Application of Threshold Models to The Detection of IgG and IgM Antibodies

Louis Munyakazi, Biostatistics Department
Amgen Inc. Thousand Oaks, CA
louism@amgen.com

1. Brief Description:

The EIA screening assay was developed to detect IgG and IgM antibodies in human serum. Samples are loaded into a plate at 50 μL/well in duplicate and incubated for 1 hour at room temperature on a rotator. Antibodies are then detected using goat anti-human IgG and goat anti-human IgM HRP conjugated antibodies.

The optical density (OD) of each well is determined using a plate reader set at a dual wavelength setting, 450-650 nm. An average OD value is calculated for each set of samples and controls, and the ratio of the mean post-treatment OD to the mean pretreatment OD is calculated.

The ratio of the OD ratios for a subject *i* relative to another subject *j* ($ratio_i/ratio_j$) do not necessarily translate to an equivalent proportion of antibodies present between the two subjects. Moreover, the nature of the responses is such that a decision to re-assay a particular sample is justified only for ratios that exceed some threshold values. Both arguments support the assumed ordinal nature for the ratios of OD measurements.

2. Objectives:

The primary use of this assay is for screening. The objective is to obtain reasonable cutoff points that can be used to separate samples that didn't produce antibodies from those that are atypical and worthy of further investigation.

3. Model of latent variable:

Threshold models are used when there are reasons to believe that an observed response can not occur below a certain critical value. Throughout the remainder, the threshold concept is used in which it is assumed that the observed ordered category is determined by the value of a latent response. Latent variables are defined as underlying unobservable continuous responses generally used to describe probabilities that observations fall into some pre-defined categories (Snell, 1964). In modeling such a latent variable, we consider the generalized linear *fixed effects* model although the model can be extended to incorporate additional random components in the case replicates on the same subject are available. The model has the form:

$$Y_i = \eta_i + \varepsilon = \mathbf{x}_i'\beta + \varepsilon \qquad (1)$$

where $\mathbf{x}$ is a design matrix, β is a vector of unknown fixed effects, and ε is a vector of independent normal residuals with mean μ equal to 0 and variance given by σ^2. The vector ε can be drawn from distributions other than the normal distribution. They are presented in section 4 and Table 1.

In general the random variable Y_i is *not observed* but its class, say Z_j, (j = 1, . . ., J) is. The additivity assumptions implicit in model (1) seem more reasonable when the model is applied to the underlying continuous responses than when it applied to the assigned categories (Harville and Mee, 1984). Therefore Y is

referred to as an underlying latent variable because it underlines the generation of the ordinal response. If we define the threshold by γ $(\gamma_1, \gamma_2, \ldots, \gamma_J)$, a given response is assigned to category j when

$$\gamma_{j-1} < Y_i \leq \gamma_j. \quad \Leftrightarrow \quad Z_j = J \qquad (2)$$

In applying the threshold model, the objective is to make inferences about various functions of γ. The current methods to estimate the vector γ are based on computing the probability of a given sample to fall into one category (Ezzet and Whitehead, 1991; Hedeker and Gibbons, 1994; Gibbons and Hedeker, 1997; Valen and James, 1999; SAS, 1999). Consequently, the appropriate models are restricted to those models that provide predicted probabilities that lie within the interval (0 1). This reason and many others not mentioned here constitute the driving force behind the development of the generalized linear models.

4. Link functions:

When the probability of an event such as the presence or absence of antibodies in a given sample is of interest, the prediction of that event is related to a linear function of exploratory variables (see model (1)) via a *link function*. For a single response measured on the ordinary scale, the most common choices for the latent distribution functions for ε, thus Z, are (Kyungman, 1995).

- The Normal distribution function
- The Logistic distribution function
- The Extreme value distribution function

The distributions are non-linear link functions because they literally *link* the linear function of covariates in (1) to a probability estimate. The link functions are such that the computed probabilities are guaranteed to always lie between 0 and 1. Their properties are given in the table 1.

Using model (1), we assumed

$$(Y - \mathbf{x}_i'\beta)/\sigma \qquad (3)$$

are independent with a common distribution above, say F. In general σ is a constant, thus for convenience it is set to one and β is a vector of regression coefficients. In terms of the cumulative probabilities μ, one can write the regression model as

$$F^{-1}\{\mu_j(\gamma, \beta; \mathbf{x}_i')\} = \gamma_j - \mathbf{x}_i'\beta \qquad (4)$$

In this form, the *logit model* is

$$\text{Log}(\mu_j/(1-\mu_j)) = \gamma_j - \mathbf{x}_i'\beta, \qquad (5)$$

the *probit model* is

$$\Theta^{-1}(\mu_j) = \gamma_j - \mathbf{x}_i'\beta, \qquad (6)$$

and the *complementary log-log model* is

$$\text{Log}\{-\text{Log}(1-\mu_j)\} = \gamma_j - \mathbf{x}_i'\beta. \qquad (7)$$

All three models provide similar predicted probabilities (Berkson, 1951; Chambers and Cox, 1967; Valen and James, 1999).

