Universitá degli studi di Napoli Federico II

Dottorato in Statistica (XXVI ciclo)

Diagnostic measures for Multinomial Distance Model

Maria Maddalena Giugliano

Promotor:

Prof. dr. Roberta Siciliano Universitá degli studi di Napoli Federico II

March, 2014

CONTENTS

1 INTRODUCTION

Qualitative data are more and more present in any field of research. For example, in medicine one can be interested in predicting an illness based on some symptoms, e.g. presence/absence of physical characteristics, (see *JAMA Internal medicine, www.archinte.jamanetwork.com*) , in psychology one can be interested in classifying mental status based on human behaviors, (e.g. Spinhoven et all., 2012, May 7), or in economy firms are interested in splitting customers into different groups based on their purchasing preferences to address marketing researches (e.g. Day et all., 1979). Many techniques are developed to handle these type of data. Qualitative data can be organized in contingency table, where each entry contains the joint frequency of subjects that have category g of one variable and category k of the other one. If more predictor variables are available, row or column categories can be define as a combination of categories of more than one variable.

Suppose that we are interested in the question whether predictors discriminate among response groups. One can chose a parametric analysis to test which main effect of and/or interactions between predictor variables are statistically significant in discriminating between the groups. But with this parametric approach it is difficult to understand the relations between the predictor patterns and group criteria due to the fact that there can be many predictors and combinations of them, so it is difficult to figure out all combinations and patterns related to all group criteria. On the other hand, one can apply multidimensional procedures, like multidimensional scaling or correspondence analysis, to obtain a graphical representation of the relations between predictor pattern and group criteria. In the latter case, we do not have a detailed model evaluation. Ideal Point Discriminant Analysis, proposed by

Takane (Takane et all., 1987), allows both a detailed model evaluation and a graphical representation of the data. It is a technique to classify subjects according to some criteria. This method has a lot of advantages. First, it allows a mixture of continuous and categorical predictor variables; second, it can be applied in conditional, joint and separate sampling procedures; third, it is justified under a wide class of distributional assumptions on predictor variables. Fourth, it maps together subject points and class points in the same Euclidean space and the probability of a subject to belong to class g is a decreasing function of the relative Euclidean distance between that subject and that class, compared to the distances towards the other classes. In maximum dimensionality, IPDA is equal to Multinomial Logistic Regression, but it also allows dimension reduction such as in Canonical Discriminant Analysis but without any assumptions on predictors. In the paper of 1998 about IPDA visualization (Takane, 1998), however, Takane highlighted that IPDA has some weaknesses in the visualization aspects. Mark de Rooij in the paper about the visualization problem of IPDA (De Rooij, 2009), proposed a modification of the model that overcomes those weaknesses. We chose to name this model Multinomial Distance Model. Nevertheless this model presents itself as a simple and good tool in discrimination problems, it suffers the lack of diagnostic statistics to evaluate not only the goodness of fit, but also the influential and leverage points.

The aim of this work has been to find tools to evaluate the Multinomial Distance Model, so that it can be come more popular and comparable with more famous models like the baseline category logit model. To understand the Multinomial Distance Model, the next section is devoted to explain Ideal Point Discriminant Analysis. Chapter 2 presents the Multinomial Distance Model in more details and gives a general overview on others models designed for categorical response variables. Chapter 3 is about diagnostics of generalized linear models and, in particular, it focuses on diagnostics of multinomial baseline category logit model. Chapter 4 presents diagnostic tools for Multinomial Distance Model while chapter 5 is about some applications on both simulated and real datasets. Finally, in chapter 6 there are some discussions about our findings.

1.1 Background: Ideal Point Discriminant Analysis.

Ideal Point Discriminant Analysis is based on three assumptions:

1. Subjects are mapped in a multidimensional Euclidean space and their coordinates are given by a linear combination of predictors;

2. Groups are represented by points and they are mapped together with subjects in the multidimensional Euclidean space;

3. The probability of a subject to belong to a response group is a decreasing function of the distance between the corresponding points and an increasing function of the prior probability of that criterion group.

Let N be the number of subjects and G the number of response categories (or groups). Let A denote the dimensionality of the representation space. According to Takane's model, the conditional probability that subject k belongs to group g, given the set of observations on the predictor variables is given by

$$p_g(\mathbf{x}_k) = \frac{w_g \exp(-\delta^2_{kg})}{\sum_{h=1}^{G} w_h \; exp(-\delta^2_{kh})}, \tag{1.1}$$

where w_g is the bias parameter for group g and δ is the Euclidean distance between subject k and category g:

$$\delta_{kg} = \left\{ \sum_{a=1}^{A} (\eta_{ka} - z_{ga})^2 \right\}^{\frac{1}{2}}, \tag{1.2}$$

where $\eta_{ka} = \mathbf{x}_k^\top \mathbf{b}_a$ is the coordinate of subject k on dimension a and z_{ga} is the coordinate of response category g on dimension a. These latter coordinates, can either be free, then we need to estimate them, or one can chose to apply centroid restriction, that means that z_{ga} is in the centroid of all subjects belonging to group g. In this way, z_{ga} ($g = 1, \ldots, G$) are a function of the parameters $\mathbf{b}_a$. The bias parameter w_g of the model is a sort of prior probability to belong to category g. To remove scale indeterminancy in w_g the restriction

$\sum_g w_g = 1$ is imposed. So, in the model (1.1) the probability is proportional to w_g for fixed δ_{kg} and proportional to $\exp(-\delta^2_{kg})$ for a fixed w_g. Model (1.1) is a special form of Coombs'(1964) unfolding model combined with Luce's (1959) individual choice model (Takane et all., 1987). The special feature is that the coordinates of subjects are contrained to be a linear function of the predictor variables.

The (conditional) likelihood of the model defined so far is

$$L = \prod_{k=1}^{N} \prod_{g=1}^{G} (p_g(\mathbf{x}_k))^{y_{kg}} \, , \tag{1.3}$$

where $y_{kg} = 1$ if the subject belongs to category g, $y_{kg} = 0$ otherwise. Fisher's scoring algorithm is applied to maximize the likelihood with respect to all parameters.

Model (1.1) can also be justified in other sampling situations, like joint sampling or separate sampling procedures, as long as the distribution of the predictor variables belong to the exponential family. In the same way, the likelihood is still valid if the distribution of the predictior variables leads to the conditional probability stated in equation (1.1).

1.2 Notation.

For further reference we supply some notational rules that we will follow. Bold uppercase symbols indicate matrices: **Y** is the response matrix, **X** is the predictor matrix. With **B** we indicate the coefficient matrix while **b** is the coefficient vector and **Z** indicates the class coordinate matrix. **P** is the probability matrix and p is a single probability. **H** is the generalized hat matrix while $\mathbf{M} = \mathbf{I} - \mathbf{H}$, where **I** is the Identity matrix. Bold lowercase symbols indicate vectors. Indices are denoted by lowercases: k is the subject index that goes from 1 to N, thus N is the sample size; $a = 1, \dots, A$ indicates the number of dimensions, $q = 1, \dots, Q$ is the number of predictors while $g = 1, \dots, G$ is the number of response categories. Greek cases indicate single coefficients. δ is the Euclidean distance. All other used symbols will be explained from time to time. Table 1.1 is a list of the main symbols used.

symbols	
Y	response variable
X	predictor variable
k	subject index
g	category index
a	dimension index
q	predictor index
N	sample size
G	number of response categories
A	number of dimensions
Q	number of predictors
v	number of parameters
z	class coordinate
α	intercept
β	coefficient
p	probability
δ	Euclidean distance
h	leverage value
r	standardized Pearson residual
e	studentized Pearson residual
rd	deviance residual
d	individual deviance
Δ	dfbeta
Δ^*	dfbetas
c	approximation to Cook's distance
$\bar{c}$	change in confidence interval measure
$\Delta\mathrm{D}$	goodness of fit sensitivity measure
$\Delta\mathrm{d}$	neighboring effect measure
$\Delta^*\mathrm{d}$	one step approximation of neighboring effect measure
$\widetilde{\alpha}$	*pseudo*-intercept
$\widetilde{\beta}$	*pseudo*-coefficient
$\mathbf{y}$	response vector
$\mathbf{x}$	predictor vector
$\mathbf{b}$	coefficient vector
$\mathbf{z}$	class coordinate vector
$\mathbf{p}$	probability vector
$\mathbf{r}$	standardized Pearson residual vector
$\mathbf{e}$	studentized Pearson residual vector
$\mathbf{Y}$	response matrix
$\mathbf{X}$	predictor matrix
$\mathbf{B}$	coefficient matrix
$\mathbf{Z}$	class coordinate matrix
$\mathbf{P}$	probability matrix
$\mathbf{H}$	hat matrix
$\mathbf{I}$	identity matrix
$\mathbf{M}$	influence matrix
$\widetilde{\mathbf{S}}$	*Pseudo*-coefficient matrix
$\widetilde{\mathbf{X}}$	*pseudo*-design matrix
$\mathrm{L}(.)$	likelihood function

Table 1.1: Table of symbols

Multinomial Distance Model 2

2.1 Categorical Data

Categorical variables are omnipresent in social sciences and biomedical sciences. A categorical variable is a measurement which consisting in a set of categories. For example, they arise in education, e.g. student responses to an exam, in marketing, e.g. consumer preferences among brands, in behavioral sciences, e.g. types of mental illness, and so on. In statistical science we distinguish between response categorical variable and explanatory categorical variable. The former is the output of an experiment measured on subjects of a sample while the latter is a feature of those subjects and we would predict, for example, the output based on some categorical and/or numerical features. Moreover, we distinguish between nominal categorical variables and ordinal categorical variables. An example of nominal variable is mental illness where there is no order between categories (e.g. depressed, schizophrenic and so on). On the other hand, an example of an ordinal variable is the number of mental disorders of a subjects, e.g. one disorder, two disorders, ..., that is, a variable where for each category it is possible to estabilish wheter it is greater or smaller than the others. In statistics there are models that handle nominal response variables such as the baseline category logit model, and other methods which handle ordinal response variables such as the proportional odds model. Historically, there are more methods that handle numerical variables because they are easier to manage than qualitative data. In fact, relatively little development of models for categorical response variables occurred until 1960 (Agresti, 2002).

The main distributions to describe categorical response variables are the

Binomial distribution, the *Multinomial distribution* and the *Poisson distribution*. Let $y_1, y_2, \ldots, y_N$ be responses for N independent and identical trials where the probability that $y_k = 1$ is p and the probability that $y_k = 0$ is $1 - p$. Trials are identical due to the fact that the probability p is the same at each trial. Furthermore, trials are independent because y_k are indipendent random variables. Each trial is a *Bernoulli* variable which is a special case of *Binomial* distribution when the number of trials is 1 ($n = 1$). Let $y_+ = \sum_k y_k$ be the number of successes in N trials. Its probability mass function is

$$p(y_+) = \binom{N}{y_+} p^{y_+}(1-p)^{N-y_+},$$

with expected value equals to Np and variance equals to $Np(1-p)$. The likelihood function can be easly computed, using the product over subjects of the probability mass function. This distribution describes binary response variable, that is, variables that have as outcome 0 or 1.

When the response variable is multicategorical, that is, it has $G > 2$ categories, its distribtuion is approximated by the *multinomial* function. Then, $y_{kg} = 1$ if subject k belongs to category g and $y_{kg} = 0$ otherwise. The sum over single trial (subject) is equal to 1. Let n_g be the number of outcomes in category g,with $\sum_g n_g = N$. The counts $n_1, n_2, \ldots, n_G$ have *multinomial distribution*. Its probability mass function is

$$p(n_1, n_2, \ldots, n_{G-1}) = \left(\frac{N}{n_1! n_2! \ldots n_G!}\right) p_1^{n_1} p_2^{n_2} \ldots p_G^{n_G}.$$

We do not put n_G in the left part of the formula due to the fact that y_{kG} is redundant, being linearly dependent on the others. The expected value $E(p_g) = Np_g$ and the variance $Np_g(1-p_g)$. Note that the *binomial distribution* is a special case of the *multinomial distribution* with $G = 2$.

When there is not a fixed upper limit N for y, the categorical response variable is approximated by the *Poisson distribution*. An example is the number of calls per minute in a call center. Its probability mass function is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad \textit{with } y = 1, 2, \ldots$$

The expected value and variance are the same and equal to μ.

In the next chapters we will work with multicategorical response variables, then the distribution that we will use is the *multinomial distribution*. The likelihood function of a *multinomial distribution* is

$$L(\mathbf{p}) = \prod_g p_g^{n_g}, \qquad \textit{with } g = 1, \ldots, G \textit{ and } \sum_{g=1}^{G} p_g = 1.$$

In most cases it is easier to work with the $\log -likelihood$, which is

$$\log L(\mathbf{p}) = \sum_{g=1}^{G} n_g \log(p_g).$$

The maximum likelihood estimate of $\widehat{p}_g$ is the sample proportion n_g/N, where $\sum_g n_g = N$.

