IENG/MANE 332 lecture notes

Reference: PRODUCTION, Planning, Control, and Integration by SIPPER & BULFIN

Chapter 4- Part 1: FORECASTING

1. Introduction

"Twenty thousand tires? We'll never sell that many tires in June. The marketing people are crazy!" carol exclaimed. Obviously, she did not agree with marketing's prediction of tire sales in June. How could you determine a better estimate?

There are several ways you could get an answer to this question. One is to simply guess. You could ask Pete, the modeling room foreman. He has been here for 25 years, and his experience should give him a good feel for how many tires will be sold. Pete points out that automobile sale are expected to peak in august, so maybe the number is not so far off after all. Being a good problem solver, you might examine demand for previous months and try to estimate the demand for June.

Of course none of these methods are guaranteed to give good results. Your guess may be far off, and even Pete's experience may not be enough to get a good prediction. Car sales do not affect tire sales-every new car needs tires- but maybe not in a straightforward way. Even sophisticated models may not be always give accurate predictions, if you think otherwise, consider weather forecasting.

Determining what will happen in the future in order to make good decisions is a problem that must be faced quite often. This fact is true not only in our personal lives, but also in the business world. We use the term **forecast** to mean some definite method – rather than just a guess – of predicting future events.

In today's market-driven production system, forecasts are more important than ever. The rewards of good forecasts as well as the penalties for bad forecasts can be quite high. With the proliferation of personal computer packages, forecasting is easier and cheaper than ever. However, managers must beware of using canned packages without understanding their underlying principle. After all, the program will give an answer, even it is bad. In this chapter we present a range of forecasting techniques, an idea of the situations in which to use them, and their underlying principles. This chapter also addresses how to monitor the forecasts and adjust both the forecasts and adjust both the forecast and forecasting techniques.

We will discuss three classes of forecasting methods. The first class is composed of judgmental or qualitative methods. In their simplest form, they use "expert" opinion to get a forecast. Asking Pete is a simple example of this approach. The

second class, casual methods, tries to relate the variable being forecast to something else. Relating automobile production to tire sales is one example. Time series methods use the past to determine the future and are based on statistical principles. Studying past tire sales to get a forecast of future tire sales could be done using a time series approach. Finally, an overall view of forecasting systems, control, methods, and practice is presented. First, we view forecasting from a systems perspective.

2. The Forecasting System

As we noted in chapter 3, problem solving involves a series of steps, when "standard" problems occur, we can offer simplify the procedure. Forecasting has many standard models. We will discuss the stages of problem solving as they relate to forecasting in this section. The remainder of the chapter will elaborate on particular forecasting methodology.

2.1. Identify the Problem

Forecasts provide information to make better decisions. The first step is to identify the decision. If the decision will not be affected by the forecast, a forecast is unnecessary. The importance of the decision will dictate the effort that goes into producing a forecast. A one-time decision requires one forecast, whereas a recurring decision requires a forecast each time the decision is made. In either case, the decision will determine what to forecast, the level of detail needed, and how often the forecast will be made. Forecasts of sales, quality of raw materials, income, expenses, energy usage, or times customers arrive are commonly needed in business operations.

Suppose the decision is how many televisions to produce next year. This decision is important because it directly affects employment and raw material levels, marketing (advertising), distribution, and warehousing. The demand for the product itself will be forecast; we may not care about the particular variations of the product. Because most plants operate on a monthly or four-week plan, we should forecast monthly demand. If some planning is done on a quarterly basis, monthly forecasts can be combined. On the other hand, a shorter-range forecast may need individual variations of the product, for example 32", 42", and 50" televisions.

The decision maker is the problem owner. The problem solver is the analyst or forecaster. Most forecasts are prepared by teams that include management, marketing, the forecaster, and possibly data processing. Problem identification determines the problem mission or purpose, which is shown as forecast need in figure 4-1 and begins the design of a forecasting system.

2.2. Understand the Problem

The key to understanding forecasting problems is to understand the underlying process – for example, the process that creates demand for an item. We can never fully understand the process, so we can only hope to gain insight and

make necessary assumptions to create forecasts. To do so, we examine characteristics of the problem and analyze data, if they exist. We also state a forecast goal.

