Paper 258-26
Using SAS ® to Minimize Exposure and Maximize Compliance

Stephen Sloan, Bristol-Myers Squibb, Plainsboro, New Jersey

ABSTRACT

When presented with requests from government agencies for information to verify compliance with EEO standards, many organizations simply respond by checking average salaries by race and gender. The organizations then assume that they are out of compliance it the salaries differ. It is possible for these organizations, through judicious use of SAS software and statistical analysis, to minimize their exposure by identifying covariates which may explain the apparent discrepancies, making sure the remaining discrepancies are statistically significant, and calculating salary adjustments to eliminate only statistically significant differences. Furthermore, organizations can use these techniques proactively to detect possible sections of the organization which may require further analysis. These analyses, and the resulting actions taken, can help make the organizations less likely to be vulnerable to future audits. This application uses data in TSO® and DB2® and uses Base SAS and the SAS/STAT® procedures GLM®, GLMMOD®, REG®, and CORR®. The necessary skill level would be found in people with statistical expertise and expertise in SAS/STAT and the SAS macro facility.

INTRODUCTION

My organization receives many requests from government agencies for information about how we compensate people at different levels in the organization. These requests usually apply to specific divisions of our organization or to specific geographical areas in which operate, with a general focus on race and gender equality of compensation. As a federal contractor, as an entity operating in the United States, and as a law-abiding and ethical organization, we feel it is important to provide fairness and equity and to detect situations where problems may occur.

THE OLD PROCESS

In the past, we generally pursued the following sequence when exploring compensation issues:

1. Contact our statistical consultant to find out what was needed.
2. Use SAS to prepare a file with the information requested by the consultant.
3. E-mail the file to the consultant.
4. Receive a request for more data and different formats from the consultant.
5. E-mail the new information to the consultant.
6. Receive the analysis.
7. Pay the bill.

This process had a number of flaws, chief among them being that it was opaque and expensive and it was difficult to analyze or modify the reports. So we asked the consultant to share his logic with us.

It turned out that our statistical consultant hired a third-party statistician who used SPSS® to analyze a variety of covariates (such as age, length of service, education, and others) to see if any apparent discrepancies could be explained by factors other than race or gender. Once this analysis was complete, any discrepancies significant at the 5% level that could be not be explained by the covariates were resolved by recommended salary adjustments. Upon closer inspection, it turned out that these adjustments were calculated to eliminate the statistical significance of the discrepancy.

THE IMPROVEMENTS

My organization asked me to see if I could replicate and improve the model, and do so in-house to reduce the cost of the analysis. In addition, doing the work in-house meant that the data and the analysis were readily available for future refinement. Future reports would no longer require going back to our consultant and waiting for results. Instead, we could modify the analysis and re-run whenever we needed. Also, we could use our analytical method in other areas whenever necessary. Finally, since I was already spending a considerable amount of time preparing and reformatting data for our vendor, there would probably be no increase in the amount of resources involved if I were to perform the analysis myself instead of preparing the data for somebody else to use.

My first step was to purchase a copy of the current release of SPSS to make sure that I could replicate the analysis. The statistician hired by our consultant was very open about his procedures and, with a little bit of help from him, I was able to achieve the same results using my copy of SPSS.

My next step was to rewrite the model in SAS to make sure that I had replicated the logic and could achieve the same results. Since our database is a TSO/DB2 database, I used mainframe SAS, which in our organization is version 6.09 with TS470. The statistical consultants at the SAS Help Desk were very helpful in advising me about how to replicate the SPSS analysis. The SAS code that replicated the SPSS analysis is in Figure 1.

In order to show the value of performing the analysis in the first place, I computed the cost of raising the salaries of the affected race or gender all the way to the mean of the other group. This was then used as a benchmark to show the savings available if we only raised the salaries to the level necessary to make the difference statistically insignificant.

Finally, I decided to make two improvements which not only did not seem to be possible in SPSS, but which we had been told could not be done in an automated fashion:

Instead of guessing what would bring the salary discrepancy below the 5% level, I wrote a macro that would add $10 to the salaries of the race or gender with the lower average salary, re-test for statistical significance, and keep doing so until the discrepancy was statistically insignificant (Figure 2).

I also tested the interaction of race and gender for statistically significant differences in salary.   Beginning with the lowest average salary, I would add $10 to each salary. If there was still a statistically significant discrepancy when the sum passed the second lowest average salary, I began adding $10 to that one, also. This continued until there was no longer a statistically significant discrepancy and enabled us to detect problems that otherwise would not have been apparent.

