USE OF MULTIVARIATE EXTENSIONS OF GENERALIZED LINEAR MODELS IN THE ANALYSIS OF DATA FROM CLINICAL TRIALS

Ariel Alonso Abad[1], Olga María Rodríguez[1], Fabian Tibaldy[2] y José Cortinas Abrahantes[3]
[1]Centro Nacional Coordinador de Ensayos Clínicos, Cuba
[2]Limburgs Universitair Centrum, Belgium
[3]Centro de Sanidad Vegetal, Cuba

ABSTRACT
In medical studies the categorical endpoints are quite often. Even though nowdays some models for handling this multicategorical variables have been developed their use is not common. This work shows an application of the Multivariate Generalized Linear Models to the analysis of Clinical Trials data. After a theoretical introduction models for ordinal and nominal responses are applied and the main results are discussed.

Key words: multivariate analysis, multivariate logistic regression, multicategorical response.

RESUMEN
Las variables de respuesta categóricas son muy utilizadas en el marco de las investigaciones Biomédicas. A pesar de que varios modelos para el análisis de este tipo de variables han sido propuestos en la literatura su uso es aun muy infrecuente. El presente trabajo muestra una aplicación de los Modelos Lineales Generalizados Multivariados a los datos de un ensayo clínico internacional. Después de una introducción teórica, modelos multivariados para el análisis de variables categóricas nominales y ordinales son aplicados a los datos y los resultados son interpretados.

Palabras clave: análisis multivariado, regresión logística multivariada y respuestas multicategóricas.

MSC 62P10

1. INTRODUCTION TO MULTIVARIATE GENERALIZED LINEAR MODELS

Multinomial response models can be considered as special cases of multivariate generalized linear models. In analogy to the univariate case, multivariate generalized linear models are based on both distributional and structural assumptions. However, the response variable y_i is now a q-dimensional vector with expectation $\mu_i = E(y_i \mid x_i)$.

1. Distributional assumptions:

Given x_i, the y_i's are (conditionally) independent and have a distribution that belongs to a simple exponential family, which has the form

$$f(y_i \mid \theta_i, \phi, \omega_i) = \exp\left\{\frac{[y_i'\theta_i - b(\theta_i)]}{\phi}\,\omega_i + c(y_i, \phi, \omega_i)\right\} \tag{1}$$

2. Structural Assumptions:

The expectation μ_i is determined by a linear predictor

$$\eta_i = Z_i\beta \tag{2}$$

of the form

$$\mu_i = h(\eta_i) = h(Z_i\beta) \tag{3}$$

Let the response variable Y have possible values 1,…, k; where the numbers are mere labels for the categories, for example, neither ordering nor differences between the category numbers is meaningful. The categories refer to the several alternatives. Sometimes consideration of Y can take k different values, hiding the fact that we actually have a multivariate response variable. It becomes clearer by considering the response vector of the dummy variables $y = (y_1,...,y_q),\ q = k-1$

$$y_r = \begin{cases} 1 & \text{if} \quad Y = r \quad r = 1,...,q \\ 0 & \text{if} \quad \text{else} \end{cases}$$

Then we have $Y = r \Leftrightarrow y = (0,...,1,...,0)$

The probabilities are simply connected by $P(Y = r) = P(y_r = 1)$

Given m independent repetitions $y_1,\ldots,y_m$ (or equivalently $Y_1,...,Y_m$); it is useful to consider as a response variable the number of trials that yield outcome r. For the repetitions $(y_1,\ldots,y_m)$, the following sum of the vectors can be obtained

$$y = \sum_{i=1}^{m} y_i$$

Then the vector y is multinomially distributed with parameters

$$\pi_r = P(Y_i = r) \qquad i = 1,...,q.$$

The multinomial distribution of y is abbreviated by

$$y \sim M(m,\pi) \text{ where } \pi = (\pi_1,...,\pi_q)$$

In the multinomial case; $\pi_i = \mu_i = E(y_i \mid x_i)$ is a (q×1)-vector $\pi_i = (\pi_{i1},...,\pi_{iq})$ and the model defined in (3) has the form

$$\pi_i = h(Z_i\beta)$$

where h is a vector-valued response function, Z_i is a (q×p)-design matrix composed of x_i, and β is (p×1)-vector of unknown parameters.

