OPTIMIZATION AND ESTIMATION PROBLEMS IN AIRLINE YIELD MANAGEMENT

By

Jeffrey I. McGill
B. Sc. (Mathematics) Bishop's University
M. Sc. (Mathematics) Concordia University

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
COMMERCE AND BUSINESS ADMINISTRATION

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
December 1989

Department of Management Science, Faculty of Commerce

The University of British Columbia
Vancouver, Canada

Date December 7, 1989.

Abstract

This thesis addresses problems of optimization and estimation encountered in the process of *airline yield management*, also called *airline seat inventory control.* Optimality conditions are given for the problem of setting booking limits for multiple, stochastically independent demand classes that are booked in a nested fashion into a fixed pool of airline seats. These optimality conditions are compared with the approximations given by the *EMSR* method. Additional conditions are given for two stochastically dependent fare classes, and extensions are made that allow for incorporation of passenger goodwill and upgrades of passengers between fare classes. The model developed for the dependent demand case is also applied to the problem of determining an optimal overbooking limit in a single fare class. Finally, a methodology is developed for using multivariate multiple regression in conjunction with the EM method to estimate the parameters of demand distributions on the basis of historical demand data that have been censored by the presence of booking limits.

Table of Contents

List of Tables

List of Figures

Acknowledgement

This thesis could not have been completed without the help of many individuals and organizations. I would like to express particular thanks to the following:

Professor Shelby Brumelle, Management Science Department,UBC: For getting me interested in the problems addressed by this thesis, for supervising the work, for sharing with me a little of what he knows about stochastic processes, and for numerous contributions throughout the project.

Professors Tae Oum and Michael Tretheway, Transportation Department,UBC: For serving on my thesis committee, for financial support, for contributions to the work on the dependent demand problem, and for helpful advice and moral support.

Professor Harry Joe, Statistics Department, UBC: For serving on my thesis committee, and for careful reading and constructive criticism of the entire thesis, in particular the material on censored regression analysis.

Personnel at Canadian International Airlines: For providing me with details of modern airline operations as they relate to seat inventory control. Particular thanks are due Mike Maple, Manager of the Yield Sciences group at CAI and Ed L'Heureux, Manager of Operational Research.

The Government of Canada: For financial support early in my doctoral program through the Social Sciences and Humanities Research Council of Canada and in the later stages through the Natural Sciences and Engineering Research Council, Grant No. A4104.

Final heartfelt thanks are due my wife Helen List for unwavering confidence and encouragement despite the very real stesses and strains from my doctoral program that were, inevitably, transferred to her.

Chapter 1

Introduction

1.1 Airline Yield Management and the Seat Allocation Problem

One of the obvious impacts of the deregulation of North American airlines has been increased price competition and the resulting proliferation of discount fare booking classes. While this has had the effect of greatly expanded demand for air travel, it has presented the airlines with a tactical planning problem of considerable complexity — how to establish booking policies that result in optimal allocations of seats among the various fare classes. What is sought is the best trade-off between the revenue gained through greater demand for discount seats against revenues lost when full fare reservations requests must be turned away because of prior discount seat sales.

The component of airline planning that deals with these allocation problems has come to be known in the industry as *airline yield management.* Alternate terms are *airline seat allocation* and *airline seat inventory control.* The latter, more specific terms will be used in this thesis.

1.1.1 Components of the Seat Allocation Problem

The seat allocation problem is complicated by a number of factors, some of which are listed below.

advance booking: Bookings for flights are made over a long period prior to departure; 300 day lead times are common. The discount fare classes tend to book earlier than the full fare classes both because of the nature of the customers for the respective classes (leisure travellers in the discount classes, business travelers in full fare) and because of early booking restrictions placed on the discount classes. Thus decisions about limits to place on the number of discount fare bookings must often be made before any full fare demand is observed.

stochastic demand: Demands are highly variable and exhibit significantly random behaviour even after allowance is made for factors like season, day of week, competitor pricing, etc.. Thus, obtaining reliable forecasts of demand is a significant sub-problem of the seat allocation problem. This forecasting problem is made more difficult by the fact that historical demand data are censored by the presence of booking limits and the finite capacity of aircraft.

shared seating and nested booking: Rather than simply dividing the available seats into separate groups reserved solely for particular booking classes, many airlines are now placing some or all of the classes below first class into shared seating areas. The booking of passengers is often done in a 'nested' fashion. That is, bookings in the lowest fare class are permitted up to a certain limit, combined bookings in the two lowest classes are permitted up to a second, higher limit, and so on.

late cancellations and overbooking: Passengers with full fare tickets generally have the right to cancel their booking at any time before flight, or simply to fail to show-up at flight time without penalty. Airlines failing to take this into account can experience significant losses from underloaded planes. One solution to this problem is to overbook to compensate for expected late cancellations; however, this complicates the planning process since allowance must be made for compensation of passengers who are denied boarding on those occasions when the number of show-ups exceeds plane capacity.

passenger itineraries: Many bookings for particular legs of a flight are parts of a larger passenger itinerary involving other legs of the same flight or other flights. Thus, in assessing the revenue impact of accepting or rejecting a booking for one leg, allowance should be made for potential revenues from other legs of the passenger's itinerary. Also, allowance should be made for potential revenues from all other itineraries that might require the same flight leg.

full fare passenger spillage: Airlines typically place goodwill value on full fare passenger bookings above and beyond the higher fare value of the bookings. This is because full fare travellers are predominantly business travellers who can be expected to be repeat customers. Excessive sales of discount seats may result in significant numbers of full fare passengers being refused bookings or 'spilled'. There is some concern that regular customers refused reservations might be lost to competitors.

demand dependencies: The demands in different fare classes may be stochastically dependent. The existence of such dependencies should permit revision of probability distributions of final demand in one fare class on the basis of booking levels in a different fare class.

dynamics: Airlines with modern reservations control systems monitor the bookings for flights as the day of flight departure approaches. Changes to booking limits for different fare classes can be made several times prior to departure of the flight as more information becomes available concerning eventual demands. Thus the setting of booking policies is, in practice, a dynamic problem.

problem size: A typical airline may handle traffic volumes of one to two thousand flights per day and may make booking control decisions many times during the lead time before each flight departs. One airline (American Airlines) estimates that it must make approximately 100 million control decisions per year [1, (1988)]. Thus any method designed to aid booking control decisions cannot involve computation times of more than a fraction of a second per decision.

1.1.2 Two Approaches to the Seat Allocation Problem

Prior work on this problem has tended to fall into one of two categories. First, attempts have been made to encompass some or all of the above-mentioned problem components with large-scale mathematical programming models [3, 38, 57, 95, 150]. In those cases where implementation of such models has been attempted, it has been necessary to make significant compromises in order to make it computationally feasible to solve realistic versions of the problem. For example, in their dynamic programming treatment of the two-fare overbooking problem for a single flight, Alstrup, Boas, Madsen and Vidal [3, (1986)] found it necessary to aggregate seats into blocks in order to reduce computation time from an estimated 100 hours per flight to a more reasonable 9 seconds. F. Glover, R. Glover, J. Lorenzo and C. McMillan [57, (1982)] suppressed the stochastic elements of the problem in order to deal with the network elements (i.e. the interaction of various passenger itineraries). These approaches achieved a measure of success in providing approximate solutions to realistic versions of the problem but they did not offer insights into the nature of optimal solutions nor facilitate study of the effects of changes in the parameters of the problem.

In the second category, the one most relevant to this thesis, elements of the problem have been studied in isolation under restrictive assumptions. These studies have produced easily applied rules that provide some insight into the nature of good solutions. These rules are suboptimal when viewed in the context of the overall problem, but they can point the way to useful approximation methods. For example, Littlewood [89, (1972)] proposed a simple rule for a restricted version of the two fare allocation problem. Variations on this rule were later applied to multiple fare classes and multiple flight legs by Wang [145, (1983)], Simpson [126, (1985)] and Belobaba [13, (1987)]. These generalizations were proposed on the basis of intuitive rather than rigorous reasoning, and it will be shown in this thesis that they are not optimal, even for restricted versions of the problem. However, there is evidence that they do provide reasonable approximations, and they are particularly easy to implement.

1.2 Objectives of this Thesis

This thesis addresses the seat allocation problem in the manner described in the second category above. That is, components of the problem are studied in isolation using analytical rather than computational approaches. In contrast to the prior work, however, the analysis is rigorous, and optimal solutions are sought for the problem components. The emphasis is on finding structural solutions specifying policies that maximize revenues. Where appropriate, these structural solutions are used to examine the qualitative behaviour of optimal solutions. Simplifications and/or approximations of optimal solutions are sought that facilitate implementation, and consideration is given to the problems of integrating solutions to individual components into the overall reservations control framework.

The components of the seat allocation problem that are studied are:

1) seat allocation among multiple, independent fare classes sharing the same seating pool on one leg of a flight when reservations are nested and occur in order of fare class,

2) seat allocation between two nested fare classes when the demands between the two classes are stochastically dependent, and

3) estimation of the parameters of the joint demand distribution for two dependent fare classes on the basis of data that have been censored by the presence of booking limits.

More detailed statements of the objectives are furnished in the chapter summaries given later in this introduction. The next section provides an overview of past research on airline seat allocation.

1.3 A Survey of Past Research

The objective of this section is to survey the large body of previous research on components of the airline seat inventory control problem so that the findings presented in this thesis can be placed in context. Published work on components of the seat management problem dates back to the 1950's when the overbooking problem first received attention from researchers. Since that time there has been a steady stream of results reported in technical journals, proceedings of airline professional conferences and internal and external company reports. Emphasis is given here to reviewing previous research that is either significant in the field or directly relevant to this thesis, thus many publications are not directly discussed. The bibliography contains a listing of all work that turned up during the literature search for this thesis, whether cited or not. It is hoped that this will provide a source list useful to future researchers in the area.

This survey will be organized under the subheadings "Overbooking", "Single Flight Leg Seat Allocation" and "Multiple Flight Leg Seat Allocation."

1.3.1 Overbooking

The overbooking problem has received by far the most attention of the problems discussed above. An early, non-dynamic optimization model for overbooking is that of

Beckmann [11, (1958)]. Shlifer and Vardi [123, (1975)] provide a similar model extended to allow for two fare classes and a two-leg flight. Statistical models of various levels of sophistication are described in Thompson [135, (1961)], Taylor [134, (1962)], Rothstein and Stone [118, (1967)], Martinez and Sanchez [94, (1970)], and Littlewood [89, (1972)]. The objective of most of these models is to permit controlled overbooking of flights so that the probability that passenger show-ups exceed seating capacity on a flight is kept within limits set by the airline or external regulating bodies. None of these allow for the dynamics of the passenger cancellation and reservation process subsequent to the overbooking decision. The Rothstein and Stone model was implemented at American Airlines, and trial implementations were reported by other authors (e.g. Deetman [35, (1964)]); however, it is unclear from the literature whether or not this model or any of the others are in use today. (The airlines are understandably reticent about revealing what methods they do or do not use for yield management.)

A number of researchers have developed dynamic optimization approaches to the airline overbooking problem (as well as the similar problem in the hotel/motel industry). The usual objective in these formulations is to determine a booking limit for each time period before flight departure that maximizes expected revenue, where allowance is made for the dynamics of cancellations and reservations in subsequent time periods and for penalties for oversold seats. Kosten [76, (1960)] develops a continuous time approach to this problem, but this approach requires solution of a set of simultaneous differential equations that make implementation impractical. Rothstein [117, (1968)], in his Ph.D. thesis, describes the first dynamic programming model for overbooking and reviews the results of test runs of the model at American Airlines. It is not clear whether or not American or any other airline carried out a full scale implementation; however, Andersson [6, (1972)] reports that, to his knowledge, Rothstein's model had not been implemented as of that time. A dynamic programming analysis similar to Rothstein's but developed for the hotel/motel industry and extended to two fare-classes is described in Ladany [80, (1976)], [79, (1977)]. Liberman and Yechiali [87, (1977)],[88, (1978)] employ a similar analysis to obtain a control-limit type structural solution to the (one-class) hotel overbooking problem. Discussions of policy issues relating to passenger overbooking and equitable 'bumping' are found in Simon [125, (1968)],[124, (1972)]; Falkson [51, (1969)]; Bierman and Thomas [19, (1975)]; Rothstein [114, (1971)],[112, (1975)], [116, (1985)]; Vickrey [143, (1972)]; Nagarajan [98, (1979)]; and Ruppenthal [119, (1983)]. Belobaba, in a recent Ph.D. dissertation [13, (1987)], discusses the problem of overbooking in multiple fare classes and suggests a heuristic approach to solving the problem.

1.3.2 Single Flight Leg Seat Allocation

The problem of optimal allocation of seats to fare classes, which is more central to the yield management problem than the overbooking problem, has been addressed by a number of researchers. Littlewood [89, (1972)] is apparently the first to present (without proof) the *simple seat allotment rule* that will be discussed in chapter 2 of this thesis. Derivations are given by Bhatia and Parekh [18, (1973)] and, by a different method, Richter [108, (1982)]. Mayer [95, (1976)] performs some sensitivity analysis on the simple model and offers evidence, based on trial runs, that the model, if used more than once before flight departure, can perform as well as a more complex dynamic programming model. He also suggests that the seat allotment and overbooking analyses can be done independently. Titze and Greisshaber [136, (1983)] describe the results of a simulation study that suggests that the simple seat allotment rule remains approximately optimal in spite of departures from the assumption (implicit in the formula) that all low fare passengers book before high fare passengers. In the thesis mentioned above, Belobaba [13, (1987)] generalized the simple seat allotment rule to obtain approximate formulas for more than two fare classes. Pfeifer [106, (1989)] presents an analysis of the two fare class problem that allows for the possibility of upgrades from the lower fare class to the higher. The same result is proposed without proof in Belobaba's thesis. (Another derivation of this result is provided in this thesis to illustrate the generality of a general model of seat allocation between dependent demands.)

1.3.3 Multiple Flight Leg Seat Allocation

Research that incorporates the interaction of different legs of the same or other flights into the seat allocation problem has been reported since the mid-1970's. Shlifer and Vardi [123, (1975)] develop a simple, non-dynamic overbooking model for a two leg flight with two fare classes. Buhr [24, (1982)] describes a computer program for seat allocation for a two-sector flight with one fare class. He suggests that, under normal circumstances, the multileg allocation problem can be solved independently of the fare-class allocation problem. Hersh and Ladany [67, (1978)] use Bayesian updating of demand forecasts in a dynamic programming model for a two-sector flight with one fare class. F. Glover, R. Glover, Lorenzo and McMillan [57, (1982)] outline a non-dynamic, minimum cost network flow approach. Related network approaches are presented by Dror, Trudeau and Ladany [38, (1988)] and Wollmer [148, (1986)]. Wang [145, (1983)], Simpson [126, (1985)] and Belobaba [13, (1987)] propose extensions of the simple seat allotment approach to multiple leg flights.

1.4 An Overview of this Thesis

This section provides a summary of the objectives and main results of the remaining chapters of the thesis.

Summary of Chapter 2: Multiclass Allocation when Demands are Independent

Current airline reservations control systems may admit as many as eight main fare classes. The problem of controlling bookings into multiple classes is thus of considerable practical interest. As discussed earlier, the approach followed in this thesis is to seek simple rules based on restricted versions of the seat allocation problem and then investigate practical ways of implementing these rules in more realistic settings.

This chapter addresses multiple fare class allocation when the conditions listed below are assumed to hold.

1) Demands for different booking classes occur in a sequential manner with lower fare classes booking before higher.
2) The demands for the different booking classes are stochastically independent.
3) At any stage of the booking process, the only information available about demands is the prior probability distributions of demand and the number of current bookings in all fare classes.
4) No allowance is made for overbooking.

These assumptions underly all previous attempts to derive simple optimality conditions for control of bookings into two or more fare classes. Their implications will be discussed later.

A new, rigorous formulation of the multiple fare class allocation problem is presented here, and a number of new results are obtained. In particular, chapter 2:

1) presents a recursive formulation of the revenue function for multiple booking classes,
2) characterizes the problem of maximizing expected revenues as a series of

monotone optimal stopping problems,

3) presents conditions under which the expected revenue function is concave and derives optimality conditions,

4) proves the equivalence of a simple set of probability statements to the optimality conditions for the continuous demand case, and

5) demonstrates the non-optimality of the EMSR method proposed by Belobaba [13, (1987)] and presents numerical comparisons of the EMSR versus optimal solutions.

Summary of Chapter 3: Allocation Between Two Classes when Demands are Dependent

It is a common perception among airline personnel[1] that demands in different fare classes are not independent even after allowance is made for factors that influence overall demand such as the season, fare pricing, or the day-of-week. There are at least two reasons that dependency between demands might arise. The first is that customers refused a booking in one discount fare class may elect to *upgrade* to a higher fare class. This introduces a positive dependency between the observed demands for the two classes. The second reason is that scheduled events such as conferences permit advanced bookings by budget-conscious attendees. Such events stimulate demand across several booking classes, and this can again introduce positive dependency between demands.

Previous efforts to obtain optimality conditions in the two fare class case have assumed that demands are independent. This chapter addresses the problem without this assumption. Specific contributions follow.

1) A general model for the revenue in the two fare class case is introduced that

---

[1] personal communication, E. R. L'Heureux, Canadian Airlines International

admits dependency between classes. Optimization of this revenue function is shown to be equivalent to solution of a monotone optimal stopping problem.

2) A generalization to Littlewood's formula is shown to be optimal for the dependent demand case as long as demands are *monotonically associated* (a condition introduced here).
3) A rigorous proof is provided of an optimality condition proposed by Belobaba [13, (1987)] for the case that dependency arises because of upgrades.
4) It is shown that a variant of the simple allotment rule can allow for the perceived extra value of full fare passengers.
5) Numerical comparisons of airline seat allocations with and without dependency are given.
6) It is proven that monotonic association is satisfied by positively correlated bivariate normal demands and that the discount class booking limit decreases as correlation increases.

Summary of Chapter 4: A Simple Overbooking Model

As mentioned previously, the overbooking problem has received by far the greatest amount of attention of all of the components of the seat inventory control problem, and it remains an area of active research. As discussed in the literature survey above, the appoaches have ranged from simple heuristics for multiple fare classes like that of Belobaba [13, (1987)] through static models like those of Shlifer and Vardi [123, (1975)] to computationally intensive, dynamic programming approaches like that of Rothstein [117, (1968)].

This chapter addresses a primitive version of the overbooking problem from a new perspective. Specifically, the problem of determining an optimal overbooking limit for one fare class is handled with a variant of the general model for seat allocation for two dependent demands that is developed in Chapter 3.

Specific contributions of this chapter are:

1) It is shown that the general revenue model employed to analyze the dependent demand allocation problem in Chapter 3 can be applied to the single fare overbooking problem.
2) A simple optimality condition is derived for overbooking under the (reasonable) assumption that passenger show-ups occur according to a Bernoulli process.
3) An approximation is derived for the optimality condition, and it is shown that this condition is essentially equivalent to the solution developed by Shlifer and Vardi but much easier to apply.
4) Conditions are determined under which simply dividing the capacity by the confirmation probability yields a near optimal overbooking level for a single fare class.
5) A numerical example is provided.

Summary of Chapter 5: Estimation of Dependent Demands from Jointly Censored Data

Application of the optimality condition obtained in Chapter 3 for dependent demands requires knowledge of probability distributions of full fare demand conditioned on the observed demand in the discount fare class. One convenient way of obtaining these distributions is by estimating the parameters of the joint distribution of demand for the two fare classes. The estimation of these parameters is greatly complicated by the fact that historical demand data for different booking classes on flights are *censored* by booking limits and the finite capacity of aircraft. That is, in any set of observations of demand on past flights, many observations come from flights on which one or both fare classes exceeded their booking limits. In such cases it is known that demand exceeded the booking limit but not by how much. Failure to take such censorship into account when estimating demand parameters can produce estimates that are so seriously biased as to be unusable.

Chapter 5 provides a methodology for estimating the parameters of a bivariate demand distribution from data that is censored in the manner described above. Specific contributions are:

1) A censored, bivariate, multiple regression model is presented for estimating the parameters of the joint demand distribution.

2) The details of an iterative maximum likelihood estimation procedure based on the EM method of Dempster, Laird and Rubin [36, (1977)] are presented.

3) A computer implementation of the estimation procedure is described.

4) Numerical examples are provided which demonstrate the accuracy and efficiency of the method.

1.4.1 Notation

The following notational conventions are followed in this thesis. Exceptions, when they occur, are noted.

1. Scalar constants and non-random variables are represented by small italic roman or greek letters; e.g. $x, x_1, \mu, \rho \ldots$. An exception to this is the symbol for aircraft capacity, C.

2. Scalar random variables are represented by capital roman italic letters; e.g., X, Y, $X_1, X_2, \ldots$.
3. Vectors and matrices are represented by boldface roman or greek letters; e.g. $\mathbf{p}, \mathbf{H}, \boldsymbol{\beta}, \ldots$. No notational distinction is made between random and non-random vectors or matrices.
4. $\Pr[A]$ denotes the probability of event A.
5. The expectation of the random variable X will be denoted by $\mathrm{E}\{X\}$ or $\mathrm{E}[X]$. Expected revenue with respect to a random variable X as a function of parameters in $\mathbf{p}$ is written $ER[\mathbf{p}; X]$.
6. Maximum, minimum and positive part are represented as follows: $a \vee b \equiv \max(a, b)$, $a \wedge b \equiv \min(a, b)$, and $a^+ \equiv a \vee 0$.
7. The indicator operator I is defined for logical propositions as 1 if the proposition is true and 0 otherwise. For example,

$$I_{[\alpha<\beta]} = \begin{cases} 1 & \text{if } \alpha < \beta, \\ 0 & \text{if } \alpha \geq \beta. \end{cases}$$

8. The end of proofs will be marked with the symbol '■' at the right-hand margin.

Chapter 2

Multiclass Allocation when Demands are Independent

2.1 Introduction

This chapter deals with the airline seat allocation problem when multiple fare classes are booked into a common seating pool in the aircraft and when the demands for the fare classes are assumed to be stochastically independent. The following additional assumptions are made:

1. *single flight leg:* Bookings are made on the basis of a single departure and landing. No allowance is made for the possibility that bookings may be part of larger trip itineraries.

2. *low before high booking:* The lowest fare reservations requests arrive first, followed by the next lowest, etc..

3. *nested booking:* Bookings are done in a nested fashion as follows: A fixed upper limit is set for bookings in the lowest fare class, a second, higher limit is set for the total bookings in the two lowest classes and so on, up to the highest fare class which is limited only by the total number of seats available. Any fare class can be booked into seats not taken by bookings in lower fare classes. (Equivalently, a fixed *protection level* of seats is set for the highest fare class, a second protection level for the total of two highest fare classes, and so on.)

4. *no cancellations:* Cancellations, 'no-shows' and overbooking are not considered.

5. *no revision of probabilities:* Probability distributions of demand are not revised on the basis of information gained during the booking process.

As discussed in Chapter 1, these assumptions are restrictive when compared to the actual decision problem faced by airlines, but analysis of this simplified version can both provide insights into the nature of optimal solutions and serve as a basis for approximate solutions to more realistic versions.

The first useful result on this seat allocation problem was presented in 1972 by Littlewood [89, (1972)] for two fare classes. He proposed that an airline should continue to reduce the protection level for class 1 (full fare) seats as long as the fare for class 2 (discount) seats satisfied

$$f_2 \geq f_1 \Pr[X_1 > p_1], \tag{2.1}$$

where f_i denotes the fare or average revenue from the i-th fare class, $\Pr[\cdot]$ denotes probability, X_1 is full fare demand, and p_1 is the full fare protection level. The intuition here is clear—accept the immediate return from selling an additional discount seat as long as the discount revenue equals or exceeds the *expected* full fare revenue from the seat.

The following interpretation of (2.1) will prove useful in the sequel. If the combined lower fare demands reach the limit $C - p_1$ on every flight, then the probability $\Pr[X_1 > p_1]$ is the expected proportion of flights on which some full fare demand is turned away, or *spilled.* The *actual* proportion of flights on which such spillage occurs is termed the *flight spill rate,* thus the probability above represents the highest possible flight spill rate given the distribution of the Y demand (highest because discount bookings might not reach the booking limit $(C - p_1)$ on every flight). This is referred to henceforth as the *maximum flight spill rate.* Clearly, with independent demands the maximum flight spill rate will increase as p_1 decreases. Littlewood's rule specifies that the optimal booking limit is the smallest value of p_1 for which this maximal rate does not exceed the ratio of discount to full fare.

A continuous version of Littlewood's rule was derived by Bhatia and Parekh [18, (1973)] in 1973. Richter [108, (1982)] in 1982 gave a marginal analysis which proved that (2.1) gives an optimal allocation.

More recently, Belobaba [13, (1987)] proposed a generalization 0f (2.1) to more than two fare classes called the *expected marginal seat revenue* (EMSR) method. In this approach, the protection level for the highest fare class, p_1, is obtained from

$$f_2 = f_1 \Pr[X_1 > p_1]. \tag{2.2}$$

This is just Littlewood's rule expressed as an equation, and it is appropriate as long as it is reasonable to approximate the protection level with a continuous variable and to attribute a probability density to the demand X_1. The total protection for the two highest fare classes, p_2, is obtained from

$$p_2 = p_2^1 + p_2^2, \tag{2.3}$$

where p_2^1 and p_2^2 are two individual protection levels determined from

$$f_3 = f_1 \Pr[X_1 > p_2^1] \tag{2.4}$$

and

$$f_3 = f_2 \Pr[X_2 > p_2^2]. \tag{2.5}$$

The protection for the three highest fare classes is obtained by summing three individual protection levels, and so on. This process is continued until nested protection levels p_k, are obtained for all classes except the lowest. The booking limit for any class k is then just $(C - p_{k-1})$, where C is the total number of seats available. It was originally proposed that this method yielded optimal protection levels for the seat allocation problem defined above, but, as will be shown here, this is not the case. While the idea of comparing the expected marginal revenues from future bookings with current marginal revenues is valid, the method outlined above leads to a correct assessment of expected future revenues only for the highest fare class. The analysis provided in this thesis corrects this problem and produces an exact EMSR solution. To avoid confusion, the EMSR approximation described above will henceforth be referred to as the EMSRa method.

In this chapter it is shown that an optimal set of protection levels $p_1^*, p_2^*, \ldots$ must satisfy the conditions

$$\delta_k^+ ER_k[p_k^*] \leq f_{k+1} \leq \delta_k^- ER_k[p_k^*] \text{ for each } k = 1, 2, \ldots; \tag{2.6}$$

where $ER_k[p_k]$ is the expected revenue from the k highest fare classes when p_k seats are protected for those classes, and $\delta_k^+$and δ_k^- denote the right and left derivative with respect to p_k, respectively. These conditions are just an expression of the usual first-order result — a change in p_k away from p_k^* in either direction will produce a smaller increase in expected revenues than the immediate increase of f_{k+1}. Of course the sufficiency of these conditions must be confirmed by a demonstration of the concavity of the expected revenue function.

It is further shown that these optimal protection levels can be obtained by finding $p_1^*, p_2^*, \ldots$ that satisfy

$$\begin{aligned} f_2 &= f_1 \Pr[X_1 > p_1^*] \\ f_3 &= f_1 \Pr[X_1 > p_1^* \cap X_1 + X_2 > p_2^*] \\ &\vdots \\ f_{k+1} &= f_1 \Pr[X_1 > p_1^* \cap X_1 + X_2 > p_2^* \cap \cdots \cap X_1 + X_2 + \cdots + X_k > p_k^*]. \end{aligned} \tag{2.7}$$

Note that the first of these equations is identical to the first in the EMSRa method, so the EMSRa method does derive the optimal protection level for the highest fare class.

This chapter is organized as follows. In the next section notation and assumptions are presented. In section 2.3 the revenue function and its directional derivatives are given. In the following section concavity properties of the expected revenue function are established and results (2.6) and (2.7) are obtained. In the final section numerical comparisons of the EMSRa and optimal solutions are provided.

2.2 Notation and Assumptions

The demand for fare class k is X_k, $(k = 1, 2, \ldots)$, where X_1 is the highest fare class. The vector of demands is $\mathbf{X} = (X_1, X_2, \ldots)$. It is assumed that these demands are stochastically independent. Each booking of a fare class k seat generates average revenue of f_k where $f_1 > f_2 > \cdots$.

Demands for the lowest fare class arrive first, and seats are booked for this class until a fixed time limit is reached, bookings have reached some limit, or the demand is exhausted. Sales to this fare class are then closed, and sales to the class with the next lowest fare are begun, and so on for all fare classes. It is assumed that any time limits on bookings for fare classes are pre-specified. That is, the setting of such time limits is not part of the problem considered here. It is possible, depending on the airplane capacity, fares, and demand distributions that some fare classes will not be opened at all.

A *booking policy* is a set of rules which specify at any point during the booking process whether a fare class that has not reached its time limit should be available for bookings. In general, such policies may depend on the pattern of prior demands or be randomized in some manner (any stopping rule for fare class k which is measurable with respect to the sigma field generated by $[X_k > x]$, for $x = 0, 1, \ldots$ is admissible). However, we restrict attention to a class of booking policies, denoted by $\mathcal{P}$, that can be described by a vector of fixed protection levels $\mathbf{p} = (p_1, p_2, \ldots)$, where p_k is the number of seats to be protected for fare classes 1 through k. If at some stage in the process described above there are s seats available to be booked and there is a fare class k demand, then the seat will be booked if s is greater than the protection level p_{k-1} for the higher fare classes[1]. The initial number of classes that are open for any bookings is, of course, determined by setting s equal to the capacity of the aircraft or compartment. It will be shown formally that the class $\mathcal{P}$ contains a policy that is optimal over the class of all admissible policies.

