CHAPTER 2

FORECASTING AND DEMAND MODELING

2.1 INTRODUCTION

Demand forecasting is one of the most fundamental tasks that a business must perform. It can be a significant source of competitive advantage by improving customer service levels and by reducing costs related to supply–demand mismatches. In contrast, biased or otherwise inaccurate forecasting results in inferior decisions and thus undermines business performance.

For example, the toy retailer Toys "R" Us made a huge mistake in demand forecasting for the 2015 Christmas season. For several days, the actual number of online orders was more than twice the company's forecasts, and the company's distribution centers were overwhelmed. As a result, the company was forced to throttle demand by terminating some online sales, resulting in lower demand and lower revenue (Ziobro 2016).

The goal of the forecasting models discussed in this chapter is to estimate the quantity of a product or service that consumers will purchase. Most classical forecasting techniques involve time-series methods that require substantial historical data. Some of these methods are designed for demands that are stable over time. Others can handle demands that exhibit trends or seasonality, but even these require the trends to be stable and predictable. However, products today have shorter and shorter life cycles, in part driven by rapid technology upgrades for high-tech products. As a result, firms have much less historical data available to use for forecasting, and any trends that may be evident in historical data may be unreliable for predicting the future.

Fundamentals of Supply Chain Theory, Second Edition. Lawrence V. Snyder and Zuo-Jun Max Shen.

Companion website: www.wiley.com/go/Snyder/SupplyChainTheory

In this chapter, we first discuss some classical methods for *forecasting* demand, in Sections 2.2 and 2.3. Next, in Section 2.4, we discuss more recent approaches to forecasting demand using machine learning when we have large quantities of historical data available. In Sections 2.5–2.8, we discuss several methods that can be used to predict demands for new products or products that do not have much historical data. To distinguish these methods from classical time-series–based methods, we call them *demand modeling* techniques.

The methods that we discuss in this chapter are quantitative. They all involve mathematical models with parameters that must be calibrated. In contrast, some popular methods for forecasting demand with little or no historical data, such as the *Delphi method*, rely on experts' qualitative assessments or questionnaires to develop forecasts.

Demand processes may exhibit various forms of nonstationarity over time. These include the following:

- *Trends*: Demand consistently increases or decreases over time.
- *Seasonality*: Demand shows peaks and valleys at consistent intervals.
- *Product life cycles*: Demand goes through phases of rapid growth, maturity, and decline.

Moreover, demands exhibit *random error*—variations that cannot be explained or predicted—and this randomness is typically superimposed on any underlying nonstationarity.

2.2 CLASSICAL DEMAND FORECASTING METHODS

Classical forecasting methods use prior demand history to generate a forecast. Some of the methods, such as moving average and (single) exponential smoothing, assume that past patterns of demand will continue into the future, that is, no trend is present. As a result, these techniques are best used for mature products with a large amount of historical data. On the other hand, regression analysis and double and triple exponential smoothing can account for a trend or other pattern in the data. We discuss each of these methods next.

In each of the models that follow, we use $D_1, D_2, \ldots, D_t, \ldots$ to represent the historical demand data, i.e., the realized demands in periods 1, 2, …, t, …. We also use y_t to denote the forecast of period t's demand that is made in period $t-1$.

2.2.1 Moving Average

The *moving average* method calculates the average amount of demand over a given interval of time and uses this average to predict the future demand. As a result, moving average forecasts work best for demand that has no trend or seasonality. Such demand processes can be modeled as follows:

$$D_t = I + \epsilon_t, \tag{2.1}$$

where I is the mean or "base" demand and ϵ_t is a random error term.

A moving average forecast of order N uses the N most recent observed demands. The forecast for the demand in period t is simply given by

$$y_t = \frac{1}{N} \sum_{i=t-N}^{t-1} D_i. \tag{2.2}$$

Table 2.1 Monthly historical demand of books and CDs for Examples 2.1–2.5.

	Demand (Thousands)		
Month	*An Inventory Story*	*The TSP Mystery*	CDs
1	10.61	12.61	10.21
2	12.01	16.01	23.01
3	9.77	15.77	10.97
4	10.19	18.19	14.59
5	9.44	19.44	29.44
6	11.40	23.40	16.80
7	9.66	23.66	18.86
8	9.90	25.90	38.90
9	9.01	27.01	18.61
10	10.20	30.20	24.20
11	10.90	32.90	48.90
12	8.98	32.98	22.78

That is, the forecast is simply the arithmetic mean of the previous N observations. This is known as a *simple moving average forecast of order* N.

A generalization of the simple moving average forecast is the *weighted moving average*, which allows each period to carry a different weight. For instance, if more recent demand is deemed more relevant, then the forecaster can assign larger weights to recent demands than to older ones. If w_i is the weight placed on the demand in period i, then the weighted moving average forecast is given by

$$y_t = \frac{\sum_{i=t-N}^{t-1} w_i D_i}{\sum_{i=t-N}^{t-1} w_i}. \tag{2.3}$$

Typically, the weights decrease by 1 in each period: $w_{t-1} = N$, $w_{t-2} = N - 1$, ..., $w_{t-N} = 1$.

□ **EXAMPLE 2.1**

A book store has historical demand data for the book *An Inventory Story* for the past 12 months, as shown in Table 2.1. From a quick look, it is clear that the demand is relatively stable, fluctuating around the value 10, which makes it suitable for the moving average method. Suppose the book store manager wants to predict the demand of this book for the next month. Using an order of $N = 5$, the forecast is given by

$$y_{13} = \frac{D_8 + D_9 + D_{10} + D_{11} + D_{12}}{5} = 9.80.$$

□

2.2.2 Exponential Smoothing

Exponential smoothing is a technique that uses a weighted average of all past data as the basis for the forecast. It gives more weight to recent information and smaller weight to

observations in the past. Single exponential smoothing assumes that the demand process is stationary. Double exponential smoothing assumes that there is a trend, while triple exponential smoothing accounts for both trends and seasonality. These methods all require user-specified parameters that determine the relative weights placed on recent and older observations when predicting the demand, trend, and seasonality. These three weights are called, respectively, the *smoothing factor*, the *trend factor*, and the *seasonality factor*. We discuss each of these three methods next.

2.2.2.1 Single Exponential Smoothing Define $0 < \alpha \leq 1$ as the smoothing constant. Then, we can express the current forecast as the weighted average of the previous forecast and most recently observed demand value:

$$y_t = \alpha D_{t-1} + (1-\alpha) y_{t-1}. \tag{2.4}$$

Note that α is the weight placed on the demand observation and $1-\alpha$ is the weight placed on the last forecast. Typically, we place more weight on the previous forecast, so α is closer to 0 than to 1.

Since each forecast depends on the previous forecast, we need a way to get the process started. One simple way to do this is to set $y_1 = D_1$. Note that this method requires one historical demand observation D_1; the first "real" forecast, i.e., the first forecast that uses (2.4), is y_2.

Using (2.4), we can write

$$y_{t-1} = \alpha D_{t-2} + (1-\alpha) y_{t-2},$$

so

$$y_t = \alpha D_{t-1} + \alpha(1-\alpha) D_{t-2} + (1-\alpha)^2 y_{t-2}.$$

We can continue the substitution in this way and eventually obtain

$$y_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i D_{t-i-1} = \sum_{i=0}^{\infty} \alpha_i D_{t-i-1},$$

where $\alpha_i = \alpha(1-\alpha)^i$. The single exponential smoothing forecast includes all past observations, but since $\alpha_i < \alpha_j$ for $i > j$, the weights are decreasing as we move backward in time, as illustrated in Figure 2.1. Moreover,

$$\sum_{i=0}^{\infty} \alpha_i = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i = 1$$

by (C.50) in Appendix C. These weights can be approximated with an exponential function $f(i) = \alpha e^{-\alpha i}$. This is why this method is called exponential smoothing.

□ **EXAMPLE 2.2**

Suppose that the book store manager from Example 2.1 wishes to use exponential smoothing to forecast next month's demand for *An Inventory Story*. Using $\alpha = 0.2$, we first initialize $y_1 = D_1$, and then obtain

$$y_2 = 0.2 D_1 + 0.8 y_1 = 10.61$$

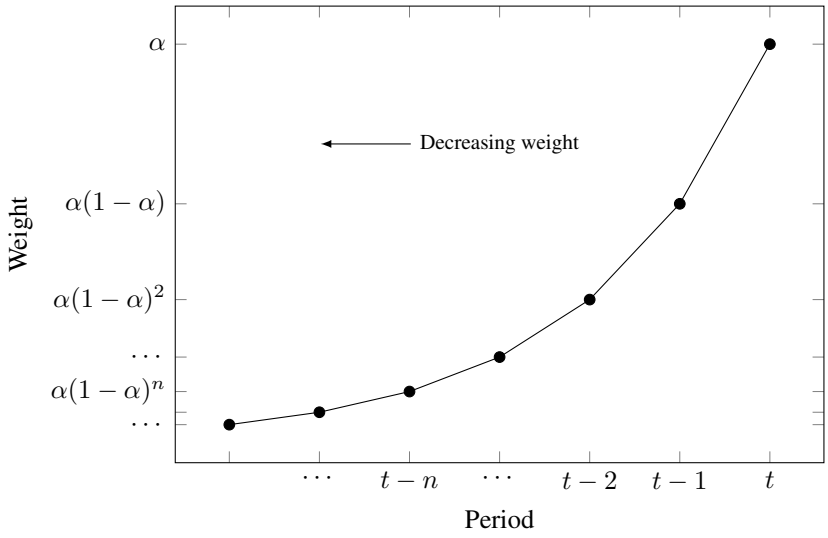

Figure 2.1 Weight distribution for single exponential smoothing.

$$y_3 = 0.2D_2 + 0.8y_2 = 10.89.$$
$$y_4 = 0.2D_3 + 0.8y_3 = 10.67$$

Continuing in this manner, we eventually get

$$y_{13} = 0.2D_{12} + 0.8y_{12} = 9.98.$$

□

2.2.2.2 Double Exponential Smoothing Double exponential smoothing can be used to forecast demands with a linear trend. Such demands can be modeled as follows:

$$D_t = I + tS + \epsilon_t, \tag{2.5}$$

where I is the base demand, S is the slope of the trend in the demand, and ϵ_t is an error term. The forecast for the demand in period t is the sum of two separate estimates from period $t-1$: one of the *base signal* (the value of the demand process) and one of the *slope*. That is,

$$y_t = I_{t-1} + S_{t-1}, \tag{2.6}$$

where I_{t-1} is the estimate of the base signal and S_{t-1} is the estimate of the slope, both made in period $t-1$. I_{t-1} represents our estimate of where the demand process fell in period $t-1$; in period t, the process will be S_{t-1} units greater. The estimates of the base signal and slope are calculated as follows:

$$I_t = \alpha D_t + (1-\alpha)(I_{t-1} + S_{t-1}) \tag{2.7}$$
$$S_t = \beta(I_t - I_{t-1}) + (1-\beta)S_{t-1}, \tag{2.8}$$

where α is the smoothing constant and β is the trend constant. Equation (2.7) is similar to (2.4) for single exponential smoothing in the sense that α is the weight placed on the most

recent actual demand D_t and $1 - \alpha$ is the weight on the previous forecast. Equation (2.8) can be explained similarly: It places a weight of β on the most recent estimate of the slope (obtained by taking the difference between the two most recent base signals) and a weight of $1 - \beta$ on the previous estimate. Note that, if the trend is downward-sloping, then S_t will (usually) be negative.

As with single exponential smoothing, we need a way to initialize the process. This time, we need two historical demand observations to initialize the forecasts, and we typically set $I_1 = D_1$ and $S_1 = D_2 - D_1$ (then $y_2 = I_1 + S_1 = D_2$). The first "real" forecast (using (2.7)–(2.8) to get values for (2.6)) is y_3.

This particular version of double exponential smoothing is also known as *Holt's method* (Holt 1957).

□ **EXAMPLE 2.3**

Suppose that the bookstore manager from Example 2.1 now turns her attention to another book (*The TSP Mystery*), with a different set of historical demand data, as presented in Table 2.1. In contrast to the stable demand of *An Inventory Story*, *The TSP Mystery*'s monthly demand data exhibits an increasing trend. Therefore, moving averages and single exponential smoothing may not accurately predict the demand of *The TSP Mystery* in the next month. For example, if we use a simple moving average of order $N = 5$, we get $y_{13} = 29.80$, which is much smaller than the demands in months 11 and 12. This may not be a good forecast, as we expect the demand in month 13 to continue to increase over time.