We will consider the widely used canonical link for multicategorical response, the logit model that is given by

$$P(Y_i = r) = \frac{e^{\beta_{r0} + z_i'\beta_r}}{1 + \sum_{s=1}^{q} e^{\beta_{s0} + z_i'\beta_s}}$$

which can be written equivalently as

$$\log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + z_i'\beta_r$$

where z_i is the vector of covariables determining the log odds for category r with respect to the reference category k.

Response variables that have more than two categories often are ordinal. This implicates that the events, described by the category numbers 1,…,k can be considered ordered. In this section the following two models for ordinal responses will be discussed; the cumulative model and the cumulative logistic models or proportional odds model.

The cumulative model assumes that the observable variable Y is merely a categorized version of a latent continuous variable U. The latter is primarily used for the construction of the cumulative model. Although interpretation is simpler when the latent variable takes the model into account, interpretation is also possible without referring to the underlying continuous variable.

For a given vector x, consisting of explanatory variables; the model postulates that the observable variable $Y \in \{1,..k\}$ and that the unobservable latent variable U is connected by

$$Y = r \Leftrightarrow \theta_{r-1} < U \leq \theta_r \quad r = 1,...,k$$

where $-\infty = \theta_0 < \theta_1 ... < \theta_k = +\infty$. This means that Y is a categorized version of U $\theta_1,...,\theta_{k-1}$.

Furthermore, the model assumes that the latent variable U is determined by the explanatory variables in the linear form

$$U = -x'\gamma + \varepsilon$$

where $\gamma = (\gamma_1,...,\gamma_p)$ is a vector of coefficients and ε is a random variable with distribution function F.

From these assumptions it follows immediately that the observed variable Y is determined by the model

$$P(Y \leq r \mid x) = F(\theta_r + x'\gamma)$$

Specific choices of the distribution function lead to specific cumulative models. A common choice of the distribution function is the logistic distribution function $F(x) = \frac{1}{1+e^{-x}}$. Consequently, the cumulative logistic model has the form

$$P(Y \leq r \mid x) = \frac{e^{\theta_r + x'\gamma}}{1+e^{\theta_r + x'\gamma}} \quad \text{where } r = 1,...,q = k - 1$$

which can be written equivalently as

$$\log \frac{P(Y \leq r \mid x)}{P(Y > r \mid x)} = \theta_r + x'\gamma$$

The common parameter γ, describes the effect of x on the log odds of the response in the category r or below. In this formula γ does not have a r subscript, therefore the model assumes an identical effect of x for all q-1 collapsing of the response into binary outcomes.

2. APPLICATION OF MULTIVARIATE EXTENSIONS OF GENERALIZED LINEAR MODELS TO THE DATA FROM AN INTERNATIONAL CLINICAL TRIAL

A multi-center randomized study was performed in four different health centers to evaluate four treatments. This clinical trial was made in four different hospitals of Santiago de Chile and Valparaiso (Chile).

The treatments Placebo, drug A, drug B, and a combination of drug A and drug B were used to treat gastric-ulcers. Patients were classified at the beginning of the study as having some pain or lot of pain (initial status). Treatments were randomly assigned and administered during four weeks. The patients were classified at the end of the study in terms of their improvements in three different categories: high response (H), medium response (M) and nonresponse (NONE).

