Final Report

West Lafayette Public Library

October 28, 2004

Submitted by STATCOM, Department of Statistics, Purdue University

Team Members:
Alexander Lipka, Joseph Nolan, Surya Tokdar

Introduction

In April 2004, the West Lafayette Public Library approached STATCOM with various problems. One of these problems was to assess the "popularity" of a book, given some information about that particular book; i.e., given its price, year of publication, category, and last time it was checked out, we wanted to be able to calculate the popularity of that book.

Description of the Data

In May 2004, the West Lafayette Public Library gave us a CD Rom containing details on all of their books, CDs, cassettes, magazines, DVDs, and videos. This information was given in the form of 30 Excel files; each file corresponding to a particular category. The details given for each item include title, author, item ID, price, year of publication, prefix/call number, the last time the book was checked out, and the total number of checkouts of that book. The data include 101,988 items.

Conditioning the Data

There were a couple of issues about the data that concerned us. The first issue was that some books are said to be published in the year 0, which seems to be either a typing error or a missing observation. Also, we felt that we should recalculate the price of each item in terms of 2004 dollars in order to have price accurately predict popularity.

In order to partially adjust for the first problem, the items that were said to be published in the year 0 were assigned to the mean year of publication of the remaining books in that category. As for the second above-mentioned issue, we did a Consumer Price Index (CPI) adjustment to the price to convert it into 2004 dollars.

There were four categories that did not have a lot of observations. To solve this problem, we merged similar categories together. Namely, we merged Children's Nonfiction DVD and Children's Nonfiction Video together. Also, in a separate category, we merged Children's Books on CD and Children's Books on Cassette.

Analysis of Data

We decided to construct a formula in which one would plug in price and category, and the formula would predict the popularity of the book in terms of the expected number of checkouts per year. We also took into account the number of years the books have been in the library, since we are more concerned about the rate of checkout than total number of checkouts. The year of publication was implicitly used to carry out this calculation, and thus was not separately used in the model. We felt that the information about item ID, prefix/call number, and the last time a book was checked out were irrelevant to the total number of checkouts of an item, and hence they were not used in the model. Although title and author are relevant, we are interested in assessing the popularity of a book per category, and thus did not use these two variables. In order to make our formula describe the data more accurately, we created four categories of price (less than $13.00, $13.00 to $20.00, $20.00 to $30.00, and above $30.00). We chose the cut off values for the categories in a way such that each category had approximately the same number of books.

We summarized the results of the formula in Table I. Each row signifies a category of book, and each column signifies one of the four price categories. The number inside each cell is the expected number of checkouts per year of a book which belongs to the corresponding price and category. For example, if we wanted to determine the expected number of checkouts for a fiction book that costs $17.00 in 2004, then the formula yields 1.5 as the average number of checkouts per year.

Table II shows the mean and standard deviation of the total number of checkouts and the price of the 10% most popular books within a category.

Discussions

.

- In general, our formula predicts that DVDs, Videos, and Books on CD tend to be more popular than the books.
- Based on the results of our formula, we have determined that the most popular category/price combination is DVDs in the price range of $14.00 to $20.00, with a predicted rate of 7.7 checkouts per year.
- The least popular category is Magazines. More specifically, magazines less than $13.00 are predicted to have an average of 0.1 checkouts per year, and magazines greater than $13.00 are predicted to have an average of 0 checkouts per year.
- One can explain the unpopularity of magazines by noticing that magazines are most frequently read within one month of their publication, and most patrons probably read the magazines in the library and subsequently do not check them out.
- Other unpopular categories include Children's Fiction, Nonfiction, Nonfiction Video, and Science Fiction.

- Thus, it appears that the popular things the library gives out are not books, and, with the exception of Nonfiction Video, the least popular things the library gives out are nonfiction in general.
- There are a few categories (for example, Children's Books on CD) in which the items that cost over $30.00 are more popular than those that cost less than $30.00. Therefore, one can conclude that it is "worth it" to buy more expensive Children's Books on CD. Similar inferences can be made for CD Roms and Children's CD Roms.
- We would, however, like to mention that a low number of checkouts do not always mean low popularity. For example, books placed high on a shelf or near the floor may not be as frequently checked out as a book that is placed on a shelf at eye level. A better promotion of the items (in terms of displays or advertisements) might improve their checkout rates.
- A more accurate model can be calculated if data on the dates of each time a book is checked out are provided.

