CHAPTER 1

Introduction

Regression and Model Building

- **Definition**

  A statistical methodology that utilizes the relation between two or more quantitative variables so that a response variable or outcome variable can be predicted from the other or others.

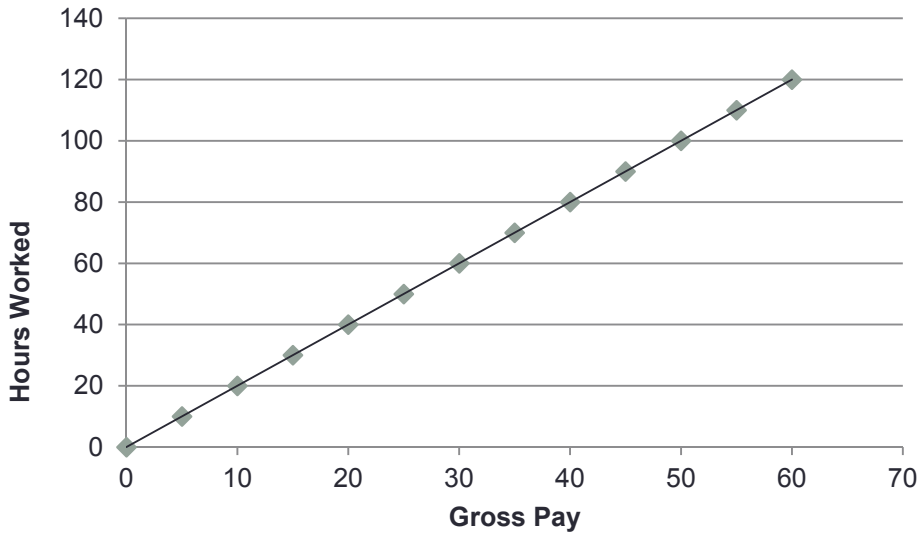
- **Examples**
  - Sales Volume
  - Employee Performance
  - Child Development
  - Hospital Stay
  - Sports Performance

- **Simple Linear Regression (SLR)**

  A single predictor (independent) variable is used for predicting the response or outcome (dependent) variable.

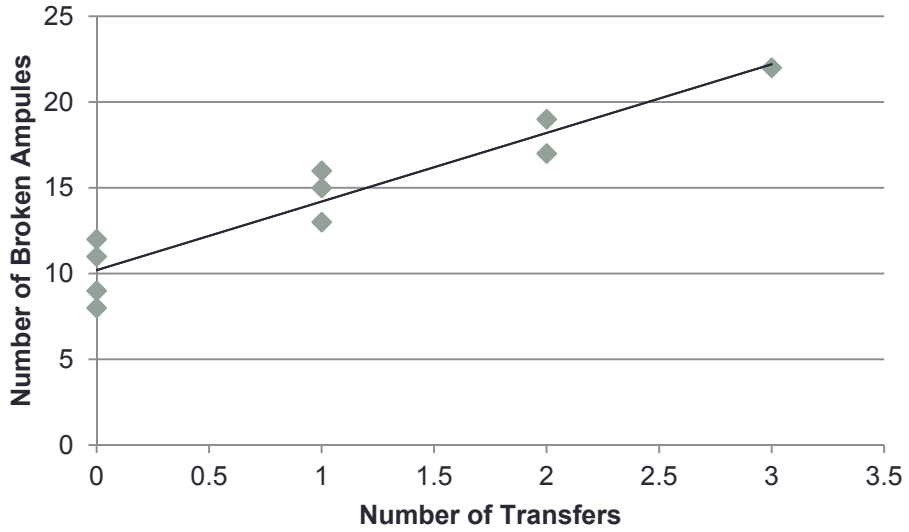
Regression and Model Building

- **Functional Relationship**
  - Expressed as a mathematical formula
  - $y = f(x)$
  - Example: Pay based on hours worked
    - ✓Rate of pay is $25 per hour
    - ✓$y = 25(x)$



Regression and Model Building

- **Statistical Relationship**
  - Not a perfect relationship
  - $y = f(x)$
  - Example: Air freight breakage
    - ✓ x is the number of times carton was transferred
    - ✓ $y = 4(x) + 10.2$



