the least squares criterion which, in simple English, means that the error terms which represent the differences between the actual and fitted values, when squared and added up will be minimized. The reason for using the squared terms is so that the positive and negative residuals do not cancel each other out. For our purposes, it will suffice to state that the computer will generate these values for us by using StatGraphics Plus.

Diagnostic Checking

The third phase is called the diagnostic checking phase and basically involves answering the question:

Is the model adequate?

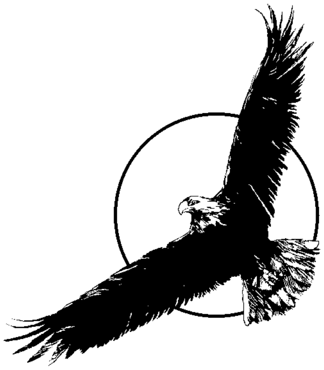
If the answer to the above question is ***no***, then something about the model needs modification and the builder returns to the specification phase and goes through the entire three phase process again. If the answer to the above question is ***yes***, then the model is ready to use.

When in the processes of discerning whether the model is adequate, a number of attributes about the model that needs to be considered:

1. How well does the model fit the data?
2. Do the residuals (actual - fitted) from the model contain any information that should be incorporated into the model? (i.e. is there information in the data that has been ignored in the creation of the model.)
3. Does the model contain variables that are useless and hence should be eliminated from the model?
4. Are the estimates derived from the estimation phase influenced disproportionately by certain observations (data)?
5. Does the model make economic sense?
6. Does the model produce valid results?

As stated previously, when the model builder is able to answer affirmatively to each of the above questions, *and only then*, they are able to use the model for their desired purpose.

[intentionally left blank]



REGRESSION ANALYSIS

In our discussion of regression analysis, we will first focus our discussion on simple linear regression and then expand to multiple linear regression. The reason for this ordering is not because simple linear regression is so ***simple***, but because we can illustrate our discussion about simple linear regression in two dimensions and once the reader has a good understanding of simple linear regression, the extension to multiple regression will be facilitated. It is important for the reader to understand that simple linear regression is a special case of multiple linear regression. Regression models are frequently used for making statistical predictions -- this will be addressed at the end of this chapter.

Simple Linear Regression

Simple linear regression analysis is used when one wants to explain and/or forecast the variation in a variable as a function of another variable. To simplify, suppose you have a variable that exhibits variable behavior, i.e. it fluctuates. If there is another variable that helps explain (or drive) the variation, then regression analysis could be utilized.

An Example

Suppose you are a manager for the Pinkham family, which distributes a product whose sales volume varies from year to year, and you wish to forecast the next years' sales volume. Using your knowledge of the company and the fact that its marketing efforts focus mainly on advertising, you theorize that sales might be a linear function of advertising and other outside factors. Hence, the model's mathematical function is:

$$\mathbf{SALES_t = B_0 + B_1\ ADVERT_t + Error}$$

Where: SALES_t represents Sales Volume in year t
ADVERT_t represents advertising expenditures in year t
B_0 and B_1 are constants (fixed numbers)
and Error_t is the difference between the actual sales volume value in year t and the fitted sales volume value in year t

Note: the Error_t term can account for influences on sales volume other than advertising.

Ignoring the error term one can clearly see that what is being proposed is a linear equation (straight line) where the SALES$_t$ value depends on the value of ADVERT$_t$. Hence, we refer to SALES$_t$ as the dependent variable and ADVERT$_t$ as the explanatory variable.

To see if the proposed linear relationship seems appropriate we gather some data and plot the data to see if a linear relationship seems appropriate. The data collected is yearly, from 1907 - 1960, hence, 54 observations. That is for each year we have a value for sales volume **<u>and</u>** a value for advertising expenditures, which means we have 54 pairs of data.

Year	**Advert**	**Sales**
1907	608	1016
1908	451	921
.	.	.
.	.	.
.	.	.
.	.	.
1959	644	1387
1960	564	1289

To get a feel for the data, we plot (called a ***scatter plot)*** the data as is shown as Figure 1. (Hereafter, the scatter plot will be called *plot*.)

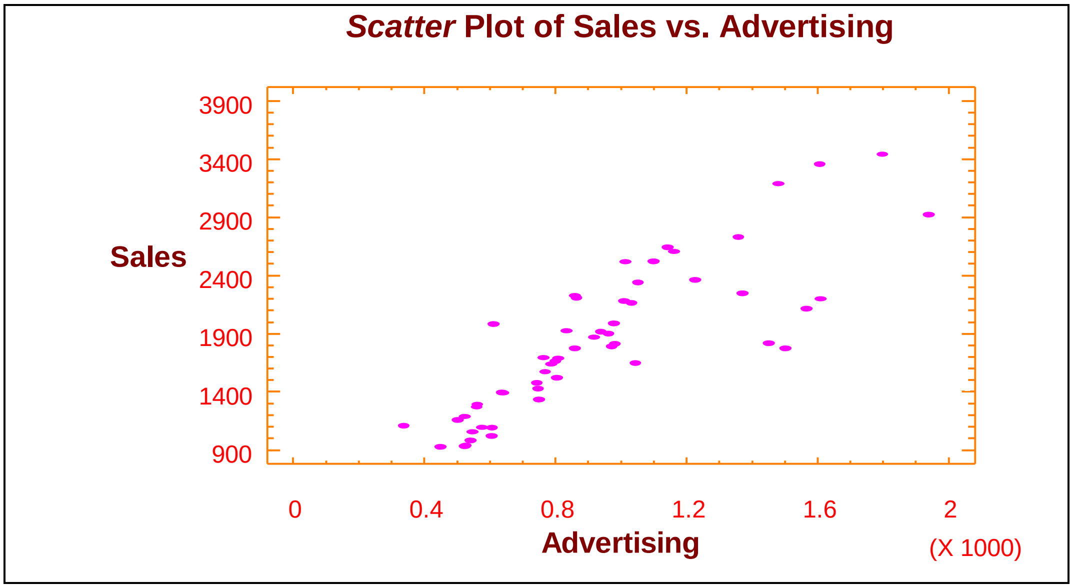


Figure 1. ***Scatter*** **Plot of Sales vs. Advertising**

As can be seen, there appears to be a fairly good linear relationship between sales (SALES) and advertising (ADVERT) (at least for advertising less than 1200 ~ note scaling factor for ADVERT x 1000). At this point, we are now ready to conclude the specification phase and move on to the estimation phase where we estimate the ***best*** fitting line.

Summary: For a simple linear regression model, the functional relationship is: $\mathbf{Y_t = B_0 + B_1 X_t + E_t}$ and for our example the dependent variable Y_t is $SALES_t$ and the explanatory variable is $ADVERT_t$. We suggested our proposed model in the example based upon theory and confirmed it via a visual inspection of the scatter plot for $SALES_t$ and $ADVERT_t$. Note: In interpreting the model we are saying that SALES depends upon ADVERT in the ***same time period*** and some other influences, which are accounted for by the ERROR term.

Estimation

We utilize the computer to perform the estimation phase. In particular, the computer will calculate the "best" fitting line, which means it will calculate the estimates for B_0 and B_1. The results are

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: sales
Independent variable: advert
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        488.833        127.439        3.83582         0.0003
Slope            1.43459       0.126866        11.3079         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                  1.50846E7       1    1.50846E7     127.87       0.0000
Residual               6.13438E6      52     117969.0
-----------------------------------------------------------------------------
Total (Corr.)           2.1219E7      53

Correlation Coefficient = 0.843149
R-squared = 71.0901 percent
Standard Error of Est. = 343.466
```

Table 1.

Since B_0 is the intercept term and B_1 represents the slope we can see that the fitted line is:

$$\mathbf{SALES_t = 488.8 + 1.4\ ADVERT_t}$$

The rest of the information presented in Table 1 can be used in the diagnostic checking phase that we discuss next.

Diagnostic Checking

Once again the purpose of the diagnostic checking phase is to evaluate the model's adequacy. To do so, at this time we will restrict our analysis to just a few pieces of information in Table 1.

First of all, to see how well the estimated model fits the observed data, we examine the R-squared (R^2) value, which is commonly referred to as the coefficient of determination. The R^2 value denotes the amount of variation in the dependent variable that is explained by the fitted model. Hence, for our example, 71.09 percent of the variation in SALES is explained by our fitted model. Another way of viewing the same thing is that the fitted model does not explain 28.91 percent of the variation in SALES.

A second question we are able to address is whether the explanatory variable, $ADVERT_t$, is a significant contributor to the model in explaining the dependent variable, $SALES_t$. Thus, for our example, we ask whether $ADVERT_t$ is a significant contributor to our model in terms of explaining $SALES_t$. The mathematical test of this question can be denoted by the hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

which makes sense, given the previous statements, when one remembers that the model we proposed is:

$$\mathbf{SALES_t = B_0 + B_1\ ADVERT_t + ERROR_t}$$

Note: If $B_1 = 0$, (i.e. the null hypothesis is true), then changes in ADVERT_t will ***not*** produce a change in SALES_t. From Table 1, we note that the p-value (probability level) for the hypothesis test, which resides on the line labeled slope, is 0.00000 (truncation). Since the p-value is less than α =. 05, we reject the null hypothesis and conclude that ADVERT_t is a significant explanatory variable for the model, where SALES_t is the dependent variable.

An Example

To further illustrate the topic of simple linear regression and the model building process, we consider another model using the same data set. However, instead of using advertising to explain the variation in sales, we hypothesize that a good explanatory variable is to use sales lagged one year. Recall that our time series data is in yearly intervals, hence, what we are proposing is a model where the value of sales is explained by its amount one time period (year) ago. This may not make as much theoretical sense [to many] as the previous model we considered, but when one considers that it is common in business for variables to run in cycles, it can be seen to be a valid possibility.

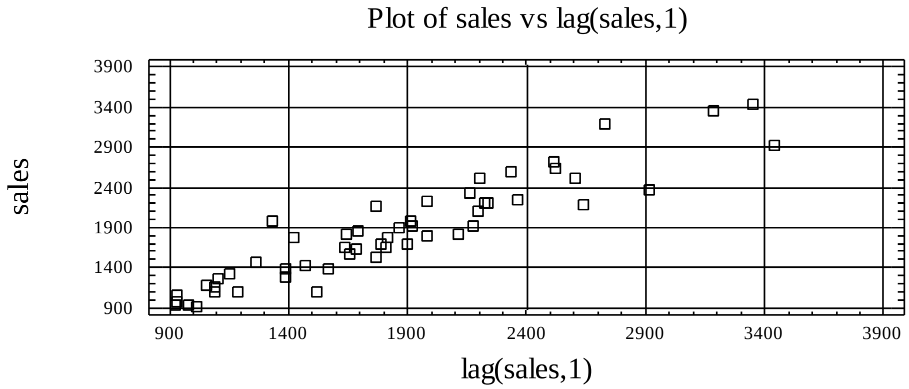


Figure 2. Plot of Sales vs. Lag(Sales,1)

Looking at Figure 2 as shown above, one can see that there appears to be a linear relationship between sales and sales one time period before. Thus the model being specified is:

$$\mathbf{SALES_t = B_0 + B_1\ SALES_{t-1} + Error_t}$$

Where: $SALES_t$ represents sales volume in year t
$SALES_{t-1}$ represents sales volume in year t-1
B_0 and B_1 are constants (fixed numbers)
and $Error_t$ is the difference between the actual sales volume value in year t and the fitted sales volume value in year t

Estimation

Using the computer, (StatGraphics software), we are able to estimate the parameters B_0 and B_1 as is shown in Table 2.

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: sales
Independent variable: lag(sales,1)
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        148.303          98.74        1.50196          0.1393
Slope           0.922186       0.050792        18.1561          0.0000
-----------------------------------------------------------------------------



                           Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    1.77921E7      1    1.77921E7     329.64       0.0000
Residual                 2.75265E6     51      53973.5
-----------------------------------------------------------------------------
Total (Corr.)            2.05447E7     52

Correlation Coefficient = 0.9306
R-squared = 86.6017 percent
Standard Error of Est. = 232.322
```

hence, the fitted model is:

$$\mathbf{SALES_t = 148.30 + 0.92\ SALES_{t-1}}$$

Diagnostic Checking

In evaluating the attributes of this estimated model, we can see where we are now able to fit the variation in sales better, as R^2, the amount of explained variation in sales, has increased from 71.09 percent to 86.60 percent. Also, as one probably expects, the test of whether $SALES_{t-1}$ does not have a significant linear relationship with $SALES_t$ is rejected. That is, the p-value for

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

is less than alpha (.00000 < .05). There are other diagnostic checks that can be performed but we will postpone those discussions until we consider multiple linear regression. Remember: *simple linear regression is a specific case of multiple linear regression.*

Update

At this point, we have specified, estimated and diagnostically checked (evaluated) two simple linear regression models. Depending upon one's objective, either model may be utilized for explanatory or forecasting purposes.

Using Model

As discussed previously, the end result of regression analysis is to be able to **explain** the variation of sales and/or to **forecast** value of $SALES_t$. We have now discussed how both of these end results can be achieved.

Explanation

As suggested by Table 1 and 2, when estimating the simple linear regression models, one is calculating estimates for the intercept and slope of the fitted line (B_0 and B_1 respectively). The interpretation associated with the slope (B_1) is that for a unit change in the explanatory variable it represents the respective change in the dependent variable along the forecasted line. Of course, this interpretation only holds in the area where the model has been fitted to the data. Thus usual interpretation for the intercept is that it represents the fitted value of the dependent variable when the independent (explanatory) variable takes on a value of zero. This is correct ***only*** when one has used data for the explanatory variable that includes zero. When one does not use values of the explanatory variable near zero, to estimate the model, then it does not make sense to even attempt to interpret the intercept of the fitted line.

Referring back to our examples, neither data set examined values for the explanatory variables ($ADVERT_t$ and $SALES_{t-1}$) near zero, hence we do not even attempt to give an economic interpretation to the intercepts. With regards to the model:

$$\mathbf{SALES_t = 488.83 + 1.43\ ADVERT_t}$$

the interpretation of the estimated slope is that a unit change in ADVERT ($1,000) will generate, ***on the average***, a change of 1.43 units in $SALES_t$ ($1,000). For instance, when $ADVERT_t$ increases (decreases) by $1,000 the average effect on SALESt will be an increase (decrease) of $1,430. One caveat, this interpretation is only valid over the range of values considered for ADVERT, which is the range from 339 to 1941 (i.e., minimum and maximum values of ADVERT).

Forecasting

Calculating the point estimate with a linear regression is a very simple process. All one needs to do is substitute the specific value of the explanatory variable, which is being forecasted, into the fitted model and the output is the point estimate.

For example, referring back to the model:

$$\mathbf{SALES_t\ =\ 488.8 + 1.4\ ADVERT_t}$$

if one wishes to forecast a point estimate for a time period when ADVERT will be 1200 then the point estimate is:

$$2168.8 = 488.8 + 1.4\,(1200)$$

Deriving a point estimate is useful, but managers usually find more information in confidence intervals. For regression models, there are two sets of confidence intervals for point forecasts that are of use as shown in Figure 3 on the next page.

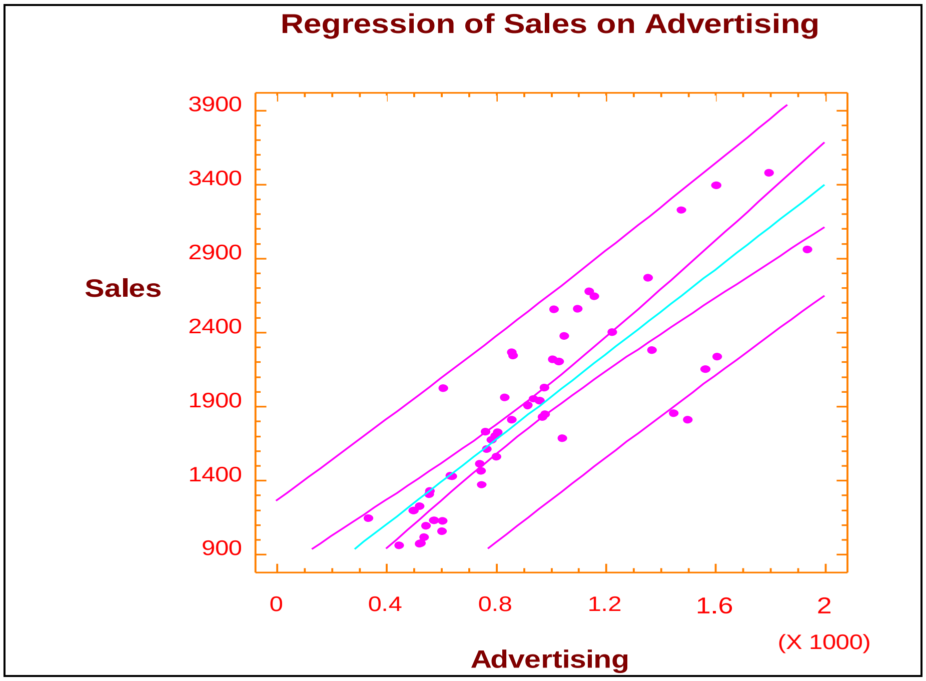


Figure 3. Regression of Sales on Advertising

Viewing Figure 3 as shown[1], one can see two sets of dotted lines, each set being symmetric about the fitted line. The inner set represents the limits (upper and lower) for the mean response for a given input, while the other set represents the limits of an individual response for a given input. It is the outer set that most managers are concerned with, since it represents the limits for an individual value. For right now, it suffices to have an intuitive idea of what the confidence limits represent and graphically what they look like. So for an ADVERT value of 1200 (input), one can visually see that the limits are approximately 1500 and 2900. (The values are actually 1511 and 2909.) Hence, when advertising is $1,200 for a time period (ADVERT_t = 1,200) then we are 95 percent confident that sales volume (SALES_t) will be between approximately 1,500 and 2,900.

[1] Figure 3 was obtained by selecting Plot of Fitted Line under the Graphical Options icon.

MARKET MODEL - Stock Beta's

An important application of simple linear regression, from business, is used to calculate the ß of a stock[2]. The ß's are measures of risk and used by portfolio managers when selecting stocks.

The model used (specified) to calculate a stock ß is:

$$\mathbf{R_{j,t} = \alpha + \beta\, R_{m,t} + \varepsilon_t}$$

Where: $R_{j,t}$ is the rate of return for the j^{th} stock in time period_t
$R_{m,t}$ is the market rate of return in time period_t
ε_t is the error term in time period_t
α and β are constants

To illustrate the above model, we will use data that resides in the data file ***SLR.SF3***. In particular, we will calculate β's for Anheuser Busch Corporation, the Boeing Corporation, and American Express using the New York Stock Exchange (NYSE - Finance) as the "market" portfolio. The data in the file ***SLR.SF3*** has already been converted from monthly values of the individual stock prices and dividends to represent the monthly rate of returns (starting with December 1986).

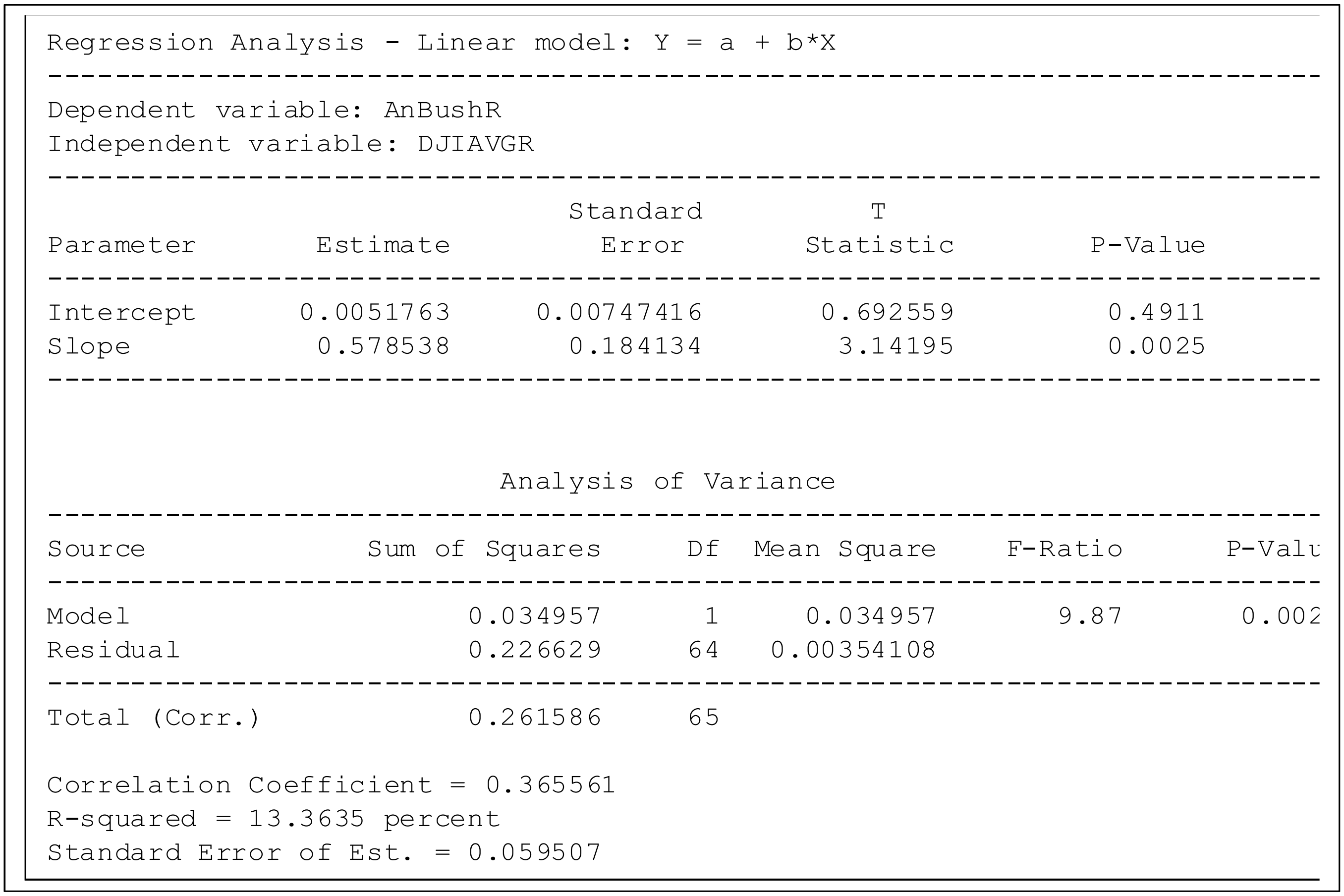
For all three stocks, the model being specified and estimated follows the form stated in the equation shown above, the individual stocks rate of returns will be used as the dependent variable and the NYSE rate of returns will be used as the independent variable.