The patients included in this clinical trial were considered as a representative random sample from some corresponding large target population defined for all the possible combinations of the explanatory variables (center × initial status × treatment group). Each of the patient's response can also be assumed to be independent of other patient responses. An exploratory analysis of the data was done using graphical techniques and descriptive tables. From Figure 1 it can be seen that the treatments were presented in the four centers and within each hospital the treatments were almost balanced.

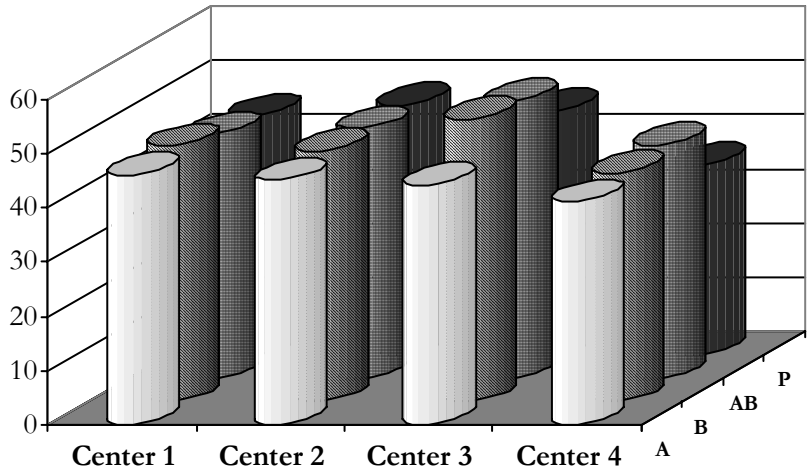

Figure 1. Distribution of patientes by treatment and center.

Figure 2 illustrates the distribution of patients taking into account the initial status and the treatment. It can be seen that the treatments were homogeneous with respect to the initial status. Moreover, the number of patients with some pain at the beginning of the study was similar for each treatment group, and the same for the other status.

Figure 2. Distribution of patients taking into account the initial status and the treatment.

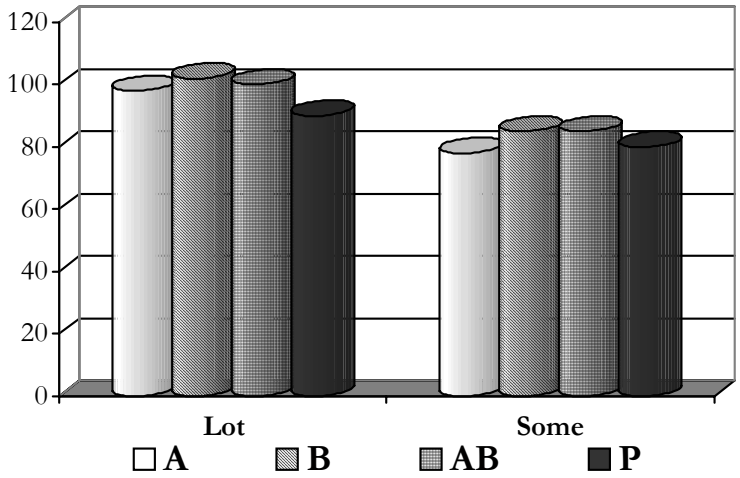

Table 1 illustrates that within each center, the best treatment was AB and the worst was A (without considering the placebo). It is remarkable that the treatment differences change among the hospitals. This fact gives us some evidence of interaction between center and treatment.

Table 1. Distribution of the response by treatment and center.