Regression and Model Building

- **Delivery Time Example**

If we let y represent delivery time and x represent delivery volume, then the equation of a straight line relating these two variables is

$$y = \beta_0 + \beta_1 x \qquad (1.1)$$

Regression and Model Building

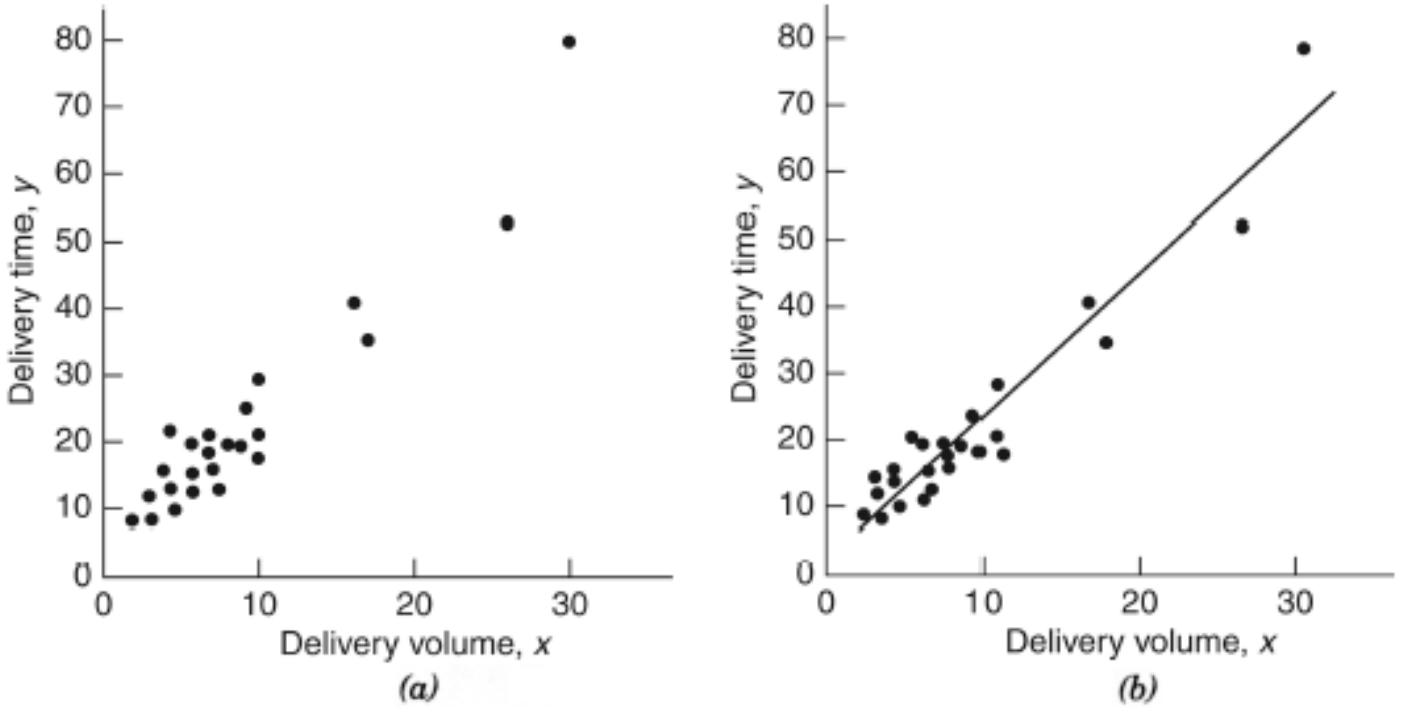


Figure 1.1 (*a*) Scatter diagram for delivery volume. (*b*) Straight-line relationship between delivery time and delivery volume.

Regression and Model Building

- **Simple Linear Regression Model**

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1.2)$$

where y – dependent (response) variable

x – independent (regressor/predictor) variable

β_0 - intercept

β_1 - slope

ε - random error term

~ N(0, σ^2)

Regression and Model Building

- **Normal Error Regression Model**
  - ε_i is normally distributed with a mean of zero and a variance of σ^2
  - Y_i are independent normal variables with a mean of $\beta_0 + \beta_1 X_i$ and a variance of σ^2
  - Normality assumption for the error term is justifiable in many situations since the error represents the effects of many factors omitted from the model.

