DIAGNOSTIC METHODS IN THE APT MODEL FOR ORDERED CATEGORICAL DATA

Hamasaki, Toshimitsu
Biostatistics, Pfizer Global Research and Development, Tokyo Laboratories, Pfizer Japan Inc.,

Goto, Masashi
Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University

https://doi.org/10.5109/13516

出版情報：Bulletin of informatics and cybernetics. 35 (1/2), pp.1-18, 2003-12. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 35, No. 1-2, 2003

DIAGNOSTIC METHODS IN THE APT MODEL FOR ORDERED CATEGORICAL DATA

By

Toshimitsu HAMASAKI* and **Masashi GOTO†**

Abstract

We describe diagnostic methods for assessing the influence of a single outlier or individual case on the estimates of the asymmetric power transformation (APT) model for ordered categorical response. The APT model includes a wide range of the probability curves of the response including both of symmetric and asymmetric curves. We also discuss methods for assessing the influence of covariate selection. Furthermore, we mention methods for assessing the appropriateness of combing the categories of the response. Two examples are used to illustrate the proposed method.

Key Words and Phrases: Cumulative logit model; Complementary log-log transformation model; Influential case; Likelihood distance; Covariate Selection; Category combining.

1. Introduction

A situation frequently encountered in data analysis of medical research studies is to relate more than one variable or covariate to an ordered categorical response. For example, in a clinical trial for the treatment of allergic rhinitis caused by mountain cedar pollen, the response variable 'sneezing' is observed on a four-point scale (no/mild/moderate/serve symptoms) when a subject receives any treatments, and then the relationship between the response and treatments is investigated, allowing for the effects of several covariates such as baseline severity, age, or gender (Lunn, Walkefield and Racine-Poon, 2001). In this situation, a common method of analysis is to invoke the concept of an unobserved response corresponding to the ordered categorical one and then to assume that a liner combination of the variables determines the probability of response through a specific link function (McCullagh, 1980; Cox and Snell, 1987; McCullagh and Nelder, 1989). Thus, models for ordered categorical responses are often specified in the term of cumulative probabilities rather than individual category responses. Cumulative logit, probit and complementary log-log models are well-known as standard models.

The cumulative logit and probit models assume the symmetric probability curve of a response. In practice, however, real data does not often satisfy this assumption. To extended scope of the standard models to asymmetric probability curves and to improve the fit in the noncentral probability regions, more flexible or data-adaptive methods have been proposed by introducing families of the response curves indexed

* Biostatistics, Pfizer Global Research and Development, Tokyo Laboratories, Pfizer Japan Inc., 3-22-7 Yoyogi, Shibuya, Tokyo 151-8589 Japan. E-mail: toshimitsu.hamasaki@japan.pfizer.com

† Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Osaka 560-8531 Japan. E-mail: gotoo@sigmath.es.osaka-u.ac.jp

by one or more shape parameters (Stukel, 1988; Taylor, 1988). Prentice (1974, 1976), Pregibon (1980) and Stukel (1988) consider families with two shape parameters. For fitting a binary response, Prentice (1974, 1976) modeled the expected probability curve with the distribution function of a log-gamma distribution. The family of the log-gamma distribution contains the logistic, normal, extreme minimum and maximum, exponential, Laplace, and reflected exponential distribution as special cases, so that this model can handle many nonstandard situations. Goto and Inoue (1987) extended the Prentice model to more general situations where a response is given in ordered multi-categories. Pregibon (1980) defined a family of link functions, including the logit link as a special case, to examine the adequacy of hypothesized link functions for fitting binary response in the context of a generalized linear model.

Mosteller and Tukey (1977), Aranda-Ordaz (1981), Guerrero and Johnson (1982), Morgan (1983, 1992) and Whittemore (1983) presented one-parameter families for fitting a binary response within the framework of simple analogies of Box and Cox power-transformation model for fitting a continuous response when working with a binary response. Aranda-Ordaz (1981) introduced two separate one-parameter models for symmetric and asymmetric departures respectively, from the logit model. The symmetric power-transformation (SPT) model contains the logit model, and the asymmetric power-transformation (APT) model includes the logit and complementary log-log transformation models. Goto, Inoue and Tsuchiya (1986) discussed the extensions of the APT model to responses with ordered multi-categories. Guerrero and Johnson (1982) suggested a one parameter Box and Cox power-transformation to the odds, which contains logit transformation. Morgan (1985) presented a one-parameter cubit logit model to model the symmetric departures from the logistic curve. The Morgan model is a first-order approximation to the SPT model. Copenhaver and Mielke (1977) modeled the expected probability curve with the distribution function of a two-parameter omega distribution. The omega-distribution includes logistic and uniform distributions as special cases. A similar approach was employed by Van Montford and Otten (1976). Their model uses the distribution function of a lambda-distribution. The advantage of these flexible models is the potential improvement in the fit to the data. Also, these models may be used to examine symmetric or asymmetric deviation from the logit or probit model.

The maximum likelihood estimate of the shape parameter in the flexible models can be obtained by using an iteratively reweighted least square algorithm. As well as in the Box and Cox power transformation model, however, it is well-known that outliers or influential cases have unacceptable effect on the maximum likelihood estimates when they exist and the estimation of the shape parameter may be influenced by the cases. It is surely important to find if the evidence for the estimated shape parameter is spread evenly throughout the data or rests within only a few cases. This problem requires special diagnostic methods because the cases that influence a shape parameter may not be distinguished in the subsequence analysis.

In this paper, we develop diagnostic methods for assessing the influence of a single outlier or individual case on the estimates of the APT model. There are two reasons why we discuss the APT model here. One is that the APT model includes a wide range of the probability curves of the response including both of symmetric and asymmetric curves. The other is that the APT model has less restriction in the parameter estimation compared with similar flexible models. We also discuss methods for assessing the influ-

ence of covariate selection and combing the categories of the response on the parameters. The methods developed in this paper can be easily extended to the other models. This paper is structured as follows: in Section 2 we briefly establish the definition and the parameter estimation of the APT model. In Section 3 we develop the diagnostic method for the shape parameter on deletion of individual cases. We also consider the diagnostic method for the other parameters on deletion of individual cases. In section 4, we discuss further diagnostic methods for assessing the influence of covariate selection and combing the categories of the response on the parameters. We show two examples to illustrate the proposed methods. In Section 5 we provide our concluding comments.