Starting from multicategorical response variables, the next sections introduce some methods to model nominal or ordinal response variables. Section 2.2 introduces the Multinomial Distance Model while section 2.3 is about the Multinomial Logit Models.

2.2 *Multinomial Distance Model*

Starting from IPDA model, many modifications are allowed to generalize it and make it to be versatile. Among the possible choices, it is possible to modify

1. the bias parameters w_g (e.g. $w_g = 1$);
2. the group coordinates z_{ga} (e.g. z_{ga} can be set to be free);
3. the distance δ^2_{kg} (e.g. the squared Euclidean distance in the exponential may be replaced by the simple Euclidean distance or a Mahalanobis distance).

In the paper of Mark De Rooij about the visualization problem of IPDA (De Rooij, 2009), it is shown that when the dimensionality is $A = G - 1$, the prior probability (the bias parameter w_g) can be incorporated in the distance part of the model. In this way, the model has a much clearer interpretation because the decision boundaries are solely based on distance, thus they are orthogonal to the line joining two class points and through their centroid. In the simplest case, with binary response variable Y and one predictor X, a distance model can be built in one dimensional Euclidean space ($A = 1$). Let z_0 and z_1 be the coordinates of response categories. Let $\eta_k = \alpha + x_k\beta$ be the coordinate of subject k. Now we can define two distances: one between subject k and category 1, and another one between subjct k and category 0. The corresponding squared Euclidean distances are

$$\begin{aligned}\delta_{k1}^2 &= (\eta_k - z_1)^2\,,\\ \delta_{k0}^2 &= (\eta_k - z_0)^2\,. \end{aligned} \tag{2.1}$$

For subject k, we can write down the probability to belong to category 1

$$p_1(x_k) = \frac{\exp(-\delta_{k1}^2)}{\exp(-\delta_{k1}^2) + \exp(-\delta_{k0}^2)},$$

and the probability to belong to category 0

$$p_0(x_k) = \frac{\exp(-\delta_{k0}^2)}{\exp(-\delta_{k1}^2) + \exp(-\delta_{k0}^2)}. \tag{2.2}$$

These probabilities are decreasing functions of the relative squared Euclidean distances. This is, the probability of subject k to belong to category 1 is inversely related to the distance between subject k and category 1. So, if the distance towards category 1 is greater than the distance towards category 0, a subject with observed value x_k has larger probability to belong to category 0 rather than to belong to category 1. As any distance model, also multinomial distance model has the idetification problem. To fix it, in one dimensional Euclidean space two restrictions on the group points are needed and we can chose $z_1 = 1$ and $z_0 = 0$. In general, multinomial distance model, in more than one dimensions, could also have others problems like translation indeterminacy that can be solved putting a specific class in the origin of the space, fixing $z_1a = 0$. There is a rotation indeterminacy. Rotation keeps the distances the same, thus the probabilities and the likelihood should be the same. This problem can be fixed setting the upper triangular part of the class coordinate matrix equals to 0. To see if restrictions are needed, an empirical approach could be used. One can fit the model without restrictions and storing log-likelihoood and parameter estimates. Fitting again the model but from different starting point. If the log-likelihood and parameter estimates are the same, the model does not need identifications. Otherwise, we have to use an identification restriction.

To make the interpretation easier, we can explicit the model in terms of log odds instead of probabilities. Under the distance model with binary response variable and one predictor variable X, for subject k we have

$$\begin{aligned}\log\left(\frac{p_1(x_k)}{p_0(x_k)}\right) &= \delta_{k0}^2 - \delta_{k1}^2 \\ &= 2\eta_k(z_1 - z_0) + z_0^2 - z_1^2 \\ &= 2(\alpha + x_k\beta)(z_1 - z_0) + z_0^2 - z_1^2 \\ &= 2\alpha(z_1 - z_0) + 2x_k\beta(z_1 - z_0) + z_0^2 - z_1^2.\end{aligned} \tag{2.3}$$

This formulation highlights the role of the group point coordinates. In fact, the term $2x_k\beta(z_1 - z_0)$ indicates the change of log odds for one unit increasing in X. If the the distance between two group points is large, then the change in log odds is large.

Distance model as defined so far, can be easly generalized to polytomous response variables. Let Y be a polytomous response variable with G response categories. The probability to belong to one of the G categories, given the predictor X, is given by

$$p_g(x_k) = \frac{\exp(-\delta_{kg}^2)}{\sum_{h=1}^{G}\exp(-\delta_{kh}^2)}, \qquad \textit{with } g = 1, \ldots, G, \tag{2.4}$$

where ${\delta_{kg}}^2$ is the squared Euclidean distance between the subject coordinate k and the class point g. In terms of log odds and assuming an unidimensional solution we have

$$\begin{aligned}\log\left(\frac{p_g(x_k)}{p_G(x_k)}\right) &= \delta_{kg}^2 - \delta_{kG}^2 \\ &= 2\alpha(z_g - z_G) + 2x_k\beta(z_g - z_G) + z_G^2 - z_g^2,\end{aligned} \tag{2.5}$$

Again, z_g represents the coordinate for category g in a one-dimensional Euclidean space, with $g = 1, \ldots, G$. For identification one restriction $z_G = 0$ is needed.

The probabilities can also be expressed in an alternative form

$$\begin{aligned} p_g(x_k) &= \frac{\exp\left(2\eta(z_g) - z_g^2\right)}{1 + \sum_{g=1}^{G-1} \exp(2\eta(z_g) - z_g^2)}, \\ &= \frac{\exp(u_g)}{1 + \sum_{h=1}^{G-1} \exp(u_h)}, \qquad \textit{with } g = 1, \ldots, G \end{aligned} \tag{2.6}$$

where:

$$u_g = 2\eta(z_g) - z_g^2,$$

that is different from Euclidean distance defined before in (2.1). If we consider the last category as reference category on which we put the restriction, consequently the formula (2.6) is simplified, too.

This model deals with ordinal response variable too, but there is no constrain to ensure the ordinality of the response categories. In Proportional Odds Model the latent variable justification and the fixed effect β give ordered categories, but this model uses the cumulative probabilities that are more difficult to interpret than the single probabilities. Adjacent category logit model in its proportional odds form, ensures that the model accounts for the ordinality, but the proportional assumption does not often hold in real situations.

Similar to the contrain on ϕ parameters in the Stereotype model, to ensure the ordinality of the response categories in the Multinomial Distance model, we introduce the same contrain on the group point coordinates

$$z_1 \leq z_2 \leq ... \leq z_G = 0. \tag{2.7}$$

If some z are equal it means that the correspondent categories are not distinguishable by the predictors, so it is better to collapse them into one single category.

2.2.1 Likelihood, Estimation Parameters and Model Assessment

It is assumed that the responses of subjects are independent multinomial distributed, so that the log-likelihood is:

$$\log - L = \sum_k \sum_g y_{kg} \log p_g(x_k). \tag{2.8}$$

Equation (2.8) is maximized with respect to model parameters $(\alpha, \beta, \mathbf{z})$, subject to the identification constraints, using a Quasi-Newton algorithm. Once we have parameter estimates, the probability to belong to each of the response categories can be computed for a new subject k^* with observed predictor value x_{k*}. Finally, subject k^* can be assigned to the group with highest probability, i.e. $\widehat{y}_{k^*g} = max_g p_g(x_k)$. In the Appendix all R-code to estimate the model described so far are supplied.

Once the likelihood and parameters are estimated, the general goodness of fit of the model needs to be evaluated. Therefore information criteria, like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be computed to assess the goodness of fit with respect to the complexity of the model. Phenomena are complex, and a statistical model is a general representation of them. Obviously, the more complex the model is, the better the model represents phenomena but it is more complicated to find and to explain the relationships between the components of such a phenomenon. This problem is called curse of dimensionality, that is a good model should find the best trade off between the model complexity and the model power to explain phenomena.

AIC and BIC give a relative assessment about how good the model is based on its complexity in terms of the number of parameters to be estimated, with respect to others models. Both statistics are likelihood based. AIC is given by

$$AIC = -2LL + 2v,$$

this is, minus two times the log-likelihood plus two times the number of parameters v. We choose the model with smallest AIC value.
Bayesian Information Criterion is given by

$$BIC = -2LL + log(N)v,$$

where v is the number of parameters and N is the number of observations. BIC has a larger complexity penalty than AIC, due to the fact that it uses the logarithm of N multiplied by the number of parameters. Again, the model with the smallest BIC value is preferred.

When a model fits poorly it is useful to look at residuals and the configuration of the points in the space spanned by predictors to find where the fit is poor. As we have said,for the Multinomial Distance model there is no literature about diagnostic statistics. In chapter 4 we provide some diagnostic tools suitable for the Multinomial Distance model to assess residuals as well as influential and leverage points.

2.3 *Multinomial Logit Models*

Classification problems can be conceived as regression problems where the response variable is categorical. Thus, we can fit Logistic Regression for binary categorical response variable or Multinomial Logistic Regression for multi-categorical response variable. In many real problems, especially in social science, response variable can be ordered and one can also be interested in the order of the response groups. In this section we describe both nominal and ordinal regression models.

2.3.1 *Baseline Category Logit Model*

Let Y a nominal response variable with G categories. Let $p_g = P(Y = g|X)$ be the probability to belong to category g given fixed predictor variable X. From $\mathbf{p} = (p_1, \dots, p_G)$ we can form $G(G-1)/2$ set of odds which are

$$\frac{P(Y=g)}{P(Y=h)} = \frac{p_g(x_k)}{p_h(x_k)}, \qquad \textit{with } h \neq g = 1, \dots, G$$

Choosing $G-1$ odds the others are redundant because they can be computed from the formers (Agresti, 2002). The baseline category logit model compares each category with a baseline category, usually the last one. Then, the model is

$$\log\left(\frac{p_g(x_k)}{p_G(x_k)}\right) = \alpha_g + \beta_g x_k, \tag{2.9}$$

From equation (2.9) probabilities are

$$p_g(x_k) = \frac{\exp(\alpha_g + \beta_g x_k)}{1 + \sum_{h=1}^{G-1} \exp(\alpha_g + \beta_g x_k)}. \tag{2.10}$$

The denominator of equation (2.10) is the same for each probability and the numerators for all $G-1$ sum up to the denominator, therefore $\sum_g p_g(x_k) = 1$.

To estimate those probabilities we use maximum likelihood theory. From equation (2.9) we obtain $G-1$ regression equations to solve simultaneously. For a sample of size N, let $y_k = (y_{k1}, y_{k2}, \ldots, y_{kG})$ be the multinomial trial for subject k. Let $y_{kg} = 1$ if subject k belongs to category g and $y_{kg} = 0$ otherwise. Thus, $\sum_g y_{kg} = 1$. The $\log -likelihood$ of the data are

$$\log \prod_{k=1}^{N} \left(\prod_{g=1}^{G} p_g(x_k)^{y_{kg}} \right) = \\ \sum_{k=1}^{N} \left[\sum_{g=1}^{G-1} y_{kg}(\alpha_g + \beta_g x_k) - \log \left(1 + \sum_{g=1}^{G-1} \exp\left(\alpha_g + \beta_g x_k\right) \right) \right]. \quad (2.11)$$

Using iterative procedures (like Newton-Raphson method) we obtain estimates of α_g and β_g which maximize the $\log -likelihood$.

