Paper 260-25

Product Development and Applied Research: A Formalized Approach Using SAS® and JMP®

John A. Wass, Abbott Laboratories, Abbott Park, IL

Abstract

In a competitive, industrial environment it is imperative that products be brought to market with minimal expenditure of resources and maximal robustness. This may be accomplished with statistical consultation, but increasingly utilizes extensive input from the bench scientist. This paper delineates a simple, formalized approach that utilizes the simplicity and widespread use of JMP for exploratory analysis and initial identification of important factors. This is followed by a more detailed analysis in SAS which will characterize main effects and interactions, and assists in the elucidation of a final model that will most closely reflect physical reality.

In the example discussed, a detailed model of an immunochemical interaction is developed from an initial screening design in JMP. The product, a reagent system for measurement of an HIV serotype, is developed in SAS from this initial analysis. Specifically, the roles of the main effects are reexamined and re-evaluated, the role of interaction is extensively probed, and a variance component analysis is performed. By utilizing the more powerful and flexible SAS PROC's such as REG, STEPWISE, MIXED, and GLM, a final model is constructed that will supply the chemist and life scientist with an accurate reflection of the true physical process.

Introduction

This paper describes a formal approach to developing immunochemical reagents using elements of SAS/STAT and JMP. The objective is to develop a series of statistical tests that will lead the applied scientist to those input parameters that may be most useful in targeting values and properties of the final output. In this case, the product is an HIV reagent used on an automated immunochemistry analyzer. The input parameters include elements of the chemical and physical processes that make up the final master lot reagent pack. By utilizing the richly graphic, exploratory data capabilities of JMP in conjunction with the extensive diagnostic capabilities of SAS, the consulting statistician can formulate a system whereby the more important input parameters are identified and product variability is minimized by use of this information.

Description of the Process

The input parameters are listed in Table 1. with brief descriptions of functionality. The underlying process is to coat plastic beads with a series of antigens, small protein units that will react with antibodies to HIV, circulating in the blood of infected patients. The antigen coating is a muti-step process and involves the addition of several antigens to the beads, changes of buffering agents, and varying incubation times and temperatures. The coated bead will ultimately reside in a reagent pack that will be mixed with a small sample of patient sera. If HIV antibodies are present in this sera, binding to the antigen coated, fluorescently- conjugated bead will occur and finally, a resultant fluorescent reaction will occur and its intensity measured by clinical instrumentation. In the specific case described, this intensity can be directly measured as relative light units (RLU's) or as a processed signal (s/co).

Table 1. Process Inputs

Abbreviation	Description
Day	designator for day of experiment
Inst	designator for specific instrument used
Run	experimental unit of basic mixture; each new preparation of reagents defines a single run

% solids	concentration of beads
Trig. Inc.	amount of chemical trigger in the incubation mixture
Ag1-Ag4	specific antigens for various HIV groups and subtypes
Between	time between sequential addition of specific proteins
Coating	time of incubation after protein addition
Heat Temp	temperature of activation
Heat Time	incubation time at the activation temperature
Buffer1-Buffer3	number of buffer changes for each buffer
pH	final pH of combined reagents

The outputs for this process are the readings for two specific HIV types, each as RLU or s/co. They are designated as follows: HIV1s, HIV1rlu, HIV2s, and HIV2RLU.

Preliminary Analysis With JMP

The first step is to plot the output data to get an idea of their distributions and the occurrence of any outliers or trending. This is easily accomplished in JMP with the "Distribution of Y" button.

The next step includes a small internal control. As the results between the same samples should correlate between types of data reads, i.e., between the RLU's and the s/co's a quick correlation may be done with the "Correlation of Y's" button.