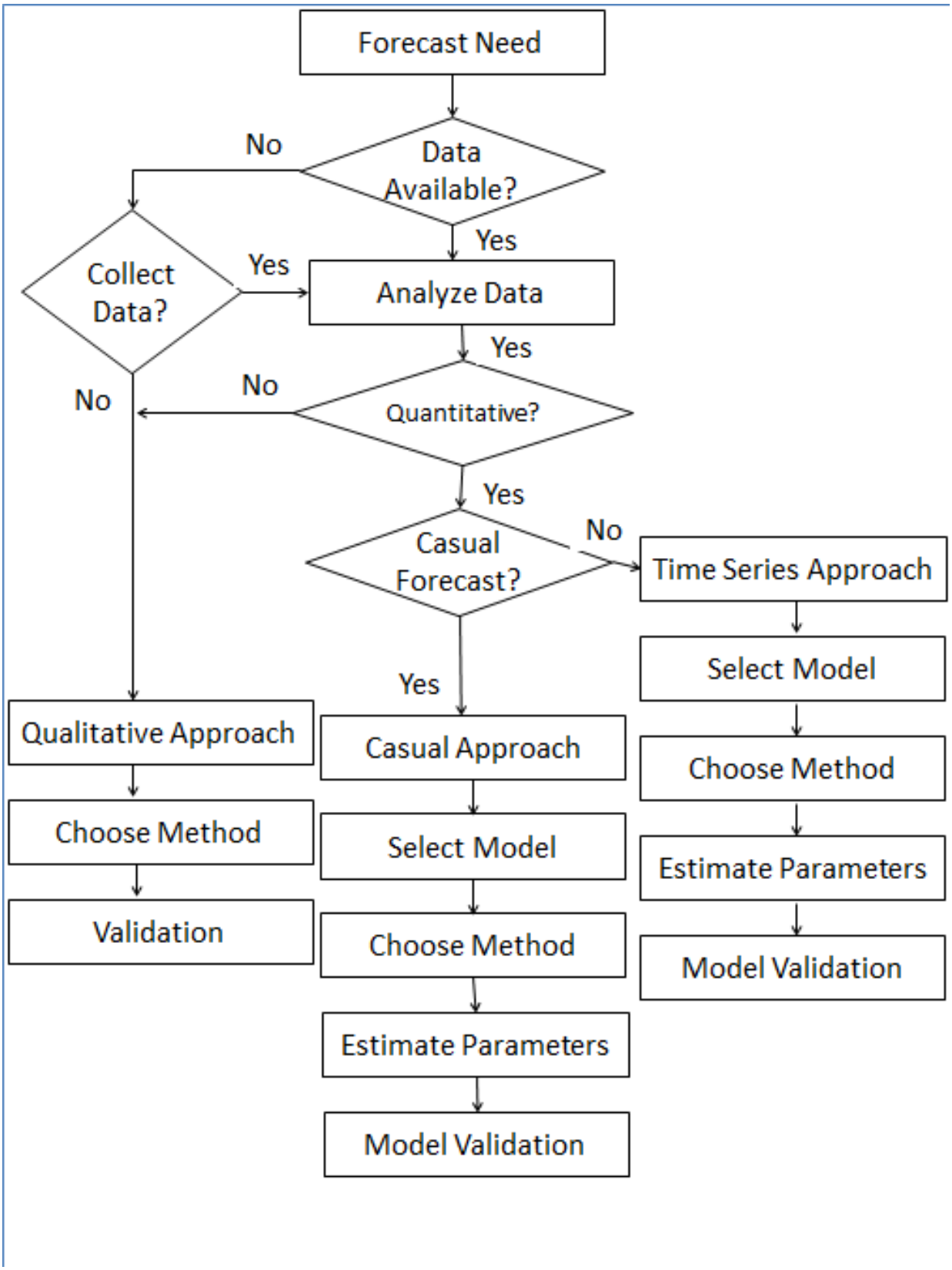


Figure 4.1. Designing a forecasting system

2.2.1. Problem Characteristics

The main characteristics of a forecasting problem are time frame, level of detail, accuracy needed, and number of items to forecast. We give examples of these by time frame.

In production systems, we are most often interested in forecasting demand for our product or services in order to decide how much to produce. Long-term decisions, such as opening new plants or adding capacity to existing plants, often depend on a forecast of demand. Here, we are not so much interested in individual products, but in overall volume. Therefore, we might use dollars as an aggregate measure of sales. A typical time frame for these types of these types of decisions would be three to five years. Long-term decisions do not require exact forecasts; the decision to build a new plant will be based on the trend of forecasts for several successive years rather than a single demand estimate. Thus very forecasts are unnecessary- a number in the ballpark will do. Typically, long-term forecasts are made for only one item. Causal and qualitative methods are often used for these forecasts.

A mid-range decision might be to allocate plant capacity to groups of products. Again, we might not need to know the demand for each individual item, but rather for groups of items that share certain production facilities. An example would be a monthly forecast of tires made in a plant; individual sizes are not important for determining gross capacity. Typical measures might be units, production hours, gallons, or pounds of an aggregate product. The time frame for these decisions is from three months to one or two years, and greater accuracy is required. Mid-range decisions typically require forecasts for only a few items. Quantitative methods, including causal and time series, are frequently used for mid-range forecasts.

The most common Short-term decision could be how much we should produce. In this case, the actual number of units of the product is needed. This decision could be weekly, monthly, or possibly quarterly. Because short-term decisions are based on these forecasts, they need to be reasonably accurate. Time series methods are most often used for short-term forecasting, but in some situations, causal and qualitative methods are also useful. Short-term decisions may require forecasts of hundreds or thousands of items.

2.2.2. Data

Examining data, when they exist, can provide more insight. Data may come from company records or commercial or government sources. Company records includes sales and purchase information. Commercial services have access to such things as database and surveys and can provide raw data or reports on specific topics; one example is *A Graphic Guide to Consumer Markets,* published yearly by the National Industrial Conference Board. The government also provides many types of data, such as population and demographic

information, and the Department of *Commerce Publishes Survey of Business* monthly.

Make sure, however, that the data reflects the true situation; for example, a record of actual sales may not be include customers who wanted to buy the item but were unable because it was not available.

If no data exist, we must collect them or use a forecasting approach that does not use data. If data are unavailable or too costly to collect, we choose a qualitative approach. Qualitative methods, the left branch of Figure 4.1, are discussed in section 3.

Data are affected by factors that are either external or internal. External factors are beyond our control, but we can influence internal factors.

A good example of an external factor is the economy. If the economy experiences a downturn, demand for goods and services usually declines too. A variety of economic indicators have been defined that may help us understand demand behavior. Other external factors include competitors' actions, complementary products, and customer choices.

Internal factors include the product quality and price, delivery time, advertising, and rebates. If more advertising is done, demand will likely increase. Rebates are also used to increase demand. Poor quality, long waits for the item, or a high price will usually reduce demand.

Data should be analyzed to see if causal factors exist. A causal factor is something that influences the data in a known way and can be helpful in forecasting. Demand data for tires provide an example of a casual factor. If most tires are sold to an automobile manufacturer, knowing the number of cars to be produced will indicate the demand for tires; that is, the production of cars causes a demand for tires. Plotting tires sold versus cars produced will give an indication of the validity of this assumption. Of course tires are also sold to consumers as replacements on older cars. Casual is shown in the middle path in Figure 4-1; detailed discussion is given in section 4. Selecting a model for a casual approach is similar to selecting a model for the time series approach, so we will combine the discussion. Time series approaches are discussed in section 5.

If data are available, we plot them to see if there is a pattern. Figure 4.2. shows weekly demand for a toothpaste for the last two years. We use this data to explain time series data analysis. Causal data analysis is similar, but rather than a plot, say, demand versus time, we would plot demand versus causal variable. When the plot is examined, it appears to be roughly level with some variations, which is typical type of a **constant process**. Because the population is relatively stable, at least in the short run, it seems reasonable that toothpaste sales would also be approximately constant. The weekly variation is caused by a **random** or **noise** component that we cannot control. For a constant underlying process, the noise component should have a mean of zero; if not, it is not noise, but part of underlying process. Similarly, the likelihood of observing a value above the

constant component should be the same as a value below it. If the variance changes over time, then our assumption of a constant process is not valid, so we will assume that the variance of the noise is constant. Thus it is reasonable to assume that the noise follows a symmetric probability distribution with mean zero and variance σ_ε^2. For now we assume the noise component is normally distributed; we see in section 7.2 that this assumption is robust. Similar arguments can be made regarding the distribution of noise for underlying processes that are not constant.

We should have some reason to assume a process is constant. Demand for many items would seem to follow a constant process; toothpaste, milk, bread, and socks are mature items and are used regularly. Constant processes may be useful even for products that are not always used regularly. Over a short horizon, many things are approximately constant. In the mature stage of a product life cycle many products exhibit stable sales. Also, models for constant processes are good introductions to more complicated models.

Some things are, by nature, not constant. During a product life cycle there is a growth stage in which sales increase. Similarly, there is a decline or phase-out stage when sales are decreasing. Assuming a constant process in either case can be disastrous. These processes are examples of a **trend process**. The rapid growth of personal computers and related equipment is a good example. The upper line of figure 4-3 is an example of increasing trend process.

The line connecting points have no meaning, but we added to emphasize the pattern in the data. This growth appears to be linear. As in the constant process, the line is not smooth but has many little jumps in it, which are caused by the random component. Again, just because the line appears to go up is not reason enough to assume a trend process; there should be some way to explain it. Trends could also be nonlinear, but for simplicity, we will restrict our discussion to linear trends.