2. The APT Model

2.1. Preliminary

Let Y_i $(i = 1, ..., n)$ denote the response classified into one of k $(k \geq 2)$ ordered categories, and let Y_i be observed with r covariates. As described above, one of the most effective approaches to construct a model for the ordered categorical response is to invoke the concept of an unobserved respons Z_i. The actual recorded response Y_i is envisaged as a crude manifestation of the unobserved variables in such a way that the relationship is monotone, namely

$$\theta_{s-1} < Z_i \leq \theta_s \Leftrightarrow Y_i = s \tag{2.1}$$

where θ_s is a cutoff parameter on the underlying distribution of ordered categories, and $\theta = (\theta_1, ..., \theta_{k-1})^{\mathrm{T}}$ is a $(k-1) \times 1$ vector of the cutoff parameters. Incidentally, θ_s is given by

$$-\infty = \theta_0 < \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1} \leq \theta_k = +\infty.$$

We assume that the dependence of the unobserved response on the covariates may be specified by means of a linear model. Then, we have

$$Z_i = \mathbf{x}_i^{\mathrm{T}} \beta + \varepsilon_i$$

where $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_r)^{\mathrm{T}}$ is an $(r+1) \times 1$ vector of unknown (regression) parameters, $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ir})^{\mathrm{T}}$ is an $(r+1) \times 1$ vector of covariates and ε_i is a random error with a cumulative distribution function F. Then, the probability $\Pr(Z_i \leq z_i)$ is $F(z_i - \mathbf{x}_i^{\mathrm{T}} \beta)$. The relationship (2.1) between the unobserved variable and the response gives the implied model for Y_i in the form

$$p_{is} = \Pr(Y_i \leq s) = \Pr(Z_i \leq \theta_s) = F(\theta_s - \mathbf{x}_i^{\mathrm{T}} \beta)$$

or in linearized form

$$F^{-1}(p_{is}) = \theta_s - \mathbf{x}_i^{\mathrm{T}} \beta.$$

If $F(z_i) = \exp z_i / (1 + \exp z_i)$, implying that ε_i has the logistic distribution, then this scheme produces the cumulative logit model. The cumulative probit model arises if has the normal distribution. For further details of the standard ordered categorical models, see McCullagh (1980), Agresti (1984, 1999), McCullagh and Nelder (1989) and Johnson and Albert (1999).

2.2. Definition

We now suppose that the response probability p_{is} satisfies the relationship

$$T_\lambda(p_{is}) = \theta_s - \mathbf{x}_i^{\mathrm{T}}\beta$$

through a transformation function $T_\lambda(p_{is})$ with one shape parameter λ. The APT model is an analogy of the Box and Cox power-transformation model not direct to p_{is}, but to a corrected odds $1/(1-p_{is})\,(=p_{is}/(1-p_{is})+1)$. The transformation of the probability $T_\lambda(p_{is})$ for the APT model is defined by

$$T_\lambda(p_{is}) = \begin{cases} \log\dfrac{1}{\lambda}\left(q_{is}^{\lambda}-1\right), & \lambda \neq 0, \\ \log\left(\log q_{is}\right), & \lambda = 0 \end{cases} \tag{2.2}$$

where $q_{is} = 1/(1-p_{is})$ (Aranda-Ordaz, 1981; Goto *et al.*, 1986). Then, $T_\lambda(p_{is})$ coincides with a logit transformation for $\lambda = 1$ and with a complementary log-log transformation if $\lambda = 0$. For this model, by the inverse-transformation of (2.2), we have

$$p_{is} = F(\theta_s - \mathbf{x}_i^{\mathrm{T}}\beta) = \begin{cases} 1-\left\{1+\lambda\exp(\theta_s-\mathbf{x}_i^{\mathrm{T}}\beta)\right\}^{-1/\lambda}, & \lambda \neq 0, \\ 1-\exp\left\{-\exp(\theta_s-\mathbf{x}_i^{\mathrm{T}}\beta)\right\}, & \lambda = 0. \end{cases}$$

Fig.1 provides a diagram showing how the response probabilities for the APT model vary the shape of distribution. Fig.2 provides the behaviors of skewness and kurtosis of the response varying with $\lambda(>0)$ from the APT model. From the two figures, it is clear that the probability curve from the APT model is not symmetric except for $\lambda = 1$, and that it becomes more log-tailed distribution as λ increases. Then, the APT model includes a wide range of the probability curves of the response including both of symmetric and asymmetric curves. This feature is very useful to assess symmetric and asymmetric departures from the logit and complementary log-log transformation models. Also, though the APT has a very similar form with the STP and Guerrero and Johnson models (Guerrero and Johnson, 1982), the APT model has less restriction in the parameter estimation compared with the models (Goto *et al.*, 1986). Therefore, in practice, the APT model may be one of most powerful tools to analyzing the ordered categorical response.

2.3. Strategy for Parameter Estimation

The estimates of the parameters λ, β and θ for the APT model can be obtained by using the maximum likelihood method. Suppose that observations $\{(\mathbf{x}_i, Y_i), i = 1, ..., n\}$ are given for n individuals. Also, let n_s be the number of individuals having Y_i in the interval $[\theta_{s-1}, \theta_s]$, and N_s be a random variable whose observation is n_s $(s = 1, ..., k-1)$, where $n_1 + \cdots + n_{k-1} = n$. Then, the joint occurrence probability of an event $\{n_1, ..., n_{k-1}\}$ has a multi-nominal distribution given by

$$\Pr(N_1 = n_1, ..., N_{k-1} = n_{k-1}) \propto \prod_{s=1}^{k-1} \Pr(Y_s = s|\,\mathbf{x}_i)$$

and the probability is given by

$$\Pr(Y_i = s|\,\mathbf{x}_i) = p_{is} - p_{i,s-1}$$

where p_{is} and $p_{i,s-1}$ are the response probabilities of the i-th individual that Y_i is less than or equal to category s and $s-1$ respectively. Therefore, the log-likelihood for λ, β and θ based on n observations is given by

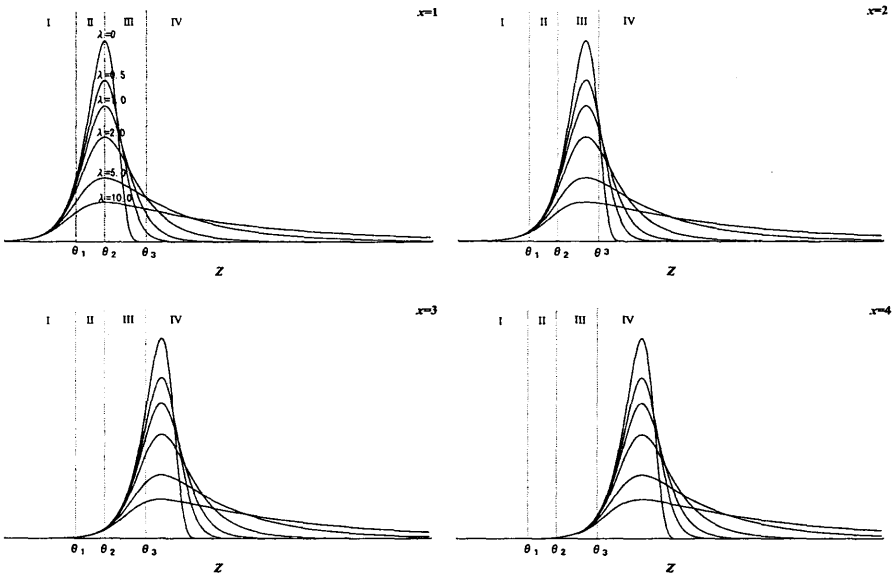

Figure 1: Diagram showing how the response probabilities for the APT model vary with $\lambda(>0)$ when $1 \leq x \leq 4$, $\beta_0 = 0$ and $\beta_1 = 2$. Response categories are represented as four continuous interval of the Z-axis