The final probe is a brief screening analysis, performed to reduce the number of significant input variables for the more detailed analyses in SAS. This will assist in both statistically simplifying the problem and identifying those variables actually driving the reaction. With JMP, the "Fit Model" button is used and a screening model with main effects only is generated. Although the importance of interaction is recognized with immunochemical reactions, statistically we will not pursue them in JMP as i) there are too few degrees of freedom to estimate all of the main effects plus interactions in a design of this size and ii) only a preliminary estimate of importance is desired at this stage. The point is to highlight those main effects that are important and to later consider interactions using only those main effects that are either statistically or chemically significant.

Extended Analysis with SAS

Based upon the results of the preliminary JMP analysis, outliers, distributions and significant main effects have been identified. These results may be quickly confirmed and extended in PROC UNIVARIATE. We look to see if the moments, quantiles and extremes as well as the simple stem-and-leaf and boxplot diagrams for the HIV1 s/co show close agreement with the previous description. Confirmations such as these are often performed for cross-platform software validation and are becoming more important when assembling documentation.

The analyst's task now shifts to the detailed analysis which includes modeling, variance component analysis and possible covariate effects. Depending upon the data, simple linear regression(SLR), multivariate regression, polynomial regression or nonlinear regression may be utilized. In many cases SLR may suffice and is the first model attempted. Experience with this regression demonstrates a utility for the immunochemical reactions of the type dealt with in this paper.

For multivariate regression the general linear model is utilized:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

where y, β , X, and ε denote the response variable, the unknown parameters , the independent (input) variables, and the random error, respectively. The SLR will assist in obtaining least squares estimates of the parameters, variance associated with error, standard error of parameter estimates, predict values of the dependent variable, and evaluate the model lack of fit. Prior to the detailed analysis, results from the JMP and regression analysis are used to confirm the a priori

assumptions for this model, i.e., error expected values are zero, existence of homoscadasticity, and that the errors are uncorrelated and normally distributed. When this is completed, the analysis is performed in PROC REG. The following abbreviations are used: antigen1 - a, antigen2 - b, antigen3 - c, antigen4 - d, buffer1 - aex, buffer2 - bex, and buffer3 - cex. The initial model will contain only the main effects. The chemists are then consulted for the important interactions and these are modeled in the second pass. It is known from experience that at least several of these interactions are important, and main effects that are initially deemed (statistically) unimportant in a "main effects only model" may be very important in an interactive effect. The code for these first steps are:

```
Data hiv;
Input day inst run solids trigger
a b c d betwen coat htemp
 htime aex bex cex ph hiv1s hiv2s
hiv1rlu hiv2rlu;

ac=a*c;
abet=a*betwen;
acex=a*cex;
aph=a*ph;
cbet=c*betwen;
ccex=c*cex;
cph=c*ph;
betcex=betwen*cex;
betph=betwen*ph;
cexph=cex*ph;
cards;
```

(data)

```
proc univariate data=hiv plot;
/* Descriptive Statistics */
 var hiv1s hiv2s hiv1rlu hiv2rlu;
 title 'Summary of HIV Data';
run;


/* Simple Linear Regression of */
proc reg data=hiv;
/* main effects only           */
 model hiv1s= day run a c betwen
ph / r influence acov;
 title 'HIV1 S/CO Partial
Regression';
       run;

/* Simple Linear Regression of */
/* the full model              */
proc reg data=hiv;
 model hiv1s= day run a c betwen
ph ac abet acex aph cbet ccex cph
betcex betph cexph
 / r influence acov;
 title 'HIV1 S/CO Full
Regression';
run;
```

There are a number of excellent diagnostics available with PROC REG and several are chosen for their specific information: "r" will yield standard errors for the predicted and residual values, the Studentized residuals, and Cook's D statistic which is a measure of the influence of individual observations upon the parameter estimates; "influence" will give the influence of each observation on the predicted values (among other things); "acov" yields the estimated asymptotic covariance matrix of the parameter estimates under heteroscedasticity. These estimates give the analyst a preliminary snapshot of model adequacy and may indicate problems with the data or form of the model.