Also plotted in figure 4.3 is a **seasonal process**. Every four months, the pattern seems to repeat, but random fluctuations are still present. An example of this type of process would be passenger miles for an airline. The term seasonal is used because the weather is often an underlying cause; ice cream and soft drinks are more popular in summer than in winter.

Sometimes cyclical processes are defined, a typical one being the number of telephone calls during a day; they peak at mid-morning and mid-afternoon. However, we will make no distinction between seasonal and cyclical. Again, there should be an underlying justification for assuming a seasonal process.

When data are plotted, the choice of scale is very important. If the wrong scale is chosen, data from a constant process may look seasonal due to the random fluctuations. When trend and seasonality are present, the data must be decomposed to see the effects of each. Also, outliers should be removed before analyzing the data. An example would be sales that were extremely high or low due to an unusual occurrence such as a strike or earthquake. Retailers often

remove special seasons, e.g., Christmas, from the time series data and handle them as exceptions.
The result of data analysis is to understand the process that causes demand. There will always be some part that is unexplainable – the random component. However, the model we sue will be direct result of the process we assume.

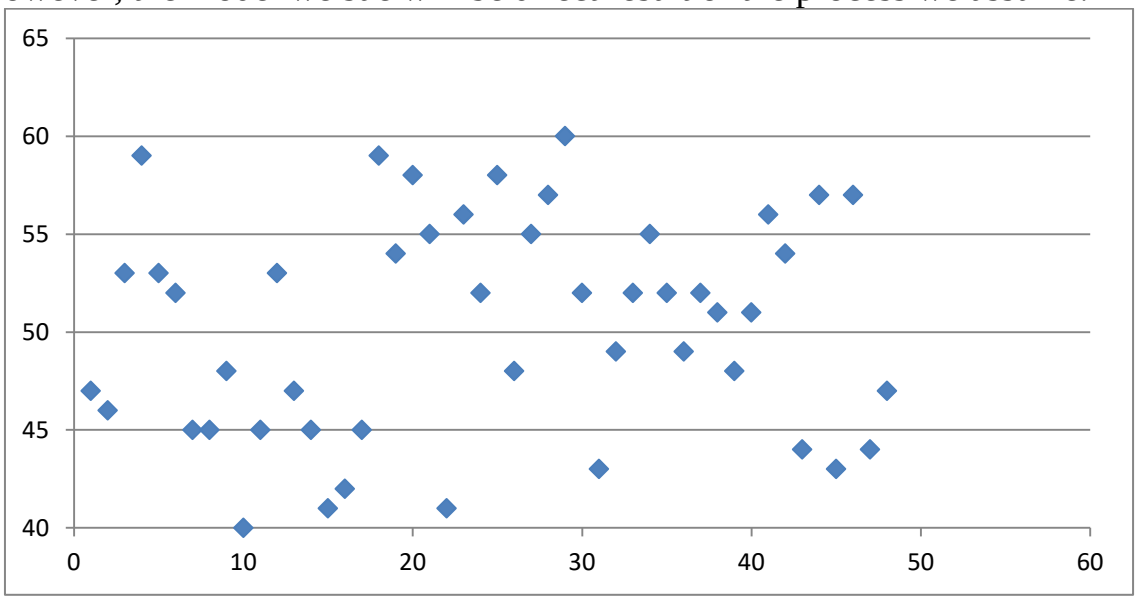


Figure 4.2. A constant process

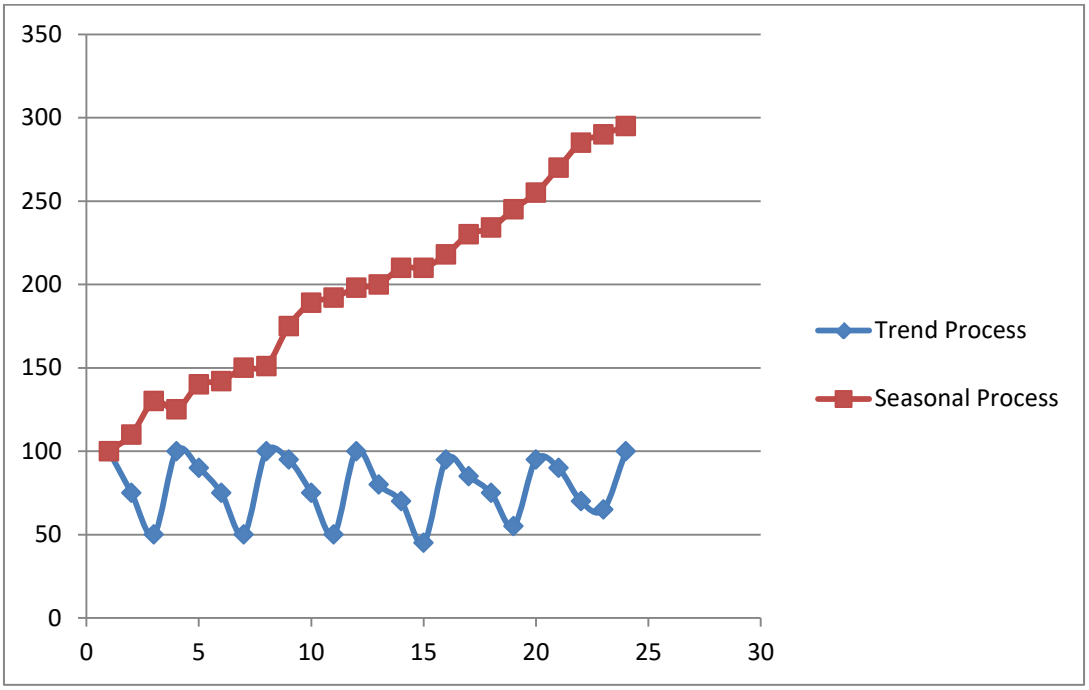


Figure 4.3. Typical demand patterns

2.2.3. FORECAST GOAL

The goal of any forecasting system is to provide forecasts of the appropriate accuracy in a timely manner and at a reasonable cost. A timely forecast is determined by its use. The basic trade off in forecasting is between response to a

change and stability; that is, if we experience a demand that is abnormally high this week, we have to decide if we need to make more products next week. If the high demand reflected a change in the demand pattern, we should increase production, but if it was just a random fluctuation, we should not. A good forecasting system will react to actual changes but ignore chance variations.

2.3. Develop a Model

Once processes are identified, they determine the form of the model. Qualitative forecasting does not use easily stated models. Causal models depend on the particular situation but generally have the form;

$$d_t = f(x_{t-k}) + \varepsilon_t$$

Where d_t represents the dependent variable, e.g., demand, x_t the independent variable (or causal factor), and ε_t the noise component at time t. The independent variable at time t is ideally a function of independent variable at time $t - k, k \geq 1$. The k-period time lag allows us to know the value of the independent variable before we need to forecast the dependent variable; if there is no time lag, we must forecast the independent variable to get a forecast of the dependent variable. The functional relationship between d and x is represented by f and could be linear, quadratic, or some other mathematical relationship. There could be more than one casual factor.