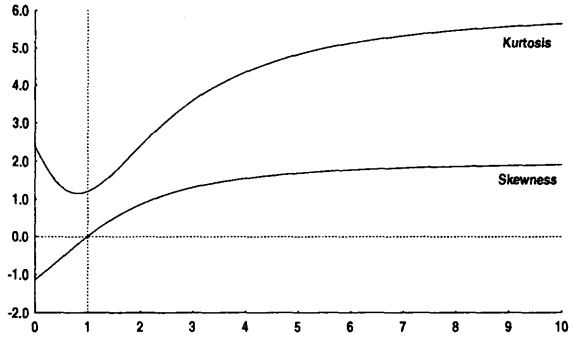

Figure 2: The behaviors of skewness and kurtosis of the distribution for the response

$$\log L(\lambda, \beta, \theta) = \sum_{i=1}^{n} \log \{g(x_i, y_i : \lambda, \beta, \theta)\} \tag{2.3}$$

where

$$g(x_i, y_i : \lambda, \beta, \theta) = \begin{cases} F(\theta_{y_i} - \mathbf{x}_i^{\mathrm{T}}\beta), & y_i = 1, \\ F(\theta_{y_i} - \mathbf{x}_i^{\mathrm{T}}\beta) - F(\theta_{y_{i-1}} - \mathbf{x}_i^{\mathrm{T}}\beta), & 2 \le y_i \le k-1, \\ 1 - F(\theta_{y_{i-1}} - \mathbf{x}_i^{\mathrm{T}}\beta), & y_i = k. \end{cases}$$

The maximum likelihood estimates of the parameters λ, β and θ can be obtained by maximizing the log-likelihood (2.3) over the parameters. In general, it is well-known that the estimates of λ, β or θ can be highly correlated, so that the marginal variance of the estimated β or θ can be hugely inflated by not knowing λ when λ, β and θ are estimated simultaneously. To avoid this, the estimates of the parameters are found in the two stages. First, for fixed λ, the log-likelihood (2.3) is maximized with respect to β and θ. Then, log-likelihood (2.3) for fixed λ is the log-likelihood for an iteratively reweighted least square problem with an ordered categorical response. Therefore, if we denote the maximum likelihood estimates of β and θ for the fixed λ by $\hat{\beta}(\lambda)$ and $\hat{\theta}(\lambda)$, then substitution of $\hat{\beta}(\lambda)$ and $\hat{\theta}(\lambda)$ into the log-likelihood (2.3) yields the log-likelihood respect with respect to λ,

$$\log L(\lambda) = \sum_{i=1}^{n} \log \left\{ g(\mathbf{x}_i, y_i : \lambda, \hat{\beta}(\lambda), \hat{\theta}(\lambda)) \right\}. \tag{2.4}$$

Secondly, the log-likelihood (2.4) is maximized with respect to λ to obtain the maximum likelihood estimates $\hat{\lambda}$ of λ. Consequently, we have the maximum likelihood estimates $\hat{\lambda}$, $\hat{\beta}(\hat{\lambda})$ and $\hat{\theta}(\hat{\lambda})$. Thus, $\hat{\beta}(\hat{\lambda}) \to \beta$ and $\hat{\theta}(\hat{\lambda}) \to \theta$ if $\hat{\lambda} \to \lambda$ as $n \to \infty$.

As a regression model for ordered categorical data is overparameterized if there are responses with k categories and $k-1$ unknown cutoff parameters $\theta_1, ..., \theta_{k-1}$, any set of k probabilities can be obtained but not in a unique way (Jansen, 1991; Johnson and Albert, 1999; Lunn *et al.*, 2001). In other way, if we add a constant to every cutoff value and subtract the same constant from the intercept in the regression function, the values of $\theta_s - \mathbf{x}_i^{\mathrm{T}}\beta$ used to define the category probabilities are unchanged. To make the model identifiable, there are two approaches. The first approach is to introduce a reference point on the latent scale. Usually this is done by fixing one of the cutoff parameter at zero, or alternatively fixing $\mathbf{x}_i^{\mathrm{T}}\beta$ at some arbitrary value. The second approach is to specify a prior distribution on the vector of category cutoff points $\theta_1, ..., \theta_{k-1}$ (Johnson and Albert, 1999).

In this paper, we compute the parameter estimates by using the first approach, and actually fix $\mathbf{x}_i^{\mathrm{T}}\beta$ by centerized $\mathbf{x}_i$ to the mean with $\beta_0 = 0$. In addition, the choice of starting values is important in an iteratively reweighted least square algorithm. Then, letting $\beta^0 = (\beta_1^0, ..., \beta_r^0)^{\mathrm{T}}$ and $\theta^0 = (\theta_1^0, ..., \theta_{k-1}^0)^{\mathrm{T}}$ by starting values of $\beta = (\beta_1, ..., \beta_r)^{\mathrm{T}}$ and $\theta = (\theta_k, ..., \theta_{k-1})^{\mathrm{T}}$ respectively, we use $\beta_1^0 = \cdots = \beta_r^0 = 0$ and

$$\theta_s^0 = \log \left\{ \left(\sum_{t=1}^{s} n_t + 0.5 \right) \Big/ \left(n - \sum_{t=1}^{s} n_t + 0.5 \right) \right\} \quad (s = 1, ..., k-1) \ .$$

3. Diagnostics for Assessing the Influence of an Individual Case on the Parameters

3.1. Influence of an Individual Case on the Shape Parameter

Here we develop two diagnostic methods for assessing the influence of an individual case on the estimates of the shape parameter λ in the APT model. One is the change in λ on deletion of the i-th individual case and the other is the likelihood distance which is the changes in log-likelihood on deletion of the i-th individual case. Both the two diagnostics require the calculation of n maximum likelihood estimates of λ or likelihood when a single influential case is considered. Such a computation may be prohibitively expensive in moderate to large data sets. To reduce this amount of computation, we develop the one-step estimates of λ.

First we develop the diagnostic for λ on deletion of the i-th individual case. Let $\log L_{[i]}(\lambda)$ and $\hat{\lambda}_{[i]}$ denote the log-likelihood and the maximum likelihood estimate of λ when the i-th case is deleted from the fit, respectively. Thus, we can approximate $\log L_{[i]}(\lambda)$ by the second-order Taylor series expansion of $\log L_{[i]}(\lambda)$ about $\hat{\lambda}$. Then we write

$$\log L_{[i]}(\lambda) \approx \log L_{[i]}(\hat{\lambda}) + (\lambda - \hat{\lambda}) \log L'_{[i]}(\hat{\lambda}) + \frac{(\lambda - \hat{\lambda})^2 \log L''_{[i]}(\hat{\lambda})}{2} \tag{3.1}$$

where $\log L'_{[i]}(\hat{\lambda})$ and $\log L''_{[i]}(\hat{\lambda})$ are the first and second order partial derivatives of the log-likelihood (2.4) with respect to λ when the i-th case is deleted from the fit. If $\hat{\lambda}$ is sufficiently close to λ , the remainder terms will be small relative to the other terms and then ignored. In fact, it is well-known that $\hat{\lambda} \to \lambda$ as $n \to \infty$. If $\left|\log L''_{[i]}(\lambda)\right| \neq 0$, the case-deletion log-likelihood (3.1) can be maximized in which the estimate is given as

$$\hat{\lambda}^*_{[i]} = \hat{\lambda} - \left\{\log L''_{[i]}(\hat{\lambda})\right\}^{-1} \log L'_{[i]}(\hat{\lambda})$$

where $\hat{\lambda}^*_{[i]}$ is the one-step estimate of $\hat{\lambda}_{[i]}$.

