SIMPLE SAS® ANALYSES THAT ADD VALUE TO CORPORATE DATABASES

Richard Sue, Ontario Hydro

Introduction

In order to meet future electricity demand and respect the sensitivities of the environment, Ontario Hydro has embraced the energy management challenge for the 1990s and beyond; Ontario Hydro believes in energy conservation and is working to make Ontario more energy efficient. The task at hand is to make electricity users aware of their behaviour and attitudes towards energy use and to get them to actively participate in the wise and efficient use of electricity.

The implementation of energy management is essentially a marketing issue. Relatively little is known about our market from the perspective of energy management. Therefore, in addition to collecting primary data through interviews and surveys, it is important to identify any existing data resource and to "mine" that data resource to extract relevant information. Since our department – the Market Research Department – has no control over how these secondary data resources are compiled, both the format and content of each data resource can vary considerably. To effectively process and analyze these data files the Market Research Department relies heavily on the capabilities of the SAS® System.

Much of the work in the Market Research Department is project-oriented, and geared towards providing fast response to information requests. For some of our projects, we are provided with the necessary data from another department, and due to the fast turn-around time required, it is not possible to develop elaborate reporting schemes which produce final reports by directly tapping into these databases. Instead, SAS software is used to perform the data analysis functions, while other software packages are used for reporting the findings.

Background

In January 1989 Ontario Hydro introduced the Energy Efficient Lighting Incentives Program, which offers financial incentives to Commercial and Industrial customers to replace their building's lighting equipment with more energy efficient alternatives. Customers wishing to participate in the Program must complete an application form describing the intended work to be undertaken and the expected electricity demand (kW) savings that would result. As part of the administration and monitoring of this Program, the data from these application forms are stored electronically in a DBASE IV database. The main purpose of the database is to track applications through the approval and payment process. However, the data can also be used to understand and improve the program's processes.

The Market Research Question

In order to make improvements to the Program so as to increase customer participation and target new customer groups, it is important to know who is already participating. The database for the Lighting Program was identified as a convenient source of valuable information about program participants.

From a market research perspective, the objective of analyzing the data in the Lighting Program database was to identify any trends and patterns among various categories of program applicants. Following are some of the research questions which needed to be answered:

- Which segments of the market are applying for incentives?
- Where are these participants located?
- Which energy efficient lighting technology is the most popular?
- What is the average length of time taken to complete the projects specified in the applications?
- What is the average amount of incentives paid to each applicant?

Database Conversion

One of the first tasks to perform before converting a DBASE database into a SAS dataset is to ensure that the variable names will survive the conversion. DBASE field names can have up to 10 characters, while SAS version 6.04 variable names can have a maximum of 8 characters. In the conversion process, DBASE field names with more than 8 characters are truncated to 8 characters for the SAS variable name. If the truncation results in two or more variables having the same name, then the conversion will not work properly. To correct the matter, the database field names must be changed using the DBASE software so that all resulting SAS variable names are unique.

PROC DBF is used to perform the conversion. The DB3~ option can be used to convert numeric, character, date, and logical fields of DBASE IV databases into variables of a SAS dataset. The following program statements show how to convert the DBASE IV database named eel_inf.dbf into a permanent SAS dataset called lighting.eel_inf.

LIBNAME lighting 'c:\lighting';
PROC DBF DB3=eel_inf OUT=lighting.eel_inf;
RUN;

If the original database was not designed for data analysis purposes, it may be necessary to create additional variables in the SAS dataset to accommodate subsequent analyses. In our case, to be able to classify applications
by the year and month in which they were received, it was most convenient to have two numeric variables – one for the year and one for the month – instead of a single date variable containing the date the application was received.

To create new variables in the SAS dataset, a DATA step is needed. The following program statements show how the year and month variables (dr_year and dr_month, respectively) were created from the date variable dat_rcvd using the SAS functions YEAR() and MONTH().

```sas
DATA lighting.eel_inf;
  SET lighting.eel_inf;
  dr_year = YEAR(dat_rcvd);
  dr_month = MONTH(dat_rcvd);
RUN;
```

**Data Analysis**

Advanced analytical methods are not needed to obtain useful information from existing databases. By simply producing counts of records, sums of numeric variables, and averages of numeric variables, valuable information can be extracted from corporate databases about trends and patterns in the marketplace. However, not all counts and averages are meaningful. You still need to selectively choose which variables or factors are worth examining. A clear understanding of the market research questions and the information needs of your client, together with insights you have about the market, will help determine which variables to analyze and which to ignore.

**I. Defining Classification Variables**

Variables which are suitable as classification variables are those which describe the market environment from the perspective of your corporation. For Ontario Hydro, we operate in five geographic regions: Central, Western, Eastern, Northeastern and Northwestern. We divide our market into five major sectors: Residential, Commercial, Industrial, Agriculture and Transportation. Each of these sectors, in turn, contain several segments. These variables (region, sector and segment) make ideal classification variables.