For time series approaches, the common models discussed are constant, linear trend, and seasonal, or combinations of these.

$$d_t = a + \varepsilon_t \qquad \text{(Constant)}$$

$$d_t = a + bt + \varepsilon_t \qquad \text{(Linear trend)}$$

$$d_t = ac_t + \varepsilon_t \qquad \text{(Seasonal)}$$

Where a represents the constant portion, b the trend, c_t the seasonal factor for period t, and ε_t the random or noise component.

Models should be as simple as possible. In forecasting, try to use as few components in model as possible. These are the most common models, but others are used.

Recall from chapter 3 that models should be as simple as possible. In forecasting, try to use as few components in a model as possible. A complicated function may “fit” the data but also tends to obscure important relationships. If there are many components, the effect of each becomes small and may be indistinguishable from the noise. As an example, consider demand for CD players. Sales most of the year is relatively constant, but gifts during December cause sales to peak. Rather than using a complicated model to capture the December sales peak, use one simple model for all of the year except December and another model for December.

2.4. Solve the Model

The first step in solving the model is to choose a method. If we have a causal model, the method will be regression. For time series models, there may be several methods available, even for the same process. For example there are many methods to forecast a constant time series.

Given the model, if we know the coefficients, we could simply plug the right numbers to get a forecast. Because we do not know the actual parameters of the model equation, we must estimate them. The method used determines how they are estimated; usually they are estimated to minimize the difference between the forecast and the actual value over some set of historical data. Once parameters are estimated, applying the model to the appropriate numbers provides a forecast.

2.5. Interpret and Implement the Solution

Interpreting the solution is the major task of operating the forecasting system. Figure 4-4 shows the steps involved. As new data become available, we update the forecast. Also, we can compare the previous forecast to what actually happened to get feedback on the quality of the forecasting procedure. If the quality is acceptable, we say that the procedure is in control. If the procedure is out of control, we need to return to the design phase; either we need to re-estimate the parameters of the current model or change the model itself. If the forecasting system is in control, we forecast for a future period. This forecast is examined by a manager (owner) and judgment is used to accept, modify, or reject the forecast.

It is important to incorporate judgment into the system, especially when statistical methods are used. These methods have the underlying assumption that we are dealing in a stable environment, which is not always the case. For example, if a labor strike is forthcoming, the forecast should be modified to reflect this fact. It is important that any modification be done by the person making the decision and that is done within the system. If no provision is made for modifications, either the forecast will not be used or at some point it will mislead the decision maker.

If a demand forecast is low, more units than expected are demanded and shortages, called stock outs, occur. On the other hand, a forecast higher than actual demand, which results in too many units, creates inventory that is marked down or scrapped. Fisher et al, (1994) point out that stock out plus markdown costs incurred for a particular product often exceed the manufacturing cost of the product. If the cost of an excess unit (inventory) is not equal to the cost of a unit short, then the decision maker will likely adjust the forecast accordingly-making the forecast larger if the shortage cost is higher or decreasing it if the inventory cost is higher.

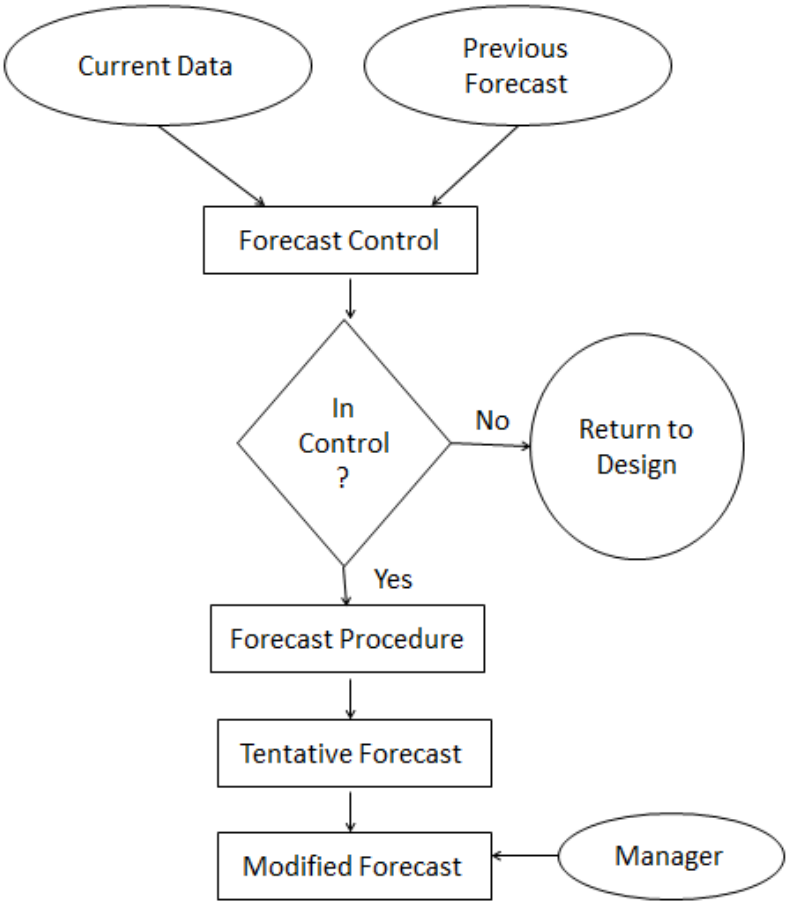


Figure 4.4. Operating a forecasting system

2.6. Caveats

There are two very important facts to remember; forecasts almost never give the exact answer, and the farther into the future we look, the less accurate forecast will be. It would be highly unlikely that the weekly sales forecast of doughnuts would be equal to the actual sales. Fortunately, exact forecasts are usually not required; we just need to be in the ballpark.

A more reasonable approach might be to forecast a range of values or the probability of a group of values. We might forecast selling 250 dozen doughnuts next week but really mean sales should be between 230 and 270 dozen. The weather service does not say it will rain today, but that there is a 60 percent chance of rain. That the accuracy of a forecast depends on how far into the future we look can easily be seen from weather forecasts; the forecast for tomorrow is usually more accurate than the forecast for a week from today.

3 QUALITATIVE FORECASTING

3.1. Market Survey

A market survey consists of several steps. First, a questionnaire must be developed that should contain questions whose answers provide the information needed to determine forecast. This information about the customer might be things such as age and income as well as an indication of whether or not the customer will purchase the product. If the customer is a retail outlet, the information might include size of the store and the projected number of items that will be purchased. A survey to analyze demand for autofocus 350mm cameras is given in Table 4-1. Along with survey design, a method to analyze the results must be determined.

