

Some SAS Techniques Used in Insurance Business

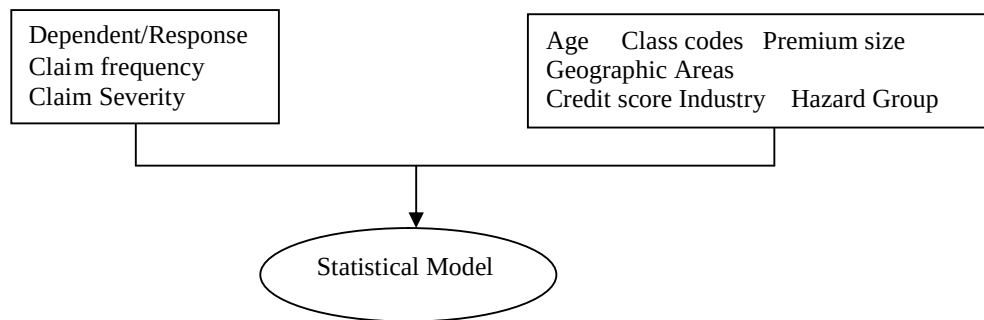
Wei Yu

Abstract

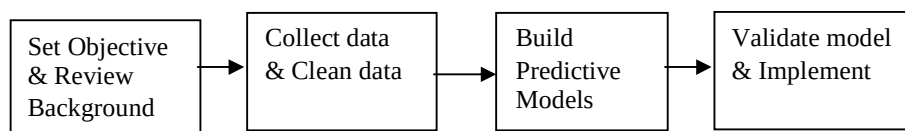
There are many statistical techniques used in predictive modeling. In this paper, I present the selection procedures for building the predictive models. The difference between regression and GLM, the difference between PROC GLM and PROC GENMOD are explained in this paper. Some techniques for manipulating and cleaning the data are also presented. In insurance business, the factor of territory plays more and more important role today. We need to know the geographic distribution of our data. The quarterly updated SAS zip code file is a good resource to help locate the distribution of data. In this paper, I present how I code to visually present the distribution of data in a map, not only using the different colors to show the distribution of data, but also labeling the name of fields in the map.

Introduction

Predictive modeling is an idea whose time has come. Simply put, predictive modeling is a process by which one uses statistical of data to make predictions about future events or otherwise gain actionable insights about multidimensional problems. The basic structure for a predictive model is as follows:



The basic procedures will be



During these procedures, we need to select predictive variables, decide the model options, use some SAS techniques to clean the data and run the statistical models.

The basics for PROC REG, PROC GLM, and PROC GENMOD

Regression models and Generalized Linear Models (GLM) are widely used in predictive modeling. GLM theory is a significant advance beyond linear regression theory. A major part of this advance comes from allowing a broader family of distributions to be used for the error term, rather than just the Normal (Gaussian) distribution as required in linear regression.

1) Using PROC REG for Simple Regression Analysis

```
PROC REG DATA=TEST;  
      MODEL WEIGHT=HEIGHT;  
      PLOT WEIGHT*HEIGHT;  
RUN;
```

2) Using PROC REG for Multiple Regression Analysis

There are 3 methods to select independent variables into models: stepwise, backward and forward.

```
PROC REG DATA=APTDATA;  
MODEL RENT = AREA BEDS BATHS DISTANCE SHUTTLE AGE  
      / SELECTION=STEPWISE;  
RUN;
```

3) Using PROC GLM for analysis of variance

PROC GLM is mainly used for analysis of variance. For balanced data – where there are equal numbers of observations in each classification, PROC ANOVA can be used. For unbalanced data, use PROC GLM.

```
PROC GLM DATA=HEALTH;  
CLASS GENDER;  
MODEL WEIHT= GENDER HEIGHT;  
RUN;
```

4) Using PROC GENMOD for Statistical Modeling, especially for complicated modeling

PROC GENMOD can be used for above mentioned regression model, GLM model; it can also be used for Logistical Model.

The following codes are an example for logistical Model:

```
PROC GENMOD DATA=LOSS;  
CLASS TYPE;  
MODEL LOSSFLAG=TYPE SEVERITY SIZE/LINK=LOGIT DIST=BINOMIAL;  
RUN;
```

Based on the data and the model goal, sometimes you may change the parameters of link and distributions. For example, if the dependent variable is the loss frequency, then the Poisson distribution will apply instead of Binomial distribution.

SAS Mapping tool

Territory factor now becomes an important factor in business analysis. Sometimes we want to know the geographic distributions of our business. In SAS, there is a SAS zip code file in library SASHELP and an updated zip file can be obtained via SAS web site. Some SAS data files in Library Maps are very helpful too. The following is an example that I used to find the distribution of insurance policyholders. I want to show a map with all counties in Texas with the county name labeling on the map and also show the number of policies in each county.

Step 1:

```
/* Create a map data set of Texas counties. */  
data tx;  
    set maps.counties;  
    where state=48; /* FIPS code for Texas */  
run;
```

Step 2:

```
/* my data file.*/  
data Policy1;  
set policy;  
keep state county countynm number;  
run;  
  
/*Sort my data.*/  
proc sort data=Policy1 out=Policy2;  
by state county;  
run;
```

Step 3:

```
/* Macro variable for labeling.*/  
%annomac;  
%maplabel(tx, policy2, anno, countynm, state county,  
font=centb, color=black, size=.7);  
run;
```

Step 4:

```

/*format */
proc format;
value mnfmt 1-100='Less than 100'
             100<-500='100 to 500'
             500<-1000='500 to 1000'
             1000<-100000='Greater than 1000';
run;

```

Final Step:

```

ods pdf file='h:\map\map1.pdf';
options reset=all colors=( white lime yellow pink);

proc gmap data=policy2 map=tx all;
  id state county;
  choro number / coutline=black legend=legend1 discrete
                 cemty=black annotate=anno;
legend label=(position=(top center) f=centb h=3pct c=black
                  'Number')
         value=(c=black) frame;
title c=black f=centb h=3pct 'The distribution of Policies
in Texas';
run;
ods pdf close;
quit;

```