Next, we develop the likelihood distance, that is the diagnostic for the changes in log-likelihood on deletion of the i-th individual case. Let $\log L(\hat{\lambda})$ denote the maximized log-likelihood at $\hat{\lambda}$ when all cases are used for the fit, and let $\log L(\hat{\lambda}_{[i]})$ be the maximized log-likelihood at $\hat{\lambda}_{[i]}$ when the i-th case is deleted from the fit. Thus, the likelihood distance between $\log L(\hat{\lambda})$ and $\log L(\hat{\lambda}_{[i]})$ is given as

$$LD_i(\hat{\lambda}) = 2\left\{\log L(\hat{\lambda}) - \log L(\hat{\lambda}_{[i]})\right\}. \tag{3.2}$$

Then, the second-order Taylor series expansion of $\log L(\hat{\lambda}_{[i]})$ about $\hat{\lambda}$ yields the approximation

$$\log L(\hat{\lambda}_{[i]}) - \log L(\hat{\lambda}) \approx (\hat{\lambda}_{[i]} - \hat{\lambda}) \log L'(\hat{\lambda}) + \frac{(\hat{\lambda}_{[i]} - \hat{\lambda})^2 \log L''(\hat{\lambda})}{2}$$

where $\log L'(\hat{\lambda})$ and $\log L''(\hat{\lambda})$ are the first and second order partial derivatives of the log-likelihood (2.4) with respect to λ. Thus, if we assume $\log L'(\hat{\lambda}) = 0$ and use $\hat{\lambda}^*_{[i]}$ as an approximation of $\hat{\lambda}_{[i]}$, the likelihood distance (3.2) can be written as

$$LD_i^*(\hat{\lambda}) = (\hat{\lambda}_{[i]}^* - \hat{\lambda})^2 \left\{ -\log L''(\hat{\lambda}) \right\}$$

which can be compared to the percentiles of a chi-square distribution with one degree of freedom. From these results, by using the index plots of $\hat{\lambda}_{[i]}^*$ and $LD_i^*(\hat{\lambda})$ versus case number, it is possible visually to assess the influence of an individual case on the estimate of the shape parameter λ and likelihood in the APT model.

3.2. Influence of an Individual Case on the Regression and Cutoff Parameters

Similarly as in the previous section, we develop two diagnostics for assessing the influence of an individual case on the estimates of the regression parameter β and cutoff parameters θ for fixed the shape parameter $\lambda = \lambda_0$ in the APT model.

First we develop the diagnostic for changes in β and θ on deletion of the i-th individual case. If we rewrite the estimates $\gamma = (\hat{\beta}(\lambda_0),\ \hat{\theta}(\lambda_0))^{\mathrm{T}}$ for $\lambda = \lambda_0$, we denote the case-deletion log-likelihood $\log L_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)$ and maximum likelihood estimate $\hat{\gamma}_{[i]} = (\hat{\beta}_{[i]}(\lambda_0),\ \hat{\theta}_{[i]}(\lambda_0))^{\mathrm{T}}$ of $\hat{\gamma} = (\hat{\beta}(\lambda_0),\ \hat{\theta}(\lambda_0))^{\mathrm{T}}$ when the i-th case is deleted from the fit, respectively. Thus, we can approximate $\log L_{[i]}(\gamma|\,\lambda = \lambda_0)$ by the second-order Taylor series expansion of $\log L_{[i]}(\gamma|\,\lambda = \lambda_0)$ about $\hat{\gamma}$. Then we write

$$\begin{aligned} &\log L_{[i]}(\gamma|\,\lambda = \lambda_0) \approx \log L_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0) + \\ &(\gamma - \hat{\gamma})^{\mathrm{T}} \log L'_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0) + \frac{(\gamma - \hat{\gamma})^{\mathrm{T}} \log L''_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)(\gamma - \hat{\gamma})}{2} \end{aligned} \tag{3.3}$$

where $\log L'_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)$ and $\log L''_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)$ are the $(r + k - 1) \times 1$ gradient vector and the $(r + k - 1) \times (r + k - 1)$ Hessian matrix with elements of the first and second order partial derivatives of the log-likelihood (2.4) with respect to β and θ when the i-th case is deleted from the fit. If $\left|\log L''_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)\right| \neq 0$, we have a one-step estimates of the case-deletion estimates

$$\hat{\gamma}_{[i]}^* = \hat{\gamma} - \left\{ \log L''_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0) \right\}^{-1} \log L'_{[i]}(\hat{\gamma}|\,\lambda = \lambda_0)$$

where $\hat{\gamma}_{[i]}^* = (\hat{\beta}_{[i]}^*(\lambda_0),\ \hat{\theta}_{[i]}^*(\lambda_0))^{\mathrm{T}}$.

Next, we develop the diagnostic for the changes in log-likelihood on deletion of the th individual case. Let $\log L(\hat{\gamma}|\,\lambda = \lambda_0)$ denote the maximized log-likelihood at $\hat{\gamma}$ when all cases are used for the fit, and let $\log L(\hat{\gamma}_{[i]}\big|\,\lambda = \lambda_0)$ be the maximized log-likelihood at $\hat{\gamma}_{[i]} = (\hat{\beta}_{[i]}(\lambda_0),\ \hat{\theta}_{[i]}(\lambda_0))^{\mathrm{T}}$ when the i-th case is deleted from the fit. Thus, the likelihood distance between $\log L(\hat{\gamma}|\,\lambda = \lambda_0)$ and $\log L(\hat{\gamma}_{[i]}\big|\,\lambda = \lambda_0)$ is given as

$$LD_i(\hat{\gamma}) = 2\left\{ \log L(\hat{\gamma}|\,\lambda = \lambda_0) - \log L(\hat{\gamma}_{[i]}\big|\,\lambda = \lambda_0) \right\}. \tag{3.4}$$

If the profile of the case deletion log-likelihood (3.3) is assumed to be elliptic with respect to β and θ as it is the quadric function of β and θ, the second-order Taylor series expansion of $\log L(\hat{\gamma}_{[i]}\big|\,\lambda = \lambda_0)$ about β and θ yields the approximation

$$\begin{aligned} &\log L(\hat{\gamma}_{[i]}\big|\,\lambda = \lambda_0) - \log L(\hat{\gamma}|\,\lambda = \lambda_0) \\ &\approx (\hat{\gamma}_{[i]} - \gamma)^{\mathrm{T}} \log L'(\hat{\gamma}|\,\lambda = \lambda_0) + \frac{(\hat{\gamma}_{[i]} - \hat{\gamma})^{\mathrm{T}} \log L''(\hat{\gamma}|\,\lambda = \lambda_0)(\hat{\gamma}_{[i]} - \hat{\gamma})}{2} \end{aligned}$$