2.3.2 *Proportional Odds Model*

Suppose we have a multi-categorical ordinal response variable Y. We may be interested in modeling Y as a function of a predictor variable X. Furthermore, the observed scale scores on Y are assumed to be discretized measurements on an continuous latent response variable Y^*. Suppose that $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < ... < \alpha_G = \infty$ are cutpoints of the continuous scale such that the observed response Y satisfies

$$Y = g \qquad \text{if } \alpha_{g-1} < Y^* < \alpha_g.$$

Thus, Y falls in category g when the latent variable assumes values in the interval defined by α_{g-1} and α_g. The general form of the probability model is

$$P(Y \leq g|x) = P(Y^* \leq \alpha_g|x) = \Phi(\alpha_g - \beta' x), \quad (2.12)$$

where Φ is some invertible function. If we assume that

$$Y^* = \beta x + \epsilon$$

and we specify Φ as standard logistic function for ϵ, and apply its inverse function, that is logit link, to the probability, then we obtain the Proportional

Odds Model (McCullagh, 1980)

$$logit[P(Y \leq g|x)] = \alpha_g + \beta' x, \qquad with\ g = 1, ...G-1, \tag{2.13}$$

that is the logit of the probability to belong to one of the categories less or equal to g. The complement of the latter probability is the probability to belong to a category greater than category g. Each of cumulative logits has its own intercept, that is the estimated cutpoint. Given its formulation, the logit is an increasing function of the probability to belong to one of the categories less or equal to g. The reason is that α_g increases in g due to the fact that the latter probability increases in g for fixed x. All logits share the same effect β and then the response curves have the same shape, but they are shifted by $(\alpha_h - \alpha_g)/\beta$ in the x direction, for categories $g < h$. In terms of odds ratios, for fixed category g and two different values of a predictor variable x_1 and x_2 we have

$$\begin{aligned} logit[P(Y \leq g|x_1)] - logit[P(Y \leq g|x_2)] &= \log \frac{P(Y \leq g|x_1)/P(Y > g|x_1)}{P(Y \leq g|x_2)/P(Y > g|x_2)} \\ &= \beta(x_1 - x_2). \end{aligned} \tag{2.14}$$

The odds of making response $\leq g$ at $x = x_1$ is $\exp[\beta(x_1 - x_2)]$ times the odds at $x = x_2$. Because the β parameter is invariant to the cutpoints, the odds ratios are the same over the $g-1$ cumulative probabilities. Equation (2.14) shows that odds ratios are proportional to the distance between the values of x, i.e., the same proportionality constant applies to each logit (Agresti, 2002). The proportional odds model assumes that the covariate effects are invariant to the cutpoints, thus impying proportionality in the odds ratios. Often this assumption does not hold in real problems, that means this kind of model does not suit them. Many alternatives have been proposed such as Unconstrained Partial Proportional Odds Model (Peterson and Herrell, 1980) which estimates two set of parameters, one for proportional odds, and the other one for non-proportional odds.

2.3.3 Adjacent Category Logits Model

An alternative way to overcome the proportionality assumption is to fit the adjacent categories logit model (Simon, 1974; Goodman, 1983). Let p_g $(g = 1, \ldots, G)$ be the probability to belong to response category g with multinomial distribution, the adjacent categories logits are

$$logit[P(Y = g|Y = g \quad or \quad Y = g+1)] = \log \frac{p_g(x_k)}{p_{g+1}(x_k)}, \qquad g = 1...G-1.$$

Let x be a predictor variable, the general adjacent categories logit model is

$$\log \frac{p_g(x_k)}{p_{g+1}(x_k)} = \alpha_g + \beta_g x_k. \tag{2.15}$$

Equation (2.15) can also be viewed as a different parametrization of baseline category logit model. Consider the baseline category logits

$$\log \frac{p_1(x_k)}{p_G(x_k)}, \log \frac{p_2(x_k)}{p_G(x_k)}, ..., \log \frac{p_{G-1}(x_k)}{p_G(x_k)},$$

each baseline category logit can be expressed in terms of adjacent categories logits

$$\log \frac{p_g(x_k)}{p_G(x_k)} = \log \frac{p_g(x_k)}{p_{G+1}(x_k)} + \log \frac{p_{g+1}(x_k)}{p_{g+2}(x_k)} + ... + \log \frac{p_{G-1}(x_k)}{p_G(x_k)}.$$

Thus, the baseline category logit model can be expressed in terms of model (2.15) as

$$\begin{aligned} \log \frac{p_g(x_k)}{p_G(x_k)} &= \sum_{h=g}^{G-1} \alpha_h + \left(\sum_{h=g}^{G-1} \beta_g \right) x_k \\ &= {\alpha_g}^* + {\beta_g}^* x_k, \end{aligned} \tag{2.16}$$

with $g = 1...G-1$. In this case, no common effect is assumed for each g, thus the model does not utilize the ordinality of Y.

One could also assume that a predictor variable has the same effects over response categories. Thus, we obtain a model similar to (2.15), but with fixed β, that is $\beta_1 = \beta_2 ... = \beta_g = \beta$. The model will be

$$\log \frac{p_g(x_k)}{p_{g+1}(x)} = \alpha_g + \beta x_k. \tag{2.17}$$

This model has proportional odds like Proportional Odds Model and both models fit well in similar situations due to the fact that they assume stochastically ordered distributions of Y at different predictor values (Agresti, 2010, p.89). In fact, the odds ratios are the same for each pair of adjacent categories, thus they do not depend on g.

About the interpretation, for a fixed predictor X, the estimated odds of the lower instead of the higher of two adjacent categories $p_g(x_k)/p_{g+1}(x_k)$ multiplies by $\exp(\beta)$ for every one unit increase in X. If we consider the adjacent categories logit model with common effects β, the equivalent baseline category logit model is

$$\begin{aligned}\log \frac{p_g(x_k)}{p_G(x_k)} &= \sum_{h=g}^{G-1} \alpha_h + (G-g)\beta x_k \\ &= {\alpha_g}^* + \beta u_g,\end{aligned} \tag{2.18}$$

where $u_g = (G-g)x_k$. So, the Adjacent categories logit model corresponds to a baseline category model with an adjusted model matrix. This model takes into account the ordinality of Y, using a single common parameter β for each predictor variable and letting the predictor variable itself incorporates a distance measure $G-g$ between each category g and the baseline category G (Agresti, 2010). This connection is important for ML estimate of parameters in adjacent categories logit model. In fact, the parameter estimates of the adjancent category logit model can be obtained from the estimate parameters of baseline category logit model. It can be shown that

$$\widehat{\beta}_g = \widehat{\beta}^*_g - \widehat{\beta}^*_{g+1}\,,$$

where $\widehat{\beta}^*_g$ are the estimated parameters of baseline category logit model.

2.3.4 *Continuation-Ratio Model*

Another alternative logit model is the continuation-ratio logit model. Consider the continuation-ratio log odds for each category relative to the higher categories

$$\log \frac{p_g(x_k)}{p_{g+1}(x_k) + \cdots + p_G(x_k)}, \qquad \textit{with } g = 1, \ldots, G-1, \tag{2.19}$$

or the log odds for each category relative to the lower categories

$$\log \frac{p_{g+1}(x_k)}{p_1(x_k) + \cdots + p_g(x_k)}. \tag{2.20}$$

Equation (2.19) is the ordinary logit of the probabilities

$$\omega_g = P(Y = g|Y \geq g) = \frac{p_g}{p_g + \cdots + p_G}, \qquad \textit{with } g = 1, \ldots, G-1.$$

Thus, sequential logits ca be defined (Agresti, 2010, p.97)

$$\log\left(\frac{\omega_g}{1-\omega_g}\right),$$

with explanatory variables, the continuation-ratio logit model using sequential logits is

$$logit\left[\omega_g\left(x_k\right)\right] = \alpha_g + \beta_g x_k. \tag{2.21}$$

If we assume proportionality for odds, then we have the same model (2.21) but with common parameter β for all response categories. This model is useful when a sequential process determines the response variable. An example is the survival of a subject after receiving a medical treatment. As in proportional odds model, the continuation-ratio logit model find a motivation in a latent variable underlying the observed ordinal response variable (Tutz, 1991). It is assumed that latent variable Y^* satisfies

$$Y^* = \beta x_k + \epsilon,$$

where ϵ that follows a cumulative distribution function Φ. For a set of thresholds (α_g), the observed ordinal response variable satisfies

$$Y = g \qquad \textit{given} \quad Y \geq g, \qquad \textit{if} \quad Y^* \leq \alpha_g.$$

The sequential mechanism assumes a binary decision at each step. Only the final resulting category is observable. The general model will be:

$$P(Y = g|Y \geq g) = \Phi(\alpha_g - \beta x_k). \tag{2.22}$$

An important feature of this model is the multinomial factorization with sequential probabilites. Let x_k be the value for subject k on predictor X. Let

$(y_{kg}, g = 1, \ldots, G)$ be the response vector of subject k, with $y_{kg} = 1$ if the subject belongs to category g and 0 otherwise. Then $\sum_g y_{kg}$=1. Let $b(n, y; \omega)$ be the binomial probability of y successes in n trials with parameter ω in each trial. The multinomial mass function of a single observation $(y_{k1}, y_{k2}, ...y_{kG})$ can be factorize in

$$b\left[1, y_{k1}; \omega_1\left(x_k\right)\right] b\left[1 - y_{k1}, y_{k2}; \omega_2\left(x_k\right)\right] \cdots \\ b\left[1 - y_{k1} - \cdots - y_{k,G-2}, y_{k,G-1}; \omega_{G-1}\left(x_k\right)\right]. \quad (2.23)$$

The log likelihood is the sum of the logarithms of all multinomial mass functions for different values of x_k, such that different ω_g enter into different terms.

2.3.5 *Stereotype Model*

For all the models discussed so far, some problems arise. In fact, when proportionality for odds does not hold, adjacent categories model, continuation-ratio model and cumulative model assuming constant β fit poorly (Agresti, 2010, p.103). One can use adjacent categories model in its general form which allows for different effects for each response category. But this general model corresponds to the baseline category model which treats the response variable as nominal. Furthermore, the number of parameters increases in G or with the number of the predictors. Anderson (1984) proposed a category logit model, called stereotype model, which is nested between the adjacent categories logit model with the proportional odds and its general form (2.15). The stereotype model is

$$\log \frac{p_g(x_k)}{p_G}(x_k) = \alpha_g + \phi_g \beta x_k, \qquad \textit{with } g = 1, \ldots, G-1. \quad (2.24)$$

In terms of response probabilities we have

$$p_g(x_k) = \frac{\exp(\alpha_g + \phi_g \beta x_k)}{\sum_{g=1}^{G} \exp(\alpha_g + \phi_g \beta x_k)}, \quad (2.25)$$

with restrictions $\alpha_G = 0$, $\phi_G = 0$ and $\phi_1 = 1$. For a one unit increase in predictor X, the odds of response g instead of response G is $\exp(\phi_g \beta)$ times larger. This model is more parsimonious then the models described so far. Compared with model (2.16) Anderson's model has less parameters to estimate. In fact, here $G-1$ intercepts, $G-2$ ϕ parameters and one β for each predictor variable considered in the model need to be estimate.

The model can easly be write in baseline category logit model. In fact, setting

$$\beta_g^* = \phi_g \beta,$$

for all categories, we can write the model as

$$\log \frac{p_g(x_k)}{p_G(x_k)} = \alpha_g + \beta_g^* x_k$$

The stereotype model can model ordinal data. The parameters ϕ can be viewed as scores for the response categories. The constraint

$$1 = \phi_1 \geq \phi_2 \geq ... \geq \phi_G = 0.$$

allows the model to treat Y as ordinal. The monotonicity of the ϕ parameters also implies that the effect of a single predictor has the same direction for each pairs of categories. Thus, a given predictor X has uniformely positive or negative local log odds ratios with Y. Anderson noted that the higher the value of βx the more the distribution of Y moves to the low of the response scale. Thus, to make sure that for a positive values of β correspond to a positive effect of the predictor, one can write the model as

$$\log \frac{p_g(x_k)}{p_G(x_k)} = \alpha_g - \phi_g \beta x_k. \tag{2.26}$$

Furthermore, this model allows to verify if response categories are distinguishable with respect to the predictor variables. In fact, if two ϕ parameters are equal, the corresponding categories can be collapsed into one, and then the model can be refitted.

So far, we decribed models to discriminate among response groups based on a set of predictor variables. As we could see, all these models allow for detailed effect evaluations, but they do not give a graphical representation of the data. The Multinomial distance model deals with this weakness, providing both detailed model evaluation and a graphical representation.

DIAGNOSTICS IN GENERALIZED LINEAR MODELS 3

3.1 *Generalized Linear Models*

A Generalized Linear Model (GLM) extends an ordinary regression model to cover non normal response distributions. GLMs consist of three components

1. A *random component*, specifying the conditional distribution of the response variable Y, given the values of the explanatory variables. These distributions come from the exponential family which has density

$$f(y_k|\theta_k,\varphi) = \exp\left[\frac{y_k\theta_k - b(y_k)}{a(\varphi)} + c(y_k,\varphi)\right].$$

   θ_k is the canonical parameter and represent the location while φ is the dispertion parameter and represents the scale. Some important distributions come from the exponential family, like the Gaussian, the Binomial and the Poisson specifying functions a, b and c. GLMs are also extended to the multivariate exponential family, like the multinomial distribution. If φ is known, the function can be simplified to

$$f(y_k|\theta_k) = a(\theta_k)b(y_k)\exp\left[y_kQ(\theta_k)\right].$$

2. *Systematic component*, which is the linear predictor η_k given by

$$\eta_k = \sum_{q=1}^{Q} \beta_q x_{kq}, \qquad k = 1,\ldots,N$$

   with $q = 1,\ldots,Q$ predictor variables.