Regression and Model Building

- **Historical Origins**
  - Developed by Sir Francis Galton in late 1800's
  - Studied the relationship between the heights of parents and their children
  - Noted the heights of children revert or regress to the mean
- **Basic Concepts**
  - Response variable, *Y*, varies with the predictor variable, *X*, in a systematic fashion
  - For a given value of *X*, there is variance in the value of *Y*

Regression and Model Building

- **The mean response** at any value, *x*, of the regressor variable is

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta x + \varepsilon) = \beta_0 + \beta_1 x$$

- The **variance** of y at any given *x* is

$$\mathrm{Var}(y|x) = \sigma^2_{y|x} = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

Regression and Model Building

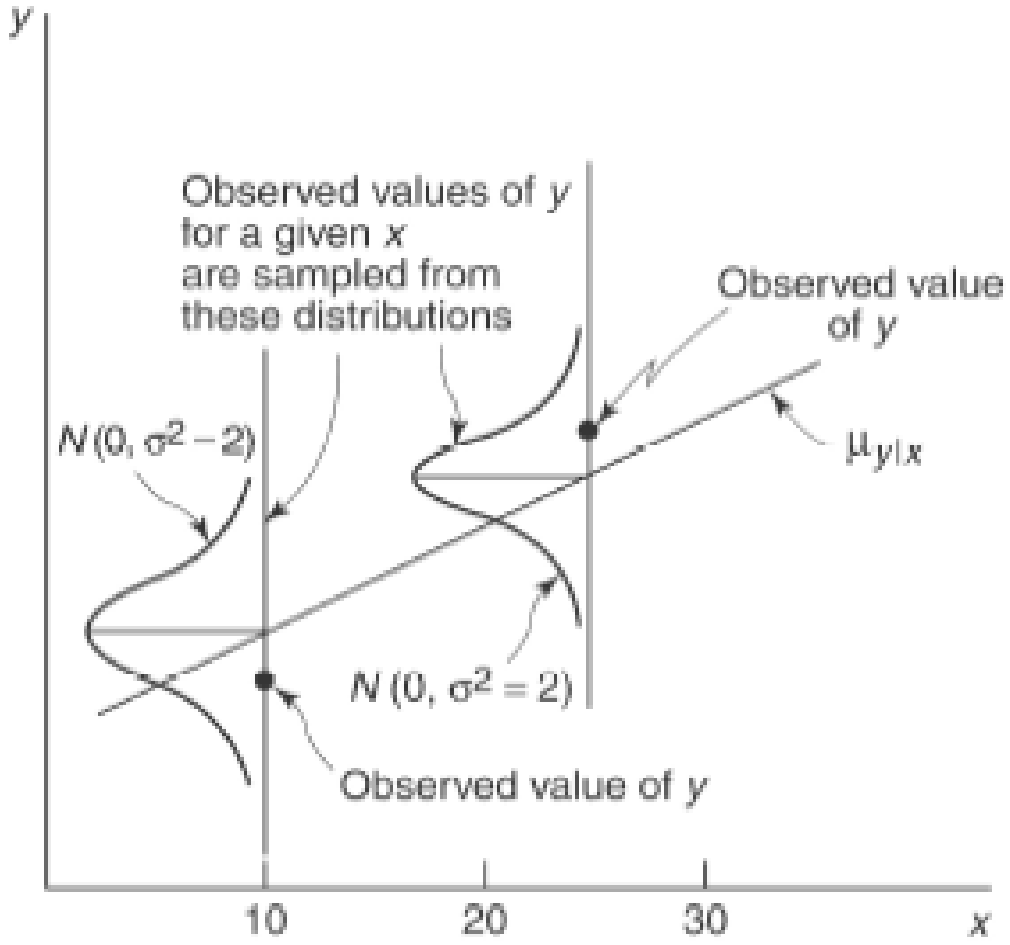


Figure 1.2 How observations are generated in linear regression.