Cent.	Treatment											
	A			B			AB			Placebo		
	H	M	N	H	M	N	H	M	N	H	M	N
1	7(15.2)	22(47.8)	17(36.9)	12(25.5)	26(55.3)	9(19.1)	18(40.0)	26(57.7)	1(2.2)	9(20.4)	20(45.4)	15(34.1)
2	16(35.5)	24(53.3)	5(11.1)	23(50.0)	19(41.3)	4 (8.7)	26(56.5)	18(39.3)	2(4.3)	13(28.2)	14(30.4)	19(41.3)
3	14(31.8)	16(36.3)	14(31.8)	31(59.6)	6(11.5)	15(28.8)	30(58.8)	14(27.4)	7(13.7)	12(26.6)	17(27.7)	16(35.5)
4	23(56.1)	12(29.2)	6(14.6)	27(64.2)	14(33.3)	1(2.3)	28(65.1)	10(23.2)	5(11.6)	14(40.0)	10(28.5)	11(31.4)
Total	60	74	42	93	65	29	102	68	15	48	51	41

The numbers between brackets are the percentage calculated by treatment and center (Cent.)

As a first approach, it was decided to fit a model for nominal responses without considering the natural order of the responses. Five different models were fitted of which the result is summarized in the following table (Table 2).

Table 2. Results of the model fitting without considering the natural order of the responses.

	MODEL	LIKELIHOOD RATIO	DF	P
1	Main Effects	61.66	48	0.0891
2	Main Effects + Treat*Initial	57.33	42	0.0576
3	Main Effects + Center*Initial	55.36	42	0.0811
4	Main Effects + Center*Initial + Treat*Initial	50.88	36	0.0512
5	Main Effects + Center*Initial + Treat*Initial + Center*Treat	16.43	18	0.5624

Main Effects = Center + Initial + Treat

The interactions in Models 2,3 and 4 were not significant. Only the interaction between center and treatment was significant in Model 5. It is also important to point out that Model 5 was the only one which had a good fit. Hence, it was decided to use the following model

$$\log\frac{\text{P(High response)}}{\text{P(None response)}} = \beta_{10} + \beta_{1C}\cdot\text{Center} + \beta_{1I}\cdot\text{Initial} + \beta_{1T}\cdot\text{Treat} + \beta_{1CT}\cdot\text{Center * Treat}$$

$$\log\frac{\text{P(Medium response)}}{\text{P(None response)}} = \beta_{20} + \beta_{2C}\cdot\text{Center} + \beta_{2I}\cdot\text{Initial} + \beta_{2T}\cdot\text{Treat} + \beta_{2CT}\cdot\text{Center * Treat}$$

A summary of the results of the model fitting can be found in Table 3 and 4.

Table 3. Maximum-Likelihood Analysis-of-Variance Table.

SOURCE	DF	CHI-SQUARE	PROB
INTERCEPT	2	54.43	0.0000
CENTER	6	51.26	0.0000
TREAT	6	56.83	0.0000
INITIAL	2	47.29	0.0000
CENTER*TREAT	18	30.19	0.0356
LIKELIHOOD RATIO	30	25.94	0.6782

Table 4. Estimated coefficients and standard errors for the Multivariate Model.

	LOGIT (High/None)		LOGIT (Medium/None)	
Variable	**Coefficient**	**Standard Error**	**Coefficient**	**Standard Error**
Intercept	10.854	0.1492	0.9922	0.1504
Center	-0.4935	0.2704	0.4168	0.2426
	0.3654	0.2542	-0.5654	0.1999
	0.3658	0.2457	-0.9043	0.2068
Treatment	-0.5648	0.2185	0.2450	0.2748
	-0.2019	0.2084	12.678	0.2916
	0.5994	0.2700	0.8808	0.2942
Initial	-0.7781	0.1184	-0.3753	0.1170
Treatment*Center	-0.8093	0.3924	0.0756	0.4361
	-0.7749	0.3384	0.0885	0.5115
	-0.7403	0.4205	0.0906	0.5141
	-0.3927	0.3972	0.0407	0.3257
	12.490	0.6354	0.3193	0.3163
	11.926	0.6251	-0.3000	0.3458
	0.4622	0.3951	-11.315	0.3959
	0.5202	0.3781	-0.1867	0.3881
	-0.0866	0.4343	-0.1343	0.4035

It can be observed that the model fitted the data quite well and that the main effects were significant. An unexpected result was that the interaction between center and treatment was found to be significant. Therefore it was decided to explorer graphically the magnitude of this interaction. In Figure 3 the percentage of high and medium responses were plotted for each treatment in the different centers and the points were jointed with lines.