[2] For an additional explanation on the concept of stock beta's, refer to the Appendix.

1. **Anheuser Busch Co. (AnBushr)**

Using the equation, the model we specify is **$AnBushR_t = \alpha + \beta$ DJIAVGRt $+ \varepsilon_t$.**

The estimation results are shown below in Table 3:



```
Regression Analysis - Linear model: Y = a + b*X
----------------------------------------------------------------------------
Dependent variable: AnBushR
Independent variable: DJIAVGR
----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
----------------------------------------------------------------------------
Intercept      0.0051763     0.00747416       0.692559          0.4911
Slope           0.578538       0.184134        3.14195          0.0025
----------------------------------------------------------------------------


                           Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Valu
----------------------------------------------------------------------------
Model                     0.034957      1     0.034957       9.87       0.002
Residual                  0.226629     64   0.00354108
----------------------------------------------------------------------------
Total (Corr.)             0.261586     65

Correlation Coefficient = 0.365561
R-squared = 13.3635 percent
Standard Error of Est. = 0.059507
```

Table 3

As shown in the estimation results, the estimated β for Anheuser Busch Co. is 0.578. Note that with a p-value of 0.00254, the coefficient of determination, R-squared, is only 13.36 percent, which indicates a poor fit of the data. However, at this point we only wish to focus on the estimated β.

2. The Boeing Co.

The model we specify, using equation (1) is $\mathbf{BoeingR_t = \alpha + \beta\ \ DJIAVGRt + \varepsilon_t}$

The results appear below in Table 4.

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: BoeingR
Independent variable: DJIAVGR
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept     0.00502472     0.00946959       0.530616         0.5975
Slope           0.903706       0.233293        3.87369         0.0003
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   0.0852952      1    0.0852952      15.01       0.0003
Residual                 0.363793     64   0.00568426
-----------------------------------------------------------------------------
Total (Corr.)            0.449088     65

Correlation Coefficient = 0.435809
R-squared = 18.993 percent
Standard Error of Est. = 0.0753941
```

Table 4

Note that the estimated β for The Boeing Co. is 0.904 while the R^2 value is only 18.99 percent.

3. American Express

The model we specify, using equation is as follows:

$$\mathbf{AmExpR_t = \alpha + \beta\ DJIAVGRt + \varepsilon_t}$$

which can be estimated using StatGraphics

With the results appearing as Table 5:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: AmExpR
Independent variable: DJIAVGR
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept     -0.0103298     0.00993636        -1.0396         0.3024
Slope            1.07964       0.244793        4.41043         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    0.121738      1     0.121738      19.45       0.0000
Residual                 0.400541     64   0.00625845
-----------------------------------------------------------------------------
Total (Corr.)            0.522279     65

Correlation Coefficient = 0.482795
R-squared = 23.3091 percent
Standard Error of Est. = 0.0791104
```

Table 5

The estimation results indicate that the β is 1.07964, with an R-squared value of 23.31 percent.

Summary

Using monthly value from December 1986, we utilized simple linear regression to estimate the β‘s of Anheuser Busch Co. (0.579), the Boeing Co. (0.904), and American Express (1.080). Note that the closer the β‘s are to 1.0, the closer the stocks move with the market. What does that imply about Anheuser Busch Corporation, the Boeing Corporation, and American Express?

The risk contribution to a portfolio of an individual stock is measured by the stock's beta coefficient. Analysts review the market outlooks - if the outlook suggests a market decline, stocks with large positive coefficients might be sold short. Of course, the historical measure of β must persist at approximately the same level during the forecast period. (Additional discussion about stock betas appears in the Appendix.)

Multiple Linear Regression

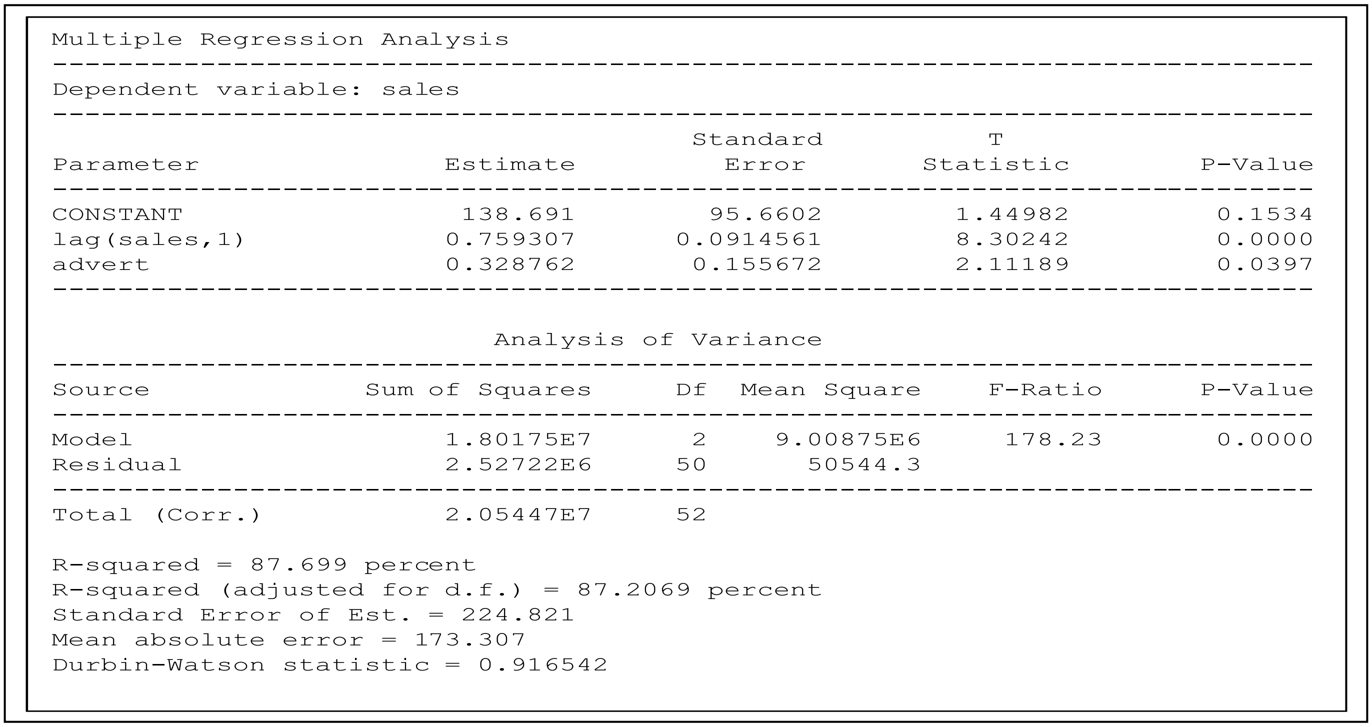
Referring back to the Pinkham data, suppose you decided that ADVERT_t contained information about SALES_t that lagged value of SALES_t (i.e. SALES_{t-1}) did not, and vice versa, and that you wished to regress SALES_t or both ADVERT_t and SALES_{t-1}; the solution would be to use a multiple regression model. Hence, we need to generalize our discussion of simple linear regression models by now allowing for more than one explanatory variable, hence the name multiple regression. [Note: *more than one* explanatory variable, hence we are not limited to just two explanatory variables.]

Specification: Going back to our example, if we specify a multiple linear regression model where SALES_t is again the dependent variable and ADVERT_t and SALES_{t-1} are the explanatory variables, then the model is:

$$\mathbf{SALES_t = B_0 + B_1\ ADVERT_t + B_2\ SALES_{t-1} + ERROR_t}$$

where: $B_{0,}$ $B_{1,}$ and B_2 are parameters (coefficients).

Estimation: To obtain estimates for B_0, $B_{1,}$ and B_2 via StatGraphics, the criterion of least squares still applies, the mathematics employed involves using matrix algebra. It suffices for the student to understand what the computer is doing on an **intuitive** level; i.e. the best fitting line is being generated. The results from the estimation phase are shown in Table 7.



```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: sales
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 138.691        95.6602        1.44982         0.1534
lag(sales,1)            0.759307      0.0914561        8.30242         0.0000
advert                  0.328762       0.155672        2.11189         0.0397
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   1.80175E7      2    9.00875E6     178.23       0.0000
Residual                2.52722E6     50      50544.3
-----------------------------------------------------------------------------
Total (Corr.)           2.05447E7     52

R-squared = 87.699 percent
R-squared (adjusted for d.f.) = 87.2069 percent
Standard Error of Est. = 224.821
Mean absolute error = 173.307
Durbin-Watson statistic = 0.916542
```

Table 7

Diagnostic Checking

We still utilize the diagnostic checks we discussed for simple linear regression. We are now going to expand that list and include additional diagnostic checks, some require more than one explanatory variable but most also pertain to simple linear regression. We waited to introduce some of the checks [that also pertain to simple linear regression] because we didn't want to introduce too much at one time and most of the corrective measures involve knowledge of multiple regression as an alternative model.

The first diagnostic we consider involves focusing on whether any of the explanatory variables should be ***removed*** from the model. To make these decision(s) we test whether the coefficient associated with each variable is significantly different from zero, i.e. for the ith explanatory variable:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

As discussed in simple linear regression this involves a t-test. Looking at Table 7, the p-value for the tests associated with determining the significance for $SALES_{t-1}$ and $ADVERT_1$ are 0.0000 and 0.0397, respectively, we can ascertain that neither explanatory variable should be eliminated from the model. If one of the explanatory variables had a p-value greater than α =. 05, then we would designate that variable as a candidate for deletion from the model and go back to the specification phase.

Another attribute of the model we are interested in is the R^2 adjusted value that in Table 7 is 0.8721, or 87.21 percent. Since we are now considering multiple linear regression models, the R^2 value that we calculate represents the amount of variation in the dependent variable ($SALES_t$) that is explained by the fitted model, which includes **all** of the explanatory variables **jointly** ($ADVERT_t$ and $SALES_{t-1}$). At this point we choose to ignore the adjusted (ADJ) factor included in the printout.

Since we have already asked the question if anything should be deleted from the model the next question that should be asked if there is anything that is missing from the model, i.e. should we add anything to the model. To answer this question we should use theory but from an empirical perspective we look at the residuals to see if they have a pattern, which as we discussed previously would imply there is information. If we find missing information for the model (i.e. a pattern in the residuals), then we go back to the specification phase, incorporate that information into the model and then cycle through the 3 phase process again, with the revised model. We will illustrate this in greater detail in our next example. However, the process involved is very similar to that which we employed earlier in the semester. We illustrate the residual analysis with a new example.

Example

The purpose behind looking at this example is to allow us to work with some cross sectional data and also to look in greater detail at analyzing the residuals. The data set contains three variables that have been recorded by a firm that presents seminars. Each record focuses on a seminar with the fields representing:

- number of people enrolled (ENROLL)
- number of mailings sent out (MAIL)
- lead time (in weeks) of 1st mailing (LEAD)

The theory being suggested is that the variation in the number of enrollments is an approximate linear function of the number of mailings and the lead-time. As recommended earlier, we look at the scatter plots of the data to see if our assumptions seem valid. Since we are working with two explanatory variables, a three dimensional plot would be required to see all three variables simultaneously, which can be done in StatGraphics with the PLOTTING FUNCTIONS, X-Y-Z LINE and SCATTER PLOT options (note the dependent variable is usually Z). See Figure 7 for this plot.

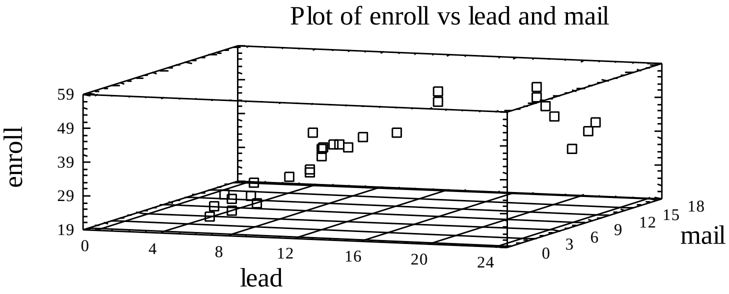


Figure 7. Plot of Enroll vs. Mail & Lead

This plot provides some insight, but for beginners, it is usually more beneficial to view multiple two-dimensional plots where the dependent variable ENROLL is plotted against the different explanatory variables, as is shown in Figures 8 and 9.

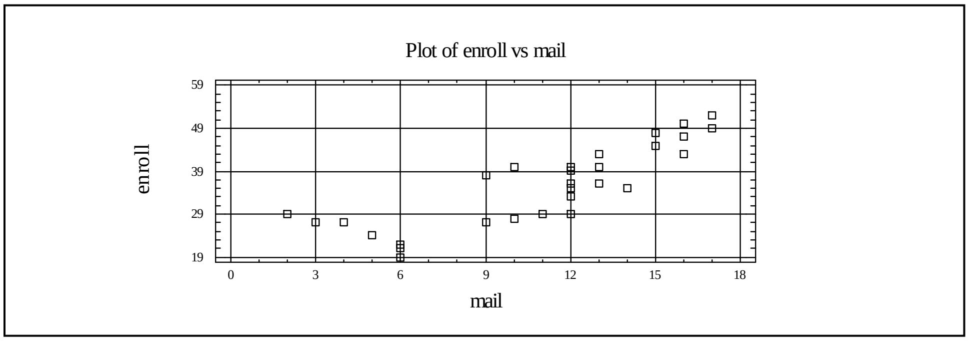


Figure 8. Plot of Enroll vs. Mail

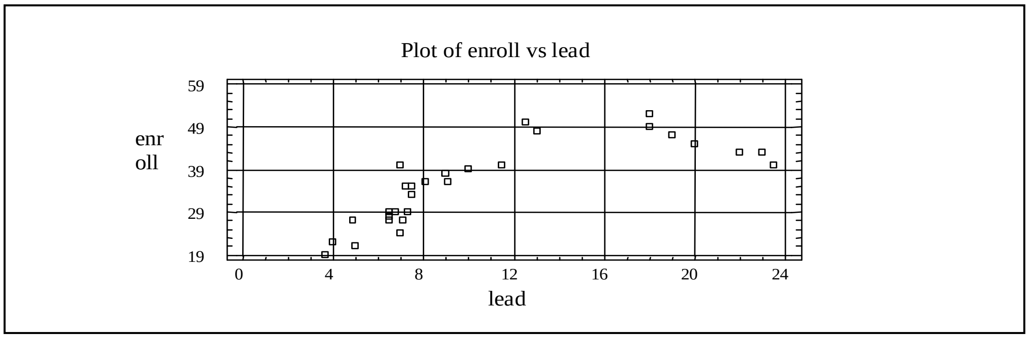


Figure 9. Plot of Enroll vs. Lead

Looking at Figure 9, which plots ENROLL against LEAD, we notice that there is a dip for the largest LEAD values which may economically suggest diminishing returns i.e. at a point the larger lead time is counterproductive. This suggest that ENROLL and LEAD may have a parabolic relationship. Since the general equation of a parabola is:

$$\mathbf{y = ax^2 + bx + c}$$

we may want to consider including a squared term of LEAD in the model. However, at this point we are not going to do so, with the strategy that if it is needed, we will see that when we examine the residuals, as we would have ignored some information in the data and it will surface when we analyze the residuals. (In other words we wish to show that if a term should be included in a model, but is not identified, one should be able to identify it as missing when examining the residuals of a model estimated without it.)

Specification

Thus the model we tentatively specify is:

$$\mathbf{ENROLL_i = B_0 + B_1\ MAIL_i + B_2\ Lead_i + ERROR_i}$$

Estimation

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: enroll
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  14.8523         2.1596       6.87733         0.0000
lead                     0.627318       0.165436       3.79191         0.0008
mail                      1.27378       0.233677       5.45103         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     1985.35      2      992.674      56.87       0.0000
Residual                  453.824     26      17.4548
-----------------------------------------------------------------------------
Total (Corr.)             2439.17     28

R-squared = 81.3943 percent
R-squared (adjusted for d.f.) = 79.9631 percent
Standard Error of Est. = 4.17789
Mean absolute error = 3.33578
Durbin-Watson statistic = 1.03162
```

Table 8

Note that MAIL and LEAD are both significant, since their p-values are 0.0000 and 0.0008, respectively. Hence, there is no need at this time to eliminate either from the model. Also, note that R^2_{adj} is 79.96 percent.

To see if there is anything that should be added to the model, we analyze the residuals to see if they contain any information. Utilizing the graphics options icon, one can obtain a plot of the To see if To see if there is anything that should be added to the model, we analyze the residuals to see if they contain any information. Utilizing the graphics options icon, one can obtain a plot of the standardized residuals versus lead (select residuals versus X). Plotting against the predicted values is similar to looking for departures from the fitted line. For our example since we entertained the idea of some curvature (parabola) when plotting ENROLL against LEAD, we now plot the residuals against LEAD. This plot is shown as Figure 10.

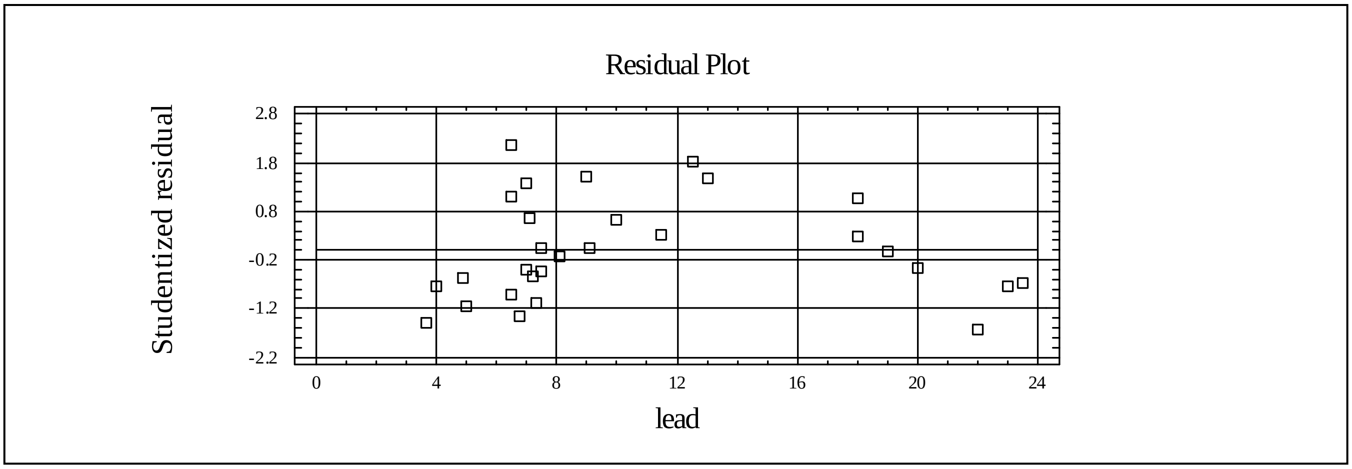


Figure 10. Residual Plot for Enroll against Lead

What we are looking for in the plot is whether there is any information in LEAD that is missing from the fitted model. If one sees the curvature that still exists, then it suggests that one needs to add another variable, actually a transformation of LEAD, to the model. Hence we go back to the specification phase, based upon the information just discovered, and specify the model as:

$$\mathbf{ENROLL_i = B_0 + B_1\ MAIL_i + B_2\ Lead + B_3\ (LEAD)^2{}_i + ERROR_i}$$

The estimation of the revised model generates the output presented in Table 9.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: enroll
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                0.226184        2.89795      0.0780495         0.9384
lead                     4.50131       0.675669        6.66201         0.0000
mail                    0.645073       0.189375        3.40633         0.0022
lead * lead            -0.132796       0.022852       -5.81115         0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     2246.12      3      748.707      96.96       0.0000
Residual                  193.053     25      7.72211
-----------------------------------------------------------------------------
Total (Corr.)             2439.17     28

R-squared = 92.0853 percent
R-squared (adjusted for d.f.) = 91.1356 percent
Standard Error of Est. = 2.77887
Mean absolute error = 2.081
Durbin-Watson statistic = 1.121
```

Table 9

Diagnostic Checking

At this point we go through the diagnostic checking phase again. Note that all three explanatory variables are significant and that the R^2_{adj} value has increased to 91.13 percent from 79.96 percent. For our purposes at this point, we are going to stop our discussion of this example, although the reader should be aware that the diagnostic checking phase has not been completed. Residual plots should be examined again, and other diagnostic checks we still need to discuss should be considered.

Before we proceed however, it should be pointed out that the last model is still a multiple ***linear*** regression model. Many students think that by including the squared term, to incorporate the curvature, that we may have violated the linearity condition. This is not the case, as when we say "linear" it is linear with regards to the coefficients. An intuitive explanation of this is to think like the computer, all LEAD^2 represents is the squared values of LEAD, therefore, the **calculations** are the same as if LEAD^2 was another explanatory variable.

The next three multiple regression topics we discuss will be illustrated with the data that was part of a survey conducted of houses in Eugene, Oregon, during the 1970's. The variables measured (recorded), for each house, are sales price (price), square feet (sqft), number of bedrooms (bed), number of bathrooms (bath), total number of rooms (total), age in years (age), the house has an attached garage (attach), and whether the house has a nice view (view).