As a result of bringing the analysis in-house, doing the analysis in SAS instead of SPSS, and refining the analytical procedures as mentioned above, our organization was able to save a considerable amount of money both in salary adjustments and in consulting fees. In addition, we have been able to be more pro-active in ferreting out potential discrepancies and correcting them, instead of waiting for audits and then trying to explain problems that surfaced during the audit. Finally, we have historical analyses that we can use to detect changes in our operating environment that may not be otherwise apparent.

We intend to continue refining the model. Our goals are to improve our processes, reduce race and gender discrepancies, and gain more understanding of current statistical methods and processes.   We would also like to see if we can carry the process further. To this end, we are also looking at the SAS Data Mining product, the Enterprise Miner®, to see if it can help us reduce the cost of testing and analysis when using our model. We could use the different features of the Enterprise Miner to find other covariates that may account for seeming race or gender discrepancies in salary. We would also use the Enterprise Miner to pinpoint the demographic traits that would be most likely to occur in a cohort with statistically significant salary discrepancies by race or gender. We could then save time and expense by quickly applying our model to test for discrepancies in those areas.

CONCLUSION

In conclusion, replacing purchased SPSS analysis with in-house SAS analysis has allowed us to detect and correct race and gender discrepancies more accurately and quickly and at a lower cost to the organization.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stephen Sloan
Bristol-Myers Squibb
777 Scudders Mill Road
Plainsboro, New Jersey 08536
609-897-3550
Stephen.Sloan@BMS.COM

TRADEMARK CITATION

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Figure 1 – Testing for discrepancies

```
  PROC GLM DATA= FILE NOPRINT
OUTSTAT=GLMOUT;
  BY DIVISION GRADE MIDPOINT; /*** MIDPOINT IS
THE SALARY MIDPOINT
                                                  FOR THE
GRADE ***/
    CLASS GENDER RACE;
    MODEL SALARY=GENDER RACE GENDER*RACE
              ACHEV EXCEL OTHER PROMO
              SERVICE TIP TIG / SOLUTION;
    LSMEANS GENDER RACE GENDER*RACE /
ADJUST=TUKEY
                                   OUT=LSMOUT;

  DATA LSMOUT2;
    SET LSMOUT;
    BY DIVISION GRADE MIDPOINT;
    IF FIRST.MIDPOINT THEN DO;
       MI=.; NONMI=.; FEMALE=.; MALE=.;
       MINFEM=.;   MINMALE=.; NONMINF=.;
NONMINM=.;
    END;
    IF RACE=0 AND GENDER=' ' THEN
MINORITY+LSMEAN;
    IF RACE=0 AND GENDER=0 THEN
MINFEM+LSMEAN;
    IF RACE=0 AND GENDER=1 THEN
MINMALE+LSMEAN;
    IF RACE=1 AND GENDER=' ' THEN
NONMI+LSMEAN;
    IF RACE=1 AND GENDER=0 THEN
NONMINF+LSMEAN;
    IF RACE=1 AND GENDER=1 THEN
NONMINM+LSMEAN;
    IF RACE=' ' AND GENDER=0 THEN
FEMALE+LSMEAN;
    IF RACE=' ' AND GENDER=1 THEN
MALE+LSMEAN;
    IF LAST.MIDPOINT;

  PROC SORT DATA=GLMOUT;
    BY DIVISION GRADE MIDPOINT;
    WHERE _TYPE_='SS3';

  DATA GLMOUT2;
    SET GLMOUT;
    BY DIVISION GRADE MIDPOINT;
    IF FIRST.MIDPOINT THEN DO;
       SERVICE=.;  TIG=.; TIP=.; AGE=.;
       SEX=.; RACE=.; INTERACT=.;
    END;
    IF _SOURCE_='GENDER' THEN SEX+PROB;
    IF _SOURCE_='RACE' THEN RACE+PROB;
    IF _SOURCE_='GENDER*RACE' THEN
INTERACT+PROB;
    IF _SOURCE_='AGE' THEN AGE+PROB;
    IF _SOURCE_='SERVICE' THEN SERVICE+PROB;
    IF _SOURCE_='TIG’ THEN TIG+PROB; * Time in
grade;
    IF _SOURCE_='TIP' THEN TIP+PROB; * Time in
position;
    IF LAST.MIDPOINT;

  PROC SUMMARY;
    BY DIVISION GRADE MIDPOINT;
    CLASS GENDER RACE;
     VAR SALARY;
    OUTPUT OUT=COUNTS MEAN=;

  DATA COUNTS2;
    SET COUNTS;
    BY DIVISION GRADE MIDPOINT;
    IF FIRST.MIDPOINT THEN DO;
       MIN=0; NONMIN=0; FEM=0; MAL=0;
       MINSAL=0; NONMSAL=0; FEMSAL=0;
MALSAL=0;
    END;
    IF _TYPE_=1 THEN DO;
       IF RACE=0 THEN MIN+_FREQ_;
       IF RACE=0 THEN MINSAL+SALARY;
       IF RACE=1 THEN NONMIN+_FREQ_;
       IF RACE=1 THEN NONMSAL+SALARY;
  END;
    IF _TYPE_=2 THEN DO;
     IF GENDER=0 THEN FEM+_FREQ_;
     IF GENDER=0 THEN FEMSAL+SALARY;
     IF GENDER=1 THEN MAL+_FREQ_;
     IF GENDER=1 THEN MALSAL+SALARY;
   END;
   IF LAST.MIDPOINT;

  * ** NOW CALCULATE RACE AND GENDER
EXPOSURE ***;

  DATA  REPORT;
    MERGE  GLMOUT2(DROP=_TYPE_)
COUNTS2(DROP=_TYPE_ _FREQ_)
           LSMOUT2(KEEP=DIVISION GRADE
                             MIDPOINT MINORITY NONMI MALE
FEMALE);
    BY DIVISION GRADE MIDPOINT;
    IF .<SEX<.05 THEN DO;
       IF MALE<FEMALE THEN
GEXPOSE=MAL*(FEMSAL-MALSAL);
       ELSE GEXPOSE=FEM*(MALSAL-FEMSAL);
    END;
    IF .<RACE<.05 THEN DO;
       IF MI<NONMI THEN REXPOSE=MIN*(NONMSAL-
MINSAL);
       ELSE REXPOSE=NONMIN*(MINSAL-NONMSAL);
    END;
```