2.3 The Revenue Function

The function $R_k[s;\mathbf{p};\mathbf{x}]$ is the revenue generated by the k highest fare classes when s seats are available to satisfy all demand from these classes, when $\mathbf{x} = (x_1, x_2, \ldots)$ is the demand vector, and when $\mathbf{p} = (p_1, p_2, \ldots)$ is the vector of protection levels. We define the revenue function recursively by

$$R_1[s;\mathbf{p};\mathbf{x}] = \begin{cases} f_1 s & \text{for } 0 \le s < x_1 \\ f_1 x_1 & \text{for } x_1 \le s \end{cases} \tag{2.8}$$

$$R_{k+1}[s;\mathbf{p};\mathbf{x}] = \begin{cases} R_k[s;\mathbf{p};\mathbf{x}] & \text{for } 0 \le s < p_k \\ (s - p_k) f_{k+1} + R_k[p_k;\mathbf{p};\mathbf{x}] & \text{for } p_k \le s < p_k + x_{k+1} \\ x_{k+1} f_{k+1} + R_k[s - x_{k+1};\mathbf{p};\mathbf{x}] & \text{for } p_k + x_{k+1} \le s, \end{cases} \tag{2.9}$$

for $k = 1, 2, \ldots$.

For convenience of notation, a dummy protection level p_0 will be introduced; its value will be identically zero throughout. There is no limit to the number of fare classes or to the corresponding lengths of the protection and demand vectors; however, the revenue from the k highest fares depends only on the protection levels $(p_0, p_1, \ldots, p_{k-1})$ and the demands $(x_1, x_2, \ldots, x_k)$. The symbols $\mathbf{p}$ and $\mathbf{x}$ will be used to denote vectors of lengths which vary depending on context, as in

$$R_k[s;\mathbf{p};\mathbf{x}] = R_k[s;(p_0, p_1, \ldots, p_{k-1});(x_1, x_2, \ldots, x_k)].$$

---

[1] Restriction to this class of policies is implicit in previous research.

The objective is to find a vector $\mathbf{p}$ that maximizes the expected revenue $ER_k[s;\mathbf{p};\mathbf{X}]$ for all k. If s is viewed as a real-valued variable, the function $ER_k[s;\mathbf{p};\mathbf{X}]$ is continuous and piecewise linear on $s > 0$ and not differentiable at the points $s = p_k$. Maximization of this function can be accomplished either by treating all variables as integer-valued and using arguments based on first differences, or by treating all quantities except demands as continuous and using standard tools of nonsmooth optimization. Both approaches are equivalent for this problem and yield the same optimality conditions. The second approach will be used here because it permits greater economy of notation and terminology.

2.3.1 Marginal Value of an Extra Seat

This section develops the first-order properties of the revenue function. The notation and terminology used here and in what follows are consistent with Rockafellar [109, (1970)]. Let δ^+ and δ^- denote the left and right derivatives with respect to the first argument of the revenue or expected revenue functions. Thus $\delta^- ER[s;(p_0,\ldots,p_{k-1});\mathbf{X}]$ is the left derivative of $ER[\cdot]$ with respect to s. (This slightly unconventional notation is required because s, the number of seats remaining, will sometimes be replaced by p_k when the argument is being viewed as a discretionary quantity.) For fixed $\mathbf{p}$ and $\mathbf{x}$, the derivatives for the revenue function are easily computed from (2.8) and (2.9) to be

$$\delta^+ R_1[s;\mathbf{p};\mathbf{x}] = \begin{cases} f_1 & \text{for } s < x_1 \\ 0 & \text{for } s \geq x_1 \end{cases} \tag{2.10}$$

$$\delta^- R_1[s;\mathbf{p};\mathbf{x}] = \begin{cases} f_1 & \text{for } s \leq x_1 \\ 0 & \text{for } s > x_1 \end{cases} \tag{2.11}$$

and

$$\delta^+ R_{k+1}[s;\mathbf{p};\mathbf{x}] = \begin{cases} \delta^+ R_k[s;\mathbf{p};\mathbf{x}] & \text{for } 0 \leq s < p_k \\ f_{k+1} & \text{for } p_k \leq s < p_k + x_{k+1} \\ \delta^+ R_k[s - x_{k+1};\mathbf{p};\mathbf{x}] & \text{for } p_k + x_{k+1} \leq s. \end{cases} \tag{2.12}$$

$$\delta^- R_{k+1}[s;\mathbf{p};\mathbf{x}] = \begin{cases} \delta^- R_k[s;\mathbf{p};\mathbf{x}] & \text{for } 0 < s \leq p_k \\ f_{k+1} & \text{for } p_k < s \leq p_k + x_{k+1} \\ \delta^- R_k[s - x_{k+1};\mathbf{p};\mathbf{x}] & \text{for } p_k + x_{k+1} < s. \end{cases} \tag{2.13}$$

Any continuous, piecewise-linear function $G[s]$ is concave on $s > 0$ if and only if the right derivative is less than or equal to the left derivative for any s. This condition can be extended to the point $s = 0$ by defining $\delta^- G[0] = +\infty$. The *subdifferential,* $\delta G[s]$, is then defined for any $s \geq 0$ as the closed interval from $\delta^+ G[s]$ to $\delta^- G[s]$. Given concavity, $G[\cdot]$ will be maximized at any point s for which $0 \in \delta G[s]$.

2.4 Optimal Protection Levels

This section establishes the optimality within the class $\mathcal{P}$ of protection levels determined by the first-order conditions given in (2.6).

We first consider a point in the booking process when s seats remain unbooked, fare class $k+1$ is being booked, and the decision of whether or not to stop booking that class is to be made. That is, a decision on the value of the protection level p_k for the remaining fare classes is to be made. The following lemma establishes a condition under which concavity of the expected revenue function with respect to s is ensured, conditional on the value of X_{k+1}. This leads to an argument by induction that concavity of the conditional expected revenue function will be satisfied if condition (2.6) is satisfied for all of the higher protection levels. Finally, it is shown that condition (2.6) also guarantess optimality of p_k.

Lemma 2.4.1 *If some policy,* $\mathbf{p}$*, makes* $ER_k[s;(p_0,\ldots,p_{k-1});\mathbf{X}]$ *concave on* $s \geq 0$ *and if* p_k^* *satisfies*

$$f_{k+1} \in \delta ER_k[p_k^*;(p_0,\ldots,p_{k-1});\mathbf{X}], \tag{2.14}$$

then $E\left\{R_{k+1}[s;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\}$ *is concave on* $s \geq 0$ *with probability 1.*

Proof: It follows from the definition of the revenue function in (2.9) and the hypothesized concavity of ER_k that $\mathrm{E}\left\{R_{k+1}[s;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\}$ is continuous on $s > 0$ and concave on the three intervals $0 \leq s < p_k$, $p_k \leq s < p_k + X_{k+1}$, and $p_k + X_{k+1} \leq s$.

To complete the proof, it is enough to verify that

$$\begin{aligned} &\delta^+\mathrm{E}\left\{R_{k+1}[s;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} \\ &\quad\leq \ \delta^-\mathrm{E}\left\{R_{k+1}[s;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} \end{aligned} \tag{2.15}$$

at the two points $s = p_k^*$ and $s = p_k^* + X_{k+1}$. From (2.12) and (2.13) the left and right derivatives at $s = p_k^*$ are

$$\delta^-\mathrm{E}\left\{R_{k+1}[p_k^*;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} = \delta^- ER_k[p_k^*;\mathbf{p};\mathbf{X}] \tag{2.16}$$

and

$$\delta^+\mathrm{E}\left\{R_{k+1}[p_k^*;(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} = f_{k+1}. \tag{2.17}$$

By the hypothesis of the lemma, inequality (2.15) must be satisfied.

Again applying (2.12) and (2.13), the left and right derivatives at $s = p_k^* + X_{k+1}$ are

$$\delta^-\mathrm{E}\left\{R_{k+1}[p_k^* + X_{k+1};(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} \ = \ f_{k+1} \tag{2.18}$$

and

$$\begin{aligned} &\delta^+\mathrm{E}\left\{R_{k+1}[p_k^* + X_{k+1};(p_0,\ldots,p_{k-1},p_k^*);\mathbf{X}] \mid X_{k+1}\right\} \ = \\ &\qquad\qquad \delta^+ ER_k[p_k;(p_0,\ldots,p_{k-1});\mathbf{X}]. \end{aligned} \tag{2.19}$$

By the hypothesis of the lemma, inequality (2.15) must be satisfied at $s = p_k^* + X_{k+1}$. ■

Corollary 2.4.2 *If, for some $k \in \{1, 2, \ldots\}$ the conditions of the lemma hold, then*

$$ER_{k+1}[s; (p_0, \ldots, p_{k-1}, p_k^*); \mathbf{X}]$$

is concave on $s \geq 0$.

Proof: We have

$$ER_{k+1}[s; (p_0, \ldots, p_{k-1}, p_k^*); \mathbf{X}] = \mathrm{E}\left[\, \mathrm{E}\{R_{k+1}[s; (p_0, \ldots, p_{k-1}, p_k^*); \mathbf{X}] \,|\, X_{k+1}\} \,\right]$$

It follows from the concavity of the conditional expectation on the right-hand side that

$$\delta^+ ER_{k+1}[s; (p_0, \ldots, p_{k-1}, p_k^*); \mathbf{X}] \leq \delta^- ER_{k+1}[s; (p_0, \ldots, p_{k-1}, p_k^*); \mathbf{X}]$$

(The expectation operator E and the differential operators δ^+ and δ^- can be interchanged since R_1 is bounded by $f_1 s$ for all policies $\mathbf{p}$ and demand $\mathbf{x}$.) ∎

Theorem 2.4.3 *Let $\mathbf{p}$ be any policy satisfying*

$$f_{k+1} \in \delta ER_k[p_k; (p_0, \ldots, p_{k-1}); \mathbf{X}] \tag{2.20}$$

for $k = 1, 2, \ldots$. Then $E\{R_{k+1}[s; \mathbf{p}; \mathbf{X}] \,|\, X_{k+1}\}$ is concave on $s \geq 0$ for $k = 1, 2, \ldots$. Moreover, it is optimal to stop the sales of fare class $k+1$ whenever p_k seats remain unsold.

Proof: From (2.10) and (2.11),

$$\mathrm{E}\{\delta^+ R_1[s; \mathbf{p}; \mathbf{X}] \,|\, X_1\} \;=\; f_1 I_{[X_1 > s]}$$

and

$$\mathrm{E}\{\delta^- R_1[s; \mathbf{p}; \mathbf{X}] \,|\, X_1\} \;=\; f_1 I_{[X_1 \geq s]},$$

where $I_{[A]} = 1$ if condition A holds, and $I_{[A]} = 0$ otherwise. Then $\delta^+ \mathrm{E}\{R_1[s; \mathbf{p}; \mathbf{X}] \,|\, X_1\} \leq \delta^- \mathrm{E}\{R_1[s; \mathbf{p}; \mathbf{X}] \,|\, X_1\}$. Thus, $\mathrm{E}\{R_1[s, \mathbf{p}; \mathbf{X}] \,|\, X_1\}$ is concave in s for any policy $\mathbf{p}$, and, given condition (2.20), the concavity assertion in the theorem follows from Lemma 2.4.1 by induction.

To prove optimality of the protection level p_k it is necessary to examine the behaviour of $ER_{k+1}[s;(p_0,\ldots,p_k);\mathbf{X}]$ as a function of p_k for any s. Denote the left derivative, right derivative and subdifferential with respect to p_k by δ_k^-, δ_k^+, and δ_k, respectively.

From (2.9),

$$\delta_k^+ R_{k+1}[s;\mathbf{p};\mathbf{x}] = \begin{cases} 0 & \text{for } 0 \le s \le p_k \\ -f_{k+1} + \delta^+ R_k[p_k;\mathbf{p};\mathbf{x}] & \text{for } p_k < s < p_k + x_{k+1} \\ 0 & \text{for } p_k + x_{k+1} \le s. \end{cases} \tag{2.21}$$

$$\delta_k^- R_{k+1}[s;\mathbf{p};\mathbf{x}] = \begin{cases} 0 & \text{for } 0 < s < p_k \\ -f_{k+1} + \delta^- R_k[p_k;\mathbf{p};\mathbf{x}] & \text{for } p_k \le s \le p_k + x_{k+1} \\ 0 & \text{for } p_k + x_{k+1} < s. \end{cases} \tag{2.22}$$

Recall that $R_k[p_k;\mathbf{p};\mathbf{x}]$ is independent of x_{k+1}. Taking the expectations of these derivatives and reversing the order of differentiation and expectation yields

$$\begin{aligned} \delta_k^+ ER_{k+1}[s;\mathbf{p};\mathbf{X}] &= (-f_{k+1} + \delta^+ ER_k[p_k;\mathbf{p};\mathbf{X}])\Pr[X_{k+1} > s - p_k] & (2.23) \\ \delta_k^- ER_{k+1}[s;\mathbf{p};\mathbf{X}] &= (-f_{k+1} + \delta^- ER_k[p_k;\mathbf{p};\mathbf{X}])\Pr[X_{k+1} \ge s - p_k] & (2.24) \end{aligned}$$

Now $\delta^+ ER_k[p_k;\mathbf{p};\mathbf{X}] \le \delta^- ER_k[p_k;\mathbf{p};\mathbf{X}]$ for all p_k by concavity of $ER_k[s;\mathbf{p};\mathbf{X}]$ with respect to s. But then $\delta_k^+ ER_{k+1}[s;\mathbf{p};\mathbf{X}] \le \delta_k^- ER_{k+1}[s;\mathbf{p};\mathbf{X}]$ from (2.20), (2.23) and (2.24); and hence $ER_{k+1}[s;(p_0,\ldots,p_k);\mathbf{X}]$ is concave with respect to p_k. Furthermore, condition (2.20) implies

$$\delta_k^+ ER_{k+1}[s;\mathbf{p};\mathbf{X}] \le 0 \le \delta_k^- ER_{k+1}[s;\mathbf{p};\mathbf{X}];$$

that is, $0 \in \delta_k ER_{k+1}[s;\mathbf{p};\mathbf{X}]$, and p_k maximizes $ER_{k+1}[s;\mathbf{p};\mathbf{X}]$, as required. ∎

2.4.1 Evaluation of the Protection Levels

In this section a more explicit expression for computing the optimal protection levels is derived.

Lemma 2.4.4 *If* $\mathbf{p}$ *satisfies*

$$f_1 \Pr[X_1 > p_1 \cap X_1 + X_2 > p_2 \cap \cdots \cap X_1 + X_2 + \cdots + X_k > p_k] = f_{k+1}, \tag{2.25}$$

for all k, then with probability 1 for $k = 1, 2, \ldots$ and $s \geq p_k$

$$\begin{aligned} \delta^+ E[R_{k+1}[s;\mathbf{p};\mathbf{X}] \,|\, X_{k+1}] &= f_1 \Pr[X_1 > p_1 \cap \cdots \\ &\cap X_1 + X_2 + \cdots + X_k > p_k \cap X_1 + \cdots + X_{k+1} > s \,|\, X_{k+1}] \end{aligned} \tag{2.26}$$

Proof:

Assume that $\mathbf{p}$ satisfies the hypothesis of the Lemma. For $s \geq p_k$, we can obtain the following expression from (2.12) by taking the expectation and interchanging E and δ^+ :

$$\begin{aligned} \delta^+ \mathrm{E}\{R_{k+1}[s;\mathbf{p};\mathbf{X}] \,|\, X_{k+1}\} &= f_{k+1} I_{[s<p_k+X_{k+1}]} \\ &+ \delta^+ \mathrm{E}\{R_k[s - X_{k+1};\mathbf{p};\mathbf{X}] \,|\, X_{k+1}\} \, I_{[s \geq p_k + X_{k+1}]}. \end{aligned} \tag{2.27}$$

Using (2.25) to substitute for f_{k+1}, the right-hand side of this expression can be rewritten as

$$\begin{aligned} &f_1 \Pr[X_1 > p_1 \quad \cap \cdots \cap X_1 + \cdots + X_k > p_k \cap s < p_k + X_{k+1} \,|\, X_{k+1}] \\ &\qquad + \delta^+ \mathrm{E}\{R_k[s - X_{k+1};\mathbf{p};\mathbf{X}] \,|\, X_{k+1}\} \, I_{[s \geq p_k + X_{k+1}]}. \end{aligned} \tag{2.28}$$

For $k = 1$, using (2.10) and the fact that $[X_1 + X_2 > s \cap s \geq p_1 + X_2] \Rightarrow [X_1 > p_1]$, (2.27) becomes

$$\begin{aligned} \delta^+\mathrm{E}\{R_2[s;\mathbf{p};\mathbf{X}] \mid X_2\} &= f_1 \Pr[X_1 > p_1 \cap X_1 + X_2 > s \cap s < p_1 + X_2 \mid X_2] \\ &\quad + f_1 \Pr[X_1 > p_1 \cap X_1 + X_2 > s \cap s \geq p_1 + X_2 \mid X_2] \\ &= f_1 \Pr[X_1 > p_1 \cap X_1 + X_2 > s \mid X_2]. \end{aligned} \tag{2.29}$$

Thus the Lemma holds for $k = 1$.

The proof is completed by induction. Using the induction hypothesis that the Lemma holds for k, substitute for $\delta^+ R_k$ in the last term of (2.28).

$$\begin{aligned} & E\{\delta^+ R_{k+1}[s;\mathbf{p};\mathbf{X}] \mid X_{k+1}\} \\ &= f_1 \Pr[X_1 > p_1 \cap \cdots \\ &\quad \cap X_1 + \cdots + X_k > p_k \cap s < p_k + X_{k+1} \mid X_{k+1}] + \\ &\quad f_1 \Pr[X_1 > p_1 \cap \cdots \\ &\quad \cap X_1 + \cdots + X_k > s - X_{k+1} \cap s - X_{k+1} \geq p_k \mid X_{k+1}] \\ &= f_1 \Pr[X_1 > p_1 \cap \cdots \\ &\quad \cap X_1 + \cdots + X_k > p_k \cap X_1 + \cdots + X_{k+1} > s \mid X_{k+1}], \end{aligned} \tag{2.30}$$

completing the proof of the lemma. ■

Corollary 2.4.5 *If* $\mathbf{p}$ *satisfies* (2.25), *then for* $s \geq p_k$

$$\begin{aligned} & \delta^+ E R_{k+1}[s;\mathbf{p};\mathbf{X}] = \\ & \quad f_1 \Pr[X_1 > p_1 \cap \cdots \cap X_1 + \cdots + X_k > p_k \cap X_1 + \cdots + X_{k+1} > s]. \end{aligned} \tag{2.31}$$

Theorem 2.4.6 *If* $\mathbf{p}$ *satisfies* (2.25), *then* $\mathbf{p}$ *is optimal.*

Proof: By Lemma 2.4.4 if **p** satisfies (2.25), then

$$\begin{aligned} f_{k+1} &= f_1 \Pr[X_1 > p_1 \cap \ldots \cap X_1 + \cdots + X_k > p_k] \\ &= \delta^+ ER_k[p_k; \mathbf{p}; \mathbf{X}]. \end{aligned} \tag{2.32}$$

By Theorem 2.4.3, **p** is thus optimal. ■

2.4.2 Monotone Optimal Stopping Problems and the Optimality of Fixed Protection Level Booking Policies

In this section it is established that the fixed protection levels **p** defined by condition (2.20) are optimal over the set of all admissible policies, not just over the set of fixed policies $\mathcal{P}$. To this end, first consider the problem of stopping bookings in fare class 2 when there are s seats remaining and $X_2 \geq x_2$ has been observed, where $x_2 \geq 0$. If the protection level p_1 is to be chosen by any admissible policy, it may in general depend on the number of seats remaining and the value of x_2. That is, *a priori* p_1 must be regarded as a random variable on the sigma-field generated by $\{X_2 \geq x_2\}$. In the usual terminology of stochastic processes, p_1 is classified as a *stopping time* for the X_2 booking process. The problem of finding an optimal policy for choosing p_1 belongs to a class of stochastic optimization problems known as optimal stopping problems. In the discrete case, it has been shown by Derman and Sacks [37, (1960)] and Chow and Robbins [26, (1961)] that optimal stopping problems defined as *monotone* have particularly simple solutions. The seat allocation problem is in reality a discrete problem, so a proof of monotonicity could be obtained by using these results. However, to avoid departing from the continuity assumptions and notation used throughout this chapter, use is made instead of a continuous generalization reported by Ross [111, (1971)]. To do this, the additional stipulation must be made that the probability distribution functions of the demands are continuous.

In the context of the booking problem for fare class 2, Ross's conditions for monotonicity will be satisfied if:

1. There is a p_1^* such that

$$\delta^{+}{}_1 ER_2[s;(p_1);\mathbf{X}] \geq 0 \text{ for } p_1 < p_1^*,$$

and

$$\delta^{+}{}_1 ER_2[s;(p_1);\mathbf{X}] \leq 0 \text{ for } p_1 \geq p_1^*.$$

   (Recall that $\delta^{+}{}_1$ denotes the right derivative with respect to p_1.)

2. $R_2[s;(p_1);\mathbf{x}]$ and $\delta^{+}{}_1 ER_2[s;(p_1);\mathbf{X}]$ are bounded and continuous in p_1.

3. p_1^* is finite with probability 1.

Now condition 1 follows from the proof of theorem 2.4.3. For condition 2, $R_2[s;(p_1);\mathbf{x}]$ is continuous in p_1 by inspection of (2.9) and bounded if the fares and demands are bounded. Also, from (2.23), $\delta^{+}{}_1 ER_2[s;(p_1);\mathbf{X}] = (-f_1 + f_2 \Pr[X_1 \geq p_1]) \Pr[X_2 > s - p_1]$. This directional derivative will be bounded if the fares are bounded, and is continuous by the continuity assumption on the demand distribution. Finally, p_1^* will be finite if fares and demands are bounded, by inspection of the expression for $\delta^{+}{}_1 ER_2[s;(p_1);\mathbf{X}]$ given above.

If the model is monotone the expected revenue will be maximized by protecting p_1^* seats for X_1 demand; that is, a fixed-limit policy will be optimal for the protection level p_1. As demonstrated previously, choice of the protection level p_1^* ensures concavity of $ER_2[s;(p_1);\mathbf{X}]$ on $s > 0$, hence the stopping problem for fare class three will be monotone and p_2^* will be optimal for the protection level p_2. The optimality of the fixed-limit policies for the remaining fare classes follows by induction using the concavity results proved in the previous section.

The significance of this result in the context of airline seat allocation is that static protection levels defined by condition (2.25) will be optimal as long as no change in the probability distributions of demand is forseen. In other words, no *ad hoc* adjustment of protection levels is justified unless a shift in the demand distributions is detected. This is not a particularly suprising result, given the assumed independence of demand and other restrictive assumptions of the underlying model; however, it is less obvious that a similar result holds in the case of two classes of demand that are stochastically dependent. This result and others are the subject of chapter 3.

2.4.3 Implementation of Independent Demand Optimal Protection Levels

The problem of solving for the optimal protection levels is reduced to finding a solution $\mathbf{p}^* = (p_1^*, p_2^*, \ldots)$ to (2.25) for $k = 1, 2, \ldots$. A condition which guarantees the solvability is that the demand distribution have a density function. If an empirical distribution for integer demand is being used, then the above equations can likely be solved to within the statistical error of the demand distribution.

Empirical studies have shown that the normal probability distribution gives a good continuous approximation to airline demand distributions [13, 123]. If normality is assumed, solutions to (2.25) can be obtained with straightforward numerical methods. It is important to note, however, that with a large number of fare classes solution of (2.25) could become time-consuming because of the need to perform a large number of numerical integrations.

There is a way in which the optimality conditions (2.25) can be used to monitor the past performance of seat allocation decisions given historical data on seat bookings for a series of flights. Detailed discussion of this idea is deferred until after the optimality conditions for dependent demands are obtained in chapter 3; however, the approach will be outlined here. For simplicity the discussion will assume three fare classes; the method generalizes easily to an arbitrary number of classes.

With three fare classes, conditions (2.25) can be written

$$\Pr[X_1 > p_1] = \frac{f_2}{f_1} \tag{2.33}$$

$$\Pr[X_1 > p_1 \cap X_1 + X_2 > p_2] = \frac{f_3}{f_1} \tag{2.34}$$

Given a series of past flights, the probability $\Pr[X_1 > p_1]$ can be interpreted as the proportion of flights on which class 1 demand exceeded its protection level. Then (2.33) specifies that this proportion should be close to the ratio f_2/f_1. Similarly, (2.34) specifies that the proportion of flights on which both class 1 demand exceeded its protection level and the total of class 1 and 2 demands exceeded their protection level should be close to the ratio f_3/f_1. If allocation decisions are being made optimally, these conditions should be approximately satisfied in a sufficiently long series of past flights. Severe departures from these ratios would be symptomatic of sub-optimal allocation decisions. The appealing aspect of this approach is its simplicity — no modeling of the demand distributions and no numerical integrations are required.

2.5 Comparison of EMSRa and Optimal Solutions

The EMSRa method determines the optimal protection level for the full fare class but is not optimal for the remaining fare classes. However, the EMSRa equations are particularly simple to implement because they do not involve joint probability distributions. It is thus of interest to examine the performance of the EMSRa method relative to the optimal solutions given above. Note that neither the EMSRa nor optimal equations give explicit solutions for the optimal protection levels, so analytical comparison of the revenues produced by the two methods is difficult unless unrealistic demand distributions are assumed. Numerical comparison of the two methods can, however, give some indication of relative performance.

Table 2.1: Comparison of EMSRa versus Optimal - Three Fare Classes

example number	f_3	f_2	p_1	p_2 EMSRa	p_2 optimal	% error revenue
1	0.6	0.7	32	70	80	0.37 %
2	0.6	0.8	27	80	87	0.32 %
3	0.6	0.9	19	86	91	0.19 %
4	0.7	0.8	27	64	75	0.41 %
5	0.7	0.9	19	73	82	0.45 %
6	0.8	0.9	19	57	70	0.50 %

Table 2.2: Effect of Capacity on EMSRa Error

capacity	% error revenue
82	0.54 %
100	0.45 %
120	0.35 %
140	0.24 %
160	0.14 %

This section gives the results of numerical comparisons of EMSRa versus optimal solutions in a three fare-class problem[2]. Table 2.1 presents the results of six examples in which cabin capacity is fixed at 100 seats and fares f_i are varied. Fares are expressed as proportions of full fare; thus, $f_1 = 1$ throughout. The '% error revenue' column gives the loss in revenues incurred from using the EMSRa method as a percentage of optimal revenues. In Table 2.2, the fares are held constant at levels $f_3 = 0.7$ and $f_2 = 0.9$, and cabin capacity is varied.

Discrete approximations to the normal probability distribution were used for all demand distributions. The nominal mean demands for fare classes 1, 2 and 3 were 40, 60 and 80, and the nominal standard deviations, 16, 24 and 32, respectively. These figures are nominal because the discretization procedure introduced small deviations from the exact parameter values. These parameters correspond to a coefficient of variation of 0.4; i.e., the standard deviation is 40% of the mean. This is slightly higher than the 0.33 that Belobaba [13, (1987)p143] mentions as a common airline 'k-factor' for total demand.

Remarks:

In this set of examples the EMSRa method produces seat allocations that are significantly different from optimal allocations, but the loss in revenue associated is not great. Specifically:

1. In these examples, the EMSRa method consistently underestimates the number of seats that should be protected for the two upper fare classes. The discrepancy is 19% in the worst case (example #6). It will be shown below with a counterexample that the EMSRa method is not guaranteed to underestimate in this way.
2. In the worst case the discrepancy between EMSRa and optimal solutions with respect to revenues is approximately 1/2 percent.
3. The error appears to increase as the discount fares approach the full fare; however, the sample is much too small here to justify any general conclusion of this nature.
4. The error decreases as the aircraft capacity increases. This effect is to be expected since allocation policies have less impact when the capacity is able to accomodate

---

[2] These calculations were carried out on a microcomputer. Computations for four or more fare classes were not carried out because of the excessive running times that would have been required with realistic aircraft capacities. (The size of the joint demand distributions that must be dealt with are of the order C^{k-1}, where k is the number of fare classes.)

most of the demands.

On the basis of these examples, a decision of whether or not to use the EMSRa approach rests on whether or not a potential revenue loss in the order of 1/2 percent or less (with three fare classes) is justified by the simpler implementation of the method relative to the optimal method. Further work is needed to determine the relative performance of the EMSRa method with a larger number of fare classes or under other conditions.

EMSRa Underestimation of Protection Levels — A Counterexample

As mentioned above, the EMSRa method consistently underestimated the protection level p_2 for the two upper fare classes in all of the numerical trials. It is thus reasonable to conjecture that the approximation will always behave in this way. This is not true for all demand distributions, as shown by the following counterexample using exponentially distributed demands. It remains an open question whether or not the conjecture holds true for normally distributed demands.

For convenience, let the unit of demand be 100 seats, and introduce the relative fares $r_2 = f_2/f_1$ and $r_3 = f_3/f_1$. Now suppose that X_1 and X_2 follow identical, independent exponential distributions with mean 1.0 (100 seats). That is $\Pr[X_i > x_i] = e^{-x_i}$ for $i = 1, 2$. It is not suggested that the exponential distribution has any particular merit for modeling airline demands, although it could serve as a surrogate for a severely right-skewed distribution if the need arose. Its use here is purely as a device for establishing a counterexample to a general conjecture.

Let p_i^a denote protection levels obtained with the EMSRa method. Then with the above distributional assumptions and equations (2.2) through (2.5), we have $p_1^a = -\ln(r_2)$, and $p_2^a = -\ln(r_3) - \ln(r_3/r_2)$.