Dummy Variables

Prior to this current example, all the regression variables we have considered have been either ratio or interval data, which means they are non-qualitative variables. However, we now want to incorporate qualitative variables into our analysis. To do this we create dummy variables, which are binary variables that take on values of either zero or one. Hence, the dummy variable (attach) is defined as:

attach =	1	if garage is attached to house
	0	otherwise (i.e. not attached)

and

view =	1	if house has a nice view
	0	otherwise

Note that each qualitative attribute (attached garage and view) cited above has two possible outcomes (yes or no) but there is only 1 dummy variable for each. That is because ***there must always be, at maximum, one less dummy variable than there are possible outcomes for particular***

qualitative attribute. We mention this because there are going to be situations, for other examples, where one wants to incorporate a qualitative attribute that has more than two possible outcomes in the analysis. For example, if one is explaining sales and has quarterly data they might want to include the season as an explanatory variable. Since there are four seasons (Fall, Winter, Spring, and Summer) there will be three (four minus one) dummy variables. To define these three dummy variables, we arbitrarily select one season to "withhold" and create dummy variables for each of the other seasons. For example, if summer was "withheld" then our three dummy variables could be

D_1 =	1	if Fall
	0	otherwise
D_2 =	1	if Winter
	0	otherwise
D_3 =	1	if Spring
	0	otherwise

Now, what happens when we withhold a season is not that we ignore the season, but the others are being compared with what is being withheld.

Outliers

When an observation has an undue amount of influence on the fitted regression model (coefficients) then it is called an outlier. Ideally, each observation has an equal amount of influence on the estimation of the fitted lines. When we have an outlier, the first question one needs to ask is "Why is that observation an outlier?" The answer to that question will frequently dictate what type of action the model builder should take.

One reason an observation may be an outlier is because of a recording (inputting) error. For instance, it is easy to mistakenly input an extra zero, transpose two digits, etc. When this is the cause, then corrective action can clearly be taken. ***Don't always assume the data is correct!*** Another source is because of some extra ordinary event that we do not expect to occur again. Or the observation is not part of the population we wish to make interpretation/forecasts about. In these cases, the observation may be "discarded."

If the data is cross-sectional, then the observation may be eliminated, thereby decreasing the number of observations by one. If the data is times series, by "discarding the impact" of the observation one does not eliminate observations since doing so may effect lagging relationships, however one can set the dummy variable equal to one (1) for that observation, zero (0) otherwise.

At other times, the outcome, which is classified as an outlier, is recorded correctly, may very well occur again, and is indeed part of the concerned population. In this case, one would probably want to leave the observation in the model construction process. In fact, if an outlier or set of outliers represents a source of specific variation then one should incorporate that specific variation into the model via an additional variable. Keep in mind, just because an observation is an outlier does not mean that it should be discarded. These observations contain information that should not be ignored just so "the model looks better."

Now that we have defined what an outlier is and what action to *take/not take* for outlier, the next step is to discuss how to determine what observations are outliers. Although a number of criteria exist for classifying outliers, we limit our discussion to two specific criteria - standardized residuals and leverage.

The theory behind using standardized residuals is that outlier are equated with observations which have large residuals. To determine what is large, we standardize the residuals and then use the rule that any standardized residual outside the bounds of -2 to 2 is considered an outlier. [Why do we use -2 and 2? Could we use -3 and 3?]

The theory behind the leverage criteria is that a large residual may not necessarily equate with an outlier. Hence, the leverage value measures the amount of influence that each observation has on the set of estimates. It’s not intuitive, but can be shown mathematically, that the sum of the leverage points is equal to the number of B coefficients in the model (P). Since there are N observations, under ideal conditions each observation should have a leverage value of P/N. Hence, using our criteria of large being outside two standard deviation, the decision rule for declaring outlier by means of leverage values is to declare an observation as a potential outlier if its leverage value exceeds 2*P/N. StatGraphics employs a cut off of 3* P/N.

To illustrate, identifying outliers, we estimate the model:

$$\textbf{Price}_i = \textbf{B}_0 + \textbf{B}_1\ \textbf{SQFT}_i + \textbf{B}_2\ \textbf{BED} + \textbf{Error}$$

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: price
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                -15.4038        7.34394       -2.09749         0.0414
sqft                     3.52674       0.269104        13.1055         0.0000
bed                      7.64828        2.78697         2.7443         0.0086
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     29438.5      2      14719.3     140.65       0.0000
Residual                  4918.52     47      104.649
-----------------------------------------------------------------------------
Total (Corr.)             34357.0     49

R-squared = 85.6841 percent
R-squared (adjusted for d.f.) = 85.0749 percent
Standard Error of Est. = 10.2298
Mean absolute error = 7.19612
Durbin-Watson statistic = 1.682
```

Table 10

With the results being shown in Table 10, in our data set of houses, clearly some houses are going to influence the estimate more than others. Those with undue influences will be classified as potential outliers. Again, the standardized residuals outside the bounds -2, +2 (i.e. absolute value greater than 2), and the leverage values greater than 3 3/50 (P = 3 since we estimated the coefficient for two (2) explanatory variables and the intercept and n = 50 since there were 50 observations) will be flagged. After estimating the model we select the "unusual residuals" and "influential points" options under the tabular options icon. Note that from tables 11 and 12 observations 8, 42, 44, 47, 49 and 50 are classified as outliers.

```
                          Unusual Residuals
--------------------------------------------------------------
                           Predicted                Studentized
Row                Y               Y      Residual     Residual
--------------------------------------------------------------
    44         111.3          85.482        25.818         2.73
    47         115.2         92.1828       23.0172         2.40
    49         129.0         89.2508       39.7492         5.03
--------------------------------------------------------------
```

Table 11

```
                 Influential Points
------------------------------------------------
                        Mahalanobis
Row        Leverage        Distance        DFITS
------------------------------------------------
     8    0.0816156         3.28611     0.560007
    42     0.144802         7.14775      0.58652
    49    0.0947427         4.04401      1.62728
    50     0.339383         23.6798    0.0932134
------------------------------------------------
Average leverage of single data point = 0.06
```

Table 12

Once the outliers are identified one then needs to decide what, if anything, needs to be modified in the data or model. This involves checking the accuracy of the data and/or determining if the outliers represent a specific source of variation. To ascertain any sources of specific variation one looks to see if there is anything common in the set, or subset, of observations flagged as outliers. In Table 11[3] one can see that some of the latter observations (42, 44, 47, 49, and 50) were flagged. Since the data (n = 50) was entered by ascending price, one can see that the higher priced homes were flagged. As a result, for this example, the higher priced homes are receiving a large amount of influence. Hence, since this is cross-sectional data, one might want to split the analysis into two models - one for "lower" priced homes and the second for "higher" priced homes.

Multicollinearity

When selecting a set of explanatory variables for a model, one ideally would like each explanatory to provide unique information that is not provided by the other explanatory variable(s). When explanatory variables provide duplicate information about the dependent variable, then we encounter a situation called multicollinearity. For example, consider our house data again, where the following model is proposed:

$$\textbf{Price} = \textbf{B}_0 + \textbf{B}_1\ \textbf{SQFT} + \textbf{B}_2\ \textbf{BATH} + \textbf{B}_3\ \textbf{TOTAL} + \textbf{ERROR}$$

Clearly there is a relationship among the three (3) explanatory variables. What problems might this create? To answer this, consider the estimation results, which are shown on the following page.

[3] StatGraphics also used two other techniques for identifying outliers (Mahalanobis Distribution and DIFTS), which we have elected not to discuss since from an intuitive level they are similar to the standardized residual/leverage criteria.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: price
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                -42.6274       9.50374       -4.48533          0.0000
sqft                     3.02471      0.296349        10.2066          0.0000
bath                    -10.0432       3.49189       -2.87614          0.0061
total                    10.7836       2.06048        5.23351          0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     30780.2      3      10260.1     131.95       0.0000
Residual                  3576.84     46      77.7575
-----------------------------------------------------------------------------
Total (Corr.)             34357.0     49

R-squared = 89.5892 percent
R-squared (adjusted for d.f.) = 88.9102 percent
Standard Error of Est. = 8.81802
Mean absolute error = 5.89115
Durbin-Watson statistic = 1.53269
```

Table 13

If one were to start interpreting the coefficients individually and noticed the bath has a negative coefficient, they might come to the conclusion that one way to increase the sales price is to eliminate a bathroom. Of course, this doesn't make sense, but it does not mean the model is not useful. After all, when the BATH is altered so are the TOTAL and SQFT. So a problem with multicollinearity is one of ***interpretation*** when other associated changes are not considered. One important fact to remember, is that just because multicollinearity exists, does not mean the model can not be used for meaningful forecasting, provided the forecasts are within the data region considered for constructing the model.

Predicting Values with Multiple Regression

Regression models are frequently used for making statistical predictions. A multiple regression model is developed, by the method of least squares, to predict the values of a dependent, response, variable based on two or more independent, explanatory variables.

Research data can be classified as ***cross-sectional data*** or as ***time series data***. Cross-sectional data has no time dimension, or it is ignored. Consider collecting data on a group of subjects. You are interested in their age, weight, height, gender, and whether they tend to be left-handed. The time dimension in collecting the data is not important and would probably be ignored; even though researchers tend to collect the data within a reasonably short time period.

Time series data is a sequence of observations collected from a process with equally spaced periods of time. For example, in collecting sales data, the data would be collected weekly with the time (the specific week of the year) and sales being recorded in pairs.

Using Cross-sectional Data for Predictions

When using regression models for making predictions with cross-sectional data, it is imperative that you use only the relevant range of the predictor variable(s). When predicting the value of the response variable for a given value of the explanatory variable, one may *interpolate* within the range of the explanatory variables. However, contrary to when using time series data, one **may not *extrapolate* beyond the range of the explanatory variables.** (To predict beyond the range of an explanatory variable is to assume that the relationship continues to hold true below and/or above the range -- something that is not known nor can it be determined. To make such an interpretation is meaningless and, at best, subject to gross error.)

An Example: Using a Regression Model to Predict

Consider the following research problem - a real estate firm is interested in developing a model to predict, or forecast, the selling price of a home in a local community. Data was collected on 50 homes in a local community over a three week period.

The data can consist of both **qualitative** and **quantitative** values. *Quantitative variables are measurable whereas qualitative variables are descriptive.* For example: your height, a quantitative value, is measurable whereas the color of your hair, a qualitative variable, is descriptive.

For our real estate example, the dependent variable (selling price) and the explanatory variables (square feet, number of bathrooms, and total number of rooms) with their corresponding ranges for quantitative variables (low to high). None of the data are qualitative variables.

Table 13. Variable With Range of Values

Variables	Range of Values
Price (selling) ($1000)	30.6 - 165
Square feet (100 ft^2)	8 - 40
Number of Bathrooms	1 - 3
Total number of rooms	5 - 12

As a review, the multiple regression model can be expressed as:

$$\mathbf{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i}$$

The slope, β_i, known as a **net regression coefficient**, represents the unit change in Y per unit change in X_i taking into account (or, holding constant) the effect of the remaining explanatory variables. In our real estate problem, b_1 , where X_1 is in square feet, represents the unit change selling price per unit change in square feet, taking into account the effect of number of bedrooms, and total number of rooms.

The resulting model fitting equation is shown in Table 14.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: price
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 -42.6274       9.50374      -4.48533          0.0000
sqft                      3.02471      0.296349       10.2066          0.0000
bath                     -10.0432       3.49189      -2.87614          0.0061
total                     10.7836       2.06048       5.23351          0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     30780.2      3      10260.1     131.95       0.0000
Residual                  3576.84     46      77.7575
-----------------------------------------------------------------------------
Total (Corr.)             34357.0     49

R-squared = 89.5892 percent
R-squared (adjusted for d.f.) = 88.9102 percent
Standard Error of Est. = 8.81802
Mean absolute error = 5.89115
Durbin-Watson statistic = 1.53269
```

Table 14

Multiple regression analysis is conducted to determine whether the null hypothesis, written as H_o: $\beta_i = 0$ (with i = 0 - 3), can be rejected. If the null hypothesis can be rejected, then there is sufficient evidence of a relationship (or, an association) between the response variable and the explanatory variables in the sample. Table 14 also displays the resulting analysis of variance (ANOVA) for the multiple regression model using the explanatory variables listed in Table 12.

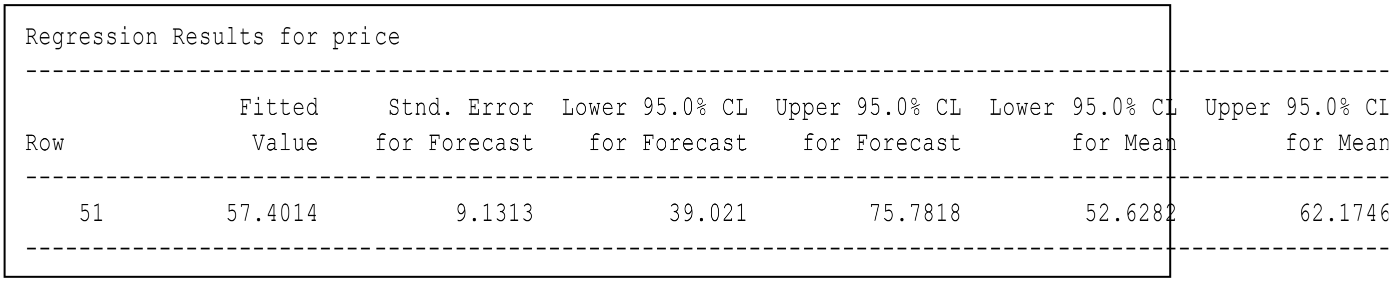
The ANOVA for the full multiple regression shows a p-value equal to 0.0000, thus H_o can be rejected (because the p-value is less than α of 0.05). Since the null hypothesis may be rejected, there is sufficient evidence of a relationship (or, an association) between selling price and the three explanatory variables in the sample of 50 houses.

> **CAUTION:** As stated, when using regression models for making predictions with cross-sectional data, use only the relevant range of the explanatory variable(s). To predict outside the range of an explanatory variable is to assume that the relationship continues to hold true below and/or above the range -- something that is not known nor can be determined. To make such an interpretation is meaningless and, at best, subject to gross error.

Suppose one wishes to obtain a point estimate, along with confidence intervals for both the individual forecasts and the mean, for a home with the following attributes

1500 square feet, 1 bath, 6 total rooms.

To do this using Statgraphics, all one needs to do is add an additional row of data to the data file (HOUSE.SF). In particular one would insert a 15 in the sqft column (remember that the square feet units is in 1,000 's), a 1 in the bath column and a 6 in the total column. We leave the other columns blank, especially the price column, since Statgraphics will treat it as a missing value and hence estimate it. To see the desired output, one runs the regression, using the additional data point, goes to the tabular options icon and selects the "report" option. Table 15 shows the forecasting results for our example.



```
Regression Results for price
------------------------------------------------------------------------------------------------------------------
               Fitted     Stnd. Error  Lower 95.0% CL  Upper 95.0% CL  Lower 95.0% CL  Upper 95.0% CL
Row             Value    for Forecast    for Forecast    for Forecast        for Mean        for Mean
------------------------------------------------------------------------------------------------------------------
    51        57.4014          9.1313          39.021         75.7818         52.6282         62.1746
------------------------------------------------------------------------------------------------------------------
```

Table 15

Summary

In the introduction to this section, **cross-sectional data** and **time series data** were defined. With cross-sectional data, the time dimension in collecting the data is not important and can be ignored; even though researchers tend to collect the data within a reasonably short time period. When predicting the value of the response variable for a given value of the explanatory variable with cross-sectional data, a researcher is restricted to *interpolating* within the range of the explanatory variables. However, a researcher may not *extrapolate* beyond the range of the explanatory variables because it cannot be assumed that the relationship continues to hold true below and/or above the range since such an assumption cannot be validated. Cross-sectional forecasting is stationary, it does not change over time.

On the other hand, time series data is a sequence of observations collected from a process with equally spaced periods of time. Contrary to the restrictions placed on cross-sectional data, when using time series data a major purpose of forecasting is to *extrapolate* beyond the range of the explanatory variables. Time series forecasting is dynamic, it does change over time.

Practice Problem

> As part of your job as personnel manager for a company that produces an industrial product, you have assigned the task of analyzing the salaries of workers involved in the production process. To accomplish this, you have decided to develop the “best” model, utilizing the concept of parsimony, to predict their weekly salaries. Using the personnel files, you select, based on systematic sampling, a sample of 49 workers involved in the production process. The data, entered in the file ***COMPANY***, corresponds to their weekly salaries, lengths of employment, ages, gender, and job classifications.

a. $\hat{y}$ = ____

b. H_0: ______________________ H_1: _______________________

p-value: ___________________ Decision: __________________

c. In the final model, state the value and interpret for R^2_{adj}. R^2_{adj}: ________ %

d. In the final model, state the value and interpret for b_1 . b_1 = ________

e. Predict the weekly salaries for the following employees:

Category	Employee #1	Employee #2
Length of employment (in months)	10	125
Age (in years)	23	33
Gender	female	male
Job classification	technical	clerical

Employee	95% LCL	$\hat{y}$	95% UCL
# 1			
# 2			

[Check documentation on file to ascertain gender coding for female and male. Also check for proper coding for job classification.]

Stepwise Regression

When there exists a large number of potential explanatory variables, a good *exploratory* technique one can utilize is known as stepwise regression. This technique involves introducing or deleting variables one at a time. There are two general procedures under the umbrella of stepwise regression -- forward selection and backwards elimination. A hybrid of both forward selection and backwards elimination exists and is generally just known as stepwise.

In the sections below, we describe the three (3) procedures cited above. In order to follow the discussion, we first need to review the t- test for regression coefficients. Recall that for the model

$$\mathbf{Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots\ldots\ldots + \beta_k X_{k,i} + \varepsilon_t}$$

the t-test for: $H_0 : \beta_k = 0$
$H_1 : \beta_k \neq 0$

actually tests whether the variable X_k should be included in the model. If one rejects $H_{0,}$ then the decision is to keep X_t in the model, whereas if one does not reject H_0 the decision is to eliminate X_t from the model. Since rejecting H_0 is usually done when either $t \leq -2.0$ or $t \geq 2.0$, one can see that having a variable in the model is equated to having a t-value with an absolute value greater than 2. Likewise, if a variable has a corresponding t-value, which is equal to or less than 2 in absolute terms, it should be eliminated from the model.

To simplify the programming for the stepwise procedures, the software packages generally rely on the fact that squaring a distribution gives one a F distribution. Hence, the discussion above about the t value and whether to keep or eliminate the corresponding variable can be expressed as:

> If the F-statistic ($F = t^2$) is greater than 4.0 , then the corresponding variable should be included in the model. If the F-statistic is less than 4.0, then the corresponding variable should not be included in the model.

Given this background information, we now discuss the three (3) stepwise procedures.

Forward Selection

This procedure starts with no explanatory variables in the model, only a constant. It then calculates an F-statistic for each variable and focuses its attention on that variable with the highest F-value. If the *highest* F-value is greater than 4.0, then the corresponding variable is inserted into the model. If the *highest* F-value is less than 4.0, then the process stops. Assuming the first variable is inserted in the model, an F-statistic is then calculated for each of the variables not in the model, conditioned upon the fact that the first variable selected is in the model. The procedure then focuses on the variable with the highest F-value and asks whether the F-value is greater than 4.0. If the answer is *yes***,** the associated variable is inserted into the model and the process continues by calculating an F-statistic for each of the variables not included in the model, conditioned upon the fact that the first two variables selected are included in the model. Once again, the procedure focuses attention on that variable with the largest F-value and determines whether it is larger than 4.0. If the answer is *yes* the associated variable is inserted into the model and the process continues by calculating an F-statistic for each of the variables not included in the model, conditioned upon the fact that the first three variables selected are included in the model. This process continues on until finally either all of the variables have been included in the model or none of the remaining variables are significant.

Backward Elimination

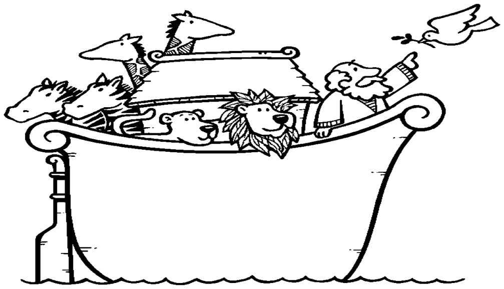
This procedure starts with all of the explanatory variables in the model and successively drops one variable at a time. Given all of the explanatory variables in the model, the “full” regression is run and an F-statistic for each explanatory variable is calculated. The attention now focuses on that variable with the smallest F-value. If the F-value is less than 4.0, then that variable is eliminated from the model and a new regression model is estimated. From this “smaller regression” F-statistics are examined and again the attention now focuses on that variable with the smallest F-value. If the F- value is less than 4.0, then that variable is eliminated from the model and a new regression model is estimated. This process continues on until either all of the explanatory variables have been eliminated from the model or all of the remaining explanatory variables are significant.

Stepwise

This procedure is a hybrid of *forward selection* and *backwards elimination*. It operates the same as forward selection, except at each stage the possibility of deleting a variable, as in backward elimination is considered. Hence, a variable that enters at one stage may be eliminated at a later stage (due to multicollinearity)

Summary

Generally all three stepwise procedures will provide the same model. Under extreme collinear conditions (explanatory variables) the final results may be different. Keep in mind that stepwise procedures are good exploratory techniques, to provide the model builder with some insight. One should not be fooled into thinking that stepwise models are the best because the "computer generates the models." Stepwise procedures fail to consider things such as outliers, residual patterns, autocorrelation, and theoretical considerations.



RELATIONSHIPS BETWEEN SERIES

When building models one frequently desires to utilize variables that have significant linear relationships. In this section we discuss **correlation** as it pertains to cross sectional data, **autocorrelation** for a single time series (demonstrated in the previous chapter), and **cross correlation**, which deals with correlations of two series. Hopefully, the reader will note the relationship between correlation, autocorrelation, and cross correlation.

Correlation

As we mentioned previously, when we talk of statistical correlation we are discussing a value which measures the linear relationship between two variables. The statistic

$$r_{xy} = \frac{\sum\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{Sx\ Sy}$$

where S_y and S_x represent the sample standard deviation of Y and X respectively, measures the strength of the **linear** relationship between the variables Y and X. Again we are not going to dwell on the mathematics, but will be primarily concerned with the interpretation.

To interpret the correlation coefficient, it is important to note that the denominator is included so that values generated are not sensitive to the choice of metrics (i.e. inches vs. feet, ounces vs. pounds, cents vs. dollars, etc.). As a result, the range of possible values for the correlation coefficients range from -1.0 to 1.0.

Since the denominator is always a positive value, one can interpret the **sign of the correlation coefficient** as the indicator of relationship of how X and Y move together. For instance, if the correlation coefficient is positive, this indicates that positive (negative) changes in X tend to accompany positive (negative) changes in Y (i.e. X and Y move in the same direction). Likewise, a negative correlation value indicates that positive (negative) changes in X tend to accompany negative (positive) changes in Y (i.e. X and Y move in opposite directions).

The **absolute value** of the correlation coefficient indicates how strong of a linear relationship two variables have. The closer the absolute value is to 1.0 the stronger the linear relationship.

To summarize we consider the plots in Figure 1, where we show five different values for the correlation coefficient. Note that (1) the sign indicates whether the variables move in the same direction and (2) the absolute value indicates the strength of the linear relationship.

Autocorrelation

As indicated by its name, the **autocorrelation** function will calculate the correlation coefficient for a series and itself in previous time periods. Hence, when analyzing one series and determining how (linear) information is carried over from one time period to another, we will rely on the autocorrelation function.

The autocorrelation function is defined as:

$$r(k) = \frac{\sum\left[\left(x_t - \overline{x}\right)\left(x_{t-k} - \overline{x}\right)\right]}{Sx_t \;\; Sx_{t-k}}$$

where again Sx and Sx(t-k) are the sample standard deviations of X_t and X_{t-k}; which you think about it are the same value. Hence when you substitute X_t and X_{t-k} into the correlation equation for Y and X you can see the similarity. The one difference is with the time element component and hence the inclusion of k. What k represents is the “lag” factor. So when one calculates r(1) that is the sample autocorrelation of a time series variable and itself 1 time period ago, r(2) is the sample autocorrelation of a time series variable and itself 2 time periods ago, r(3) is the sample autocorrelation of a time series variable and itself 3 time periods ago, etc.

To illustrate the value of the autocorrelation function, consider the series ***TSDATA.BUBBLY*** (StatGraphics data sample), which represents the monthly champagne sales volume for a firm. The plot of this series shows a strong seasonality component as shown on the next page in Figure 2.

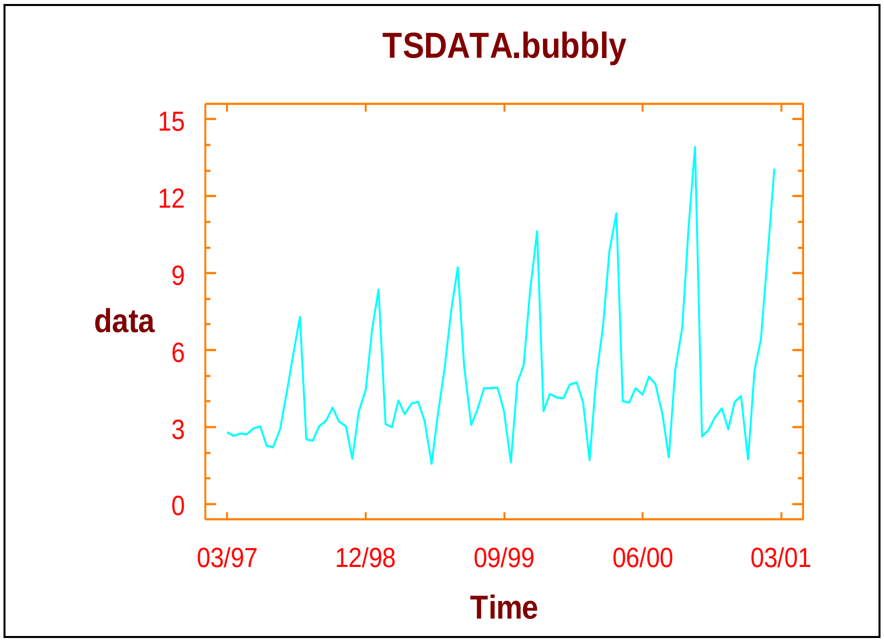


Figure 2. Time Sequence Plot for Bubbly Data

The autocorrelation function can be displayed numerically, Table 1, below:

```
Table 1.    Estimated autocorrelations for TSDATA.bubbly
----------------------------------------------------------------
Lag     Estimate   Stnd.Error     Lag     Estimate   Stnd.Error
----------------------------------------------------------------
 1       .48933      .10911        2       .05787      .13269
 3      -.15498      .13299        4      -.25001      .13512
 5      -.03906      .14052        6       .03647      .14065
 7      -.03773      .14076        8      -.24633      .14088
 9      -.18132      .14592       10      -.00307      .14858
11       .37333      .14858       12       .80455      .15935
13       .40606      .20200       14       .02545      .21150
15      -.17323      .21153       16      -.24418      .21322
17      -.05609      .21652       18       .02920      .21669
19      -.03339      .21674       20      -.20632      .21680
21      -.14682      .21913       22      -.01295      .22029
23       .27869      .22030       24       .60181      .22446
----------------------------------------------------------------
```

The autocorrelation function can also be displayed numerically graphically (where dotted lines -- symmetric about 0 -- represent the significance limits) as shown in Figure 3.

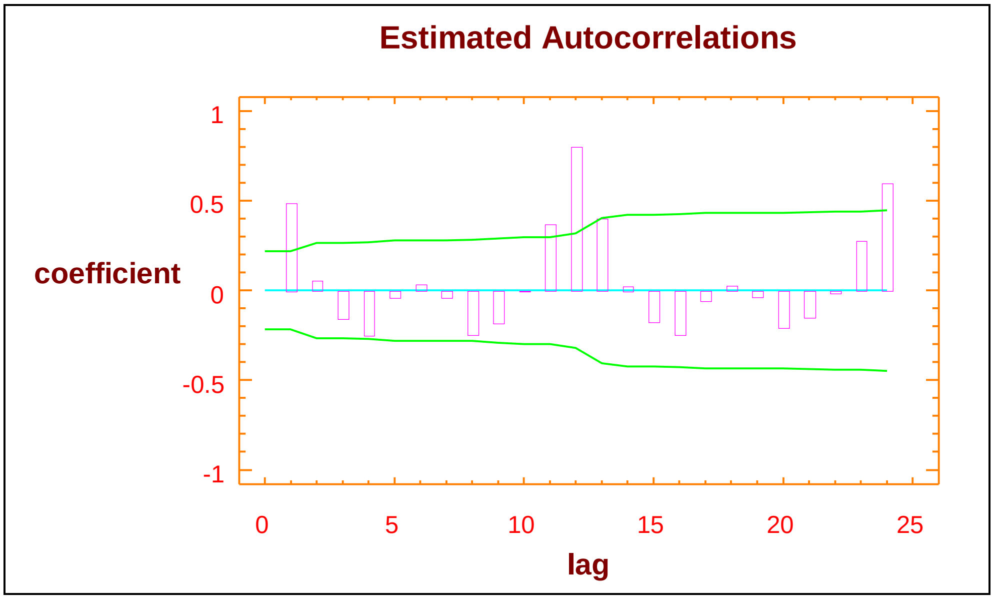


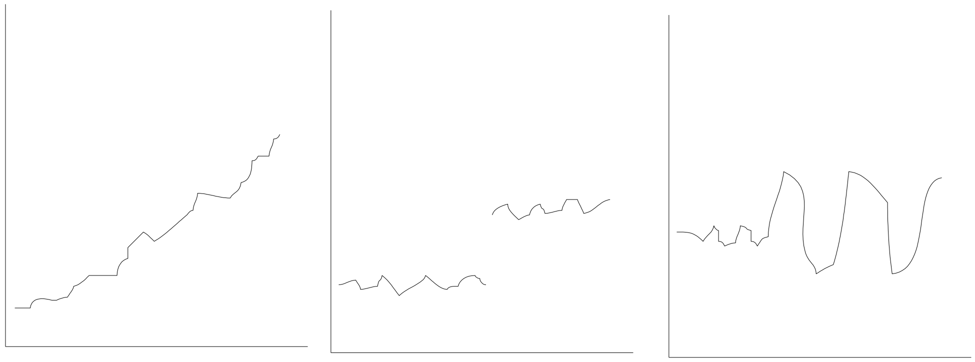
Figure 3. Estimated Autocorrelations

By analyzing the display, the autocorrelation at lags 1, 11, 12, 13, and 24 are all significant (α = 0.05). Hence, one can conclude that there is a linear relationship between sales in the current time period and itself and 1, 11, 12, 13, and 24 time periods ago. The values at 1, 11, 12, 13, and 24 are connected with a yearly cycle (every 12 months).

Stationarity

The next topic we wish to discuss in this section is the cross correlation function, which will be used to examine the relationship between two series displaced by k time periods. This will allow us to begin identifying ***leading indicators.*** However in order to discuss the cross correlation function, we first need to review what it means for a series to be stationary. This discussion is necessary because the interpretation of the cross correlation function only makes useful sense if both series involved are stationary.

Recall, a series is stationary if it has a constant mean and variance. Common departures from stationarity (i.e. non-stationary series) are shown below:



When a series is nonstationary because of a changing variance, one can treat this problem by taking logs of the data [logs in this course will be natural logs (Ln), not common logs (base 10)]. When a series is nonstationary due to a changing mean then one can take differences to treat that problem. If seasonality exists then one may in addition to taking differences of consecutive time periods, take seasonal differences.

If a nonstationary series has a nonconstant mean **and** a nonconstant variance then differences and logs may both be required to achieve a transformation to a stationary series. When taking both logs and differences one must take the logs first (i.e. treat the nonconstant variance and the attack the nonconstant mean). Why?

Cross Correlation

With the knowledge discussed in the autocorrelation section and the stationarity section, we are now prepared to discuss the cross correlation function, which as we said before is designed to measure the linear relationship between two series when they are displaced by k time periods. The cross correlation function is shown below. (The formula is shown on extra large type to highlight the components of the formula.)

$$r_{xy}(k) = \frac{\sum\left[\left(Y_t - \overline{Y}\right)\left(X_{t-k} - \overline{X}\right)\right]}{S_{Y_t}\ S_{X_{t-k}}}$$

To interpret what is being measured in the cross correlation function one needs to combine what we discussed about the correlation function and the autocorrelation function. Again note, like in the autocorrelation function, that k can take on integer values, only now k can take on positive **and** negative values.

For instance, let Y represent SALES and X represent ADVERTISING for a firm. If k = 1, then we are measuring the correlation between SALES in time period t and ADVERTISING in time period t-1. i.e. we are looking at the correlation between SALES in a time period and ADVERTISING in the previous time period. If k = 2, we would be measuring the correlation in SALES in time period t and ADVERTISING two time periods prior. What if k = 3, k = 4, ...? Note that when k is zero we are considering the relationship of ADVERTISING in the same time periods.

When k takes on negative values then our interpretations are the same as above, except that now we are looking at cases were Y (SALES) are leading indicators for X (ADVERTISING). This is the "opposite" of what we were doing with the positive values for k. Note the cross correlation function is not symmetric about 0. i.e.

$$r_{xy}(k) \neq r_{xy}(-k) \quad \text{for all } x,y,\ k \neq 0$$

An Example

To illustrate the cross correlation function, we consider the data TSDATA.units and TSDATA.leadind. This data is sample data from Statgraphics and resides on the network.

The joint plot of **units** and **leadind,** is shown in Figure 4 on the following page. Note how **leadind** "leads" **units**. And how both series are nonstationary. Given at least one of the series is nonstationary, the cross correlation function will be meaningless if it is applied to the original data. Since both series can be transformed to stationary series by simple differences (verify this), we will apply the cross correlation function to the differenced series for both series.

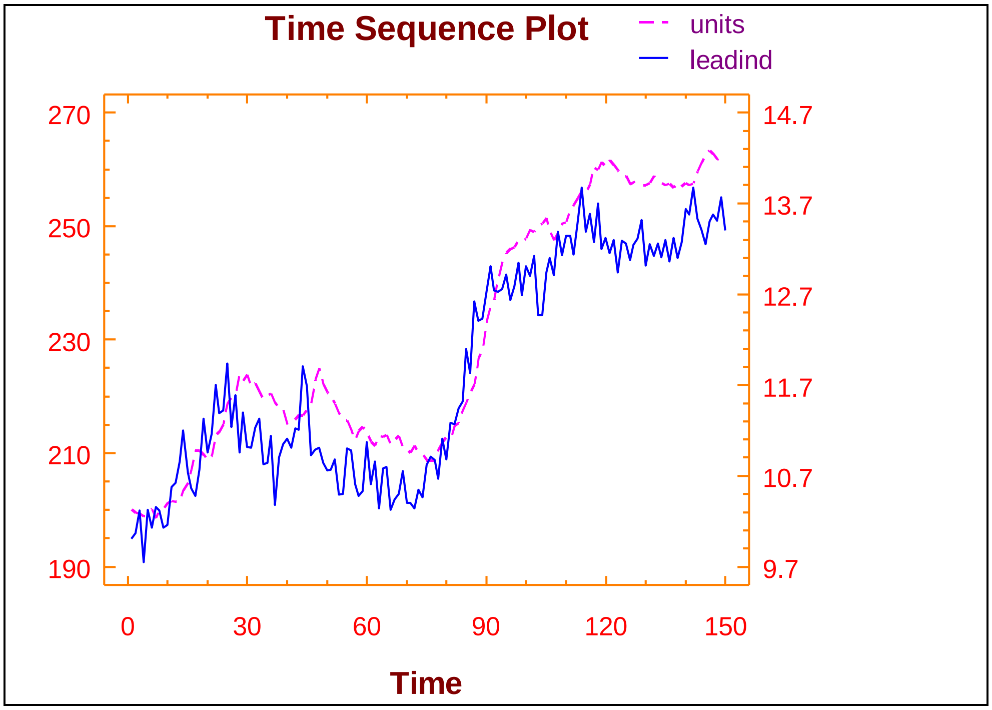


Figure 4. Time Sequence Plot of Lead and Lag Indicators

Looking at the CCF (cross correlation plot) plot displayed in Figure 6 on the next page, we can see significant cross correlation values at lags 2 and 3. Given **leadind** was the input (X_{t-k}) value and **units** is the output (Y_t) value, we can conclude that **leadind** is a leading indicator of **units** by 2 and 3 time periods. So a change in **leadind** will result in a change in **units** two and three time periods later. Note it takes two time periods for a change in **leadind** to show up in **units.**

(Note: for a situation where it is of interest to determine whether advertising leads sales, then advertising would be the ***input*** and sales would be the ***output***.)

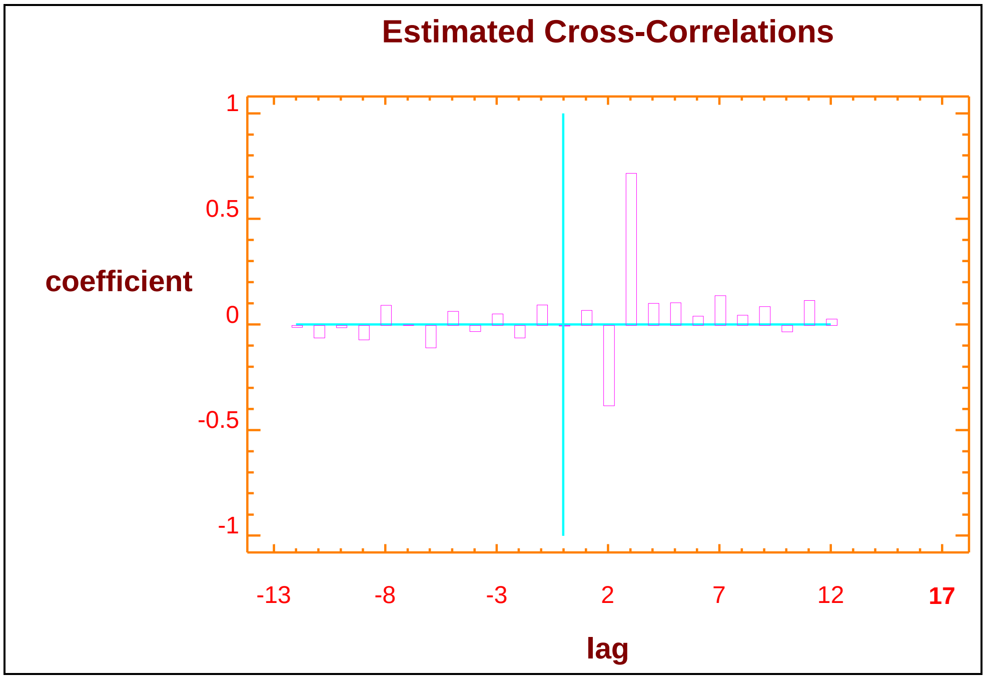


Figure 5. Estimated Cross-Correlations

Questions:

- Does **units** lead **leadind?**
- What do you think would be the relationship between sales and advertising for a firm?
- In the **units/leadind** example, what does the CCF value for k = 0 mean?

Mini-Case

Herr Andres Lüthi owns a bank in Bern, Switzerland. One of Herr Lüthi's requirements of his employees is they must continually solicit unnumbered accounts from foreign investors. Herr Lüthi prefers to call such accounts "CDs" because they have time limits similar to certificates of deposits used in the United States.

Being very computer literate, Herr Lüthi created a file, ***CD***, to store his data. In this historical file, he maintains data of the sales volume of CDs, *volume*, for his bank. All the data is maintained on a monthly basis. Included in the data set are *call* (the number of cold calls Herr Lüthi's employees made each month during the period January 1990 through July 1995), *rate* (the average rate for a CD), and *mail* (the number of mailings Herr Lüthi sent out to potential customers). Because of the excellent services provided by the bank, it is the norm for customers roll their CDs over into new CDs when their original CDs expire.

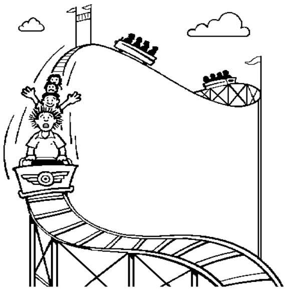
It should be noted that several years ago Herr Lüthi took many of his employees on an extended ski vacation. Records show that the ski vacation was in 1992, February through May. The few non-skiers, who opted to take their holidays in Spain, continued soliciting CDs. They were, of course, credited with any walk-in traffic and any roll over accounts.

You were recently offered a position at Herr Lüthi's bank. As part of your responsibilities, you are to construct a regression model that can be used to analyze the bank's performance with regards to selling CDs. When the Board of Directors met last week, they projected the following for next

Number of Cold Calls	900
Average Rate for a CD	3.50
Number of Mailings	4,500

month:

Prepare your analysis for Herr Lüthi.



INTERVENTION ANALYSIS

In this section we will be introducing the topic of intervention analysis as it applies to regression models. Besides introducing intervention analysis, other objectives are to review the three-phase model building process and other regression concepts previously discussed. The format that will be followed is a brief introduction to a case scenario, followed by an edited discussion that took place between an instructor and his class, when this case was presented in class. The reader is encouraged to work through the analysis on the computer as they read the narrative. (The data resides in the file FRED.SF.

As you work through the analysis, keep in mind that the sequence of steps taken by one analyst may be different from another analysis, but they end up with the same result. What is important is the ***thought process*** that is undertaken.

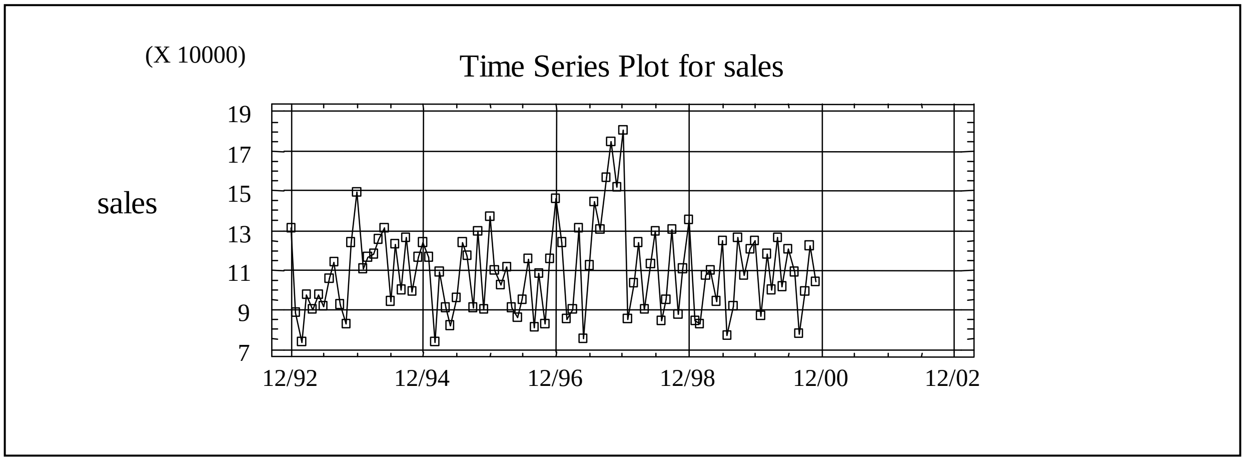
> ***Scenario: You have been provided with the monthly sales (FRED.SALE) and advertising (FRED.ADVERT) for Fred's Deli, with the intention that you will construct a regression model which explains and forecasts sales. The data set starts with December 1992.***

Instructor: What is the first step you need to do in your analysis?
Students: *Plot the data.*

Instructor: Why?
Students: *To see if there is any pattern or information that helps specify the model.*

Instructor: What data should be plotted?
Students: *Let's first plot the series of sales.*

Instructor: Here is the plot of the series first for the sales. What do you see?



Students: *The series seems pretty stationary. There is a peak somewhere in 1997. It is a little higher and might be a pattern.*

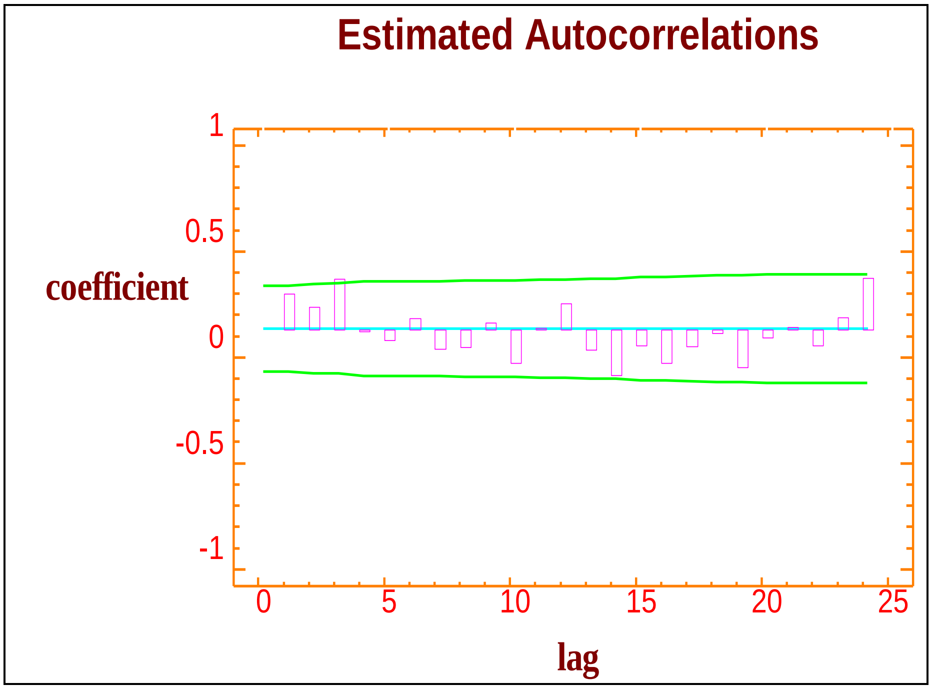
Instructor: What kind of pattern? How do you determine it?
Students: *There may be a seasonality pattern.*

Instructor: How would you see if there is a seasonality pattern?
Students: *Try the autocorrelation function and see if there is any value that would indicate a seasonal pattern.*

Instructor: OK. Let's go ahead and run the autocorrelation function for sales. How many time periods would you like to lag it for?
Students: *Twenty-four.*

Instructor: Why?
Students: *Twenty-four would be a two years worth in a monthly value.*

Instructor: OK, let's take a look at the autocorrelation function of sales for 24 lags.



Instructor: What do you see?

Students: *There appears to be a significant value at lag 3, but besides that there may also be some seasonality at period 12. However, it's hard to pick it up because the values are not significant. So, in this case we don't see a lot of information about sales as a function of itself.*

Instructor: What do you do now?

Students: *See if advertising fits sales.*

Instructor: What is the model that you will estimate or specify?

Students: $Sales_t = \beta_0 + \beta_1\ Advert_t + \varepsilon_1$.

Instructor: What is the time relationship between sales and advertising?

Students: *They are the same time period.*

Instructor: OK, so what you are hypothesizing or specifying is sales in the current time period is a function of advertising in the current time period, plus the error term, correct?

Students: *Yes.*

Instructor: Let's go ahead and estimate the model. To do so, you select model, regression, and let's select a simple regression for right now. The results appear on the following page.

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: sales
Independent variable: advert
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept       113462.0         9049.5        12.5379         0.0000
Slope          -0.119699       0.250226      -0.478362         0.6335
-----------------------------------------------------------------------------



                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   1.08719E8      1    1.08719E8       0.23       0.6335
Residual               4.46602E10     94    4.75108E8
-----------------------------------------------------------------------------
Total (Corr.)          4.47689E10     95

Correlation Coefficient = -0.0492793
R-squared = 0.242845 percent
Standard Error of Est. = 21797.0
```