TABLE 4-1
A sample market survey form

Please check the appropriate boxes
☐ I do not own a 35 mm camera
☐ I own an SLR 35 mm camera
☐ A own an autofocus 35 mm camera
☐ I plan to purchase a new SLR 35 mm camera in the next two years
☐ I plan to produce a new autofocus 35 mm camera in the next two years
☐ I do not plan to produce a new 35 mm camera in the next two years

The next step is carrying out the survey, which may be done by mail, FAX, e-mail, telephone, a tear-out postcard in a magazine, or in person. How the survey is carried out can affect the number as well as the quality of responses. The number, location and individual customers to be surveyed should be carefully planned in conjunction with the purpose of the study.

After the survey is conducted, the results should be tabulated and analyzed. Care must be taken in interpreting these results. Response rates may be low, the questions answered incorrectly, or factors not considered in the questionnaire may affect the actual outcome of events. Statistical analysis can also be time consuming. Results of the camera survey might be that 75 percent of the respondents owned a 35 mm camera; 35 percent owned an SLR model and 50 percent owned an autofocus model. The percentages do not add up to 100 percent because some people own both. Of the 35 percent who own SLR cameras, 75 percentages indicate that they will buy an autofocus camera within the next two years. Extrapolating the intentions of SLR and autofocus owners surveyed to the general population can give a forecast of the number of autofocus cameras demanded for the next two years. Of course other factors, e.g., the economy, may play a large role in actual purchases.

Mastio (1994) discusses a survey based on 24 important needs and criteria to determine the factors blow molders use to select a machine supplier. Results indicate that machine dependability and spare parts availability are the most

important needs to below molders. Good technical service and machinery that is easy to operate are other key criteria when selecting a machine. Although important, price is not the most criteria factor. About 65 percent of the blow molders said that they plan to purchase new equipment in the next year. A manufacturer of below molding machinery can make aggressive expansion plans by providing these customers what they seek. Other examples of market surveys for forecasting purposes include demand for computer networks in industry and demand for power transmission products.

Recently, extensions and modifications of market survey techniques have been proposed. One is to use existing customer data base information to augment the survey modify the procedure to account for cross effects on demand for similar products when a new product is made available. Using market surveys to provide insight for management to modify quantitative forecasts has been proposed by Cook (1995).

3.2. Expert Opinion and the Delphi Technique

A different method would be to ask an expert for an opinion on projected sales. This opinion is based on experience and knowledge of the particular situation. Sales and marketing personnel are prime examples of “experts” for forecasting a new product. A variation might be to ask several experts and use a combination of the results, say the median or average, as the forecast. This method is easy to implement but may be inaccurate.

A more formal variation of expert opinion is called the Delphi techniques, named after the Oracle at Delphi of Greek mythology who predicted the future events. A committee of “experts” corresponds to the Oracle in this technique, and a facilitator determines the participants, writes questionnaires, and analyzes the results. Committee members may be expert in quite different fields. For example, one may be sales oriented, but another may be economist.

They give a variety of views and consider many different factors in the process. The person from sales has a good idea of the company’s history in selling other items, whereas the economist may have good information regardless the overall economic picture. Both of these factors affect the sale of a new product.

Committee members are asked to submit an anonymous forecast of specific events, and more importantly, their reasons why they picked that forecast. A simple example of an initial questionnaire is given in Table 4-2. Statements should be unambiguous and simple. Rather than the questionnaire asking if sales will be large, it should ask if they will be above a given value. Questions should have a single answer; if multiple responses are needed to use a separate question for each.

Responses are summarized, and the questionnaire is modified and returned to committee members, who are then asked to repeat the process. The questionnaire for each round should reflect the results of previous rounds; summary statistics, e.g., mean, median, and range, are included in the updated

questionnaire. Table 4-3 shows a questionnaire for intermediate rounds. The procedure continuous until reasonable agreement among committee members is reached –usually three or four rounds are enough to reach consensus- and results are summarized, reported to the participants, and used to make decisions.

Delphi can be used to forecast demand for products. Vickers (1992) uses Delphi to examine the European automobile market, and Stocks (1990) forecasts CD-ROM demand in Australian libraries. Demand for services can also be forecast using Delphi. The numbers of families needing financial aid and tourists who visit a region are examples.

Delphi has several advantages, among which are that it includes involving a variety of people, even those in different locations, and it prohibits domination by strong personalities, giving everyone an equal chance to participate; anonymity allows freer expression of ideas. It also keeps attention focused on the task; written responses are often better thought out than verbal ones. Probably its greatest advantage, however, is in generating and evaluating a large number of ideas for the forecast, many of which might be overlooked in face-to-face meetings.

The biggest disadvantage is the time needed to carry out a Delphi study, often more than a month; electronic methods may speed the process, however. It is also time consuming for the participants, and it may be difficult to keep them totally involved. Written ideas may need clarification or risk being misunderstood; a typical Delphi study is a type of voting procedure, sometimes compromise agreements are not reached.

3.3. Comments on Qualitative Forecasting Methods

Qualitative methods are often used in industry. Sometimes, expert opinion is used because it is “close enough”, fast, and easy to do, and it is particularly adept at quickly sensing trends in market. On the other hand, market surveys and the Delphi method are both time consuming and costly. However, for new product introduction and forecasting technological advances, they may be the only choice. If all steps of the method are followed, the results are usually fairly accurate. One major benefit of judgmental forecasting may be that it forces a commitments from the responsible parties; that is, if the head of sales quotes a figure, he or she may work extra hard to see that sales are not less than forecast. Market surveys give good results, but the time required to do them makes them less appropriate for short-term forecasting. With the growth of networks and customers with access to interactive computing, market surveys may become more timely and accurate. However, the cost must be weighed against the benefit the survey provides.

4. Causal Forecasting with Regression

Mary carter is manager of the plumbing department for the Columbia store of Home Sales, a leading retailer of building supplies. Each month, she must place

an order for bathroom plumbing fixtures. If she orders more than the number sold, the excess fixtures represent money the company cannot use elsewhere. If she orders too few, sales are lost to the store's competitors.

Mary has been thinking about how she might anticipate the demand for fixtures. She knows that most of the fixtures sold at her store are for new houses; replacement fixtures account for less than 6 percent of total sales. Plumbing fixtures are installed after the roof and walls of the house are on, typically about a month after the building permit is issued. Because each new house built requires a building permit, the number of building permits issued last month may help her determine the number of fixtures she should order this month.

This scenario is typical of many forecasting problems. We wish to forecast a dependent variable –sales of plumbing fixtures in our example- and the value of the dependent variable is related to an observable value of one or more independent variables –housing starts in the example. We call this process causal forecasting, because the value of the dependent variable is often caused by, or at least highly correlated with, the value of the independent variable.