where $\log L'(\hat{\gamma}|\,\lambda=\lambda_0)$ and $\log L''(\hat{\gamma}|\,\lambda=\lambda_0)$ are the 2×1 gradient vector and the 2×2 Hessian matrix with elements of the first and second order partial derivatives of the log-likelihood (2.4) with respect to β and θ. Thus, if we assume $\log L'(\hat{\gamma}|\,\lambda=\lambda_0)=\mathbf{0}$ and use $\hat{\gamma}^*_{[i]}=(\hat{\beta}^*_{[i]}(\lambda_0),\ \hat{\theta}^*_{[i]}(\lambda_0))^{\mathrm{T}}$ as an approximation of $\hat{\gamma}_{[i]}=(\hat{\beta}_{[i]}(\lambda_0),\ \hat{\theta}_{[i]}(\lambda_0))^{\mathrm{T}}$, the likelihood distance (3.4) can be written as

$$LD^*_i(\hat{\gamma})=(\hat{\gamma}^*_{[i]}-\hat{\gamma})^{\mathrm{T}}\left\{-\log L''(\hat{\gamma}|\,\lambda=\lambda_0)\right\}(\hat{\gamma}^*_{[i]}-\hat{\gamma})$$

which can be compared to the percentiles of a chi-square distribution with $(r+k-1)$ degree of freedom. The index plots of $(\hat{\beta}^*_{[i]}(\lambda_0),\ \hat{\theta}^*_{[i]}(\lambda_0))$ and $LD^*_i(\hat{\gamma})$ versus case number may provide useful information on the influence of an individual case on the estimates $\hat{\beta}$ and $\hat{\theta}$, and likelihood in the APT model.

3.3. Examples

In this section, to illustrate the diagnostics developed in the previous section, we analyze two data. The first example is the nodal involvement data in cancer patient taken from Brown (1980). The second is the arthritis pain data from Koch and Edwards (1988). Our goals here are (i) to see the type of data-analytic information that can come from the diagnostics, and (ii) to see how well the diagnostics detect influential cases.

Example 1 Nodal Involvement Data in Cancer: The first data, taken from Brown (1980), is the 53 patients receiving surgical treatment for cancer of the prostate. A critical question in determining treatment for patients is whether the cancer has spread to the neighboring lymph nodes and whether this can be predicted from variables observed before surgery, in particular X-ray reading (x_1), stage of tumor assessed by palpation (x_2), grade of tumor as determined by biopsy (x_3), age of patient at diagnosis (x_4) and level of serum aid phosphatase (x_5). A nodal involvement, determined at surgery, is a binary response. Several authors have analyzed this example and the several approaches have been proposed (Cox and Snell, 1989; Goto, Isomura and Hamasaki, 2002). Here we follow the approach discussed by Cox and Snell (1989). The model discussed here includes the five variables x_1, x_2, x_3, $\log x_5$ and $x_6=x_2\times x_3$.

Table 1 provides the diagnostics results for the APT model. The maximum likelihood estimate of the shape parameter is 2.622. This optimal value suggests that the asymmetric model with the distribution having a more long tail, rather than the logistic distribution, is suitable for describing the data.

Fig.3 provides the index plot of case-deletion likelihood distance $LD^*_{[i]}(\hat{\lambda})$. This plot shows that the $LD^*_{[i]}(\hat{\lambda})$ has an extreme value when Case 26 is deleted, where Case 26 is the patient with the values of $x_1=0$, $x_2=0$, $x_3=0$, $\log x_5=4.407$, $x_6=0$ and $y=1$ (improved). Of the 17 patients with the values of $x_1=0$, $x_2=0$, $x_3=0$ and $x_6=0$, only this case is the patient with "improved". The rest of cases are the patients with "not improved". This suggests that Case 26 has a substantial influence on $\hat{\lambda}$. Fig.4 provides the index plot of case-deletion estimate $\hat{\lambda}^*_{[i]}$. This plots shows that the one step case-deletion estimate $\hat{\lambda}^*_{[i]}$ has an extreme value when Case 24 is deleted, in addition to the deletion of Case 26, where Case 24 is the patient with the values of $x_1=0$, $x_2=0$, $x_3=0$, $\log x_5=4.407$, $x_6=1$ and $y=0$ (not improved). Though the coefficient for level of serum aid phosphatase suggests that the patient should be improved if level of

Table 1: Nodal involvement data in cancer. Diagnostics for the APT model

	All Data		Deleted Case 26		Deleted Cases 24,26	
	$\hat{\lambda} = 2.622$		$\hat{\lambda}_{[26]} = -0.278$		$\hat{\lambda}_{[24,26]} = -0.285$	
ML	−19.598		−15.038		−15.014	
AIC	51.197		42.075		42.027	
MR	20.8%		15.4%		13.7%	
Paramter	Estimate	Std.Err.	Estimate	Std.Err.	Estimate	Std.Err.
β_1	3.769	1.567	2.223	0.751	2.076	0.712
β_2	4.634	2.225	2.891	0.621	2.707	0.612
β_3	5.785	2.786	2.946	0.797	2.758	0.789
β_5	−6.830	3.398	−3.415	1.016	−3.216	1.007
β_6	3.895	2.035	1.853	0.856	1.857	0.970
θ_1	2.182	0.939	0.064	0.264	0.019	0.257

ML: Maximum likelihood
MR: Misclassification rate

serum aid phosphatase is a positive high value, this patients with the highest value of level of serum aid phosphatase, is not improved. This suggests that both Cases 24 and 26 have the substantial influences on $\hat{\lambda}$.

Without Case 26, the maximum likelihood estimate of the shape parameter is −0.278, which is closer to zero compared with that for all data. This optimal value suggests that the complementary log-log model is reasonable to describe the data. For this optimal value, the misclassification rate is reduced from 20.8% to 15.4%. Otherwise, without both Cases 24 and 26, the maximum likelihood estimate of the shape parameter is −0.285, which is very close to the value when without only Case 26, and still suggests the appropriateness of the complementary log-log model for this data. For this optimal value, the misclassification rate is reduced to 13.7%, which is slightly smaller than the

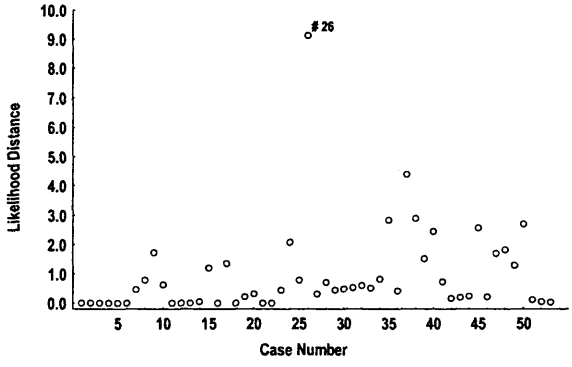

Figure 3: Nodal involvement data in cancer. Index plot of vase-deletion likelihood distance $LD^*_{[i]}(\hat{\lambda})$ versus case number

value when without only Case 26.