For the optimal solutions, condition (2.7) gives $p_1 = -\ln(r_2) = p_1^a$, and

$$\begin{aligned} r_3 &= \Pr[X_1 > p_1 \cap X_1 + X_2 > p_2] \\ &= \Pr[X_1 > p_2] + \Pr[p_1 < X_1 \leq p_2 \cap X_2 > p_2 - X_1] \\ &= e^{-p_2} + \int_{p_1}^{p_2} \Pr[X_2 > p_2 - x_1] e^{-x_1}\, dx_1 \\ &= e^{-p_2}(1 + p_2 - p_1). \end{aligned} \tag{2.35}$$

Suppose that $r_2 = 1/2$ and $r_3 = 1/4$. Then $p_1^a \cong 0.69$ and $p_2^a \cong 2.08$ (69 and 208 seats, respectively). Given p_1, a simple line search using (2.35) produces the optimal $p_2 \cong 2.37$ from the equation above. Thus, for this example, the EMSRa method underestimates p_2 by 29 seats. This behavior is consistent with the conjecture.

Now suppose instead that $r_2 = 4/10$ and $r_3 = 1/10$. Then $p_1^a \cong 0.92$ and $p_2^a \cong 3.69$. In this case, however, $p_2 \cong 3.61$, and the EMSRa method *overestimates* p_2 by 8 seats. It is not difficult to show that for these demand distributions, the EMSRa method will overestimate p_2 whenever $r_2/r_3 > 3.51$, approximately.

2.6 Summary — Independent Demands Case

This chapter provides a rigorous formulation of the revenue function for the multiple fare class seat allocation problem and demonstrates conditions under which the expected revenue function is concave. It is shown that a booking policy that maximizes expected revenue can be characterized by a simple set of conditions on the subdifferential of the expected revenue function. These conditions are further simplified to a set of conditions relating the probability distributions of demand for the various fare classes to their respective fares. It is proven that the optimal fixed protection limit policies are optimal over the class of all policies that depend only on observed demands. A numerical comparison is made of the optimal solutions with the approximate solutions yielded by the expected marginal seat revenue (EMSRa) method. A tentative conclusion on the basis of this restricted set of examples is that the EMSRa method produces seat allocations that are significantly different from optimal allocations with an associated loss in revenue of the order of 1/2 percent.

Chapter 3

Allocation Between Two Classes when Demands are Dependent

This chapter deals with the airline seat allocation problem in the case that there is stochastic dependency between the demands for two fare classes, henceforth referred to as the *discount fare* class, and the *full fare* class. The other assumptions are the same as those of the previous chapter (see page 16). The reduction in the number of fare classes might seem to limit the usefulness of this analysis, given that airlines typically offer eight or more major fare classes. However, the fares appearing in this and the previous analysis are only assumed to be *average* fares, so in the present analysis one can consider the discount fare class to be composed of several sub-classes with different fares. If full fare demand is correlated in some way with aggregate demand for the discount sub-classes, the analysis of this chapter provides a useful estimate of an appropriate protection level for the full fare class. More importantly, the present analysis provides a means of evaluating the general significance of demand dependencies and gives some indication of the appropriate response to such dependencies.

Any discussion of protection levels for the full fare class must inevitably involve the so-called *spill rate* for full fare passengers — the rate at which full fare passengers are refused bookings because of prior sales of discount seats. This chapter shows that a simple modification to the optimality formula for the full fare protection level can allow for additional control of full fare spillage or, conversely, for estimation of the revenue impact of a particular spill rate policy.

With some exceptions to be noted later, previous work on the two fare seat allocation problem has assumed stochastic independence between discount and full fare demands. However, there are at least two reasons that such independence might fail to hold. First, scheduled events such as conferences can be expected to stimulate demand for both fare classes since there is generally sufficient time for budget-conscious travellers to book in the early-booking discount class. The occurrence of many such stimuli would lead to a positive correlation between the demands for the fare classes, even after such effects as day-of-the-week, season, fares, etc. had been allowed for. Second, a proportion of customers seeking discount fare bookings can be expected to *upgrade* to full fare if they find that all discount seats have been sold. Such behaviour also causes positive dependency between discount demand and the ultimate full fare demand. In this case, the strength of the dependency is infuenced in part by the booking limit set for the discount fares — the lower the booking limit, the higher the number of upgrades and the higher the apparent full fare demand. Other arguments can be put forward for the existence of dependencies, both positive and negative, in special cases.

Belobaba [13, pp143-150] discussed the possible impact of demand dependencies on booking limits and showed that, in a three fare class problem, the booking limit for the lowest fare class is reduced as the correlation between demands for the two upper fare classes increases. This is a simple consequence of the increase in the variance of the total demand for the two higher fare classes that results from increasing correlation. He did not examine the problem of determining the booking limit *between* two dependent classes (the problem examined here). Belobaba also proposed a seat allocation formula for the case that demand dependency arises because of upgrades. A formal proof of the correctness of that result is provided here. A similar result using different methods has been obtained by Pfeifer [106, (1989)].

Before proceeding with a detailed analysis of the dependent demand case, we offer the following brief intuitive argument. Recall that Littlewood's rule for two independent fare classes, if approximated by an equation, specifies that an optimal booking limit ℓ^* for the discount fare class satisfies

$$f_2 = f_1 \Pr[X_1 > C - \ell^*].$$

The case considered here is much the same as that considered in deriving Littlewood's rule except that, here, the full fare demand distribution must be modified as each discount demand occurs because of the dependency between the demands. That is, after observing $X_2 \geq \ell$ the full fare demand distribution becomes $\Pr[X_1 > C - \ell \,|\, X_2 \geq \ell]$. It seems reasonable to conjecture that the optimal booking limit can be obtained simply by replacing the probability in Littlewood's formula with this conditional probability. It is shown here that this conjecture is valid as long as the discount and full fare demands are *monotonically associated.* This condition is precisely defined later, but loosely speaking, it implies that demand distributions must be such that the full fare spill rate increases as more seats are sold to discount customers.

This chapter is organized as follows. Section 3.1 describes a general seat allocation model that forms the basis for later analyses. In Section 3.2 the allocation model is used to derive a generalization of Littlewood's rule for the dependent demand case. It is shown that a variant of the rule is valid when the discount and full fare demands are monotonically associated. A numerical example is provided in Section 3.2.2. In Section 3.2.4 a minor adjustment to the same model accommodates the *goodwill costs* incurred when a full fare passenger is unable to obtain a reservation on his or her preferred flight. Section 3.2.5 deals with the case that the dependency (between full and discount fare demands) arises because of the tendency for a proportion of discount fare customers to *upgrade* to full fare if no more discount seats are available. The general model is used to prove optimality of a seat allocation formula that incorporates upgrades. In Section 3.3 it is proven that the monotonic association condition is satisfied when demands follow a bivariate normal distribution with positive correlation. It is also proven that the optimal full fare protection level increases as the correlation increases. Finally, section 3.4 summarizes the main results of this chapter.

3.1 A Generic Seat Allocation Model with Dependent Demands

This section presents a general model for the seat allocation problem that serves as a basis for the specific analyses of later sections. As with the independent demand model of Chapter 2, it is similar in structure to *optimal stopping* models described in Chow *et al.* [27, (1971)], and Derman and Sacks [37, (1960)], and this correspondence is once again used to characterize instances of the problem for which a simple rule yields an optimal solution — the class of monotone optimal stopping problems.

The notational conventions of Chapter 2 are maintained here except that since only two fare classes are under consideration, subscripts are dispensed with, and the highest fare demand is denoted Y and the discount demand, X. The corresponding average fares are f_Y and f_X respectively. It is more convenient in the two fare class case to consider the *booking limit* for the discount fare class to be the decision variable instead of the protection level used previously. This limit is denoted ℓ, and the corresponding full fare protection level is $(C - \ell)$.

As with the multiple fare class problem, the present problem can be handled either with the assumption of discrete decision variables or continuous ones. With only two fare classes and one decision variable to consider, neither approach is particularly favoured over the other. Since the application considered here is in fact discrete, the analysis in this chapter assumes discrete quantities. The gain function defined below takes the place of the left derivative used previously.

In the general model, the revenue resulting from the policy of booking $(X \wedge \ell)$ discount passengers is $f_X(X \wedge \ell)$. Define $F(\ell)$ to be the seats remaining after discount bookings are closed, so that $F(\ell) = C - (\ell \wedge X)$. It is assumed that there is now an additional demand for a total of $Y(\ell)$ full fare seats. Note that the distribution of the full fare demand might depend on the decision variable ℓ, as is the case when a proportion of customers denied discount bookings elect to *upgrade* to full fare bookings. Moreover, it is not assumed that X and $Y(\ell)$ are independent.

By satisfying as much of the demand $Y(\ell)$ as possible, an additional revenue of $f_Y(Y \wedge F(\ell))$ is generated. In the case that a goodwill cost or penalty is incurred for turning away full fare demands, the unsatisfied portion of this demand will incur a total cost of $f_G(Y - F(\ell))^+$. Combining the above revenues and costs gives the net revenue function

$$R(\ell) \quad = \quad f_X(X \wedge \ell) + f_Y(Y(\ell) \wedge F(\ell)) - f_G[Y(\ell) - F(\ell)]^+, \tag{3.1}$$

whose expectation is to be maximized as a function of ℓ. Since it is not possible to allocate more than the available capacity, $R(\ell)$ is only defined for ℓ such that

$$0 \leq \ell \leq C. \tag{3.2}$$

Now suppose $\ell - 1$ requests have been satisfied from the discount demand, and an additional discount request is received; (i.e. $X \geq \ell$). If bookings stop at $\ell - 1$, then the expected revenue is $E[R(\ell - 1) \,|\, X \geq \ell)]$.[1] If the additional request is satisfied, expected revenue is $E[R(\ell) \,|\, X \geq \ell)]$. It is useful to write the expected incremental gain, $G(\ell)$, of satisfying an additional request.

$$G(\ell) = E\left[R(\ell) \,|\, X \geq \ell\right] - E\left[R(\ell - 1) \,|\, X \geq \ell\right]$$

---

[1] The optimality rule being sought can depend on the amount of discount fare demand that has been observed up to the stopping time. Thus the expected revenue function must be maximized conditional on the event $\{X \geq \ell\}$.

It follows from (3.1) that

$$\begin{aligned} G(\ell) &= f_X \\ &+f_Y E\left[Y(\ell) \wedge (C-\ell) - Y(\ell-1) \wedge (C-\ell+1) \,|\, X \geq \ell\right] \\ &+f_G E\left[(Y(\ell)-(C-\ell))^+ - (Y(\ell-1)-(C-\ell+1))^+ \,|\, X \geq \ell\right], \end{aligned} \tag{3.3}$$

provided $\Pr[X \geq \ell] > 0$ so that the conditional expectations are defined. If X is less than ℓ then the decision as to whether the ℓ-th request should be satisfied can never arise and $G(\ell)$ is not defined. The domain of G is also limited to that of $R(\ell)$, as specified by condition (3.2).

The gain function $G(\ell)$ is just the first difference of the expected revenue function conditional on the information that $X \geq \ell$. Clearly, a booking limit of ℓ is preferred to $\ell - 1$ whenever $G(\ell)$ is positive. Furthermore, if $G(\ell)$ is nonnegative for all ℓ up to some ℓ^*, and nonpositive thereafter, then ℓ^* is optimal.

The following sections of the chapter will consider applications of the above model to specific allocation problems which are monotone in the sense defined in Chapter 2.

3.2 Specific Seat Allocation Problems

This section specializes the above general model to three variants of the seat allocation problem with dependent demands. In the first variant, it is assumed that there are no penalties for refused bookings and that full fare demand is not influenced by the discount booking level ℓ. The second considers the loss of goodwill associated with full fare passenger spillage by introducing a penalty for refused bookings. The third deals with the upgrades case in which ultimate full fare demand is influenced by the discount booking level.

3.2.1 A Simple Seat Allocation Model with Dependent Demands

The model analyzed here is the usual seat allocation model [14, 18, 95, 108, 136] except that the demands of the two fare classes, Y and X are allowed to be stochastically dependent.

It is assumed that the full fare demand is not influenced by the booking limit assigned to discount fares, so that $Y(\ell) = Y$, for $\ell = 1, \ldots, C$. Note that since demand is integer, $Y > C - \ell$ is the same as $Y \geq C - \ell + 1$.

Using these properties, the gain associated with increasing the booking limit from $\ell - 1$ to ℓ, given by (3.3), can be simpified to

$$\begin{aligned} G(\ell) &= f_X + f_Y E[(Y \wedge (C - \ell)) - \\ &\qquad (Y \wedge (C - \ell + 1)) \,|\, Y > C - \ell, X \geq \ell] \Pr[Y > C - \ell \,|\, X \geq \ell] \\ &= f_X - f_Y \Pr[Y > C - \ell \,|\, X \geq \ell]. \end{aligned} \tag{3.4}$$

This expression has a simple intuitive interpretation: when an additional seat is sold to a discount customer, there is a certain gain of one discount fare, and if the full fare demand exceeds the new lower protection level, there is a loss of one full fare.

The expected gain is positive whenever $G(\ell) > 0$, or equivalently whenever

$$\Pr[Y > C - \ell \,|\, X \geq \ell] < \frac{f_X}{f_Y}. \tag{3.5}$$

If it is the case that

$$\Pr[Y > C - \ell \,|\, X \geq \ell] \text{ is nondecreasing in } \ell, \tag{3.6}$$

then $G(\ell)$ is nonincreasing in ℓ, and the problem is monotone.

Henceforth, property (3.6) is referred to as the *monotonic association* property. Loosely speaking, this property specifies that as the discount booking limit increases, the full fare spill rate tends to increase. (Recall that this is always true when demands are independent.) It is not easy to imagine realistic demand distributions for which this property would not hold if X and Y demands are positively associated. (For example, as is shown later, the property holds if discount and full fare demands follow a bivariate normal distribution with nonnegative correlation.) However, if for some reason the demands are negatively associated, the property might well fail to hold.

A suitable ℓ to satisfy the definition of an optimal solution in a monotone problem is

$$\begin{aligned} \ell^* &= \max\{\ell : G(\ell) > 0\} \\ &= \max\{0 \le \ell \le C : \Pr[Y > C - \ell \,|\, X \ge \ell] < \frac{f_X}{f_Y}\}, \end{aligned} \tag{3.7}$$

where we will adopt the convention that $\ell^* = 0$ if $\Pr[Y > C] \ge \frac{f_X}{f_Y}$ so that the maximum is over the empty set. (Recall that the domain of G consists of those ℓ between 0 and C such that $\Pr[X \ge \ell] > 0$.) It is thus optimal to sell at most ℓ^* seats to customers requesting discount fares.

The probability $\Pr[Y > C - \ell \,|\, X \ge \ell]$ can be interpreted as the *maximal flight spill rate* as was the corresponding term in Littlewood's rule (2.1). But then (3.7) is just a generalization of the fact that this rate should be just less than the discount/full fare ratio. The optimality rule can also be expressed as follows: The discount booking limit should be set sufficiently high that, when discount seats sell out, full fare seats also sell out roughly $(f_X/f_Y) \times 100\%$ of the time.

If the demands are independent, then (3.6) clearly holds, and the optimality condition becomes

$$\ell^* = \max\{0 \le \ell \le C : \Pr[Y > C - \ell] < \frac{f_X}{f_Y}\}. \tag{3.8}$$

Figure 3.1 illustrates a possible expected revenue function.

Note that in this case there is not a unique ℓ^* which is optimal. The ℓ^* defined by (3.7) is the smallest. The largest optimal discount booking limit is obtained by permitting

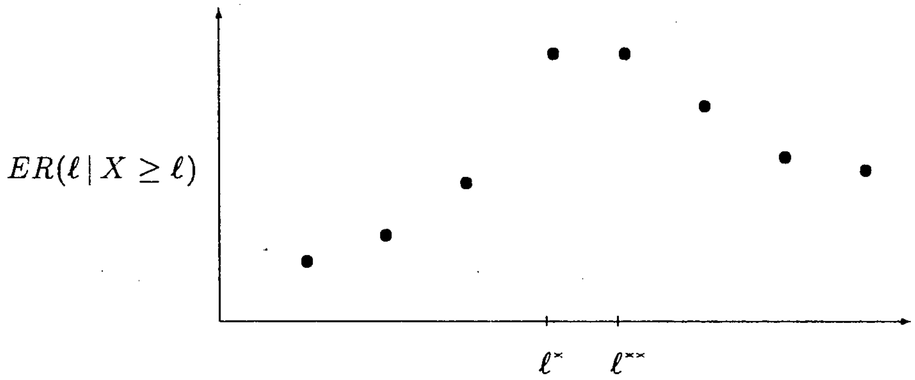


Figure 3.1: Expected Revenue Function

equality in (3.7):

$$\ell^{**} = \max\{0 \leq \ell \leq C : \Pr[Y > C - \ell \mid X \geq \ell] \leq \frac{f_X}{f_Y}\}. \tag{3.9}$$

This is just Littlewood's rule (2.1) except that now dependency between discount and full fare demands is allowed, subject to the monotonic association property (3.6). The following section illustrates the effect of such dependency.

3.2.2 Example: Seat Allocation with Dependent Demands

Table 3.1 on page 47 presents an example of optimal discount seat booking limits for a range of cabin capacities and for both independent and dependent demands. For this example, the discount fare was fixed at 60% of the full fare, and discrete approximations to bivariate normal distributions were used to model the discount/full joint probability functions. The mean combined demand was 100 seats in all calculations. In the dependent demand cases, correlations of $\varrho = 0.5$ and $\varrho = 0.9$ between discount and full fare demands were used.

Table 3.1: Effect of Demand Dependency on Discount Seat Booking Limits

	cabin capacity					
	46	60	80	100	120	140
discount booking limit: $\ell(\varrho = 0)$[a]	19	33	53	73	93	113
full fare protection: $C - \ell$	27	27	27	27	27	27
discount booking limit ($\varrho = 0.5$)	19	32	51	68	86	103
full fare protection	27	28	29	32	34	37
% revenue increase[b]	0%	0.04%	0.15%	0.30%	0.32%	0.18%
discount booking limit ($\varrho = 0.9$)	19	32	49	65	81	97
full fare protection	27	28	31	35	39	43
% revenue increase	0%	0.08%	0.54%	1.25%	1.27%	0.71%

---

[a] independent: correlation=0. For all calculations, mean demands were 70 discount and 30 full, and standard deviations were nominally 26.5 discount and 11.5 full. The standard deviations varied slightly between cases because of the discretization procedure.

[b] revenue increase achieved by allowing for dependency

With reference to Table 3.1, note that in the independent demand case, the discount seat booking limits correspond to a fixed *protection* level of 27 seats for full fare passengers at all cabin capacities. In this case (equation (3.8)), the protection level $(C - \ell)$ is determined solely by the discount/full fare ratio, which is held constant in this example. Viewed another way, the discount booking limits are increased as capacity increases in order to keep the maximum flight spill rate for full fares in balance with the discount/full fare ratio, as discussed earlier. Since the mean demands are being held constant for all cabin capacities, it appears that increased capacity is being allocated exclusively to discount demands. Recall, however, that unsold discount seats can be sold to full fare passengers. In the capacity=140 case, for example, the majority of flights will have discount demands of less than 113 seats, and full fare seating capacity will be accordingly larger than 27 seats most of the time. It is only when discount demands reach 113 seats that the marginal revenue considerations expressed by equation (3.8) dictate closing down discount sales.

In these examples, the optimal full fare protection level increases with capacity when discount and full fare demands are dependent (the $\varrho = 0.5$ and $\varrho = 0.9$ cases). The same spill rate balancing considerations are acting here; however, because of the positive correlation between demands, the discount booking limits are not increased as much as in the independent case. (The information that discount demand has exceeded some value should imply an increased probability of higher full fare demand and should lead to higher protection levels for full fare seats.) It is shown later that with bivariate normal demand distributions the optimal booking limit never increases as correlation increases.

With small cabin capacities relative to demand (capacities of 46 seats or fewer), the booking limits in the dependent cases are the same as the those in the independent case, as there is no revenue benefit from taking dependency into account. This is because with small capacities the discount demand is almost certain to exceed the discount seat booking limit, and so $\Pr[Y > C-\ell \mid X \geq \ell] \cong \Pr[Y > C-\ell]$. For large capacities relative to demand (e.g. above 140 seats in the above example), the optimal discount booking limit in the independent case will be substantially lower than that in the dependent case; however, the corresponding revenue benefits are small as there is ample space for both fare classes under most realizations of the demand process.

3.2.3 Implementation of Dependent Demand Booking

The optimal booking rule for the dependent demand case (3.7) is simple to implement as a planning tool if some joint distribution such as the bivariate normal is assumed to hold for the demands. In this case it is straightforward to calculate the conditional distribution $\Pr[Y > C-\ell \mid X \geq \ell]$ for enough values of ℓ to solve the optimality condition. It is then possible to study the impact of hypothesized shifts in the demand distribution or in other parameters in much the same way as in the example above.

Implementation of the dependent demand booking rule as a control tool in a reservations system is also possible, but less straightforward. Estimation of the conditional demand distributions can be done, as above, by using a joint demand distribution. In this case, however the parameters of the distribution must be obtained by fitting to historical data and, possibly, by adusting for anticipated market conditions. This fitting process is not straightforward since 1) demand data from a history of previous flights will be censored whenever demand reaches a booking limit or the capacity of the aircraft, and 2) the parameters of the demand distribution depend on external factors like fares, competition and time of year. The same problems are present in the independent demand case, but the estimation is simpler since correlation between demand classes need not be considered. This estimation problem is dealt with in chapter 5.

As mentioned in the previous chapter, the spill rate interpretation of the optimal allocation rules suggests a second implementation scheme. The optimal allocation rule in either the independent or dependent demand cases specifies that the *observed maximal spill rate* should be as close as possible to, without exceeding, the discount/full fare ratio. If the observed proportion is too high, the booking limit should be adjusted downward; if it is too low, it should be adjusted upward. This approach has two significant advantages. First, there is no requirement for modeling the demand distribution; and second, there is little computational difference between the independent and dependent demand cases. To see the second point note that the observed maximal spill rate in the independent case is the proportion of flights on which the full fare demand exceeded the protection level $(C - \ell)$. In the dependent case, it is the proportion of those flights on which the discount booking limit was reached for which the full fare protection level was also exceeded. This type of adaptive control strategy has the disadvantage that it is based entirely on aggregate historical data, not on forecasts of future demand for individual flights; however, it provides an easily implemented way of monitoring past performance relative to theoretically optimal booking limits.

3.2.4 Full Fare Passenger Goodwill and Spill Rates

Airlines are justifiably concerned about the impact of discount seat allocation policies on the proportion of full fare reservations requests that must be turned away. This proportion, expressed as a percentage, is often referred to as the *passenger spill rate*, or simply *spill rate*. A related concept is the proportion of flights on which one or more reservations requests are turned away, or the *flight spill rate* discussed earlier. The following section examines these two spill rates in more detail and gives a simple relationship connecting the two rates when discount demand can be assumed to be high; i.e., $\Pr[X > \ell] \approx 1$.

Flight and Passenger Spill Rates

Let Z denote the number of full fare booking requests spilled given full fare and discount fare demands Y and X respectively, discount fare booking limit ℓ, and capacity C. That is,

$$Z = \begin{cases} [Y-(C-\ell)]^+ & \text{if } X > \ell \\ [Y-(C-X)]^+ & \text{if } X \leq \ell. \end{cases} \tag{3.10}$$

The *flight spill rate* has been defined above as the proportion of flights on which at least one full fare passenger is refused a booking. The expected value of this proportion is simply the probability that full fare demand exceeds the number of seats remaining after discount sales have stopped. Let r_F denote the flight spill rate. Then

$$r_F = \Pr[Y > C-\ell \,|\, X > \ell]\Pr[X > \ell] + \Pr[X \leq \ell \cap X + Y > C]. \tag{3.11}$$

The *passenger* spill rate can be viewed as the long-run proportion of full fare requests that are turned away; that is, the total number of refused requests in a long series of flights divided by the total number of requests. Let r_P denote the passenger spill rate. With the demand distribution and parameters as defined above, this quantity can be written

$$r_P = \frac{\mathrm{E}[Z]}{\mathrm{E}[Y]}. \tag{3.12}$$

Let $F(\cdot)$ be the cumulative distribution function for X-demand and define the protection level $p = (C-\ell)$. Then

$$\begin{aligned} \mathrm{E}[Z] &= \mathrm{E}\{[Y-p]^+ \,|\, X > \ell\}\Pr[X > \ell] + \int_0^\ell \mathrm{E}\{[Y-(C-x)]^+ \,|\, X = x\}\, dF(x) \\ &= \mathrm{E}[Y \,|\, Y > p \cap X > \ell]\Pr[Y > p \cap X > \ell] - p\Pr[Y > p \cap X > \ell] \qquad (3.13) \\ &\quad + \int_0^\ell \mathrm{E}\{[Y-(C-x)]^+ \,|\, X = x\}\, dF(x). \end{aligned}$$

If (X,Y) are adequately represented as bivariate normal random variables, then both r_F and r_P can be calculated for any particular set of distribution parameters. The calculation of E[Z] can be accomplished numerically with the aid of expansions for the conditional expectations in equation (3.13) that are given in the appendix (equation (A.33)) and in chapter 5 (equation (5.24)).

For monotonically associated demands, the full fare spill rate is most severe when discount fare demand is high. In the extreme case that $\Pr[X > \ell] \approx 1$, the spill rates become (approximately)

$$r_F \approx \Pr[Y > p] \tag{3.14}$$

and

$$r_P \approx (1/\mathrm{E}[Y])\{(\mathrm{E}[Y \mid Y > p] - p)\Pr[Y > p]\}. \tag{3.15}$$

With a bivariate normal distribution for (X,Y), the calculation of the spill rates becomes essentially equivalent to that for a single fare class with cabin capacity p and normal demand distribution given by the marginal distribution for Y. Harmer [62, (1976)] derived a simple relationship between the flight and passenger spill rate for a single fare class with normally distributed demand. It is straightforward to obtain this result using some of the properties of the normal distribution given in the appendix. Denote the mean and standard deviation of the marginal Y distribution by μ_{Y} and σ_{Y} respectively. From appendix equation (A.8),

$$\mathrm{E}[Y \mid Y > p] = \mu_{\mathrm{Y}} + \sigma_{\mathrm{Y}}(\phi(z_p)/\Pr[Y > p]);$$

where $z_p = (p - \mu_{\mathrm{Y}})/\sigma_{\mathrm{Y}}$. Then, from (3.14) and (3.15),

$$\begin{aligned} r_P &= (1/\mu_{\mathrm{Y}})[\mu_{\mathrm{Y}} r_F + \sigma_{\mathrm{Y}}\phi(z_p) - p r_F] \\ &= (\sigma_{\mathrm{Y}}/\mu_{\mathrm{Y}})(\phi(z_p) - z_p r_F). \end{aligned} \tag{3.16}$$

An Example

In the following example, it is assumed for simplicity that discount fare demand is sufficiently high that the discount booking limit is always reached. If an optimal seat allocation rule is used (in either the independent or dependent demand case), the flight spill rate is close to the discount/full fare ratio. For example, consider the independent demand case with a plane capacity of 100 seats in Table 3.1. If mean low fare demand is significantly higher than 70 seats so that the discount booking limit of 73 seats is reached most of the time, and full fare mean demand remains at 30 seats, then the flight spill rate is approximately 60%, since $f_X/f_Y = 0.60$. In this example, the full fare passengers are essentially being booked into a fixed allocation of 27 seats. From equation (3.16), the corresponding passenger spill rate is 21%.

It is difficult to obtain reliable data on actual airline passenger spill rates, but it is hard to imagine that airline managers would tolerate turning away 21% of their best customers, even given the high demand for discount fares assumed in the example.[2]

Goodwill Premiums

There thus appears to be a substantial discrepancy between spill rates corresponding to optimal booking limits and the spill rates that would be tolerated by airlines. Possible explanations for this discrepancy include:

1. Optimal allocation rules may simply not be used by many airlines.

2. The airlines may be compensating for demand dependencies, either deliberately or on a trial-and-error basis, by lowering discount booking limits below those specified by the simple allotment rule.

---

[2] In the Boeing report cited above, all sample calculations were presented with *flight* spill rates of 5% or lower.

3. The discount and full fare demands may overlap in time to a sufficient degree, that the observed full fare demand can be used to adjust the discount booking limit.

4. Voluntary 'bumping' of discount passengers may be used to permit high overbooking levels for full fare passengers, thus reducing the effective full fare spill rate.

5. The discount booking limits may be adjusted downward in an ad hoc fashion to compensate for the perceived extra value of full fare passengers above and beyond their higher fares. (Full fare passengers are predominantly composed of business travellers who can be expected to travel more frequently than the discount, predominantly leisure, travellers. Low spill rates can be seen then as a way of promoting future earnings from these customers by maintaining passenger *goodwill.*)

The latter case, which recognizes the goodwill benefits associated with serving the full fare passenger, is now examined.