The relationship between the dependent and independent variables, however, is not always clear-cut. For example, total sales for a company may vary in a pattern similar to that of some general economic indicators. In this case, aggregate sales could be the dependent variable and various economic indicators, such as the prime interest rate, would be the independent variables. To estimate the relationship, regression techniques are often useful. Let us examine Mary's problem more closely to see how this process is done.

4.1. Simple Linear Regression

The first thing Mary should do is make a scatter plot of her data. Table 4-4 gives the number of housing permits issued in the Columbia area and the number of plumbing fixtures sold in her store for each month of last two years. Notice that the permits for a given month are aligned with the fixture sales for the following month because the is a one month lag between the permit and the sale; i.e., sales in February depend on permits from January. A scatter plot is given in figure 4-5. This plot has housing permits, sorted in ascending order, as its x-axis and the corresponding fixture sales as its y-axis. The scatter plot seems to show a linear relationship between housing starts and fixture demand, so simple linear regression is appropriate. The underlying method is

$$d_t = a + bh_t + \varepsilon_t \qquad t = 1,2,\dots,n$$

Where; d_t = the value of dependent variable at time t

h_t = the value of independent variable at $t-1$

ε_t = random error in the model

a = intercept of the straight line relating d_t and h_t

b = slope of the line

n = the total number of periods of the available data

TABLE 4-4

Housing permits and plumbing fixture sales

Data Point	Month of Permit	Number of Permits	Month of Fixture Sale	Number of Fixtures
1	Jan-94	22	Feb-94	72
2	Feb-94	16	Mar-94	44
3	Mar-94	24	Apr-94	80
4	Apr-94	95	May-94	191
5	May-94	84	Jun-94	187
6	Jun-94	13	Jul-94	57
7	Jul-94	114	Aug-94	238
8	Aug-94	147	Sep-94	283
9	Sep-94	96	Oct-94	204
10	Oct-94	59	Nov-94	144
11	Nov-94	35	Dec-94	10
12	Dec-94	41	Jan-95	109
13	Jan-95	28	Feb-95	63
14	Feb-95	21	Mar-95	50
15	Mar-95	18	Apr-95	67
16	Apr-95	46	May-95	109
17	May-95	145	Jun-95	304
18	Jun-95	122	Jul-95	239
19	Jul-95	108	Aug-95	223
20	Aug-95	85	Sep-95	173
21	Sep-95	107	Oct-95	211
22	Oct-95	53	Nov-95	104
23	Nov-95	17	Dec-95	59
24	Dec-95	12	Jan-96	24

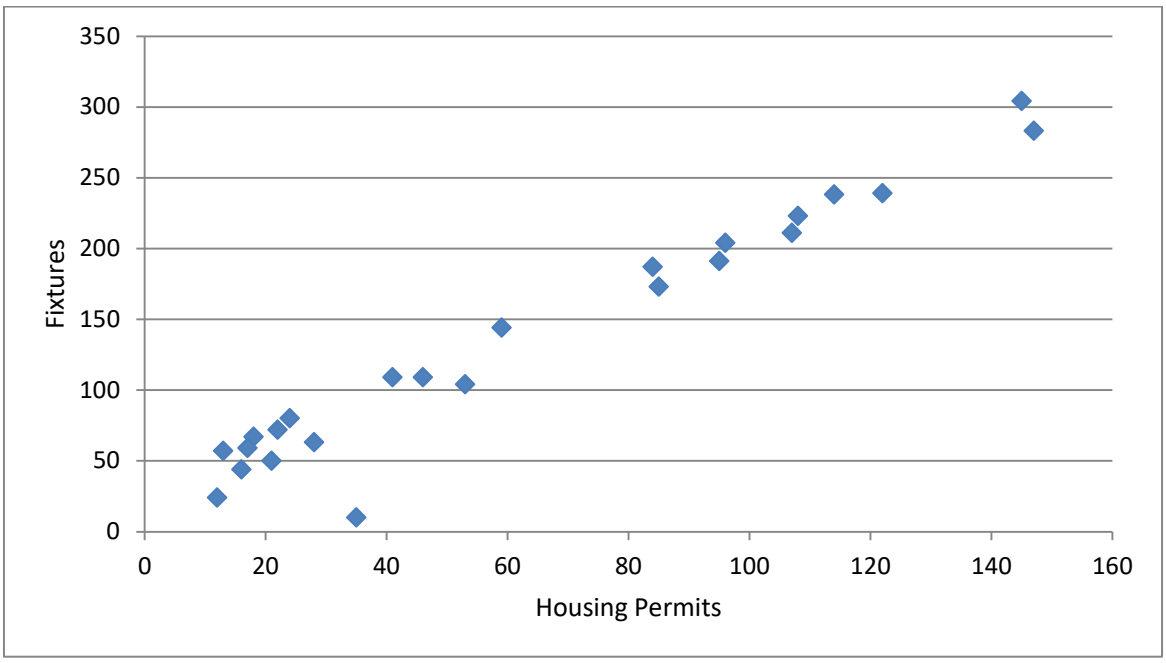


Figure 4.5. Scatter plot of permits and sales

We wish to choose estımate of a and b say $\hat{a}$ and $\hat{b}$, so that a straight line fits the data as closely as possible. To do so, we minimize the sum of the squared differences between the actual sales and the sales indicated by the model. The difference is the "error" of the forecast. By squaring the difference we ensure that the value is nonnegative, penalizing both underestimates and overestimates. Squaring the difference also causes more weight to be placed on larger differences than on smaller differences. Thus we are minimizing the error in our prediction, with larger errors weighted more heavily.
From any basic statistics text, we find

$$\hat{b} = \frac{n\sum_{t=1}^{n} h_t d_t - \sum_{t=1}^{n} h_t \sum_{t=1}^{n} d_t}{n\sum_{t=1}^{n} h_t^2 - (\sum_{t=1}^{n} h_t)^2} \qquad \text{(Slope)}$$

$$\hat{a} = \frac{\sum_{t=1}^{n} d_t}{n} - \hat{b}\frac{\sum_{t=1}^{n} h}{n} \quad \text{(Intercept)}$$

To calculate $\hat{b}$ we need $\sum_{t=1}^{n} h_t d_t$, $\sum_{t=1}^{n} h_t$. $\sum_{t=1}^{n} d_t$, $\sum_{t=1}^{n} h_t^2$. Using a spreadsheet, we calculated these values for Mary's problem; they are given in table 4-5, from these we compute

$$\hat{b} = \frac{(24 * 294095) - (1508 * 3337)}{(24 * 140928) - (1508)^2} = 1.828252$$

$$\hat{a} = \frac{3337}{24} - 1.83\left(\frac{1508}{24}\right) = 24.16647$$

This results in the regression equation

$$\hat{d} \cong 24.17 + 1.83h$$

Where $\hat{d}$ is the estimate for the number of plumbing fixtures sold in a month given that there were h housing starts the previous month. If there are 23 housing starts in January of 1995, we would expect to sell about 24.17+1.83*23=66 fixtures in February.