Regression and Model Building

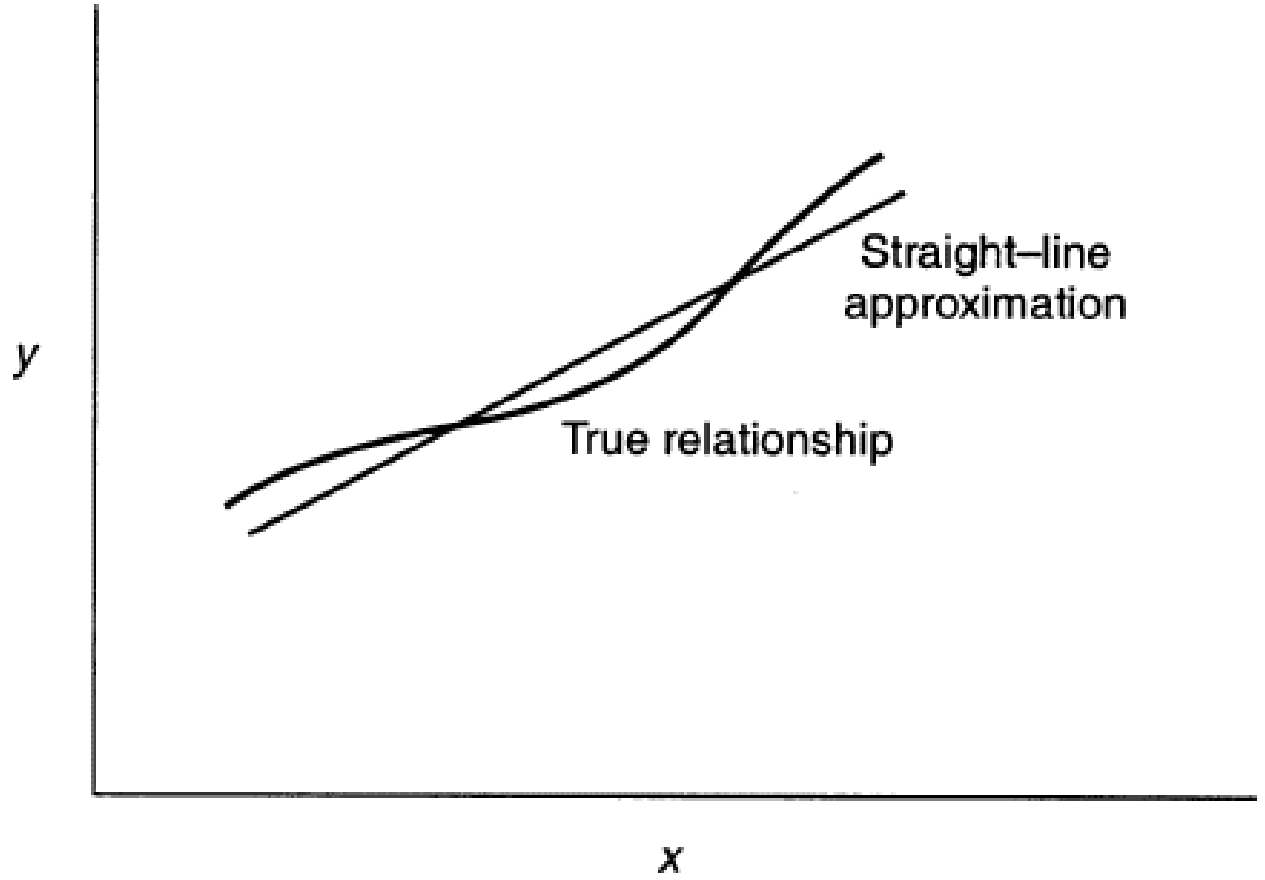


Figure 1.3 Linear regression approximation of a complex relationship.