The effect of not being able to accommodate a full fare passenger can be viewed in two ways. First, a *premium* of f_G can be included in the full fare. Alternatively, the revenue derived from a full fare can be kept at f_Y, and a *loss* of f_G can be incurred for each full fare customer not accommodated. The argument used in section 3.2.1 can be applied to this version of the revenue model to derive the optimality condition

$$\begin{aligned} \ell^* &= \max\{\ell \geq 0 : G(\ell) > 0\} \\ &= \max\{0 \leq \ell \leq C : \Pr[Y > C - \ell \,|\, X \geq \ell] < \tfrac{f_X}{f_Y + f_G}\}. \end{aligned} \tag{3.17}$$

It is clear from equation (3.7) with f_Y replaced by $f_Y + f_G$, and from (3.17), that the optimal allocation is identical with either interpretation of goodwill. In either case, the incorporation of goodwill considerations will increase the full fare protection level and reduce the full fare spill rate. To illustrate one implication of the optimality condition, consider an airline that wishes to limit its passenger spill rate to 3%. From formula

(3.16), using the same assumptions as the example given above, a passenger spill rate of 3% corresponds to a flight spill rate of 15%, and this in turn corresponds to a goodwill premium of $f_G \approx 3f_Y$ (that is, the solution to: $0.15 = 0.6/(1+x)$). Thus a goodwill premium of three times the full fare would be required to justify restricting the passenger spill rate to 3%. It is not clear whether such a high premium is justified. Such a justification would depend upon an airline's assessment of the proportion of their full fare customers who might be lost permanently to competitors after failing to obtain a booking. [3] Perhaps one of the chief uses of equation (3.17) would be, as in this example, to impute the goodwill premium implied by a particular spill rate policy.

3.2.5 Upgrades

We now examine the case in which the dependency between discount and full fare demands arises because of a tendency for some discount fare customers to upgrade to full fares if denied a discount reservation. In this context, it is assumed that the upgrading tendency is the *only* source of dependency and that the initial X and Y demands (i.e. before upgrading) are independent. Under these circumstances, the *ultimate* Y demand will depend both on the X demand and on the booking limit set for the X demand. It is this dependency on the booking limit that necessitates an analysis separate from and more involved than that for the dependent demand case discussed in section 3.2.1. Note that the optimality condition derived here was previously proposed without formal proof by Belobaba [13, page 130, equation 5.53] and that a similar result has been obtained independently by Pfiefer [106, (1989)] using different methods. The purpose here is to provide a formal proof of the result within the context of a general model for the seat

---

[3] Goodwill premiums can be justified, in part, by the pattern of airline demands. For example, Tretheway [141] finds that frequent flyers represent only 3% of the travelling public but account for over 40% of airline revenues. One anonymous referee for a paper based on the present chapter suggests on the basis of experience with carriers that goodwill premiums in the range 10% to 20% are used.

allocation problem.

To model the upgrading, define

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th customer would upgrade if denied a discount fare,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

Assume that $\{D_1, D_2, \ldots\}$ are independent and identically distributed with $ED_i = \gamma$ being the probability that a customer denied a discount fare will upgrade. Also assume independence of the process $\{D_1, D_2, \ldots\}$ of upgrades, the demand X for discount fares, and the demand Y for full fares *exclusive of the upgrades.* Let $U(\ell)$ denote the total number of upgrades when the discount booking limit is ℓ; that is,

$$U(\ell) = \sum_{i=\ell+1}^{i=B} D_i. \tag{3.19}$$

This quantity is, of course, zero if $X \leq \ell$. Identification of this model with the general revenue model (3.1) is the same as in section 3.2.1 except that now

$$Y(\ell) = Y + U(\ell) \tag{3.20}$$

is the sum of the full fare demand and any upgrades.

To motivate the optimality condition, marginal analysis can be used as in Belobaba [13, page 130, equation 5.53]. If a discount fare customer is booked, then the revenue is f_X. If a discount fare customer cannot be booked, then with probability γ there is an upgrade generating revenue f_Y, and with probability $1-\gamma$ there is no upgrade. In the latter case, the booking decision will have no impact on revenue if $X \leq \ell$. However, if $X > \ell$, then additional revenue f_Y is obtained if the seat being considered is used either by some other upgrade or by a full fare customer. This analysis leads one to conjecture that it is optimal to book a discount fare customer if

$$f_X > \gamma f_Y + (1-\gamma)\Pr[(Y + U(\ell)) > C - \ell \,|\, X \geq \ell]. \tag{3.21}$$

To verify this optimality condition, compute $G(\ell)$ from (3.3). Let $H(\ell) = [Y(\ell) \wedge (C-\ell)] - [Y(\ell-1) \wedge (C-\ell+1)]$. To evaluate H consider two cases. First suppose that $Y(\ell) > C-\ell$. Then $Y(\ell-1) \geq C-\ell+1$ and $H(\ell) = -1$. Second, suppose that $Y(\ell) \leq C-\ell$. Then $Y(\ell-1) \leq C-\ell+1$ and $H(\ell) = Y(\ell) - Y(\ell-1) = -D_\ell$. Thus the equation for the gain (3.3) reduces to

$$\begin{aligned} G(\ell) &= f_X - f_Y \Pr[Y(\ell) > C-\ell \,|\, X \geq \ell] - \\ &\quad f_Y \Pr[Y(\ell) \leq C-\ell \,|\, X \geq \ell] E[D_\ell] \qquad (3.22) \\ &= f_X - (1-\gamma) f_Y \Pr[Y(\ell) > C-\ell \,|\, X \geq \ell] - \gamma f_Y, \end{aligned}$$

where the assumption that D_ℓ is independent of X and of Y is used to obtain the first equation.

It remains to be shown that the problem is monotone by establishing that $G(\ell)$ is nonincreasing in ℓ. Using the fact that $D_i \leq 1$ gives

$$\begin{aligned}
&\Pr[Y(\ell-1) > C-\ell+1 \,|\, X \geq \ell-1] \\
&\qquad = \Pr[(Y + \textstyle\sum_{i=\ell}^{i=B} D_i) > C-\ell+1 \,|\, X \geq \ell-1] \\
&\qquad \leq \Pr[(Y + \textstyle\sum_{i=\ell+1}^{i=B} D_i) > C-\ell \,|\, X \geq \ell-1] \\
&\qquad = \Pr[Y(\ell) > C-\ell \,|\, X \geq \ell-1].
\end{aligned} \tag{3.23}$$

By conditioning on whether $X = \ell - 1$ or $X \geq \ell$, and manipulating the conditional probabilities, $\Pr[Y(\ell) > C - \ell \,|\, X \geq \ell - 1]$ can be rewritten as

$$\begin{aligned}
&\Pr[Y(\ell) > C-\ell \,|\, X \geq \ell] \\
&\quad + \Pr[X = \ell-1 \,|\, X \geq \ell-1]\Big\{ \Pr[Y(\ell) > C-\ell \,|\, X = \ell-1] \\
&\qquad\qquad - \Pr[Y(\ell) > C-\ell \,|\, X \geq \ell] \Big\}.
\end{aligned} \tag{3.24}$$

The difference in the last term cannot be positive since

$$\begin{aligned}
\Pr[Y(\ell) > C-\ell \,|\, X \geq \ell] &\geq \Pr[Y > C-\ell \,|\, X \geq \ell] \\
&= \Pr[Y(\ell) > C-\ell \,|\, X = \ell-1],
\end{aligned} \tag{3.25}$$

where the assumption that Y and X are independent and the observation that $U(\ell) = 0$ if $X = \ell - 1$ are used to obtain the last equation. Replacing the difference in (3.25) by 0, and using the inequality (3.23), shows that

$$\Pr[Y(\ell-1) > C-\ell+1 \,|\, X > \ell-1] \leq \Pr[Y(\ell) > C-\ell \,|\, X > \ell], \tag{3.26}$$

and so $G(\ell)$ is nonincreasing. Then, from (3.19) and (3.22), $G(\ell)$ is positive as long as

$$\Pr[(Y + U(\ell)) > C-\ell \,|\, X \geq \ell] < \frac{f_X - \gamma f_Y}{(1-\gamma) f_Y}, \tag{3.27}$$

which is equivalent to (3.21). Define ℓ^* to be the largest ℓ ($0 \leq \ell \leq C$) satisfying (3.27). As with optimality condition (3.7), set $\ell^* = 0$ if no ℓ can satisfy (3.27). This is the case, for example, when γ is sufficiently large that the right hand side of (3.7) is nonpositive. This ℓ^* satisfies the condition in the definition of a monotone problem and $|\, Y(\ell) - Y(\ell - 1)\, | \leq 1$. Hence the problem is monotone and it is optimal to book discount fares up to ℓ^*.

Implementation of Upgrades Formula

The comments made earlier regarding implementation of the dependent demand solution apply again here. In the present case estimation of the joint distribution of $Y + U(\ell)$ and X is somewhat easier since Y and X can be estimated independently and then Y adjusted by the binomial distribution $U(\ell)$ for each ℓ. Alternatively, the spill rate control approach could be applied with no change except for adjustment of the discount/full fare ratio as indicated in (3.27).

A numerical example of the use of the upgrades formula is provided in Belobaba [13, pp138–139].

3.2.6 Overbooking

A final application of the general dependent demand model is in calculation of optimal overbooking levels in a single fare class. Discussion of that application will be deferred until the next chapter.

3.3 Monotonic Association Between Bivariate Normal Random Variables

This section establishes properties claimed above that hold when demands follow a bivariate normal distribution. Specifically, it is shown that if the demands X and Y have a joint bivariate normal distribution with correlation ρ, then: 1) the monotonic association condition will be satisfied if $\rho \geq 0$, and 2) the discount booking limit given by equation (3.7) decreases as ρ increases, for $-1 \leq \rho \leq 1$. In order to establish these properties we must first present some basic results on bivariate association found in Lehmann [84, (1955)], Lehmann [85, (1966)], Slepian [129, (1962)], Esary and Proschan [49, (1972)], and Tong [138, (1980)].

3.3.1 Some Basic Results in Bivariate Association

We start with some definitions and abbreviations for properties of association between random variables X and Y(not necessarily normally distributed):

positive regression dependency (PRD) Y is positive regression dependent on X if

$$\Pr[Y > y \mid X = x] \text{ is nondecreasing in } x \text{ for every } y. \tag{3.28}$$

right tail increasing (RTI) Y is right tail increasing with respect to X if

$$\Pr[Y > y \mid X > x] \text{ is nondecreasing in } x \text{ for every } y. \tag{3.29}$$

Connections among these properties as well as some other properties not mentioned here are delineated in a theorem in Tong[138, Theorem 5.1.1]. Theorem 3.3.1 below presents the section of that result relevant to the monotonic association condition.

Theorem 3.3.1 *For any random variables X and Y for which the covariance $Cov(X,Y)$ exists, the following sequence of implications holds:*

$$PRD \Rightarrow RTI \Rightarrow [Cov(X,Y) \geq 0] \tag{3.30}$$

The following corollary is relevant to the monotonic association condition.

Corollary 3.3.2 *If (X,Y) are bivariate normal, Y is RTI with respect to X if and only if $[Cov(X,Y) \geq 0]$.*

Proof: From Theorem 3.3.1, we need only point out that for normally distributed random variables $[\mathrm{Cov}(X,Y) \geq 0] \Rightarrow$ PRD, which is a standard result in regression analysis. ■

A second useful result concerns the behaviour of joint tail probabilities as correlation increases.

Theorem 3.3.3 (Slepian, 1962) *Let* $\mathbf{Z}_R$ *and* $\mathbf{Z}_T$ *be multivariate normal random vectors of the same dimension with correlation matrices* $\mathbf{R}$ *and* $\mathbf{T}$, *respectively, and with zero mean vectors. Let* $\mathbf{a}$ *be any constant vector of the same dimension. Then*

$$[\mathbf{R} \geq \mathbf{T}] \Rightarrow \Pr[\mathbf{Z}_R > \mathbf{a}] \geq \Pr[\mathbf{Z}_T > \mathbf{a}]. \tag{3.31}$$

Furthermore,if $\mathbf{R}$ *and* $\mathbf{T}$ *are positive definite, the inequality is strict.*

The following corollary is a direct consequence of the theorem for bivariate normal (X,Y).

Corollary 3.3.4 *If* (X,Y) *are bivariate normal with correlation* $-1 < \rho < 1$, *and* k *and* x *are constants, then* $\Pr[Y > k - x \cap X > x]$ *increases as* ρ *increases.*

3.3.2 Application to Seat Allocation

We are now in a position to establish the properties mentioned above relating to monotonic association and the optimal allocation ℓ^*.

Theorem 3.3.5 *If* X *and* Y *are bivariate normal with nonnegative correlation* ρ *then* X *and* Y *are monotonically associated.*

Proof: From corollary 3.3.2, nonnegative correlation implies that $\Pr[Y > y \,|\, X > \ell]$ is nondecreasing in ℓ for every y. Choose any ℓ and $\hat{\ell} > \ell$. Then

$$\Pr[Y > C - \ell \,|\, X \geq \ell] \leq \Pr[Y > C - \ell \,|\, X \geq \hat{\ell}] \leq \Pr[Y > C - \hat{\ell} \,|\, X \geq \hat{\ell}];$$

that is, $\Pr[Y > C - \ell \,|\, X \geq \ell]$ is nondecreasing in ℓ. ■

Theorem 3.3.6 *If discount and full fare demands,* X *and* Y *are drawn from the bivariate normal family of distributions, then the optimal discount booking limit* ℓ^* *decreases as the correlation of the demands increases.*

Proof: For ease of reference we repeat the optimality condition for ℓ^* given in equation (3.7):

$$\ell^* = \max\{0 \le \ell \le C : \Pr[Y > C - \ell \,|\, X \ge \ell] < \frac{f_X}{f_Y}\}. \tag{3.32}$$

The theorem is proved if it can be established that $\Pr[Y > C - \ell \,|\, X \ge \ell]$ is increasing in ρ, since then the maximum in condition (3.7) is attained at decreasing values of ℓ as ρ increases. But this follows immediately from Corollary 3.3.4 since

$$\Pr[Y > C - \ell \,|\, X \ge \ell] = \Pr[Y > C - \ell \cap X \ge \ell] \Pr[X \ge \ell].$$

■

3.4 Summary — Dependent Demands Case

This chapter has presented a simple resource allocation model and applied it to seat allocation problems. For ease of reference, the main results are summarized below:

1. When discount and full fare demands are bivariate normal with arbitrary correlation, the optimal discount booking limit will decrease as the correlation increases. In particular, if the correlation is positive, the optimal booking limit is less than that specified by Littlewood's rule (independent demand).

2. With monotonically associated discount and full fare demands X and Y, respectively, cabin capacity C, discount fare f_X, full fare f_Y, and full fare goodwill premium f_G; it is optimal to limit discount fare bookings to ℓ^* seats, where:

$$\ell^* = \max\{0 \le \ell \le C : \Pr[Y > C - \ell \,|\, X \ge \ell] < \frac{f_X}{f_Y + f_G}\}. \tag{3.33}$$

   In particular, this optimality condition will hold when demands are positively correlated bivariate normal random variables. Again, this will result in a lower discount seat booking limit.

3. When the discount and *initial* full fare demands are independent but the presence of upgrades creates a dependency between discount and *ultimate* full fare demand, results (3.17) and (3.27) can be combined to obtain the following optimal discount seat allocation:

$$\begin{aligned} \ell^* &= \max\{0 \leq \ell \leq C : \\ &\quad \Pr[Y + U(\ell) > C - \ell \mid X \geq \ell] < \tfrac{f_X - \gamma(f_Y + f_G)}{(1-\gamma)(f_Y + f_G)}\}, \end{aligned} \tag{3.34}$$

   where γ is the upgrade probability, and $U(\ell)$ is the total number of upgrades given discount allocation ℓ. Once again, this implies lower discount seat booking limits.

4. If demands are monotonically associated then, regardless of the actual demand distributions, booking limits should be controlled so that, in a long series of flights, $(f_X/f_Y) \times 100\%$ of the times that discount seats sell out, the full fare seats should also sell out.

It has been shown that these conditions are optimal among all policies that use only the information $X > \ell$. Given stable fares, the only possible justification for changing an optimal booking limit is a perceived shift in the joint demand distribution for discount and full fares. Thus, for example, the occurrence of a sudden 'flurry' of discount demand at some point in the booking process cannot in itself justify a change in the booking limit unless it can be validly associated with a change in the joint demand distribution. If it is decided that such a change has occurred, a reasonable response is to simply recalculate the optimal booking limit on the basis of the new joint demand distribution and the seat capacity remaining for the flight. More sophisticated dynamic modelling is required to optimally account for the possibility of periodic revision of the joint demand distribution on the basis of more information than $X > \ell$.

The three variants of optimal booking conditions given above all suggest lower discount booking limits than those implied by Littlewood's rule for independent demands.

This is important since these results are more easily reconciled with reasonable full fare passenger spill rates. Numerical examples suggest that the revenue gains from application of these conditions may be modest (e.g. 0.32% in the 0.5 correlation case in Table 3.1). However, given the largely fixed cost, low margin nature of airline operations in competitive markets, such revenue gains represent almost pure profit and thus are greatly magnified in terms of profit impact.

Chapter 4

A Simple Overbooking Model

This chapter examines a basic version of the problem that has received the most attention from airline researchers over the years — that of determining suitable overbooking levels in a one or more fare classes. This analysis was motivated by the recognition that the seat allocation problem for two dependent fare classes and the overbooking problem for a single fare class had very similar structures. Section 4.1, below, exploits this similarity in developing a simple model for the overbooking problem. In section 4.2, specification of the passenger confirmation process as a Bernoulli process leads to a simple condition for an optimal overbooking level that is very similar to Littlewood's rule for the seat allocation problem. This condition is further simplified through the use of the normal approximation to the binomial distribution, and it is shown that under certain circumstances the simple ratio of cabin capacity to the confirmation probability gives a good estimate of the optimal overbooking level. Section 4.3 shows that the simple overbooking model can be easily modified to handle the case that bookings occur in groups. Section 4.4 shows the connection between the optimality condition derived here and that of Shlifer and Vardi [123, (1975)] and presents a numerical comparison of results from the various models. Implementation is discussed in section 4.5, and a summary and conclusions are provided in the final section.

4.1 The Dependent Demand Model and Overbooking in One Fare Class

This section applies the generic model for two dependent demand classes to the problem of determining an optimal overbooking level for a single fare class. One of the interesting features of this simple overbooking model is that it has virtually the same structure as the simple seat allocation model discussed in section 3.1. The initial demand for seats in the overbooking model can be identified with discount fare demand in the seat allocation model, and the number of customers who show up at flight time can be identified with full fare demand. The differences lie in the assessment of the average revenues from each of these 'fare classes' and in the calculation of the demand distributions.

The published work most relevant to the simple overbooking problem considered here is that of Beckmann [11, (1958)], and of Shlifer and Vardi [123, (1975)]. Both of these analyses start with a model of the aggregate cancellations for a flight. In contrast, the work described here starts at the level of individual passenger cancellations. Beckmann's analysis assumes that the booking level will not be much larger than the capacity of the airplane (an assumption that was reasonable given the tight restrictions on overbooking that prevailed at the time of his article).

For ease of reference, the net revenue model for the dependent demand seat allocation case, equation (3.1), is repeated here:

$$R(\ell) = f_X(X \wedge \ell) + f_Y(Y(\ell) \wedge F(\ell)) - f_G[Y(\ell) - F(\ell)]^+, \tag{4.1}$$

It will now be shown that with a suitable re-interpretation of the components, this same model is applicable to the one fare class overbooking problem.

In the simple overbooking model, there are C seats available for one fare class. Denote the demand for bookings in this fare class by X, and identify this demand with discount demand in the generic model.[1] Once a customer has booked a seat, he or she will either

---

[1] This identification relates only to the fact that the discount demand arrives first in the generic model

show up at flight time (henceforth, confirm) or fail to show up (henceforth, cancel). Cancellations which occur early enough in the booking process for the seats to be sold to different passengers will not be considered[2]. That is, the demand for bookings X occurs first, then the cancellation process occurs.

To model this process, associate with the i-th customer booked a random variable

$$B_i = \begin{cases} 1 & \text{if the customer confirms,} \\ 0 & \text{if the customer cancels.} \end{cases} \tag{4.2}$$

If ℓ seats are booked, then $N(\ell) = \sum_{i=1}^{\ell} B_i$ seats are confirmed.

Now identify the number of confirmations given bookings of ℓ with the second period demand in the generic model; that is, let $Y(\ell) \equiv N(\ell)$. The limits (3.2) become $0 \leq \ell$ with no upper bound on ℓ. Booking a customer is assumed to generate no revenue until that customer confirms so that $f_X = 0$. The revenue from a confirmed booking is f_Y, and the penalty for "bumping" a confirmed customer for whom no seat is available is $f_O \equiv f_G$.

The revenue as a function of the number of seats booked, ℓ, is thus

$$R(\ell) = f_Y[N(\ell) \wedge C] - f_O[N(\ell) - C]^+.$$

or

$$R(\ell) = f_Y N(\ell) - (f_Y + f_O)[N(\ell) - C]^+. \tag{4.3}$$

The expected revenue function conditional on the booking level, ℓ, being reached is

$$\mathrm{E}[R(\ell)|X \geq \ell] = f_Y \mathrm{E}[N(\ell)] - (f_Y + f_O)\mathrm{E}[(N(\ell) - C)^+]. \tag{4.4}$$

---

and not to the fare class. In fact, it is more meaningful to consider the fare class under consideration to be the full fare class since full fare passengers generally exhibit the highest cancellation rates.

[2] For a dynamic programming treatment which allows for early cancellations, see Rothstein [117, 1968].

As before, the optimal booking level can be determined from the gain or incremental revenue from an additional booking. If the booking level has been reached (i.e. $X \geq \ell$), then

$$\begin{aligned} R(\ell) - R(\ell-1) &= f_Y(N(\ell) - N(\ell-1)) - \\ & \quad (f_Y + f_O)\left[(N(\ell) - C)^+ - (N(\ell-1) - C)^+\right] \\ &= f_Y B_\ell - (f_Y + f_O) Z_\ell, \end{aligned} \tag{4.5}$$

where

$$Z_\ell = \begin{cases} B_\ell & \text{if } N(\ell-1) \geq C, \\ 0 & \text{if } N(\ell-1) < C. \end{cases}$$

Now the gain in expected revenue associated with an increase in booking limit from $\ell - 1$ to ℓ can be computed, as follows:

$$G(\ell) = f_Y \mathrm{E}[B_\ell] - (f_Y + f_O)\mathrm{E}[B_\ell | N(\ell-1) \geq C] \Pr[N(\ell-1) \geq C]. \tag{4.6}$$

To make further progress in the analysis and ensure that the problem is monotone, the cancellation process must be made more explicit. This is done in the following section.

4.2 A Bernoulli Cancellation Process

The simplest model of passenger cancellations is obtained by assuming that cancellations occur according to a Bernoulli process. That is, assume that $\{B_1, B_2, \ldots\}$ are independent and identically distributed and that $EB_i = \alpha$ is the probability that a customer confirms. If ℓ seats are booked, then $N(\ell) = \sum_{i=1}^{\ell} B_i$ seats are confirmed and the distribution of $N(\ell)$ is binomial.

The possibility of a variable confirmation probability will be considered at the end of this section. The independence of customer confirmations is a reasonable assumption as long as bookings do not occur in groups. The next section will extend the analysis to group bookings, which will allow for the distribution of the total number of confirmations to have more variance and fit a broader range of data.

To derive the optimality condition, evaluate the incremental revenue function (4.6), which simplifies to

$$G(\ell) = f_Y\alpha - (f_Y + f_O)\alpha \Pr[N(\ell - 1) \geq C]. \tag{4.7}$$

This gain is clearly decreasing in ℓ (strictly decreasing if $\alpha > 0$), so that the problem is monotone. Thus, additional customers should be booked as long as $G(\ell)$ is positive. That is, define

$$\begin{aligned} \ell^*(C) &= \max\{\ell \geq 0 : G(\ell) > 0\} \\ &= \max\{\ell : \Pr[N(\ell - 1) \geq C] < \tfrac{f_Y}{f_Y + f_O}\}. \end{aligned} \tag{4.8}$$

and book up to $\ell^*(C)$. For a random X-class demand with C seats available, it is optimal to book $X \wedge \ell^*(C)$ seats.

The solution to (4.8) is particularly easy if: 1) there is a 100 percent overbooking penalty so that $f_Y = f_O$, and 2) the confirmation rate is 50 percent so that $\alpha = 0.5$. In this case, $\ell^*(C) = C/\alpha - 1$, or $\ell^*(C) \cong C/\alpha$. This might seem to be only of incidental interest since the confirmation rate is so extreme; however, the following analysis shows that C/α is a robust approximation to the optimal policy over a wide range of confirmation probabilities as long as the overbooking penalty and fare are approximately equal.

To examine the error in this approximation, note that $N(C)$ is approximately normal with mean $C\alpha$ and variance $C\alpha(1-\alpha)$ for values of C and α that are reasonable in many situations[3]. The normal approximation to the tail of a binomial is quite good as long

---

[3] The standard conditions for normal approximations of the binomial distribution are: $C \geq 50$, $C\alpha \geq 5$, and $C(1-\alpha) \geq 5$.

as the tail being evaluated is not too extreme. Since in our case, we are evaluating the $f_Y/(f_Y + f_O)$-th percentile, the approximation should be very accurate.

Note that $\Pr[N(\ell-1) \geq C] = \Pr[N(\ell-1) > C-1]$. For convenience, define $\ell' = \ell - 1$ and $C' = C - 1$. Let z be defined by

$$\Pr[Z > z] = \frac{f_Y}{f_Y + f_O}, \tag{4.9}$$

where Z is distributed as a standard normal random variable. Since $(N(\ell') - \ell'\alpha)/\sqrt{\ell'\alpha(1-\alpha)}$ is approximately standard normal, (4.8) can be approximated by

$$\ell^*(C) \;\cong\; \max\{\ell \geq 0 : \Pr[Z > \frac{C' - \ell'\alpha}{\sqrt{\ell'\alpha(1-\alpha)}}] < \frac{f_Y}{f_Y + f_O}\}. \tag{4.10}$$

Using the definition of z at (4.9) and the fact that Z is a continuous random variable, $\ell^*(C)$ is approximately the solution to

$$z = \frac{C' - \ell'\alpha}{\sqrt{\ell'\alpha(1-\alpha)}}. \tag{4.11}$$

This quadratic equation can be easily solved for ℓ to obtain

$$\ell^* \cong \begin{cases} 1 + \frac{C'}{\alpha} + \xi - \sqrt{(\xi + \frac{C'}{\alpha})^2 - (\frac{C'}{\alpha})^2} & \text{if } \frac{f_Y}{f_Y + f_O} \leq \frac{1}{2}, \\ \\ 1 + \frac{C'}{\alpha} + \xi + \sqrt{(\xi + \frac{C'}{\alpha})^2 - (\frac{C'}{\alpha})^2} & \text{if } \frac{f_Y}{f_Y + f_O} > \frac{1}{2}, \end{cases} \tag{4.12}$$

where

$$\xi = \frac{z^2}{2}\frac{1-\alpha}{\alpha}. \tag{4.13}$$

The appropriate sign for the radical was determined by noting that $\ell^* > C'/\alpha$ if $f_Y/(f_Y + f_O) > 1/2$ and $\ell^* < C'/\alpha$ if $f_Y/(f_Y + f_O) < 1/2$.

For many realistic values of the parameters, ξ is negligible compared with C'/α, and in this case, the approximation in (4.12) simplifies to $\ell^* \cong \frac{C}{\alpha} + 1 - \frac{1}{\alpha}$. Finally, since $0 \leq 1 - \frac{1}{\alpha} \leq 1$ for $1 \geq \alpha \geq .5$, we have:

$$\ell^* \cong C/\alpha. \tag{4.14}$$

For example, this approximation will be very good whenever $f_Y \cong f_O$, and the conditions mentioned above for normal approximations of the binomial distribution apply.

Shlifer and Vardi [123, page 102] note that "the cancellation probability varies from 0.8 for reservations on record a few months in advance to 0.3 when on record one to two weeks before take-off.". This variation of the cancellation probability can easily be modelled in our setting. Let $EB_i = \alpha_i$. The analysis leading to (4.8) is still valid, although now the distribution of $N(\ell)$ is rather complicated. However, as long as the $\alpha's$ do not differ by too much and ℓ is reasonably large, $N(\ell)$ will be approximately normal with mean $\sum_1^\ell \alpha_i$ and variance $\sum_1^\ell \alpha_i(1-\alpha_i)$.

4.3 Group Cancellations

Shlifer and Vardi [123, (1975)] cite a study by the El-Al airline company which indicates that the number of confirmations $N(\ell)$ behaves as though customers cancelled in groups of two. The above analysis is still appropriate for groups of a fixed size, say two, if the model is simply reinterpreted in terms of groups instead of individuals. So $B_i = 1$ if the i-th group confirms and $B_i = 0$ otherwise, f_Y and f_O are the revenue and penalty associated with a group of two, and $N(\ell)$ is the number of groups which confirm if ℓ groups are booked. The capacity of the plane C is the number of groups which it can hold.

If the variance of the number of individuals confirming in the Bernoulli (group of one) case and the group of two case are compared, it will be noticed that the variance for the group of two is twice that for the group of one. A more general analysis of the gain function (4.6) can be carried out in the spirit of Shlifer and Vardi by assuming that the group sizes are small relative to ℓ and that the confirmations are independent among groups. With these assumptions, the event $N(\ell - 1) \geq C$ provides little information about the value of B_ℓ. In this case,

$$\mathrm{E}[B_\ell | N(\ell - 1) \geq C] \cong \mathrm{E}[B_\ell] = \alpha, \tag{4.15}$$

and the optimality condition (4.8) is still approximately correct. As Shlifer and Vardi point out, $N(\ell)$ will still be approximately normal and the mean and variance of $N(\ell)$ should be proportional to ℓ. Let α and β be, respectively, the constants of proportionality. The analysis of the Bernoulli case based on the normal approximation can now be applied and (4.12) is still valid except with

$$\xi = \frac{z^2}{2} \frac{\beta^2}{\alpha^2}. \tag{4.16}$$

4.4 A Comparison with the Shlifer and Vardi Model

It is interesting to compare the optimality condition (4.8) with the condition proposed by Shlifer and Vardi [123, (1975)]. They approximate the passsenger show-up distribution with a normal distribution with mean $\ell\alpha$ and variance $\ell\beta$, where α is the confirmation probability, as before, and β is an adjustment for the variance. The parameter β can either be determined by fitting the normal distribution to historical data or derived by assuming, as we do, that the passenger confirmation process is Bernoulli. As mentioned in the previous section, Shlifer and Vardi reported that airline data examined by them was consistent with a Bernoulli process in which bookings and cancellations occur in groups of two. It is easy to show that in this case $\beta = 2\alpha(1 - \alpha)$. This expression for β is used in the numerical example given below.