5. The basic of threshold estimation:

The above methods focus on estimating the thresholds that separate the ordered OD ratios and at the same time model the influence of the exploratory variable $\mathbf{x}_i$ on the thresholds. Instead of modeling the Z categories themselves, a common strategy is to observe if the response Y falls into one of the following intervals:

$$\{(\gamma_0\ \gamma_1), (\gamma_1, \gamma_1+\gamma_2), \ldots, (\gamma_1+\gamma_2+\ldots, \gamma_k)\}$$

where all γ_i are positive. In addition, $\gamma_0 = -\infty$ and γ_k, the upper threshold for the last category, is unbounded i.e., $\gamma_k \equiv +\infty$. Moreover, γ_1 is set to zero for model identifiability. Model identifiability means that in the process of fitting model (1), the intercept β_0, e.g., the model overall mean, is included. The set-to-zero restriction on γ_1 bears the same idea as in overparameterized linear model (Searle, 1971; Little, R.C. et al 1991).

To solve model (1) and obtain an estimate of the solution vector β, we rely on the concept of *likelihood*. More specifically, given the observed data and an assumed distribution, the likelihood function gives the probability of observing what was actually observed. The estimator that makes the observed OD data most likely is the *maximum likelihood estimator*.

6. Application to actual data set:

In order to accommodate for the above model formulation, the data is first dichotomized into distinct categories as in Table 3. In the case of IgG and IgM, three meaningful categories were feasible. As a consequence of the data classification, samples were regrouped into classes of unequal sizes. This dichotomization is not necessary required, however, as one can exploit non-categorized data directly.

There is no known mathematical strategy available to optimally collapse elements of a given data set into groups (McCullagh and Nelder, 1989). From practical considerations, a minimum of 2 observations per class was required in order to include a category in the analysis. A more meaningful number is 5 (Hoel, 1971). As expected, smaller sample sizes are found at higher scored categories (3 and/or 4), meaning that those particular samples are atypical and therefore worthy of further investigation.

The hope was that the nature of the conclusion will not be affected by the choice of the categories given the fact that, in each case, we are measuring the same parameters however many categories are selected. It is a well-known fact that the interpretation of the slope parameter β is the same regardless of the number of categories (Stiger et al, 1999).

For the present problem of working with a maximum of four distinct categories (1, 2, 3, and 4), a total of (# classes - 1) thresholds must be introduced. The partition of the line is as follow:

$\{(\gamma_0, \gamma_1), (\gamma_1, \gamma_1+\gamma_2), (\gamma_1+\gamma_2, \gamma_1+\gamma_2+\gamma_3,), (\gamma_1+\gamma_2+\gamma_3, \gamma_4)\}$

However, since the parameter γ_0, γ_1, and the last parameter γ_4 are generally set to $-\infty$, 0, and $+\infty$ respectively, only two threshold values, γ_2 and γ_3, are of practical interest and therefore must be estimated. In cases where three categories are possible, only γ_2 is estimable.

The probit model was chosen for implementation because of the following reasons:

(1) similarity of the predicted probabilities obtained from model (5), (6), and (7), (see Table 6, and Fig. 1-6 of SAS output).
(2) wide range of applications,
(3) the data is heavily concentrated in the tail (J=1, 2) and few at $J \geq 3$,

7. Exploratory Variables

With regard to cancer patients' data, it is equally important to evaluate and monitor directly which

of the exploratory variables denoted by x_i (stage of the disease, patient age etc. . .) affect the probability for the development of antibodies.

Table 2. summarizes the different steps that lead to the final model and parameter estimates. Table 3. provides examples of a working frequency distribution for selected categories or classes. The effect of model on the cutpoint estimates is presented in Tables 4 and 5 for IgG and IgM respectively. Findings obtained from the IgG and IgM data and recommendations for future experiments are presented in the final report.

The statistical methods described herein were implemented using the SAS system (SAS, 1999) on PC. The corresponding SAS codes and partial output are in *Appendix I.*

7. References:

Amemiya, T (1981). Qualitative Response Models: A survey. *Journal of Economic Literature*, **19** 1483-1536.

Chambers E.A and Cox D.R (1967). Discrimination between Alternative Binary response models. *Biometrika*, **54** 573-578.

Hoel, P.G. (1971). Introduction to mathematical statistics. Fourth Ed. New York: Wiley

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics* **43**, 439-456.

Agresti A. (1990) - Categorical data analysis. John Wiley & Sons Inc.

Wolfinger, R. (1999). The Nlmixed Procedure (draft). Chap. 3.

Harville, D.A. and Robert W. Mee (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, **40**, 393-408.

Maura E. Stokes et al (1995). Categorical data analysis using the SAS System, Cary NC. SAS Inc. 499 pp.

Ezzet F and Whitehead J. (1991). A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* **10**, 901-907.

Hedeker D and Gibbons R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.

Ulm K (1991) - A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine* **10**, 341-349.

Little R.C et al (1991). SAS System for Linear Models, Third Ed. Cary, NC: SAS Institute Inc.

Siger et al (1999). Testing proportionality in the proportional odds model fitted with GEE. *Statistics in Medicine* **18**, 1419-1433.

Snell, E.J (1964). A scaling procedure for ordered categorical data. *Biometrics* **20**, 592-607.

Kyungmann K (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine* **14**, 1341-1352.

Valen E. Johnson, J.H. Albert (1999). Ordinal Data Modeling 258pp Springer-Verlag New York

[1]Louis Munyakazi Ph.D.
Amgen Inc.
One Amgen Center Drive
Thousand Oaks, CA 91320-1789
louism@amgen.com