3. *Link function* $g(.)$, which transforms the expectation of the response variable $\mu_k = E(Y_k)$ to the linear predictor

$$g(\mu_k) = \eta_k = \sum_{q=1}^{Q} \beta_q x_{kq}$$

   The link function must be invertible, such that $\mu_k = g^{-1}(\eta_k)$. The inverse link $g^{-1}(.)$ is also called the mean function.

When the conditional distribution of the response variable is binomial and the link function is the *logistic* function then we have classical logistic regression. When the distribution of the response is multinomial and the link function is again *logistic* it is a multinomial logit model.

Maximum likelihood theory is used to estimate parameters. In general, for a GLM we have

$$\mu_k = E(Y_k) = b'(\theta_k) \qquad\qquad var(Y_k) = b''(\theta_k)a(\phi)$$

where $b'(\theta_k)$ is the first derivative of the function $b(.)$ and $b''(.)$ is the second derivative. In practice, several functions do not have a closed form, then they are maximized using iterative procedures like Fisher scoring algorithm or Newton-Raphson algorithm.

After fitting the model, we have to evaluate it. When a model fits poorly it is useful to look at residuals to find where the fit is poor. It is important to distinguish between outliers, leverage and influential points. An outlier is an observation whose response value is unusual, given the value of predictor variable. There are three different cases:

1. the outlier has predictor value in the center of the predictor distribution. In this case, deleting the outlier has low impact on regression results, that means that this observation has low leverage and a little influence;

2. the outlier has predictor value far from the predictor mean. This outlier has high leverage and substantial influence on the regression results. It is a regression outlier;

3. the outlier has a predictor value far from the mean but it is in line with the rest of the data. This observation has large leverage but does not have influence on the regression analysis.

In the Generalized Linear Model framework some unusual and influential diagnostic measures have been built (Pregibon, 1981). Lesaffre and Albert (1989) extended univariate diagnostic tools to multiple group logit models. The next two sections propose an overview on both cases.

3.2 Diagnostics for Univariate Generalized Linear Models

For a simple logistic regression model, to evaluate the fit and to identify outliers and influential points, the residual vector and a projection matrix are needed. To assess the leverage we need the so called *generalized hat values*. The name is due to the fact that we can express the fitted values $(\widehat{y}_k)$ in terms of the observed values y_k. In matrix notation we have

$$\widehat{\mathbf{Y}} = \mathbf{HY}$$

where **H** is the *generalized hat* matrix, or projection matrix given by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{1/2}, \tag{3.1}$$

where **W** is a diagonal matrix with elements $w_{kk} = \widehat{p}_k(1-\widehat{p}_k)$ (for ungrouped data). The square matrices **H** or $\mathbf{M} = \mathbf{I} - \mathbf{H}$, where **I** is the identity matrix, are idempotent and symmetric. Furthermore, if the Pearson residuals are multiplied by **M**, the result is again Pearson residuals (Lesaffre and Albert, 1989). Taking the diagonal values of **H** we obtain a measure of leverage for each subject, ranging from 0 to 1. The *generalized hat values* close to 1 have high leverage, this is, they are extreme points in the design space. However, because **H** depends on both the design matrix and the fit, extreme points in design space do not necessary have high value of h_{kk} $(k = 1, \ldots, N)$. The same considerations are also valid for matrix **M**, but in this case leverage points have values of m_{kk} close to 0. Plots of residuals and h_{kk} against subject indexes are useful to detect outliers.

Unlike linear regression model, where the residuals are uniquely defined, in logistic regression it is possible to define different types of residuals, based on several scales (Pregibon, 1981). The most useful residuals are the Pearson residuals and the Deviance residuals. Let y_k be the observed response variable, with $k = 1...N$. Let $\widehat{p}_k$ be the estimate of $P(Y = 1|X = x_k)$, then $\widehat{p}_k$ is the fitted response variable. The *Standardized Pearson residuals* are

$$r_k = \frac{y_k - \widehat{p}_k}{\sqrt{\widehat{p}_k(1-\widehat{p}_k)}}.$$

The numerator of the above expression is called the *raw residual*. When $N \to \infty$ the covariance of *raw residuals* $y_k - \widehat{p}_k$ can be approximated by

$$cov(y_k - \widehat{p}_k) = \widehat{p}_k(1-\widehat{p}_k)(1-h_{kk}) = \widehat{p}_k(1-\widehat{p}_k)m_{kk},$$

where h_{kk} are the diagonal elements of the *generalized hat matrix*. Therefore, the asymptotic covariance of *Standardized Pearson residuals* is 1. Dividing the raw residuals by their asymptotic covariances, we obtain *Studentized Pearson residuals*

$$e_k = \frac{(y_k - \widehat{p}_k)}{\sqrt{\widehat{p}_k(1-\widehat{p}_k)(1-h_{kk})}} = \frac{r_k}{\sqrt{(1-h_{kk})}}.$$

Absolute values of e_k larger than 2 or 3 provide evidence of lack of fit (Agresti, 2002).

Deviance residuals measure the agreement between the observed and fitted log-likelihoods of subject k. They are given by

$$rd_k = \sqrt{(d_k)} \times sign(y_k - \widehat{p}_k),$$

where

$$d_k = -2LL_k = -2\left(y_k \log \widehat{p}_k + (1-y_k)\log\left(1-\widehat{p}_k\right)\right),$$

that is, the deviance for subject k. A plot of residuals against predictor variables may detect lack of fit.

The residuals and the projection matrix help to identify outliers, but they do not indicate the extent to which they affect the parameter estimates. To appraise the influence of outliers case deletion methods can be used. Essentially, it is possible to compute some measures which evaluate the influence of a case, by deleting that case from the analysis and comparing the estimates of full model with the estimates of the model fitted without that case. If the difference between the estimate of such a parameter based on full data and the estimate after deleting observation k is large, it means that case k has influence on the estimation process. This can be done for each observation and each parameter. However, this becomes computationally intensive. Pregibon (1981) proposed to approximate the estimate of parameters after deleting case k, using the so-called one step estimate. The vector of estimated parameters after deleting the kth observation, is obtained from the estimation equation of the iterative procedure (e.g. Newton-Raphson method), using the estimated parameters based on the full data as starting point and terminating after one step. Let $\widehat{b}_q$ be the estimated coefficient of predictor q. The change in individual coefficients, when dropping subject k, is measured by

$$\Delta_{qk} = \widehat{b}_q - \widehat{b}_{q(-k)} \qquad \textit{with } k = 1, \ldots, N \textit{ and } q = 1, \ldots, Q,$$

which is called *dfbeta*. The larger the values are, the more influence the case k has on the coefficients. A standardized version of Δ_{qk}, it is obtained by dividing it by its standard error, called *dfbetas*. The problem related to this statistic is that the number of *dfbetas* grows with the number of subjects and the number of predictors.

An overall discrepancy measure between $\widehat{b}_q$ and $\widehat{b}_{q(-k)}$ is *Generalized Cook's distance*. It is a sort of test for the hypotesis that $\widehat{b}_q = \widehat{b}_{q(-k)}$. An one step approximation to the *Generalized Cook's distance* is given by

$$c_k = \frac{r_k^2 h_{kk}}{(1 - h_{kk})^2}.$$

There are many interpretations of c_k. We prefer interpreting it as a measure of the change of the confidence region of plausible values for parameters, computed including subject k. Graphically, it can be represented as a circle with radius equal to the Cook's value. Another useful plot is the index plot obtained by plotting subject indexes versus c_k to see for what points Cook's distance is larger. Note that c_k is a mixture between a discrepancy measure, the standardized Pearson residuals, and the leverage value.

A similar measure of c_k can be computed which is given by

$$\bar{c}_k = \frac{r_k^2 h_{kk}}{1 - h_{kk}}.$$

It express the same diagnostic of c_k but it indicates how the confidence interval changes including case k. Pregibon (1981) showed that the one step approximation of $\bar{c}_k$ is better than c_k.

To evaluate the sensitivity of the goodness of fit, another diagnostic to evaluate the influence of subject k on the global goodness of fit of the model is

$$\Delta_k D = d_k^2 + \bar{c}_k$$

where d_k^2 is the individual deviance for subject k. The interpretation of this statistic is the change in deviance attributable to deleting subject k. If the value of $\Delta_k D$ for subject k is large it means that by deleting that case the fit gets worse.

Finally, we can also assess the effect of each subject on the classification of the other subjects. Pregibon (1981) proposed another tool that evaluates the *neighboring effects* by measuring the difference between the probability of subject j and the same probability computed after deleting another subject k. Considering the individual deviances, we can write

$$\Delta_k d_j^2 = d_j^2 - d_j^2(-k)$$

and its one step approximation is given by

$$\Delta_k^* d_j^2 = \frac{2 r_j h_{kj} r_k}{1 - h_{kk}} + \frac{r_k^2 h_{kj}^2}{(1 - h_{kk})^2}.$$

When $\Delta_k^* d_j^2 > 0$ the fit of case j gets worse if we delete case k. It it is equal to 0 the fit is the same and if it is smaller than 0 the fit of case j gets better. It is noteworthy that $\Delta_k^* d_j^2 \neq \Delta_j^* d_k^2$. To have a summary measure of this effect, it is possible to sum over subjects and obtain $\sum_{j \neq k} \Delta_k^* d_j^2$. If this sum is smaller than 0 it means that by deleting case k the fit should improve.

3.3 *Diagnostics for Multivariate Generalized Linear Models*

Univariate GLMs diagnostics can be easily extended to multicategorical response variables (Lesaffre and Albert, 1989; O'Connell and Liu, 2011). The *generalized hat* matrix or projection matrix in multicategorical case, is given by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^{\top}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{W}^{1/2}, \tag{3.2}$$

where **W** is a block-diagonal matrix and each $(G \times G)$ block is given by $\mathbf{W}_{kk} = p_g(x_k)(\lambda_{gh} - p_g(x_k))$, where λ_{gh} is the Kronecker delta, with $(g, h = 1, \ldots, G)$. The matrix **H** $(NG \times NG)$ is a multiblock matrix, with N the number of subject and G the number of response categories and where $\mathbf{H}_{kk}$ is the $(G \times G)$ diagonal block. The $\det|\mathbf{H}_{kk}|$ or $tr(\mathbf{H}_{kk})$ of these submatrices can be used as a measure of leverage for subject k. Also in the multicategorical case, high values of $\det|\mathbf{H}_{kk}|$ indicate leverage points. The same considerations are also valid for matrix **M**, but in this case leverage points have values of $\det|\mathbf{M}_{kk}|$ close to 0.

If we have grouped data and a multicategorical response variable Y, the *Standardized Pearson residuals* are given by

$$\mathbf{r}_k = \widehat{\mathbf{W}}_k^{-1/2}\widehat{\mathbf{o}}_k$$

where $\mathbf{W}_k$ is a diagonal matrix of $\widehat{p}_g(x_k)$ and $\widehat{\mathbf{o}_k}$ is the raw residual vector of length G given by $\mathbf{y}_k - \mathbf{p}_k$. Also in this case, large value of the above statistic indicates poor fit for that subject.

By analogy with the binary case, the covariance matrix of raw residuals is $\mathbf{M} = \mathbf{I} - \mathbf{H}$, where **H** is the generalized hat matrix and **I** is the identity matrix. Thus, raw residuals are divided by this covariance obtaining *Studentized Pearson Residuals*

$$\mathbf{e}_k = \mathbf{M}_{kk}^{-1/2}\mathbf{r}_k.$$

Multiplying $\mathbf{e}_k$ by itself, that is, $\mathbf{e}_k^{\top}\mathbf{e}_k$ we obtain a score statistic but differently from results of Pregibon (1982), is not a χ^2.

The deviance in the multinomial logit model is

$$d_k = -2LL_k = -2\sum_{g=1}^{G} y_{kg} \log p_g(x_k),$$

which measures the agreement between the observed and fitted log-likelihoods of subject at x_k. This statistic can also be used to detect outliers. Large values of the individual deviance indicate that the model does not fit well for that subject. Summing over subjects the deviance of the model is obtained, that measure the difference between the log-likelihood of the fitted model, and the log-likelihood of the saturated model that fit the data perfectly (Nelder and Wedderburn, 1972). .

Finally, the case deletion methods can be extended to multicategory logit models. Let Δ_{qk} the impact on each coefficient of deleting each observation in turn

$$\mathbf{\Delta}_{qk} = \mathbf{b}_q - \mathbf{b}_{q(-k)},$$

where $\mathbf{b}_q$ is the vector of coefficients of predictor q with legnth $G-1$, and $\mathbf{b}_{q(-k)}$ is the same coefficient vector computed after that observation k is deleted. $\mathbf{\Delta}_{qk}$ is the multicategorical version of the $dfbeta$. To standardize $\mathbf{\Delta}_{qk}$ it is useful to divide it by the deleted coefficient standard errors $SE_{-k}(\mathbf{b}_k)$, obtaining the multicategorical version of the so-called $dfbetas$. If these values are large it indicates that those observations affect the coefficient estimates.