Shlifer and Vardi use the normal approximation to derive the following expression (in our notation) for $D(\ell)$, the expected number of passengers denied boarding given ℓ seats booked:

$$D(\ell) = \mathrm{E}[(N(\ell) - C)^+] = \sqrt{\ell\beta}(z(\ell)\Phi[z(\ell)] + \phi[z(\ell)]), \tag{4.17}$$

where

$$z(\ell) = \frac{\ell\alpha - C}{\sqrt{\ell\beta}},$$

and $\Phi[\cdot]$ and $\phi[\cdot]$ denote the cumulative standard normal distribution and its density, respectively. They then derive a condition equivalent to the following for the optimal booking level, ℓ^* :

$$D'(\ell^*) = \frac{\alpha f_Y}{f_Y + f_O}, \tag{4.18}$$

where $D'(\cdot)$ denotes the first derivative of $D(\ell)$ with respect to ℓ. They point out that a solution to this equation is obtainable numerically (presumably by approximating the derivative for a range of ℓ's and then choosing the value closest to the right hand side ratio), and provide a table of typical solutions for various values of the fare and bumping penalty.

Note that by carrying out the derivative in (4.18) one obtains the more basic expression:

$$\Phi[z(\ell^*)] + \frac{\sqrt{\ell^*\beta}}{2\ell^*\alpha}\phi[z(\ell^*)] = \frac{f_Y}{f_Y + f_O}, \tag{4.19}$$

where

$$z(\ell^*) = \frac{\ell^*\alpha - C}{\sqrt{\ell^*\beta}}.$$

Condition (4.19) is a continuous version of the normal approximation (4.10) with the addition of the second term on the left hand side. This term arises because Shlifer and Vardi employ the normal approximation before optimizing while we employ it after. In practice, it has a small effect for realistic values of the parameters. Table 4.1 provides a

Table 4.1: Overbooking Levels for 200 Seat Capacity

f_O/f_Y	3		1		1/3	
α	0.5	0.9	0.5	0.9	0.5	0.9
C/α	400	222	400	222	400	222
binomial[a]	380	218	400	222	420	226
normal approx.[b]	382	218	400	222	420	228
Shlifer & Vardi[c]	380	218	399	222	419	227

---

[a] our optimality condition (4.8)
[b] our normal approximation(4.12)
[c] Shlifer and Vardi, our equation (4.18)

comparison of optimal booking levels calculated with our optimality condition (4.8), the approximations (4.12) and (4.14), and Shlifer and Vardi's condition (4.18). Note that, as expected, the approximation C/α is very good for $\alpha = 0.5$ when $f_O = f_Y$ and that the other approximations are all in close agreement for all parameter values. Note also that it is computationally much easier to calculate booking levels with (4.8) than with (4.18).

4.5 Implementation

The term 'simple' overbooking model has been used repeatedly throughout this chapter for good reason. The actual problem faced by airlines is considerably more complex than that described here. Overbooking must be done over multiple fare classes, and overbooking penalties vary between classes. The existence of various types of 'voluntary bumping' schemes in which passengers on an oversold flight are offered varying levels of compensation to accept a delay to a later flight further complicate the assessment of overbooking penalties. Cancellations occur throughout the booking period so that, ideally, overbooking levels should be adjusted dynamically as the time of flight departure approaches. As discussed in the literature review of chapter 1, some of these complications have been addressed by others (e.g. Alstrup *et al.* [3, (1986)], Rothstein [117, (1968)]), but the resulting formulations were rather unwieldy dynamic programs. Heuristic approaches to the multiple fare overbooking problem such as that of Belobaba [13, (1987)] are easily implemented but of uncertain accuracy.

Notwithstanding these considerations, the results of this chapter can be useful in two ways:

1. The optimality condition is very simple to solve and can be used to provide a good estimate of an optimal overbooking level in those instances when only one fare class is being dealt with; e.g., small aircraft serving remote areas or the first class cabin in larger aircraft. As always, an approximate method for dealing with the dynamics of the booking and cancellation process is to periodically recalculate the overbooking level as the time of flight departure approaches.

2. For airlines lacking sophisticated facilities for setting and monitoring overbooking levels, the approximation C/α gives a simple rule of thumb for setting a nominal total overbooking level for a flight. If the average overbooking penalty and average fare are approximately equal, the nominal overbooking level should be close to C/α; if the average penalty is less than the average fare, the limit should be higher than the ratio, and so on. This heuristic suffers from the same "uncertain accuracy" mentioned above with regard to Belobaba's multiple fare class overbooking heuristic, but it provides a useful reference figure that can be adjusted in an *ad hoc* fashion as the booking process proceeds.

4.6 Summary and Conclusions – Simple Overbooking Model

This chapter has shown that a variant of the general model for determining optimal seat allocations for two dependent fare classes can be applied to the problem of obtaining an overbooking level for a single fare class. Assumption of a Bernoulli process for confirmations leads to an optimality condition requiring only the calculation of quantiles of the binomial probability distribution. This condition is further simplified to one requiring only quantiles of the standard normal distribution, and the connection between this approximation and that of Shlifer and Vardi [123, (1975)] is demonstrated. It is shown that in the event that the overbooking penalty and fare are equal, the simple ratio of capacity to confirmation probability yields a very good approximation to the optimal overbooking level. Extension of the overbooking model to allow for group bookings is also discussed.

The problem of determining an approach to overbooking that accounts for the usual complexities of the process, is feasible to implement, and has known accuracy remains a difficult open problem.

Chapter 5

Estimation of Dependent Demands from Jointly Censored Data

5.1 Introduction

Chapter 3 has shown that it is possible to determine optimal discount booking limits when discount and full fare demands are stochastically dependent. To do so requires estimates for conditional probabilities of the form $\Pr[Y > C - \ell \,|\, X \geq \ell]$ for a sufficient number of values of ℓ that the optimality condition (3.7) can be solved. If the joint distribution of the demands can be estimated, the calculation of these conditional probabilities is straightforward. This chapter develops a methodology for estimating the parameters of the joint demand distribution on the basis of past observations of demand and other, related variables.

It is assumed that the joint distribution of the demands is bivariate normal[1], so the estimation problem reduces to that of estimating the means, variances and correlation of the two classes of demand. This would be a routine statistical exercise were it not for two complications: 1) the joint demand distribution is sensitive to numerous external economic and other factors, and 2) demand cannot be observed once booking limits have been reached. Specifically:

external factors: Airline demands can be highly sensitive to such factors as prevailing economic conditions, fare structures, promotional activities, season, day of week,

---

[1] As mentioned earlier, there is evidence [13, 123] that the normal distribution provides a good approximation for the marginal distributions of demand. Normality of marginal distributions does not guarantee normality of the joint distribution; however, this is a reasonable assumption in the present context.

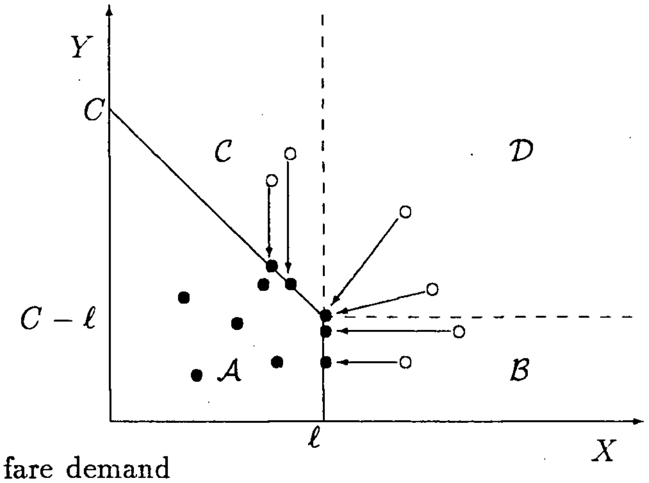
and time of day. Such factors and their effect on passenger choice behaviour have been discussed at some length by Belobaba [13, (1987)]. Overall estimates of the means and variances of demand can be based on aggregate data without regard for these externalities; however, such estimates will have little use in estimating the demand distribution that might be expected to prevail in some particular series of future flights. In particular, it is likely that spurious correlations between two demand classes will be found since demands for all classes can be expected to move together under the influence of most of the factors listed above[2]

censored data: In current airline reservations systems there is no way of capturing the information that a customer has requested a booking on a flight but has been turned away. In a record of total bookings for past flights, the number of flights on which demand reached a booking limit can be determined but not the amount by which the limit was exceeded. Data obtained under such circumstances are described as *censored* in the statistical literature. Figure 5.1 depicts a number of observations of demand for a discount fare class X and a full fare class Y that have been booked in a nested fashion. There is an upper limit of ℓ[3] on X-demand and a maximum capacity limit of C for both demands. Any capacity remaining after X-demand has been satisfied or has reached ℓ can be applied to Y-demand. Demand occurring after the capacity limits have been reached is not recorded, so in such cases it is known that capacity was exceeded but not by how much. In practice, both the

---

[2] It can be argued that some factors might operate in such a way as to produce negative correlation between classes. For example, a general improvement in the economy might increase the number of passengers willing to pay more to avoid the restrictions of discount fare reservations. This might have the effect of increasing full fare demand while reducing discount fare demand. However, it seems likely that the very real increase in both types of traffic experienced between different days of the week and between seasons would overwhelm such an effect.

[3] In this context, ℓ refers either to a limit on the number of discount seats that could be sold or to the number of discount seats that had been sold before the time limit for bookings in the fare class was reached.



Y : full fare demand
X : discount fare demand
C: capacity
ℓ: discount class booking limit
● : denotes observations
○ : denotes censored observations (mapped onto observations as shown)
$\mathcal{A}$: $\{i : x_i \leq \ell \text{ and } y_i \leq C - \ell\}$
$\mathcal{B}$: $\{i : x_i > \ell \text{ and } y_i \leq C - \ell\}$
$\mathcal{C}$: $\{i : x_i \leq \ell \text{ and } y_i > C - \ell\}$
$\mathcal{D}$: $\{i : x_i > \ell \text{ and } y_i > C - \ell\}$

Figure 5.1: Censored Demand With Nested Fare Classes

booking limit and plane capacity may vary from flight to flight; however, this in no way interferes with the analysis described here. No generality is lost by assuming that both parameters are constant.

The approach taken here to the estimation problem is to account for the effects of the external factors with a bivariate multiple regression model in which the two demand classes are dependent variables and the external factors are regressors. The censorship of the dependent variables is handled by an adaption of the so-called EM method of

Dempster, Laird and Rubin [36, (1977)], to be described later.

This chapter is organized as follows. The next section discusses two examples of prior approaches to analysis of censored airline demand data, and very briefly surveys the extensive literature on censored data analysis in other areas. Subsection 5.2.1 provides a summary of censored regression analysis and the EM method that forms the basis of the subsequent analysis. Section 5.3 shows that the EM method can be applied to the particular problem presented by airline demand data. The final section presents the results of numerical trials of the method on simulated airline demand data[4].

5.2 Background: Estimation from Censored Data

The simplest approaches to estimation with censored data are 1) to simply ignore the censorship and perform the estimation on the data as is, or 2) to discard any censored observations and perform the estimation on the remaining data (this is equivalent to estimating on the basis of a sample from a truncated distribution). Neither of these strategies are particularly desirable. Approach 1) will lead to badly biased estimates when more than a small proportion of observations are censored, and approach 2) is even worse since no information at all is retained about data beyond the censoring limits. As will be shown later, there are far better approaches to analysis of censored data

Two Examples of Airline Demand Estimation

Two examples of previous work on the estimation of airline demands are discussed here. The first, described in a technical report by D. L. Harmer of the Boeing Company [62, (1976)], dealt with the problem of estimating a single, stable normal demand distribution on the basis of censored data. The second, which appeared in Belobaba's 1987 Ph.D.

---

[4] The author was unable to obtain actual airline data in time for these numerical trials.

dissertation [13, (1987)], addressed the problem of detection of correlation among four fare classes.

In the Boeing report, Harmer [62, (1976)] discussed the problem of dealing with censored data in estimation of the mean and variance of a univariate normal demand distribution. The underlying demand distribution was assumed to be stable, so complication 1), above, was not a problem. He performed a least-squares fit of the observed demands against their percentiles (on normal probability paper) and estimated the mean as the 'predicted value' for the 0.50 quantile. He estimated the standard deviation by subtracting the estimated mean from the the predicted value for the 0.8413 quantile. To avoid biasing the least squares fit by the observations clustered around the censorship point, he discarded those observations. This approach bears some resemblance to *linear unbiased* estimation in the statistical literature (see, for example, Nelson and Schmee [103, (1979)] or Sarhan and Greenberg [121, (1956)]). The discarding of censored data in this case is not as serious as it is in the direct estimation case because some of the information contained in the censored data is reflected in the quantiles assigned to the uncensored data points. This approach is apparently still in use (see Belobaba [15, (1989)]).

Belobaba [13, pp147–150,(1987)] described empirical tests for correlations among demands in four fare classes on flights of Western Airlines occurring in a six month period. To avoid complication 1) above, he restricted sampling to flights that could be regarded as homogeneous with regard to the external factors that might influence demand. To handle the problem of censored observations, he restricted sampling to flights occurring during a low-demand season and discarded any observations in which one or other demand was censored. Belobaba found no compelling evidence of correlation between discount and full fare demands.

Discarding censored observations in this way corresponds to truncation of data. It is known that the correlation coefficient of the truncated bivariate normal distributed underestimates the correlation coefficient of the underlying population (see Johnson and Kotz [71, page 112ff,(1972)]). Belobaba does not report the number of observations that were discarded, so it is not clear whether this bias was significant. Assuming that the bias is not significant, an appropriate conclusion from that study is that there is little evidence of correlation between discount and full fare demands when market conditions are such that both demands are low.

Analysis of Censored Data in Other Areas

There has been a great deal of work done on censored data problems in the areas of reliability testing, lifetime estimation, biomedical statistics, and econometrics. This type of problem arises, for example, in reliability or lifetime testing when it is necessary to terminate an experiment before all test items have failed. In econometric modelling, it is often the case that values of one or more dependent variables cannot occur or cannot be observed outside of a particular range, and this must be taken into account when fitting a model to the data.

The literature in these areas is extensive. For good surveys of related results in the lifetime and reliability area, see the books by Lawless [83, (1982)] and Nelson [101, (1982)]. For a brief general survey of the statistical literature on censored data analysis, see the article by McCool in Kotz and Johnson [96, (1982)]. For an overview of work on censored data in the econometric area, see the book by Maddala [92, (1983),Chapters 6 and 7]. A survey article by Amemiya [5, (1984)] deals exclusively with regression analysis on censored data in *Tobit* models of econometrics and contains a good bibliography.

The particular problem encountered with airline data is called *Type I censorship* in the statistical literature because the number of censored obsevations is not determined in advance. (Type II censorship occurs chiefly in lifetime testing when it is decided in advance that an experiment will be terminated when a fixed number of test items have failed.)

5.2.1 Censored Regression Analysis and the EM Method

This section briefly describes the censored regression problem in a general setting, places the problem of airline demand estimation in that context, reviews some of the methods available for analysis, and introduces the method chosen for the present analysis — the EM method. Most of the general discussion here is based on the survey article by Amemiya [5, (1984)].

The earliest discussion of the censored regression problem in the econometric literature is that of Tobin [137, (1958)] who examined a model of the ratio of durable-goods expenditures to disposable incomes by households as a function of age of the head of household and the ratio of liquid assets to disposable income. The data on the dependent variable (expenditures) were censored from below by the lowest available price of a durable good; that is, there were observations made of households whose disposable incomes were below the minimum price, so no expenditures were made. Tobin described a regression model that accounted for the censorship and worked out the details of a maximum likelihood procedure for estimating the parameters of the model. The type of model dealt with by Tobin eventually came to be known as the *Tobit* model (an amalgam of Tobin and Probit).

There have been many subsequent studies describing variants of the Tobit model. Amemiya [5, (1984)] classifies the reported work into five categories depending upon the number of dependent variables and the nature of the censorship. Each of the categories involves from one to three dependent variables, and in every case the censorship on all dependent variables is determined by one of the dependent variables. That is, the dependent variables were censored or not censored accordingly as variable 1 was censored or not censored. The problem considered here does not fit into any of these categories since censorship of either of the two demand variables in airline data can occur independently of censorship of the other variable[5]. As a consequence, the likelihood function for the airline data is more complex than that of any of the five categories of Tobit models discussed by Amemiya. Notwithstanding this distinction, the general options available for analysis of airline demand data are the same as those for the Tobit models.

There are a number of options available for estimating censored regressions, including direct maximization of the likelihood function using Newton or Quasi-Newton optimization techniques, a 'two-step' procedure due to Heckman [66, (1976)], and an iterative approach to missing data problems proposed by Hartley [65, (1958)] and others, and generalized as the *EM method* of Demster, Laird and Rubin [36, (1977)]. Direct maximization is rejected as an option here because of the complexity of the likelihood function (to be described later). The EM method was chosen in preference to the Heckman procedure because of its greater generality; that is, in the initial stages of the work described here the EM method seemed to offer greater potential adaptability to the peculiarities of the problem. It is possible that the Heckman procedure could be used on this problem, but that possibility was not explored.

The EM Method for General Incomplete Data Problems

For purposes of illustration, this summary will assume that the parameters of some univariate distribution are to be estimated. The basic principles of the method apply without change to the bivariate airline demand case discussed later.

In its most general form, the EM method deals with situations in which there is a vector $\tilde{\mathbf{x}}$ of observed *incomplete* data (wholly or partially censored demands in our case) corresponding to a vector $\mathbf{x}$ of unobserved *complete* data from some sample space (true

---

[5] In an earlier article cite[1974]amemiya74, Amemiya dealt with a bivariate simultaneous equations model with censorship similar to that considered here. However, he was able to exploit special structure to transform the model to one with simpler censorship.

demands in our case). The same $\tilde{\mathbf{x}}$ could arise from many different $\mathbf{x}$ vectors, and we know how the $\mathbf{x}$ are mapped onto incomplete data points. (In the airline demand problem, the booking limits that are causing the censorship are known.) Realizations in the sample space occur according to some probability function $f(\mathbf{x}|\boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ is to be estimated. This determines a probability function $g(\cdot|\boldsymbol{\theta})$ for the observed data through

$$g(\tilde{\mathbf{x}}|\boldsymbol{\theta}) = \int_{\mathcal{S}} f(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}; \tag{5.1}$$

where $\mathcal{S}$ is the subset of the sample space that maps onto the observation $\tilde{\mathbf{x}}$.

The objective is to estimate the parameters $\boldsymbol{\theta}$ by maximizing the log likelihood function $\tilde{\mathcal{L}}(\boldsymbol{\theta}|\tilde{\mathbf{x}}) = \log g(\tilde{\mathbf{x}}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. This function will often be difficult to maximize because of its complex form. The log likelihood function for the unobserved data $\mathbf{x}$ is $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \log \mathrm{f}(\mathbf{x}|\boldsymbol{\theta})$, and this function often has a much more tractable form for maximization with respect to $\boldsymbol{\theta}$. However, in the absence of exact knowledge of $\mathbf{x}$, $\mathcal{L}(\boldsymbol{\theta}|\mathbf{X})$ is a random variable over the sample space.

Now suppose some prior estimate $\boldsymbol{\theta}^o$ of the parameters is available, and the incomplete data $\tilde{\mathbf{x}}$ have been observed. Then it is possible, in principle, to calculate the expected value of the random likelihood function, *conditional* on $\boldsymbol{\theta}^o$ and $\tilde{\mathbf{x}}$. That is, it is possible to calculate

$$\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^o) = \mathrm{E}[\mathcal{L}(\boldsymbol{\theta}|\mathbf{X})|\tilde{\mathbf{x}}, \boldsymbol{\theta}^o]. \tag{5.2}$$

This is the expectation step of the EM method — the 'E' in EM. The maximization, or 'M' step, involves maximizing the expectation in (5.2), thereby obtaining a new estimate for $\boldsymbol{\theta}$. A complete execution of the EM method involves finding an initial estimate for $\boldsymbol{\theta}$ and then iterating between the E and M steps until convergence is achieved to some value of $\boldsymbol{\theta}$.

Application of the EM method to any particular problem involves finding a convenient way of computing the expectation of the E-step and of carrying out the maximization. Generally, for the method to be practical, it is required that the expectation be expressed in some convenient functional form and that there exist some efficient way of maximizing that function.

It is not at all obvious that this process will lead to the maximization of the original log likelihood function $\tilde{\mathcal{L}}(\boldsymbol{\theta})$; however, Dempster *et al.* demonstrate that $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is nondecreasing over the sequence of parameter estimates generated by the EM algorithm. Thus, if the likelihood is bounded, the algorithm is guaranteed to converge to some fixed value $\tilde{\mathcal{L}}^*$. In general, however, there is no guarantee that $\tilde{\mathcal{L}}^*$ will be a local maximum or even a stationary point of $\tilde{\mathcal{L}}(\boldsymbol{\theta})$. Thus convergence properties of the method must be checked for each application. Discussion of convergence properties of the EM method with censored airline data will be deferred until after the details of this application have been presented in the next section.

5.3 A Jointly Censored Bivariate Multiple Regression Model

Let X and Y denote random demands which are jointly related to a set of r variables $w_1, \ldots w_r$. It is assumed that a set of n demand pairs $\{(X_i, Y_i) : i = 1, \ldots, n\}$ satisfies the linear system

$$X_i = \boldsymbol{\alpha}\mathbf{W}_i + \delta_i \tag{5.3}$$

$$Y_i = \boldsymbol{\beta}\mathbf{W}_i + \epsilon_i, \tag{5.4}$$

where for each i, $\mathbf{W}_i = (1, w_{i1}, \ldots, w_{ir})'$ is a column vector of regressors, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_r)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)$ are row vectors of regression coefficients. For every i, the errors δ_i and ϵ_i are assumed to be jointly distributed according to a bivariate normal distribution with zero means, variances σ^2 and τ^2 respectively, and correlation

ρ.[6]

Under the normality assumption, each value of the vector $\mathbf{W}$ determines a joint density function for X and Y through (5.3) and (5.4). This joint density can be written:

$$f(x,y) = \left[2\pi\sigma\tau\sqrt{1-\rho^2}\right]^{-1} \exp[-Q((x-\boldsymbol{\alpha}\mathbf{W}),(y-\boldsymbol{\beta}\mathbf{W}))], \tag{5.5}$$

where

$$Q(u,v) = [2(1-\rho^2)]^{-1}\left(\left(\frac{1}{\sigma^2}\right)u^2 - \left(\frac{2\rho}{\sigma\tau}\right)uv + \left(\frac{1}{\tau^2}\right)v^2\right). \tag{5.6}$$

Let $\boldsymbol{\theta}$ denote the vector of $2r+5$ parameters $(\alpha_0,\ldots,\alpha_r;\beta_0,\ldots,\beta_r;\sigma,\tau,\rho)$ in (5.5). The objective is to estimate $\boldsymbol{\theta}$ on the basis of data $\mathbf{x} = (x_1,\ldots,x_n)', \mathbf{y} = (y_1,\ldots,y_n)'$ and $\mathbf{W}_1,\ldots,\mathbf{W}_n$. In the absence of censorship of the data, the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$ is easily obtained. This calculation is summarized here for later reference.

From (5.5), the data have the log likelihood function

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x},\mathbf{y}) = -n\log(2\pi\sigma\tau\sqrt{1-\rho^2}) - \sum_{i=1}^{n} Q((x_i-\boldsymbol{\alpha}\mathbf{W}_i),(y_i-\boldsymbol{\beta}\mathbf{W}_i)). \tag{5.7}$$

It is well known[7] that the MLE's for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the least squares estimators (LSE's) $\hat{\boldsymbol{\alpha}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}$, where $\mathbf{X} = (X_1,\ldots,X_n)'$ and $\mathbf{Y} = (Y_1,\ldots,Y_n)'$, and $\mathbf{W}$ is the n by $r+1$ matrix with ith row $\mathbf{W}_i'$. Once $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ have been obtained, maximum likelihood estimates for the remaining parameters can be calculated from

$$\hat{\sigma}^2 = (1/n)\sum_1^n (x_i - \hat{\boldsymbol{\alpha}}\mathbf{W}_i)^2, \tag{5.8}$$

$$\hat{\tau}^2 = (1/n)\sum_1^n (y_i - \hat{\boldsymbol{\beta}}\mathbf{W}_i)^2, \tag{5.9}$$

---

[6] Constant variance is a strong assumption, as in all regression work. Belobaba [13, (1987),p143-144] gives full fare demand statistics for 21 flight/day-of-week combinations that show quite stable standard deviations when allowance is made for experimental variation. Problems with non-constant variance can be handled with weighted least squares methods, but that possibility will not be discussed further here.

[7] See,for example, Johnson and Wichern [73, (1982), p324].

and

$$\hat{\rho} \;=\; (1/n\hat{\sigma}\hat{\tau})\sum_{1}^{n}(x_i - \hat{\alpha}\mathbf{W}_i)(y_i - \hat{\beta}\mathbf{W}_i). \tag{5.10}$$

It is important to note here that the maximization of $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})$ can be done first over $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and then, separately, over (σ,τ,ρ). That is, the log likelihood is *separable* with respect to these two groups of parameters.

Now suppose that the data are censored in the manner depicted in Figure 5.1. A description of this type of censorship was given earlier [page 80]. As mentioned there, both C and ℓ may be different for each observation without affecting the validity of the following analysis. They are kept constant here to avoid unnecessary subscripts.

Given demands X and Y, the *observed* demands $\tilde{X}$ and $\tilde{Y}$ will be

$$\tilde{X} \;=\; \begin{cases} X & \text{if } X \le \ell \\ \ell & \text{if } X > \ell. \end{cases} \tag{5.11}$$

$$\tilde{Y} \;=\; \begin{cases} Y & \text{if } Y \le C - \tilde{X} \\ C - \tilde{X} & \text{if } Y > C - \tilde{X} \end{cases} \tag{5.12}$$

With censorship of the data in this manner, the log likelihood function corresponding to the data $\tilde{\mathbf{x}},\tilde{\mathbf{y}}$ and $\mathbf{W}$ is now:

$$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{\theta};\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \;=\; & \sum_{i\in\mathcal{A}} \log f_i(\tilde{x}_i,\tilde{y}_i) \\ & + \sum_{i\in\mathcal{B}} \log \int_{\ell}^{\infty} f_i(u,\tilde{y}_i)\,du \\ & + \sum_{i\in\mathcal{C}} \log \int_{C-\tilde{x}_i}^{\infty} f_i(\tilde{x}_i,v)\,dv \\ & + \sum_{i\in\mathcal{D}} \log \int_{\ell}^{\infty}\int_{C-\ell}^{\infty} f_i(u,v)\,du\,dv, \end{aligned} \tag{5.13}$$

where $f_i(\cdot,\cdot)$ denotes the density given by (5.5) corresponding to data points $\tilde{x}_i, \tilde{y}_i$, and $\mathbf{W}_i$; and $\mathcal{A},\ldots,\mathcal{D}$ are the index sets described in Figure 5.1.

The MLE for $\boldsymbol{\theta}$ can be obtained, in principle, by direct maximization of $\tilde{\mathcal{L}}(\cdot)$ with respect to the parameters. However, the presence of the integral expressions in the last three terms of (5.13) make it impossible to derive tractable first order conditions for the maximum. Thus it is necessary to resort to numerical methods. (Note that even in simpler censored regressions where first order conditions can be obtained, it is still necessary to use numerical methods to solve the first order conditions.)

5.3.1 Maximization of the Likelihood Function with the EM Method

A variety of methods have been devised for obtaining MLE's for the parameters in censored regressions (see, for example, Amemiya [5, (1984)] and Maddala [92, (1983)]). All of the practical methods involve iterative refinement of estimates starting from an initial set. This section describes an application of the EM method to the bivariate censored regression described above.

The log likelihood function for the actual (uncensored) demands is given in (5.7), but this function is not available because some of the data have been censored. However, if some prior estimate $\boldsymbol{\theta}^o \equiv (\boldsymbol{\alpha}^o;\boldsymbol{\beta}^o;\sigma_o,\tau_o,\rho_o)$ is available for the parameters, it is possible to calculate the *expected value* of the log likelihood conditional upon the observed (censored) demands $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. That is, it is possible to calculate

$$\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^o) = \mathrm{E}[\mathcal{L}(\boldsymbol{\theta};X,Y)|\tilde{\mathbf{x}},\tilde{\mathbf{y}},\boldsymbol{\theta}^o]. \tag{5.14}$$

This calculation corresponds to the expectation step of the EM method, and the maximization step involves maximizing $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^o)$ over $\boldsymbol{\theta}$.

