

CRITICAL STEPS TO MEANINGFUL CAUSAL DATA ANALYSIS

Deborah Wall, Providian Insurance, Philadelphia, PA

ABSTRACT

Sophisticated statistical procedures aside, what are the fundamental, necessary steps to assuring that your data analysis will be meaningful, add value and achieve its objectives? This paper outlines a helpful checklist and gives examples of SAS code and procedures to assure that your analysis or model is not a GIGO project (garbage in, garbage out). Such simple PROCS as PRINT, FREQ, MEANS, CORR, FACTOR, and TABULATE can help turn a mediocre analysis into a true gem. The examples are inclusive of a causal business analysis designed to add value through providing answers to the key question: HOW DO I ENHANCE THE PROFITABILITY OF MY COMPANY? A target marketing/predictive modeling angle is emphasized.

INTRODUCTION

SAS® is an invaluable tool to the knowledgeable analyst. However, knowing the key steps to setting up a meaningful causal analysis and coupling it with efficient coding techniques can make your task much easier, providing your company with relevant, insightful decision support.

REMEMBER: FANCY STATISTICAL TECHNIQUES AND COMPLEX, ELABORATE CODE ARE NO SUBSTITUTE FOR A WELL-PLANNED ANALYSIS AND BASIC, EFFICIENT CODE.

WHY CAUSAL ANALYSIS?

Causal data analysis links a series of independent data (descriptor variables) to a dependent "event" (predicted variable). The value in using causal analysis is that we can infer, or predict, the outcome of a future event based on a selected set of key independent variables. Examples of PROCS used for casual data analysis include: linear regression: PROC REG (a continuous numeric prediction), logistic regression: PROC LOGISTIC (a yes/no probability prediction), discriminant analysis: PROC DISCRIM and PROC STEPDISC (classification of an event), to name some of the more common techniques.

Some typical causal data analyses answer questions such as:

- Who is likely to buy my product?
- How much of a profit margin can I expect?
- Which customers are the most profitable?
- What is the expected revenue?
- Who is likely to stop paying or lapse their account?

The following steps are a handy checklist to ensure that you are heading in the right direction in order to be able to provide solid, confident answers to these questions.

CHECKLIST OF CRITICAL STEPS TO MEANINGFUL CAUSAL DATA ANALYSIS:

STEP 1: ACCURATELY DEFINE THE QUESTION THE ANALYSIS SHOULD ANSWER:

This is the most important, yet most neglected step in the analysis. Address business leverage points: revenue expansion, profit improvement, expense reduction.

KEEP THE ANALYSIS QUESTION DEFINITION SIMPLE. FILL IN THE ANCILLARY DETAILS LATER. BREAK THE QUESTION INTO SEVERAL PARTS IF NECESSARY.

A good example of a clear problem definition is as follows:

This analysis will provide an answer to what defines a profitable customer. This consists of two parts:

A) What factors contribute to the likelihood of a purchase, and differentiate a buyer from a non-buyer?

B) Of the buyers, what factors can tell us how much of a profit to expect from each?

SAS® is needed to verify that your analysis question definition is set up correctly from a data analysis point of view. Which brings us to step 2.

STEP 2: DATA FILE PREPARATION:

In other words: KNOW YOUR DATA. Since this is a causal data analysis, it is critical that the dependent (predicted) variable(s) be properly defined and populated on the file.

Reading in the analysis file: If you are unfamiliar with the file structure or variable meanings, dump a small

portion of the file. REMEMBER: WHEN IN DOUBT, DUMP IT OUT.

Read in 100 observations of all variables which you think you'll need, and all of those of which you're not sure. Data exploration can help weed out the ones you won't use.

REMEMBER: THE INDEPENDENT VARIABLES YOU USE FOR PREDICTION MUST BE READILY AVAILABLE AND CONSISTENTLY CODED EACH TIME YOU RUN YOUR PREDICTION ANALYSIS.

Use list input to read in all variables on the file if you're unfamiliar with the variables.

Use meaningful variable names.

Use the following code for the "dump":

```
OPTIONS OBS=100;
DATA DUMP1;
INFILE IN1;
BUYFLAG 1.
PROFIT 5.2
AGE 3.
INCOME 7.
REGION $CHAR2.
GENDER $CHAR1.
HOMEOWN $CHAR1.
HOMEVALU 8.
;
PROC PRINT;
```

Look for unusable symbols or a totally unpopulated field. If this occurs in a critical variable field, your file needs to be recreated, and you should go no further with the analysis until the file is correct. You will be wasting your time and the company's money if you do.