Regression and Model Building

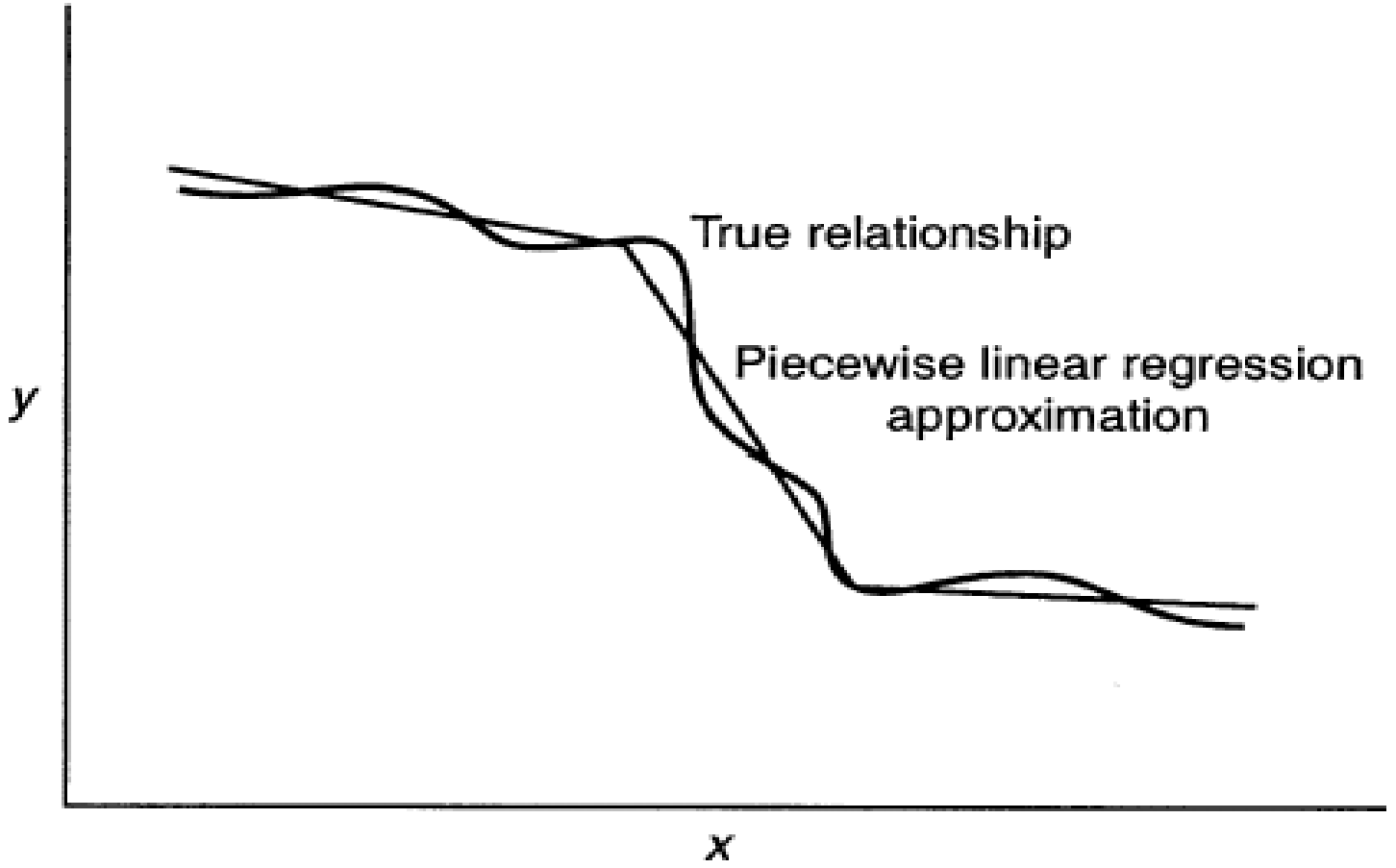


Figure 1.4 Piecewise linear approximation of a complex relationship.