Instead, we will use double exponential smoothing for *The TSP Mystery*, with $\alpha = \beta = 0.2$. We initialize $I_1 = D_1 = 12.61$ and $S_1 = D_2 - D_1 = 3.40$. Then we have

$$\begin{aligned} y_2 &= I_1 + S_1 = 16.01 \\ I_2 &= 0.2D_2 + 0.8(I_1 + S_1) = 16.01 \\ S_2 &= 0.2(I_2 - I_1) + 0.8S_1 = 3.40 \\ y_3 &= I_2 + S_2 = 19.41 \\ I_3 &= 0.2D_3 + 0.8(I_2 + S_2) = 18.68 \\ S_3 &= 0.2(I_3 - I_2) + 0.8S_2 = 3.25. \end{aligned}$$

We continue this process and finally obtain

$$\begin{aligned} I_{12} &= 0.2D_{12} + 0.8(I_{11} + S_{11}) = 35.65 \\ S_{12} &= 0.2(I_{12} - I_{11}) + 0.8S_{11} = 1.94. \end{aligned}$$

So the forecast from double exponential smoothing is $y_{13} = I_{12} + S_{12} = 37.59$, which coincides with the increasing trend. □

2.2.2.3 Triple Exponential Smoothing Triple exponential smoothing can be used to forecast demands that exhibit both trend and seasonality. Seasonality means that the demand series has a pattern that repeats every N periods for some fixed N. N consecutive periods are called a *season*. (If the demand pattern repeats every year, for example, then a

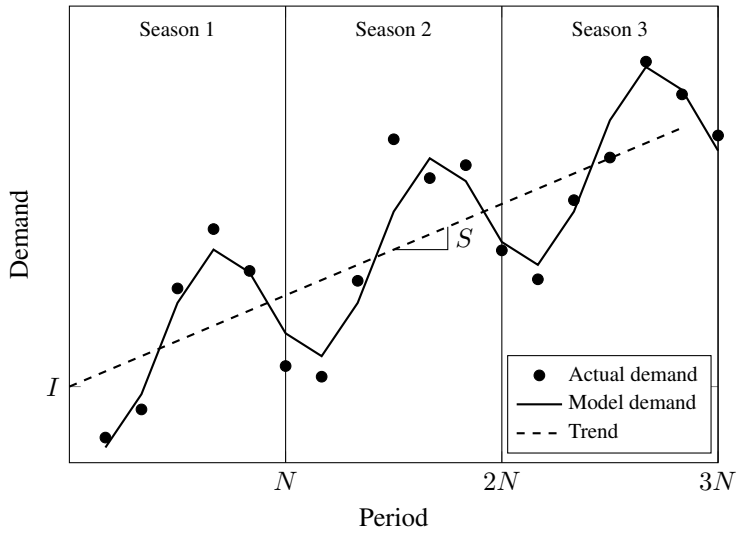

Figure 2.2 Random demands with trend and seasonality.

season is one year. This is different from the common usage of the word "season," which would refer to a portion of the year.)

To model the seasonality, we use a parameter c_t, $1 \leq t \leq N$, to represent the ratio between the average demand in period t and the overall average. (Thus, $\sum c_t = N$.) For example, if $c_6 = 0.88$, then on average, the demand in period 6 is 12% below the overall average demand. The c_t are called *seasonal factors*. We assume that the seasonal factors are unknown but that they are the same every season. The demand process can be modeled as follows:

$$D_t = (I + tS)c_t + \epsilon_t, \tag{2.9}$$

where I is the value of base signal at time 0, S is the true slope, and ϵ_t is a random error term. (See Figure 2.2.)

The forecast for period t is given by

$$y_t = (I_{t-1} + S_{t-1})c_{t-N}, \tag{2.10}$$

where I_{t-1} and S_{t-1} are the estimates of the base signal and slope in period $t-1$ and c_{t-N} is the estimate of the seasonal factor one season ago.

The idea behind smoothing with trend and seasonality is basically to "de-trend" and "de-seasonalize" the time series by separating the base signal from the trend and seasonality effects. The method uses three smoothing parameters, α, β, and γ, in estimating the base signal, the trend, and the seasonality, respectively:

$$I_t = \alpha \frac{D_t}{c_{t-N}} + (1-\alpha)(I_{t-1} + S_{t-1}) \tag{2.11}$$

$$S_t = \beta(I_t - I_{t-1}) + (1-\beta)S_{t-1} \tag{2.12}$$

$$c_t = \gamma \frac{D_t}{I_t} + (1-\gamma)c_{t-N}. \tag{2.13}$$

Equations (2.11) and (2.12) are very similar to (2.7) and (2.8) for double exponential smoothing, except that (2.11) uses the deseasonalized demand observation, D_t/c_{t-N}, instead of D_t, to average it with the current forecast. In (2.13), I_t is our estimate of the base signal, so D_t/I_t is our estimate of c_t based on the most recent demand. This is averaged with our previous estimate of c_t (made N periods ago) using weighting factor γ.

Initializing triple exponential smoothing is a bit trickier than for single or double exponential smoothing. To do so, we usually need at least two entire seasons' worth of data ($2N$ periods), which will be used for the initialization phase. One common method is to initialize the slope as

$$S_{2N} = \frac{1}{N}\left(\frac{D_{N+1} - D_1}{N} + \frac{D_{N+2} - D_2}{N} + \cdots + \frac{D_{2N} - D_N}{N}\right). \tag{2.14}$$

In other words, we take the per-period increase in demand between periods 1 and $N + 1$, and the per-period increase between periods 2 and $N + 2$, and so on; and then we take the average over those N values. To initialize the seasonal factors c_t, we estimate the seasonal factor for each period in the first two seasons, and then average them over those two seasons to obtain the initial seasonal factors:

$$c_{N+t} = \frac{1}{2}\left(\frac{D_t}{\sum_{j=1}^{N} D_j/N} + \frac{D_{N+t}}{\sum_{j=1}^{N} D_{N+j}/N}\right) \tag{2.15}$$

for $t = 1, \ldots, N$. Each denominator is the average demand in one season of the available data, so the fractions in the parentheses estimate the seasonal factor for the tth period in each season. The right-hand side as a whole averages these estimates over the two seasons. Finally, we estimate the base signal as $I_{2N} = D_{2N}/c_{2N}$. The first "real" forecast is y_{2N+1}.

This method is also sometimes known as *Winters's method* or the *Holt–Winters method* (Winters 1960).

□ **EXAMPLE 2.4**

The book store described in Example 2.1 also sells CDs. The total monthly demand of all CDs in the last year is given in Table 2.1. Note that in addition to the increasing trend, the monthly demand has a seasonal pattern with seasons of one quarter: the demand in the first and third months of a quarter is about half of that in the second month of the same quarter. This observation motivates us to use triple exponential smoothing for demand forecasting.

Since the observed pattern repeats quarterly, i.e., every 3 months, we choose $N = 3$. To initialize the seasonal factors, we extract the average over the first two quarters:

$$c_4 = \frac{1}{2}\left(\frac{D_1}{(D_1 + D_2 + D_3)/3} + \frac{D_4}{(D_4 + D_5 + D_6)/3}\right) = 0.71$$
$$c_5 = \frac{1}{2}\left(\frac{D_2}{(D_1 + D_2 + D_3)/3} + \frac{D_5}{(D_4 + D_5 + D_6)/3}\right) = 1.51$$
$$c_6 = \frac{1}{2}\left(\frac{D_3}{(D_1 + D_2 + D_3)/3} + \frac{D_6}{(D_4 + D_5 + D_6)/3}\right) = 0.79.$$

The base signal and slope are initialized with the first two quarters as

$$I_6 = D_6/c_6 = 21.36$$

$$S_6 = \frac{1}{3}\left(\frac{D_4 - D_1}{3} + \frac{D_5 - D_2}{3} + \frac{D_6 - D_3}{3}\right) = 1.85.$$

Then, we forecast D_7 and update the signals and seasonality with $\alpha = \beta = \gamma = 0.2$ as follows:

$$\begin{aligned} y_7 &= (I_6 + S_6)c_4 = 16.39 \\ I_7 &= 0.2\frac{D_7}{c_4} + 0.8(I_6 + S_6) = 23.91 \\ S_7 &= 0.2 \times (I_7 - I_6) + 0.8S_6 = 1.99 \\ c_7 &= 0.2 \times \frac{D_7}{I_7} + 0.8c_4 = 0.72. \end{aligned}$$

Repeating this procedure for the subsequent periods, we ultimately obtain the final results:

$$\begin{aligned} I_{12} &= 33.09 \\ S_{12} &= 1.86 \\ c_{12} &= 0.75 \quad c_{11} = 1.51 \quad c_{10} = 0.74 \\ y_{13} &= (I_{12} + S_{12})c_{10} = 25.90. \end{aligned}$$

So, our forecast for the demand in month 13 is 25.90. □

2.2.3 Linear Regression

Historical data can also be used to forecast demands by determining a cause–effect relationship between some independent variables and the demand. For instance, the demand for sales of a brand of laptop computer may heavily depend on the sales price and the features. A regression model can be developed which describes this relationship. The model can then be used to forecast the demand for laptops with a given price and a given set of features.

In linear regression, the model specification assumes that the dependent variable, Y, is a linear combination of the independent variables. For example, in *simple linear regression*, there is one independent variable, X, and two parameters, β_0 and β_1:

$$Y = \beta_0 + \beta_1 X. \tag{2.16}$$

Here, X and Y are random variables. For any given pair of observed variables x and y, we have

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{2.17}$$

where ϵ is a random error term. The objective of regression analysis is to estimate the parameters β_0 and β_1.

To build a regression model, we need historical data points—observations of both the independent variable(s) and the dependent variable. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n paired data observations for a simple linear regression model. The goal is to find values of β_0 and β_1 so that the line defined by (2.16) gives the best fit of the data. In particular, β_0 and β_1 are chosen to minimize the sum of the squared residuals, where the residual for

data point i is defined as the difference between the observed value of y_i and the predicted value of y_i obtained by substituting $X = x_i$ in (2.16). That is, we want to solve

$$\underset{\beta_0,\beta_1}{\text{minimize}} \sum_{i=1}^{n} \hat{e}_i^2 = \underset{\beta_0,\beta_1}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2, \tag{2.18}$$

where $\hat{e}_i$ is the residual for data point i. The optimal values of β_0 and β_1 are given by

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x} \tag{2.19}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.20}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x_i and y_i, respectively; r_{xy} is the sample correlation coefficient between x and y; and s_x and s_y are the sample standard deviations of x and y, respectively (see, e.g., Tamhane and Dunlop (1999)).

If the demands exhibit a linear trend over time, then we can use regression analysis to forecast the demand using the time period itself (rather than, say, price or features) as the independent variable. In this case, it can be shown (see, e.g., Nahmias (2005, Appendix 2-B)) that the optimal values of β_0 and β_1 are given by:

$$\hat{\beta}_1 = \frac{A_{xy}}{A_{xx}} \tag{2.21}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} D_i - \frac{\beta_1(n+1)}{2}, \tag{2.22}$$

where $D_1, \ldots, D_n$ are the observed demands and

$$A_{xy} = n\sum_{i=1}^{n} iD_i - \frac{n(n+1)}{2}\sum_{i=1}^{n} D_i \tag{2.23}$$

$$A_{xx} = \frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}. \tag{2.24}$$

According to the comparison by Carbonneau et al. (2008), linear regression often achieves better performance than moving average and trend methods.

□ **EXAMPLE 2.5**

Return to the sales data for *The TSP Mystery* in Table 2.1. Rather than using double exponential smoothing to forecast the demand for period 13, as we did in Example 2.3, we can instead use linear regression. Using either (2.19)–(2.20) or (2.21)–(2.22), we get $\hat{\beta}_0 = 10.88$ and $\hat{\beta}_1 = 1.89$. Therefore, the forecast for the demand in period 13 is

$$10.88 + 13 \cdot 1.89 = 35.46.$$

The observed data and the best-fit line are plotted in Figure 2.3.

□

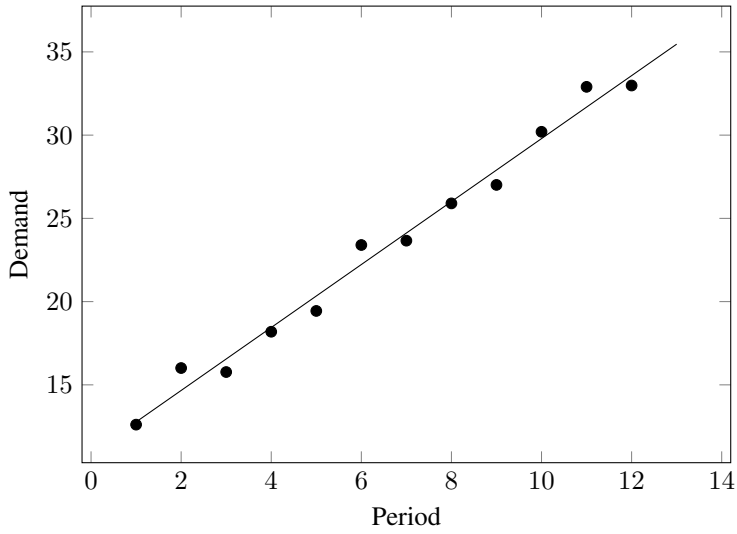

Figure 2.3 Observed demands for *The TSP Mystery* and best-fit line for Example 2.5.