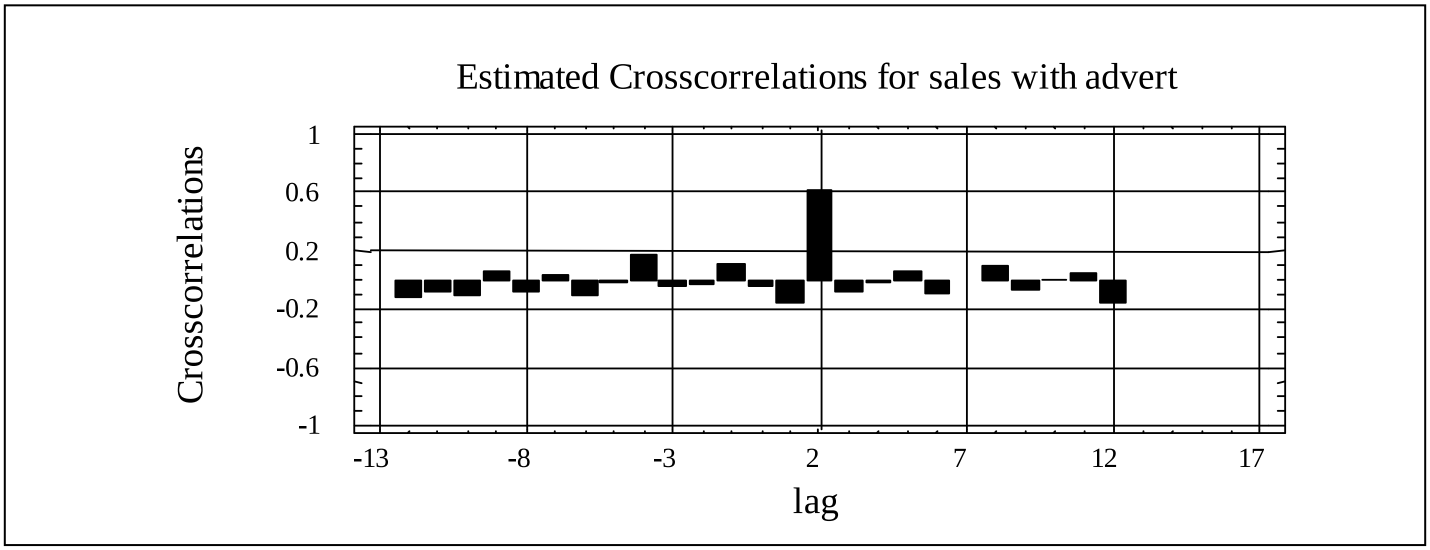
Instructor: What do you see from the result? What are the diagnostic checks you would come up with?
Students: *Advertising is not significant.*

Instructor: Why?
Students: *The p-value is 0.6335; hence, advertising is a non-significant variable and should be thrown out. Also, the R-squared is 0.000, which indicates advertising is not explaining sales.*

Instructor: OK, what do we do now? You don't have any information as its past for the most part, and you don't have any information as advertising as current time period, what do you do?
Students: *To see if the past values of advertising affects sales.*

Instructor: How would you do this?
Students: *Look at the cross-correlation function.*

Instructor: OK. Let's look at the cross-correlation between the sales and advertising. Let's put in advertising as the input, sales as the output, and run it for 12 lags - one year on either side. Here is the result of doing the cross-correlation function, what do you see?



Students: *There is a large "spike" at lag 2 on the positive side. What it means is that there is a strong correlation (relationship) between advertising two time periods ago and sales in the current time period.*

Instructor: OK, then, what do you do now?

Students: *Run a regression model where sales is the dependent variable and advertising lagged two (2) time periods will be the explanatory variable.*

Instructor: OK, this is the model now we are going to specify

$$\mathbf{Sales_t = \beta_0 + \beta_1\ Advert_{t-2} + \varepsilon_1}$$

What we are seeing here is that the sale is a function of advertising two time periods ago. So, at this point this is the model that you have specified. Going to the three-phase model building process, let's now estimate the model, and then we will diagnostically check it. The estimation results for this model are as follows:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: sales
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 56480.4        7184.33         7.8616         0.0000
lag(advert,2)            1.51372       0.199747        7.57819         0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                  1.68543E10      1   1.68543E10      57.43       0.0000
Residual               2.70002E10     92     2.9348E8
-----------------------------------------------------------------------------
Total (Corr.)          4.38544E10     93

R-squared = 38.4323 percent
R-squared (adjusted for d.f.) = 37.7631 percent
Standard Error of Est. = 17131.3
Mean absolute error = 10777.8
Durbin-Watson statistic = 1.1601
```

Instructor: Looking at the estimation results, we are now ready to go ahead and do the diagnostic checking. How would you analyze the results at this point from the estimation phase?

Students: *We are getting 2 lag of advertising as being significant, since the p-value is 0.0000. So, it is extremely significant and the R-squared is now 0.3776.*

Instructor: Are you satisfied at this point?

Students: *No.*

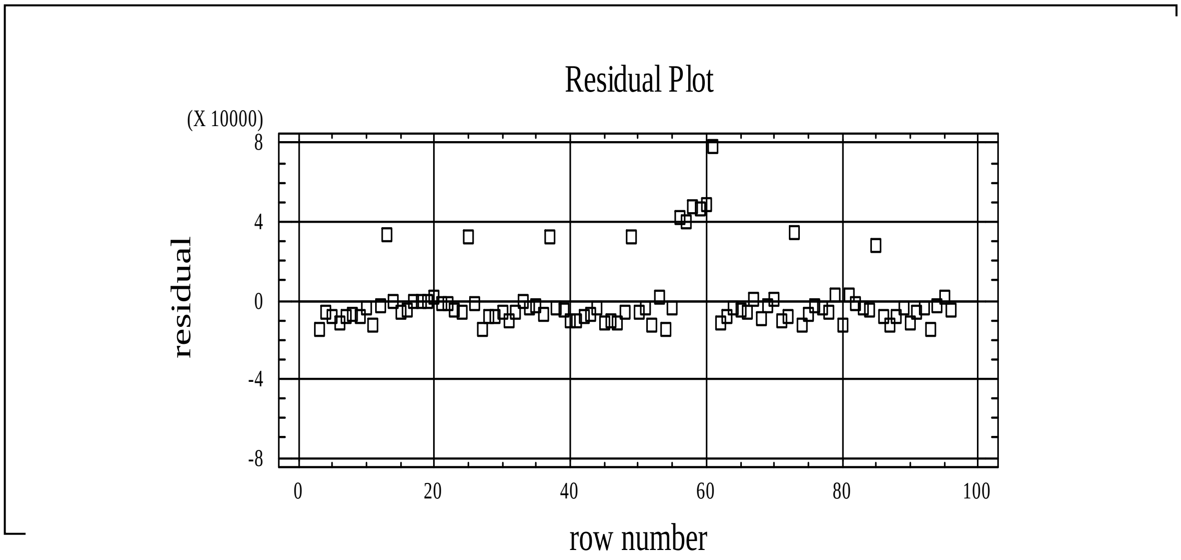
Instructor: What would you do next?

Students: *Take a look at some diagnostics that are available.*

Instructor: Such as what?

Students: *We can plot the residuals, look at the influence measures, and couple other things.*

Instructor: OK. Let’s go ahead and first of all plot the residuals. What do residuals represent? Remember that the residuals represent the difference between the actual values and the fitted values. Here is the plot of the residuals against time (the index):



Instructor: What do you see?
Students: *There is a clear pattern of points above the line, which indicates some kind of information there.*

Instructor: What kind of information?
Students: *It depends on what those values are.*

Instructor: Let us take a look at a feature in Statgraphics. When one maximizes the pane, which displays the residual graph versus time (row), one is then able to click on any point (square) and find out which observation it is by looking above the graph in the "row" box.

We are now able to identify each of the points by lining up the plus mark on each point and clicking. If you do that for the first point, you will notice that X is 13, the second point, X is 25 and the third point, X is 37. The fourth point that is out by itself is 49.

As you see what is going on there, you have a pattern of every 12 months. Recall that we started it off in December. Hence each of the clicked points is in December. Likewise, if you see the cluster in the middle, you will notice that those points correspond to observations 56, 57, 58, 59, 60, and the 61. Obviously, something is going on at observation 56 through 61.

So, if you summarize the residuals, you have some seasonality going on at the month 13, 25,.... i.e. every December has a value, plus something extra happen starting with 56th value and continues on through the 61st value. We could also obtain very similar information by taking a look at the "Unusual Residuals" and "Influential Points"

Instructor: To summarize from our residuals and influential values, one can see that what we have left out of the model at this time are really two factors. One, the seasonality factor for each December, and two, an intervention that occurred in the middle part of 1997 starting with July and lasting through the end of 1997. This may be a case where a particular salesperson came on board and some other kind of policy/event may have caused sales to increase substantially over the previous case. So, what do you do at this point? We need to go back to incorporate the seasonality and the intervention.

Students: *The seasonality can be accounted for by creating a new variable and assigning "1"for each December and "0" elsewhere.*

Instructor: OK, what about the intervention variable?

Students: *Create another variable by assigning a "1" to the months 56, 57, 58, 59, 60, and 61. Or we figure out the values for July through December in 1997. i.e. "1" for the values from July 97 to December 1997 inclusive, and zero elsewhere.*

Instructor: Very good. So, what we are going to do is to run a regression with these two additional variables. Those variables are already included in the file. One variable is called FRED.INTERVENT and if you look at it, it has "1" for the values from 56 to 61, and "0" elsewhere. Other variable FRED.DEC has values of "1" only for December values, "0" elsewhere. So, what is the model we are going to estimate?

Students: $Sales_t = \beta_0 + \beta_1\ Advert_{t-2}\ \beta_2\ Dec_t + \beta_3\ Intervent_t + \varepsilon_1$.

Instructor: What does this model say in words at this point?

Students: *Sales in the current time period is a function of advertising two time periods ago, a dummy variable for December and intervention variable for the event occurred in 1997.*

Instructor: Good. Let's summarize what we have done.

You started off with a model that has advertising two time periods ago as explanatory variable, but you say some information was not included in that model. That is, we are missing some information that is included in the data. Then, we looked at the residuals and the influence values, and we came up with two new variables that incorporated that missing information. Having re-specified the model, we are now going to re-estimate, and to diagnostic check the revised model. The estimation for the revised model are shown on the following page:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: sales
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 42539.4        1632.99          26.05         0.0000
lag(advert,2)            1.73904      0.0447512        38.8602         0.0000
december                 38262.6        1511.04         25.322         0.0000
intervent                50700.6        1614.95        31.3946         0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                  4.2551E10      3   1.41837E10     979.39       0.0000
Residual                1.3034E9     90    1.44822E7
-----------------------------------------------------------------------------
Total (Corr.)        4.38544E10     93

R-squared = 97.0279 percent
R-squared (adjusted for d.f.) = 96.9288 percent
Standard Error of Est. = 3805.55
Mean absolute error = 3015.79
Durbin-Watson statistic = 1.90283
```

Instructor: Given these estimation results, how would you analyze (i.e. diagnostically check) the revised model?

Students: *All the variables are significant since the p-values are all 0.0000 (truncation). In addition, R-squared value has gone up tremendously to 0.969 (roughly 97 percent). In other words, R^2 has jumped from 37 percent to approximately 97 percent, and the standard error has gone down substantially from 17000 to about 3800. As a result, the model looks much better at this time.*

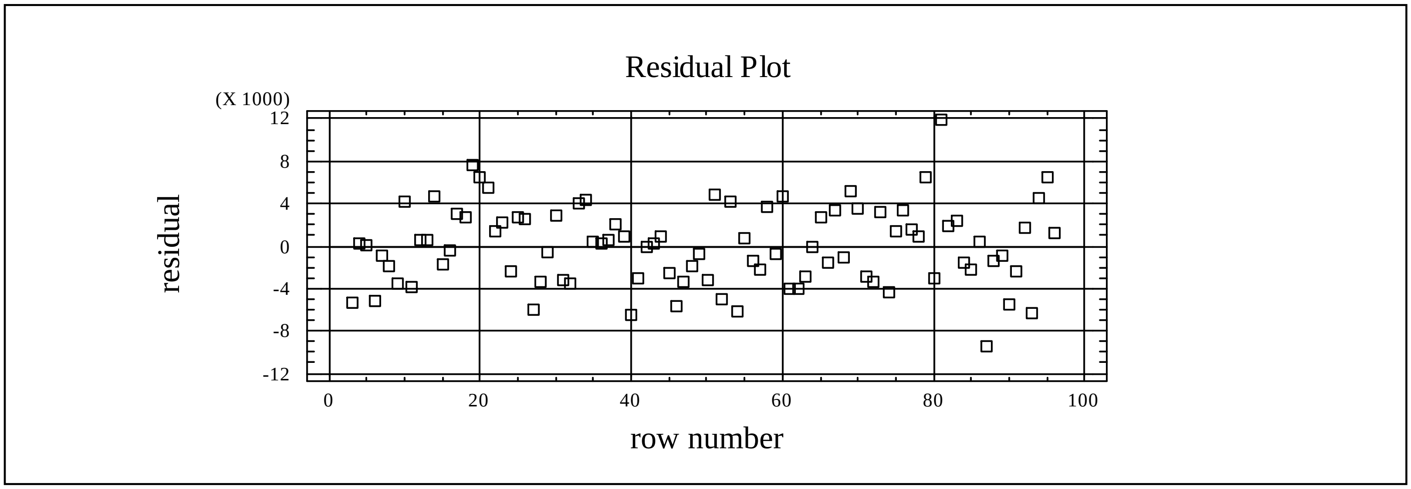
Instructor: Is there anything else you would do?

Students: *Yes, we will go back to diagnostic check again to see if this revised model still has any information that has not been included, and hence can be improved.*

Instructor: What is some diagnostic checking you would try?

Students: *Look at the residuals again, and plot it against time.*

Instructor: OK, here is the plot of the residual against time. Do you see any information?



Students: *No, the pattern looks pretty much random. We cannot determine any information left out in the model with the series of the structure.*

Instructor: OK, anything else you would look at?
Students: *Yes, let us look at the influence measures.*

Instructor: OK, when you look at the "Unusual Residuals" and "Influential Points" options, what do you notice about these points.
Students: *They have already been accounted for with the December and Intervention variables.*

Instructor: Would you do anything differently to the model at this point?
Students: *We don't think so.*

Instructor: **Unless you are able to identify those points with particular events occurred, we do not just keep adding dummy variables in to get rid of the values that have been flagged as possible outliers. As a result, let us assume that we have pretty much cleaned things up, and at this point, you can be satisfied with the model that you have obtained.**

Summary

The objectives in this section, once again are to introduce the concepts of intervention analysis, and review the three-phase model building process. To do this, we look at the situation where we have sales and advertising, in particular, we have monthly values starting in December 1997.

The three-phase model building process talks about specifying, estimating, and diagnostic checking a model. In our analysis the first step we did was try to decide what would be an appropriate model to specify, that was what variable or variables helped explain the variation in sales. As we saw, advertising the current time period did not affect sales. When we used the cross-correlation function, however, we were able to see that advertising two time periods prior had an effect on sales. Thus, we ran the simple linear regression of sales against advertising two time periods prior. From this regression, we look at the diagnostic checks and noticed that a fair amount of information had been left out of the model. In particular, we had left out two factors. The first one was the seasonality factor that occurred in each December, and the second one was an intervention that happened in the last half of 1997, from July to December 1997. To incorporate these two factors into the model, we set up two additional variables. The revised model increased R-squared substantially and reduced the means-squared. Thus, the revised model was our final model.

SAMPLING

As an efficient method to obtain information about a population, one frequently needs to sample from a population. There are many different probabilistic sampling methods. In addition to random sampling, two other frequently used techniques are stratified sampling and systematic sampling[1]. The type of sampling method appropriate for a given situation depends on the attributes of the population being sampled, sampling cost, and desired level of precision.

Random Sampling

A simple random sample is a sample in which every group of size n has an equal chance of being selected. In order to conduct a random sample, one needs the frame (listing of all elements) and then either by "drawing from the hat" or using a random number table[2] one obtains the elements selected for the sample.

Stratified Sample

A stratified sample is appropriate to use when the population of concern has subpopulations (strata) that are homogenous within and heterogeneous between each other, with regards to the parameter of concern. The reason it may be appropriate to use stratified sampling, as opposed to simple random sampling of the whole population, is that each subgroup will have relatively smaller variances than the overall population. Hence, when we combine the results from the different subgroups, the aggregated variance (standard error) will be smaller, than the same size sample from the entire population using simple random sampling.

For example, assume our desires to estimate the average number of hours business students study per week. One could use a simple random sample. However, if one were to stratify based upon

[1] There are many other techniques available but we will restrict our discussion to these.

[2] Many software packages, such as Stat Graphics, have random number generators.

concentrations[3], take a random sample from each concentration, then the aggregated result would probably be more precise (smaller confidence interval) than the one from a random sample (same sample size). The greater precision would come from the aggregation of strata (subpopulations) whose individual variances are less than the variance of the entire population.

Systematic Sample

Systematic sampling is a widely used technique when there is no pattern to the way in which the data set is organized. The lack of pattern is important since a systematic sample involves selecting every nth observation. For example, one may be selecting every 4th observation. Clearly, the technique could provide a biased estimate if there is a periodicity (seasonality) to the data and the sampling interval is a multiple of the period.

Comparison Of Survey Sampling Designs

Design	**How to Select**	**Strengths/Weaknesses**
Simple Random	Assign numbers to elements using random numbers table.	Basic, simple, often costly. Must assign a number to each element in target population.
Stratified	Divide population into groups that are similar within and different between variable of interest. Use random numbers to select sample from each Stratum.	With proper strata, can produce very accurate estimates. Less costly than simple random sampling. Must stratify target population correctly.
Systematic	Select every kth element from a list after a random start.	Produces very accurate estimates when elements in population exhibit order. Used when simple random or stratified sampling is not practical [e.g.: population size not known]. Simplifies selection process. Do not use with periodic populations.

CROSSTABULATIONS

[3] Other discriminating variables could be used, such as, age, premajor vs. upper division, etc...

In this section we will be focusing our attention on a technique frequently used in analyzing survey results, cross tabulation. The purpose of cross tabulation is to determine if two variables are independent or whether there is a relationship between them.