Once you have determined that the file is usable, and have read in the entire analysis universe of necessary variables, use **PROC FREQ** to check the categorical or limited level discrete variables (under 100 levels). Use **PROC UNIVARIATE** and **PROC MEANS** to examine continuous numeric variables. The following examples of efficient code may help:

```
PROC FREQ DATA=IN1.DATA; TABLES BUYFLG
GENDER HOMEOWN AGE REGION;
```

```
%MACRO CONTVAR
PROFIT AGE INCOME HOMEVALU
%MEND CONTVAR;
```

```
PROC UNIVARIATE DATA=IN1. DATA; VAR
%CONTVAR;
```

```
PROC MEANS DATA=IN1.DATA MEAN MIN MAX
N NMIS; VAR %CONTVAR;
```

NOTE : If your dataset contains over 200,000 records, an appropriately large random sample is strongly suggested, as these procedures are CPU intensive, and may abend on a large dataset.

Check to make sure the levels and distribution of the dependent and independent variables are what occurs in "the real world". For example, the ratio of buyers to non-buyers should agree with sales performance report results, or the profit margin ranges should reflect those seen on a balance sheet.

STEP 3: INDEPENDENT VARIABLE DATA REDUCTION:

Try to reduce your independent variable universe to less than 60 variables, or about 20 main factors. You have a better chance of finding the most important predictor (independent) variables if you reduce your universe by eliminating redundant variables which measure the same factor. Principle components analysis using **PROC FACTOR** is the perfect technique for this task. Remember, the data must be in numeric format. And do not include the dependent variable within your factor analysis.

Again, if your analysis universe is large, take a random sample. Limiting the number of factors (NFACT=50) keeps the procedure from forming small, insignificant factors which may be weak or somewhat redundant, and limiting the Eigen Value criteria to 1 or more helps to eliminate any factors which don't have at least one strong variable contributing to the population variance explanation.

```
PROC FACTOR DATA=IN1.DATA
ROTATION=VARIMAX NFACT=50 SCREE
MINEIGEN=1 S C ; VAR NUMVAR10 - -
NUMVAR120;
```

Use the scree plot to determine how many critical factors your population contains: the bend in the elbow shows where the point of diminishing returns begins. A good rule of thumb to follow is not to include factors which fall after a major bend in the elbow.

Within the chosen number of factors, select the top one to five variables which load on the factor. Do not select variables which are weak loadings on a factor, i.e. have a loading of less than .6.

KEEP ONLY THE INDEPENDENT VARIABLES YOU NEED, PLUS THE DEPENDENT VARIABLES. USE THE DROP OR KEEP OPTION IN THE SET STATEMENT TO CREATE A NEW, SMALLER DATASET.

STEP 4: RECODE/TRANSFORM THE FINAL SET OF INDEPENDENT VARIABLES:

In order to transform insignificant variables into new significant variables which may better link the cause to the effect, it is necessary to collapse or recategorize the independent variable distribution. PROC FREQ and PROC UNIVARIATE can help with collapsing of variables into fewer categories. A good rule of thumb is to collapse at even intervals: i.e. 10% or 20% groupings, depending on how the independent variable distributes.

- Code for missing values by either mean plugging, or using a unique numeric value
- All categorical variables must be transformed into multiple dummy variables, or into numerically ranked equivalents to their symbolic rankings.
- Try creating interaction variables between, for example, two significant bilevel variables.
- Use mathematical transformations such as logs and square roots to create a better fitting variable.

REMEMBER, CREATIVITY COMBINED WITH A KNOWLEDGE OF THE BUSINESS ANALYZED, IS IMPORTANT IN CREATING NEW TRANSFORMED VARIABLES. TEST OUT YOUR HYPOTHESES.

STEP 5: CHOOSE THE CORRECT PREDICTIVE TECHNIQUE AND USE THE STEPWISE OPTION TO GET YOUR PREDICTION EQUATION POSSIBILITIES:

PREDICTION DESIRED USE:

BIVARIATE	PROC
PROBABILITY: LOGISTIC/	
FROM 0 TO 1	SELECTION=
	STEPWISE

Note: Due to the CPU intensive nature of the PROC LOGISTIC STEPWISE option, the user may want to

consider using the PROC REG STEPWISE option instead. Results are similar for both techniques.