2.3 FORECAST ACCURACY

2.3.1 MAD, MSE, and MAPE

At some point after a forecast is computed, the actual demand is observed, providing us with an opportunity to evaluate the quality of the forecast. The most basic measure of the forecast accuracy is the *forecast error*, denoted e_t, which is defined as the difference between the forecast for period t and the actual demand for that period:

$$e_t = y_t - D_t, \tag{2.25}$$

where y_t is a forecast obtained using any method and D_t is the actual observed demand.

Since the forecast and the demand are random variables, so is the forecast error; let μ_e and σ_e^2 denote its mean and variance, respectively. If the mean of the forecast error, μ_e, equals 0, we say the forecasting method is *unbiased*: It does not produce forecasts that are systematically either too low or too high. However, even an unbiased forecasting method can still be very inaccurate. One way to measure the accuracy is using the variance of the forecast error, σ_e^2. To compute μ_e or σ_e^2, however, we need to know the probabilistic process that underlies both the demands and the forecasts. Typically, therefore, we use performance measures based on sample quantities rather than population quantities.

Two of the most common such measures are the *mean absolute deviation (MAD)* and the *mean squared error (MSE)*, defined as follows:

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{2.26}$$

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2. \tag{2.27}$$

MSE is identical to the sample variance of the random forecast error e_t except for the denominator of the coefficient. MAD is sometimes preferred to MSE in real applications because it avoids the calculation of squaring, though modern spreadsheet and statistics packages can compute either performance measure easily. When the forecast errors are normally distributed, their standard deviation is often estimated as

$$\sigma_e \approx 1.25\text{MAD}. \tag{2.28}$$

This is useful when σ_e is required (e.g., for inventory optimization models—see Section 4.3.2.7), since, as previously noted, we do not typically know σ_e directly.

Note that both MAD and MSE are dependent on the magnitude of the values of demand; if we express the demands in different units (e.g., tons vs. pounds), the performance measures will change. By comparison, the *mean absolute percentage error (MAPE)* is independent of the magnitude of the demand values:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{D_t}\right| \times 100. \tag{2.29}$$

□ **EXAMPLE 2.6**

Table 2.2 gives the hypothetical actual demands for periods 13–24 for *An Inventory Story* from Example 2.1. It also gives the moving average forecasts for these periods (using $N = 5$), the single exponential smoothing forecasts for these periods (using $\alpha = 0.2$), and the corresponding forecast errors. Finally, at the end of the table are the performance measures (MAD, MSE, and MAPE) for each of the forecasting methods. In this case, the moving average has slightly smaller values of the performance measures and is therefore slightly more accurate. □

2.3.2 Forecast Errors for Moving Average and Exponential Smoothing

Assume that the demand is generated by the process

$$D_t = \mu + \epsilon_t, \tag{2.30}$$

where $\epsilon_t \sim N(0, \sigma^2)$. Since the demand process is stationary, either moving average or exponential smoothing is an appropriate forecasting method.

In a moving average of order N, the forecast y_t is given by (2.2). It follows that

$$\mu_e = \mathbb{E}[y_t - D_t] = \frac{1}{N}\sum_{i=t-N}^{t-1}\mathbb{E}[D_i] - \mathbb{E}[D_t] = \frac{1}{N}N\mu - \mu = 0.$$

Therefore, moving-average forecasts are unbiased when the demand is stationary.

We can also derive the variance of the forecast error, which can be expressed as

$$\begin{aligned}\sigma_e = \sqrt{\text{Var}[y_t - D_t]} &= \sqrt{\text{Var}[y_t] + \text{Var}[D_t]} \\ &= \sqrt{\frac{1}{N^2}\sum_{i=t-N}^{t-1}\text{Var}[D_i] + \text{Var}[D_t]}\end{aligned}$$

Table 2.2 Demands (D_t), forecasts (y_t), and forecast errors (e_t) for *An Inventory Story*, periods 13–24, for Example 2.6.

		Moving Average		Exponential Smoothing	
t	D_t	y_t	e_t	y_t	e_t
13	10.98	9.80	−1.18	9.98	−1.00
14	12.07	10.01	−2.06	10.18	−1.89
15	11.45	10.63	−0.82	10.56	−0.89
16	9.39	10.88	1.49	10.74	1.35
17	10.59	10.57	−0.02	10.47	−0.12
18	8.43	10.90	2.47	10.49	2.06
19	11.78	10.39	−1.39	10.08	−1.70
20	7.71	10.33	2.62	10.42	2.71
21	7.86	9.58	1.72	9.88	2.02
22	8.38	9.27	0.89	9.47	1.09
23	4.11	8.83	4.72	9.26	5.15
24	12.88	7.97	−4.91	8.23	−4.65
MAD			2.02		2.05
MSE			6.13		6.26
MAPE			25.97		26.85

$$= \sqrt{\frac{1}{N^2} N\sigma^2 + \sigma^2}$$
$$= \sigma \sqrt{\frac{1+N}{N}}.$$

Note that the second equality uses the fact that the forecast and demand in period t are statistically independent.

If forecasts are instead performed using exponential smoothing, one can show (see Problem 2.12) that

$$\mu_e = 0 \tag{2.31}$$

$$\sigma_e = \sigma \sqrt{\frac{2}{2-\alpha}}. \tag{2.32}$$

2.4 MACHINE LEARNING IN DEMAND FORECASTING

2.4.1 Introduction

We are in the age of big data. The huge volume of data generated every day, the high velocity of data creation, and the large variety of sources all make today's business information environment different than it was only a decade ago. Using data intelligently is key to business decision-making. A 2012 *Harvard Business Review* article notes: "Data-driven decisions are better decisions—it's as simple as that. Using big data enables managers to

decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management" (McAfee and Brynjolfsson 2012).

Fortunately, many businesses have access to large volumes of historical demand data that can help when forecasting future demands. In this section, we introduce some of the main machine learning techniques for demand forecasting. Compared with classical forecasting methods such as the time series methods discussed in Section 2.2, machine learning models often significantly increase prediction accuracy.

2.4.2 Machine Learning

In general, *machine learning* (ML) refers to a set of algorithms that can learn from and make predictions about data. These algorithms take data as inputs and generate predictions or decisions as outputs. Machine learning is closely related to *statistical learning*, which refers to a set of tools for modeling and understanding complex data sets (James et al. 2013). Machine learning and statistical learning have developed rapidly in recent years. Both techniques fall into the overall field of *data science*, which covers a wider range of topics, including database design and data visualization techniques.

One category of ML algorithms is called *supervised learning*, in which the historical data contain both inputs and outputs, and the learning algorithm learns to predict an output for a given set of inputs. For example, we might have historical data that contains the outdoor temperature and the number of glasses of lemonade that were sold on each day. The learning algorithm tries to infer the relationship between the two, so that for a given temperature, it can predict the number of glasses of lemonade that will be sold. Regression is a simple example. In contrast, *unsupervised learning* explores relationships and structures within the data without any known "ground truth" labels or outputs. For example, if we wish to partition consumers into market segments, we might use a clustering algorithm, which is a type of unsupervised learning. (See Friedman et al. (2001) or James et al. (2013) for further discussion of this dichotomy.) Demand forecasting falls into the category of supervised learning since we need to predict future demands (outputs) using historical demand data and other market information (inputs).

Common supervised learning methods include linear regression (and its nonlinear extensions), kernel methods, tree-based models, support vector machines (SVMs), and neural networks. Graphical models involving hidden Markov models (or, in their simplest form, mixture models) and Markov random fields also receive considerable attention. In the following subsections, we discuss the learning methods that are most commonly applied to demand forecasting.

2.4.2.1 Linear Regression

Linear regression is a very simple supervised learning method. It assumes that the output Y is linear in the inputs $X_1, X_2, \ldots, X_p$, where p is the number of distinct input variables (also called *predictors* or *features*):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2.33}$$

For particular values of the inputs and outputs, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \tag{2.34}$$

where ϵ is a random error term. The β_js are coefficients that need to be estimated from data. If $p = 1$, then we have *simple linear regression*, which we discuss in Section 2.2.3.

Table 2.3 Snippet of historical data on demand for baseball jerseys for Examples 2.7–2.9.

Batting Avg.	Games Won	Years in Majors	Demand (Cases)
0.274	68	1	14.3
0.332	150	11	28.7
0.262	79	12	17.6
0.396	127	8	26.0
0.262	156	4	27.1
0.280	142	7	26.0
0.112	75	10	14.7
0.429	82	0	19.2
0.259	88	7	18.1
0.302	95	6	19.4

(In Section 2.2.3, we focused on the use of time as the independent variable in order to predict demands as a function of time. Here, our independent variables can be any feature.)

The most common way to obtain the β_js is *least squares*, which seeks to find the minimizer of the sum of the squares of the residuals. (Recall from Section 2.2.3 that the residual for data point i is the difference between the observed and predicted values of y_i.) The derived estimated coefficients are denoted $\hat{\beta}_j$. Then we can make predictions on new inputs by using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p, \tag{2.35}$$

where $\hat{y}$ is our predicted value for the output, given the observed values $\{x_1, x_2, \ldots, x_p\}$ of the inputs.

□ **EXAMPLE 2.7**

A sports apparel retailer sells jerseys (T-shirts) with the names of major league baseball players stitched onto the back. The retailer believes that the demand for a given player's jersey depends on his batting average last year, the number of games his team won last year, and the number of years the player has been playing in the major leagues. Therefore, the retailer keeps careful records of these statistics, as well as the demand for jerseys, for each player. Last year's records for 300 players can be found in the file `jerseys.xlsx`, the first few rows of which are reproduced in Table 2.3. Demands are expressed in cases sold last year. (In baseball, batting averages are expressed as decimals between 0 and 1.)

The retailer wishes to predict the demand for this year's jerseys using the historical data. Let X_1 represent batting average, X_2 represent games won, and X_3 represent years in majors. Using regression, we find that

$$\hat{\beta}_0 = -0.0651$$
$$\hat{\beta}_1 = 18.0430$$
$$\hat{\beta}_2 = 0.1403$$
$$\hat{\beta}_3 = 0.1831.$$

For example, if Roy Hobbs had a 0.292 average last year, his team won 95 games, and he has been in the major leagues for 4 years, the demand for his jersey this year can be predicted as

$$\hat{y} = -0.0651 + 18.0430 \times 0.292 + 0.1403 \times 95 + 0.1831 \times 4 = 19.2644.$$

□

Although the linear regression model assumes a linear relationship between the output and the inputs, we can model nonlinear relationships by introducing basis functions and splines. When the number of predictors is large, we can utilize shrinkage methods such as least absolute shrinkage and selection operator (LASSO) and ridge regression. In general, linear regression is a simple but strong learning method.

2.4.2.2 Tree-based models Tree-based models use decision trees to make predictions for a given set of inputs. They can be applied both to regression problems (in which the outputs are continuous) and to classification problems (in which the outputs are categorical). The trees used for these two types of problems are referred to as *regression trees* and *classification trees*, respectively. In demand forecasting, regression trees have received more attention because of their simplicity and interpretability.

A regression tree divides the space of input variables, i.e., the set of possible values of $X_1, X_2, \ldots, X_p$, into distinct and nonoverlapping regions and assigns a single output, c_k, to each region k. If a given input $x_1, x_2, \ldots, x_p$ falls into region k, then the demand forecast y for that input is equal to c_k. The c_k values are determined simply by averaging the observations in the historical data that fall into that region.

The goal is to choose the partition strategy that minimizes the sum of squares of the residuals, similar to linear regression. However, in practice, the number of possible partitions may be too large to enumerate. Therefore, it is common to use a binary splitting method called *recursive partitioning*, which generates two regions from the original region at each iteration. For the purposes of prediction, the size of the tree is limited by a pruning process. A single tree may not perform well due to high variance of the forecast, so researchers have developed methods that combine several trees to enhance the prediction performance. These include *random forests*, *bagging*, and *boosting*.