To illustrate cross tabulation assume that a survey has been conducted in which the following questions were asked:

-- What is your age
____ less than 25 years ____ 25-40 _____ more than 40

-- What paper do you subscribe to
____ Chronicle ____ BEE ___ Times

-- What is your annual household gross income
____ < $15,000 ____ $15,000 - $40,000 ___ >$40,000

Letting the first response for each question be recorded as a 1, the second as a 2 and the third as a 3[4], the file ***CLTRES.SF*** contains 200 responses.

We will first consider the hypothesis test generally referred to as *a test of dependence*:

H_0: AGE and PAPER are **not** dependent
H_1: AGE and PAPER are dependent.

To perform this test via Statgraphics, we first pull up the data file CLTRES.SF, then we go to the main menu and select

Describe
Categorical Data
Cross tabulation

and fill one of the variables as the row variable and the other as the column variable. For our example we will select Age as the row variable and paper as the column variable. For the desired output we go to the tabular options and select the Chi-square and frequency table options.

[4] For example for the second question about the paper, we will create a variable called PAPER, with Chronicle = 1, BEE = 2 and Times = 3.

The chi-square option gives us the value of the chi-square statistic for the hypothesis (see Figure 2). This value is calculated by comparing the actual observed number for each cell (combination of levels for each of the two variables) and the expected number under the assumption that the two variables are independent.

Figure 2

```
Chi-Square Test
------------------------------------------
      Chi-Square        Df         P-Value
------------------------------------------
            2.63         4          0.6218
------------------------------------------
```

Since the p-value for the chi-square test is 0.6218, which exceeds the value of α = 0.05, we conclude that there is not enough evidence to suggest that AGE and PAPER are dependent. Hence it is appropriate to conclude that age is *not* a factor in determining who subscribes to which paper. Selecting the **frequency option** provides us with the following output (pane):

Figure 3

```
Frequency Table for age by paper

                                                          Row
             1            2            3                 Total
            --------------------------------------
1           |        13 |        44 |        14 |         71
            |      6.50 |     22.00 |      7.00 |      35.50
            --------------------------------------
2           |        22 |        52 |        10 |         84
            |     11.00 |     26.00 |      5.00 |      42.00
            --------------------------------------
3           |        11 |        27 |         7 |         45
            |      5.50 |     13.50 |      3.50 |      22.50
            --------------------------------------
Column               46          123           31          200
Total             23.00        61.50        15.50       100.00
```

Note that the top entry for each cell represents the actual number of responses for the cell from the survey. The bottom entry in each cell represents the cell's percentage for the entire sample (array). By right clicking on the output pane displayed in Figure 3, one can choose the option pane analysis and select either column or row percentages, for the lower entry. If the hypothesis test had concluded that AGE and PAPER were dependent, then a comparison of the cell's percentage with the value in the Column Total column would suggest how the variables are related. If, looking at Figure 1, we had selected the *row option*, then one should substitute row in the above discussion for column.

THE READER IS ENCOURAGED TO ANALYZE WHETHER PAPER AND INCOME ARE RELATED.

Practice Problem

A survey was administered to determine whether various categories describing a student were independent. Part of the survey questionnaire appears below:

PLEASE PROVIDE THE REQUESTED INFORMATION BY CHECKING (ONCE).

What is your:

- age ____ < 18 ____ 18 - 26 ____ > 26
- gender ____ male ____ female
- course load ____ < 6 units ____ 6 - 12 units ____ > 12 units
- gpa __ < 2.0 __ 2.0 - 2.5 ___ 2.6 - 3.0 __ 3.1 - 3.5 __ > 3.5
- annual income ___ < $10,000 ___ $10,000 - $20,000 ___ > 20,000

The information is coded and entered in the file ***STUDENT.SF*** by letting the first response be recorded as a 1, the second as a 2, etc.

a. Test whether a relationship exists between the categories “age” and “gpa.”
H_0: ____
H_1: ____
p-value: _________ Decision: __________________________

b. Test whether a relationship exists between the categories “gender” and “income.”
H_0: ____
H_1: ____
p-value: _________ Decision: __________________________

c. Describe your observation of the table display for the categories “gender” and “income.”

THE ANALYSIS OF VARIANCE

In this section we will study the technique of **analysis of variance** (ANOVA), which is designed to allow one to test whether the means of more than two qualitative populations are equal. As a follow up, we will discuss what interpretations can be made should one decide that the means are statistically different.

We will discuss two different models (experimental designs), one-way ANOVA and two-way ANOVA. Each model assumes that the random variable of concern is continuous and comes from a normal distribution[5] and the sources of specific variation are ***strictly qualitative***. A one-way ANOVA model assumes there is only one (1) possible source of specific variation, while the two-way ANOVA model assumes that there are two (2) sources of specific variation. Each model assumes that the populations, defined by the specific estimate of the amount of common variation that exists, and compares this value with an estimate of the variation due to the source(s) of specific variation. This comparison, which will be done via an F test, will allow one to determine if there is a significant difference between the means of the populations, as defined by the specific source of variation.

One-Way Analysis Of Variance

As stated, the one-way ANOVA model assumes that the variation of the random variable of concern is made up of common variation and one possible source of specific variation, which is qualitative. The purpose of the one-way ANOVA analysis is to see if the population means of the different populations as defined by the specific source of variation are equal or not. For example, assume you are the utility manager for a city and you want to enter into a contract for a single supplier of streetlights. You are currently considering four possible vendors. Since their prices are identical, you wish to see if there is a significant difference in the mean number of hours per streetlight.[6]

Design

The design we employ, randomly assigns experimental units to each of the populations. In the street light example we will randomly select light bulbs from each population and then randomly assign them to various streetlights. When there are an equal number of observations per population, then the design is said to be a balanced design. Most texts when introducing an one-way ANOVA discuss a balanced design first, since the mathematical formulas that result are easier to present for a balanced design than an unbalanced design. Since our presentation will not discuss the formulas, what we present does not require a balanced design, although our first example will feature a balanced design.

> Going back to our example, we randomly selected 7 light bulbs from each of the populations and recorded the length of time each bulb lasted until burning out. The results are shown below where value recorded is in 10,000 hours.

GE	DOT	West	Generic
2.29	1.92	1.69	2.22
2.50	1.92	1.92	2.01
2.50	2.24	1.84	2.11
2.60	1.92	1.92	2.06
2.19	1.84	1.69	2.19
2.29	2.00	1.61	1.94
1.98	2.16	1.84	2.17

One can easily calculate the ***sample means*** (XBARS) for each population with the results[7] being 2.34, 2.00, 1.79 and 2.10 for GE, DOT, West, and Generic respectively. Recall that our objective is to determine if there is a statistically significant difference between the four ***population means***, not

[5] The results of the ANOVA models are robust to the assumption of normality (i.e. one need not be concerned about the normality assumption).

[6] In this example the random variable is the number of hours per light and the source of specific variation is the different vendors (qualitative).

the sample means. To do this, note that there is variation within each population and between the populations. Since we are assuming that the within variations are all the same, a significant between population variance will be due to a difference in the population means. To determine if the between population variation is significant, we employ the following Statgraphics steps so that we can conduct the hypothesis test:

H_0: All of the population means are the same
H_1: Not all population means are the same via an F statistic.

Create a StatGraphics file [LGHTBULB -- notice spelling, 8 letters] with three variables. The first variable [HRS] represents the measured value (hours per light bulb in 10,000 hours) and the second variable [BRAND] indicates to which population the observation belongs. This can be accomplished by letting GE be represented by a 1, DOT by a 2, West by a 3, and Generic by a 4. We will also created a third variable [Names] which is unnecessary for the Windows version of Statgraphics[8].

```
ROW        HRS        BRAND      NAMES
-----      -------    -------    ---------
  1          2.29        1        GE
  2          1.92        2        DOT
  3          1.69        3        WEST
  4          2.22        4        GENERIC
  .            .         .
  .            .         .
 25          1.98        1
 26          2.16        2
 27          1.84        3
 28          2.10        4
```

Using the created the data file ***LGHTBULB.SF,*** as shown above, we are now able to select the one way ANOVA option in Statgraphics by going to the main menu and selecting:

Compare
    Analysis of Variance
        One-way ANOVA

[7] There is some rounding.

[8] The data file accessed from the WWW page, was released prior to our converting over this file to the Windows version of Stat Graphics.

and declaring **hours** as the dependent variable, along with **brand** as the factor variable.

The resulting output pane, when selecting the ANOVA table option, under tabular options, is

```
ANOVA Table for hours by brand

                            Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Between groups             1.08917      3     0.363057      15.62       0.0000
Within groups             0.557714     24    0.0232381
-----------------------------------------------------------------------------
Total (Corr.)              1.64689     27
```

Table 1. Output for One-Way ANOVA

From this output we can now conduct the hypothesis test:

H_0: All four population means are the same
H_1: Not all four population means are the same

by means of the F test. Note that the F-ratio is the ratio of the between groups (populations) variation and the within groups (populations) variation[9]. When this ratio is *large enough,* then we say there is significant evidence that the population means are not the same. To determine what is large enough, we utilize the p-value (Sig. level) and compare it to alpha. Setting $\alpha = 0.05$, we can see for our example that the p-value is less than alpha. This indicates that there is enough evidence to suggest that the population means are different and we reject the null hypothesis.

To go one step further and see what kind of interpretation one can make about the population means, when it is determined that they are not all equal, we can utilize the means plot option under the graphics option icon. The resulting pane is shown below

[9] The mean square values are estimates of the respective variances.

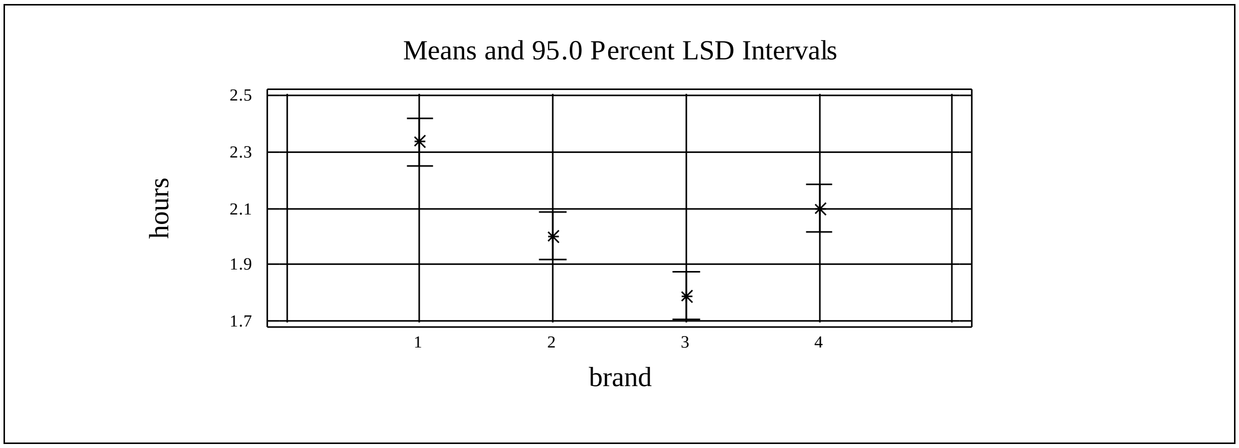


Figure 1. Intervals for Factor Means

To interpret the means plot, note that the vertical axis is numeric and the figures depicted for each brand covers the confidence intervals for the respective population mean. When the confidence intervals overlap then we conclude the population means are not significantly different, when there is no overlap we conclude that the population means are significantly different. The interpretations are done taking the various pair-wise comparisons. Interpreting Figure 1, one can see that GE (brand 1) is significantly greater than all of the other three brands, WEST (brand 3) is significantly less than all of the others and that DOT (brand 2) and GENERIC (brand 4) are not significantly different. The means table provides the same information but in a numerical format .

Practice Problems

1. A consumer organization was interested in determining whether any difference existed in the average life of four different brands of walkmans. A random sample of four batteries of each brand was tested. Using the data in the table, at the 0.05 level of significance, is there evidence of a difference in the average life of these four brands of Walkman batteries? [Create the file ***WALKBAT.SF***.]

Brand 1	**Brand 2**	**Brand 3**	**Brand 4**
12	19	20	14
10	17	19	21
18	12	21	25
15	14	23	20

2. A toy company wanted to compare the price of a particular toy in three types of stores in a suburban county: Discount toy stores, specialty stores, and variety stores. A random sample of four discount toy stores, six specialty stores, and five variety stores was selected. At the 0.05 level of significance, is there evidence of a difference in the average price between the types of stores? [Create the file ***TOY.SF***.]

Discount Toy	**Specialty**	**Variety**
12	15	19
14	18	16
15	14	16
16	18	18
	18	15
	15	

Two-Way Analysis Of Variance

Given our discussion about the one-way ANOVA model, we can easily extend our discussion to a two way ANOVA model. As stated previously, the difference between a one-way ANOVA and two-way ANOVA depends on the number ***qualitative sources of specific variation*** for the variable of concern.

The two-way ANOVA model we will consider has basically the same assumptions as the one- way ANOVA model presented previously. In addition we will assume the factors influence the variable of concern in an additive fashion. The analysis will be similar to the one-way ANOVA, in that each factor is analyzed.

To illustrate the two-way ANOVA model we consider an example where the dependent variable is the sales of Maggie Dog Food per week. In its pilot stage of development Maggie Dog Food is packaged in four different colored containers (blue, yellow, green and red) and placed at different shelf heights (low, medium, and high). As the marketing manager you are interested in seeing what impact the different levels for each of the two factors have on sales. To do this you randomly assign different weeks to possess the different combinations of package colors and shelf height. The results are shown below:

		Shelf Height		
		Low	**Med**	**High**
Can Color	**Blue**	125	140	152
	Yellow	112	130	124
	Green	85	105	93
	Red	85	97	98

Given this design, we can test two sets of the hypotheses.

H_0: The population means for all four colors is the same
H_1: The population means for at least two colors are different

and

H_0: The population means for the different shelf heights are the same
H_1: The population means for at least two of the shelf heights are different

To conduct this analysis using Statgraphics we enter the data into a file called ***DOG.SF*** as shown in Table 2 below:

Table 2.

SALES	COLOR	HGT
125.	B	L
112.	Y	L
85.	R	L
85.	G	L
140.	B	M
130.	Y	M
105.	R	M
97.	G	M
152.	B	H
124.	Y	H
93.	R	H
98.	G	H

Now that the data is entered into the file ***DOG.SF***, we are ready to have Statgraphics generate the required output. To accomplish this we escape back to the main menu and select