To avoid cumbersome expressions in what follows, a tilde, $\sim$, over an operator (e.g. $\tilde{\mathrm{E}}$, $\widetilde{\mathrm{Var}}$ and $\widetilde{\mathrm{Cov}}$) will denote conditioning on the observations $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ given the prior parameter estimate $\boldsymbol{\theta}^o$. Thus, for example,

$$\begin{aligned}
\tilde{\mathrm{E}}[X_i] &= \mathrm{E}[X_i|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \boldsymbol{\theta}^o] \\
\tilde{\mathrm{E}}[X_i] &= \begin{cases} \tilde{x}_i & \text{for } i \in \mathcal{A}, \mathcal{C} \\ \mathrm{E}[X_i|X_i > L, Y_i = \tilde{\mathbf{y}}, \boldsymbol{\theta}^o] & \text{for } i \in \mathcal{B} \\ \mathrm{E}[X_i|X_i > L, Y_i > C - L, \boldsymbol{\theta}^o] & \text{for } i \in \mathcal{D}, \end{cases} \\
\widetilde{\mathrm{Var}}[X_i] &= \begin{cases} 0 & \text{for } i \in \mathcal{A}, \mathcal{C} \\ \tilde{\mathrm{E}}[(X_i - \tilde{\mathrm{E}}[X_i])^2] & \text{for } i \in \mathcal{B}, \mathcal{D}, \end{cases} \\
\text{and} & \\
\widetilde{\mathrm{Cov}}[X_i, Y_i] &= \begin{cases} 0 & \text{for } i \in \mathcal{A}, \mathcal{B}, \mathcal{C} \\ \tilde{\mathrm{E}}[(X_i - \tilde{\mathrm{E}}[X_i])(Y_i - \tilde{\mathrm{E}}[Y_i])] & \text{for } i \in \mathcal{D}. \end{cases}
\end{aligned}$$

From (5.7) and (5.14)

$$\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^o) = -n\log(2\pi\sigma\tau\sqrt{1-\rho^2}) - \sum_{i=1}^{n} \tilde{\mathrm{E}}[Q((X_i - \boldsymbol{\alpha}\mathbf{W}_i), (Y_i - \boldsymbol{\beta}\mathbf{W}_i))]. \qquad (5.15)$$

Expansion of the $\tilde{\mathrm{E}}[Q(\cdot,\cdot)]$ terms in (5.15) requires conditional expectations of sums of squares and cross products of the deviations $(X_i - \boldsymbol{\alpha}\mathbf{W}_i)$ and $(Y_i - \boldsymbol{\beta}\mathbf{W}_i)$. These can be decomposed as follows:

$$\begin{aligned}
\tilde{\mathrm{E}}[(X_i - \boldsymbol{\alpha}\mathbf{W}_i)^2] &= \mathrm{E}[(X_i - \boldsymbol{\alpha}\mathbf{W}_i)^2|\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \boldsymbol{\theta}^o] \\
&= (\tilde{\mathrm{E}}[X_i] - \boldsymbol{\alpha}\mathbf{W}_i)^2 + \widetilde{\mathrm{Var}}[X_i], \qquad (5.16)
\end{aligned}$$

$$\tilde{\mathrm{E}}[(Y_i - \boldsymbol{\beta}\mathbf{W}_i)^2] = (\tilde{\mathrm{E}}[Y_i] - \boldsymbol{\beta}\mathbf{W}_i)^2 + \widetilde{\mathrm{Var}}[Y_i], \qquad (5.17)$$

and

$$\tilde{\mathrm{E}}[(X_i - \boldsymbol{\alpha}\mathbf{W}_i)(Y_i - \boldsymbol{\beta}\mathbf{W}_i)] = (\tilde{\mathrm{E}}[X_i] - \boldsymbol{\alpha}\mathbf{W}_i)(\tilde{\mathrm{E}}[Y_i] - \boldsymbol{\beta}\mathbf{W}_i) + \widetilde{\mathrm{Cov}}[X_i, Y_i]. \qquad (5.18)$$

Calculation of the conditional moments in the right-hand sides of these expressions is discussed in the next section.

Now from (5.6) and (5.15) through (5.18),

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ) &= -n\log(2\pi\sigma\tau\sqrt{1-\rho^2}) - \sum_{i=1}^{n} Q((\tilde{\mathrm{E}}[X_i]-\boldsymbol{\alpha}\mathbf{W}_i),(\tilde{\mathrm{E}}[Y_i]-\boldsymbol{\beta}\mathbf{W}_i)) \\ &\quad -[2(1-\rho^2)]^{-1}\sum_{i=1}^{n}\left(\frac{\widetilde{\mathrm{Var}}[X_i]}{\sigma^2}\right)+\left(\frac{2\rho\widetilde{\mathrm{Cov}}[X_i,Y_i]}{\sigma\tau}\right)+\left(\frac{\widetilde{\mathrm{Var}}[Y_i]}{\tau^2}\right).\end{aligned} \tag{5.19}$$

The maximizer $\hat{\boldsymbol{\theta}}$ of $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ)$ can be obtained by direct calculation, but it is simpler,[8] instead, to exploit the similarities between the expressions for $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ)$ given above and the log likelihood for uncensored data given in (5.7). The key similarity is that $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ)$ is also separable with respect to $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and (σ,τ,ρ).

First, note that the final summation term in (5.19) is independent of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Aside from this term, (5.7) and (5.19) are identical except that in (5.19) expected values are substituted for censored observations. But then $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ)$ will be maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by replacing all censored observations with their expected values and finding the LSE's $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, as was done with (5.7).

Now compare (5.7) with (5.15), above. The two expressions are identical except that in (5.15) the sums of squares and cross products of deviations are replaced with their expected values. Furthermore, because of the separability property, the optimal expected values are calculated with $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Then, from (5.8),(5.9) and (5.10), $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^\circ)$ is maximized with respect to σ,τ and ρ by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathrm{E}}[(X_i-\hat{\boldsymbol{\alpha}}\mathbf{W}_i)^2] \tag{5.20}$$

$$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathrm{E}}[(Y_i-\hat{\boldsymbol{\beta}}\mathbf{W}_i)^2] \tag{5.21}$$

and

---

[8] This argument parallels that for the univariate multiple regression case. See, for example, [5, pp21–23].

$$\hat{\rho} = \frac{1}{n\hat{\sigma}\hat{\tau}} \sum_{i=1}^{n} \tilde{\mathrm{E}}[(X_i - \hat{\boldsymbol{\alpha}}\mathbf{W}_i)(Y_i - \hat{\boldsymbol{\beta}}\mathbf{W}_i)]; \tag{5.22}$$

where the $\tilde{\mathrm{E}}[\cdot]$ terms are obtained by substituting $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ in (5.16),(5.17) and (5.18).

Thus one iteration of the EM method applied to the present problem involves the following steps:

1. Use the prior parameter estimate $\boldsymbol{\theta}^o = (\alpha^o; \beta^o; \sigma_o, \tau_o, \rho_o)$ and the observations $(\tilde{x}_i, \tilde{y}_i)$ to calculate the expected values $\tilde{\mathrm{E}}[X_i]$ and $\tilde{\mathrm{E}}[Y_i]$ for all censored observations.

2. Replace the censored observations with their expected values and perform a least squares analysis on the resulting data. This will yield the new regression coefficient vectors $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ as well as the sums of squares and cross products: $\sum_{i=1}^{n}(\tilde{\mathrm{E}}[X_i] - \boldsymbol{\alpha}\mathbf{W}_i)^2, \sum_{i=1}^{n}(\tilde{\mathrm{E}}[Y_i] - \boldsymbol{\beta}\mathbf{W}_i)^2$ and $\sum_{i=1}^{n}(\tilde{\mathrm{E}}[X_i] - \boldsymbol{\alpha}\mathbf{W}_i)(\tilde{\mathrm{E}}[Y_i] - \boldsymbol{\beta}\mathbf{W}_i)$.

3. Use $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \sigma_o, \tau_o$ and ρ_o to calculate the variance terms $\widetilde{\mathrm{Var}}[X_i]$, $\widetilde{\mathrm{Var}}[Y_i]$ and $\widetilde{\mathrm{Cov}}[X_i, Y_i]$ for each censored observation. Then use these values together with the sums of squares and cross products from step 2 to obtain new variance and correlation estimates $\hat{\sigma}^2, \hat{\tau}^2$ and $\hat{\rho}$ using equations (5.20),(5.21) and (5.22) along with (5.16), (5.17) and (5.18).

4. Replace $\boldsymbol{\theta}^o$ with the new estimates and return to step 1.

It remains to provide formulae for the conditional expectations, variances and covariance $\tilde{\mathrm{E}}[X_i], \tilde{\mathrm{E}}[Y_i], \widetilde{\mathrm{Var}}[X_i], \widetilde{\mathrm{Var}}[Y_i]$ and $\widetilde{\mathrm{Cov}}[X_i, Y_i]$.

5.3.2 Conditional Expectations, Variances and Covariance

A complete derivation of conditional moments for truncated multivariate normal random variables is provided in Appendix (A). Determination of expressions for the conditional moments required here can be accomplished by substitution into the expressions provided in the appendix. For brevity, expansions will be given only for $\tilde{\mathrm{E}}[X_i]$, $\widetilde{\mathrm{Var}}[X_i]$ and $\widetilde{\mathrm{Cov}}[X_i, Y_i]$.

Note first that the appropriate expansion for an observation depends on the censorship region in which the observation falls. Thus, for example, for an observation falling in regions $\mathcal{A}$ or $\mathcal{C}$ (see Figure 5.1), the X-demand is not censored, so $\tilde{\mathrm{E}}[X_i] = \tilde{x}_i$. For an observation falling in region $\mathcal{B}$, the X-demand is censored but not the Y-demand, so the appropriate expansion for $\tilde{\mathrm{E}}[X_i]$ is an expectation based on a truncated conditional distribution from the bivariate normal distribution (Appendices A.1.2 and A.2.1). Finally, for an observation falling in region $\mathcal{D}$, both X and Y-demands are censored, so the appropriate expansion is a conditional expectation from a truncated bivariate normal distribution (Appendix A.2.3). Thus,

$$\tilde{\mathrm{E}}[X_i] = \begin{cases} \tilde{x}_i & \text{for } i \in \mathcal{A}, \mathcal{C} \\ \mathrm{E}[X_i | X_i > L, Y_i = \tilde{y}_i] & \text{for } i \in \mathcal{B} \\ \mathrm{E}[X_i | X_i > L, Y_i > C - L] & \text{for } i \in \mathcal{D} \end{cases} \tag{5.23}$$

Now from Appendix A.1.2, equation (A.9) and Appendix A.2.1, equations (A.17) and (A.18), we have

$$\mathrm{E}[X_i | X_i > L, Y_i = \tilde{y}_i] = \mathrm{E}[X_i | \tilde{y}_i] + \sigma_o \sqrt{1 - \rho_o^2}\, \mathrm{H}\left(\frac{L - E[X_i | \tilde{y}_i]}{\sigma_o \sqrt{1 - \rho_o^2}} \right), \tag{5.24}$$

where $\mathrm{E}[X_i|\tilde{y}_i] = \hat{\alpha}\mathbf{W}_i + (\rho_o\sigma_o/\tau_o)(\tilde{y}_i - \hat{\beta}\mathbf{W}_i)$, and $\mathrm{H}(z) = \phi(z)/\Phi(-z)$ is the *hazard rate* of the univariate normal distribution evaluated at z. (The standard univariate normal density and distribution are denoted $\phi(\cdot)$ and $\Phi(\cdot)$, respectively.)

The expressions for observations falling in region $\mathcal{D}$ are cumbersome, so the following abbreviations will be employed:

$$x_L = (L - \hat{\alpha}\mathbf{W}_i)/\sigma_o$$

$$\begin{aligned}
y_L &= (C - L - \hat{\beta}\mathbf{W}_i)/\tau_o \\
\Phi_x &= \phi(x_L)\,\Phi\left(\frac{\rho_o x_L - y_L}{\sqrt{1-\rho_o^2}}\right) \\
\Phi_y &= \phi(y_L)\,\Phi\left(\frac{\rho_o y_L - x_L}{\sqrt{1-\rho_o^2}}\right) \\
\phi &= \phi(x_L, y_L) \\
\Phi &= \Phi(-x_L, -y_L),
\end{aligned}$$

where $\phi(\cdot,\cdot)$ and $\Phi(\cdot,\cdot)$ denote the bivariate normal joint density and distribution for standardized variables with correlation ρ_o.

From Appendix A.2.3, equations (A.28), (A.29) and (A.33), we have

$$\mathrm{E}[X_i | X_i > L, Y_i > C - L] = \hat{\alpha}\mathbf{W}_i + \sigma_o\left(\frac{\Phi_x + \rho_o\Phi_y}{\Phi}\right).$$

The expansion for $\widetilde{\mathrm{E}}[Y_i]$ is similar.

The conditional variance for X is given by

$$\widetilde{\mathrm{Var}}[X_i] = \begin{cases} 0 & \text{for } i \in \mathcal{A}, \mathcal{C} \\ \mathrm{Var}[X_i | X_i > L, Y_i = \tilde{y}_i] & \text{for } i \in \mathcal{B} \\ \mathrm{Var}[X_i | X_i > L, Y_i > C - L] & \text{for } i \in \mathcal{D}. \end{cases} \tag{5.25}$$

Then, using Appendix A.1.2, equation (A.11) and Appendix A.2.1, equation (A.17), we have

$$\mathrm{Var}[X_i | X_i > L, Y_i = \tilde{y}_i] = \sigma_o^2(1-\rho_o^2)\left(1 - \mathrm{H}'\left(\frac{L - E[X_i|\tilde{y}_i]}{\sigma_o\sqrt{1-\rho_o^2}}\right)\right),$$

where $\mathrm{H}'(\cdot)$ denotes the first derivative of the hazard rate evaluated at $(\cdot)$. From Appendix A.2.3, equations (A.28), (A.29) and (A.35), we have

$$\mathrm{Var}[X_i | X_i > L, Y_i > C - L] = \\ \sigma_o^2\left[1 + \frac{x_L\Phi_x + \rho_o^2 y_L \Phi_y}{\Phi} + \rho_o(1-\rho_o^2)\frac{\phi}{\Phi} - \left(\frac{\Phi_x + \rho_o\Phi_y}{\Phi}\right)^2\right].$$

Finally, the conditional covariance is given by

$$\widetilde{\mathrm{Cov}}[X_i] = \begin{cases} 0 & \text{for } i \in \mathcal{A}, \mathcal{B}, \mathcal{C} \\ \mathrm{Cov}[X_i, Y_i | X_i > L, Y_i > C - L] & \text{for } i \in \mathcal{D}. \end{cases} \tag{5.26}$$

Then, from Appendix A.2.3, equations (A.37) (A.28) and (A.29) we have

$$\mathrm{Cov}[X_i, Y_i | X_i > L, Y_i > C - L] =$$
$$\sigma_o \tau_o \left[\rho_o + \rho_o \frac{x_L \Phi_x + y_L \Phi_y}{\Phi} + (1 - \rho_o^2) \frac{\phi}{\Phi} - \frac{(\Phi_x + \rho_o \Phi_y)(\rho_o \Phi_x + \Phi_y)}{\Phi^2} \right].$$

These expressions are daunting in appearance but are easy to evaluate with the aid of a computer. Standard routines are available for calculating the univariate normal probabilities $\Phi(\cdot)$ and the bivariate tail probability $\Phi(\cdot, \cdot)$.

5.3.3 Convergence Properties

As mentioned previously, the EM method always converges to some fixed value of the likelihood function. There is, however, no guarantee that the fixed value will be a local maximum or even a stationary point of the likelihood, or that the sequence of parameter estimates will converge. These properties must be checked in each particular application of the method.

Let $\tilde{\mathcal{L}}^*$ be the limit of the censored log likelihood function over the sequence of iterates of the EM method. Wu [151, (1983)] gives conditions for: 1) $\tilde{\mathcal{L}}^*$ to be a stationary point of $\tilde{\mathcal{L}}(\boldsymbol{\theta})$, and 2) the sequence of estimates of $\boldsymbol{\theta}$ to converge. A sufficient condition for property 1) to hold is that $\mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\theta}^o)$ is continuous in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^o$. This is true here over the parameter space defined by $\boldsymbol{\alpha} \in \Re^{r+1}, \boldsymbol{\beta} \in \Re^{r+1}, \sigma > 0, \tau > 0, \rho^2 < 1$, thus the EM method will converge to a stationary point of the likelihood function (5.13). (Convergence to any of the boundary points $\sigma = 0, \tau = 0$, or $\rho^2 = 1$ indicates mis-specification of the model.)

An example given by Murray [97, (1977)] shows that the EM method may converge to a saddle point of the likelihood for incomplete data from a bivariate normal distribution. The censored regression described above involves estimation from censored bivariate normal data, so it is conceivable that the same condition arise here. It is also conceivable, though very unlikely, that the sequence of parameter estimates 'cycle' through a set of separate stationary points all having the same value of the likelihood. Wu [151, (1983),p102] points out that convergence to a stationary value, local maximum or global maximum depends on the choice of starting points and recommends that several EM iterations be tried with different starting points that are representive of the parameter space.

These concerns are primarily of a technical nature. From a practical standpoint, convergence problems become less and less likely as sample size increases. In the case of estimation of airline demands, sample sizes can be expected to be large and, given a correctly specified model, convergence of the EM method to the MLE of the parameters can be anticipated in most cases. If the sample size is small or if censorship of the data is extreme, care should be taken to try several starting points for the algorithm and, ideally, to check the properties of the likelihood function in the vicinity of the solution. In the unlikely event that cycling of the parameter estimates occurs, this will become immediately evident during execution of the algorithm.

5.3.4 Problems with More than Two Dependent Variables

The expressions for the moments of the truncated multivariate normal distribution that are provided in the appendix apply to distributions of arbitrary dimension. Also the development of the EM algorithm for the bivariate censored regression can be generalized to regressions with arbitrary numbers of dependent variables with no new conceptual framework required. Thus it is possible to apply these methods to the analysis of airline data from more than two fare classes.

The number of censorship regions that must be allowed for is 2^k, where k is the number of dependent variables. (The number of different types of censorship is $k + 1$.) Thus the analysis required to implement the method does become more complex as the number of dependent variables increases. Furthermore, each added variable increases by one the highest dimension of the integrals of the multivariate normal distribution that must be computed. Since the computational complexity of these integrals can be expected to rise exponentially with dimension, it can be conjectured that the running times of multivariate censored regressions will rise exponentially with the number of dependent variables. There is thus a definite practical limit to the number of dependent variables that can be analyzed. The implication for modeling of airline demands is that, subject to experimental verification, it may be feasible to consider dealing with as many as eight fare classes in a single model, but not many more than this.

5.4 Numerical Example

This section describes a computer implementation of the EM algorithm for censored bivariate regressions and summarizes the results of a series of test runs on simulated airline demand data.

5.4.1 The Bivariate Censored Regression Program

The algorithm described in section 5.2.1 was implemented as a program in the FORTRAN77 computer language. The main operations of the program are:

1. **INPUT:** Read in the data and initial values for the standard deviations and correlation σ, τ and ρ (supplied by user).

2. **INITIALIZE:** Perform univariate regressions on all data points that were not censored in the X-direction (regions $\mathcal{A}$ and $\mathcal{B}$ in Figure 5.1), to establish an initial value for $\boldsymbol{\alpha}$. Similarly compute an initial estimate for $\boldsymbol{\beta}$ on the basis of data falling in regions $\mathcal{A}$ and $\mathcal{C}$.

3. **EXPECTATION:** Use the current estimates of σ, τ and ρ to compute the conditional expected values, variances and covariance of all censored observations with the methods provided in section 5.3.2.

4. **REGRESS:** Perform a bivariate regression on the uncensored (X, Y) data along with the expected values of the censored (X, Y) data obtained in step 3. This produces new estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as well as the sum of squares and cross-products (SSCP) matrix based on expected values of the censored observations.

5. **ESTIMATE MOMENTS:** Combine the SSCP matrix obtained in step 4 with the conditional variances and covariance obtained in step 3 to obtain new estimates for σ, τ and ρ using equations (5.16) through (5.18) and equations (5.20) through (5.22).

6. **TEST:** Compare the new and old parameter estimates. If one or more have changed by more than a pre-set tolerance, then return to the EXPECTATION step; otherwise, stop.

A number of the basic statistical functions were accomplished through calls to subroutines and functions in the Integrated Mathematical Subroutine Library (IMSL) supplied by IMSL, Inc.. Among these were the routines 1) RGIVN for multivariate regression, 2) DMILLR for Mill's ratio (the univariate hazard rate), 3) DBNRDF for the bivariate normal distribution function, and 4) DNORDF for the univariate normal distributon function. The program, including comments and reporting routines, is approximately 800 lines long.

No extra efforts were made to optimize the performance of this program, the main objective being to test the viability of the EM method for this application. Despite this, the performance was generally quite good, as will be seen in the next section. One obvious area for possible improvement is in finding initial values for the regression parameters (INITIALIZE step). Here, it might be better to perform univariate *censored* regressions based on more of the data so as to obtain initial values that are likely to be closer to the MLE's. Note however, that estimates at least as good as these will be available after one execution of the REGRESS step in any case.

5.4.2 Test Runs

This section summarizes the results of a series of test runs done on simulated airline demand data.[9] The objectives of these runs were 1) to determine whether the EM method was viable for this application, 2) examine the effect of sample size on the efficiency of the method, 3) examine the effect of the degree of censorship on both accuracy and efficiency, and 4) determine the effects of the underlying correlation on accuracy and efficiency. These runs focussed on the following performance measures: 1) accuracy of the estimates, 2) number of iterations (henceforth, *steps*) required, and 3) cpu running time.

An initial series of trials was conducted to establish sensible ranges for sample size and degreees of censorship for the test runs. General conclusions from these trials were:

---

[9] In the present case, the lack of actual airline data is not a serious drawback. Regression analysis is already an accepted adjunct to the forecasting process in modern airlines. If it can be shown that the present method produces accurate regression estimates in reasonable time on problems of realistic size, then the method will be no less relevant than standard regression analysis. In fact, it will be probably be a great deal more relevant since it handles the censorship problem encountered in most airline demand data.

1. For small models (under six independent variables), the running time of the program is relatively insensitive to the number of independent variables in the model. This was expected since the number of independent variables effects mainly the time for each execution of the REGRESS step in the program. Since this step involves a routine bivariate regression with highly optimized code (the IMSL subroutine) it was anticipated that it would execute quickly relative to other parts of the program for any reasonable number of variables.

2. Running time is quite sensitive to the sample size. This also was expected since each censored sample point requires calculation of conditional moments which in turn require either univariate or bivariate integrals of normal distributions.

3. For sample sizes of 100, the program performed well on data in which roughly 30% of the observations were censored in one or both directions. Occasional problems with slow convergence occurred when censoring exceeded 70%. These two values were selected as representative of 'low' and 'high' censorship. (Note that these values are rough indicators only — 70% censoring of a sample of size 50 is far more severe from the point of view of accuracy than the same percentage in a sample of size 500.)

4. With two independent variables, performance is unreliable with sample sizes of 50 or less when censoring is high. If we take as a rough estimate that one degree of freedom is lost for each censored observation (a 'rule of thumb' often used in censored data situations), 70% censorship implies loss of all but 12 degrees of freedom; i.e., $50 - 35$(censored) $- 3$(vector parameters). This is actually an overly conservative estimate in the bivariate regression case since in a portion of the censored observations one or other variable is not censored. Also, if ρ is nonzero, an observation of one variable conveys information about the other.

5. A convergence tolerance of 0.1% is adequate to ensure convergence of the program. That is, in all test runs the program exhibited stable behaviour once the maximum change in any parameter estimate from one step to the next was 0.1% or less.

On the basis of conclusion 1, above, it was decided to carry out the remaining runs with a small model (two independent variables) and to focus on other factors that might influence performance. Sample sizes of 50, 100, and 500 were selected as adequate for purposes of measuring the impact of sample size. Correlations of 0 and 0.8 were selected for examining the effect of correlation. The 0.1% convergence tolerance was used for all subsequent trials. This is a fairly coarse criterion since it only guarantees stability to the third significant figure in all estimates, but it minimized the time and expense of the trials. (In practice, most estimates exhibited stability to the fourth significant figure.)

The following model was used in all test runs:

$$\begin{aligned} X &= 100 - 100w_1 + 10w_2 + \delta \\ Y &= 1 + 100w_1 + 1w_2 + \epsilon; \end{aligned}$$

where δ and ϵ were sampled from a bivariate normal distribution with mean vector $\mathbf{o}$, and standard deviations 60 and 20 respectively. The correlation was set to 0 in half of the runs, and to 0.80 in the other half. Values for variables w_1 and w_2 were sampled from uniform distributions on the intervals [0.2,0.6] and [4.0,14.0] respectively. The parameters in the model and in the uniform distributions could have been chosen arbitrarily — any reasonable sets of parameters would have served for testing the algorithm — however, these particular values were chosen so that the simulated values of X and Y would be representative of airline demand data. (The overall mean values of X and Y based on this model are 150 and 50 repectively, and there is a low probability of negative values of either variable.)

Given a set of randomly generated values of X and Y from this model, a plane capacity C and an X-demand booking limit ℓ, censorship of the data was accomplished in the manner described in equations (5.11) and (5.12). To achieve a specified degree of censorship it was necessary to do a few trial-and-error runs with different values for C and ℓ. It was found that low (30%) censoring could be achieved with $C = 245$ and $\ell = 200$, and high (70%) censoring with $C = 145$ and $\ell = 116$. These parameters were kept constant in all test runs. Consequently, the 30% and 70% censorship figures are *nominal* — different random data sets produced different degrees of censorship. For example, in 40 repeated runs on data with correlation 0 and sample size 50, the average 'low' and 'high' degrees of censorship were 27% and 79%, respectively.

A typical test run was conducted as follows:

1. n pairs of values of X and Y were generated from the model above. The uniform variates were generated with the IMSL routine RNUNF, and the bivariate normal variates with IMSL CHFAC (Cholesky factorization of the desired correlation matrix) in combination with RNMVN (bivariate normal random variates).
2. A bivariate regression was run on the data to determine the LSE's for the parameters with 0% censorship.
3. Approximately 30% of the data was censored as described above and the censored regression program was run on it.
4. Step 3 was repeated with 70% censorship.
5. Steps 1 through 4 were repeated if required.

Results of the test runs are summarized in tables 5.1, 5.2 and 5.3. These tables will be described briefly below, and comments and conclusions based on the test runs will follow.

Table 5.1 gives the results of single runs that typify the estimates obtained for each sample size, correlation and degree of censorship. Its purpose is to illustrate the kind of accuracy that can be expected under the different conditions. For purposes of assessing accuracy, the appropriate reference values are the parameter estimates with 0% censoring rather than the base model parameter values. To improve readability, the number of significant figures displayed in the tables is between one and three. Recall that the convergence criterion used in the test runs guaranteed stability to the third significant figure.

Table 5.2 gives typical running times for each of the test conditions. These, of course, are highly dependent on the particular computer used, the coarseness of the convergence criterion, and the degree of optimization of the censored regression program. This set of running times was obtained on a VAX 11/785 minicomputer running at approximately 1.5 MIPS (million instructions per second) under the VMS operating system at the University of Denver.

Finally, Table 5.3 presents the means and standard errors of parameter estimates for forty replications of the sample size 50 case. Its purpose is to indicate the variability that can be experienced when the sample size is small.

Conclusions from Test Runs

The test runs described above were limited in scope; however, some general conclusions can be drawn relating to the original objectives of the numerical trials.

1. The EM method is an effective and practical technique for censored bivariate regression as long as the sample size is reasonable and the censorship is not too severe. In the full series of test runs, the method performed well in many cases in which the censorship was over 80%. (Average high censorship was 79% in the forty

Table 5.1: Typical Parameter Estimates for Selected Sample Sizes, Correlations and Degrees of Censorship

					base model parameters[a]							
					60	20	100	-100	10	1	100	1
n	ρ	%cen	steps	$\hat{\rho}$	$\hat{\sigma}$	$\hat{\tau}$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
50	0	0%		-0.03	62	19	143	-166	9	-10	76	3
		low	10	-0.06	61	19	134	-142	8	-11	79	3
		high	40	-0.11	52	23	128	-66	4	-12	77	4
50	0.8	0%		0.75	60	18	114	-101	9	13	73	1
		low	11	0.79	59	19	119	-118	9	15	73	1
		high	41	0.54	56	14	63	-53	12	10	53	1
100	0	0%		0.10	56	16	85	-66	10	-7	103	2
		low	7	0.10	51	17	81	-57	10	-6	105	2
		high	64	0.01	47	12	66	-18	9	2	72	2
100	0.8	0%		0.80	65	19	94	-92	10	1	106	1
		low	14	0.78	61	17	98	-85	9	2	107	0
		high	52	0.68	62	15	119	-103	8	3	102	0
500	0	0%		-0.06	59	20	75	-58	11	2	89	1
		low	10	-0.08	59	20	69	-48	11	2	89	1
		high	37	-0.09	55	19	74	-55	11	6	83	1
500	0.8	0%		0.81	62	20	115	-132	10	9	83	1
		low	14	0.80	63	21	114	-136	10	10	79	1
		high	48	0.78	59	20	122	-132	8	11	69	1

---

[a] Base Model: $\mathrm{E}[X] = \alpha_0 + \alpha_1 w_1 + \alpha_2 w_2$, $\mathrm{E}[Y] = \beta_0 + \beta_1 w_1 + \beta_2 w_2$

Table 5.2: Typical Computer Run Times

Table entries: cpu time in seconds

	$\rho = 0$		$\rho = 0.8$	
sample size	low%	high %	low%	high %
50[a]	2	14	3	17
100	6	67	11	52
500	34	188	50	243

---

[a] times for sample size 50 are averages from 40 runs.