Assuming a quadratic approximation of the log-likelihood around $\widehat{\mathbf{b}}_q$ yields an approximate *generalized Cook's distance* for multicategorical case given by

$$c_k = {\mathbf{r}_k}^\top {\mathbf{M}_{kk}}^{-1}\mathbf{H}_{kk}{\mathbf{M}_{kk}}^{-1}\mathbf{r}_k$$

Removing case k also affects the interval estimates. As in the univariate case, we can compute a similar measure of c_k which is

$$\bar{c}_k = \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1}\mathbf{H}_{kk}\mathbf{r}_k,$$

that indicates the contribution of case k to the confidence region of $\widehat{\mathbf{b}}(k)$.

From $\bar{c}_k$ we can construct a tool to assess the sensitivity of the goodness of fit. We have

$$\Delta_k D = d(\widehat{\mathbf{b}}) - b_k[\widehat{\mathbf{b}}(k)] = d_k^2 + \bar{c}_k$$

which is the approximation to the change in goodness of fit deleting subject k. Clearly, if the magnitude of $\Delta_k D$ is large the corresponding case is influential for the global goodness of fit.

Finally, also in multicategorical case, we can assess neighboring effects. On the logarithmic scale we have

$$\Delta_k d_j^2 = 2 \log \left\{ \frac{\widehat{p}_{gj}}{\widehat{p}_{gj}(-k)} \right\}, \qquad k \neq j = 1, \ldots, N$$

where $\widehat{p}_{gj}$ is the estimated probability of class g and subject j including subject k and the denominator is the same estimated probability without case k. Lesaffre and Albert 1989 proposed the following one-step approximation

$$\Delta_k^* d_j^2 = 2\mathbf{r}_j^\top \mathbf{H}_{jk} \mathbf{M}_{kk}^{-1} \mathbf{r}_k + \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1} \mathbf{H}_{kj} \mathbf{H}_{jk} \mathbf{M}_{kk}^{-1} \mathbf{r}_k.$$

The interpretation of this diagnostic is the same as in univariate case. Also in this case, it is useful to sum over j to obtain a summary measure which is easier to interpret.

O'Connell and Liu in their paper about the model diagnostics for the Proportional Odds Models and Partial Proportional Odds Models proposed some graphical representations of the diagnostic measures described so far, to detect faster and more easily outliers and influencial points. The general feature is to create index plots of those diagnostics to figure out which points are far from the rest. For more details see O' Connell and Liu 2011.

DIAGNOSTICS FOR MULTINOMIAL DISTANCE MODEL 4

4.1 Model implementation

In chapter 2 we presented the Multinomial Distance Model like an extention of Ideal Point Discriminant Analysis. Given G response categories, N subjects and Q predictor variables $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_Q)$ the Multinomial Distance Model in one dimension is given by:

$$p_g(\mathbf{x}_k) = \frac{\exp(-\delta_{kg}^2)}{\sum_{h=1}^{G} \exp(-\delta_{kh}^2)},$$

for $g = 1, ..., G$, $k = 1, ..., N$ and δ_{kg}^2 is the squared Euclidean Distance between subject k and response category g. Equation (2.5) expresses the model in terms of log-odds.

To implement the model we need the response matrix $\mathbf{Y}$

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 0 \end{bmatrix}$$

that has N rows and G columns. In each row there is 1 if the subject k belongs to category g and 0 otherwise. Each subject can only belong to one of the response categories. The $\mathbf{X}$ matrix will be

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1Q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2Q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{NQ} \end{bmatrix}$$

where the first column is for the intercept. Once the deviance function is optimized (see the R code in the Appendix), we obtain the estimated coefficients. In particular we have

$$\mathbf{b} = [\alpha, \beta_1, \ldots, \beta_Q]^\top$$

and

$$\mathbf{z} = [z_1, z_2, \ldots, z_{G-1}]^\top .$$

Note that the coordinate of the last category is set equal to 0 for the identification of the model. We can avoid to compute the Euclidean distances using equation (2.6) to compute the probabilities. Thus, we have to multiply the $\beta's$ coefficients with each of the z_g and subtract from the intercepts the squared of the group coordinates. Then, we have $G-1$ *pseudo*-intercepts $\widetilde{\alpha}_g$ and $(G-1)Q$ *pseudo*-coefficients $\widetilde{\beta}_{gq}$

$$\textit{Pseudo-}\widetilde{\mathbf{S}} = \begin{bmatrix} 2\alpha(z_1) - z_1^2 & 2\alpha(z_2) - z_2^2 & \ldots & 2\alpha(z_{G-1}) - z_{G-1}^2 & 0 \\ 2\beta_1(z_1) & 2\beta_1(z_2) & \ldots & 2\beta_1(z_{G-1}) & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 2\beta_Q(z_1) & 2\beta_Q(z_2) & \ldots & 2\beta_Q(z_{G-1}) & 0 \end{bmatrix}$$

Each column of the above matrix contains the *pseudo*-coefficients for each response category. These *pseudo*-coefficients are not the coefficients of the multinomial distance model, but derived coefficients such that the multinomial distance model can be written as a baseline category logit model (see later). For example, the log-odds of category 1 compared to category G is

$$\log\left(\frac{p_1(\mathbf{x}_k)}{p_G(\mathbf{x}_k)}\right) = 2\alpha(z_1) - z_1^2 + 2\beta_1(z_1)x_{k1} + \cdots + 2\beta_Q(z_1)x_{kQ}.$$

Indeed, the probability of subject k to belong to category 1 is equal to:

$$\begin{aligned} p_1(\mathbf{x}_k) &= \frac{\exp\left[2\alpha(z_1) - z_1^2 + 2\beta_1(z_1)x_{k1} + \cdots + 2\beta_Q(z_1)x_{kQ}\right]}{1 + \sum_{h=1}^{G-1} \exp\left[2\alpha(z_h) - z_h^2 + 2\beta_1(z_h)x_{k1} + \cdots + 2\beta_Q(z_h)x_{kQ}\right]} \\ &= \frac{\exp\left[2(\mathbf{x}_k^\top \mathbf{b})(z_1) - z_1^2\right]}{1 + \sum_{h=1}^{G-1} \exp\left[2(\mathbf{x}_k^\top \mathbf{b})(z_h) - z_h^2\right]} \end{aligned}$$

where $\mathbf{x}_k^\top$ is the k-th row vector of the matrix $\mathbf{X}$ of length $Q+1$.

Starting from equation (2.6), we also noticed that the one-dimensional multinomial distance model can be written as a constrained baseline category logit model, that is

$$\begin{aligned}\log\left(\frac{p_g(\mathbf{x}_k)}{p_G(\mathbf{x}_k)}\right) &= 2\alpha(z_g) + 2\mathbf{x}_k^\top \mathbf{b}(z_g) - z_g^2 \\ &= \alpha_g^* + x_k\beta_g^*,\end{aligned} \tag{4.1}$$

where

$$\alpha_g^* = 2\alpha(z_g) - z_g^2, \qquad \beta_{gq}^* = 2\beta_q(z_g),$$

are the intercept and slope of predictor X_q for category g respectively. The model does not use the proportional assumption because each predictor variable has its own effect on each category, due to the fact that we multiply the parameters by the coordinate of each category. Here, we consider the last category G as a baseline category and we set its coordinate equal to 0. Thus, the log-likelihood of the model is

$$\begin{aligned}\log -L(\mathbf{b};\mathbf{y}) &= \sum_{k=1}^{N}\sum_{g=1}^{G} y_{kg}\log\left(p_g(\mathbf{x}_k)\right) \\ &= \sum_{k=1}^{N}\left\{\sum_{g=1}^{G-1} y_{kg}\log\left[\frac{p_g(\mathbf{x}_k)}{1-\sum_{g=1}^{G-1}p_g(\mathbf{x}_k)}\right] + \log\left[1-\sum_{g=1}^{G-1}p_g(\mathbf{x}_k)\right]\right\},\end{aligned} \tag{4.2}$$

where the $\log$ of the first term in the square brackets is the logit $(\alpha_g^* + \mathbf{x}_k^\top \mathbf{b}_g^*)$ and the second term is the probability of the last (baseline) category. We, therefore, seek estimates $\widehat{\mathbf{b}}$ such that the gradient of the function is equal to 0. A closed form of the maximum likelihood estimate, except in trivial cases, does not exist. Thus, some form of iterative procedure is required. The Quasi-Newton update formula to estimate $\mathbf{b}$ is

$$\mathbf{b}_{(t+1)} = \mathbf{b}_{(t)} - \alpha_{(t)}\mathbf{H}(\mathbf{b}_{(t)})^{-1}\nabla(\mathbf{b}_{(t)})$$

with $\mathbf{H}(\mathbf{b}_{(t)})$ an approximation of Hessian matrix computed at $\mathbf{b}_{(t)}$ and $\nabla(\mathbf{b}_{(t)})$ the gradient of the function computed at $\mathbf{b}_{(t)}$ and for some α that satisfies the Wolfe conditions (Wolfe, 1969) which ensure that the objective function is minimized at each step. The first partial derivative of the log-likelihood respect to β_q is given by

$$\frac{\partial L(\mathbf{b};\mathbf{y})}{\partial \beta_q} = \sum_{k=1}^{N} \left[\sum_{g=1}^{G-1} y_{kg}(2x_{kq}z_g) - \left(\frac{\sum_{g=1}^{G-1}(2x_{kq}z_g)\exp(2\mathbf{x}_k^\top \mathbf{b} z_g - z_g^2)}{1+\sum_{g=1}^{G-1}\exp(2\mathbf{x}_k^\top \mathbf{b} z_g - z_g^2)} \right) \right]$$

and for each z_g is

$$\frac{\partial L(\mathbf{b};\mathbf{y})}{\partial z_g} = \sum_{k=1}^{N} \left[y_{kg}(2\mathbf{x}_k^\top \mathbf{b} - 2z_g) - \frac{(2\mathbf{x}_k^\top \mathbf{b} - 2z_g)\exp(2\mathbf{x}_k^\top \mathbf{b} - z_g^2)}{1+\sum_{g=1}^{G-1}\exp(2\mathbf{x}_k^\top \mathbf{b} z_g - z_g^2)} \right].$$

Therefore, $\nabla(\mathbf{b}_{(t)})$ has $Q+G-1$ rows. We can also simplify by writing

$$\nabla(\mathbf{b}_{(t)}) = \begin{cases} \sum_{k=1}^{N}\sum_{g=1}^{G-1}(2x_{kq}z_g)(y_{kg}-\widehat{y}_{kg}) \\ \quad\vdots \\ \sum_{k=1}^{N} 2(\mathbf{x}_k^\top \mathbf{b} - z_1)(y_{k1}-\widehat{y}_{k1}) \\ \quad\vdots \\ \sum_{k=1}^{N} 2(\mathbf{x}_k^\top \mathbf{b} - z_{G-1})(y_{k(G-1)}-\widehat{y}_{k(G-1)}) \end{cases}$$

Even if this algorithm is not necessary to estimate the one dimensional multinomial distance model, it is faster than the algorithm that optimize model (2.4).

If we have a multinomial distance model in more dimensions we have to sum over the dimensions to get probabilities. Suppose we have A dimensional multinomial distance model. After fitting we get:

$$\mathbf{B} = \begin{bmatrix} \alpha_1 & \dots & \alpha_A \\ \beta_{11} & \dots & \beta_{1A} \\ \beta_{21} & \dots & \beta_{2A} \\ \vdots & \vdots & \vdots \\ \beta_{Q1} & \dots & \beta_{QA} \end{bmatrix} \qquad \mathbf{Z} = \begin{bmatrix} z_{11} & \dots & z_{1A} \\ z_{21} & \dots & z_{2A} \\ z_{31} & \dots & z_{3A} \\ \vdots & \vdots & \vdots \\ z_{G1} & \dots & z_{GA} \end{bmatrix}$$

Setting $z_{1a} = 0$ we fix translation problem and $z_{ga} = 0, \forall g \leq a$ we fix rotation issue. To obtain probabilities, formula (2.4) has to be computed. Also for more than one dimensions, we can computed the *pseudo*-$\widetilde{\mathbf{S}}$ matrix

$$\begin{bmatrix} 2\alpha_1(z_{11})+\cdots+2\alpha_A(z_{1A})-z_{11}^2-\cdots-z_{1A}^2 & \dots & 2\alpha_1(z_{G1})+\cdots+2\alpha_A(z_{GA})-z_{G1}^2+\cdots+z_{GA}^2 \\ 2\beta_{11}(z_{11})+\cdots+2\beta_{1A}(z_{1A}) & \dots & 2\beta_{11}(z_{G1})+\cdots+2\beta_{1A}(z_{GA}) \\ \vdots & \ddots & \vdots \\ 2\beta_{Q1}(z_{11})+\cdots+2\beta_{QA}(z_{1A}) & \dots & 2\beta_{Q1}(z_{G1})+\cdots+2\beta_{QA}(z_{GA}) \end{bmatrix}$$

Therefore, the probability of class G is given by

$$p_G(x_k) = \frac{\exp[2\alpha_1(z_{G1})+\cdots+2\alpha_A(z_{GA})-z_{G1}^2-\cdots-z_{GA}^2+2\beta_{11}(z_{G1})+}{\sum_{h=1}^{G}\exp[2\alpha_1(z_{h1})+\cdots+2\alpha_A(z_{hA})-z_{h1}^2+\cdots+z_{hA}^2+2\beta_{11}(z_{h1})+}$$
$$\frac{\cdots+2\beta_{1A}(z_{GA})+\ldots+2\beta_{Q1}(z_{G1})+\cdots+2\beta_{QA}(z_{GA})]}{\cdots+2\beta_{1A}(z_{hA})+\ldots+2\beta_{Q1}(z_{h1})+\cdots+2\beta_{QA}(z_{hA})]}.$$

Thus, models in more than one dimensions can be computed and for each of them it is possible to compute the above *Pseudo*-$\widetilde{\mathbf{S}}$ coefficient matrix.