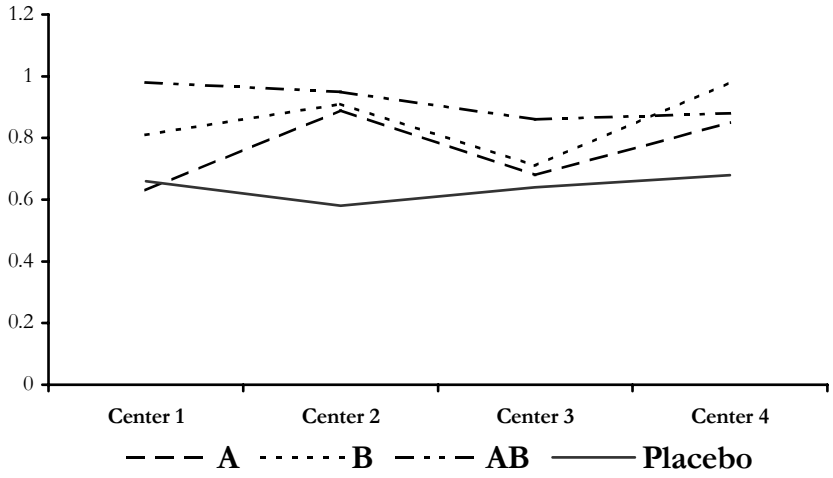

Figure 3. Interaction graph between center and treatment.

The graph illustrates a general pattern that was also observed in Figure 3; namely that treatment AB was the best in almost all the centers, followed by treatment B, treatment A and finally the placebo. Moreover there is evidence of the interaction between center and treatment. For instance, the difference between treatment B and A was much larger in Center 1 and 4 than in the other two centers. Also the difference observed between treatment A and placebo in Center 2 was large, whereas in Center 1 no difference could be found between the two treatments.

To analyze the performance of the different treatment in each of the four Centers the following graphs were made (see Figure 4):