Compare
Analysis of Variance
Multifactor ANOVA

then select **SALES** as the dependent variable and for the factors we select **COLOR** and **HEIGHT.** {We do not choose to consider a covariate for this model). When selecting the tabular option ANOVA Table, we get the following pane:

```
Analysis of Variance for sales - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:color                      4468.33      3        1489.44      47.75     0.0001
 B:height                     654.167      2        327.083      10.49     0.0110

RESIDUAL                      187.167      6        31.1944
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             5309.67     11
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Table 4

Looking at the two-way ANOVA table (Table 4) one can see that the total variation is comprised of variation for each of the two factors (height and color) and the residual. The F-ratios for the factors are significant as indicated by their respective p-values. Hence, one can conclude that there is enough evidence to suggest that the means are not all the same for the different colors and that the means are not all the same for the different shelf heights.

To determine what one can conclude about the relationship of the population means for each of the factors we look at the mean plot (table) for each of the factors.[see Graphics options] The mean plot for shelf height and color are shown in Figures 5 and 6 [10].

[10] To change the means plot from one variable to the other, one needs to right click on the pane and choose the appropriate Pane Option(s).

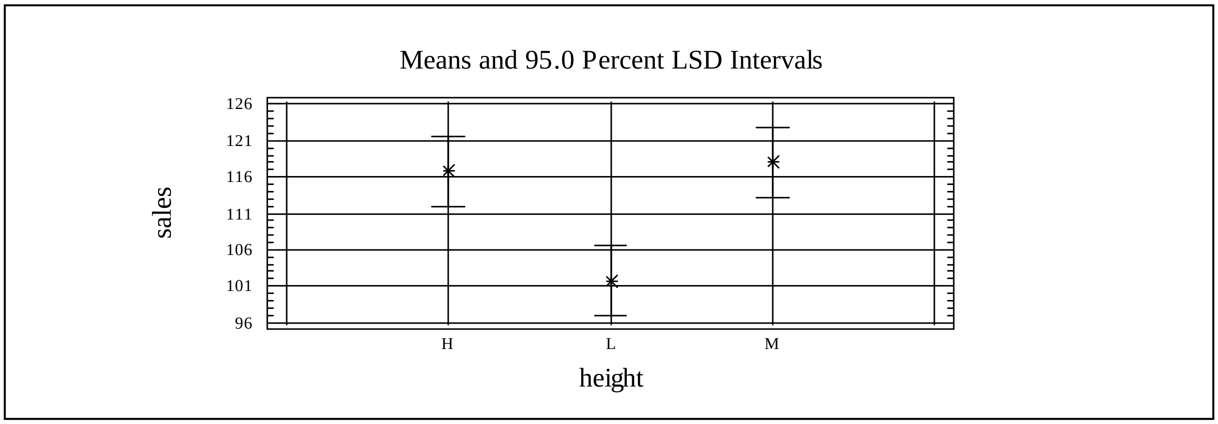


Figure 5

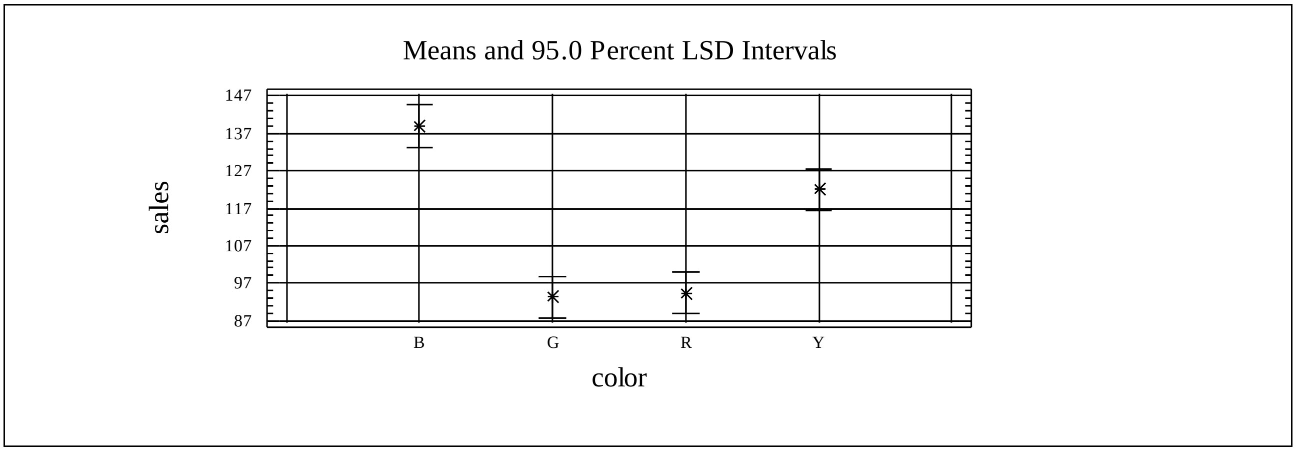


Figure 6

Interpreting the means plots just like we did for the one-way ANOVA example, we can make the following conclusions. With regards to shelf height, the low shelf height has a lower population mean than both the medium and high shelf heights, while we are unable to detect a significant difference between the medium and high shelf heights. With regards to the colors, the blue population mean is greater than the yellow population mean which is greater than both the green population mean and the red population mean and that we are unable to detect a significant difference between the green and red population means. The mean tables (tabular options) provide the same results as the means plots, just that it is given in numerical format.

Practice Problems

1. The Environmental Protection Agency of a large suburban county is studying coliform bacteria counts (in parts per thousand) at beaches within the county. Three types of beaches are to be considered -- ocean, bay, and sound -- in three geographical areas of the county -- west, central, and east. Two beaches of each type are randomly selected in each region of the county. The coliform bacteria counts at each beach on a particular day were as follows:

		Geographic Area	
Type of Beach	**West**	**Central**	**East**
Ocean	25 20	9 6	3 6
Bay	32 39	18 24	9 13
Sound	27 30	16 21	5 7

Enter the data and save as the file ***WATER.SF***.

At the 0.05 level of significance, is there an

a. effect due to type of beach?

H_0: ________________________ H_1: ________________________

p-value: ____________________ Decision: __________________

b. effect due to type of geographical area?

H_0: ________________________ H_1: ________________________

p-value: ____________________ Decision: __________________

c. effect due to type of beach and geographical area? ***OPTIONAL***

H_0: ________________________ H_1: ________________________

p-value: ____________________ Decision: __________________

d. Based on your results, what conclusions concerning average bacteria count can be reached?

2. A videocassette recorder (VCR) repair service wished to study the effect of VCR brand and service center on the repair time measured in minutes. Three VCR brands (A, B, C) were specifically selected for analysis. Three service centers were also selected. Each service center was assigned to perform a particular repair on two VCRs of each brand. The results were as follows:

Service Centers	Brand A	Brand B	Brand C
1	52 57	48 39	59 67
2	51 43	61 52	58 64
3	37 46	44 50	65 69

Enter the data and save as the file ***VCR.SF***

At the .05 level of significance:

(a) Is there an effect due to service centers?
(b) Is there an effect due to VCR brand?
(c) Is there an interaction due to service center and VCR brand? ***OPTIONAL***

3. The board of education of a large state wishes to study differences in class size between elementary, intermediate, and high schools of various cities. A random sample of three cities within the state was selected. Two schools at each level were chosen within each city, and the average class size for the school was recorded with the following results:

Education Level	City A	City B	City C
Elementary	32, 34	26, 30	20, 23
Intermediate	35, 39	33, 30	24, 27
High School	43, 38	37, 34	31, 28

Enter the data and save as the file ***SCHOOL.SF***.

At the .05 level of significance:

(a) Is there an effect due to education level?
(b) Is there an effect due to cities?
(c) Is there an interaction due to educational level and city? ***OPTIONAL***

4. The quality control director for a clothing manufacturer wanted to study the effect of operators and machines on the breaking strength (in pounds) of wool serge material. A batch of material was cut into square yard pieces and these were randomly assigned, three each, to all twelve combinations of four operators and three machines chosen specifically for the equipment. The results were as follows:

Operator	Machine I	Machine II	Machine III
A	115 115 119	111 108 114	109 110 107
B	117 114 114	105 102 106	110 113 114
C	109 110 106	100 103 101	103 102 105
D	112 115 111	105 107 107	108 111 110

Enter the data and save as the file ***SERGE.SF***.

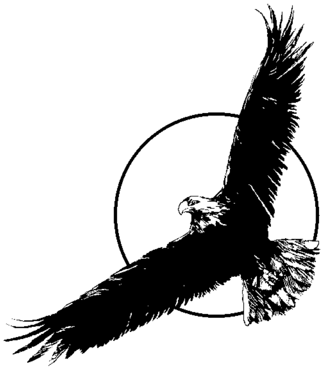
At the .05 level of significance:

(a) Is there an effect due to operator?
(b) Is there an effect due to machine?
(c) Is there an interaction due to operator and machine? ***OPTIONAL***

Appendices

Quality

The Concept of Stock Beta



QUALITY

Common Causes and Specific Causes

As stated early, and repeated here because of the concept's importance, in order to reduce the variation of a process, one needs to recognize that the total variation is comprised of **common causes** and **specific causes**. Those factors, which are not readily identifiable and occur randomly are referred to as the **common causes**, while those which have large impact and can be associated with special circumstances or factors are referred to as **specific causes**.

It is important that one get a feeling of a specific source, something that can produce a significant change and that there can be numerous common sources which individually have insignificant impact on the processes variation.

Stable and Unstable Processes

When a process has variation made up of only common causes then the process is said to be a stable process, which means that the process is in statistical control and remains relatively the same over time. This implies that the process is predictable, but does not necessarily suggest that the process is producing outputs that are acceptable as the amount of common variation may exceed the amount of acceptable variation. ***If a process has variation, which is comprised of both common causes and specific causes, then it is said to be an unstable process -- the process is not in statistical control***. An unstable process does not necessarily mean that the process is producing unacceptable products since the total variation (common variation + specific variation) may still be less than the acceptable level of variation.

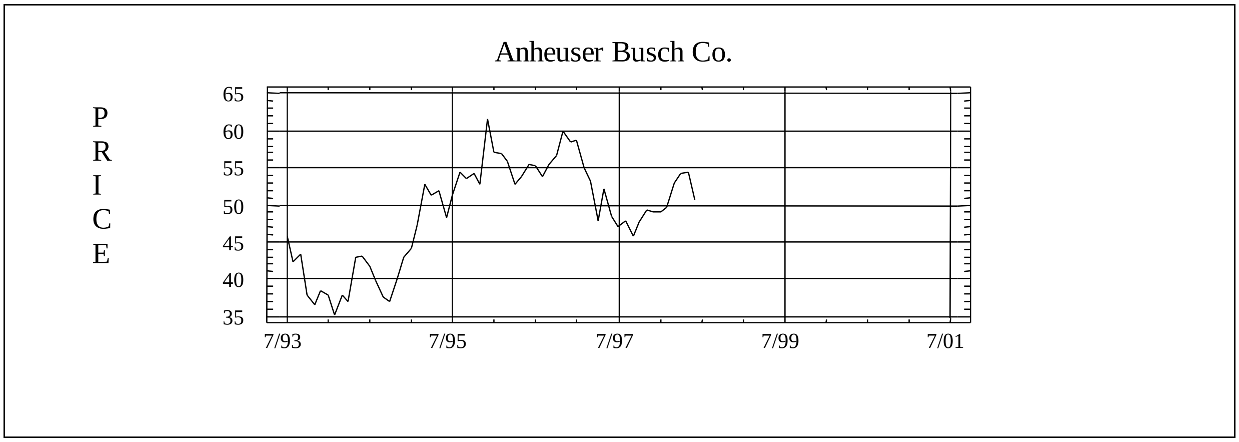
Tampering with a stable process will usually result in an increase in the variation, which will decrease the quality. Improving the quality of a stable process (i.e. decreasing common variation) is usually only accomplished by a structural change, which will identify some of the common causes, and eliminate them from the process.

Identification Tools

There are a number of tools used in practice to determine whether specific causes of variation exist within a process. In the remaining part of this chapter we will discuss how time series plots, the runs test, a test for normality and control charts are used to identify specific sources of variation. As will become evident there is a great deal of similarity between time series plots and control charts. In particular, the control charts are time series plots of statistics calculated from subgroups of observations, whereas when we speak of time series plots we are referring to plots of consecutive observations.

Time Series Plots

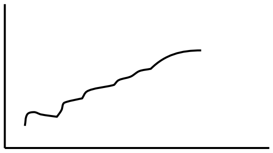
One of the first things one should do when analyzing a time series is to plot the data, since according to Confucius "A picture is worth a thousand words." ***A* time series plot *is a graph where the horizontal axis represents time and the vertical axis represents the units in which the variable of concern is measured.*** For example, consider the following series where the variable of concern is the price of Anheuser Busch Co. stock on the last trading day for each month from July 1993 to June 1998 inclusive (for data see ***STOCK03*** in StatGraphics file ***STOCK.SF***). Using the computer we are able to generate the following time series plot. Note that the horizontal axis represents time and the vertical axis represents the price of the stock, measured in dollars.



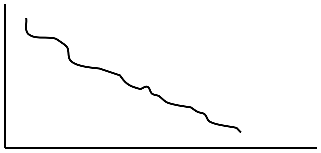
When using a time series plot to determine whether a process is stable, what on is seeking is the answer to the following questions:

1. Is the mean constant?
2. Is the variance constant?
3. Is the series random (i.e. no pattern)?

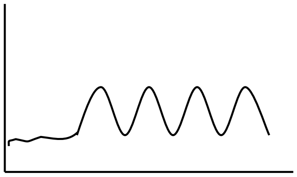
Rather than initially showing the reader time series plots of stable processes, we show examples of nonstable processes commonly experienced in practice.



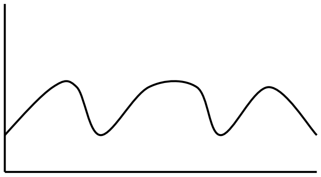
(a)



(b)



(c)



(d)

In figures (a) and (b) a change in mean is illustrated as in figure (a) there is an upward trend, while in figure (b) there is a downward trend. In figure (c) a change in variance (dispersion) is shown, while figure (d) demonstrates a cyclical pattern, which is typical of seasonal data. Naturally, combinations of these departures are examples of nonstable processes.

Runs Test

Frequently nonstable processes can be detected by visually examining their time series plots. However, there are times when patterns exist that are not easily detected. A tool that can be used to identify nonrandom data in these cases is the runs test. The logic behind this nonparametric test, is as follows:

> Between any two consecutive observations of a series the series either increases, decreases or stays the same. Defining a run as a sequence of exclusively positive or exclusively negative steps (not mixed) then one can count the number of observed runs for a series. For the given number of observations in the series, one can calculate the number of expected runs, assuming the series is random. If the number of observed runs is significantly different from the number of expected runs then one can conclude that there is enough evidence to suggest that the series is not random. Note that the runs test is a two tailed test, since there can be either too few of observed runs [once it goes up (down) it tends to continue going up (down)] or too many runs [oscillating pattern (up, down, up, down, up, down, etc..)]. To determine if the observed number significantly differs from the expected number, we encourage the reader to rely on statistical software (StatGraphics) and utilize the p-values that are generated.

Normal Distribution?

Another attribute of a stable process, which recalls lacks specific causes of variation, is that the series follows a normal distribution. To determine whether a variable follows a normal distribution one can examine the data via a graph, called a histogram, and/or utilize a test which incorporates a chi-square test statistic.

A histogram is a two dimensional graph in which one axis (usually the horizontal) represents the range of values the variable may assume and is divided into mutually exclusive classes (usually of equal length), while the other axis represents the observed frequencies for each of the individual classes. Recalling the attributes of a normal distribution

- symmetry
- bell shaped
- approximately 2/3 of the observations are within one (1) standard deviation of the mean
- approximately 95 percent of the observations are within two (2) standard deviations of the mean

one can visually check to see whether the data approximates a normal distribution. Many software packages, such as Statgraphics, will overlay the observed data with a theoretical distribution calculated from the sample mean and sample standard deviation in order to assist in the evaluation. Even so many individuals still find this evaluation difficult and hence prefer to rely on statistical test. The underlying logic of the statistical test for normality is that, like the visual inspection, the observed frequencies are compared with expected values which are a function of an assumed normal distribution with the sample mean and sample standard deviation serving as the parameters. The test statistic:

$$\chi^2 = \sum_{all\, i} \left[\frac{\left(\mathrm{O}_i - \mathrm{E}_i\right)^2}{\mathrm{E}_i} \right]$$

where:

> O_i is the number of observed observations in the ith class and E_i is the number of expected observations for the i^{th} class, follows a conditional chi-square distribution with n-1 degrees of freedom.

In particular, the null hypothesis that the series is normally distributed is rejected when the χ^2 values are too large (i.e. the observed values are not close enough to the expected). To aid in the calculations and determining what is too large, we encourage the reader to rely on the results generated by their statistical software package especially the p-values that are calculated.

Exercises

The data for these exercises are in the file ***HW.SF***. For each series determine if the series are stationary (i.e. constant mean and constant variance), normal and random. If any of the series violates any of the conditions (stationarity, normal and random); then, there is information and you only need to cite the violation.

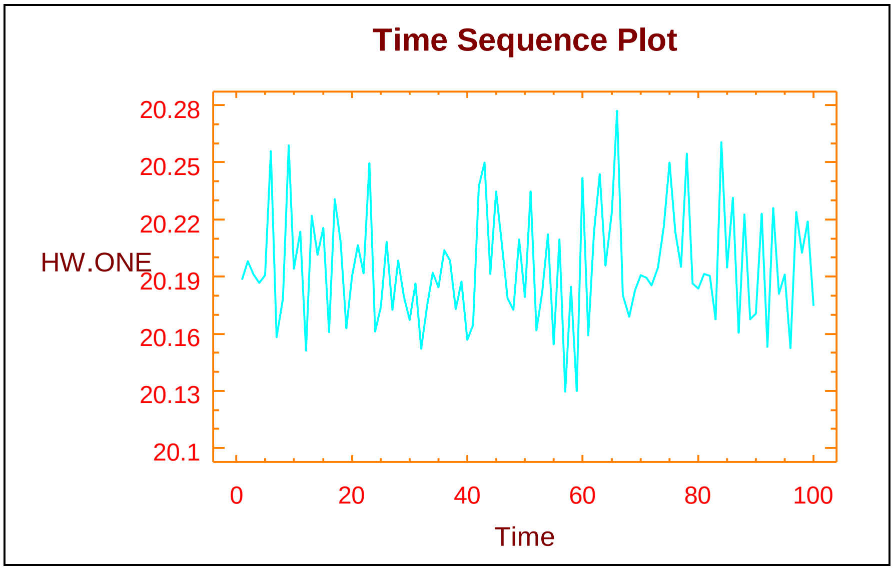
You are encouraged to examine each series ***before*** looking at the solution provided. The series are:

HW.ONE
HW.TWO
HW.THREE
HW.FOUR
HW.FIVE
HW.SIX

For each series, the time units selection is “index” since the series is not monthly, daily or workdays in particular.

1. HW.ONE

The time series (horizontal) plot shows:



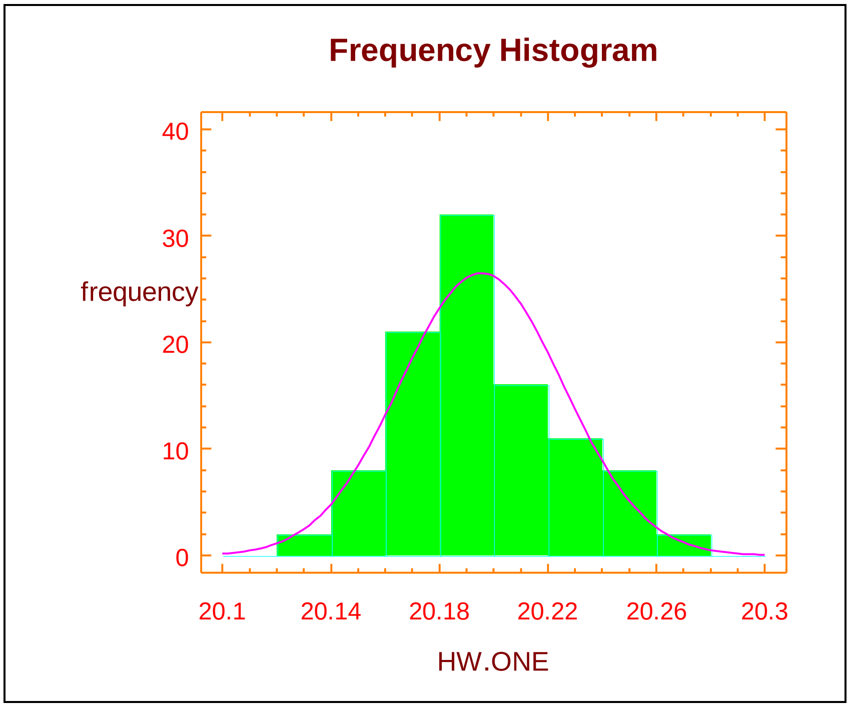
Time Series Plot: HW.ONE

Stationarity?

From the visual inspection, one can tell the series is stationary. This may not be obvious to you at this time; however, it will be with more experience. Remember, one way to determine if the series is stationary is to take snap shots of the series in different time increments, then impose them in different time intervals and see if they match up. If you do that with this series, you will indeed see that is in fact stationary.

Normality?

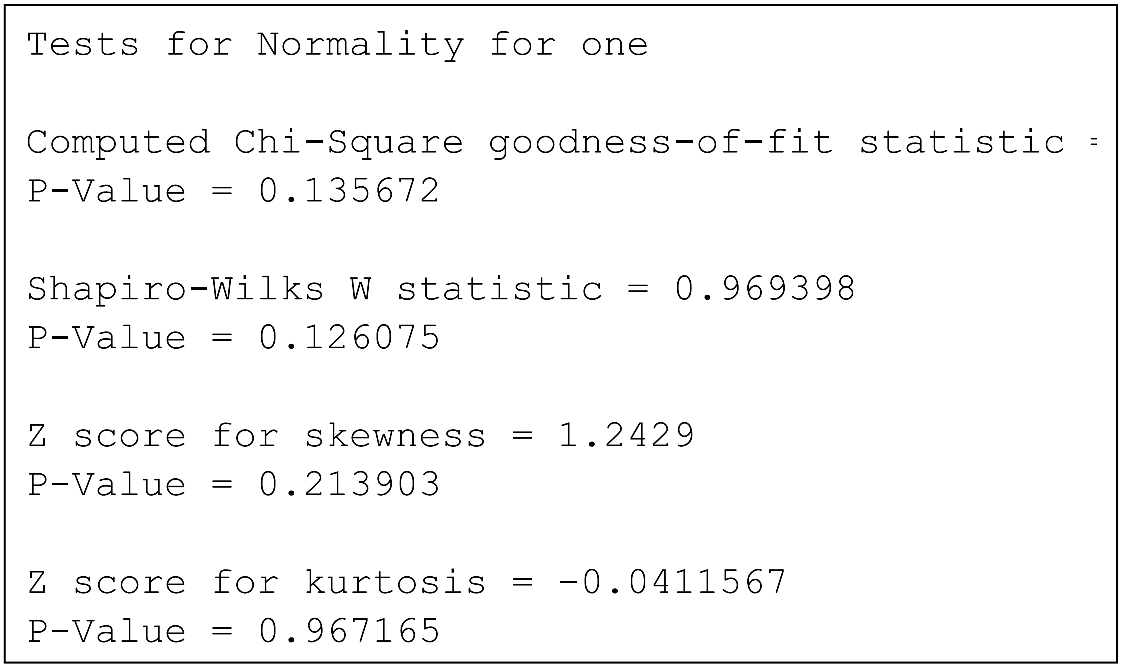
Shown below is the histogram that is generated by Statgraphics for the HW.ONE.[11]



Histogram: HW.ONE

Remember, a histogram shows the frequency with which the series occurs at different intervals along that horizontal axis. From this, one can see that the distribution of HW.ONE appears somewhat like a normal distribution. Not exactly, but in order to see how closely it does relate to theoretical normal distribution, we rely on the Chi-square test. For the Chi-square test, we go to the tabular options and select Tests for Normality. As a result, Chi-square statistic will then be calculated as displayed in the table shown on the following page.

[11] To obtain such a graph using Stat graphics, we selected the data file HW.SF and then selected Describe, Distribution Fitting, Uncensored Data, specify One (data series) , Graphical Options and Frequency Histogram.



```
Tests for Normality for one

Computed Chi-Square goodness-of-fit statistic =
P-Value = 0.135672

Shapiro-Wilks W statistic = 0.969398
P-Value = 0.126075

Z score for skewness = 1.2429
P-Value = 0.213903

Z score for kurtosis = -0.0411567
P-Value = 0.967165
```

Using the information displayed in the table, one is now able to perform the following hypothesis test:

H_0: the series is normal and

H_1: the series is *not* normal.

As we can see from the table, the p-value (significance level) equals 0.1356. Since the p-value is greater than alpha (0.05), we retain the null hypothesis, and hence we feel that there is enough evidence to say that the distribution is normally distributed. Thus, we are able to pass the series as being normally distributed at this time.

Random?

Relying upon the nonparametric test for randomness, we are now able to look at the series HW.ONE and determine if in fact we think the series is random. [12]

[12] This result was generated by the following Stat Graphics steps after selecting the data file HW.SF. Select Special, Time Series, Descriptive, specifying One as the data series, followed by Tabular Options and in particular selecting Test for Randomness.

```
Tests for Randomness of one

Runs above and below median
---------------------------
     Median = 20.1911
     Number of runs above and below median = 51
     Expected number of runs = 51.0
     Large sample test statistic z = -0.100509
     P-value = 1.08007