Table 5.3: Average Results for 40 Runs with Sample Size 50
Table entries: mean
(standard error^a)

				base model parameters^b							
				60	20	100	-100	10	1	100	1
ρ	%cen	steps	$\hat{\rho}$	$\hat{\sigma}$	$\hat{\tau}$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
0	0%		0.03 (0.03)	60 (1)	19 (0.3)	102 (6.5)	-106 (11)	10 (0.4)	2 (2)	98 (4)	1 (0.2)
	low	10 (0.3)	0.03 (0.03)	60 (1)	20 (0.4)	102 (7.1)	-102 (12)	10 (0.4	1 (3)	100 (5)	1 (0.2)
	high	59 (4)	0.05 (0.04)	68 (3)	21 (0.5)	105 (9.9)	-130 (17)	12 (0.8)	-0.1 (3)	101 (6)	1 (0.3)
0.8	0%		0.79 (0.01)	59 (1)	20 (0.3)	104 (7)	-107 (15)	10 (0.4)	2 (2)	100 (5)	0.9 (0.2)
	low	14 (1)	0.79 (0.01)	59 (1)	20 (0.4)	103 (7)	-111 (15)	10 (0.5)	2 (2)	98 (5)	0.9 (0.2)
	high	71 (8)	0.79 (0.02)	64 (2)	20 (0.8)	109 (11)	-123 (21)	11 (0.8)	4 (3)	96 (7)	1 (0.3)

---

[a] standard error: $S/\sqrt{40}$, where S is the sample standard deviation of the 40 observations.

[b] Base Model: $\mathrm{E}[X] = \alpha_0 + \alpha_1 w_1 + \alpha_2 w_2$, $\mathrm{E}[Y] = \beta_0 + \beta_1 w_1 + \beta_2 w_2$

replications at sample size 50.)

2. The effect of sample size on the performance of the algorithm was as expected. Samples of size 50 produced usable but highly variable estimates — virtually all estimates had the correct order of magnitude and the correct sign, and samples of 100 or 500 produced generally good estimates with low variability. A rough guide based on these trials is that usable estimates will be obtained if the number of uncensored observations is at least two or three times the number of parameters being estimated. Running times increased roughly linearly with sample size, again as expected, since the most time-consuming computation was in calculating the expected value of a censored observation, and this had to be done for each censored observation.

3. Increasing the degree of censorship from approximately 30% to approximately 70% increases the standard error of the estimated regression coefficients by as much as 1.5 to 2 times (see Table 5.3), increases the number of iterations required by roughly five times, and increases running time by 5 to 10 times.

4. The size of the correlation (0 or 0.8) of the error terms in the underlying model had no obvious effects on the performance of the algorithm. It had been expected that higher correlation would lead to greater accuracy and efficiency because of the greater 'information content' of the partially censored observations. A tentative explanation for the absence of this effect is that it was counterbalanced by the tendency for more observations to fall in the doubly censored region when the correlation was high.

5.5 Inferences and Regression Diagnostics

This section deals with ways in which inferences can be made from the estimated regression coefficients and discusses the detection of departures from the assumptions of the regression model.

Inferences on the Regression Coefficients

One drawback of the EM method when compared with direct maximization of the likelihood function is that an estimate of the covariance matrix for the regression coefficients is not readily available on each iteration. In the case of direct maximization, such an estimate can be obtained at any time by evaluating the negative inverse of the Hessian matrix of the log likelihood function at the current estimates of the regression parameters.[10] This assumes that the particular direct maximization method being used is already calculating and using the second derivatives in some way; e.g. Newton or Quasi-Newton methods.

With the censored regression method described here, the Hessian matrix can be calculated to an arbitrary degree of accuracy by using finite second differences of the log likelihood function. (In fact, this is the way most direct maximization methods work in the absence of closed form expressions for the second derivatives.) Thus an estimate of the covariance matrix of the coefficients can be obtained through a side calculation whenever required. It is not easy to determine the exact sampling distribution of the regression coefficient estimators because, as noted previously, the censored regression procedure can, in theory, converge to a saddle point or cycle among several stationary points. In practice, however, with reasonable sample sizes and degree of censorship the

---

[10] For readers unfamiliar with this connection between second order derivatives of the log likelihood and the variance of the parameter estimates, consider the estimation of a single parameter with the ML estimation method. In this case, the Hessian is just the second derivative of the likelihood function, and this will be increasingly negative the 'sharper' the curvature of the likelihood function at the current parameter estimate. In the general case, the negative inverse Hessian is an asymptotically unbiased estimator of the true covariance matrix.

method will converge to the MLE, and the sampling distribution of the estimators will be approximately normal. The availability of the covariance matrix and the knowledge of the approximate sampling distribution can then permit tests of the significance of the regression coefficients and other inferential procedures to be carried out.

Regression Diagnostics

Like any regression method, the censored regression described here is vulnerable to such problems as mis-specification of the model, non-constant error variances, correlated errors, multicollinearity, and so on. With uncensored regressions, methods such as analysis of residual plots, weighted least squares and ridge regression methods are available for detecting and dealing with these problems when they arise. In the censored regression case, a simple approach is to apply the same methods to the estimated uncensored data obtained in the EXPECTATION step of the algorithm. The fact that this estimated uncensored data is obtained through a normality assumption will tend to mask problems with the data, but the approach should work well with large sample sizes and/or moderate censorship. Also, because of the masking, if a diagnostic procedure indicates a significant problem with the estimated data, this can be taken as strong evidence of a problem with the underlying data or model.

5.6 Summary — Censored Regression Analysis

This chapter has developed a method for carrying out bivariate regression analysis on data that has been censored in the manner of airline demands. It has been shown that the EM method can be adapted to this problem and produces an algorithm that is both effective and efficient when consideration is given to the complexity of the estimation problem. Specific conclusions are:

1. Usable estimates of nine parameters can be obtained on the basis of as few as 50 observations when 80% or more of the observations have been partially or totally censored.

2. In particular, it is feasible to test for correlation between two demand classes (the problem that originally motivated this work) with data that has been censored by booking limits and collected under widely varying circumstances (e.g. different routes, seasons, days of week), as long as a reasonable model can be formulated for the effects of the various factors.

3. With a sample of size 500, a run of the algorithm on highly censored data may require as much as 4 cpu minutes on a minicomputer. However, with more highly optimized code, it is expected that this figure could be reduced by as much as a factor of ten. It is entirely reasonable to expect that the method could be run efficiently on a modern microcomputer.

4. The number of independent variables that can be included in the analysis is limited more by the capability of the underlying (uncensored) regression subroutine than by the censored regression algorithm. For a fixed sample size, the effect of the censored data analysis on the running time is simply to multiply the time to execute one uncensored regression on the data by the number of iterations required in the EM algorithm. Thus, if a particular model can be run with uncensored data, it should generally be possible to run it with censored data.

5. It is feasible to develop an algorithm for more than two dependent variables (fare classes), but the complexity of the calculations and the running times can be expected to rise exponentially with the number of classes.

Chapter 6

Summary and Conclusions

The objectives of this chapter are to draw together the principal results of this thesis and to discuss areas for future work. Section 6.1 briefly summarizes the contents of each of the preceding chapters and highlights the connections among them. The last section discusses directions for future analytical work in airline seat inventory control and concludes the thesis.

6.1 Principal Results

The general approach taken in this thesis was to examine restricted versions of problems of airline seat allocation with the objective of obtaining insights into the nature of optimal allocation rules. The emphasis was on finding concise, rigorously derived optimality conditions and structural solutions rather than complex mathematical programming formulations.

The problem examined in chapter 2 was that of determining optimal seat allocations among multiple, nested fare classes when the demands for those classes were stochastically independent. Optimality conditions were derived, and these were compared numerically with the EMSRa approximations of Belobaba [13, 1987]. In chapter 3, the same problem was studied for two fare classes when the independence assumption was removed. A general model for dependent demands was developed, and it was shown that a condition similar to Littlewood's simple seat allotment rule [89, (1972)] was optimal as long as the demands for the two fare classes were monotonically associated. Optimality conditions were also derived that incorporated passenger goodwill and passenger upgrades. The interpretation and use of full fare passenger spill rates in the seat allocation context were discussed. The chapter closed with an examination of conditions under which the monotonic association condition holds and of the effect of increasing correlation on the protection level for full fare seats.

Both the independent demand (Chapter 2) and dependent demand (Chapter 3) optimality conditions can be implemented if adequate estimates of demand probability distributions are available. If the demand distributions are relatively stable over the booking period prior to a flight, the fixed optimal protection levels determined by the optimality conditions cannot be improved upon by any *ad hoc* adjustment of levels based on observed demands. In the more common case that demand distributions are revised as the time of flight departure approaches, a simple approximate implementation scheme is to recalculate protection levels after each revision of demand forecasts. Note that in the case of the dependent demand model, the shift in the conditional high fare demand distribution that can be predicted on the basis of observed low fare demands *has already been accounted for in the optimality condition.* Any revision of demand distributions should be based on other, external forecasting procedures that indicate a shift in the parameters of the joint discount fare /full fare demand distribution.

A simple application of the optimality conditions that requires no estimation of demand distributions is that of monitoring historical maximal flight spill rates. Severe departures of historical spill rates from those specified by the optimality conditions will signal suboptimal allocation decisions.

Chapter 4 presented an application of the dependent demand model to the problem of determining an optimal overbooking level in a single fare class. It was shown that when passenger confirmations occur according to a Bernoulli process, the optimality rule was similar in structure to Littlewood's rule. Simple approximation formulas for the optimality rule were derived, and it was demonstrated that when the overbooking penalty and fare are equal, the simple ratio of capacity to confirmation probability yields a good approximation to the optimal overbooking level.

The analysis of seat allocation between dependent demands in chapter 3 assumed knowledge of the joint probability distribution of the demands. Chapter 5 dealt with the problem of estimating that distribution on the basis of historical demand data that was subject to the influence of external factors and censored by the presence of booking limits. It was shown that a bivariate multiple regression model could account for the influence of the external factors and that EM method could deal with the censoring of the data. The results of numerical trials of the method demonstrated both its accuracy and its practicality on data sets in which 80% or more of the observations were censored.

6.2 Directions for Future Work

As discussed in chapter 1, the seat inventory control problem faced by modern airlines is substantially more complex than any of the subproblems examined in this thesis. The dynamics of the reservations and cancellation process, the interactions of different flight legs, the consideration of passenger itineraries, and the need to consider overbookings in multiple fare classes are all elements that have not been addressed here. A general prescription for future work, then, is to incorporate any or all of these factors into analyses of the seat allocation and overbooking problems. However, as mentioned earlier, such comprehensive treatments tend to lead to mathematical programs that obscure the nature and qualitative behaviour of optimal solutions. Also, the dimensionality of the fully specified seat allocation problem is so great that it is unlikely that any practical mathematical programming approach will be able to encompass all of the complexities. It appears that the hope for progress in the area is in linking together separate programs and heuristics in such a way that seat management decisions can be enhanced, if not optimized. It is in this linking that an understanding of the nature of optimal solutions in isolated subproblems can play a role. With this in mind, then, the discussion of future directions given below will focus on areas that are in the same vein as this thesis; that is, on areas which may be amenable to analytical rather than computational treatment and which may yield simple structural rules or approximations.

6.2.1 Multiclass Allocation when Demands are Independent

A drawback of the analysis of the multiple independent fare class allocation problem given here is the restriction to a single flight leg. Such leg-based planning fails to account for the difference in expected revenues between a passenger booking a single leg and one booking several different legs as part of an itinerary. One approximate way of accounting for this factor, discussed in Belobaba [13, (1987)], is to attribute the whole fare to each leg of a passenger's itinerary, in effect introducing a new fare class for each passenger itinerary. This method, however, has the effect of exaggerating the revenue impact of an itinerary and could lead to an excessive rate of refusal of single leg bookings. An important area for further reasearch, then, is in generalizing the multiple allocation problem to more than a single flight leg.

The numerical comparison of the optimal solution versus EMSRa approximation given in chapter 2 encompassed three fare classes and a fixed set of demand distributions. Further numerical trials are required to investigate the relative performance of the methods with more fare classes and with varying demand distributions. It was demonstrated that the EMSRa method would either underestimate or overestimate optimal protection levels depending on the parameters of demand distribution in the case that demands followed an exponential distribution. It is an open question whether or not the same behaviour can occur with more realistic, normally distributed demands.

6.2.2 Allocation Between Two Classes when Demands are Dependent

An obvious direction for future work on the dependent demand problem is in extending the analysis to three or more fare classes. This work was not undertaken here partly because of time constraints and partly because the problem presented no new conceptual difficulties — the characterization of the problem as an optimal stopping problem remains valid, and optimality conditions similar to that obtained for the two-fare case can be expected. The one difficulty in this generalization would lie in extending the monotonic association condition to more fare classes and relating the new condition to easily interpreted properties of realistic demand distributions.

It was shown in chapter 3 that positive correlation of normally distributed demands was a sufficient condition for monotonic association. The case that demands are negatively correlated was not dealt with explicitly. It seems reasonable to conjecture that full fare protection levels will drop with increasingly negative correlation, and that there will exist a value of the correlation below which no seats should be protected, but this remains an area for further work.

The use of observed flight spill rates[1] from past flight data as a method of monitoring seat allocation performance was proposed in chapters 2 and 3. The appeal of this approach was that no estimation of demand distributions was required. An interesting question is whether or not such an approach could be integrated with forecasting techniques in an adaptive-control framework for seat management.

---

[1] Recall that the only information required to estimate the flight spill rate for a fare class is the proportion of flights on which that fare class reached its booking limit. It is not necessary to know the number of rejected reservations requests.

6.2.3 A Simple Overbooking Model

An immediate question prompted by the analysis of the single fare class overbooking problem is whether or not similar simple optimality conditions exist for multiple fare classes. An analysis of this problem should allow for the different penalties associated with refusing boarding to passengers in different fare classes and/or for the options airlines have for resolving oversold flight situations when they arise. Preliminary work with two fare classes has indicated that there may exist relatively simple conditions that provide upper and lower bounds on optimal overbooking levels; however, much work remains to be done on this problem.

It should be noted that the multiple fare class overbooking problem actually subsumes the multiple fare class allocation problem. There is no need in principle to first determine seat allocations and then determine overbooking limits based on those allocations. (This is the way the problem tends to be viewed in the industry for both practical and historical reasons.)

6.2.4 Estimation of Dependent Demands from Jointly Censored Data

It was pointed out in the conclusion of chapter 5 that the censored regression method can be extended to more than two dependent variables. This is a relatively routine exercise given the expressions for the moments of truncated multinormal distributions provided in the appendix and the general approach of chapter 5. The calculations will, however, be complex when more than a few fare classes are involved because of the multiplicity of the censorship regions. There is room for further analytical work to systematize these calculations; for example, to permit easy evaluation of such quantities as $\mathrm{E}[X_3 \mid X_1 = x_1, X_2 > x_2, X_3 > x_3]$.

The numerical trials of the censored regression algorithm demonstrated that it was a viable method for accomodating censorship in airline demand data. The method thus substantially expands the usefulness of regression analysis in estimating the parameters of airline demand distributions. From the standpoint of assessing the usefulness of the censored regression method, this is sufficient, since regression analysis is an accepted technique for airline demand modeling. Nonetheless, an application of the method to actual airline data is of considerable interest since there is the possibility of answering an open empirical question regarding the degree of correlation between demands for full and discount fare classes.

6.2.5 Dynamic Modeling

A final area in which analytical work might be fruitful and which applies to all of the above problems is in incorporating the dynamics of the reservations process. Given the simple fixed protection level policies that are optimal when demand distributions are assumed to be stable, there is reason to be optimistic that some form of dynamic control-limit policies will be optimal in the more general dynamic case. In fact, results along these lines have already been obtained in the hotel/motel overbooking setting by Liberman and Yechiali [88, (1978)] and in a more general context by Gerchak, Parlar and Yee [56, (1985)]. There is thus hope that cumbersome direct dynamic programming methods might be replaced with more practical methods involving control limit policies conditional on current demand forecasts.

6.2.6 Conclusion

It can be seen from the foregoing discussion that there is more work to be done in the area of airline seat management than has been done to date. The author hopes that this thesis contributes in some measure to ongoing efforts in the area.

Bibliography

[1] American Airlines. *Yield Management Systems and Dinamo Demonstration.* Technical Report, American Airlines, Operations Research, February 1988.

[2] J. Alstrup and S. Boas. *Booking Strategies for Flights with Two Passenger Types (in Danish).* Master's thesis, IMSOR — Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, 1984.

[3] J. Alstrup, S. Boas, O.B.G. Madsen, and R.V.V. Vidal. Booking policy for flights with two types of passengers. *European Journal of Operational Research*, 27:274–288, 1986.

[4] Takeshi Amemiya. Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica*, 42:999–1012, 1974.

[5] Takeshi Amemiya. Tobit models: a survey. *Journal of Econometrics*, 24(1/2):999–1012, 1984.

[6] Sven-Eric Andersson. Fares and bookings. In *Proceedings 12th AGIFORS Symposium*, pages 130–137, American Airlines, New York, 1972.

[7] Sven-Eric Andersson. *No-Show Cost Function for a Mixed C/M Cabin.* unpublished report 1982-12-28, SAS Management Consultants, Stockholm, Sweden, 1982.

[8] Sven-Eric Andersson. *Theory for the Control Level System.* unpublished report 830609, SAS Management Consultants, Stockholm, Sweden, 1983.

[9] V. B. Ashok and C. P. Suresh. Optimal allocation of seats by fare. 1973. (Presentation to AGIFORS reservations study group, TWA Inc.).

[10] A. P. Basu. Bivariate failure rate. *Journal of the American Statistical Association*, 66:103–104, 1971.

[11] M. J. Beckmann. Decision and team problems in airline reservations. *Econometrica*, 26:134–145, 1958.

[12] M. J. Beckmann and F. Bobkowski. Airline demand: an analysis of some frequency distributions. *Naval Logistics Research Quarterly*, 5(43):43–51, 1958.

[13] Peter P. Belobaba. *Air Travel Demand and Airline Seat Inventory Management.* PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 1987. Report R87-7, Flight Transportation Laboratory.

[14] Peter P. Belobaba. Airline yield management: an overview of seat inventory control. *Transportation Science*, 21(2):63–73, 1987.

[15] Peter P. Belobaba. Application of a probabilistic decision model to airline seat inventory control. *Operations Research*, 37(2):183–197, 1989.

[16] Peter P. Belobaba. Developments in airline seat inventory management. 1987. Presentation at ORSA/TIMS Conference, New Orleans, May 1987.

[17] Peter P. Belobaba. Mathematical models in airline reservations control. unpublished presentation slides from TIMS/ORSA Conference, Denver, CO, October 1988.

[18] A. V. Bhatia and S. C. Parekh. Optimal allocation of seats by fare. 1973. Presentation to AGIFORS Reservations Study Group.

[19] H. Bierman Jr. and J. Thomas. Airline overbooking strategies and bumping procedures. *Public Policy*, 21:601–606, 1975.

[20] Z. W. Birnbaum and P. L. Meyer. On the effect of truncation in some or all of the coordinates of a multinormal population. *Jounal of the Indian Society for Agricultural Statistics*, 5:17–28, 1953.

[21] H. W. Block. *Monotone Hazard and Failure Rates for Absolutely Continuous Multivariate Distributions.* Research Report 73-20, University of Pittsburgh, Pittsburgh, PA, 1975.

[22] M. A. Brenner. The significance of airline passenger load factors. In G. W. James, editor, *Airline Economics*, Lexington Books, Lexington, Mass., 1982.

[23] E. C. Brindley and W. A. Thompson. Dependence and ageing aspects of multivariate survival. *Journal of the American Statistical Association*, 67:822–830, 1972.

[24] Jörn Buhr. Optimal sales limits for two-sector flights. In *Proceedings 22nd AGIFORS Symposium*, pages 291–304, 1982.

[25] Gregory S. Carpenter and Dominique M. Hanssens. *Market Expansion, Cannibalization and Optimal Product-Line Pricing.* research working paper 87-AV-13, Columbia Business School, Columbia University, New York, October 1987.

[26] Y. S. Chow and H. Robbins. A martingale systems theorem and applications. In *Proceedings 4th Berkeley Symposium Math. Statist. Prob.*, University of California Press, 1961.

[27] Y. S. Chow, Herbert Robbins, and David Siegmund. *Great Expectations: The Theory of Optimal Stopping.* Houghton Mifflin Company, Boston, 1971.

[28] J. M. Cigliano. Price and income elasticities for airline travel: the north atlantic market. *Business Economics*, 15(4):17–21, 1980.

[29] A. Clifford Cohen, Jr. Restriction and selection in multinormal distributions. *Annals of Mathematical Statistics*, 28:731–741, 1957.

[30] A. Clifford Cohen, Jr. Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association*, 50:885–893, 1955.

[31] A. Clifford Cohen, Jr. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, 1(3):217–237, 1959.

[32] A. Clifford Cohen, Jr. Tables for maximum likelihood estimates: singly truncated and singly censored samples. *Technometrics*, 3(4):535–541, 1961.

[33] H. Cramér. Orthogonal expansion derived from the normal distribution. In *Mathematical Methods of Statistics*, pages 221–231, Princeton University Press, 1946.

[34] C. De Ridder. *title not known.* Master's thesis, University of Delft, 1964. This thesis was referred to by Rothstein and Stone (1967) as containing many details and proofs relating to the model of Taylor (1962). De Ridder was an employee of KLM at the time of publication.

[35] C. Deetman. Booking levels. In *Proceedings 4th AGIFORS Symposium*, American Airlines, New York, 1964.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[37] C. Derman and J. Sacks. Replacement of periodically inspected equipment. *Naval Logistics Research Quarterly*, 7:597–607, 1960.

[38] Moshe Dror, Pierre Trudeau, and Shaul P. Ladany. Network models for seat allocation on flights. *Transportation Research B*, 22B(4):239–250, 1988.

[39] Ewalds D'Sylva. *An Analysis of Reservation System Stategies.* Presentation slides, Airline Analysis, Boeing Commercial Airplane Company, Seattle, WA, February 1984.

[40] Ewalds D'Sylva. *Analysis of Reservations Strategies for a Single Path with Two Fare Classes.* Presentation slides, Airline Analysis, Boeing Commercial Airplane Company, Seattle, WA, February 1984.

[41] Ewalds D'Sylva. *A Family of Optimal Seat Assignment Problems.* Presentation slides, Sales Technology and Strategy Analysis, Boeing Commercial Airplane Company, Seattle, WA, January 1983.

[42] Ewalds D'Sylva. *Finding the Optimal Revenue Configuration.* Presentation slides, Boeing Commercial Airplane Company, Seattle, WA, July 1982.

[43] Ewalds D'Sylva. *The Optimal Partitioning of an Airplane's Seating Capacity.* Presentation slides, Sales Technology, Boeing Commercial Airplane Company, Seattle, WA, February 1982.

[44] Ewalds D'Sylva. *Passenger Choice Emulator (PACEM).* Presentation slides, Airline Analysis, Boeing Commercial Airplane Company, Seattle, WA, April 1984.

[45] Ewalds D'Sylva. *A Pilot Study of Seat Inventory Management for a Flight Itinerary.* Presentation slides, Airline Analysis, Boeing Commercial Airplane Company, Seattle, WA, February 1983.

[46] Ewalds D'Sylva. *Strategies in Seat Inventory Management.* Presentation slides, Airline Analysis, Boeing Commercial Airplane Company, Seattle, WA, July 1983.

[47] D. D. Dyer. On moments estimation of the parameters of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society, C: Applied Statistics*, 22:287–291, 1973.

[48] Björn J. Elle. *The Size of Aircraft for a Fluctuating Transport Demand.* Technical Report SAAB TN 65, SAAB Aktiebelag, Linköping, Sweden. Cited by Lyle (1970). No date given.

[49] J. D. Esary and F. Proschan. Relationships among some concepts of bivariate dependence. *Annals of Mathematical Statistics*, 43(2):651–655, 1972.

[50] M. M. Etschmaier and M. Rothstein. Operations research in the management of the airlines. *Omega - International Journal of Management Science*, 2(2):160–175, 1974.

[51] L. M. Falkson. Airline overbooking: some comments. *Journal of Transport Economics and Policy*, 3:352–354, 1969.

[52] D. J. Finney. Cumulants of truncated multinormal distributions. *Journal of the Royal Statistical Society, Series B*, 24:535–536, 1962.

[53] Y. Fukuda. Optimal disposal policies. *Naval Logistics Research Quarterly*, 8:221–227, 1961.

[54] A. V. Gajjar and K. Subramaniam. On the sample correlation coefficient in the truncated bivariate normal population. *Communications in Statistics, B:Simulation and Computation*, 7:455–477, 1978.

[55] J. L. Gasco. Reservations and booking control. In *Proceedings 17th AGIFORS Symposium*, 1977.

[56] Yigal Gerchak, Mahmut Parlar, and Tony K. M. Yee. Optimal rationing policies and production quantities for products with several demand classes. *Canadian Jounal of Administrative Sciences*, 2(1):161–176, 1985.

[57] F. Glover, R. Glover, J. Lorenzo, and C. McMillan. The passenger mix problem in the scheduled airlines. *Interfaces*, 12(3):73–79, 1982.

[58] I. R. Goodman and Samuel Kotz. Hazard rates based on isoprobability contours. In *Statistical Distributions in Scientific Work, Vol. 5*, pages 289–308, Nato Adv. Sci. Inst. Series C: Math. Phys. Sci., Reidel, Dordrecht, Boston, MA, 1980.

[59] A. K. Gupta. Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika*, 39:260–273, 1952.

[60] A. K. Gupta and D. S. Tracy. Recurrence relations for the moments of truncated multinormal distribution. *Communications in Statistics, A: Theory and Methods*, 5(9):, 1976.

[61] M. A. Hamdan. The equivalence of tetrachoric and maximum likelihood estimates of ρ in 2×2 tables. *Biometrika*, 57:212–215, 1970.

[62] D. L. Harmer. *A Description of the Surplus Seats Analysis System.* Technical Report, The Consulting Division - Boeing Computer Services Inc., Seattle, Washington, 1976.

[63] R. Harris. A multivariate definition for increasing hazard rate distribution functions. *Annals of Mathematical Statistics*, 41:713–717, 1970.

[64] H. Leon Harter and Albert H. Moore. Iterative maximum-likelihood estimation of the parameters of normal populations from singly and doubly censored samples. *Biometrika*, 53(1/2):205–213, 1966.

[65] H. O. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194, 1958.

[66] J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492, 1976.

[67] M. Hersh and S. P. Ladany. Optimal seat allocation for flights with intermediate stops. *Computers and Operations Research*, 5(1):31–37, 1978.

[68] J. B. Jennings. Booking level management. In *Proceedings 21st AGIFORS Symposium*, pages 175–196, 1981.

[69] Sa Joao. *Reservations Forecasting in Airline Yield Management.* Master's thesis, Massachusetts Institute of Technology, Boston, MA, February 1987.

[70] N. L. Johnson and S. Kotz. *Continuous Univariate Distributions 1.* Houghton Mifflin, Boston, MA, 1970.

[71] N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions.* John Wiley and Sons, New York, 1972.

[72] N. L. Johnson and Samuel Kotz. A vector multivariate hazard rate. *Journal of Multivariate Analysis*, 5:53–66, 1975.

[73] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1982.

[74] A. Kaplan. Stock rationing. *Management Science*, 15:260–267, 1969.

[75] C. G. Khatri. Estimation of parameters of a truncated bivariate normal distribution. *Journal of the American Statistical Association*, 58:519–526, 1963.

[76] L. Kosten. Een mathematisch model voor een reserveringsprobleem. *Statistica Neerlandica*, 14:85–94, 1960.

[77] G. M. Kostyrsky. The evaluation of an airline's pricing policy. In *Proceedings 11th AGIFORS Symposium*, American Airlines, New York, 1971.

[78] D. G. H. Kraft, T. H. Oum, and M. W. Tretheway. Airline seat management. *Logistics and Transportation Review*, 22(2):115–130, 1986.

[79] S. P. Ladany. Bayesian dynamic operating rules for optimal hotel reservations. *Zeitschrift Opns. Res.*, 21:B165–B176, 1977.

[80] S. P. Ladany. Dynamic operating rules for motel reservations. *Decision Sciences*, 7:829–840, 1976.

[81] S. P. Ladany and D. N. Bedi. Dynamic booking rules for flights with intermediate stop. *OMEGA*, 5(6):721–730, 1977.

[82] S. P. Ladany and M. Hersh. Non-stop versus one-stop flights. *Transportation Research*, 11(3):155–159, 1977.

[83] J. F. Lawless. *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, New York, 1982.

[84] E. L. Lehmann. Ordered families of distributions. *Annals of Mathematical Statistics*, 26:399–419, 1955.

[85] E. L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153, 1966.

[86] C. O. Lennon. Market reactions to fare constraints and its effects on passenger revenue. In *Proceedings 12th AGIFORS Symposium*, American Airlines, New York, 1972.

[87] V. Liberman and U. Yechiali. Hotel overbooking problem - inventory system with stochastic cancellations. *Advances in Applied Probability*, 9(2):p220ff, 1977.

[88] V. Liberman and U. Yechiali. On the hotel overbooking problem — an inventory system with stochastic cancellations. *Management Science*, 24:1117–1126, 1978.

[89] K. Littlewood. Forecasting and control of passengers. In *Proceedings 12th AGIFORS Symposium*, pages 95–117, American Airlines, New York, 1972.

[90] E. H. Lloyd. Least squares estimation of location and scale parameters using order statistics. *Biometrika*, 39:88–95, 1952.

[91] Christopher Lyle. A statistical analysis of the variability in aircraft occupancy. In *Proceedings 10th AGIFORS Symposium*, American Airlines, New York, 1970.

[92] G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge, U.K., 1983.

[93] Albert W. Marshall. Some comments on the hazard gradient. *Stochastic Processes and their Applications*, 3:293–300, 1975.

[94] R. Martinez and M. Sanchez. Automatic booking level control. In *Proceedings 10th AGIFORS Symposium*, pages 1–20, American Airlines, New York, 1970.

[95] Michael Mayer. Seat allocation, or a simple model of seat allocation via sophisticated ones. In *Proceedings 16th AGIFORS Symposium*, pages 103–135, 1972.

