

Analysis of Covariance: A Useful Tool For Investigating Molecular Interaction

John A. Wass

Abbott Laboratories

Abbott Park, IL 60045-3500

Introduction

Analysis of covariance (ANCOVA) is a technique whereby many regression surfaces or lines, corresponding to treatments or treatment combinations, may be compared. In this case there may be a different regression surface or line for each treatment or combination of treatments. The technique is actually a combination of analysis of variance and regression and is very useful in adjusting an analysis for those variables that could not or are not controlled by the experimenter. The experiment is designed to allow a certain number of error degrees of freedom (df). This is achieved through the covariates. The experimenter in turn, wishes to trade some of those df for covariates which will help reduce the magnitude of the error variance. The experiment usually involves randomly assigning n_i experimental units to any treatment i , then applying the treatments and measuring the dependent variables (y_{ij}) for any given group of independent variables or covariates ($x_{1ij}, \dots, x_{kij}$). For the elementary case, the covariance model consists of a linear function of the covariates: $y_{ij} = b_{0i} + b_{1i}x_{1ij} + b_{2i}x_{2ij} + \dots + b_{ki}x_{kij} + e_{ij}$.

In this example ANCOVA is utilized to examine the interaction of four theoretical molecules as they bind and yield a measurable reaction. By adding the regressor terms and comparing slopes, several interesting and possibly unexpected results are obtained. Specifically, by adjusting concentrations of these molecules, previously unsuspected relationships are uncovered and molecules thought to play a minor role assume greater importance to the end reaction.

The calculations are easily implemented in SAS but interpretation of the results requires knowledge of both the statistics involved as well as the underlying chemistry. By utilizing the class and model statements within PROC GLM the data is readily classified and subsequently operated upon. Parameter estimates as well as comparison of slopes is readily accomplished with LSMEANS and ESTIMATE statements. Graphic display of the results may be visualized in the GPLOT procedure. For the model used herein, increased importance to the reaction is ascribed to a molecule that was believed to play a lesser role while the importance of the primary molecule was slightly reduced.

Importance of the Approach

Herein is reported a simple, stepwise method for determining 1) whether molecular interaction exists in any measured system, 2) if so, the extent of the interaction and, 3) the significance of the interaction. This is accomplished by the statistical technique of ANCOVA implemented in SAS.

The importance of this method lies in its generalizability, i.e., we observe this system almost as a black box. The molecular reaction is

analyzed as if we knew neither the nature of the molecules themselves nor the nature of the reaction. The analysis is done wholly from a statistical standpoint. The data is from routine clinical chemistry experiments in the development of reagent systems. This data is analyzed in a modified form to allow comparisons of the different reagents and their interactions.

Statement of the Problem

In a system of 4 distinct molecular reactants and one measured output (the amount of product) we ask 1) does any individual molecule have an effect on the amount of product formed? 2) does the concentration of a molecule have an effect on the amount of product formed? and 3) is there a significant difference in the amount of product formed/molecule relationship for different values of the concentration?

Strategy

In this example we will begin with the simplest situation, a one-way analysis of covariance in a completely randomized design. We realize that it is important to note that the mean of an observation from the i th treatment group depends both upon the main variable as well as the covariate. The model is: $a_i + b_jx_{ij} + e_{ij}$ where $i = 1, 2, \dots, k$ $j = 1, 2, \dots, n$ the mean of y_i for any given x is $m_{y|x} = a_i + b_{ix}$ and it is assumed that $e_{ij} \sim \text{iid } N(0, s^2)$. Prior to the actual analysis we will verify that the data can be adequately described by a simple linear regression model. At this time regression diagnostics are run and the assumption of equal variance is checked. The ANCOVA techniques involve comparing the k slopes and distances between the regression lines at various values of x . These preselected values are chosen by the experimenter and represent areas of interest along the line. For the actual steps the eventual strategy (when we have many covariates) is taken from Milliken and Johnson (1) and consists of i) assuming that a simple linear regression model will describe the data, ii) test the hypothesis that the slopes are all 0, iii) assuming ii) is rejected, test the hypothesis that the slopes are equal, iv) if iii) is rejected, use the unequal slope model to determine if the slopes can be grouped and construct confidence bands around the difference of selected pairs of models.

Once this simple case is completed, we may add other covariates and recalculate the process to quantify and interpret their contributions to variability.

Covariance: Summary of Hypothesis Tests

H_{01} : Do all slopes = 0 ?

Test in SAS with: Type III (or IV) sum of square corresponding to $X^* \text{Molec}$

Test in SAS with: Type III sum of squares corresponding to *Molec

H_{03} : If the slopes are equal, are all intercepts equal?

Test in SAS with: Type III sum of squares corresponding to Molec

H_{04} : If the slopes are NOT equal, are the adjusted means (of the treatments) equal?

Test in SAS with: PDIF option with LSMEANS

H_{05} : If the slopes are NOT equal, are all of the intercepts equal?