Regression and Model Building

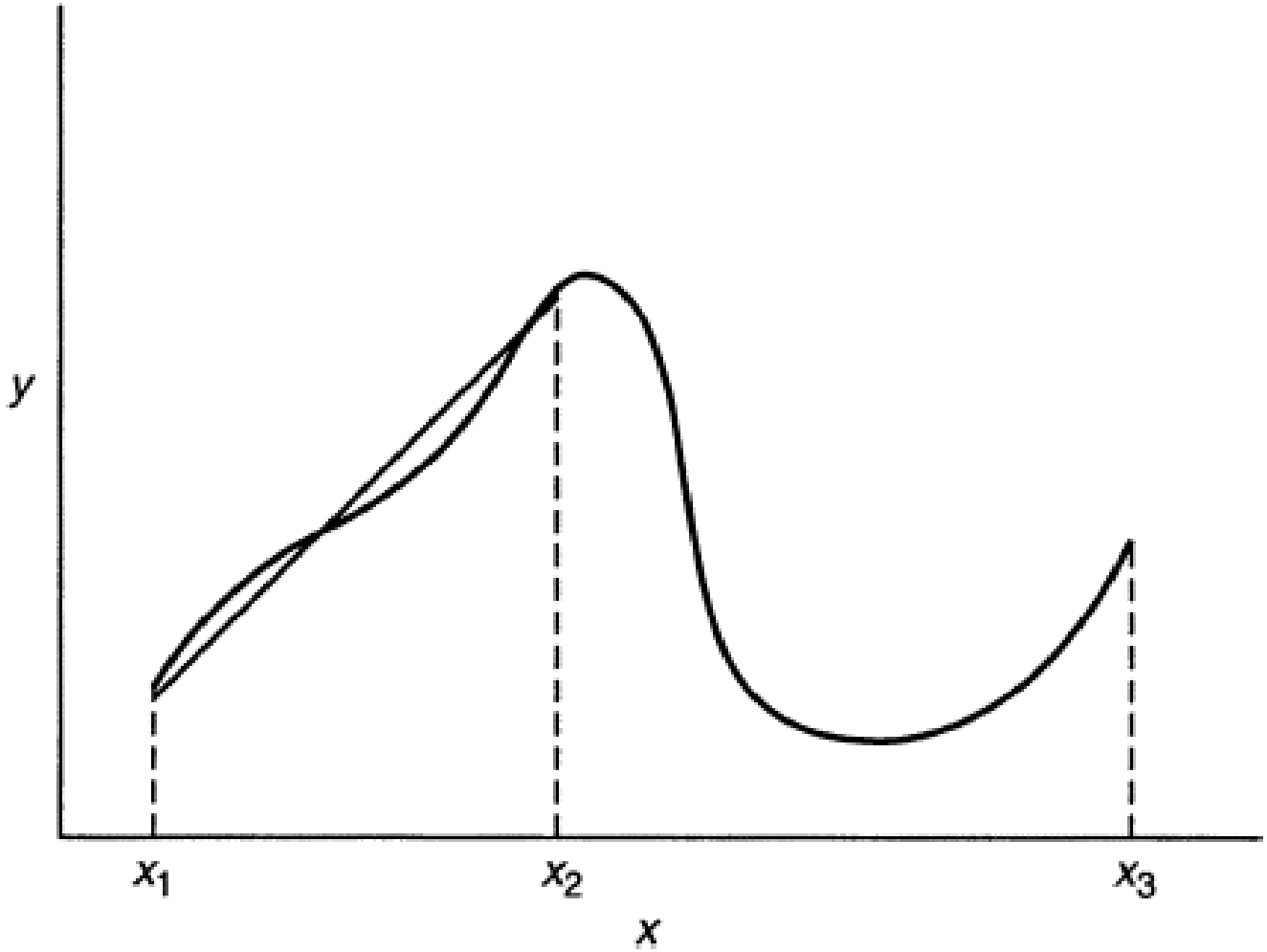


Figure 1.5 The danger of extrapolation in regression.

Regression and Model Building

- Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1.3)$$

- Estimation of Unknown Parameters
- Model Adequacy Checking
- Iterative Process
- Causation
- Insight & Understanding

Data Collection

- Analysis/model is only as good as the data
- Three different methods can be used for data collection
  1. A retrospective study based on historical data
  2. An observational study
  3. A designed experiment

Data Collection

Example 1.1

Consider an acetone–butyl alcohol distillation column. Possible factors that may influence the concentration of acetone in the distillate (product) stream are: the reboil temperature, the condensate temperature, and the reflux rate. For this column, production maintains and archives the following records:

- The concentration of acetone in a test sample taken every hour from the product stream
- The reboil temperature controller log, which is a plot of the reboil temperature
- The condenser temperature controller log
- The nominal reflux rate each hour

The nominal reflux rate is supposed to be constant for this process. Only infrequently does production change this rate.

Data Collection

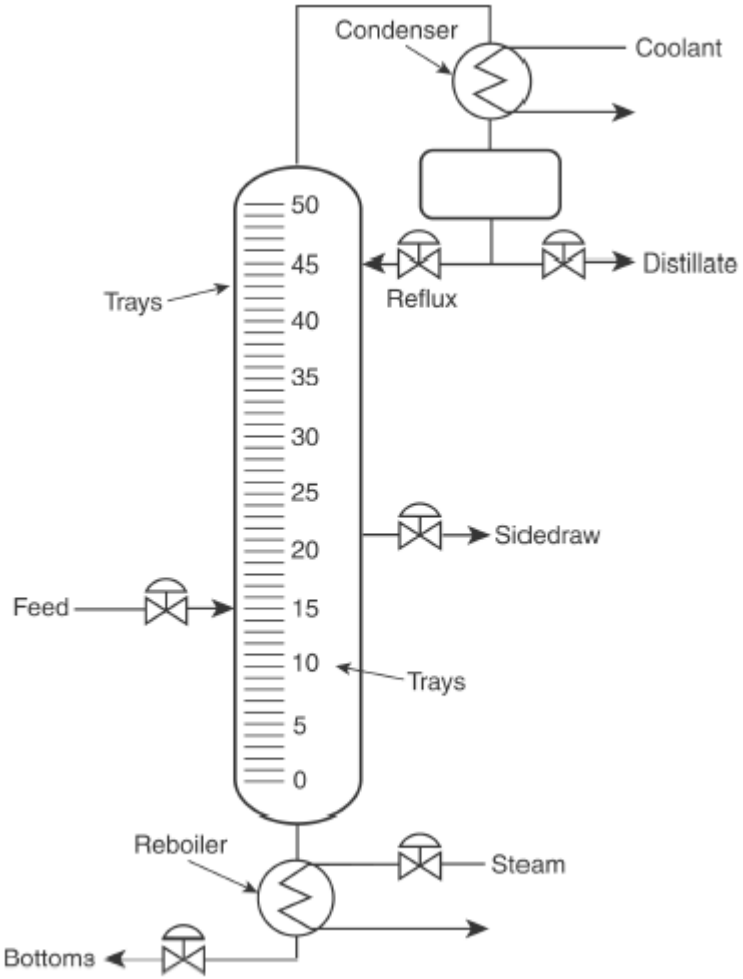


Figure 1.6 Acetone–butyl alcohol distlllation column.