Fig.5 provides the index plot of case-deletion estimates $\hat{\beta}^*_{1[i]}(\hat{\lambda})$, $\hat{\beta}^*_{2[i]}(\hat{\lambda})$, $\hat{\beta}^*_{3[i]}(\hat{\lambda})$, $\hat{\beta}^*_{4[i]}(\hat{\lambda})$, $\hat{\beta}^*_{5[i]}(\hat{\lambda})$ and $\hat{\theta}^*_{1[i]}(\hat{\lambda})$. This plot clearly shows that all of one step case-deletion estimates have extreme values when either Case 24 or 26 is deleted. This suggests that both Cases 24 and 26 have substantial influences on the estimates, which is the same conclusion as seen in Fig.4. Without Case 26 or both Cases 24 and 26, the estimates are greatly changed compared with those for all data.

Example 2 Arthritis Pain Data: The second data, taken from Koch and Edwards (1988), the 84 patients receiving an active or placebo treatment for arthritis pain. A critical question is whether the active treatment for the patients is better than placebo treatment. The response is observed on the three ordered categories having "marked improved (3)", "some improved (2)" and "none (1)", with the two variables treatment (x_1) and gender (x_2). Table 2 provides the diagnostics results for the APT model. The maximum likelihood estimate of the shape parameter is 1.013, which is very close to one. This optimal value suggests that the distribution of the response is symmetric and the cumulative logit model is suitable for describing the data.

Fig.6 provides the index plot of case-deletion likelihood distance $LD^*_{[i]}(\hat{\lambda})$. This plot shows that the $LD^*_{[i]}(\hat{\lambda})$ has an extreme value when Case 62 is deleted, where Case 62 is the patients with the values of $x_1 = 0$ (placebo), $x_2 = 1$ (Male), and $y = 3$ (Marked Improved). Of 11 male patients receiving placebo treatments, only this case is the patient with "marked improved". The rest of cases are the patients with "None". Fig.7 provides the index plot of case-deletion estimate $\hat{\lambda}^*_{[i]}$. This plots also shows that the one step case-deletion estimate $\hat{\lambda}^*_{[i]}$ has an extreme value when Case 62 is deleted. These two plots suggest that Case 62 has a substantial influence on $\hat{\lambda}$.

Without Case 62, the maximum likelihood estimate of the shape parameter is -0.179, which is close to zero. This optimal value suggests that the probability curve of the response is greatly changed from symmetric curve to asymmetric curve and the complementary log-log model is reasonable to describe the data. However, for this

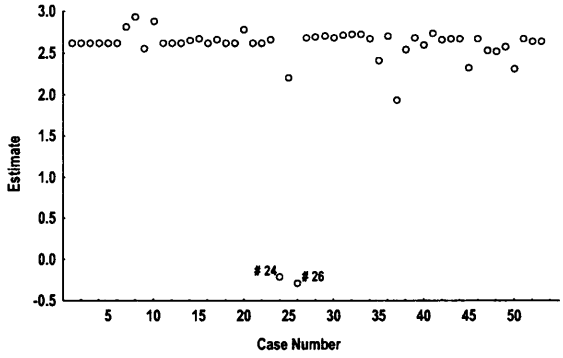

Figure 4: Nodal involvement data in cancer. Index plot of case-deletion estimate $\hat{\lambda}^*_{[i]}$ versus case number

optimal value, the misclassification rate is increased from 38.1% to 54.2%.

Fig. 8 provides the index plot of case-deletion estimates $\hat{\beta}^*_{1[i]}(\hat{\lambda})$,$\hat{\beta}^*_{2[i]}(\hat{\lambda})$, $\hat{\theta}^*_{1[i]}(\hat{\lambda})$ and $\hat{\theta}^*_{2[i]}(\hat{\lambda})$. This plot clearly shows that all of one step case-deletion estimates have extreme values when Case 62 is deleted. This suggests that Case 62 has a substantial influence on the estimates, which is the same conclusion as seen in Fig. 6 or Fig.7. Without Case 62, the estimates are greatly changed compared with those for all data.

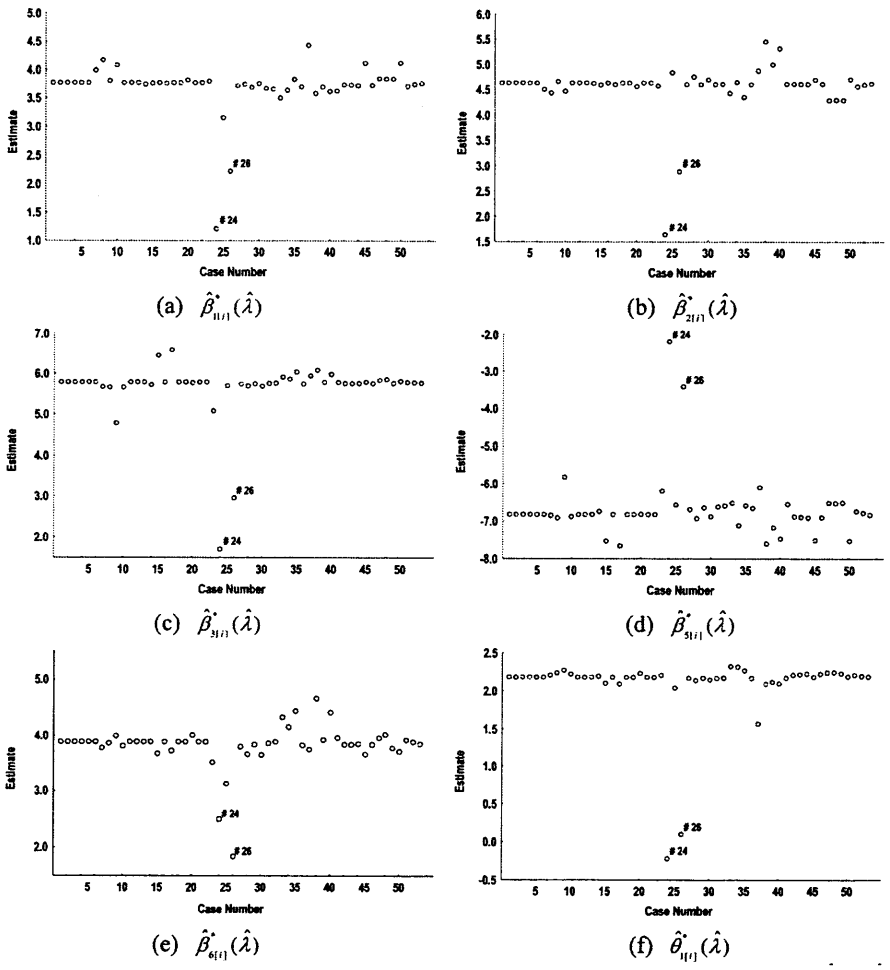

Figure 5: Nodal involvement data in cancer. Index plot of case-deletion estimates $\hat{\beta}^*_{1[i]}(\hat{\lambda})$, $\hat{\beta}^*_{2[i]}(\hat{\lambda})$, $\hat{\beta}^*_{3[i]}(\hat{\lambda})$, $\hat{\beta}^*_{5[i]}(\hat{\lambda})$, $\hat{\beta}^*_{6[i]}(\hat{\lambda})$ and $\hat{\theta}^*_{1[i]}(\hat{\lambda})$ versus case number