Tree-based models are used widely in demand forecasting for many industries. For example, Ferreira et al. (2015) apply regression trees with bagging to predict the demand of new styles for an online retailer. They show that tree-based models outperform linear regression and some nonlinear regression models consistently. Ali et al. (2009) develop regression trees to predict stock-keeping unit (SKU) sales for a European grocery retailer. They incorporate information about current promotions when constructing regression trees and show that regression trees provide better accuracy than linear regression and SVMs.

□ **EXAMPLE 2.8**

Return to the baseball-jersey data set from Example 2.7. Figure 2.4 shows one possible regression tree for this data set. For example, we would predict a demand of 23.5 cases for Roy Hobbs (who has a 0.292 batting average, has been in the major leagues for 4 years, and is on a team that has won 95 games).

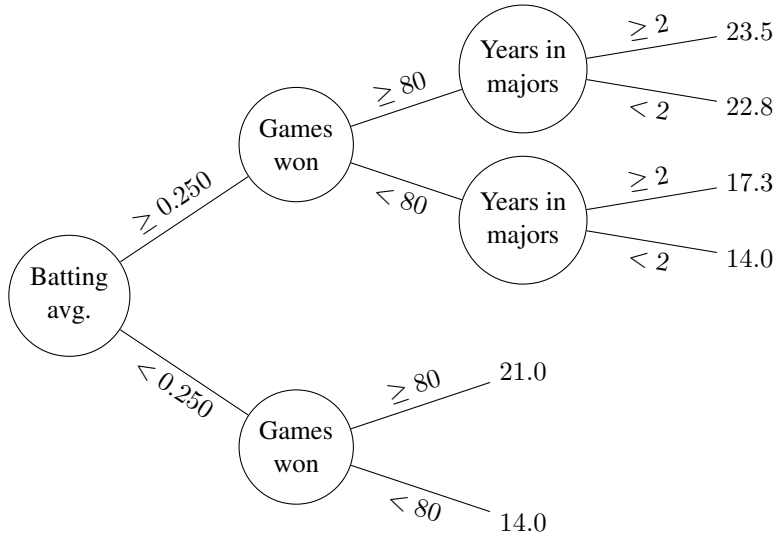

Figure 2.4 Regression tree for baseball jerseys for Example 2.8.

Of course, there are many possible regression trees for this data set, and the figure gives only one example. Better ones can be found using the recursive partitioning method.

□

2.4.2.3 Support vector machines SVMs are designed to partition the space of input variables into two regions, i.e., to make a binary prediction about a given output based on which region a given input vector falls into. The partition is accomplished by finding a separating hyperplane. In particular, assuming that the training data set is linearly separable, the optimal separating hyperplane is found by solving the following optimization problem:

$$\underset{\beta_0,\beta}{\text{minimize}} \quad ||\beta||_2^2 \tag{2.36}$$

$$\text{subject to} \quad y^i(\mathbf{x}^i \cdot \beta + \beta_0) \geq 1 \quad \forall i = 1, 2, \ldots, N, \tag{2.37}$$

where N is the number of observations, y^i is the binary output ($y^i \in \{0, 1\}$) for observation i, $\mathbf{x}^i \in \mathbb{R}^p$ is the vector of input variables for observation i, and $\cdot$ denotes dot product. This is also called a *maximum margin classifier*, where the *margin* is defined as $\frac{1}{||\beta||_2}$. The optimal values of the vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and the scalar β_0 characterize the separating hyperplane. For a given input vector $x_1, \ldots, x_p$, we predict an output value of 1 if $\mathbf{x}^i \cdot \boldsymbol{\beta} + \beta_0 > 0$ and a value of 0 otherwise.

For example, suppose we wish to predict which customers will purchase a product based on their age, income, and money spent at the store in the past year. We code each customer in the historical data with a 1 or 0 depending on whether they purchased the product, then solve (2.36)–(2.37) to find the hyperplane that does the best job of separating the 1s from

the 0s. For each new customer, we simply calculate $\mathbf{x}^i \cdot \boldsymbol{\beta} + \beta_0$ and make a prediction accordingly.

SVMs can be generalized to allow nonlinearities by mapping the input space into a high-dimensional space using *kernel functions*. In essence, this allows the region to be partitioned using a surface that is not linear, i.e., is not a hyperplane. Popular choices of kernel functions include polynomials and radial basis functions (RBFs).

Since SVMs can be used to make binary predictions, they can be used to predict whether a given customer will purchase a product. They can also be used to forecast the demand as a quantity using *support vector regression* (SVR), an adaptation of the SVM approach to regression problems using kernel functions. SVR is among the best machine learning methods for supply chain demand forecasting (Carbonneau et al. 2008).

□ **EXAMPLE 2.9**

For the baseball-jersey data set from Example 2.7, let us first use SVM to predict whether the demand for a given player's jerseys will be greater than or equal to 25 cases this year. We can label the historical data by assigning $y^i = 1$ to players whose jerseys had a demand greater than or equal to 25 and $y^i = 0$ for those who did not. Solving the SVM optimization problem[1] results in the solution $\boldsymbol{\beta} = (4.5879, 0.0745, 0.1154)$ and $\beta_0 = -12.1620$. In other words, if

$$-12.1620 + (4.5879, 0.0745, 0.1154) \cdot (x_1, x_2, x_3) > 0,$$

then we predict that the demand will be greater than or equal to 25. For Roy Hobbs, who has an input vector of $\mathbf{x}^i = (0.292, 95, 4)$, we have

$$-12.1620 + (4.5879, 0.0745, 0.1154) \cdot (0.292, 95, 4) = -3.2832,$$

so we predict that Roy will *not* sell more than 25 cases of jerseys this year.

Next, we can use an SVR model to predict the demand for Roy Hobbs jerseys explicitly. Using MATLAB's `fitrsvm` function, we obtain SVR coefficients of $\boldsymbol{\beta} = (13.8451, 0.1387, 0.1932)$ and $\beta_0 = 1.1436$. Therefore, we can predict the demand for Roy Hobbs jerseys as

$$1.1436 + (13.8451, 0.1387, 0.1932) \cdot (0.292, 95, 4) = 19.1357 \text{ cases.}$$

(Note that the SVM and SVR optimization problems are nonconvex and typically have multiple optima. Your results might differ if you use a different implementation to solve the same problem.) □

2.4.2.4 Neural Networks A neural network consists of several *nodes*, also called *neurons*, arranged into *layers*. The first layer of nodes represents the inputs (the X_i values); the last layer represents the outputs (the Y value); and one or more layers in between, called *hidden layers*, process the information from the input layer and perform the actual computation of the network. (See Figure 2.5.) Neural networks have been used extensively for classification problems such as image and speech processing, where the

[1] We did not use (2.36)–(2.37), but rather a modified formulation, since the training data set in this example is not linearly separable. We used MATLAB's `fitcsvm` function to do the optimization.

goal is to determine what sort of physical or linguistic object the inputs represent. But neural networks can and have been successfully applied to regression-type problems such as demand forecasting.

The central idea behind neural networks is that in each layer (except the first), we extract linear combinations of the inputs from the previous layer as derived features, and then model the output as a nonlinear function of these features. For example, in a typical network with a single hidden layer with M nodes, each hidden-layer node $m = 1, \ldots, M$ calculates the derived feature

$$Z_m = \sigma(\alpha_{0m} + \boldsymbol{\alpha}_m^T \mathbf{X}), \tag{2.38}$$

where $\mathbf{X}$ is the vector of inputs, α_{0m} is a scalar, $\boldsymbol{\alpha}_m$ is a vector with p elements (one per input feature), and $\sigma(\cdot)$ is a nonlinear function called the *activation function*. Note that the term inside the $\sigma(\cdot)$ is a linear combination of the inputs plus a constant. Typical activation functions include the sigmoid function and the ReLU function. The Z_m are also called *hidden units* since they are not directly observed. Once the hidden units are calculated by the hidden-layer nodes, the output Y is modeled as a function of the hidden units:

$$Y = g(Z_1, \ldots, Z_M), \tag{2.39}$$

where $g(\cdot)$ is a (possibly nonlinear) function.

The key challenge in fitting a neural network model is the determination of the weights α_{0m} and α_m. This is usually done using some sort of algorithm that modifies the weights as the network "learns" right and wrong answers. The most common such algorithm is known as *backpropagation*, which calculates gradients with respect to the weights; another method (such as *gradient descent*) is then used to update the weights. Determining these weights—sometimes referred to as training the network—can be computationally intensive. However, once the network is trained, generating an output value for a new set of inputs is extremely efficient. (For further details, see, e.g., Friedman et al. (2001).)

Some neural networks contain multiple hidden layers, not just one; this can improve the accuracy of the network's predictions but makes the network harder to train. Such *deep neural networks* have led to huge advances in machine learning, with great successes not only in classification and prediction problems such as image processing and demand forecasting, but also, when coupled with reinforcement learning (RL), in solving decision problems such as those in board games; one famous example is Google DeepMind's AlphaGo program, which beat the world-champion (human) Go player in 2016.

Carbonneau et al. (2008) test two different types of neural networks on demand forecasting and conclude that neural networks perform better than traditional methods. Venkatesh et al. (2014) combine neural networks with clustering to predict demand for cash at automatic teller machines (ATMs). They find that their model increases the prediction accuracy substantially.

2.5 DEMAND MODELING TECHNIQUES

As the pace of technology accelerates, companies are introducing new products faster and faster to stay competitive. There is a diffusion process associated with the demand for any new product, so companies need to plan the timing and quantity of new product releases

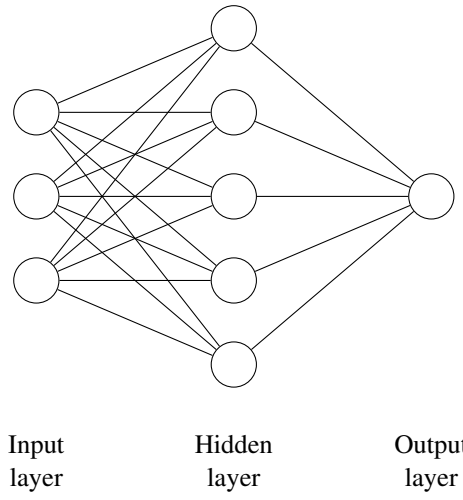

Figure 2.5 A simple neural network.

carefully to match supply and demand as closely as possible. To do so, they need to understand the life cycles and demand dynamics of their products.

One of the authors has worked with a high-tech company in China. The company was complaining about their very inaccurate demand forecasts, which led to excess inventory valued at approximately $25 million. The author was invited to give lectures on demand forecasting and inventory management. The first day's lecture focused on the classical time-series demand forecasting techniques discussed earlier in this chapter. The reaction from the company's forecasting team was lukewarm. They were already quite familiar with these techniques and had tried hard to make them work, unsuccessfully. It turns out that classical forecasting techniques did not work well with the company's highly variable, short-life-cycle products, so the firm introduced products at the wrong times in the wrong quantities. The forecasting team's reaction was quite different when the author discussed the Bass diffusion model, the leading-indicator method, and choice models, which are designed to account for short life cycles and other important factors. We discuss each of these methods in detail in the following sections. (As a postscript, the company reported more than a 50% increase in sales about one and a half years after they improved their forecasting techniques, partially due to the fact that money was being invested in a better mix of products.)

2.6 BASS DIFFUSION MODEL

The sales patterns of new products typically go through three phases: rapid growth, maturity, and decline. The *Bass diffusion model* (Bass 1969) is a well-known parametric approach for estimating the demand trajectory of a single new product over time. Bass's basic three-parameter model has proved to be very effective in delivering accurate forecasts and insights for a huge variety of new product introductions, regardless of pricing and advertising decisions. The model forecasts well even when limited or no historical data

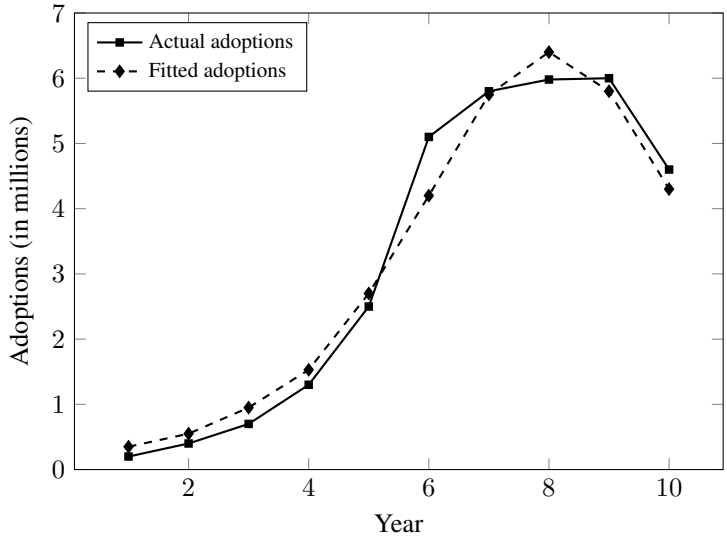

Figure 2.6 Color TVs in the 1960s: Forecasts from Bass model and actual demands. Reprinted by permission, Bass, Empirical generalizations and marketing science: A personal view, *Marketing Science*, 14(3), 1995, G6–G19. ©1995, the Institute for Operations Research and the Management Sciences (INFORMS), 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA.

are available. For example, Figure 2.6 depicts demand data (forecast and actual) for the introduction of color television sets in the 1960s.