Noting that output from a SLR may yield low r-squared, high error and misleading parameter prob-values due to mispecification of the model, further consultation is undertaken with the chemists. On the basis of discussion with staff scientists, the analyst will concentrate on the main effects plus those interactions that are believed to i) be of chemical import to the reaction and ii) be of interest to the chemist as a possibly important reaction. To refine the analysis, a stepwise regression is performed to further narrow the field of main and interactive effects as well as to discover any hitherto unsuspected important side reactions:

```
/* Stepwise regression of full */

/* model from proc reg         */
proc stepwise  data=hiv;
  model hiv1s = run day a c
betwen cex ph ac abet acex aph
cbet ccex cph betcex betph cexph
  / stepwise backward sle=0.05
sls=0.05;
  title 'HIV1 S/CO Stepwise
Regression';
```

```
run;

proc reg outest=est;
           model hiv1s = run day a
       c betwen cex ph ac abet
       acex aph cbet ccex cph
       betcex betph cexph
           / selection = rsquare
       cp best=3;
run;

/* Mallow's Cp Plot */
proc plot;
 plot _cp_*_in_ = 'C' _p_*_in_ =
'*' / overlay
  vaxis = -2 to 23 by 1 haxis = 0
to 18 by 1
  hpos = 40 vpos = 30;
 run;
```

The above code will test those effects deemed important to the physical model by i) starting with a full model and subtracting terms one-at-a-time and re-testing for significance of factors left (backward), and ii) building the model one-factor-at-a-time and again, retesting for significance after each variable entry. The entrance and exit alpha's are both set to 0.5 as a conservative first step. These values may be changed based upon agreement between the bench scientist and statistician concerning the significance of known physical factors based upon previous testing of similar chemical systems.

The second procedure, "outest", will generate an output data set that may be utilized to select a presumptive model which can be evaluated on the basis of both rsquare and Mallow's Cp, the latter will be displayed graphically by PROC PLOT. The "rsquare" and "best=n " options will output a compact table of the best n models for each subset of data, best in this case denoting minimal error mean squares. Mallow's Cp is a measure of total squared error in a model and has the useful features of indicating error variance plus bias introduced by omitting important variables in the model. It is thus an excellent indicator of over-trimming of independent variables. In product development, we wish to err (at least initially) on the conservative side and not omit the important.

Upon preliminary inspection of the resultant table and plot, the “best” n-factor model is chosen based upon maximal rsquare and minimal error. It may be possible chemically, and desirable statistically, to further eliminate terms and simplify the model. This is only done in consultation with the chemist however, and must take into account a balance between reduction in rsquare/increase in error and true elimination of less significant events. From the statistical side, further investigations can be undertaken to examine such diagnostics as the variance inflation factor.

The final steps are to examine the variance components and, in isolating an important interaction, possible covariates of this factor. This is implemented by the following:

```
/* variance component analysis */
proc mixed data=hiv;
 class run day a c betwen cex ph;
 model hiv1s= ;
 random    day run a c betwen ph
ac abet acex aph cbet ccex cph
betcex betph cexph;
run;


proc glm data=hiv;
/* ANCOVA */
 class run;
 model hiv1s = run run*day
run*betwen / solution ss4;
run;
```

It is readily apparent that with small error estimates, we may not wish to ascribe meaning to these variance components without further investigation. The ANCOVA, done with PROC GLM and utilizing the type IV error ss, indicates which factors are significantly effected by any of the other independent variables. Again, this may be taken as an additional indicator of effects requiring further study.

Conclusion

The above represents a useful template for determining important model effects, testing a model, and determining sources of variability. It

can be generalized, with sufficient chemical and statistical background, to many other systems.

Contact Information

John A. Wass, Ph.D.
Abbott Laboratories
D7CE AP31/ 4th fl
200 Abbott Park Rd.
Abbott Park, IL 60064-6199

Ph: (847) 938-3675
Fax: (847) 935-2433
Email: John.Wass@add.ssw.abbott.com