Runs up and down
---------------------------
     Number of runs up and down = 74
     Expected number of runs = 66.3333
     Large sample test statistic z = 1.71534
     P-value = 0.0862824
```

Test for Randomness

Recall again that one will reject the null hypothesis of the series is random with the alternate being the series is not random, if there are either too few **or** too many runs. Ignoring the information about the median, and just looking at what is said with regards to the number of runs of up and down, we note that for HW.ONE there are 74 runs. The expected number of runs is 66.3. We do not need to rely on a table in a book as we stated before, but again we can just look at the p-value, which in this case is 0.086 (rounded). So, since the p-value again is larger than our value of α = 0.05, we are able to conclude that we cannot reject the null hypothesis, and hence we conclude that the series may in fact be random.

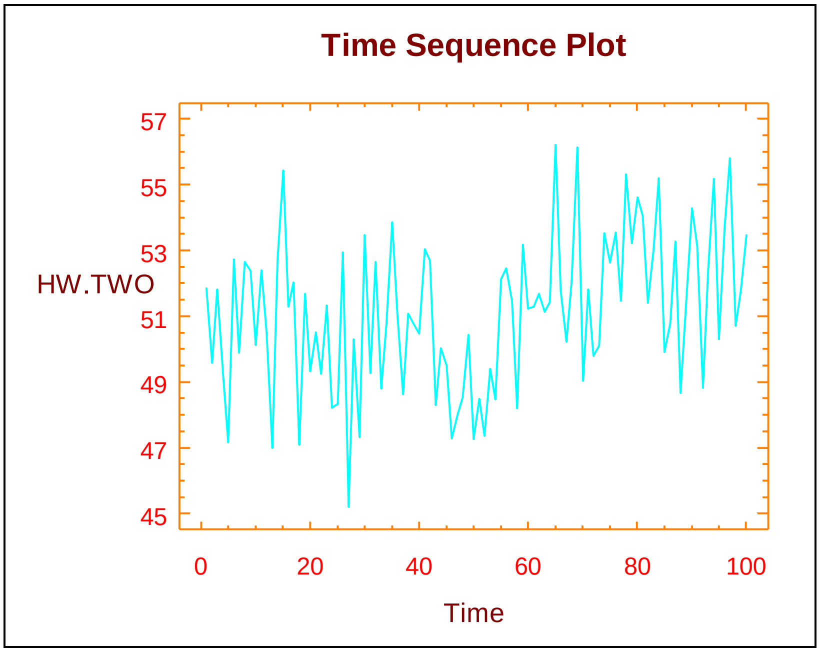
Summary

Having checked the series for stationarity, normality and randomness, and not having rejected any of those particular tests, we are therefore able to say that we do not feel there is any information in the series based upon these particular criteria.

2. HW.TWO

Stationarity?

As one can see in the horizontal time series plot shown below, HW.TWO is ***not*** stationary.



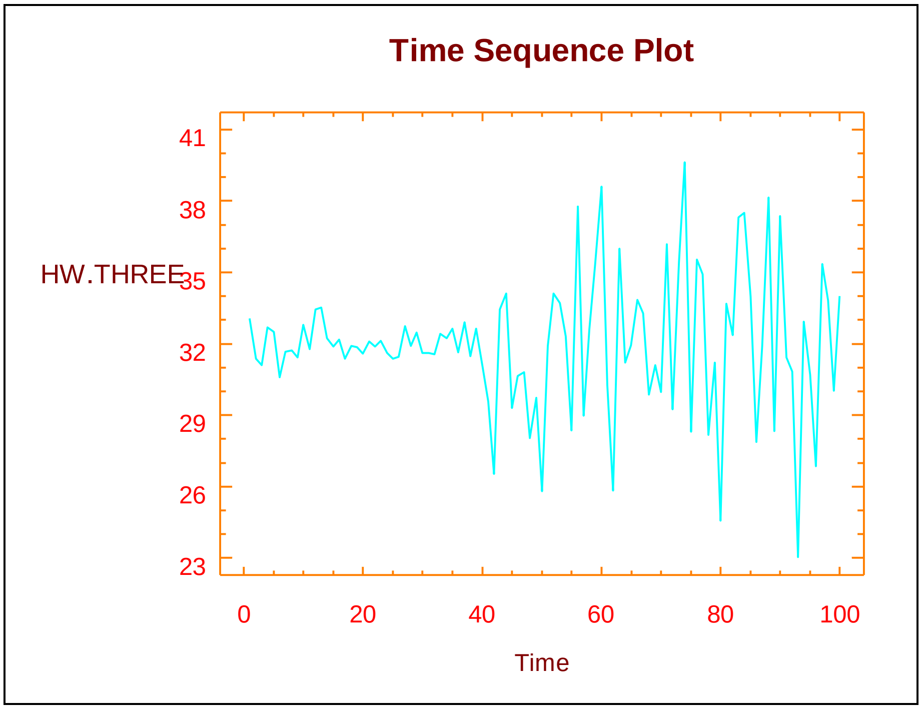
Time Series Plot: HW.TWO

In particular, if you notice at around 60, there is a shift in the series, so that the mean increases. Hence, for this process, there is information to look at the series because there is a shift in the mean. Given that piece of information, we will not go to the remaining steps checking for normality and also for randomness. If it is difficult for you to see the shift of the mean, take a snap shot for the series from say 0 to 20, and impose that on the values from 60 to 80, and you will see that there is in fact a difference in the mean itself.

3. HW.THREE

Stationarity?

The initial step of our process is once again to take a look at the visual plot of the data itself. As one can see from the plot shown on the following page, there is a change in variance after the 40th time period.



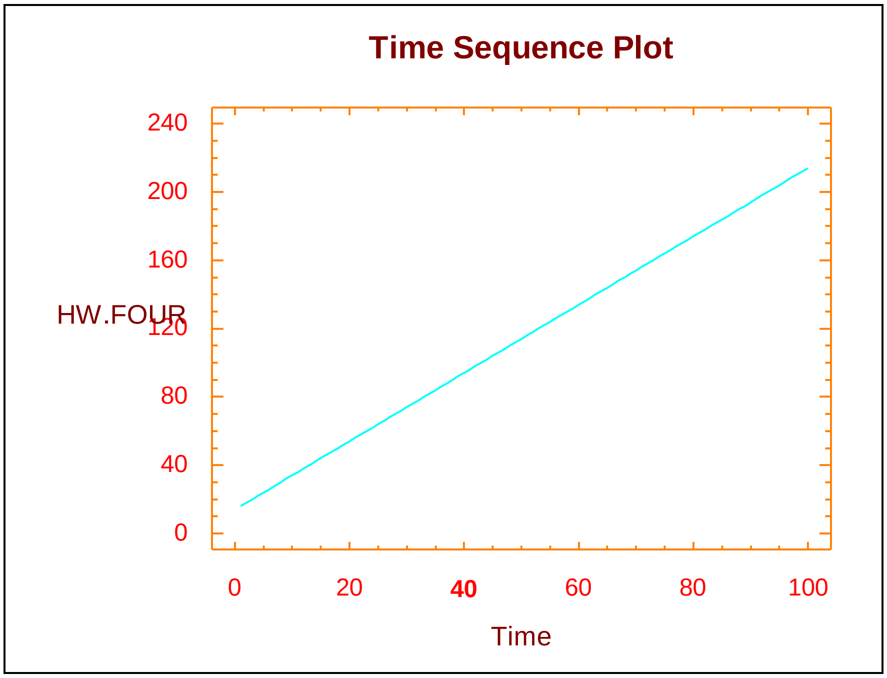
Time Series Plot: HW.THREE

In particular, the variance increases substantially when compared to the variance in the first 40 time periods. This is the source of information and once again we will not consider the test for normality or the runs test. We have acquired information about the change of variance.

If you were the manager of a manufacturing process and saw this type of the plot, you would be particularly concerned about the increase in the variability at the 40th time period. Some kind of intervention took place and one should be able to determine what caused that particular shift of variance.

4. HW.FOUR

The time series plot of HW.FOUR is appears below:



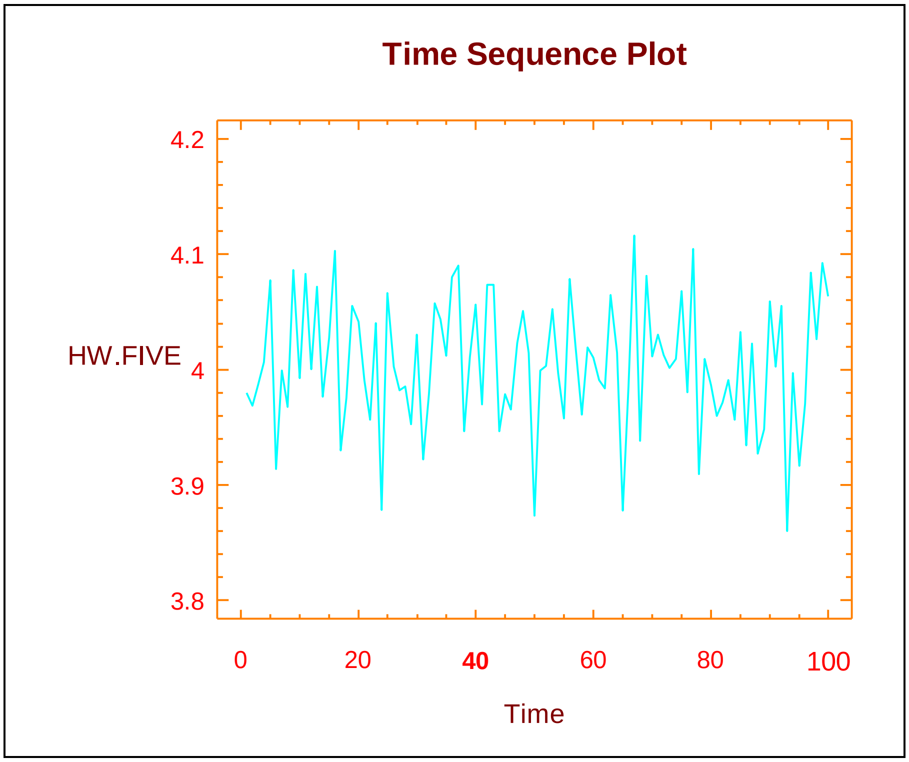
Time Series Plot: HW.FOUR

Stationarity?

As one can clearly see from this plot, the values are linear in that the values fall on a straight line. This series is clearly not stationary. Once again, if one were to take a snap shot of the values say between 0 and 40, and just shift that over so they match up between 60 and 100, you have two separate lines clearly the means are not the same. The mean is changing over time. (We will discuss this kind of series when we are applying regression analysis techniques.)

5. HW.FIVE

Shown below is the time series plot of the HW.FIVE:



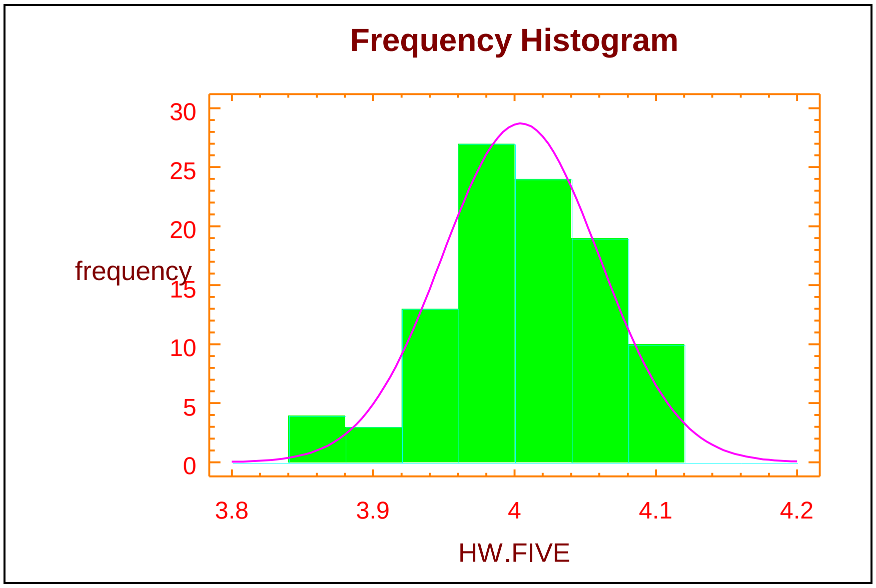
Time Series Plot: HW.FIVE

Stationarity?

The series is clearly stationary. It has a constant mean and a constant variance as we move in time. Once again, recall that one can take a snap shot of the series between a couple time periods say the 0 and 20, and that will look very similar to any other increments of 20 time periods that shown on the time series plot of the series. We now think that the series may in fact be stationary. Recall we also want to check for normality and the runs test. Hence, we now perform these two tests.

Normality?

Once again, utilizing Statgraphics options, we are able fit the series to a normal distribution. A theoretical distribution is generated using the sample mean and standard deviation as the parameters. Using those values, we can compare the frequency of our actual observations with the theoretical normal distribution. Selecting the default options provided by Statgraphics, the following figure is displayed:



Histogram: HW.FIVE

Note again that the distribution is not exactly normally distributed, but it may closely follow a normal distribution. To have an actual test, we revert back to the Chi-square test and again using the default options provided by StatGraphics, we come up with the output shown on the following page.

```
Tests for Normality for five

Computed Chi-Square goodness-of-fit statistic = 14.72
P-Value = 0.836735

Shapiro-Wilks W statistic = 0.970928
P-Value = 0.161365

Z score for skewness = 0.769262
P-Value = 0.441736

Z score for kurtosis = -0.291363
P-Value = 0.77077
```

As one can see from the information provided above, the significance level is 0.836735. Since this value is greater than 0.05, we are not able to reject the null hypothesis that the series is normal, and hence we feel the series may in fact be approximately normally distributed. We now test the series for randomness.

Random?

Again, we use the nonparametric test for randomness and what we are able to determine for this particular series is based upon information shown below:

```
Tests for Randomness of five

Runs above and below median
---------------------------
     Median = 4.00253
     Number of runs above and below median = 54
     Expected number of runs = 51.0
     Large sample test statistic z = 0.502545
     P-value = 0.615281

Runs up and down
---------------------------
     Number of runs up and down = 71
     Expected number of runs = 66.3333
     Large sample test statistic z = 0.997291
     P-value = 0.318622
```

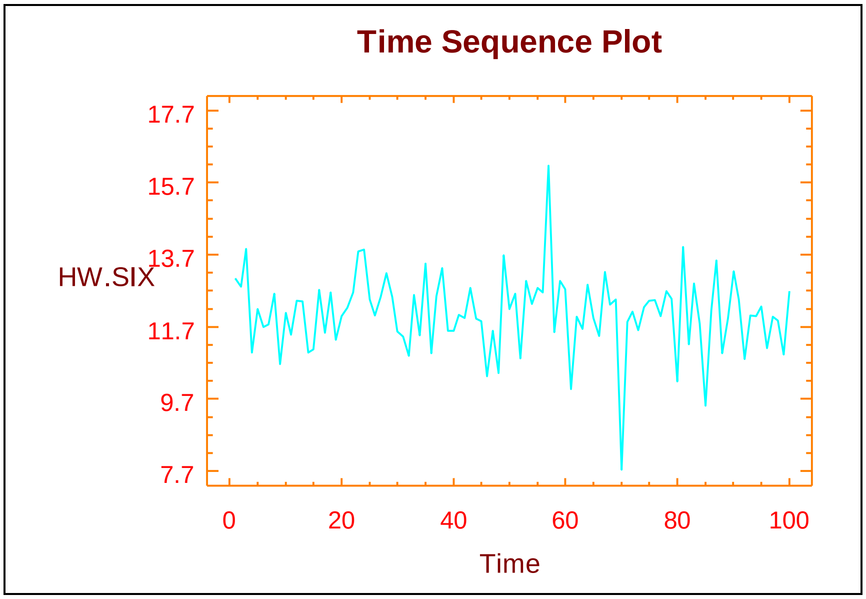
Ignoring the information about the median, we focus our attention on the area discussing the actual number of runs up and down. Noted that the actual number is 71 and the expected number is 66.3. Is that discrepancy large enough for us to conclude that there are too many runs in the series, and hence possibly a pattern? To answer that question, we rely on the z-value, which is 0.997291, and the following information, which provides the p-value, which is 0.319. Since the p-value exceeds $\alpha = 0.05$, we are *not able to reject* (or, retain) the null hypothesis that the series is random; thus we **retain** (or, **fail to reject**) the null hypothesis.

Summary

As with HW.ONE, the series we just looked at, HW.FIVE, by visual inspection is stationary, and can pass for a normal distribution, and can pass for random series. Hence, based upon these criteria again, we are not able to find any information in this particular series.

6. HW.SIX

Stationarity? Again, the first step of our investigation is to take a look at the time series plot

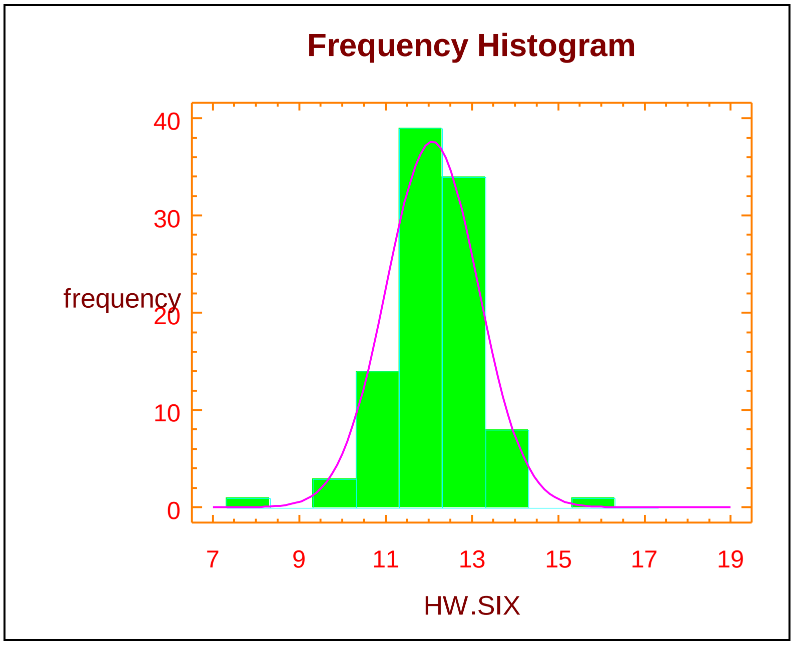


Time Series Plot: HW.SIX

From this time series plot, there are two values that stand out. We call those values outliers. They occur at approximately the 58th observation and about the 70th observation. Besides these two observations, which may have important information of themselves, the rest of the series appears to be stationary.

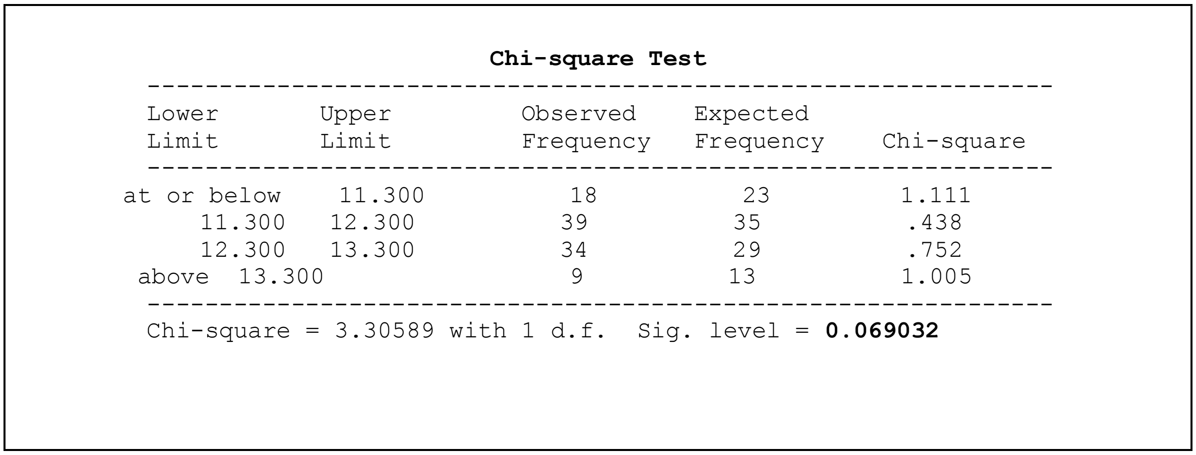
Normality?

Using the sample statistics of the mean equaling 12.0819 and standard deviation equals to 1.06121, we now compare our actual observations with the theoretical normal distribution. As one can see from the histogram displayed below, the two outliers appear on the extreme points, but the rest of the series are very closely approximate in normal distribution.



Histogram: HW.SIX

Going to the Chi-square test, and again accepting the option provided by Statgraphics, we note the following result on the next page:



```
                         Chi-square Test
   ----------------------------------------------------------------
   Lower        Upper          Observed     Expected
   Limit        Limit          Frequency    Frequency     Chi-square
   ----------------------------------------------------------------
 at or below     11.300           18           23           1.111
       11.300    12.300           39           35            .438
       12.300    13.300           34           29            .752
  above   13.300                   9           13           1.005
   ----------------------------------------------------------------
   Chi-square = 3.30589 with 1 d.f.  Sig. level = 0.069032
```

We noted that the p-value is 0.069032, again since the value is greater than α = 0.05, we conclude that the series may in fact pass for a normal distribution. The Chi-square statistic is sensitive to outlier observations as the outlier observations tend to inflate the statistics itself. Given that fact and the significance level we have obtained, we should feel comfortable that the series may in fact fall in normal distribution. {*The inflation of the Chi-square statistics would tend to make the significance level much smaller than what it would be without the outlier observations.*} Hence, we are unable to reject the null hypothesis that the series is normally distributed.

Randomness?

To determine whether the series can pass the randomness, we once again utilize the nonparametric runs test. As one can see the information provided on the next page (abbreviated format):

```
Tests for Randomness of six

Runs above and below median
---------------------------
     Median = 12.101
     Number of runs above and below median = 51
     Expected number of runs = 51.0
     Large sample test statistic z = -0.100509
     P-value = 1.08007

Runs up and down
---------------------------
     Number of runs up and down = 72
     Expected number of runs = 66.3333
     Large sample test statistic z = 1.23664
     P-value = 0.21622
```

The actual number of runs up and down is 72 verses the expected number of 66.3333. The question we need to ask now is “Is the difference significant, which would imply that we have too many runs verses the theoretical distribution?” As one can see from the p-value, which is 0.21662, we will not reject the null hypothesis that the series is random because the p-value again exceeds our stated value of alpha. Thus, we are able to conclude that we feel the series may in fact be random.

Summary

We have observed from HW.SIX that the series may in fact be stationary, normally distributed and random. We are possibly concerned about this measure with the two outlier observations numbers 58 and 70. As a manager, one will naturally want to ask a question what happen at those time periods, and see if there is information. Note, without the visual plot, we will never expect the series to have information based solely upon the normality test and the runs test. Thus, one can see that the visual plot of the data is extremely important if we are to determine information in the series itself. Of all the tests we had looked at, the ***visual plot*** is probably our most important one and one that we should always do whenever looking at a set of data.

The Concept of Stock Beta

An important application of simple linear regression, from business, is used to calculate the β of a stock. The β‘s are a measure of risk and are used by portfolio managers when selecting stocks. The model used (specified) to calculate a stock β is as follows:

$$R_{j,t} = \alpha + \beta R_{m,t} + \varepsilon_t$$

Where: $R_{j,t}$ is the rate of return for the j^{th} stock in time period $_t$
$R_{m,t}$ is the market rate of return in time period $_t$
ε_t is the error term in time period $_t$
α and β are constants

A formula for $R_{j,t}$ (the rate of return for the j^{th} stock in time period $_t$) follows:

$$R_{j,t} = ((P + D)_t - (P + D)_{t-1}) / (P + D)_{t-1}$$

Where: P is the price of the stock
D is the stock’s dividend
$(P + D)_t$ is the sum of the price and dividend for given time period
$(P + D)_{t-1}$ is the sum of the price and dividend for a previous time period (in this case, one time period prior to the “given” time period)

Generally, the average stock moves up and down with the general market as measured by some accepted index such as S&P 500 or the New York Stock Exchange (NYSE) Index. By definition, a stock has a beta of one (1.0). As the market moves up or down by one percentage point, stock will also tend to move up or down by one percentage point. A portfolio of these stocks will also move up or down with the broad market averages.

If a stock has a beta of 0.5, the stock is considered to be one-half as volatile as the market. It is one-half as risky as a portfolio with a beta of one. Likewise, a stock with a beta of two (2.0) is considered to be twice as volatile as an average stock. Such a portfolio will be twice as risky as an average portfolio.

Betas are calculated and published by *Value Line* and numerous other organizations. The beta coefficients shown in the table below appeared in *Value Line*, April 19, 1996. Most stocks have beta in the range of 0.75 to 1.50, with the average for all stocks is a beta of 1.0. Which stock is the most stable? Which stock is the most risky? Is it possible for a stock to have a negative beta (consider *gold stocks*)? If so, what industry might it represent?

Stock	Beta
Harley-Davidson	1.65
Seagate Technology	1.65
Dow Chemical	1.15
General Electric	1.15
Proctor & Gamble	1.05
Sara Lee	1.05
Chevron	0.75
Pacific Gas & Electric	0.80
Homestake Mining	0.40

In summary, the regression coefficient, β (the beta coefficient), is a market sensitivity index; it measures the relative volatility of a given stock versus the average stock, or “the market.” The tendency of an individual stock to move with the market constitutes a risk because the stock market does fluctuate daily. Even well diversified portfolios are exposed to market risk.

As asked in your text, given the β‘s for Anheuser Busch Company (0.579), The Boeing Company (0.904), and American Express (1.080), what are the implications for the three companies?

[Note: If the concept of stock risk is of special interest, please refer to any intermediate financial management text for a more in-depth explanation. The concept is critically important to financial management.]