DISCRIMINATION	PROC
BETWEEN CLASSES	STEPPDISC

CONTINUOUS	PROC
NUMERIC	REG /SELECTION=
	STEPWISE

STEP 6: USE PROC CORR IN CONJUNCTION WITH YOUR STEPWISE OUTPUT TO ELIMINATE MULTICOLLINEAR AND UNSTABLE VARIABLES. FOR A CATEGORICAL/BIVARIATE DEPENDENT VARIABLE USE PROC FREQ WITH THE CHI SQUARE OPTION TO ELIMINATE NON SIGNIFICANT VARIABLES :

Use the PROC CORR and the RANK option to eliminate any variables selected in the stepwise procedure which are highly correlated with each other. Instead, select the one which is most strongly correlated to the dependent variable. This helps stabilize the prediction equation. Use PROC CORR also to check the direction of the correlation between the dependent and independent variables. If the direction of the relationship between each independent variable and the dependent variable is not consistent between the stepwise output and the correlation matrix, that independent variable is unstable. Try to eliminate it and find a more stable replacement.

The chi-square test for independence between a bivariate dependent variable and categorical independent variables will help to identify noisy, non significant variables which may have been selected by the stepwise procedure by adding incremental explanation.

REMEMBER: IF YOU CAN BUILD A MAIN EFFECTS MODEL, WHICH IDENTIFIES STABLE, MAJOR FACTORS WHICH CONTRIBUTE SIGNIFICANTLY TO EXPLAINING THE OUTCOME YOU WISH TO PREDICT, YOUR MODEL OR ANALYSIS HAS A GREATER CHANCE OF HOLDING UP OVER TIME, AND MAKING MORE MANAGERIAL SENSE.

*****THIS PROGRAM RUNS CHI-SQUARE AND CORRELATION ANALYSES FOR BUYER PREDICTION;

```
PROC CORR RANK DATA=INI.DATA; VAR
SQRHVAL AGE65PL HOMEOWN
UNDAGE60 BUYER;
```

```
PROC FREQ DATA=INI.DATA;
TABLES (AGE65PL HOMEOWN FEMALE
MALE25KP)
*BUYER/CHISQ;
```

STEP 7: AFTER SELECTING A FINAL SET OF RECODED, STABLE VARIABLES, USE THE NON STEPWISE PREDICTIVE TECHNIQUE TO GENERATE YOUR FINAL EQUATION. USE A TABULAR GAINS CHART TO CHECK THAT INDEPENDENT VARIABLE TRENDS AGREE WITH THE COEFFICIENT IN THE EQUATION.

After settling on a group of stable, managerially significant independent variables, run the predictive technique which renders the scoring equation to predict future outcomes. Use a gains chart, built in PROC TABULATE, to check the directional trends of your independent variables. The final model should contain variables whose coefficients match the directional trends outlined in the gains chart.

```
%MACRO VARLIST;
AGE INCOME HOMEVAL MALIN25P
%MEND VARLIST;
RUN;
TOTREC=1;
PROC TABULATE; CLASS SEGMENT; VAR
TOTREC BUYER %VARLIST SCORE;

TABLE (SEGMENT ALL), ((TOTREC
BUYER)*(SUM PCTSUM<TOTREC>))
((SCORE %VARLIST)*(MEAN MIN MAX));
```

CONCLUSION

This critical steps checklist does not pretend to be a comprehensive methodology of predictive model building and causal analysis. Instead, it should provide the user with a road map for the analysis, design and construction to render a meaningful and stable answer to a critical business question.

There are many fine points within the predictive technique which have not been covered within the scope of this paper. However, these steps form the backbone of the key steps that any good analyst must follow to obtain statistically and managerially significant analyses, whatever the analysis arena. Your SAS® manuals have wonderfully written introductory chapters and examples which can provide more insight into some of the fine points. Use the SAS® manuals often. They can be an analyst's best friend.

The other best friend of an analyst is strong industry and business process knowledge. Do all you can to learn about the industry in which you are employed, and the specific process flows which support the areas of analyses upon which you focus.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks of their respective companies.