In general, $\hat{b}$ can be positive or negative. A positive value implies that the dependent variable increase as the independent variable increases or that they are positively correlated. A negative $\hat{b}$ implies the opposite. The magnitude of $\hat{b}$should reflect the amount of change in the dependent variable for a unit change in the independent variable. If either the sign or magnitude of $\hat{b}$ seems inappropriate for the situation, think carefully about the model.

Table 4.5. Regression computation for Mary's data

Data Point	h_t	d_t	h_t^2	d_t^2	$h_t d_t$
1	22	72	484	5184	1584
2	16	44	256	1936	704
3	24	80	576	6400	1920
4	95	191	9025	36481	18145
5	84	187	7056	34969	15708
6	13	57	169	3249	741
7	114	238	12996	56644	27132
8	147	283	21609	80089	41601
9	96	204	9216	41616	19584
10	59	144	3481	20736	8496
11	35	102	1225	10404	3570
12	41	109	1681	11881	4469
13	28	63	784	3969	1764
14	21	50	441	2500	1050
15	18	67	324	4489	1206
16	46	109	2116	11881	5014
17	145	304	21025	92416	44080
18	122	239	14884	57121	29158
19	108	223	11664	49729	24084
20	85	173	7225	29929	14705
21	107	211	11449	44521	22577
22	53	104	2809	10816	5512
23	17	59	289	3481	1003
24	12	24	144	576	288
Total	1508	3337	140928	621017	294095

The value of $\hat{a}$ represents the value of the dependent variable when the independent variable is zero; if zero is not a possible value of the independent variable, $\hat{a}$ may still be positive.

The coefficient of determination is defined as

$$r^2 = \frac{\sum_{t=1}^{n}(\hat{d}_t - \bar{d})^2}{\sum_{t=1}^{n}(d_t - \bar{d})^2}$$

For Mary's data, $r^2 = 0.98$, indicates an excellent fit, because the regression equation explains 98% of the variance. In practice, a coefficient of determination of 0.85 is considered quite well.

4.2. Random errors

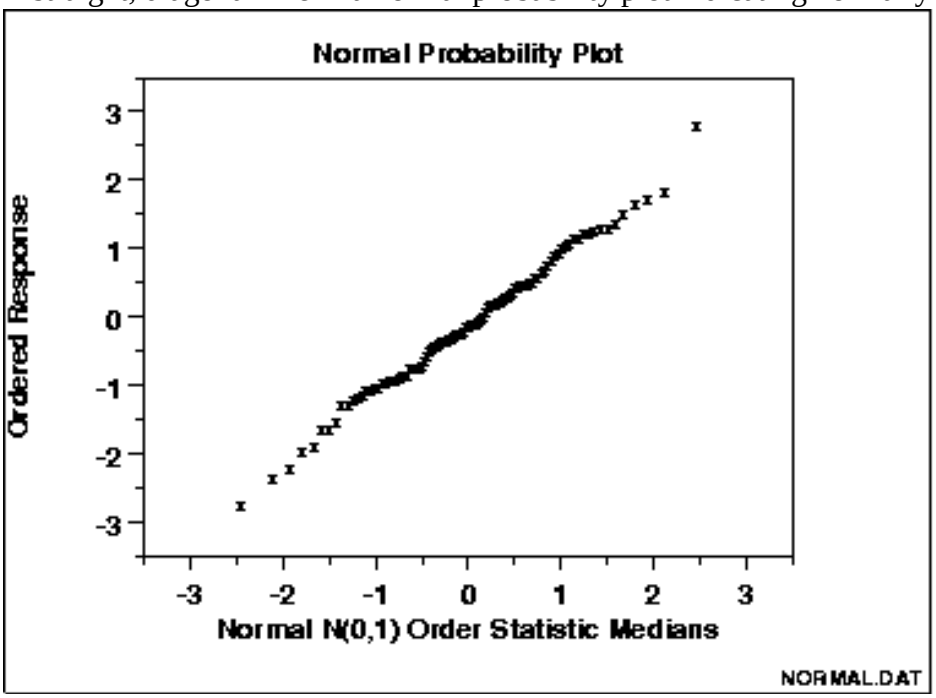
In regression model, ε_t, random error or residual is normally distributed with mean value 0 and standard deviation σ_t

$$\varepsilon_t \sim N(0, \sigma_t)$$

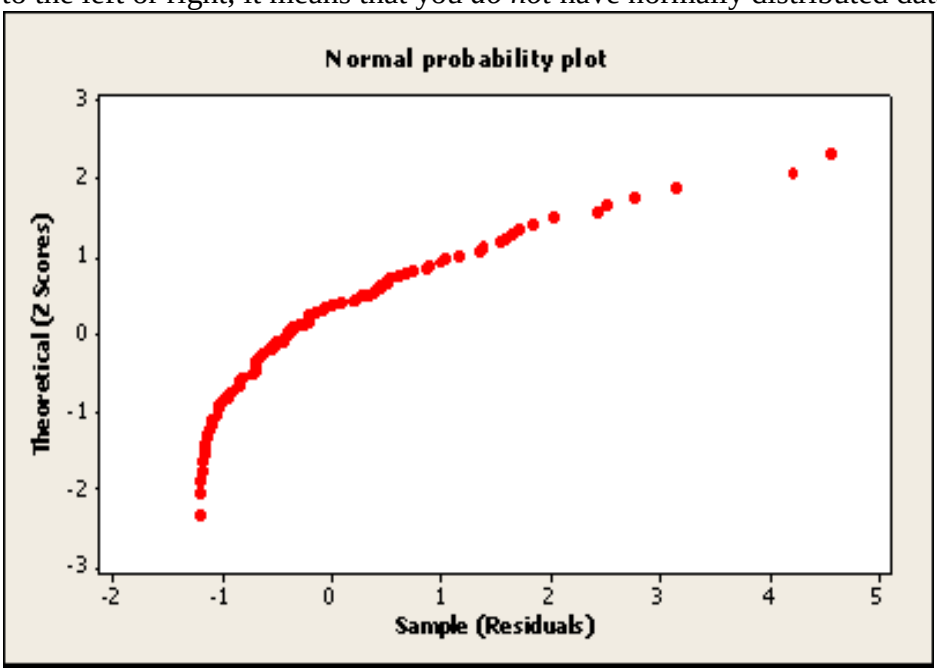
4.2.1 Test of normality

When you have a set of data that you think might have a normal distribution, a graph of your data can help you decide whether or not your data is normal. There is a specialized type of plot you can create, called a **normal probability plot**. A normal probability plot graphs z-scores (normal scores) against your data set.

A straight, diagonal line in a normal probability plot indicating normally distributed data.