The premise of the Bass model is that customers can be classified into *innovators* and *imitators*. Innovators (or *early adopters*) purchase a new product without regard to the decisions made by other individuals. Imitators, on the other hand, are influenced in the timing of their purchases by previous buyers through word-of-mouth communication. Refer to Figure 2.7 for an illustration. The number of innovators decreases over time, while the number of imitators purchasing the product first increases, and then decreases. The goal of the Bass model is to characterize this behavior in an effort to forecast the demand. It mathematically characterizes the word-of-mouth interaction between those who have adopted the innovation and those who have not yet adopted it. Moreover, it attempts to predict two important dimensions of a forecast: how many customers will eventually adopt the new product, and when they will adopt. Knowing the timing of adoptions is important as it can guide the firm to smartly utilize resources in marketing the new product. Our analysis of this model is based on that of Bass (1969).

2.6.1 The Model

The Bass model assumes that $P(t)$, the probability that a given buyer makes an initial purchase at time t given that she has not yet made a purchase, is a linear function of the number of previous buyers; that is,

$$P(t) = p + \frac{q}{m} D(t), \tag{2.40}$$

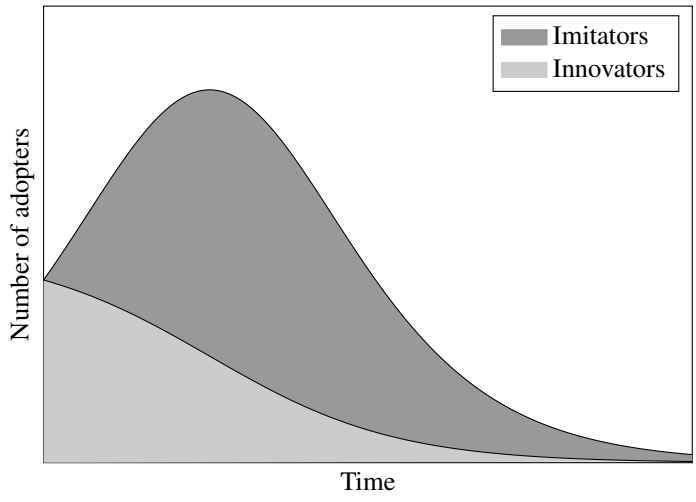

Figure 2.7 Bass diffusion curve.

where $D(t)$ is the cumulative demand by time t. Equation (2.40) suggests that two factors will influence the probability that a customer makes a purchase at time t. The first factor is the *coefficient of innovation*, denoted p, which is a constant, independent of how many other customers have adopted the innovation before time t. The second factor, $\frac{q}{m}D(t)$, measures the "contagion" effect between the innovators and the imitators and is proportional to the number of customers who have already adopted by time t. The parameters q and m represent the *coefficient of imitation* and the *market size*, respectively. We require $p < q$. In fact, usually $p \ll q$; for example, $p = 0.03$ and $q = 0.38$ have been reported as average values (Sultan et al. 1990).

We assume that the time index, t, is measured in years. Of course, any time unit is possible, but the values we report for p and q implicitly assume that t is measured in years.

Let $d(t)$ be the derivative of $D(t)$, i.e., the demand *rate* at time t. Using Bayes' rule, one can show that

$$P(t) = \frac{d(t)}{m - D(t)}. \tag{2.41}$$

(See Section 2.6.2 for a derivation of the analogous equation in the discrete-time model.) Combining (2.40) and (2.41), we have

$$d(t) = \left(p + \frac{q}{m}D(t)\right)(m - D(t)). \tag{2.42}$$

Our goal is to characterize $D(t)$ so that we can understand how the demand evolves over time. To a certain extent, (2.42) does this, but (2.42) is a differential equation; it expresses $D(t)$ in terms of its derivative. Our preference would be to have a closed-form expression for $D(t)$. Fortunately, this is possible:

Theorem 2.1

$$D(t) = m\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}} \tag{2.43}$$

$$d(t) = \frac{mp(p+q)^2 e^{-(p+q)t}}{\left(p + qe^{-(p+q)t}\right)^2} \tag{2.44}$$

Proof. Omitted. ∎

As a corollary, one can determine the time at which the demand rate peaks, and the demand rate and cumulative demand at that point:

Corollary 2.2 *The peak demand occurs at time*

$$t^* = \frac{1}{p+q} \ln\left(\frac{q}{p}\right). \tag{2.45}$$

The demand rate and cumulative demand at time t^* are given by

$$d(t^*) = \frac{m(p+q)^2}{4q} \tag{2.46}$$

$$D(t^*) = \frac{m(q-p)}{2q}. \tag{2.47}$$

Proof. Omitted; see Problem 2.17. ∎

If p is very small, then the demand growth occurs slowly, whereas if p and q are large, sales take off rapidly and fall off quickly after reaching their maximum. Note that the formulas in Corollary 2.2 are only well defined if $q > p$, which we previously assumed to be true. If, instead, $q < p$, then the innovation effects will dominate the imitation effects, and the peak demand will occur immediately upon the introduction of the product and will decline thereafter. In summary, by varying the values of p and q, we can represent many different patterns of demand diffusion.

□ **EXAMPLE 2.10**

The bookstore manager from Example 2.3 now wishes to model the demand for a third book, *The Case of the Violated Constraint*, which is expected to be a best-seller but whose sales will taper off after their peak. The bookstore's marketing department has estimated that the sales of the book will follow a Bass diffusion process with parameters $p = 0.05$, $q = 0.3$, and $m = 2700$, which are calculated assuming that the time index is measured in weeks (not years).

At what time will the sales of *The Case of the Violated Constraint* reach their peak, and what will the demand rate be at that time? How many copies of the book will have been sold by that point? What will the demand rate be at week 20, and how many copies will have been sold by that point?

From Corollary 2.2, we have

$$t^* = \frac{1}{0.05 + 0.3} \ln\left(\frac{0.3}{0.05}\right) = 5.12,$$

so the peak occurs during week 5. Moreover,

$$d(t^*) = \frac{2700(0.05 + 0.3)^2}{4 \cdot 0.3} = 27.63$$

$$D(t^*) = \frac{2700(0.3 - 0.05)}{2 \cdot 0.3} = 1125.00$$

and, from (2.43)–(2.44),

$$D(20) = 2700 \frac{1 - e^{-(0.05+0.3)\cdot 20}}{1 + \frac{0.3}{0.05} e^{-(0.05+0.3)\cdot 20}} = 2682.86$$

$$d(20) = \frac{2700 \cdot 0.05(0.05 + 0.3)^2 e^{-(0.05+0.3)\cdot 20}}{\left(0.05 + 0.3 e^{-(0.05+0.3)\cdot 20}\right)^2} = 5.97.$$

Therefore, at the time of peak demand, the demand rate will be 27.63 books per week, and 1125 books will have been sold. At week 20, the demand rate will be 5.97 books per week, and 2682.86 (or 2683) books will have been sold. □

Seasonal influence factors can be incorporated into the Bass framework. Kurawarwala and Matsuo (1996) present a growth model to forecast demand for short-life-cycle products that is motivated by the Bass diffusion model. They use α_t to denote the seasonal influence parameter at time t, given as a function with a periodicity of 12 months. Their proposed seasonal growth model is characterized by the following differential equation:

$$d(t) = \left(p + \frac{q}{m} D(t)\right)(m - D(t))\alpha_t, \tag{2.48}$$

where $D(t)$ is the cumulative demand by time t ($D(0) \equiv 0$), $d(t)$ is its derivative, and m, p, and q are the scale and shape parameters, which are analogous to the parameters in the Bass diffusion model. This is identical to (2.42) except for the multiplier α_t.

Integrating (2.48), we get the cumulative demand $D(t)$ as follows:

$$D(t) = m \left[\frac{1 - e^{-(p+q)\int_0^t \alpha_\tau d\tau}}{1 + \frac{q}{p} e^{-(p+q)\int_0^t \alpha_\tau d\tau}} \right]. \tag{2.49}$$

When $\alpha_t = 1$ for all t, (2.49) reduces to (2.43) from Bass's original model.

2.6.2 Discrete-Time Version

A discrete-time version of the Bass model is available. In this case, d_t represents the demand in period t, and D_t represents the cumulative demand up to period t. Let P_t be the probability that a customer buys the product in period t given that she did not buy it in periods $1, \ldots, t-1$. Bayes' rule says that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Here, let A represent "customer buys in t" and B represent "customer didn't buy in $1, \ldots, t-1$." Then

$$\mathbb{P}(A|B) = \frac{1 \cdot \frac{d_t}{m}}{1 - \frac{D_t}{m}} = \frac{d_t}{m - D_t}.$$

(Note the similarity to (2.41), which is for continuous time.) Then the discrete-time analogue of (2.42) is

$$d_t = \left(p + \frac{q}{m} D_{t-1}\right)(m - D_{t-1}), \tag{2.50}$$

where $D_0 \equiv 0$.

2.6.3 Parameter Estimation

The Bass model is heavily driven by the parameters m, p, and q. In this section, we briefly discuss how these parameters may be estimated.

If historical data are available, we can estimate the parameters p, q, and m by first finding the least-squares estimates of the parameters a, b, and c in the following linear regression model:

$$d_t = a + bD_{t-1} + cD_{t-1}^2 \qquad t = 2, 3, \ldots.$$

Note that this model uses the discrete-time version of the Bass model (in which we observe demands d_t and calculate cumulative demands D_t) since, in practice, we observe discrete demand quantities rather than a continuous demand function. After finding a, b, and c using standard regression analysis, the parameters of the Bass model can be determined as follows:

$$m = \frac{-b - \sqrt{b^2 - 4ac}}{2c} \tag{2.51}$$

$$p = \frac{a}{m} \tag{2.52}$$

$$q = -mc. \tag{2.53}$$

However, because the Bass model is typically used for new products, in most cases historical data are not available to estimate the parameters. Instead, m is typically estimated qualitatively, using judgment or intuition from management about the size of the market, market research, or the Delphi method. In some markets these estimates can be rather precise. For instance, the pharmaceutical industry is known for their accurate demand estimates, which derive from abundant data regarding the incidence of diseases and ailments (Lilien et al. 2007). The parameters p and q tend to be relatively consistent within a given industry, so these can often be estimated from the diffusion patterns of similar products. Lilien and Rangaswamy (1998) provide industry-specific data for a wide range of industries. (See Table 2.4 for some examples.)

2.6.4 Extensions

After more than half a century, the Bass model is still actively used in demand forecasting and production planning. Sultan et al. (1990), Mahajan et al. (1995), and Bass (2004) provide broad overviews of these applications. The original model has also been extended in a number of ways. Ho et al. (2002) provide a joint analysis of demand and sales dynamics when the supply is constrained, and thus the usual word-of-mouth effects are mitigated. Their analysis generalizes the Bass model to include backorders and lost sales and describes the diffusion dynamics when the firm actively makes supply-related decisions to influence the diffusion process. Savin and Terwiesch (2005) describe the demand dynamics of two new products competing for a limited target market, generalizing the innovation and imitation effects in Bass's original model to account for this competition. Schmidt and Druehl (2005) explore the influence of product improvements and cost reductions on the new-product diffusion process. Ke et al. (2013) consider the problem of extending a product line while accounting for both inventory (supply) and diffusion (demand). The model determines whether and when to introduce the line extension and the corresponding production quantities. Islam (2014) uses the Bass model (as well as experimental discrete

Table 2.4 Bass model parameters. Adapted with permission from Lilien and Rangaswamy, *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning*, Addison-Wesley, with permission obtained from Pearson, 1998, p. 201.

Product	p	q
Cable TV	0.100	0.060
Camcorder	0.044	0.304
Cellular phone	0.008	0.421
CD player	0.157	0.000
Radio	0.027	0.435
Home PC	0.121	0.281
Hybrid corn	0.000	0.797
Tractor	0.000	0.234
Ultrasound	0.000	0.534
Dishwasher	0.000	0.179
Microwave	0.002	0.357
VCR	0.025	0.603

choice data—see Section 2.8) to predict household adoption of photovoltaic (PV) solar cells.

