



Business Intelligence
Solutions

Head Office

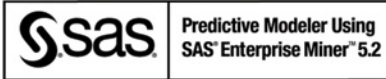
T: 267.616.1444
F: 1.905.761.1006
9 Flagstaff Place
Philadelphia, PA
19115 USA

info@bisolutions.us

Toronto

T: 647.271.1932
F: 905.761.1006
150 Borrow's Street
Thornhill, ON
L4J 2W8 Canada

www.bisolutions.us



www.BISolutions.us | Data. Knowledge. Action.

WHITE PAPER:

**JOINT REGRESSION MODELS FOR SALES ANALYSIS
USING SAS**

JUNE 2008

Eugene Brusilovskiy & Dmitry Brusilovsky



Introduction

Assume that a company generates significant profit by selling service A and service B. A customer can either buy one of these services, or both of them.

We will call the customer that purchases only service A an *A_Purchaser* and the customer that purchases only service B a *B_Purchaser*. Because customers can purchase both services, we have two correlated dependent variables: \$ sales of service A (*Service_A_Sales*), and \$ sales of service B (*Service_B_Sales*). We also assume that the list of drivers of each service is not identical, and the contribution of the two services to company profit generation is significantly different. Thus, it is inadequate to create one dependent variable by summing or otherwise combining these two variables. Yet, because the sales of both services are positively related, we need to find an appropriate approach to analyze the joint distributions of these two variables. Therefore, it is natural to consider the sales analysis as a bivariate regression problem.

Objective

The objective of this white paper is to find an acceptable approach to analyze sales data, using the SAS software (including SAS Enterprise Miner (EM)), taking into account the specificity of sales data and the approach assumptions. An adequate model will help to develop a strategy to increase sales.

Problem Statement

Four different formulations of bivariate regression sales analysis make perfect business sense:

1. Both sales variables, *Service_A_Sales* and *Service_B_Sales*, are treated as continuous variables
2. The first dependent variable, *Service_A_Sales*, is considered to be continuous, but the other, *Service_B_Sales*, is dichotomized. As a result, each *B_customer* is treated as a *B_purchaser* (if $Service_B_Sales > 0$) or *B_non-purchaser* (if $Service_B_Sales = 0$).
3. The reverse situation also makes sense: the first dependent variable, *Service_A_Sales*, is dichotomized, but the second one, *Service_B_Sales*, is treated as a continuous variable. As a result, each *A_customer* is treated as an *A_purchaser* (if $Service_A_Sales > 0$) or *A_non-purchaser* (if $Service_A_Sales = 0$).
4. Both sales variables can be treated as dichotomous (i.e., purchaser/non-purchaser).

We would like to find an adequate approach to model the sales described by two dependent variables, where each is a function of demographic, psychographic, geographic and other customer attributes.



Data Structure and Model Selection

Existing data can be in the time series, or cross-sectional, or cross sectional time series format. Based on the data format, data availability and other data properties, the appropriate model can be:

- Linear vs. Non-linear
- Parametric vs. Semi-parametric vs. Non-parametric
- Fixed effect vs. Random effect vs. Mixed effect
- Explanatory vs. Predictive

Properties of Hypothetical Data

We will consider two different hypothetical datasets. The first dataset meets all the assumptions and requirements of traditional statistical analysis. Namely, both dependent variables (*Service_A_Sales* and *Service_B_Sales*) are normally distributed, there are no missing values and outliers, the number of categorical variables (and the number of distinct categories) is small, there is no severe multicollinearity among the predictors, and the relationship between dependent and independent variables is well understood.

In the second dataset, both dependent variables have very skewed distributions. The dimensionality of the predictor space is high, and predictors are differently scaled. More than half of the predictors are categorical, and some of them have many of categories. The set of interval-scaled predictors includes several redundant variables, i.e., there is severe multicollinearity in the data. Besides, some predictors have a high percentage of outliers and missing values. The relationship between dependent and independent variables is not well understood, and it is likely that some of the independent variables are irrelevant.

Possible Approaches for Joint Modeling

Below, we present a number of approaches that can be used for joint modeling and give a brief overview of their advantages and limitations.

WITHIN SAS/STAT:

Multivariate regression, using PROC REG and Multivariate Analysis of Covariance (MANCOVA), using PROC GLM [1] assume that both dependent variables are continuous and normally distributed, observations are independent, and the relationship between dependents and independents is linear. For MANCOVA the assumption of homogeneity of variances and covariances is also important. These methods are highly sensitive to the presence of outliers and multicollinearity.

Structural Equation Modeling (Multivariate Regression) with PROC CALIS [2], assumes that both dependent variables are continuous.



Procedures REG, GLM and CALIS can be applied for the first hypothetical dataset to identify the significant predictors of sales, assuming that the relationship between dependents and independents is linear

Generalized Linear Mixed Models, using PROC GLIMMIX [3] can be employed when one dependent variable is binary and the other is interval.