A straight, diagonal line means that you have normally distributed data. If the line is skewed to the left or right, it means that you *do not* have normally distributed data.

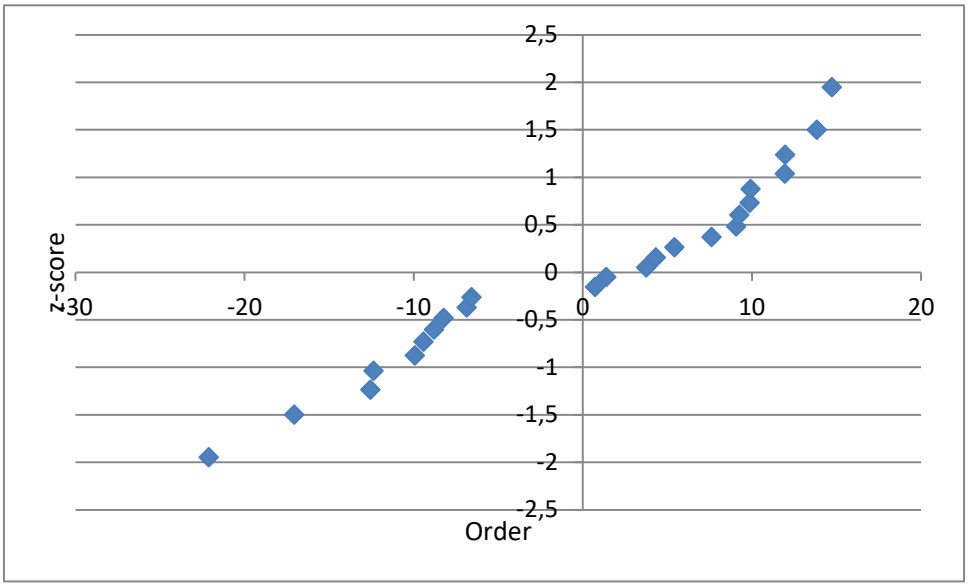


How to Draw a Normal Probability Plot By Hand

1. Arrange your x-values (errors) in ascending order.
2. Calculate $f_i = \frac{i-0.375}{n+0.25}$, where i is the position of the data value in the ordered list and n is the number of observations.
3. Find the z-score for each f_i
4. Plot your x-values on the horizontal axis and the corresponding z-score on the vertical axis.

For example of Mary's store we try to check this property.

ε_t	Order	i	$f_i = \frac{i-0.375}{n+0.25}$	z-score
7,611986	-22,1055	1	0,025773196	-1,9469
-9,4185	-17,0638	2	0,067010309	-1,49843
11,95548	-12,5598	3	0,108247423	-1,2359
-6,85041	-12,3575	4	0,149484536	-1,03865
9,260362	-9,91951	5	0,190721649	-0,87524
9,066254	-9,4185	6	0,231958763	-0,73241
5,412802	-8,78943	7	0,273195876	-0,60318
-9,91951	-8,21321	8	0,31443299	-0,48332
4,321338	-6,85041	9	0,355670103	-0,37006
11,96666	-6,56789	10	0,396907216	-0,26136
13,84471	0,733938	11	0,43814433	-0,15568
9,875198	1,382314	12	0,479381443	-0,05171
-12,3575	3,753246	13	0,520618557	0,051706
-12,5598	4,321338	14	0,56185567	0,155676
9,924994	5,412802	15	0,603092784	0,261361
0,733938	7,611986	16	0,644329897	0,370057
14,73699	9,066254	17	0,68556701	0,483324
-8,21321	9,260362	18	0,726804124	0,603176
1,382314	9,875198	19	0,768041237	0,732411
-6,56789	9,924994	20	0,809278351	0,87524
-8,78943	11,95548	21	0,850515464	1,038647
-17,0638	11,96666	22	0,891752577	1,235902
3,753246	13,84471	23	0,932989691	1,498434
-22,1055	14,73699	24	0,974226804	1,946903



Normal probability plot of residuals

4.2.2 Discussion about mean value of errors

In simple regression model, there is one dependent variable and one independent variable and we wish to estimate dependent variable by independent variable.
It is obvious that, there is difference between exact value and estimated value. This difference refers to random error.
Mean value of random errors must be 0. Elsewhere there is at least another independent variable, or we made mistake/mistakes in our calculations.
For example of Mary's store we try to check this property.

Data Point	h_t	d_t	$\hat{d}_t$	$\varepsilon_t = d_t - \hat{d}_t$
1	22	72	64,38801	7,611986
2	16	44	53,4185	-9,418502
3	24	80	68,04452	11,955482
4	95	191	197,8504	-6,85041
5	84	187	177,7396	9,260362
6	13	57	47,93375	9,066254
7	114	238	232,5872	5,412802
8	147	283	292,9195	-9,919514
9	96	204	199,6787	4,321338
10	59	144	132,0333	11,966662
11	35	102	88,15529	13,84471
12	41	109	99,1248	9,875198
13	28	63	75,35753	-12,357526
14	21	50	62,55976	-12,559762
15	18	67	57,07501	9,924994
16	46	109	108,2661	0,733938
17	145	304	289,263	14,73699
18	122	239	247,2132	-8,213214
19	108	223	221,6177	1,382314
20	85	173	179,5679	-6,56789
21	107	211	219,7894	-8,789434
22	53	104	121,0638	-17,063826
23	17	59	55,24675	3,753246
24	12	24	46,10549	-22,105494
Total	1508	3337		0,000704

$$E(\varepsilon_t) = \frac{\sum_{t=1}^{n} \varepsilon_t}{n} = \frac{0.000704}{24} = 0.00003 \cong 0$$

4.3. Comments on Regression

Regression models are very useful for forecasting when there is a strong relationship and a time lag between the dependent variable and the independent variable. If there is no time lag between dependent and independent variables, i.e., they occur in the same time period, we cannot forecast future values of the dependent value unless we use a forecast of the independent variable, which may introduce additional error in the forecast of the dependent variable.

If causal relationships do not exist, regression is not the best forecasting method.

Standard Normal Probabilities

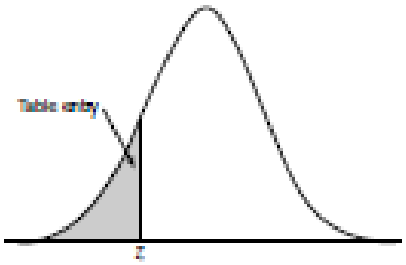


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

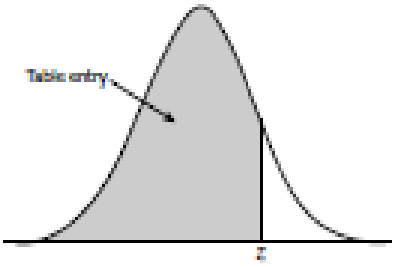


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998