2.7 LEADING INDICATOR APPROACH

Product life cycles are becoming shorter and shorter, so it is difficult to obtain enough historical data to forecast demands accurately. One idea that has proven to work well in such situations is the use of *leading indicators*—products that can be used to predict the demands of other, later products because the two products share a similar demand pattern. This approach was introduced by Aytac and Wu (2013) and by Wu et al. (2006), who describe an application of the method at the semiconductor company Agere Systems.

The approach is applied in situations in which a company introduces many related products, such as multiple varieties of semiconductors, cellular phones, or grocery items. The idea is first to group the products into clusters so that all of the products within a cluster share similar attributes. There are several ways to perform this clustering. If one can identify a few demand patterns that all products follow, then it is natural simply to group products sharing the same pattern into the same cluster. For instance, after examining demand data for about 3500 products, Meixell and Wu (2001) find that the products follow six basic demand patterns (i.e., diffusion curves from the Bass model in Section 2.6) and can be grouped into these patterns using statistical cluster analysis. Wu et al. (2006), on the other hand, focus on exogenously defined product characteristics, such as resources, technology group, or sales region, and group the products that have similar characteristics into the same cluster.

The goal is then to identify some potential leading-indicator products within each cluster. A product is a leading indicator if the demand pattern of this product will likely be approximately repeated later by other products in the same cluster. For example, Figure 2.8 depicts the demand for a leading indicator product (solid line) and the total demand for all of the products in the cluster (dashed line). If the leading indicator curve is shifted to the

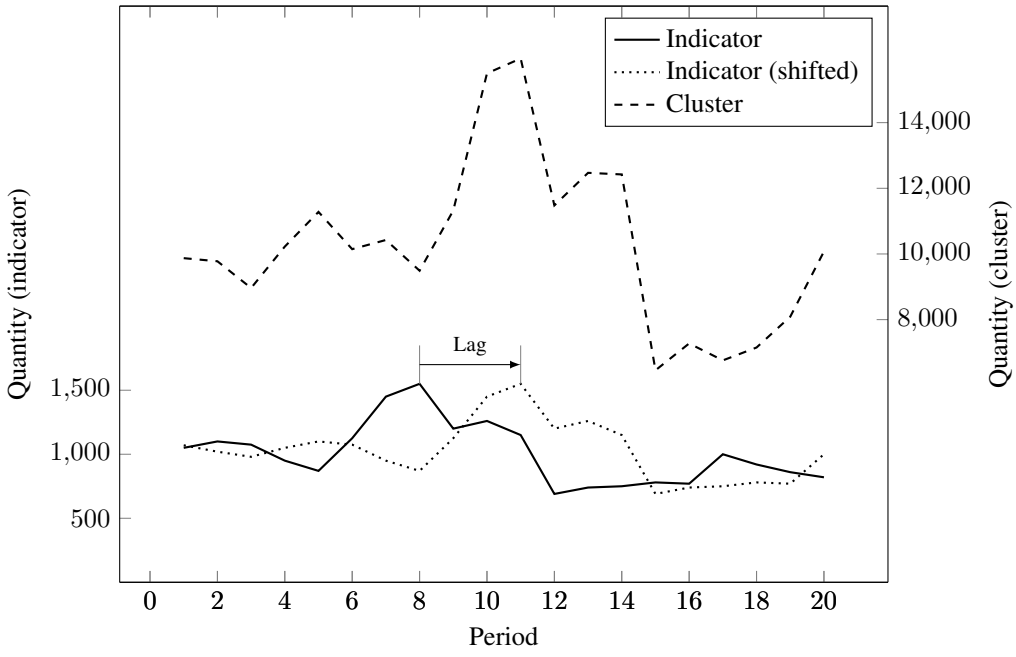

Figure 2.8 An example of a leading-indicator product.

right by three periods (the "lag"), the two curves share a similar structure. Therefore, the leading indicator product provides some basis for predicting the demand of the rest of the products in the cluster. Even though all of the products are on the market simultaneously, the lag provides enough time so that supply chain planning for the products in the cluster can take place based on the forecasts provided by the leading indicator. Of course, correctly identifying the leading indicator is critical.

Wu et al. (2006) suggest the following procedure to identify a leading indicator within a given cluster. Let C be the set of products, i.e., the cluster. Each product $i \in C$ will be treated as a potential leading indicator. Suppose we have historical demand data through period T. Let D_{it} be the observed demand for product i in period t, and let D_t be the total demand for the entire cluster in period t, $t = 1, \ldots, T$. Then leading indicators can be identified using Algorithm 2.1. In line 4 of the algorithm, the correlation ρ_{ik} measures how well the demand of item i over the time interval $[1, T-k]$ predicts the demand of the cluster over $[k+1, T]$.

Once a leading indicator i with time lag k is identified as having a satisfactory correlation coefficient ρ_{ik}, we can forecast the demand for the rest of the product cluster using the demand history from the leading indicator as follows:

1. Regress the demand time-series of product cluster C (excluding i) over $[k+1, T]$ against the time series of the leading indicator over $[1, T-k]$ using the model

$$D_t^{-i} = \beta_0 + \beta_1 D_{i,t-k} \tag{2.54}$$

and determine the optimal regression parameters β_0 and β_1.

Algorithm 2.1 Leading-indicator identification

choose $k_{\min}$, $k_{\max}$, $\rho_{\min}$ ▷ Initialization

for all $i \in C, k \in \{k_{\min}, \ldots, k_{\max}\}$ **do** ▷ Correlation calculation

shift product-i demands by k periods

calculate ρ_{ik} (correlation between "i lag k" and $C \setminus \{i\}$) as

$$\rho_{ik} \leftarrow \frac{\sum_{t=k+1}^{T}(D_{i,t-k} - \bar{D}_i)(D_t^{-i} - \bar{D}^{-i})}{\sqrt{\sum_{t=k+1}^{T}(D_{i,t-k} - \bar{D}_i)^2 \sum_{t=k+1}^{T}(D_t^{-i} - \bar{D}^{-i})^2}},$$

where D_{it} is the observed demand for product i in period t, $\bar{D}_i$ is its mean over the time interval $[k+1, T]$, D_t^{-i} is the total demand for all products in the cluster excluding i in period t, and $\bar{D}^{-i}$ is its mean over the time interval $[k+1, T]$

end for

for all $i \in C, k \in \{k_{\min}, \ldots, k_{\max}\}$ **do** ▷ Identification of leading indicators

if $\rho_{ik} \geq \rho_{\min}$ **then**

label i as leading indicator with lag k

end if

end for

if any leading indicators were found **then**

return leading indicators and corresponding clusters

else

for all C **do** ▷ Reclustering

using statistical cluster analysis, subdivide C into clusters based on statistical demand patterns; attributes can include demand mean or SD, shipment frequency, etc.

end for

go to 2

end if

2. For a given month $t > T$ (that is, a month for which we do not have historical data but whose demand we wish to forecast), generate the forecast for the cluster, $\tilde{D}_t^{-i}$, using the time series of the leading indicator i from k periods earlier:

$$\tilde{D}_t^{-i} = \beta_0 + \beta_1 D_{i,t-k}. \tag{2.55}$$

2.8 DISCRETE CHOICE MODELS

2.8.1 Introduction to Discrete Choice

In economics, *discrete choice models* involve choices between two or more discrete alternatives. For example, a customer chooses which of several competing products to buy; a firm decides which technology to use in production; or a passenger chooses which transportation mode to travel by. The set of choices is assumed to be discrete, and the corresponding models are therefore called discrete choice models. (A related set of models, called continuous choice models, assume that the range of choices is continuous. Although these models are not the focus of our discussion, many of the concepts that we describe below are easily transferable to continuous choice models. In fact, discrete choices generally reveal less information about the choice process than continuous ones, so the econometrics of discrete choice is usually more challenging.)

The idea behind discrete choice models is to build a statistical model that predicts the choice made by an individual based on the individual's own attributes as well as the attributes of the available choices. For example, a student's choice of which college to attend is determined by factors relating to the student, including his or her career goals, scholarly interests, and financial situation, as well as factors relating to the colleges, including their reputations and locations. Choice models attempt to quantify this relationship statistically. Rather than modeling the attributes (career goals, scholarly interests, etc.) as independent variables and then predicting the choice as the dependent variable, choice models are at the aggregate (population) level and assume that each decision-maker's preferences are captured implicitly by that model.

At first, it may seem that discrete choice models mainly deal with "which"-type rather than "how many"-type decisions, unlike the other forecasting and demand-modeling techniques described in this chapter. However, discrete choice models can be and have been used to forecast quantities, such as the number and duration of phone calls that households make (Train et al. 1987); the demand for electric cars (Beggs et al. 1981) and mobile telephones (Ida and Kuroda 2009); the demand for planned transportation systems, such as highways, rapid transit systems, and airline routes (Train 1978, Ramming 2001, Garrow 2010)); and the number of vehicles a household chooses to own (McFadden 1984). Choice models estimate the probability that a person selects a particular alternative. Thus, aggregating the "which" decision across the population will give answers to the "how many" questions and can be very useful for forecasting demand.

Discrete choice models take many forms, including binary and multinomial logit, binary and multinomial probit, and conditional logit. However, there are several features that are common to all of these models. These include the way they characterize the choice set, consumer utility, and the choice probabilities. We briefly describe each of these features next. (See Train (2009) for more details about these features.)

The Choice Set: The *choice set* is the set of options that are available to the decision-maker. The alternatives might represent competing products or services, or any other options or items among which the decision-maker must choose. For a discrete choice model, the set of alternatives in the choice set must be *mutually exclusive*, *exhaustive*, and *finite*. The first two requirements mean that the set must include all possible alternatives (so that the decision-maker necessarily does make a choice from within the set) and that choosing one alternative means not choosing any others (so one alternative from the set dominates all other options for the decision-maker). The third requirement distinguishes discrete choice analysis from, say, linear regression analysis in which the dependent variable can (theoretically) take an infinite number of values.

Consumer Utility: Suppose there are N decision-makers, each of whom must select an alternative from the choice set I. A given decision-maker n would obtain a certain level of *utility* from alternative $i \in I$; this utility is denoted U_{ni}. Discrete choice models usually assume that the decision-maker is a utility maximizer. That is, he will choose alternative i if and only if $U_{ni} > U_{nj}$ for all $j \in I$, $j \neq i$.

If we know the utility values U_{ni} for all $n \in N$ and all $i \in I$, then it will be very easy for us to calculate which alternative decision-maker n will choose (and therefore to predict the demand for each alternative). However, since in most cases we do not know the utility values perfectly, we must estimate them. Let V_{ni} be our estimate of alternative i's utility for decision-maker n. (The V_{ni} values are called *representative utilities*. We omit a discussion about how these might be calculated; see, for example, Train (2009).) Normally, $V_{ni} \neq U_{ni}$, and we use ϵ_{ni} to denote the random estimation error; that is,

$$U_{ni} = V_{ni} + \epsilon_{ni}. \tag{2.56}$$

Choice Probabilities: Once we have determined the V_{ni} values, we can calculate P_{ni}, the probability that decision-maker n chooses alternative i, as follows:

$$\begin{aligned} P_{ni} &= \mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) \\ &= \mathbb{P}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj} \quad \forall j \neq i) \end{aligned} \tag{2.57}$$

The V_{ni} values are constants. To estimate the probability, then, we need to know the probability distributions of the random variables ϵ_{ni}.

Different choice models arise from different distributions of ϵ_{ni} and different methods for calculating V_{ni}. For instance, the logit model assumes that ϵ_{ni} are drawn iid from a member of the family of generalized extreme value distributions, and this gives rise to a closed-form expression for P_{ni}. (Logit is therefore the most widely used discrete choice model.) The probit model, on the other hand, assumes that ϵ_{ni} come from a multivariate normal distribution (and are therefore correlated, not iid), but the resulting P_{ni} values cannot be found in closed form and must instead be estimated using simulation.