Table I – Results of the Model.

Category	Price (in Dollars)			
	0-13	14-20	21-30	Above 30
Books on Cassette	1.2	1.4	1.2	1.2
Books on CD	3.5	3.7	4.2	2.6
CD Roms	0.8	1.5	1.2	2.8
Children's Award Winners	0.7	1.0	0.8	0.2
Children's Basic	0.7	0.8	1.3	0.5
Children's Books on CD	1.4	1.3	1.4	6.3
Children's CD Roms	3.5	2.1	3.5	5.0
Children's DVD	0.9	6.8	5.2	1.7
Children's Easy Reading	1.0	1.5	1.0	0.3
Children's Fiction	0.5	0.6	0.3	0.1
Children's Holiday	0.4	0.6	0.4	0.2
Children's Music Cassette	0.4	1.1	1.3	0.5
Children's Nonfiction	0.4	0.5	0.3	0.1
Children's Books with Cassettes	1.4	0.7	0.7	0.5
Children's Nonfiction DVD	6.1	0.2	2.1	1.9
Children's Nonfiction Video	6.1	3.3	2.5	1.9
DVD	4.7	7.7	6.5	3.3
Fiction	1.0	1.5	0.4	0.1
Graphic Novels	0.4	1.5	0.9	0.3
Large Print Fiction	0.8	1.2	0.7	0.8
Large Print Nonfiction	0.6	1.2	0.7	0.6
Magazines	0.1	0.0	0.0	0.0
Music CDs	1.7	0.8	0.9	0.7
Mystery	0.7	2.3	0.5	0.2
New Books	1.5	2.6	1.6	0.4
Nonfiction	0.4	0.4	0.2	0.1
Nonfiction DVD	2.3	5.1	4.5	2.2
Nonfiction Video	0.7	0.8	0.9	0.7
Science Fiction	0.4	1.0	0.4	0.2
Video	3.4	4.1	3.3	2.3

Table II – Mean and Standard Deviation of Number of Checkouts and Price for the 10% Most Popular Books in Each Category.

Category	Measure of Total # Checkouts		Measure of Price (in Dollars)	
	Mean	Standard Deviation	Mean	Standard Deviation
Books on Cassette	43.76	8.50	23.53	14.41
CD Roms	53.17	8.91	36.59	13.64
Children's Basic	43.08	9.59	13.00	4.37
Children's Books with Cassette	33.44	8.67	10.09	0.05
Children's DVD	72.93	9.54	30.01	0.04
Children's Fiction	20.40	8.78	9.79	5.14
Children's Music Cassette	37.83	9.90	18.35	5.86
Children's Nonfiction Video	130.32	22.28	14.88	5.23
DVD	106.07	12.05	21.95	6.20
Graphic Novels	44.00	35.24	14.00	5.66
Large Print Nonfiction	31.45	9.35	17.64	6.11
Music CD	40.15	8.94	16.09	5.11
New Books	27.34	5.96	15.14	3.44
Nonfiction	40.53	10.76	21.98	7.90
Science Fiction	22.62	5.05	13.48	4.05
Books on CD	46.65	6.77	27.48	22.67
Children's Award Winners	31.29	7.36	17.46	4.64
Children's Books on CD	55.00	11.05	35.67	4.92
Children's CD Roms	83.63	8.08	27.07	5.77
Children's Easy Readers	48.38	12.23	14.19	4.43
Children's Holiday Books	16.02	4.33	13.58	3.88
Children's Nonfiction DVD	13.67	0.94	20.00	0.00
Children's Nonfiction	17.03	7.68	14.26	4.17
Fiction Books	32.62	7.59	13.02	3.82
Large Print Fiction	36.85	11.37	18.56	6.09
Magazines	4.21	2.16	5.04	0.75
Mystery	34.64	7.41	14.35	1.79
Nonfiction DVD	60.40	14.71	25.40	12.75
Nonfiction	17.46	10.33	15.24	6.38
Video	143.05	24.52	18.56	10.34