Table 2: Arthritis pain data. Diagnostics for the APT model

	All Data		Deleted Case 62	
	$\hat{\lambda} = 1.013$		$\hat{\lambda}_{[62]} = -0.179$	
ML	−75.015		−70.694	
AIC	158.029		149.388	
MR	38.1%		54.2%	
Paramter	Estimate	Std.Err.	Estimate	Std.Err.
β_1	1.326	0.070	0.721	0.381
β_2	1.805	0.019	1.098	0.322
θ_1	0.015	0.242	−0.511	0.168
θ_2	0.874	0.261	0.039	0.142

ML: Maximum likelihood
MR: Misclassification rate

4. Further Diagnostic Procedures

4.1. Influence of Covariate Selection on the Cutoff Parameters

We here discuss the diagnostics method for assessing the effect of the covariate selection on the estimates of the cutoff parameter $\hat{\theta}$. The effect of covariate selection on can be assessed by the likelihood ratio test of the null hypothesis $\mathrm{H}_0 : \beta = \mathbf{0}$ against the alternative hypothesis $\mathrm{H}_1 : \beta \neq \mathbf{0}$ for the fixed $\lambda = \lambda_0$. If we write the maximum likelihood estimates of θ under the hypotheses H_0 and H_1 by $\tilde{\theta}$ and $\hat{\theta}$ respectively, the statistics is given as

$$LD_{\text{model}} = 2\left\{\log L_{\mathrm{H}_0}(\hat{\lambda}, \mathbf{0}, \tilde{\theta}) - \log L_{\mathrm{H}_1}(\hat{\lambda}, \hat{\beta}, \hat{\theta})\right\} \tag{4.1}$$

where $\log L(\hat{\lambda}, \mathbf{0}, \tilde{\theta})$ and $\log L(\hat{\lambda}, \hat{\beta}, \hat{\theta})$ are the maximum log-likelihoods under the hypotheses H_0 and H_1, respectively. The statistics (4.1) has the chi-square distribution

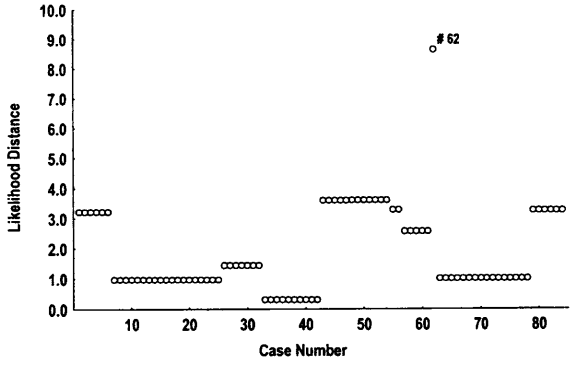

Figure 6: Arthritis pain data. Index plot of vase-deletion likelihood distance $LD^*_{[i]}(\hat{\lambda})$ versus case number

with r degrees of freedom.

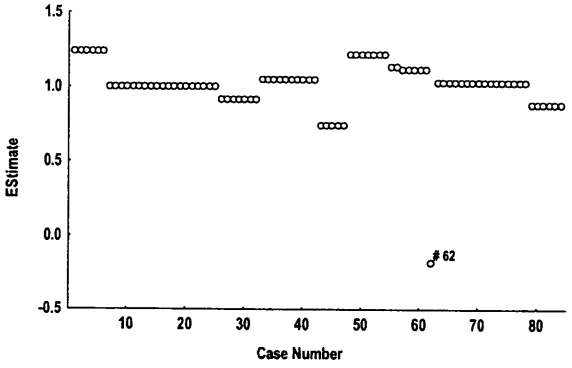

Figure 7: Arthritis pain data. Index plot of case-deletion estimate $\hat{\lambda}^*_{[i]}$ versus case number

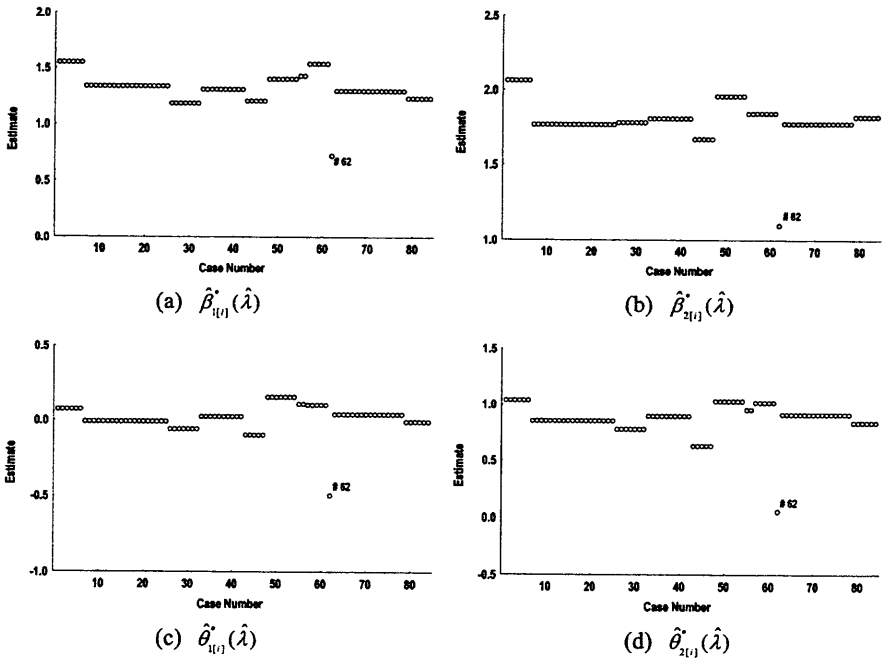

Figure 8: Arthritis pain data. Index plot of case-deletion estimates $\hat{\beta}^*_{1[i]}(\hat{\lambda})$, $\hat{\beta}^*_{2[i]}(\hat{\lambda})$, $\hat{\theta}^*_{1[i]}(\hat{\lambda})$ and $\hat{\theta}^*_{2[i]}(\hat{\lambda})$ versus case number

4.2. Appropriateness of Combing Categories

In analysis of ordered categorical response, we may often reduce the ordered categories without violating the order. For example, in a clinical trial of the treatment of migraine, the response variable is the severity of migraine observed on a four-point scale (none/mild/moderate/severe migraine), allowing for the baseline severity, a subject with a none or mild migraine after any treatments is defined as "responder" and then the other with a mild or severe migraine is as "non-responder". Combining categories would often lead to clear interpretation of results. In practice, however, it is surely important to assess whether or not combing category is more suitable for describing the data, or how much a loss of information impact on model identification. Therefore, we discuss the method for assessing the appropriateness of combing categories, within a framework of model selection. Then, in this paper, we use the AIC (Akaike, 1972) as a criterion for model selection.

Now, we consider the situation in which the category s is combined with the category $s-1$. In this situation, the number of the cutoff parameters is reduced from $k-1$ to $k-2$ and then the number of individuals observed in the interval $[\theta_{s-2},\ \theta_s]$ becomes $n_{s-1}+n_s$. Therefore, AIC statistics is given as

$$AIC = -2\log L^*_{\max} + 2(r+k-2)$$

where $\log L^*_{\max}$ is the maximum value of the log-likelihood maximized over the parameters under the re-categorization. By successively combining two adjacent categories, we can choose the category combination with the minimum value of AIC.