2.8.2 The Multinomial Logit Model

Next we derive the multinomial logit model. (Refer to McFadden (1974) or Train (2009) for further details of the derivation.) "Multinomial" means that there are multiple options from which the decision-maker chooses. (In contrast, binomial models assume there are

only two options.) The logit model is obtained by assuming each ϵ_{ni} is independently and identically distributed from the standard Gumbel distribution, a type of generalized extreme value distribution (also known as type I extreme value). The pdf and cdf of the standard Gumbel distribution are given by

$$f(x) = e^{-x}e^{-e^{-x}} \tag{2.58}$$

$$F(x) = e^{-e^{-x}}. \tag{2.59}$$

We can rewrite the probability that decision-maker n chooses alternative i (2.57) as

$$P_{ni} = \mathbb{P}(\epsilon_{nj} < V_{ni} + \epsilon_{ni} - V_{nj} \quad \forall j \neq i). \tag{2.60}$$

Since ϵ_{nj} has a Gumbel distribution, by (2.59) the probability in the right-hand side of (2.60) can be written as

$$e^{-e^{-(\epsilon_{ni}+V_{ni}-V_{nj})}}$$

if ϵ_{ni} is given. Since the ϵ are independent, the cumulative distribution over all $j \neq i$ is the product of the individual cumulative distributions:

$$P_{ni}|\epsilon_{ni} = \prod_{j \neq i} e^{-e^{-(\epsilon_{ni}+V_{ni}-V_{nj})}}.$$

Therefore, we can calculate P_{ni} by conditioning on ϵ_{ni} as follows:

$$\begin{aligned} P_{ni} &= \int (P_{ni}|\epsilon_{ni}) f(\epsilon_{ni}) d\epsilon_{ni} \\ &= \int (P_{ni}|\epsilon_{ni}) e^{-\epsilon_{ni}} e^{-e^{-\epsilon_{ni}}} d\epsilon_{ni} \\ &= \int \left(\prod_{j \neq i} e^{-e^{-(\epsilon_{ni}+V_{ni}-V_{nj})}} \right) e^{-\epsilon_{ni}} e^{-e^{-\epsilon_{ni}}} d\epsilon_{ni}. \end{aligned} \tag{2.61}$$

After some further manipulation (see Problem 2.24), we get

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}. \tag{2.62}$$

(The sum in the denominator is over *all* j, including $j = i$.) Note that the probability that individual n chooses alternative i is between 0 and 1 (as is necessary for a well defined probability). As V_{ni}, the estimate of i's utility for n, increases, so does the probability that n chooses i; this probability approaches 1 as V_{ni} approaches ∞. Similarly, as V_{ni} decreases, so does the probability that n chooses i, approaching 0 in the limit.

The expected number of individuals who will choose product i, $N(i)$, is simply given by

$$N(i) = \sum_{n=1}^{N} P_{ni}. \tag{2.63}$$

Of course, we usually don't know P_{ni} for every individual n, so instead we resort to methods to estimate $N(i)$ without relying on too much data. See Koppelman (1975) for a discussion of several useful techniques for this purpose.

Table 2.5 Estimated utilities V_{ni} for uPhone models for Example 2.11.

Model	Tech-Heads	Mainstream	Casual
10B	0.1	0.6	0.4
10W	−0.2	0.7	0.5
10+B	1.3	0.5	−0.1
10+W	1.1	0.4	0.1

Table 2.6 $\exp(V_{ni})$ values for Example 2.11.

Model	Tech-Heads	Mainstream	Casual
10B	1.11	1.82	1.49
10W	0.82	2.01	1.65
10+B	3.67	1.65	0.90
10+W	3.00	1.49	1.11

Table 2.7 Choice probabilities P_{ni} and segment sizes for Example 2.11.

Model	Tech-Heads	Mainstream	Casual
10B	0.13	0.26	0.29
10W	0.10	0.29	0.32
10+B	0.43	0.24	0.18
10+W	0.35	0.21	0.21
Segment size	0.3 M	1.7 M	0.4 M

□ **EXAMPLE 2.11**

Pear Computer is about to launch model 10 of its popular smart phone, the uPhone. The company is planning four new versions of the uPhone: the uPhone 10 white and black (abbreviated as models 10W and 10B, respectively) and the uPhone 10+ white and black (models 10+W and 10+B). The company has segmented the market into three categories, which they call Tech Heads, Mainstream Users, and Casual Users. Based on market research, Pear Computer has estimated the utilities V_{ni} of each category for each phone model as given in Table 2.5.

The company wishes to know the probability that a user of each market segment will choose each model. We will assume the estimation errors have a Gumbel distribution.

Table 2.6 lists the values of $\exp(V_{ni})$ for all n and i. From these, we can estimate the probabilities P_{ni} as shown in Table 2.7. Note that for a given market segment, the probabilities for the four models sum to 1 (except for rounding error) since we are assuming each consumer will choose exactly one of the models. If we wanted to model the situation in which a consumer may choose not to purchase any uPhone, then we could add a fifth option representing no purchase.

Table 2.7 also lists the total size of each market segment. From the information in the table, we can estimate the total number of each model sold. For example, Pear

Computer can expect to sell

$$0.13 \times 0.3 + 0.26 \times 1.7 + 0.29 \times 0.4 = 0.60$$

million units of the model 10B. Similarly, the demand forecast is 0.65 M for 10W, 0.60 M for 10+B, and 0.55 M for 10+W. □

We refer the readers to other texts (Ben-Akiva and Lerman 1985, Train 2009) for details about this and other choice models. We next give an example of how discrete choice modeling techniques can be used to estimate demand in a supply chain management setting.

2.8.3 Example Application to Supply Chain Management

Suppose there is a retailer who sells a set I of products. The retailer wishes to estimate the probability that a given customer would be interested in purchasing product i, for $i \in I$, so that he can decide which products to offer. Suppose that the customer follows a multinomial logit choice model, as in Section 2.8.2. The retailer's estimate V_i of the customer's utility U_i for product $i \in I$ is given by

$$U_i = V_i + \epsilon_i. \tag{2.64}$$

(Equation (2.64) is identical to (2.56) except that we have dropped the index n since we are considering only a single customer.) If $i = 0$, then U_i and V_i denote the actual and estimated utility of making no purchase.

For any subset $S \subseteq I$, let $P_i(S)$ denote the probability that the customer will purchase product i, assuming that her only choices are in the set S, and let $P_i(S) = 0$ if $i \notin S$. Let $P_0(S)$ denote the probability that the customer will not purchase any product. Then, from (2.62), we have

$$P_i(S) = \begin{cases} \frac{e^{V_i}}{e^{V_0} + \sum_j e^{V_j}}, & \text{if } i \in S \cup \{0\}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.65}$$

The retailer's objective is to choose which products to offer in order to maximize his expected profit. Suppose that the retailer earns a profit of π_i for each unit of product i sold. Suppose also that the retailer cannot offer more than C products. (C might represent shelf space.) Then the retailer needs to solve the following *assortment problem*:

$$\text{maximize} \quad \sum_{i \in S} \pi_i P_i(S) \tag{2.66}$$

$$\text{subject to} \quad |S| \leq C \tag{2.67}$$

$$S \subseteq I \tag{2.68}$$

(If there are multiple customers, we can just multiply the objective function by the number of customers, assuming they have identical utilities. For a discussion of handling non-homogenous customers, see Koppelman (1975).) This is a combinatorial optimization problem; the goal is to choose the subset S. This problem is not trivial to solve (though it can be solved efficiently). However, the bigger problem is that the utilities U_i, and hence the probabilities $P_i(S)$, are unknown to the retailer. One option is for the retailer to offer different assortments of products over time, estimate the utilities based on the observed demands for each assortment, and refine his assortment as his estimates improve.

Rusmevichientong et al. (2010) propose such an approach. They introduce a policy that the retailer can follow to generate a sequence of assortments in order to maximize the expected profit over time. The assortment offered in a given period depends on the demands observed in the previous periods. Rusmevichientong et al. (2010) also propose a polynomial-time algorithm to solve the assortment problem itself.

CASE STUDY 2.1 Semiconductor Demand Forecasting at Intel

Wu et al. (2010) describe a collaboration between Intel Corporation and Lehigh University researchers to apply the leading-indicator approach (Section 2.7) to forecast demands for new products in the semiconductor industry.[2] At the time of the collaboration, Intel was the largest semiconductor manufacturer in the world and produced chips for several vertical markets, such as mobile, desktop, and server devices. Forecasting demands for semiconductors is difficult due to their short life cycles, long lead times, and high demand volatility. (For another application of the leading-indicator approach in the semiconductor industry, see Wu et al. (2006).)

The approach developed by the researchers involved two key ideas. The first is that by combining forecasts from multiple diffusion models (including, possibly, the Bass model of Section 2.6), we may get better forecasts than if we simply choose a single diffusion model. The second is that leading indicators can be used to update the forecast obtained from the diffusion models using a Bayesian approach.

In particular, Wu et al. (2010) propose fitting, say, 10 different diffusion models to historical data. The Bass model is one good choice, but there are other similar models such as the Weibull, Skiadas, and simple logistic diffusion models. In particular, if we have already observed T periods of demand data for the new product, we can best-fit the parameters of each diffusion model (see Section 2.6.3) to the historical data and evaluate the accuracy of each model. The poorly performing models can be eliminated (for the Intel study, the list was narrowed down to five), and the remaining models can each be used to produce a forecast for the demands in period $T + 1$ through $T + \tau$, for some desired τ. An error term can be added to the forecast to produce a probability distribution rather than just a point forecast. This distribution is called the *prior distribution*.

Next, leading indicators are identified from older generation products or other available time series. For each leading indicator, we generate a forecast for periods $T + 1, \ldots, T + k$, where k is the lag for that leading indicator. (See Section 2.7.) Then, we fit each diffusion model to this extended time series in which periods $1, \ldots, T$ come from observed data and $T+1, \ldots, T+k$ come from the leading indicator forecast. We then use the diffusion model, with the parameters determined in the previous step, to produce a forecast distribution for periods $T + k + 1, \ldots, T + \tau$. This distribution is called the *sampling distribution*.

Finally, we perform a Bayesian update using the prior and sampling distributions to produce a *posterior distribution* for each diffusion model. These distributions are then combined by taking, for each future time period, a weighted sum of the forecasts

[2]In this and subsequent case studies, we have adapted the original notation to be consistent with the rest of the book. In some cases we have also simplified or made other minor modifications to the models, while striving to maintain the main ideas of the original models.

generated by the various diffusion models. Wu et al. (2010) show both analytically and empirically that this results in a smaller variance of forecast error than any of the individual forecasts.

The team implemented the method for 60 Intel products from the mobile, desktop, and server markets. Wu et al. (2010) report that over 10 monthly forecasting cycles, the new method reduced the 12-month forecast error, as measured by MAPE (see Section 2.3), by 9.7%. Moreover, the accuracy of the 4-month forecast, which is the most important given the products' production cycles, improved by 33%. Intel estimated that this would translate to at least $1.3 million in increased revenue per product over 4 months due to the improved forecasts leading to fewer stockouts. In addition, the decision-support system built by the team to implement this approach executes quickly, reducing the time required to generate forecasts from approximately 3 days under Intel's old aproach to 2 hours using the new system. The work described by Wu et al. (2010) was a finalist for INFORMS's prestigious Wagner Prize for Excellence in Operations Research Practice; see Butler and Camm (2010).

PROBLEMS

2.1 (Forecasting without Trend) A hospital receives regular shipments of liquefied oxygen, which it converts to oxygen gas that is used for life support. The company that sells the oxygen to the hospital wishes to forecast the amount of liquefied oxygen the hospital will use tomorrow. The number of liters of liquefied oxygen used by the hospital in each of the past 30 days is reported in the file `oxygen.xlsx`.

a) Using a moving average with $N = 7$, forecast tomorrow's demand.

b) Using single exponential smoothing with $\alpha = 0.1$, forecast tomorrow's demand.

2.2 (Forecasting with Trend) The demand for a new brand of dog food has been steadily rising at the local PetMart pet store. The previous 26 weeks' worth of demand (number of bags) are given in the file `dog-food.xlsx`.

a) Using double exponential smoothing with $\alpha = 0.2$ and $\beta = 0.1$, forecast next week's demand. Initialize your forecast by setting $I_t = D_t$ for $t = 1, 2$ and $S_2 = I_2 - I_1$.

b) Using linear regression, forecast next week's demand.

2.3 (Forecasting Cupcake Sales) Karl's Cupcakes recently launched a new variety of cupcake. The weekly demands, measured in dozens, during the first two weeks of sales were $D_1 = 47.2$ and $D_2 = 52.3$.

a) Use double exponential smoothing with $\alpha = 0.1$ and $\beta = 0.2$ to calculate y_3, the forecast made in week 2 for the demand in week 3.

b) Suppose the actual demand in week 3 is 59.4. What is y_4, the forecast made in week 3 for the demand in week 4?

2.4 (Forecasting with Seasonality) A hardware store sells potting soil, the demand for which is highly seasonal and has also exhibited a slight upward trend. The number of bags of soil sold each month for the past 40 months is reported in the file `potting-soil.xlsx`. Using triple exponential smoothing with $\alpha = 0.2$, $\beta = 0.1$, and $\gamma = 0.3$, forecast the

demand for May. Initialize your forecast by setting

$$I_t = D_t$$
$$S_t = I_t - I_{t-1}$$
$$c_t = \frac{12D_t}{\sum_{i=1}^{12} D_i}$$

for periods $t = 1, \ldots, 12$. (There are better ways to initialize this method, but this method is simpler.)