For the Energy Efficient Lighting Program in particular, additional classification variables needed to be defined. Clues for deciding which of the myriad database variables should be classification variables can be obtained by examining the market research questions you want answered. In our case, the Lighting Program offers financial incentives for several technology categories, and one of the research questions dealt with identifying which of these technologies was most popular with applicants. Therefore, we defined a classification variable called project type which identified the technology categories for which incentives were paid. Other classification variables were the year and month the application was received, and the approval status of the application.

**II. Record Counts**

Record counts provide a fast overview of your database. In our case it quickly told us how many applications were received for the Lighting Program. Comparing these application counts with population statistics provided a quick analysis of Program penetration.

There are several SAS procedures which can produce record counts by various categories: PROCs FREQ, SUMMARY, and TABULATE. The choice of procedure is a personal preference. Each procedure has advantages and disadvantages. Familiarity with all three procedures will offer the greatest flexibility with your data analysis. In this paper, examples using all three procedures will be described.

**Example 1: Number of applications received.**

To count the number of applications received from each region, the following program statements used PROC FREQ to do the counting. The variable containing the region identifier is reg_code.

```sas
PROC FREQ DATA=lighting.eel_inf;
  TABLE reg_code;
RUN;
```

Output:

<table>
<thead>
<tr>
<th>REG_CODE</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>20</td>
<td>0.4</td>
<td>20</td>
<td>0.4</td>
</tr>
<tr>
<td>CE</td>
<td>2125</td>
<td>38.7</td>
<td>2145</td>
<td>39.0</td>
</tr>
<tr>
<td>EA</td>
<td>1096</td>
<td>9.9</td>
<td>3241</td>
<td>59.0</td>
</tr>
<tr>
<td>NE</td>
<td>322</td>
<td>5.9</td>
<td>3563</td>
<td>64.8</td>
</tr>
<tr>
<td>NW</td>
<td>94</td>
<td>1.7</td>
<td>3657</td>
<td>66.6</td>
</tr>
<tr>
<td>WE</td>
<td>1838</td>
<td>33.4</td>
<td>5495</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Since PROC FREQ will list all unique values of the classification variable, any values which were incorrectly entered can be quickly identified. In the above output listing, 20 records were not assigned a proper region code. Knowing this, we could search the original database for these records and make the necessary changes.

The program statements below show how PROCs SUMMARY and TABULATE could be used to perform the record counts.

```sas
PROC SUMMARY DATA=lighting.eel_inf PRINT;
  CLASS reg_code;
RUN;
```

```sas
PROC TABULATE DATA=lighting.eel_inf;
  CLASS reg_code;
  TABLE reg_code,N;
RUN;
```

Care should be exercised when using these procedures for counting records when there are missing values for the classification variable. By default, these procedures display frequency counts for non-missing values of the classification variable. Fortunately, it is simple to get these procedures to treat missing values as a valid level
of the classification variable. In fact, by default, PROC FREQ will print the total frequency of missing values below its output table. Including missing values as a valid level of the classification variable will help you audit the quality of your database by identifying how many records have missing data for the classification variable.

The following statements include the option for specifying missing values as a valid classification level.

```sas
PROC FREO DATA=lighting.eeUnf;
    TABLE reg_code / MISSPRINT;
RUN;
```

```sas
PROC SUMMARY DATA=lighting.eeUnf PRINT MISSING;
    CLASS reg_code;
RUN;
```

```sas
PROC TABULATE DATA=lighting.eeUnf MISSING;
    CLASS reg_code;
    TABLE reg_code, N;
RUN;
```

The regional counts produced from the above can be further broken out into sub-levels to show detailed information for each region. For example, sub-levels could show the number of applications received by each region by month, or by sector, or by segment within each sector. Each of the above procedures would show these sub-levels in a different output format, your choice of which procedure to use would depend on the particular output format you desire. Since we used other software packages to chart and report the information, the output format was not an issue.

The program statements below show how sub-levels can be added to PROCs FREO, SUMMARY and TABULATE. The PROC TABULATE example includes the segment sub-level within each sector.

```sas
PROC FREO DATA=lighting.eeUnf;
    TABLE reg_code*sector / MISSPRINT;
RUN;
```

```sas
PROC SUMMARY DATA=lighting.eeUnf PRINT MISSING;
    CLASS reg_code sector;
RUN;
```

```sas
PROC TABULATE DATA=lighting.eeUnf MISSING;
    CLASS reg_code sector segment;
    TABLE reg_code*sector*segment, N;
RUN;
```

III. Sums and Averages

Sums and averages of key numeric variables can provide useful tracking and monitoring information. In the Lighting Program, some of the variables we track include the quantity of lighting products which have received financial incentives, the total incentives paid out for each application, and the electricity demand (kW) savings resulting from each application. From these variables we can provide the following information:

- Number of Units which received incentives, for each lighting technology category (e.g. compact fluorescent lamps, energy efficient fluorescent tubes, etc.),
- Total annual financial incentives paid to applicants,
- Total annual electricity demand (kW) savings,
- Average financial incentives paid per application,
- Average electricity demand (kW) savings per application, and
- Average financial incentives paid per kW saved.