Data Collection

Retrospective Study

We could pursue a retrospective study that would use either all or a sample of the historical process data over some period of time to determine the relationships among the two temperatures and the reflux rate on the acetone concentration in the product stream. In so doing, we take advantage of previously collected data and minimize the cost of the study. However, we must note several problems.

1. We really cannot see the effect of reflux on the concentration since we must assume that it did not vary much over the historical period.
2. The data relating the two temperatures to the acetone concentration do not correspond directly. Constructing an approximate correspondence usually requires a great deal of effort.
3. Production maintains both temperatures as tightly as possible to specific target values through the use of automatic controllers. Since the two temperatures vary so little over time, we will have a great deal of difficulty seeing their real impact on the concentration.
4. Within the narrow confines that they do vary, the condensate temperature tends to increase with the reboil temperature. As a result, we will have a great deal of difficulty separating out the individual effects of the two temperatures. This leads to the problem of **collinearity** or **multicollinearity**, which we discuss in Chapter 10.

Data Collection

Retrospective studies often offer limited amounts of useful information. In general, their primary disadvantages are

- Some of the relevant data often are missing
- The reliability and quality of the data are often highly questionable
- The nature of the data often may not allow us to address the problem at hand
- The analyst often tries to use the data in ways they were never intended to be used
- Logs, notebooks, and memories may not explain interesting phenomena identified by the data analysis

Data Collection

Observational Study

We could use an observational study to collect data for this problem. As the name implies, an observational study simply observes the process or population. We interact or disturb the process only as much as is required to obtain relevant data. With proper planning, these studies can ensure accurate, complete, and reliable data. On the other hand, these studies often provide very limited information about specific relationships among the data.

In this example, we would set up a data collection form that would allow the production personnel to record the two temperatures and the actual reflux rate at specified times corresponding to the observed concentration of acetone in the product stream. The data collection form should provide the ability to add comments in order to record any interesting phenomena that may occur. Such a procedure would ensure accurate and reliable data collection and would take care of problems (1) and (2) above. This approach also minimizes the chances of observing an outlier related to some error in the data. Unfortunately, an observational study cannot address problems (3) and (4). As a result, observational studies can lend themselves to problems with collinearity.

Data Collection

Designed Experiment

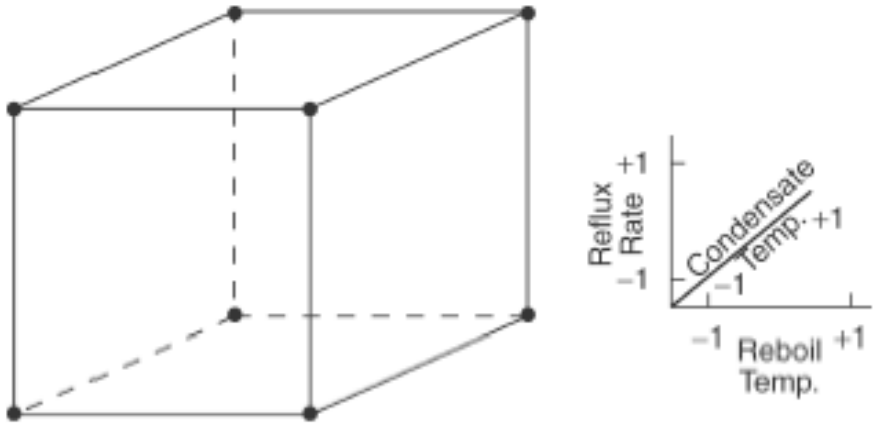


Figure 1.7 The designed experiment for the distillation column.

TABLE 1.1 The Designed Experiment for the Distillation Column

Reboil Temp.	Condensate Temp.	Reflux Rate
−1	−1	−1
+1	−1	−1
−1	+1	−1
+1	+1	−1
−1	−1	+1
+1	−1	+1
−1	+1	+1
+1	+1	+1

Uses of Regression

- There are many uses of regression, including:
  - Data description
  - Parameter estimation
  - Prediction and estimation
- Regression analysis is perhaps the most widely used statistical technique, and probably the most widely misused.