2.5 (Forecasting Melon Slicers) Matt's Melon Slicers sells specialized knives for watermelons, the demand for which is highly seasonal, with the majority of the demand occurring during the summer. The company has been selling melon slicers for three years and has calculated the following estimates of the seasonal factors, with each period representing one quarter:

Quarter	t	c_t
Winter	9	0.4
Spring	10	0.8
Summer	11	1.9
Fall	12	0.9

At the end of period 12, the company calculated the following estimates of the base signal and slope: $I_{12} = 642$, $S_{12} = 84$.

a) Calculate y_{13}, the forecast made in period 12 for the demand in period 13.

b) Suppose the demand in period 13 turns out to be 341. Calculate I_{13}, S_{13}, and c_{13}.

2.6 (Forecasting Using Regression) The demand for bottled water at football (aka soccer) matches is correlated to the outside temperature at the start of the match. The file `bottled-water.xlsx` reports the temperature (°C) and number of bottles of water sold for each home match played at a certain stadium for the past two seasons (19 home matches per season).

a) Using these data, build a linear regression model to relate the demand for bottled water to the match-time temperature. What are $\hat{\beta}_0$ and $\hat{\beta}_1$?

b) The temperatures for the next three matches are predicted to be $21.6°$, $27.3°$, and $26.6°$, respectively. Forecast the demand for bottled water at each of these matches.

2.7 (Multiple-Period-Ahead Forecasts) In this chapter, we discussed time-series methods for forecasting the demand one period ahead, i.e., in period $t-1$, we generate a forecast y_t for the demand in period t. Suppose instead that we wish to forecast multiple periods ahead, i.e., in period $t-1$, we generate a forecast $y_{t-1,t+k}$ for the demand in period $t+k$, for $k \geq 0$. Explain how to adapt each of the following methods to handle this case:

a) Moving average

b) Double exponential smoothing

c) Linear regression

2.8 (Forecasting using Machine Learning Methods) Using the data set provided in Problem 2.6, choose a learning-based forecasting method—a tree-based model, SVR, or neural networks—for forecasting bottled water given temperatures. Use your selected method to forecast the demand during matches when the temperatures are $21.6°$, $27.3°$, and $26.6°$. Compare your results with those you obtained using linear regression in Problem 2.6(b).

2.9 (Ridge Regression) Ridge regression introduces an ℓ_2-norm penalty to the objective function of linear regression. Consider a simple version in which we have only a single input ($p = 1$); then we are minimizing

$$\sum_{i=1}^{n} \left(y^i - (\beta_0 + \beta_1 x^i)\right)^2 + \lambda(\beta_0^2 + \beta_1^2),$$

where $\lambda > 0$ is the penalty parameter. Derive closed-form expressions for β_0 and β_1. You may use a matrix representation if you wish.

2.10 (Forecasting Fires) The file `nyc-fires.csv` contains the number of fires responded to by the New York City Fire Department on each day from January 1, 2013 through June 30, 2016 (NYC OpenData 2017). It also contains the high temperature (in °F) and the total precipitation (in inches) on the same days (National Oceanic and Atmospheric Administration (NOAA) 2017).

Load the data into MATLAB, Excel, or another software package of your choice. Add a variable called `IsWeekend` that indicates whether each day is a weekend day (Saturday or Sunday). Split the data into two parts, one for 2013–2015 (this will be your training data) and one for 2016 (this will be your testing data).

In this problem, you will build models to predict the number of fires on a given day using the three features (high temperature, precipitation, and weekend (Y/N)). Use only the training data when building your models.

- **a)** Build a linear regression model. Report the coefficients $\hat{\beta}_i$.
- **b)** Build a regression tree model with at most 10 branching nodes. (A branching node is a node that has child nodes.) Include a diagram of your tree.
- **c)** Build an SVR model. Report the coefficients $\boldsymbol{\beta}$ and β_0.
- **d)** For each method in parts (a)–(c), predict the number of fires on each day in the testing data. Report the predicted and actual values and the forecast error for the first 10 records in the testing data. Also report the MSE for each method for the entire testing set.

2.11 (Exponential Smoothing for Retail Sales) The file `retail-sales-data.csv` contains weekly sales data for 99 departments within 45 retail stores over approximately 3 years. This is actual data from a real company but has been anonymized (see Kaggle.com (2017)).

- **a)** Extract the sales data for store 2, department 93. Determine the most appropriate form of exponential smoothing (single, double, or triple) and apply that method to forecast the sales. Use 0.15 for all of the smoothing constants (α, β, and/or γ). Begin forecasting at the earliest period you can. (For example, in double exponential smoothing the forecasts begin in period 3.) Report the MSE, MAD, and MAPE for your forecasts. Plot the actual and forecast sales on a single plot.
- **b)** Repeat part (a) for store 3, department 60.

c) Repeat part (a) for store 1, department 16.

2.12 (Mean and Variance of Exponential Smoothing Forecast Error) Prove equations (2.31) and (2.32).

2.13 (Forecasting Simulation) Consider a product whose daily demand follows (2.30) with $\mu = 40$ and $\sigma = 6$.

a) Build a spreadsheet simulation of the demand process, as well as a moving average forecast of order 5. Simulate the system for at least 500 periods. Report the MSE and MAD of the forecast. Also calculate the standard deviation of the forecast error. How accurate is the approximation given in (2.28) for your simulated values?

b) Repeat part (a) for an exponential smoothing forecast with constant $\alpha = 0.1$.

c) Based on the results of parts (a) and (b), does one forecasting method appear to work better than the other?

2.14 (Bass Diffusion for LPhone) HCT, an Asian manufacturer of a new 4G cell phone, the LPhone 5, is planing to enter the U.S. market, and they are in the process of signing a contract with a third-party logistics (3PL) provider in which they must specify the size of the warehouse they want to rent from the 3PL. HCT wants to forecast the total sales of the LPhone 5, as well as the time at which the LPhone 5 reaches its peak sales. After some thorough market research, HCT has estimated that $p = 0.008$, $q = 0.421$, and $m = 5.8$ million. Calculate when the peak sales will occur and how many LPhone 5 the company will have sold by that point.

2.15 (Bass Diffusion for iPeel) Banana Computer Co. plans to launch its latest consumer electronic device, the iPeel, early next year. Based on market research, it estimates that the market potential for the iPeel is 170,000 units, with coefficients of innovation and imitation of 0.07 and 0.31, respectively.

a) If the iPeel is introduced on January 1, on what date will the sales peak? What will be the demand rate on that date, and how many units will have been sold?

b) On what date will 90% of the sales have occurred?

c) Plot the demand rate and cumulative demand as a function of time.

2.16 (Bass Diffusion for Books) A new novel was published recently, and the demand for it is expected to follow a Bass diffusion process. The publisher decided to print only a limited number of copies, observe the demand for the book for 20 weeks, estimate the Bass parameters, and then undertake a second printing for the remainder of the life cycle of the book using these parameters. The demand for the book during these 20 weeks is reported in the file `novel.xlsx`. Using these data, estimate m, p, and q using the method described in Section 2.6.3.

2.17 (Proof of Corollary 2.2) Prove Corollary 2.2.

2.18 (Influentials and Imitators) Suppose that potential adopters of a given product fall into two distinct segments: *influentials* and *imitators*. Each segment has its own within-segment innovation and imitation parameters and experiences its own Bass-type contagion process. In addition, the influentials can exert a cross-segment influence on the imitators, but not vice-versa. Let θ denote the proportion of influentials in the population of eventual adopters ($0 \leq \theta \leq 1$), and $\bar{\theta} = 1 - \theta$ denote the proportion of imitators. Let p_i and q_i

denote the within-segment innovation and imitation parameters, respectively, for $i = 1, 2$, where $i = 1$ represents influentials and $i = 2$ represents imitators. Let q_c denote the cross-segment imitation parameter.

a) Write a formula expressing each segment's instantaneous adoption behavior, analogous to (2.42).

b) What is special about the case in which $\theta = 0$ or $\theta = 1$?

c) If there are no pre-release purchases (i.e., $D_1(0) = D_2(0) = 0$), write a formula expressing the cumulative adoption at time t, analogous to (2.43).

2.19 (Demand Diffusion across Multiple Markets) A company plans to introduce a variety of new products to multiple vertical markets. The demands from these verticals are likely to follow different diffusion patterns. The company is interested in combining diffusion models derived from different vertical markets to help characterize the overall market demand. However, they are not sure about whether doing so would introduce additional variances and biases into the forecast. Show that combining forecasts of different diffusion models using weights that are inversely proportional to their forecast variances yields a combined forecast variance that is smaller than the forecast variance of each individual diffusion model.

2.20 (Leading Indicators) A battery manufacturer produces a large number of models of lithium-ion batteries for use in computers and other electronic devices. The products are introduced at different times and follow different demand processes. The company wishes to determine whether some of the products can serve as leading indicators for the rest of the products. The file `batteries.xlsx` contains historical demand data for 25 products for the past 26 weeks.

a) Using Algorithm 2.1 with parameters $k_{\min} = 3$, $k_{\max} = 9$, and $\rho_{\min} = 0.85$, determine all pairs (i, k) such that product i is a leading indicator with lag k. (*Note*: You should not need to recluster the products.)

b) Using one of the (i, k) you found in part (a), forecast the demand for the rest of the cluster in periods 27 and 28.

2.21 (Discrete Choice with Uniform Errors) Suppose that, in the discrete choice model, the estimation error ϵ_{ni} has a $U[-1, 1]$ distribution for all n and i. Write an expression for P_{ni}, analogous to (2.61). Your expression may include ϵ_{ni}, V_{ni}, and V_{nj}, but not ϵ_{nj}.

2.22 (Discrete Choices for Day Care) A university is in the process of choosing a location for a new day care center for its faculty's children. The two options for the location are city A, where the university is located, or city B, a neighboring city known for larger houses but a longer commute. The university wants to estimate the number of faculty with kids who are living or will live in city A during the next 10 years. To that end, the university wishes to estimate the choice probability between the two cities for a typical family. Suppose that the utility a family obtains from living in each city depends only on the average house purchase price, the distance between the city and the campus, and the family's opinion of the convenience and quality of life of each city. The first two of these factors can be observed by the researcher, but the researcher cannot observe the third. The researcher believes that the observed part of the utility is a linear function of the observed factors; in particular, the utility of living in each city can be written as

$$U_A = -0.45PP_A - 0.23D_A + \epsilon_A$$

$$U_B = -0.45PP_B - 0.23D_B + \epsilon_B,$$

where the subscripts A and B denote city A and city B, and PP and D are the purchase price and distance. The unobserved component of the utility for each alternative, ϵ_A and ϵ_B, vary across households depending on how each household views the quality and convenience of living in each city. If these unobserved components are distributed iid with a standard Gumbel distribution, calculate the probability that a household will choose to live in city A.

2.23 (Using Discrete Choice to Forecast Movie Sales) Three new movies will be shown at a movie theater this weekend. The theater wishes to estimate the expected number of people who will come to see each movie so they can decide how many screenings to offer, how large a theater each movie should be shown in, and so on. The movie studios that produced the three movies held "sneak peak" screenings of the films and conducted post-movie interviews of the attendees. Based on these interviews, they estimated the utility of each movie based on a viewer's age range. They also estimated the utility of not seeing any movie. These estimated utilities are denoted V_{ni}, although here n refers not to an individual but to a *type* of individual (based on age range). The following table lists the V_{ni} values, as well as the number of people who are considering seeing a movie at that theater this weekend.

	Age Range		
Movie	16–25	26–35	36+
Prognosis Negative	0.22	0.54	0.62
Rochelle, Rochelle	0.49	0.57	0.51
Sack Lunch	0.53	0.31	0.38
No movie	0.10	0.27	0.41
Population	700	1900	1150

a) Assume that the actual utilities U_{ni} differ from the estimated utilities V_{ni} by an additive iid error term that has a standard Gumbel distribution. Using the multinomial logit model of Section 2.8.2, calculate the expected demand for each movie.

b) Now suppose the movie theater doesn't know about the multinomial logit model and assumes that P_{ni} is simply calculated using a weighted sum of the V_{ni} values; that is,

$$P_{ni} = \frac{V_{ni}}{\sum_j V_{nj}}.$$

What are the expected demands for each movie using this method?

2.24 (Proof of (2.62)) Prove equation (2.62).