[96] John I. McCool. *Censored Data*, pages 389–396. Volume 1 of *Encyclopedia of Statistical Sciences*, John Wiley and Sons, New York, 1982.

[97] G. D. Murray. Contribution to discussion of paper by A.P. Dempster, N.M. Laird and D.B. Rubin. *Journal of the Royal Statistical Society, Series B*, 39:27–28, 1977.

[98] K. V. Nagarajan. On an auction solution to the problem of airline overbooking. *Transportation Research*, 13A:111–114, 1979.

[99] S. Nahmias and W. S. Demmy. Operating characteristics of an inventory system with rationing. *Management Science*, 11:1236–1245, 1981.

[100] G. B. Nath. Estimation in truncated bivariate normal distributions. *Journal of the Royal Statistical Society, C: Applied Statistics*, 20:313–319, 1971.

[101] Wayne Nelson. *Applied Life Data Analysis.* John Wiley and Sons, New York, 1982.

[102] Wayne Nelson and Gerald J. Hahn. Linear estimation of a regression relationship from censored data part I - simple methods and their application. *Technometrics*, 14(2):247–269, 1972.

[103] Wayne Nelson and Josef Schmee. Inference for (log) normal life distributions from small singly censored samples and blues. *Technometrics*, 21(1):43–54, 1979.

[104] K. Pearson. Mathematical contribution to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London A*, 195:1–47, 1900.

[105] K. Pearson. On the probable error of a correlation coefficient as found from a fourfold table. *Biometrika*, 9:22–28, 1913.

[106] P. E. Pfeifer. The airline discount fare allocation problem. *Decision Sciences*, V20:149–157, 1989.

[107] P. Puri and H. Rubin. On a characterization of the family of distributions with constant multivariate failure rates. *Annals of Probability*, 2:738–740, 1974.

[108] H. Richter. The differential revenue method to determine optimal seat allotments by fare type. In *Proceedings 22nd AGIFORS Symposium*, pages 339–362, 1982.

[109] R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

[110] S. Rosenbaum. Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society, Series B*, 23:405–408, 1961.

[111] Sheldon M. Ross. Infinitesimal look-ahead stopping rules. *The Annals of Mathematical Statistics*, 42(1):297–303, 1971.

[112] M. Rothstein. Airline overbooking: fresh approaches are needed. *Transportation Science*, 9:169–173, 1975.

[113] M. Rothstein. An airline overbooking model. *Transportation Science*, 5:180–192, 1971.

[114] M. Rothstein. Airline overbooking: the state of the art. *Journal of Transport Economics and Policy*, 5:96–99, 1971.

[115] M. Rothstein. Hotel overbooking as a markovian sequential decision process. *Decision Sciences*, 5:389–404, 1974.

[116] M. Rothstein. O. R. and the airline overbooking problem. *Operations Research*, 33(2):392–435, 1985.

[117] M. Rothstein. *Stochastic Models for Airline Booking Policies.* PhD thesis, Graduate School of Engineering and Science, New York University, New York, N.Y., 1968.

[118] M Rothstein and A. W. Stone. Passenger booking levels. In *Proceedings 7th AGIFORS Symposium*, American Airlines, New York, 1967.

[119] K. M. Ruppenthal and R. Toh. Airline deregulation and the no-show overbooking problem. *Logistics and Transportation Review*, 19(2):111-121, 1985.

[120] A. E. Sarhan and B. G. Greenberg. Estimation of location and scale parameters by order statistics from singly and doubly censored samples, Part II. Tables for the normal distribution for samples of size $11 \leq n \leq 15$. *Annals of Mathematical Statistics*, 29:79–105, 1958.

[121] A. E. Sarhan and B. G. Greenberg. Estimation of location and scale parameters by order statistics from singly and doubly censored samples, Part I. The normal distribution up to samples of size 10. *Annals of Mathematical Statistics*, 27:427–451, 1956.

[122] J. Schmee and G. J. Hahn. A simple method for regression analysis with censored data. *Technometrics*, 21:417–432, 1979.

[123] R. Shlifer and Y. Vardi. An airline overbooking policy. *Transportation Science*, 9:101–114, 1975.

[124] J. Simon. Airline overbooking: the state of the art - a reply. *Journal of Transport Economics and Policy*, 6:254–256, 1972.

[125] J. Simon. An almost practical solution to airline overbooking. *Journal of Transport Economics and Policy*, 2:201–202, 1968.

[126] R. W. Simpson. Setting optimal booking levels for flight segments with multi-class, multi-market traffic. In *Proceedings 25th AGIFORS Symposium*, pages 263–279, 1985.

[127] R. W. Simpson. Theoretical concepts for capacity/yield management. In *Proceedings 25th AGIFORS Symposium*, pages 281–293, 1985.

[128] N. Singh. Estimation of parameters of a multivariate population from truncated and censored samples. *Journal of the Royal Statistical Society, Series B*, 22:307–311, 1960.

[129] D. Slepian. The one-sided barrier problem for gaussian noise. *Bell Systems Technical Journal*, 41:463–501, 1962.

[130] M. Tainiter. Some stochastic inventory models for rental situations. *Management Science*, 11(2):316–326, 1964.

[131] G. M. Tallis. The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18:342–353, 1962.

[132] G. M. Tallis. The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society, Series B*, 23:223–229, 1961.

[133] C. J. Taylor. The application of the negative binomial distribution to stock control problems. *Operational Research Quarterly*, 12(2):81–88, 1961.

[134] C. J. Taylor. The determination of passenger booking levels. In *Proceedings 2nd AGIFORS Symposium*, pages 93–116, American Airlines, New York, 1962.

[135] H. R. Thompson. Statistical problems in airline reservations control. *Operational Research Quarterly*, 12:167–185, 1961.

[136] Bernhard Titze and Raimund Griesshaber. Realistic passenger booking behaviours and the simple low-fare/high-fare seat allotment model. In *Proceedings 23rd AGIFORS Symposium*, pages 197–223, 1983.

[137] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

[138] Y. L. Tong. *Probability Inequalities in Multivariate Distributions.* Academic Press, New York, 1980.

[139] D. M. Topkis. Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes. *Management Science*, 15:160–176, 1968.

[140] R. Treseder. Yield management. In *Proceedings 22nd AGIFORS Symposium*, pages 239–241, 1982.

[141] M. W. Tretheway. Frequent flyer programs: marketing bonanza or anti-competitive tool? In *Proceedings, Canadian Transportation Research Forum,* **24**, pages 433–446, 1989.

[142] J. Vardi. *An Overbooking Policy.* Technical Report, Israel Institute of Technology., April 1973. Research thesis.

[143] W. Vickrey. Airline overbooking: some further solutions. *Journal of Transport Economics and Policy*, 6:257–270, 1972.

[144] Ken Wang. Modelling the interaction between payload restriction, passenger demand and reservation booking levels. In *Proceedings 22nd AGIFORS Symposium*, pages 323–338, 1982.

[145] Ken Wang. Optimum seat allocation for multi-leg flights with multiple fare types. In *Proceedings 23rd AGIFORS Symposium*, pages 225–246, 1983.

[146] H. Weiler. Means and standard deviations of a truncated bivariate normal distribution. *Austr. Journal of Statistics*, 1:73–81, 1959.

[147] W. D. Whisler. A stochastic inventory model for rented equipment. *Management Science*, 13(9):640–647, 1967.

[148] Richard D. Wollmer. *An Airline Reservation Model for Opening and Closing Fare Classes.* unpublished company report, Douglas Aircraft Company, McDonnell Douglas Corporation, Long Beach, CA, year unknown — 1986 or later.

[149] Richard D. Wollmer. *A Hub-Spoke Seat Management Model.* unpublished company report, Douglas Aircraft Company, McDonnell Douglas Corporation, Long Beach, CA, 1986.

[150] Richard D. Wollmer. *A Seat Management Model for a Single Leg Route.* unpublished company report, Douglas Aircraft Company, McDonnell Douglas Corporation, Long Beach, CA, 1986.

[151] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Appendix A

Some Properties of Normal Probability Distributions

The normal probability distribution provides a good approximation to airline demand distributions under many circumstances, and this fact is exploited throughout this thesis. This appendix presents a number of well-known basic properties of univariate and multivariate normal probability distributions that are used either explicitly or implicitly at many points. Also included are some less well-known properties of truncated normal and multinormal distributions that are needed for the calculations relating to spill rates and, more importantly, in the censored regression analysis of chapter 5. Most of these results have been reported elsewhere; however, it is believed that the connection between the first and second moments of the truncated multinormal distribution and the multivariate hazard gradient, given in section A.2.2, is reported here for the first time.

A.1 Properties of the Univariate Normal Distribution

In what follows, X denotes a normally distributed random variable with mean μ and variance σ^2; that is, the density and distribution functions for X are given by

$$\begin{aligned} \phi(x;\mu,\sigma^2) &= [\sqrt{2\pi}\sigma]^{-1} \exp\left[-\tfrac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ \text{and} &= \\ \Phi(x;\mu,\sigma^2) &= \int_{-\infty}^{x} \phi(u;\mu,\sigma^2)\,du. \end{aligned}$$

The standard normal density and distribution, $\phi(\cdot;0,1)$ and $\Phi(\cdot;0,1)$, are abbreviated simply $\phi(\cdot)$ and $\Phi(\cdot)$ respectively.

A.1.1 Simple Properties of the Univariate Normal Distribution

A list of well known basic properties of the univariate normal distribution is provided below. Standardized values of X and x will be denoted $Z_X = (X - \mu)/\sigma$ and $z_x = (x - \mu)/\sigma$ respectively.

Property 1 (standardization of the density)

$$\phi(x; \mu, \sigma^2) = (1/\sigma)\phi(z_x) \tag{A.1}$$

Property 2 (standardization of the distribution)

$$\Phi(x; \mu, \sigma^2) = \Phi(z_x) \tag{A.2}$$

Property 3 (symmetry of the density about the mean)

$$\phi(\mu - x; \mu, \sigma^2) = \phi(\mu + x; \mu, \sigma^2) \tag{A.3}$$

Property 4 (tail probabilities)

$$\Pr[Z_X > z_x] = 1 - \Phi(z_x) = \Phi(-z_x) \tag{A.4}$$

Property 5 (derivatives)

$$\frac{d}{dx}\phi(w(x)) = -w(x)\phi(w(x))\frac{d}{dx}w(x) \tag{A.5}$$

In particular,

$$\frac{d}{dx}\phi(x; \mu, \sigma^2) = -(1/\sigma)z_x\phi(z_x). \tag{A.6}$$

A.1.2 Mean and Variance for the Truncated Normal Distribution

Let Y have the distribution of a normal random variable truncated on the left at a and let $z_a = (a - \mu)/\sigma$. Then Y has the density

$$f(y) = \begin{cases} \frac{1}{\sigma}\frac{\phi(z_a)}{\Phi(-z_a)} & \text{for } x > a \\ 0 & \text{otherwise.} \end{cases} \tag{A.7}$$

Expressions for the mean and variance of Y are well known (see, for example, Johnson and Kotz [70, (1970)]). These moments will henceforth be referred to as the *truncated mean* and *truncated variance* respectively. The truncated mean is given by

$$\mathrm{E}[X|X > a] \;=\; \mu + \sigma\left[\frac{\phi(z_a)}{\Phi(-z_a)}\right] \qquad \text{(A.8)}$$

or

$$\mathrm{E}[X|X > a] \;=\; \mu + \sigma h(z_a), \qquad \text{(A.9)}$$

where $h(z_a)$ is the *hazard rate* of the distribution evaluated at the standardized point of truncation. The truncated variance is

$$V[X|X > a] = \sigma^2[1 + z_a h(z_a) - h^2(z_a)]. \qquad \text{(A.10)}$$

It is worth noting for later reference that the variance can be written more compactly as:

$$V[X|X > a] = \sigma^2[1 - \frac{dh}{dz}(z_a)], \qquad \text{(A.11)}$$

where $\frac{dh}{dz}(z_a)$ denotes the first derivative of the hazard rate evaluated at z_a.

A.2 Properties of the Multivariate Normal Distribution

The joint density of a multivariate normal random variable $\mathbf{X} = (X_1, \ldots, X_n)'$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and variance-covariance matrix $\boldsymbol{\Sigma}$ will be denoted $\phi(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = C \exp\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\},$$

where $C = [(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}]^{-1}$, and $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix with diagonal elements σ_i^2 and off-diagonal elememts $\rho_{ij}\sigma_i\sigma_j$. Let $\Phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the joint distribution function for $\mathbf{X}$; that is,

$$\Phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \phi(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) du_n \ldots du_1,$$

or

$$\Phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \int_{(-\infty,\mathbf{x}]} \phi(\mathbf{u};\boldsymbol{\mu},\boldsymbol{\Sigma})\, d\mathbf{u}.$$

where $(-\infty,\mathbf{x}] = (-\infty,x_1] \times (-\infty,x_2] \times \cdots \times (-\infty,x_n]$, and $d\mathbf{u} = du_1 \ldots du_n$.

In the event that each of the elements of $\mathbf{X}$ is a standardized random variable with zero mean and unit variance, then the variance-covariance matrix will be $\mathbf{R}$, a correlation matrix, and the notation $\phi(\cdot;\mathbf{R})$ and $\Phi(\cdot;\mathbf{R})$ will be used for the density and distribution respectively.

Let $\nabla_{\mathbf{X}}$ denote the gradient operator $(\partial/\partial x_1,\ldots,\partial/\partial x_n)'$, so that $\nabla_{\mathbf{X}} f(\mathbf{x})$ and $\nabla_{\mathbf{X}}\nabla'_{\mathbf{X}} f(\mathbf{x})$ are the gradient and Hessian matrix respectively for any real-valued function $f(\mathbf{x})$. The abbreviation Θ^2 will be used for $\Theta\Theta'$, where Θ is any vector or vector operator. For a vector-valued transformation $\mathbf{w}(\mathbf{x})$, $\nabla[\mathbf{w}(\mathbf{x})]'$ is the matrix of first partial derivatives of the elements of $\mathbf{w}$ with ijth element $\partial w_j/\partial x_i$. Where necessary, a subscript will be used to indicate a change in the variable of differentiation, as in $\nabla_{\mathbf{W}} f(\mathbf{w}(\mathbf{x}))$. The "chain rule" can thus be written

$$\nabla_{\mathbf{X}} f(\mathbf{w}(\mathbf{x})) = \nabla_{\mathbf{X}}[\mathbf{w}(\mathbf{x})]'\nabla_{\mathbf{W}} f(\mathbf{w}(\mathbf{x})),$$

and the "product rule" for the product of a real-valued function and a vector-valued function is written

$$\nabla_{\mathbf{X}}[f(\mathbf{x})\mathbf{w}(\mathbf{x})] = f(\mathbf{x})\nabla_{\mathbf{X}}[\mathbf{w}(\mathbf{x})] + \mathbf{w}(\mathbf{x})[\nabla_{\mathbf{X}} f(x)]'.$$

For a constant vector $\mathbf{a}$, $\nabla f(\mathbf{a})$ denotes the gradient of $f(\mathbf{x})$ evaluated at $\mathbf{a}$.

A.2.1 Simple Properties of the Multivariate Normal Distribution

Property 1 (standardization of the density)

$$\phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\mathbf{S}|^{-1}\phi(\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu});\mathbf{R}), \tag{A.12}$$

where $\mathbf{S} = \mathrm{diag}(\sigma_i)$, the diagonal matrix of standard deviations of the components of $\mathbf{X}$.

Property 2 (standardization of the distribution)

$$\Phi(\mathbf{x};\boldsymbol{\mu};\boldsymbol{\Sigma}) = \Phi(\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu});\mathbf{R}) \tag{A.13}$$

Property 3 (symmetry of the density)

$$\phi(\boldsymbol{\mu}-\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \phi(\boldsymbol{\mu}+\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{A.14}$$

Property 4 (tail probabilities) For $\mathbf{Z} = \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

$$\Pr[\mathbf{Z} > \mathbf{a}] = \int_{(\mathbf{a},\infty)} \phi(\mathbf{u};\mathbf{R})d\mathbf{u} = \Phi(-\mathbf{a};\mathbf{R}) \tag{A.15}$$

Property 5 (marginal distributions)

$$X_i \text{ has density } \phi(x_i;\mu_i,\sigma_i^2) \text{ for } i = 1,\ldots,n. \tag{A.16}$$

Property 6 (conditional distribution for a bivariate normal random vector) For a bivariate normal random vector $\mathbf{X} = (X_1, X_2)$, the conditional distribution of X_1 given $X_2 = x$ is normal with mean

$$\mathrm{E}[X_1|X_2 = x] = \mu_1 + (\rho\sigma_1/\sigma_2)(x-\mu_2) \tag{A.17}$$

and variance

$$\mathrm{Var}[X_1|X_2 = x] = (1-\rho^2)\sigma_1^2. \tag{A.18}$$

Property 7 (derivatives)

$$\nabla_{\mathbf{X}}\phi(\mathbf{w}(\mathbf{x});\mathbf{R}) = -\mathbf{R}^{-1}\mathbf{w}(\mathbf{x})\nabla\mathbf{w}(\mathbf{x})\phi(\mathbf{w}(\mathbf{x});\mathbf{R}) \tag{A.19}$$

A.2.2 First and Second Moments for the Truncated Multivariate Normal Distribution

Calculations of the moments of the multinormal distribution under various forms of truncation and/or censorship have been reported in many places [20, 47, 54, 60, 75, 100, 110, 128, 146]. These calculations have been done either by direct integration (suitable only for low-dimensional multinormal distributions because of the complexity), or with the use of rather complicated recurrence relations. It is shown below that it is not difficult to calculate the moment generating function (MGF) for the truncated multinormal distribution and, from that, to obtain a very compact expression for the moments of order one and two of truncated multinormals of arbitrary dimension. The MGF has been previously obtained by Tallis [132, (1961)], and the cumulant generating function ($\log(MGF)$), by Finney [52, (1962)]. The expressions obtained by Finney for the first and second cumulants are similar to those given here for the first and second moments. The connection to the multivariate hazard rate is identified here for the first time.[1]

The MGF of the Multinormal Distribution

Now suppose that $\mathbf{Z}$ has the distribution of a standardized multivariate random variable with each element truncated on the left by values in the vector $\mathbf{a} = (a_1, \ldots, a_n)'$. Denote by $\mathcal{M}(\mathbf{t})$ the joint MGF of $\mathbf{Z}$, where $\mathbf{t} = (t_1, \ldots, t_n)'$. That is,

$$\begin{aligned} \mathcal{M}(\mathbf{t}) &= \mathrm{E}[\exp(\mathbf{t}'\mathbf{Z})] \\ &= [\Phi(-\mathbf{a};\mathbf{R})]^{-1} \int_{(\mathbf{a},\infty)} C \exp(-\tfrac{1}{2}[\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - 2\mathbf{t}'\mathbf{Z}])\, d\mathbf{Z} \\ &= [\Phi(-\mathbf{a};\mathbf{R})]^{-1} \exp[\tfrac{1}{2}(\mathbf{t}'\mathbf{R}\mathbf{t})]\Phi(\mathbf{R}\mathbf{t} - \mathbf{a};\mathbf{R}). \end{aligned} \tag{A.20}$$

We proceed to find the gradient and Hessian matrix of the MGF given in (A.20)

---

[1] The idea of the multivariate hazard rate was first introduced in the 1970's, ten years after the articles by Tallis and Finney.

and use these to find expressions for the first and second moments of the truncated distribution. To improve readability, the abbreviation $\Phi(\cdot) = \Phi(\cdot; \mathbf{R})$ is used. We have

$$\nabla_t \mathcal{M}(\mathbf{t}) = \mathbf{R}\mathbf{t}\mathcal{M}(\mathbf{t}) + [\Phi(-\mathbf{a})]^{-1}[\,\mathbf{R}\exp(\tfrac{1}{2}\mathbf{t}'\mathbf{R}\mathbf{t})\nabla_{Rt-a}\Phi(\mathbf{R}\mathbf{t} - \mathbf{a})\,]$$

and

$$\begin{aligned}\nabla_t^2 \mathcal{M}(\mathbf{t}) &= \nabla_t[\nabla_t \mathcal{M}(\mathbf{t})]' \\ &= \mathcal{M}(\mathbf{t})\mathbf{R}' + \nabla_t \mathcal{M}(\mathbf{t})[\mathbf{R}\mathbf{t}]' \\ &\quad + [\Phi(-\mathbf{a})]^{-1}\Big\{\,\mathbf{R}\exp[\tfrac{1}{2}\mathbf{t}'\mathbf{R}\mathbf{t}]\nabla^2_{Rt-a}\Phi(\mathbf{R}\mathbf{t} - \mathbf{a})\mathbf{R} \\ &\qquad\qquad + \mathbf{R}\mathbf{t}\exp[\tfrac{1}{2}\mathbf{t}'\mathbf{R}\mathbf{t}][\nabla_{Rt-a}\Phi(\mathbf{R}\mathbf{t} - \mathbf{a})]'\mathbf{R}.\,\Big\}\end{aligned}$$

Then

$$\mathrm{E}[\mathbf{Z}|\mathbf{Z} > \mathbf{a}] = \nabla_t \mathcal{M}(\mathbf{0}) = \mathbf{R}\left[\frac{\nabla\Phi(-\mathbf{a})}{\Phi(-\mathbf{a})}\right], \tag{A.21}$$

and

$$\mathrm{E}[\mathbf{Z}\mathbf{Z}'|\mathbf{Z} > \mathbf{a}] = \nabla_t^2 \mathcal{M}(\mathbf{0}) = \mathbf{R} + \mathbf{R}\left[\frac{\nabla^2\Phi(-\mathbf{a})}{\Phi(-\mathbf{a})}\right]\mathbf{R}.$$

Finally,

$$\mathrm{Var}[\mathbf{Z}|\mathbf{Z} > \mathbf{a}] = \mathbf{R} + \mathbf{R}\left[\frac{\nabla^2\Phi(-\mathbf{a})}{\Phi(-\mathbf{a})} - \left(\frac{\nabla\Phi(-\mathbf{a})}{\Phi(-\mathbf{a})}\right)^2\right]\mathbf{R}. \tag{A.22}$$

By analogy with the univariate hazard rate occurring in (A.9), define the *multivariate hazard rate*[2]:

$$\mathbf{H}(\mathbf{Z}) = \frac{\nabla\Phi(-\mathbf{Z})}{\Phi(-\mathbf{Z})}. \tag{A.23}$$

---

[2] This vector quantity, sometimes called the hazard *gradient*, has been discussed by others in the context of multivariate generalizations of monotone hazard rate properties of distributions. See, for example, Block [21, (1973)], Harris [63, (1970)], Johnson and Kotz [72, (1975)] and Marshall [93, (1975)]. Its occurrence in the present context lends support to the argument that it is the appropriate generalization of the univariate hazard rate. (There have been some competing definitions; e.g. Basu [10, (1971)], Brindley and Thompson [23, (1972)], Goodman and Kotz [58, (1980)], Puri and Rubin [107, (1974)]).

Then (A.21) and (A.22) can be expressed succinctly as

$$\mathrm{E}[\mathbf{Z}|\mathbf{Z} > \mathbf{a}] = \mathbf{RH(a)} \tag{A.24}$$

and

$$\mathrm{Var}[\mathbf{Z}|\mathbf{Z} > \mathbf{a}] = \mathbf{R} - \mathbf{R}\nabla[\mathbf{H(a)}]'\mathbf{R}. \tag{A.25}$$

For a multivariate normal random vector $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, the corresponding results are:

$$\mathrm{E}[\mathbf{X}|\mathbf{X} > \mathbf{a}] = \boldsymbol{\mu} + \mathbf{SRH}(\mathbf{S}^{-1}(\mathbf{a} - \boldsymbol{\mu})) \tag{A.26}$$

and

$$\mathrm{Var}[\mathbf{X}|\mathbf{X} > \mathbf{a}] = \boldsymbol{\Sigma} - \mathbf{SR}\nabla[\mathbf{H}(\mathbf{S}^{-1}(\mathbf{a} - \boldsymbol{\mu}))]'\mathbf{RS}. \tag{A.27}$$

A.2.3 Moments of the Truncated Bivariate Normal Distribution

In this section equations (A.26) and (A.27) are used to obtain expansions for the first and second moments about the mean of a truncated bivariate normal random vector in terms of the *univariate* standard normal density and distribution, a bivariate normal density, and the upper right 'tail' of a bivariate normal distribution. The existence of these expansions greatly facilitates computation of the moments.

Let $\mathbf{Z} = (X, Y)'$ be a bivariate normal random vector with mean vector $\boldsymbol{\mu} = (\mu_x, \mu_y)$ and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$.

In what follows the univariate normal densities and distributions can be distinguished from their bivariate counterparts by the number of arguments. Let the first, second and cross partial derivatives of the bivariate distribution be indicated with subscripts. For example, $\Phi_x(x,y) = \frac{\partial}{\partial x}\Phi(x,y)$. Then:

$$\Phi_x(x,y) = \phi(x)\Phi(\frac{y - \rho x}{\sqrt{1-\rho^2}}), \tag{A.28}$$

$$\begin{aligned} \Phi_y(x,y) &= \phi(y)\Phi(\frac{x-\rho y}{\sqrt{1-\rho^2}}), && \text{(A.29)} \\ \Phi_{xx}(x,y) &= -(\rho\phi(x,y) + x\Phi_x(x,y)), && \text{(A.30)} \\ \Phi_{yy}(x,y) &= -(\rho\phi(x,y) + y\Phi_y(x,y)), && \text{(A.31)} \end{aligned}$$

and

$$\Phi_{xy}(x,y) = \phi(x,y). \qquad \text{(A.32)}$$

Let a and b be arbitrary real constants and let $z_a = (a-\mu_x)/\sigma_x$ and $z_b = (b-\mu_y)/\sigma_y$. Furthermore, for brevity, let $\phi_2 = \phi(-z_a, -z_b)$ and $\Phi_2 = \Phi(-z_a, -z_b)$.

From (A.26) we have

$$\mathrm{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix} \middle| \begin{pmatrix} X \\ Y \end{pmatrix} > \begin{pmatrix} a \\ b \end{pmatrix}\right] = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \frac{1}{\Phi_2}\begin{pmatrix} \sigma_x & \rho\sigma_x \\ \rho\sigma_y & \sigma_y \end{pmatrix}\begin{pmatrix} \Phi_x(-z_a,-z_b) \\ \Phi_y(-z_a,-z_b) \end{pmatrix}.$$

Then, using (A.28) and (A.29),

$$\begin{aligned} &\mathrm{E}[X|X>a, Y>b] = \\ &\quad \mu_x + \frac{\sigma_x}{\Phi_2}\left[\phi(z_a)\Phi(\frac{\rho z_a - z_b}{\sqrt{1-\rho^2}}) + \rho\phi(z_b)\Phi(\frac{\rho z_b - z_a}{\sqrt{1-\rho^2}})\right] \qquad \text{(A.33)} \end{aligned}$$

and

$$\begin{aligned} &\mathrm{E}[Y|X>a, Y>b] = \\ &\quad \mu_y + \frac{\sigma_y}{\Phi_2}\left[\rho\phi(z_a)\Phi(\frac{\rho z_a - z_b}{\sqrt{1-\rho^2}}) + \phi(z_b)\Phi(\frac{\rho z_b - z_a}{\sqrt{1-\rho^2}})\right]. \qquad \text{(A.34)} \end{aligned}$$

From (A.27) we have

$$\mathrm{Var}\left[\begin{pmatrix} X \\ Y \end{pmatrix} \middle| \begin{pmatrix} X \\ Y \end{pmatrix} > \begin{pmatrix} z_a \\ z_b \end{pmatrix}\right] = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

$$+\begin{pmatrix} \sigma_x & \rho\sigma_x \\ \rho\sigma_y & \sigma_y \end{pmatrix}\left[\frac{1}{\Phi_2}\begin{pmatrix} \Phi_{xx} & \phi_2 \\ \phi_2 & \Phi_{yy} \end{pmatrix} - \frac{1}{\Phi_2^2}\begin{pmatrix} \Phi_x^2 & \Phi_x\Phi_y \\ \Phi_x\Phi_y & \Phi_y^2 \end{pmatrix}\right]\begin{pmatrix} \sigma_x & \rho\sigma_y \\ \rho\sigma_x & \sigma_y \end{pmatrix}.$$

Upon applying equations (A.30),(A.31) and (A.32) and simplifying, the following expressions for the variance and covariance terms are obtained.

$$\mathrm{Var}[X|X>a,Y>b]=$$
$$\sigma_x^2\left[1+\frac{z_a\Phi_x+\rho^2 z_b\Phi_y}{\Phi_2}+\rho(1-\rho^2)\frac{\phi_2}{\Phi_2}-\left(\frac{\Phi_x+\rho\Phi_y}{\Phi_2}\right)^2\right], \tag{A.35}$$

$$\mathrm{Var}[Y|X>a,Y>b]=$$
$$\sigma_y^2\left[1+\frac{\rho^2 z_a\Phi_x+z_b\Phi_y}{\Phi_2}+\rho(1-\rho^2)\frac{\phi_2}{\Phi_2}-\left(\frac{\rho\Phi_x+\Phi_y}{\Phi_2}\right)^2\right], \tag{A.36}$$

and

$$\mathrm{Cov}[XY|X>a,Y>b]=$$
$$\sigma_x\sigma_y\left[\rho+\rho\frac{z_a\Phi_x+z_b\Phi_y}{\Phi_2}+(1-\rho^2)\frac{\phi_2}{\Phi_2}-\frac{(\Phi_x+\rho\Phi_y)(\rho\Phi_x+\Phi_y)}{\Phi_2^2}\right]. \tag{A.37}$$

A final expansion in terms of ϕ_2 and Φ_2 and the univariate $\phi(\cdot)$ and $\Phi(\cdot)$ can then be done using (A.28) and (A.29).