Uses of Regression

- Cause and Effect Relationships
  - Caution: just because you *can* fit a linear model to a set of data, does not mean you should.
    - Some time ago, Wal-Mart decided to combine the data from its loyalty card system with that from its point of sale systems. The former provided Wal-Mart with demographic data about its customers, the latter told it where, when and what those customers bought. Once combined, the data was mined extensively and many correlations appeared. Some of these were obvious; people who buy gin are also likely to buy tonic. They often also buy lemons. However, one correlation stood out like a sore thumb because it was so unexpected.
    - On Friday afternoons, young American males who buy diapers (nappies) also have a predisposition to buy beer. No one had predicted that result, so no one would ever have even asked the question in the first place.
  - It is relatively easy to build “nonsense” relationships between variables
  - Regression does not necessarily imply causality

Model Building in Regression

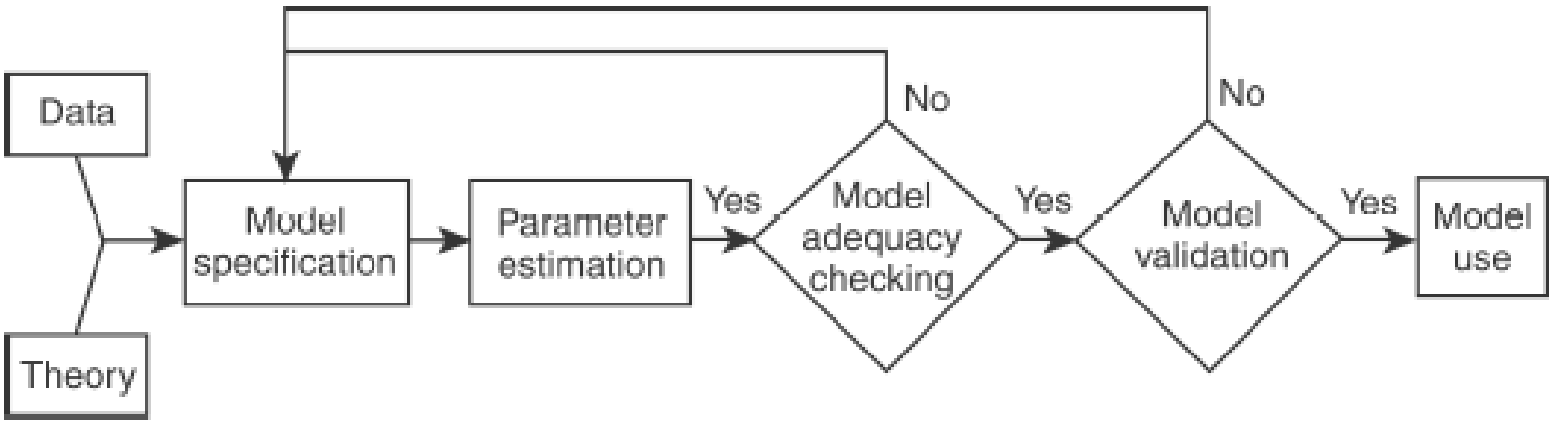


Figure 1.8 Regression model-building process.

Role of the Computer

- Necessary Tool
- Software Examples: Excel, Minitab, SPSS, SAS, and R
- Not a Substitute for Creative Thinking
- Intelligent & Artful Use

Role of the Computer

- Can use SAS or R(highly recommended) to complete assignments
- Select one
- Consider the following:
  - R is free, but a student version of SAS can be obtained at the bookstore at a discounted price
  - SAS can be accessed from the computer labs
  - SAS can be accessed through SAS On Demand

The R Software

- *A Gentle Introduction to Statistical Programming with R: Part I*
  - **Brady T. West**
  - Institute for Social Research (ISR)
  - Center for Statistical Consultation and Research (CSCAR)
  - University of Michigan –Ann Arbor
  - bwest@umich.edu
  - An **Academy Health**Webinar
- YouTube
- Syllabus

The R Software

- R is both a computing language and an environment for statistical computing and statistical graphics
- R is **FREE**, open source software
- The origins of R were in development of the S computing language

The R Software

- The R Development Project Web Page:
  - http://www.r-project.org/
- This web page contains a significant amount of information about the R software
- The actual software is downloaded from a Comprehensive R Archive Network (CRAN) mirror (note the CRAN link)

The R Software

- From the R web page, click the CRAN link on the left side of the page
- Select an appropriate mirror for your location (any U.S.A. location is fine)
- Click the appropriate operating system
- Click base(for the base R software)
- Click the link Download R x.x.x for Windows(or for any other platform)

The R Software

- This will download an executable file, and running the file will start a Wizard to guide you through the installation
- After the software has been installed, you should see an R shortcut on your desktop
- Double-click the shortcut to start R

Rstudio (highly recommended)

- https://www.rstudio.com/
- It is an IDE for R. Great tool.
- To download Rstudio, go to the website, click “Powerful IDE for R”, click “Desktop”, click “Download Rstudio Desktop”, select operating system and install it.