Once the model is estimated, we need to evaluate the global fit as well as the outliers, if there are. Unfortunately, for Mutinomial Distance Model no diagnostic statistics are available. In fact, the Multinomial Distance Model is not a generalized linear model and therefore we can not use the well known GLM theory. The main problem is that Multinomial Distance Model is multiplicative in the parameters. In fact, once we estimated the parameters, to compute probabilities or log-odds, we have to multiply the β coefficients by the estimated group coordinates z to get *pseudo*-coefficients and *pseudo*-intercepts for each category. The number of parameters to estimate is different from the number of parameters that we use to compute probabilities in the final step. Therefore, the design matrix is difficult to determine. In the estimating process, we use the simplest design matrix $\mathbf{X}$, where each row vector is the observed predictor vector plus a 1 for the intercept. But as we could see before, finally we have $(G-1)$ *pseudo*-intercepts $\widetilde{\alpha}_g$ and $(G-1)\times Q$ *pseudo*-coefficients $\widetilde{\beta}_{gq}$. Then, the design matrix should contain $(G-1)+(G-1)\times(Q)$ columns.

We showed that the Multinomial Distance Model in one-dimensional Euclidean space can be regarded as a baseline category logit model. Furthermore, here we showed that we can compute the *pseudo*-$\widetilde{\mathbf{S}}$ coefficient matrix which is equal to the coefficient matrix of a baseline category logit model. Using this assumption, here we propose to extend the diagnostics of multiple-group logistic regression (Lesaffre and Albert, 1989) to the one-dimension multinomial distance model. The next section shows this extention in details.

4.2 Extending Multiple-group Diagnostics to Multinomial Distance Model

So far, we explain how to fit the model and in the Appendix we provide R code to estimate it. As for a baseline category logit model, we have different intercepts and slopes for different classes. The total number of *pseudo*-coefficients is $(Q+1)\times(G-1)$. The *pseudo*-$\widetilde{\mathbf{S}}$ coefficient matrices computed above for one and two dimensional multinomial distance models have the same dimensionality as that of the baseline category logit model (one intercept for each category and one slope for each category and predictor). We use this analogy to adapt multiple group diagnostics to the multinomial distance model.

To construct diagnostics we need some goodness of fit measures and the classical building blocks which are the quantities:

1. $\widehat{\mathbf{b}}$, the estimated coefficient vector of length $Q+1$;
2. $\widehat{\mathbf{b}}(k)$, the estimated coefficient vector deleting the k-th observation;
3. $\mathbf{r}_k$, that is the standardized Pearson residual vector for subject k of length G;
4. $\mathbf{H}_{kj}$, which are the ($G\times G$) extra diagonal blocks of the generalized hat matrix;
5. $\mathbf{M}_{kk}$, which are the diagonal blocks of the **M=I-H** matrix.

To compute all these quantities, first of all we have to redefine the design matrix **X**. Because the Multinomial Distance Model can be regarded as multinomial logit model, suppose that we estimated $(G-1)\times(Q+1)$ coefficients. The *Pseudo* design matrix $\widetilde{\mathbf{X}}$ will be $N(G-1)\times(Q+1)(G-1)$ matrix formed by N stacked blocks where each block is

$$\widetilde{\mathbf{X}}_k = \begin{bmatrix} \mathbf{x}_k & 0 & 0 & \dots & 0 \\ 0 & \mathbf{x}_k & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{x}_k \end{bmatrix}$$

and

$$\mathbf{x}_k = [1, x_{k1}, x_{k2}, \dots, x_{kq}]$$

Moreover, let $\boldsymbol{\Sigma}$ be the covariance matrix of $\mathbf{Y}$. It is a $NG \times NG$ diagonal block matrix and each block is given by

$$\boldsymbol{\Sigma}_k = p_g(\mathbf{x}_k)[\lambda_{gh} - p_h(\mathbf{x}_k)], \qquad \textit{with } g, h = 1,\!.\,.\,.\,, G,$$

where λ is the Kronecker delta which is equal to 1 if g and h are equal. Let $\overline{\boldsymbol{\Sigma}}$ be any generalized inverse of $\boldsymbol{\Sigma}$. In the case of ungrouped data, $\overline{\boldsymbol{\Sigma}}$ is

$$\overline{\boldsymbol{\Sigma}} = \mathrm{diag}\{1/p_1(\mathbf{x}_1), \dots, 1/p_G(\mathbf{x}_1), \dots, 1/p_1(\mathbf{x}_N), \dots, 1/p_G(x_N)\}.$$

The generalized hat matrix becomes

$$\mathbf{H} = \overline{\boldsymbol{\Sigma}}^{1/2}\mathbf{Q}\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^\top\mathbf{V}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^\top\mathbf{Q}^\top\overline{\boldsymbol{\Sigma}}^{1/2},$$

where $\mathbf{Q}$ is a $NG \times N(G-1)$ block diagonal matrix, and each block is given by

$$\mathbf{Q}_k = p_g(\mathbf{x}_k)[\lambda_{gh} - p_h(\mathbf{x}_k)],$$

and $\mathbf{V}$ is a $N(G-1) \times N(G-1)$ block diagonal matrix, where each block $\mathbf{V}_k$ is equal to

$$\mathbf{V}_k = p_g(\mathbf{x}_k)[\lambda_{gh} - p_h(\mathbf{x}_k)].$$

Once $\mathbf{H}$ is computed, it is easy to compute $\mathbf{M}$. The determinant of the diagnonal blocks of $\mathbf{M}$ can be used as a diagnostic to assess the leverage of cases. If a case is a regression outlier that means that it is far from the center of the space spanned by $\widetilde{\mathbf{X}}$ combined with the fact that it has an anomalous response value. Points have large leverage, that is, they might influence regression estimates, if their response values are atypical from the others. Lesaffre

(1989) showed that the $\sum_k |\mathbf{M}_{kk}|$ is always close to $N - v$, where v is the number of parameter to estimate. Therefore, a practical rule to detect leverage points is $|\mathbf{M}_{kk}| \leq 2v/N$. We could not use this rule in the multinomial distance model because the number of parameters to estimate is different form the number of *pseudo* coefficients that we use to compute probabilities, so we will use graphical approaches to pinpoint leverage points. Even if such a case is a leverage point, it does not mean that it influences the regression estimate or prediction process. Therefore, a statistician has to analyze points with high leverage more thoroughly before to discard them. One could look at residuals, but in the multinomial distance model like in logistic regression they can be defined on several scales. The most useful residuals are the deviance residuals and the Pearson residuals. The standardized version of Pearson residuals $\mathbf{r}_k$ is given by

$$\mathbf{r}_k = \overline{\boldsymbol{\Sigma}}_k^{1/2}\mathbf{o}_k$$

where $\mathbf{o}_k$ is the residual vector given by $\mathbf{y}_k - \mathbf{p}_k$. Here $\mathbf{y}_k$ is the response row vector while $\mathbf{p}_k$ is the row vector of the estimated probabilities. It is useful to produce an index plot of $\mathbf{r}_k^\top \mathbf{r}_k$ to figure out which points have the largest residuals, that is, which points are poorly fitted by the model.

Studentized Pearson residuals are given by

$$\mathbf{e}_k = \mathbf{M}_{kk}^{-1/2}\mathbf{r}_k$$

where $\mathbf{M}_{kk}$ is the diagonal block for subject k of the $\mathbf{M} = \mathbf{I} - \mathbf{H}$ matrix. The diagnostic $\mathbf{e}_k^\top \mathbf{e}_k$ is a useful tool to assess the sensibility of the goodness of fit.

Standardized Pearson residuals can be very useful if they are combined with other quantities, like diagonal blocks of the $\mathbf{M}$ matrix (or $\mathbf{H}$ matrix), to combine leverage with poor fit in order to detect influential cases.

Influential cases can be appraised by case deletion method. If the estimated coefficients after deleting case k are substantially different from the estimates obtained considering all cases, that point is influential. The one-step approximation to obtain estimates deleting case k is given by:

$$\widehat{\mathbf{b}}(k) = \widehat{\mathbf{b}} - (\widetilde{\mathbf{X}}^\top \widehat{\mathbf{V}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}_k \widehat{\mathbf{V}}_k^{1/2}\mathbf{M}_{kk}^{-1}\widehat{\mathbf{V}}_k^{1/2}\widehat{\mathbf{r}}_k.$$

This diagnostic cannot be used in multinomial distance model due to the fact that the number of estimated coefficients is different from the number of $pseudo$-coefficients. In fact, the number of rows of the column vector $\widehat{\mathbf{b}}$ is different from the number of columns of the pseudo-design matrix $\widetilde{\mathbf{X}}$.

As an overall measure of the influence of subject k we can compute an approximation to the generalized Cook's distance:

$$c_k = \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1} \mathbf{H}_{kk} \mathbf{M}_{kk}^{-1} \mathbf{r}_k. \tag{4.3}$$

Cook's distance describes the boundary of an asymptotic confidence region for the parameter $\widehat{\beta}$. The diagnostic c_k is its one step approximation, which indicates how this region change when deleting case k. To evaluate Cook's values, there is no well defined threshold. Some authors advise to detect points with Cook's distance larger than 1. Others proposed as threshold $4/N$ or $C_k > 4/(N-q-1)$. Fox (1991) is rather cautious about defining thresholds. In fact, he advised to use graphical approaches to see which points are far from the others and to do further analysis on those points. Points detected by c_k are influential points, that is, they influence the regression estimates. However, if a point influences the estimation process that does not mean that it also influences the prediction process. Further analyses need to investigate the influential cases to see which cases influence predictions.

A similar measure to Cook's distance can be computed as

$$\overline{c}_k = \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1} \mathbf{H}_{kk} \mathbf{r}_k,$$

which express the same diagnostic as c_k, but in this case it indicates how the confidence interval changes including the k-th case. The main difference is that the one-step estimate of the latter is more accurate than the former (Pregibon, 1981). Starting from $\overline{c}_k$ another useful statistic to assess influential cases is

$$\Delta_k D = d_k^2 + \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1} \mathbf{H}_{kk} \mathbf{r}_k = d_k^2 + \overline{c}_k,$$

which indicates the change in goodness of fit by deleting case k and with d_k^2 the individual deviance. According to Williams (1987), this diagnostic should have better distributional properties than d_k^2 and $\mathbf{e}_k^\top \mathbf{e}_k$, but sustantially they measure the same thing.

As we pointed out before, if a point is influential it does not mean that it also affects the prediction. To evaluate if the case distorts the prediction rule it is possible to compute a diagnostic to assess the effect of the k-th observation on the fit of the remaining $N-1$ observations. Therefore, a diagnostic that measures the distance between the estimated probability including the k-th case and the estimated probability deleting the same case is needed. On the logarithmic scale, this difference is expressed by

$$\mathbf{d}_1 \left\{ \mathbf{p}_j, \mathbf{p}_j(-k) \right\} = \Delta_k d_j^2 = 2 * \log \left\{ p_{g(j)} / p_{g(j)}(-k) \right\}$$

and its one-step approximation is given by

$$\Delta_k^* d_j^2 = 2\mathbf{r}_j^\top \mathbf{H}_{jk} \mathbf{M}_{kk}^{-1} \mathbf{r}_k + \mathbf{r}_k^\top \mathbf{M}_{kk}^{-1} \mathbf{H}_{kj} \mathbf{H}_{jk} \mathbf{M}_{kk}^{-1} \mathbf{r}_k. \tag{4.4}$$

Properties of this diagnostic are:

1. $\Delta_k^* d_j^2 \neq \Delta_j^* d_k^2$

2. $\Delta_k^* d_j^2 > 0$ means that the fit becomes worse;

3. $\Delta_k^* d_j^2 < 0$ means that the fit becomes better;

4. $\Delta_k^* d_j^2 = 0$ means that the fit remains the same.