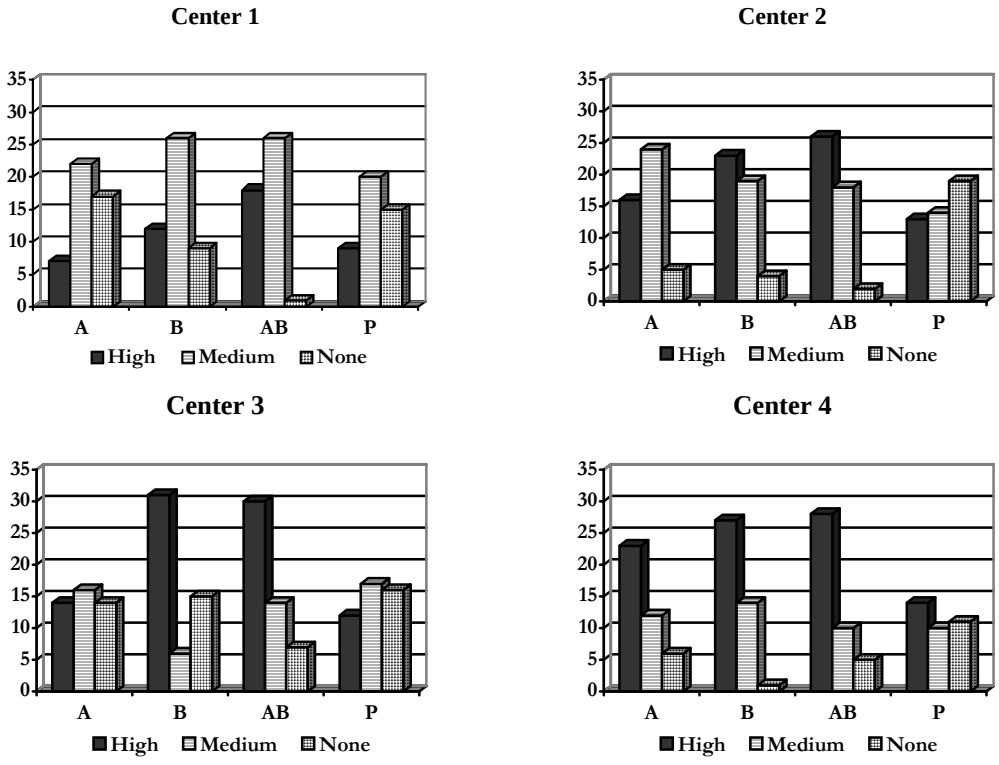

Figure 4. Performance of the treatment by center.

The 4 graphs would be very similar if there would not be an interaction between center and treatment. However, is clear from the pictures that the performance of the treatment was different in each hospital. For instance, in Center 4 the high response was more frequent than the other responses and on the other hand in Center 1 the medium response was more frequent.

At this point it can be remarked that the presence of interaction between center and treatment unables a clear interpretation of the coefficients of the model and the evaluation of the efficacy of each treatment.

We can not explain the nature of this interaction because in a clinical trial every hospital involved in the study should have similar characteristics and should work following the same protocol.

A cumulative model was also fitted, using the same factors. The result of this fit is shown below in Table 5.

Table 5 . Score Test for the Proportional Odds Assumption.

Chi-Square = 66.8102 with 16 DF (p = 0. 0001)

Model Fitting Information and Testing Global Null Hypothesis BETA = 0

Criterion	Interception Only	Interception and Covariantes	Chi-Square for Covariantes
AIC	1521.332	1411.669	.
SC	1530.485	1494.045	.
- 2 LOG L	1517.332	1375.669	141.663 with 16 DF (p = 0.0001)
Score			130.844 with 16 DF (p = 0.0001)

From the previous results it can be seen that the hypothesis of proportional odds ratio was rejected. Hence we cannot assume an identical effect of the explanatory variable for all 2 collapsings of the response into binary outcomes. This can be due to the fact that the treatments have different effects in the different centers, which was already observed in the previous model.

3. CONCLUSIONS AND SUGGESTIONS FOR FURTHER ANALYSIS

- The multicategorical variables can be analyzed, using different kind of models that consider the nature of the variables.
    - ° Models for nominal response variables.
    - ° Models for ordinal response variables.
- The Model applied to the analysis of the data gave evidence of interaction between center and treatment. As a consequence, neither the efficacy of the treatments could be evaluated, nor could a clear interpretation of the coefficients involved in the model be made.
- The second model fitted, which used the ordinal nature of the response variable, could not establish an identical effect of the explanatory variables for all 2 collapsings of the response into binary outcomes.
- The nature of the interaction should be investigated.
    - ° The four hospitals should be visited in order to know more about the implementation of the protocol.
    - ° The primary information given by the hospitals should be checked.

REFERENCES

ANDERSON, J. A. (1984): "Regression and Ordered Categorical Variables", **J.R. Statist. Soc. B**. 46, 1-30.

COX, D.R. (1970): **The Analysis of Binary Data: Methuen**, London.

FAHRMEIR, L. and G. TUTZ (1997): **Multivariate Statistical Modelling Based on Generalized Linear Models**, New York, Springer.

FIENBERG, S.E. (1980): **The Analysis of cross-classified Data**. 2nd ed. Cambridge MIT Press, Cambridge.

NELDER, J.A. and, R.W.M. WEDDERBURN (1972): "Generalized Linear Models", **Journal of Royal Statistical Society**, A 135, 370-384.

PLACKETT, R.L. (1981): **The Analysis of Categorical Data**, 2nd ed. C. Griffin, London.