This information is used to track actual costs and kW-savings against targeted goals for each market segment.

PROCs MEANS, TABULATE and UNIVARIATE can quickly calculate the sums and averages for numeric variables.

Example 2: Total and average financial incentive paid in each technology category.

The different technology categories which receive financial incentives are:
- Compact Fluorescent Lamps,
- Energy Efficient Fluorescent Tubes,
- Lighting Redesigns,
- Fluorescent Power Reducers, and
- Fluorescent Reflectors

The variables f1_comp, f1_lubes, redesign, pwr_redu, and f1_relf contain incentive data for each technology category, respectively.

In the following program statements the SAS dataset lighting.eeUnf is screened to select only those records which are approved to receive incentives; this new dataset is called lighting.approved. Then, using PROC MEANS, we obtain the total incentives and the average incentives for each technology category, classified by the year the application was received, and by market sector. We also ask for the standard deviation and the variance for each variable.

```sas
DATA lighting.approved;
    SET lighting.eeUnf;
    WHERE ap_status='A';
PROC MEANS DATA=lighting.approved
    N NMISS SUM MEAN STD VAR;
    CLASS dr_year sector;
    VAR f1_comp f1_lubes redesign pwr_redu f1_relf;
RUN;
```

Using PROC TABULATE, the program statements to calculate the same statistics are:

```sas
PROC TABULATE DATA=lighting.approved;
    CLASS dr_year sector;
    VAR f1_comp f1_lubes redesign pwr_redu f1_relf;
    TABLE dr_year*sector,
        (f1_comp f1_lubes redesign pwr_redu f1_relf)*
        (N SUM MEAN STD VAR);
RUN;
```

PROC UNIVARIATE can produce similar descriptive analyses as PROC MEANS. But PROC UNIVARIATE is
different because it can produce more detailed analyses of the distribution of a variable. To obtain separate analyses on classification groups, such as year application was received and market sector, BY variables must be used. When a BY statement is used, PROC UNIVARIATE expects the input data to be sorted in order of the BY variables.

Therefore, to perform the same analysis as above, but using PROC UNIVARIATE, a sort would have to be performed on the dataset. By default, PROC UNIVARIATE produces sum, mean, standard deviation and variance statistics.

PROC SORT DATA=lighting.approved;
  BY dr-year sector;
PROC UNIVARIATE DATA=lighting.approved;
  VAR fl_comp fl_tubes redesign pwr_redu fl_refl;
  BY dr-year sector;
RUN;

IV. Cross-Tabulations

A common way of presenting market research data is to use the cross-tabular format. This format is convenient for comparing one classification variable with a series of related variables. PROC TABULATE is a powerful tool within the SAS system for creating cross-tabulation tables of frequency data and descriptive statistics. PROC TABULATE was used in both examples above, and with proper planning, almost all your information needs could be satisfied with just a few runs of PROC TABULATE.

PROC FREQ can also be used to produce cross-tabulations, but only for frequency data. However, PROC FREQ can perform statistical tests of association to determine the relationship between two variables. A knowledge of statistical methods is needed in order to properly administer these tests and interpret their results.

Reporting

In the previous discussion about data analysis, not much emphasis was placed on the output format of the SAS procedures. This was done deliberately. Except for the TABULATE procedure, the output produced from the SAS procedures considered in this paper are not suitable for reporting purposes.

In today's computing environment, most people have a good working knowledge of mainstream word-processing, spreadsheet and graphics software, and find it easier and quicker to use these packages to do charting and reporting of research findings. By combining the analytical capabilities of SAS with the reporting features of other software packages, we can quickly produce final research reports. Results of the SAS data analyses can readily be transferred to other software packages by several methods. In most cases, the amount of data to be transferred is small and can be done, for example, by simply re-entering the numbers into a spreadsheet. Charts produced from the spreadsheet can then be easily incorporated into a report written using word-processing software. The following two charts are examples of this.

Figure 1: Number of applications received each month

Figure 2: Number of applications received each year, by region and sector.

Summary

The proliferation of internal databases in the current business environment offers a good opportunity for obtaining valuable market research information from this data by performing some simple analytical procedures. Using PROCs FREQ, MEANS and TABULATE to produce record counts, sums, averages and simple cross-tabulations, useful information about trends and patterns in the marketplace can be extracted from existing corporate databases.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.