This diagnostic is, however, not very useful because for each point you have $(N-1)$ different $\Delta_k^* d_j^2$. It is more useful to look at the sum over j of $\Delta_k^* d_j^2$ and when it is negative then case k influences the prediction process.

All diagnostics described so far can be computed without any computational efforts, except the last one $\Delta_k^* d_j^2$. Many authors advice to produce plots of the above diagnostics versus case indices (O'Connell and Liu, 2011) to see direclty which cases are dangerous for the estimating and prediction processes.

4.3 Computational Intensive Diagnostics

Influential cases can also be evaluated by case deletion, that is, computing coefficients deleting each case k in turn and assessing change in goodness of fit measures like the deviance and in estimated values. If after deleting case k the estimates are far from the estimates computed on complete data and the deviance is larger than before, it indicates that such a case influences the estimating process. For each coefficient the difference between estimated coefficent on complete data and after deleting case k can be computed. This diagnostic, called *dfbeta*, provides information concerning the effect of the kth case on the fit and can be interpreted as an influence function. It is:

$$dfbeta = \widehat{\mathbf{b}} - \widehat{\mathbf{b}}_{(-k)}.$$

It is also possible to compute the standardized version of the $dfbeta$, dividing by its standard error. There is, however, no threshold to define how different the estimate might be to define a point as influential. Pregibon (1981) and others suggested to plot this diagnostic versus the case index and see what points have the largest difference.

To assess the influence on the goodness of fit, we have to look at the change in model deviance, deleting case k. Points which produce the largest change are influential cases on the global fit of the model. However, the change in deviance could be caused by two facts:

1. the model does not fit well case k and the change is only in the single component d_k^2;
2. the point is in an extreme region of the space spanned by $\widetilde{\mathbf{X}}$ and then the change in deviance is the sum of the change in all other components (the point also affects the coefficient estimates).

However, this difference is not stressed only by the deviance, but to draw conclusions about influential cases all diagnostics should be considered. To evaluate this diagnostic, we should fit the model N times, deleting every time a subject, and finally plot the model deviance versus case index. The case with the largest change in deviance influences the global fit. Luckily, for this diagnostic Pregibon (1980) first and later Leaffre (1989), proposed its one step approximation.

Finally, to evaluate neighboring effect the influence of case k on the estimated probabilities cab ne expressed by the distance between $\widehat{\mathbf{p}}$ and $\widehat{\mathbf{p}}_{(-k)}$. That is

$$\mathbf{a}_j(k) = \widehat{\mathbf{p}}_j - \widehat{\mathbf{p}}_j(-k), \qquad \textit{with } k \neq j = 1, \ldots, N,$$

where $\widehat{\mathbf{p}}_j$ is the vector of probabilities of subject j computed considering complete data and $\widehat{\mathbf{p}}_j(-k)$ is the vector of probabilities of the same subject computed after deleting case k. Therefore, for each case k, probabilities deleting that subject and differences are computed. For each case we have $(N-1)\times(G)$ numbers. This diagnostic, even though it is a good tool to evaluate the influence on the prediction process, it is very computer intensive and it has low informative power due to the fact that for each subject we have to analyze so many numbers. Therefore, it is is more useful to look at the sum over j of these diagnostics. Again, Pregibon (1981) and Lesaffre and Albert (1989) proposed its one step approximation (what we called before $\Delta_k^* d_j^2$). Table 4.1 shows a summary of all diagnostics that can be applicable to multinomial distance model.

Diagnostic	1-step approximation	Computer Intensive
Coefficient sensitivity (dfbeta)		✓
Cook's Distance	✓	
Change in deviance	✓	✓
Goodness of fit sensitivity	✓	✓
Neighboring effects	✓	✓

Table 4.1: Diagnostic Measures: Applying to Multinomial Distance Model.

In this chapter we explained how the multinomial distance model is fitted and how it is possible to evaluate the fit. In these last two sections, we claimed that multiple group diagnostics (Lesaffre and Albert, 1989) can be extended to multinomial distance model, even though this model is not a generalized linear model. Only for *dfbeta* it turned out that it was not possible to extend the one step approximation. In the next chapter we apply those diagnostics to several datasets to show that multiple group diagnostics also work fine for multinomial distance model.

APPLICATIONS 5

This chapter concerns a simulation study and three applications on real datasets. For each section a short description about the data is supplied and then results for the multinomial distance model and the baseline category logit model are shown. The first section concerns a simulation study. A one dimensional multinomial model is applied and it is shown how multiple-group diagnostics work for that model. In the second section both an one dimensional multinomial model and a baseline category logit model were fitted on the NESDA data (Penninx et all., 2008), where the response variable is ordinal. The third section concerns with Hepatitis data (Lessaffre and Albert, 1989) which has a non-ordinal response variable and a two dimensional distance model is fitted. Last section shows a comparison between the one dimensional multinomial distance model and the proportional odds model on data analyzed by O'Connell and Liu (2011).

5.1 *Simulation Study*

Data are simulated for an ordinal categorical response variable and a single continuous predictor. The response variable has four categories. Data were simulated using the R language. We sampled a predictor x from a normal distribution with $\mu = 2$ and $\sigma^2 = 2$. Setting $\alpha = -0.6$, $\beta = 0.7$, $z_1 = 2, z_2 = 1.5, z_3 = 1$ and $z_4 = 0$ we generated Euclidean distances and then computed probabilities. We used these probabilities to get observed values by drawing from a multinomial function. Table 5.1 shows some summary statistics for the predictor variable.

In a second step four outliers were added to the data. Table 5.2 reports outlier values and Figure 5.1 shows a graphical representation of the data.

	response categories				
	1	2	3	4	total
mean	3.575	2.494	2.115	0.7721	1.989
sd	0.937	0.995	0.926	0.925	1.363
max	5.560	4.688	4.513	3.470	5.560
min	1.438	-0.441	-0.090	-1.502	-1.502

Table 5.1: Simulated data: summary measures of predictor variable.

We expect that diagnostics pinpoint these cases as outliers and influential cases.

number case	y	x
301	1	-2
302	2	7
303	3	8
304	4	8

Table 5.2: Simulated Data: outliers.

A one dimensional multinomial distance model was fitted. The deviance is 655.28 and deviance divided by the sample size is 2.15. The misclassification rate is 0.47 and Table 5.3 shows observed versus fitted values.

		Observed			
		1	2	3	4
	1	22	6	4	1
	2	12	9	13	1
Fitted	3	17	33	49	18
	4	2	7	29	81

Table 5.3: Simulated Data: observed versus fitted values.

Next we applied the diagnostics of chapter 4 to the model. In Figure 5.3 panel (a) leverage values versus individual deviance are plotted. As we can see, the diagnostic detects points $301, 302, 303$ and 304 as outliers. Points 302 and 303 have large leverage but small deviance values. Point 304 has both large individual deviance and leverage. Point 301 only has a large deviance.

Panel (b) of the Figure 5.3 is the index plot of the approximation to Cook's distance. All four outliers are detected.

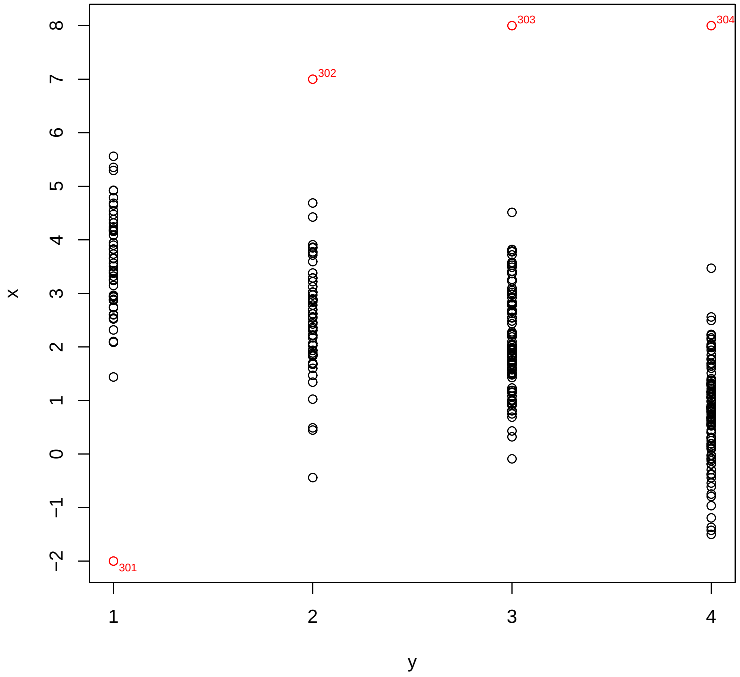


Figure 5.1: Simulated Data: outliers.

Finally, panel (c) of the Figure 5.3 presents the index plot of Studentized residuals which measures the change in goodness of fit. Points 301 and 304 have the largest residual values.

We also assessed the neighboring effect, computing the $\sum_{k \neq j} \Delta_k d_j^2$ diagnostic. Table 5.4 shows the results. When this diagnostic is negative it means that the case influences the prediction process. Therefore, if you drop the case from the analysis the misclassification rate will be smaller and the fit will be better. For the simulated data, all four outliers influence the prediction. To assess wheter the above diagnostics work well, we also applied leave-one-out computer intensive diagnostics. The model was fitted N times and at each time we discarded a case. Figure 5.2 shows the model deviance for each computed model. As can be seen, discarding cases $301, 303$ and 304 the deviance decreases, which indicates that the fit gets better. Deleting cases 106 and 204 the fit seems to improve. If we look at figure 5.1 case 106 has the

Subject number	$\sum_{k \neq j} \Delta_k d_j^2$
301	-0.9769375
302	-2.5286733
303	-1.1098528
304	-1.6916758

Table 5.4: $\Delta_k d_j^2$ diagnostic.

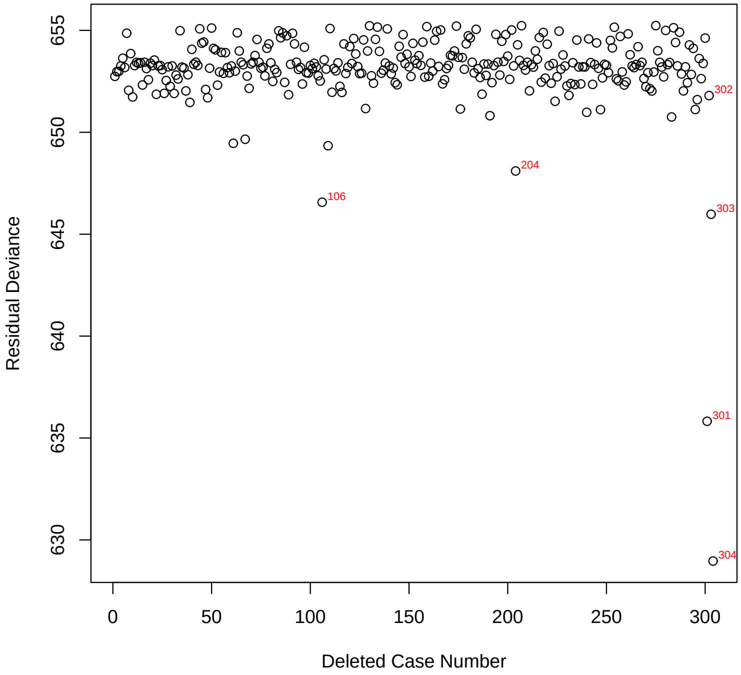


Figure 5.2: Simulated Data: LOO diagnostics. Model Deviance discarding case k.

lowest value of category 2 while case 204 has the largest value of category 4.

We also analyzed coefficient sensitivity. We compute $dfbeta$ for all coefficients. Figure 5.4 shows the plots for each coefficient. Panel (a) is the empirical influence function for the intercept. If we discard points $301, 302$ and 303 the change in the estimated intecept is large which means that those points influence its estimate. Panel (b) is the empirical influence function for β and all four outliers greatly influence its estimate. Panel (c) is the empirical influence function for z_1. Here, cases 301 and 304 influence the estimate of β.