In some cases *Mixed Models (PROC MIXED)* can be used too. There are examples of random-effect bivariate linear mixed models for longitudinal data [4, 5].

Partial Least Squares (PLS) Regression with PROC PLS [6]. In principle, any combination of different measurement scales for dependents and independents can be handled by PLS. PROC PLS is tolerant to outliers and relaxes the normality, observation independence and no multicollinearity assumptions. Therefore, PLS regression is a good choice for the second dataset.

Bivariate Logit with PROC NLMIXED [7], can be used when both dependent variables are binary. Different approaches to bivariate binary response models are described in [8, 9].

WITHIN SAS/ETS:

Bivariate Probit with PROC QLIM [10] is appropriate when both dependent variables are binary.

However, when both dependent variables are continuous, the *vector autoregressive model with exogenous variables, (VARX(p,s) modeling) with PROC VARMAX* [11] can be employed. Here, p designates the order of vector autocorrelation process, and s designates the maximum lag of the exogenous variable. This approach implies multivariate normality of residuals, homoskedasticity, stationarity of the vector time series (*Service_A_Sale, Service_B_Sales*), and the absence of correlation between residuals and exogenous variables.

Categorical variables with a large number of categories also are a problem for *VARX(p,s)* modeling. The length of time series data should be large enough in order to get reliable parameter estimates. The approach is not appropriate for the second dataset.



WITHIN SAS/EM:

It is possible to employ the *Two Stage Modeling Node*, when one dependent variable is binary and the other is interval

Multilayer Perceptron Modeling with Neural Net Node [12] is appropriate for any formulation of bivariate regression analysis.

These approaches imply observation independence and low multicollinearity and produce non-linear and non-parametric models. Therefore, they are acceptable for the analysis of the second data set. Unfortunately, the models are non-interpretable, and as a rule, it is difficult to make any business implications.

CHOOSING THE RIGHT APPROACH

Each approach has its own pros and cons, and applicability restrictions.

The selection of the “best” approach depends on the criteria used. Model performance (testing the prediction accuracy on validation data) and interpretability, as well as the actionability of the findings are important criteria. If the non-linearity of the relationship between dependents and independents is essential, then only the SAS EM approach (non-parametric and non-linear regression) can be adequate.

Conclusion

There are a number of widely used methods to analyze the formulated sales analysis problem, and trying several of them is usually appropriate. It is important to remember that a tradeoff often needs to be made between the assumption violations, the interpretability of parameter estimates and findings, and the business value of the outcome for a decision maker. Therefore, one approach might involve ignoring the non-linear relationship between the variables and the mild violations of assumptions to gain interpretability; another might entail foregoing interpretability to gain accuracy.



References

1. Gregory Carey (1998), *Multivariate Analysis of Variance (MANOVA) II: Practical Guide to ANOVA and MANOVA for SAS*,
<http://ibgwww.colorado.edu/~carey/p7291dir/handouts/manova2.pdf>
2. The CALIS procedure. Available at
<http://www.vub.ac.be/BFUCC/sas/sasdoc/stat/chap19/index.htm>
3. The GLIMMIX procedure. Available at
<http://support.sas.com/rnd/app/papers/glimmix.pdf>
4. Rodolphe Thiébaud, Hélène Jacquemin-Gadda, Geneviève Chêne, Catherine Leport and Daniel Commenges (2002), *Bivariate Linear Mixed Models Using SAS proc MIXED*. Available at
<http://arxiv.org/ftp/arxiv/papers/0705/0705.0568.pdf>
5. 5. Jabu S. Sithole and Peter W. Jones (2007), Bivariate Longitudinal Model for Detecting Prescribing Change in Two Drugs Simultaneously with Correlated Errors, *Journal of Applied Statistics*, 34 (3): 339 – 352.
6. The NLMIXED procedure. Available at
<http://www.technion.ac.il/docs/sas/stat/chap46/index.htm>
7. Paul Dickman, Annica Dominicus and Juni Palgram (2003), *Modelling bivariate binary responses with application to twin data*. Available at
http://www.pauldickman.com/teaching/seminar030515_8.pdf
8. Thomas R. Ten Have and Alfredo Morabia (1999), Mixed Effects Models with Bivariate and Univariate Association Parameters for Longitudinal Bivariate Binary Response Data, *Biometrics*, 55 (1): 85-93 (9).
9. Randall D. Tobias (2002), *An Introduction to Partial Least Squares Regression*, SAS Institute Inc., Cary, NC. Available at
<http://support.sas.com/techsup/technote/ts509.pdf>
10. The QLIM procedure. Available at
<http://support.sas.com/rnd/app/papers/qlim.pdf>
11. The VARMAX procedure. Available at
<http://support.sas.com/rnd/app/da/new/801ce/ets/chap4/index.htm>
12. Randall Matignon (2005), *Neural Network Modeling using SAS Enterprise Miner*. Published by SAS Institute Inc.