4.3. Example

In this section, to illustrate the method for assessing the appropriateness of combing categories developed in the previous section, we use the arthritis pain data again Tables 3 and 4 provide the results of coming categories for all data and Case 62 deleted, respectively. For all data, the minimum of the APT model is the value of 98.335 in which two categories "Marked" and "Some + None". For this re-categorization, the maximum likelihood estimate of the shape parameter is 1.450. This optimal value suggests the asymmetric curves of the response. For this optimal value under the new re-categorization, the misclassification rate is reduced to 27.4%. Without Case 62, the minimum of the APT model is the value of 91.653 in which two categories "Marked" and "Some + None". This re-categorization is consistent with that for all data. For this re-categorization, the maximum likelihood estimate of the shape parameter is -0.183, which is close to zero. This optimal value suggests the asymmetric curves of the response and the complementary log-log transformation model is suitable for this new re-categorization. The conclusions from these results are that the subjects with "Marked" would be the responders and then asymmetric models are suitable for describing the data.

5. Conclusions

In this paper, we develop diagnostic methods for assessing the influence of a single outlier or individual case on the estimates of the APT model. In particular, we provide the two simultaneous diagnostics, the changes in the shape parameter and the likelihood

Table 3: Arthritis pain data. Coming categories, All data

	(1+2), 3 $\hat{\lambda} = 1.450$		1, (2+3) $\hat{\lambda}_{[62]} = 0.066$	
ML	−46.167		−48.972	
AIC	98.335		103.944	
MR	27.4%		39.3%	
Paramter	Estimate	Std.Err.	Estimate	Std.Err.
β_1	1.111	0.702	1.028	0.369
β_2	2.206	0.651	1.296	0.363
θ_1	1.216	0.328	−0.401	0.176

ML: Maximum likelihood
MR: Misclassification rate

Table 4: Arthritis pain data. Coming categories, Case 62 deleted

	(1+2), 3 $\hat{\lambda} = -0.183$		1, (2+3) $\hat{\lambda}_{[62]} = -0.303$	
ML	−42.827		−45.631	
AIC	91.653		97.262	
MR	31.3%		38.6%	
Paramter	Estimate	Std.Err.	Estimate	Std.Err.
β_1	0.639	0.374	0.949	0.381
β_2	1.064	0.313	1.180	0.358
θ_1	0.027	0.139	−0.546	0.161

ML: Maximum likelihood
MR: Misclassification rate

distance on deletion of the i-th individual case. We also describe methods for assessing the influence of covariate selection on the parameters. Furthermore, we discuss the method for combing categories of responses within the framework of model selection. Finally, we illustrated and examined the proposed diagnostics by the two examples. The findings are as follows:

- The diagnostics based on the one-step estimates of the shape parameter in the APT model on deletion of the i-th individual case would provide useful information on the influential case. Especially, the index plot of the one-step estimates may be a useful tool for identifying a possible influential case or outlier. The methods developed in this paper can be also available for other flexible model.

- The proposed method would lead to a reasonable re-categorization even when there is no strong knowledge of the process of generating the data and combing the categories. Also, the method would assess whether or not combing category is more suitable for describing the data, or how much a loss of information impact on model identification.

- The APT model would provide the improvement in the fit to the data. Also the APT model is used to examine symmetric or asymmetric deviation from the logit

model.

For our purpose in this paper, index plots of case-deletion estimates $\hat{\lambda}^*_{[i]}$ and $\hat{\beta}^*_{[i]}$ or $\hat{\theta}^*_{[i]}$ are assessed separately. However, it is well-known that the estimates $\hat{\lambda}$ and $\hat{\beta}$ or $\hat{\theta}$ are highly correlated. Therefore, in practice, paying attention to this fact, we should assess the graphical information from the index plots carefully.

Acknowledgement

The authors are grateful to reviewers and Professor Takashi Yanagawa of Kyushu University for their constructive comments and helpful suggestions.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York:John Wiley & Sons.

Agresti, A. (1996). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18**, 2191-2207.

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, eds. by Petrov, B. N. and Csaki, F., 267-281, Budapest, Akademi Kaido.

Aranda-Ordaz, F.J. (1981). On two families of transformation to additivity for binary response data. *Biometrika* **68**, 357-363.

Box, G.P.E. and Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society* **B26**, 211-252.

Chunang-Stein, C. and Agresti, A. (1997). A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine* **16**, 2599-2618.

Copenhaver, T.W. and Mielke, P.W. (1977). Quantit analysis: a quantal assay refinement. *Biometrics* **33**, 175-186.

Cox D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, 2nd Edition. London:Chapman & Hall.

Goto, M. and Inoue, T. (1987). Log-gamma regression model for ordered categorical response data. *Japanese Journal of Behaviormetrics* **15**, 1-9 (in Japanese).

Goto, M., Inoue, T. and Tsuchiya, Y. (1986). Power-transformation for ordered categorical data. *Behaviormetrika* **19**, 17-39.

Goto, M., Iomura, T and Hamasaki, T. (2000). Guinea pig in statistical science. In *Proceedings of the 10th Korea and Japan Joint Conference of Statistics*, 265-285, Ohita, Japan, December 5-6.

Guerrero, V.M. and Johnson, R.A. (1982). Use of the Box-Cox transformation with binary response model. *Biometrika* **69**, 309-314.

Jansen, J. (1991). Fitting regression models to ordinal data. *Biomedical Journal* **33**, 807-815.

Johnson, V.E. and Albert, J.H. (1999). *Ordinal Data Modeling.* New York:Spriger-Verlag.

Koch, G. G. and Edwards, S. (1988). Clinical efficacy trials with categorical data. In *Biopharmaceutical Statistics for Drug Development*, ed. by Peace, K. E., New York:Marcel Dekker.

Lunn, D.J., Wakefield, J. and Racine-Poon, A. (1991). Cumulative logit models for ordinal data: a case study involving allergic rhinitis severity scores. *Statistics in Medicine* **20**, 2261-2285.

McCullagh, P. (1980). Regression models for ordinal Data (with discussion). *Journal of the Royal Statistical Society, Series B* **42**, 109-142.

McCullagh, P. and Nelder, J.A. (1987). *Generalized Linear Models*, 2nd edition. London:Chapman & Hall.

Morgan, B.J.T. (1980). Sbservations on quantit analysis. *Biometrics* **39**, 879-886.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data.* London:Chapman & Hall.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression.* New York:Addison-Wesley.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* **29**, 15-24.

Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika* **61**, 539-544.

Prentice, R.L. (1976). A generalization of the probit and logit models for dose-response curves. *Biometrics* **32**, 761-768.

Stukel. T.A. (1988). Generalized logistics models. *Journal of the American Statistical Association***83**, 426-431.

Taylor, J.M.G. (1988). The cost of generalizing logistic regression. *Journal of the American Statistical Association* **83**, 1078-1083.

Van Montford. M.A.J. and Otten, A. (1976). Quantal response analysis: enlargement of the logistic model with a kurtosis parameter. *Biometrische Zeitschrif* **18**, 371-380.

Whittemore, A.S. (1983). Transformations to linearity in binary regression. *SIAM Journal on Applied Mathematics* **43**, 703-710.

Received August 9, 2003
Revised March 15, 2004