Test in SAS with: LSMEANS and Estimate statements with x=0

H_{06} : If the slopes are NOT equal, are the distances between the lines at selected values of x comparable?

Test in SAS with: adjust= Bon or LSD option with LSMEANS plus Estimate statements

SAS Code for Covariance

```
data covar;
input molec x y ;
datalines;
1 1.87 1.88
1 1.85 1.86
1 1.82 1.84
1 1.85 1.84
1 1.89 1.89
. . .
2 19.30 19.35
2 19.27 19.19
2 19.41 19.11
2 19.27 19.37
2 19.30 19.39
. . .
3 18.99 18.95
3 19.36 18.71
3 19.22 18.65
3 19.18 37.03
3 21.22 35.67
. . .
4 99.20 102.00
4 96.8067 96.58
```

```
4 96.3337 95.62
4 96.3552 96.78
4 95.8321 96.34
```

```
proc glm;
class molec;
model y=x molec x*molec / solution;
lsmeans molec / stderr pdiff adjust=bon;
run;
```

Note: Molec signifies the type of molecule added
X is the concentration of Molec
Y is the measured product (may be concentration or rate of reaction)

Results

1) Without Interaction

Dependent Variable: Y

Source	DF	Sum of Squares
Model	4	754965.06599156
Error	115	28630.19144094
Corrected Total	119	783595.25743250
R-Square		0.963463

Mean Square	F Value	Pr > F
188741.26649789	758.12	0.0001
248.95818644		

Source	DF	Type I SS
MOLEC	3	653025.26306917
X	1	101939.80292239

Source	DF	Type III SS
MOLEC	3	1292.12421411
X	1	101939.80292239

Mean Square	F Value	Pr > F
217675.08768972	874.34	0.0001
101939.8029239	409.47	0.0001

Mean Square	F Value	Pr > F
430.70807137	1.73	0.1647
101939.80292239	409.47	0.0001

Parameter	Estimate
INTERCEPT	8.810516537 B
MOLEC	-8.505301047 B
	0.760148675 B
	-4.928245786 B
	0.000000000 B
X	0.970825886

NOTE: The XX matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

2) With Interaction

Source	DF	Type I SS
X	1	753672.94177745
MOLEC	3	1292.12421411
X*MOLEC	3	814.75017882

Source	DF	Type III SS
X	1	3797.26714766
MOLEC	3	1292.12421411
X*MOLEC	3	814.75017882

Mean Square	F Value	Pr > F
753672.94177745	3034.69	0.0001
430.70807137	1.73	0.1641
271.58339294	1.09	0.3550

Mean Square	F Value	Pr > F
3797.26714766	15.29	0.0002
305.83912410	1.23	0.3017
271.58339294	1.09	0.3550

Parameter	Estimate
INTERCEPT	10.63346002 B
X	0.96073406 B
MOLEC	1
	2
	3
	4
	0.00000000 B
X*MOLEC	1
	2
	3
	4
	0.00000000 B

NOTE: The XX matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

Interpretation

The initial analysis did not include any interactive effects. The Type I sum of squares for molec yields the between-molecule sum of squares that would be obtained for the analysis of variance model $y = \text{molec}$ and the p-value indicates a significant effect. However, the Type III sum of squares for molec yields the molecule sum of squares

adjusted for the covariate x (concentration of the molecule) and is not significant.

When an interaction term is added, although the effect of the molec term is still not significant, the parameter estimate for molecule 3 relative to the other molecules increases significantly. This is indicative of a greater effect upon the overall reaction.

Conclusions

Covariance analysis is a useful way to control or "allow for" those variables not under the explicit control of the experimenter. By comparing regression surfaces or lines, specifically slopes, intercepts and the significance of parameter estimates, the analyst can reasonably adjust least squares means for treatment groups and compare them.

Testing for heterogeneity of slopes is a natural extension of this process. These types of analyses are usually applied to data that have treatment groups and one or more covariates. Interactions between the treatment groups and covariates are reflected by the regression relationships that will differ among treatment groups. In SAS, the GLM procedure will specify and analyze these relations as an interaction (3).

In this example a simple chemical system is analyzed purely from a mathematical standpoint. The initial black box relationship is transformed into useful information about the system, based on real data, and demonstrates the necessity of following an analytic sequence that allows for interactive effects and adjustments for any covariates. This simple procedure forms the basis of a statistical strategy that may be used on more complex systems where the chemistry is at least partially unknown.

References

- 1) GA Milliken and DE Johnson. Analysis of Messy Data Vol. III: Analysis of Covariance. (In Press).
- 2) SAS/STAT[®] User's Guide, Vol. 2: GLM - VARCOMP. Version 6 Fourth Edition. SAS Institute, Inc., Cary, NC (1990).
- 3) RC Littell, RJ Freund and PC Spector. SAS System for Linear Models, 3rd Ed. SAS Institute, Inc., Cary, NC (1991).

Author Information:

John A. Wass, Ph.D.
Dept. 7CE, AP32
Abbott Laboratories
100 Abbott Park Rd.
Abbott Park, IL 60045-3500