Figure 2 – Add $10 until no longer significant

```
  %GLOBAL INCR;
  %MACRO ADD(VAR,VAL,DIV,GRADE,MIDP,NUM);
  DATA ADD&NUM;
    SET  FILE;
    LENGTH DIV $ 5;
    IF ORGN1=&DIV AND PYGRD=&GRADE AND
MIDPOINT=&MIDP;
    DIV=PUT(DIVISION,DIV.);
    GR=PUT(GRADE,GRADEF.);
    INCR=0;
    DROP DIVISION GRADE;

  %LET FLAG=N;
  %DO %UNTIL (&FLAG=Y);
  DATA ADD&NUM;
    SET ADD&NUM;
    IF &VAR=&VAL THEN DO;
       ABSALY=ABSALY+10;
       INCR=INCR+10;
       CALL SYMPUT('INCR',INCR);
  END;
  RUN;

  PROC GLM NOPRINT OUTSTAT=GLMOUT;
    CLASS GENDER MINORITY;
    MODEL ABSALY=GENDER MINORITY
GENDER*MINORITY
              ACHEV EXCEL OTHER PROMO
              SERVICE TIP TIG / SOLUTION;
  FORMAT DIVISION DIV. GRADE GRADEF.;

  DATA GLMOUT2;
    SET GLMOUT;
    IF _TYPE_='SS3' AND _SOURCE_="&VAR";
    PUT PROB=;
    IF PROB>.05 THEN DO;
       CALL SYMPUT('FLAG','Y');
    END;
  RUN;

  %END;
  DATA GLM&NUM;
    LENGTH DIVISION $ 5;
    SET GLMOUT END=EOF;
    IF _TYPE_='SS3';
    IF _SOURCE_='GENDER' THEN SEX+PROB;
    IF _SOURCE_='RACE' THEN RACE+PROB;
    IF _SOURCE_='GENDER*RACE' THEN
INTERACT+PROB;
    IF _SOURCE_='AGE' THEN AGE+PROB;
    IF _SOURCE_='SERVICE' THEN SERVICE+PROB;
    IF _SOURCE_='TIG’ THEN TIG+PROB; * Time in
grade;
    IF _SOURCE_='TIP' THEN TIP+PROB; * Time in
position;
    DIVISION=&DIV;
    GRADE=&GRADE;
    MIDPOINT=&MIDP;
    INCR=SYMGET('INCR');
  TYPE="&VAR";
    IF EOF;
  %MEND;
```