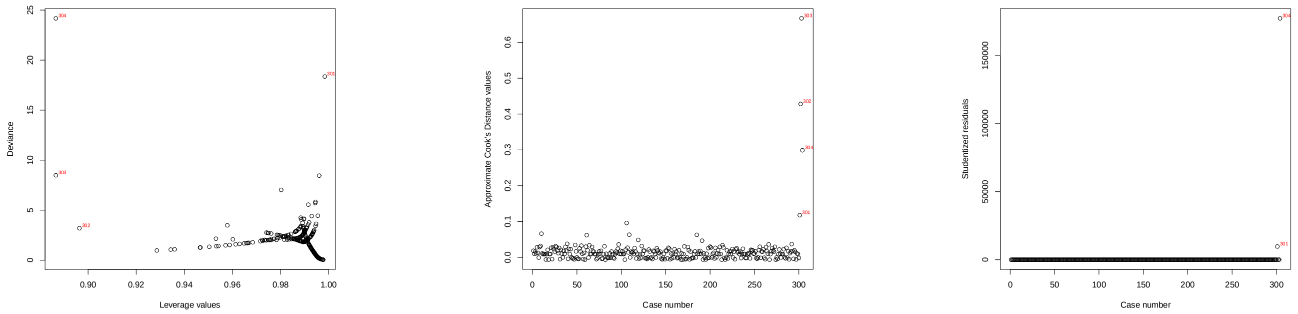
Panel (d), indeed, is the empirical influence function for z_2 and cases 302 and 303 are influential cases. Finally, panel (e) is the empirical influence function for z_3 and again cases $302, 303$ and 304 are influential.

To confirm that cases $301, 302, 303$ and 304 are outliers, we also computed $a_j(k)$ and sum over j to see if those cases also influence prediction rule. Table 5.5 shows the results which are the same as in Table 5.4, based on the one-step approximation. Therefore, all four outliers influence the prediction process, as we expected.

Subject number	$\sum_{k \neq j} a_j(k)$
301	-4.6853531
302	-0.4157954
303	-1.8862226
304	-6.1184455

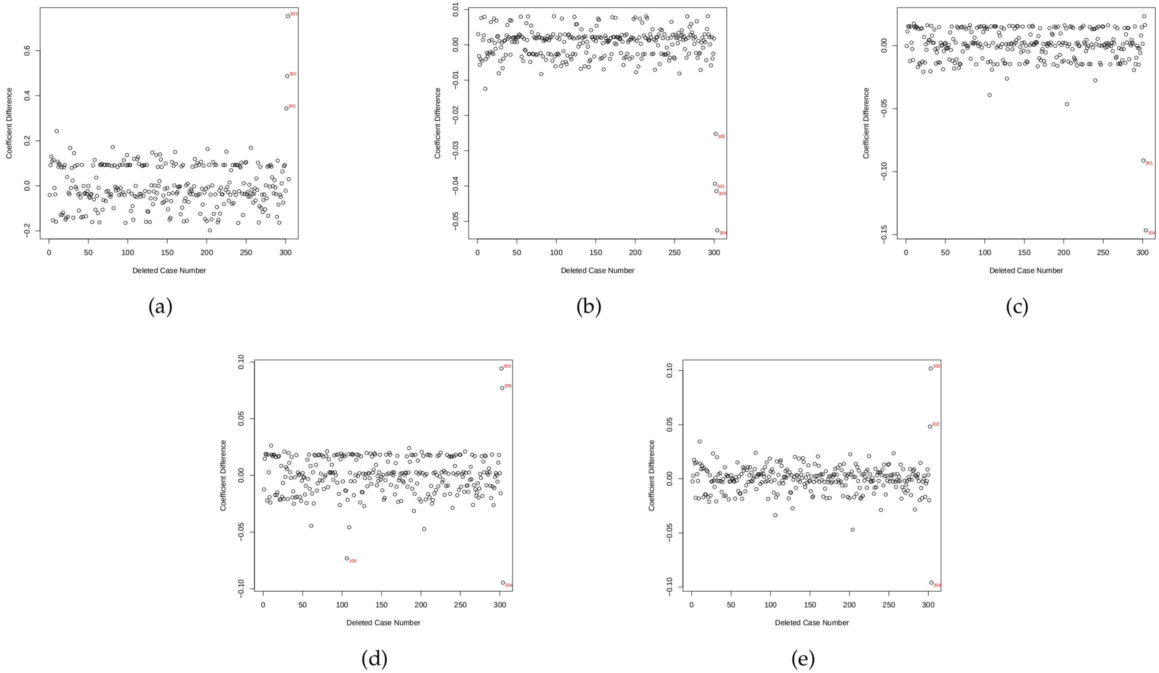
Table 5.5: $a_j(k)$ diagnostic.

We can claim that cases $301, 302, 303$ and 304 are outliers and influence both model fitting and the prediction rule. Therefore, the above simulated analysis shows that even when the likelihood space of the multinomial distance model has a different dimensions from the likelihood space of the baseline category logit model, multiple group diagnostics can be applied and those point out the correct influential cases.



(a) (b) (c)

Figure 5.3: Simulated Data:Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of Studentized residuals.



(a) (b) (c)

(d) (e)

Figure 5.4: Simulated Data: Empirical Influential Functions: Panel (a) for α. Panel (b) for β. Panel (c) for m_1. Panel (d) for m_2 and panel (e) for m_3.

5.2 NESDA data

To illustrate that the multiple-group diagnostics also work well for Multinomial Distance Model, we use the data set from NESDA study (Penninx et all., 2008). Data are composed by a six level response variable which indicates the number of mental disorders a participant has, ranging from 0 (no disorder) to 5 (five number of disorders) and two predictor variables, $gender$ and $extraversion$. The first is categorical with two categories ($0 = female, 1 = male$) while the latter is numerical variable ranging from 15 to 55 which expresses a personality aspect of a subject. For practical reasons, we use a random subsample of 408 from the total available one (2938 observations). From the 408 observations of our subsample 160 have no disorder, 116 have one disorder, 68 have two disorders, 38 have three disorders, 20 have four disorders and 6 subjects have five disorders. Figure 5.5 shows the distribution of the data. As we can see, there are some subjects with very low values for extraversion variable, given the categories.

We fit both Multinomial Distance model and Baseline Category Logit model, compute for both the diagnostic measures and then compare results.

5.2.1 Multinomial Distance Model Analysis

For the Multinomial Distance model, setting the coordinate of the last category equal to 0, we have to estimate 8 parameters: 5 category coordinates, 1 intercept and 2 predictor coefficients. The analysis is run with R using the $BFGS$ Quasi-Newton method (see R code in the Appendix). Estimates of the coordinates are $\widehat{z}_0 = 2.1979, \widehat{z}_1 = 1.5731, \widehat{z}_2 = 1.1365, \widehat{z}_3 = 0.8314, \widehat{z}_4 = 0.4996$. Estimates of the intercept is $\widehat{\alpha} = -0.1324$ and the two regression weights are $\widehat{\beta}_1 = -0.2308$ and $\widehat{\beta}_2 = 0.0639$. To applay the results of chapter 4, first we have to compute the matrix of *pseudo*-coefficients, that is:

$$\textit{Pseudo-}\widetilde{\mathbf{S}} = \begin{bmatrix} 2(-0.1324)(2.1979) - (2.1979)^2 & \dots & 2(-0.1324)(0.4996) - (0.4996)^2 & 0 \\ 2(-0.2308)(2.1979) & \dots & 2(-0.2308)(0.4996) & 0 \\ 2(0.0639)(2.1979) & \dots & 2(0.0639)(0.4996) & 0 \end{bmatrix}$$

where the first row contains the intercepts for each response category, the second row indicates the effect of gender on the log-odds of each category

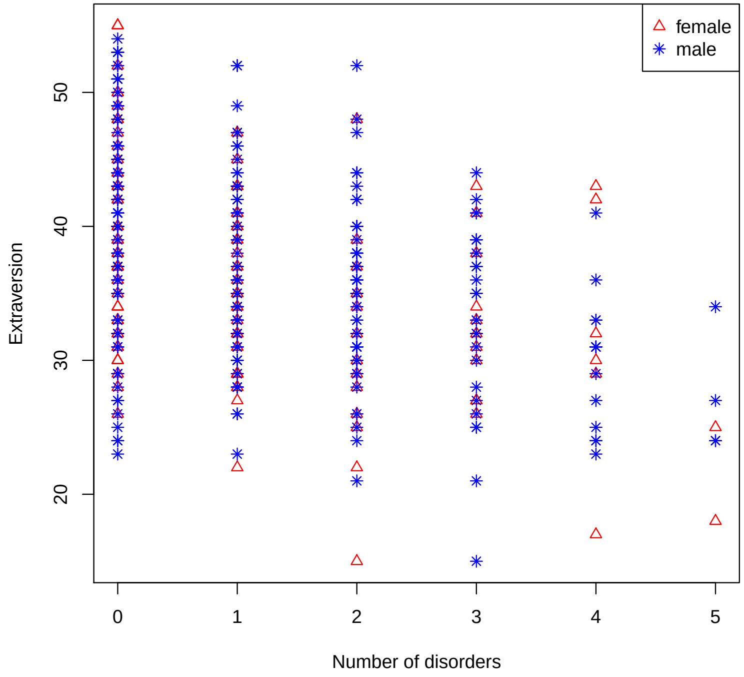


Figure 5.5: Number of disorders versus Extraversion, given Gender.

compared with the last category and the last row expresses the effect of extraversion on the log-odds of each category compared to the last one, for a one unit increase in $extraversion$. For example, the log-odds of no disorder compared to having five disorders decreases with $2 \times (-0.2308) \times (2.1979) = -1.0148$ for a male subject compared to a female subject while it increases with $2 \times (0.0639) \times (2.1979) = 0.2812$ for a one unit increase in extraversion. The deviance is 1097.841 and the deviance divided by the sample size is 2.69. Table 5.6 shows classification of observed and fitted number of disorders. The misclassification rate is 0.56.

Let us go deeper in the analysis evaluating the diagnostic measures. After the generalized hat matrix is computed (see R code in Appendix), we obtain the determinants of the $\mathbf{M}_{kk}$ matrices to detect outliers. Table 5.7 shows leverage points for our data.

Figure 5.6 (a) shows the individual deviance versus leverage value for

		Fitted					
		0	1	2	3	4	5
Observed	0	128	27	5	0	0	0
	1	68	44	4	0	0	0
	2	33	28	5	2	0	0
	3	18	15	3	1	1	0
	4	5	10	4	1	0	0
	5	0	3	3	0	0	0

Table 5.6: Number of disorders for Multinomial Distance Model: observed versus fitted values.

Subject number	Gender	Extraversion	Numb.of disorders
161	0	17	4
245	0	15	2
257	1	15	3
279	0	18	5

Table 5.7: Subjects with high leverage

each subject. The points on the left hand side (subjects $161, 245, 257$, and 279) have high leverage because they have values of $extraversion$ far from the mean. Points in the right corner of the plot are also outliers but they have low leverage because their predictor values are not far from the mean.

Figure 5.6 panel (b) shows the quadratic approximation to Cook's distance for each subject. Subjects $161, 245, 257, 207$ and 277 have large values for this diagnostic.

Figure 5.6 panel (c) is the index plot of studentized residuals. Points that produce the largest residuals are $128, 137, 347$ and 265. They are subjects with the highest individual deviance, too.

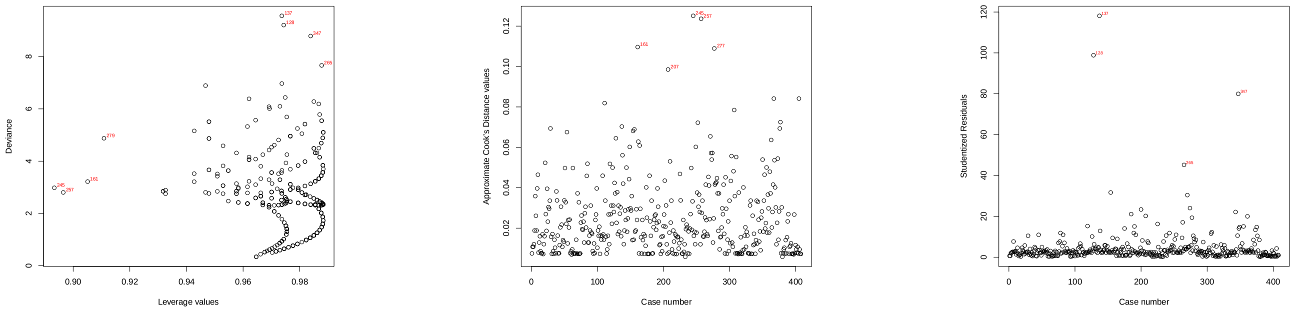
Before to get to any decision about detected outliers, we have to examine these points further. In fact, a point that influences the coefficient estimates does not also necessarily affect the prediction. Therefore, it is important to know whether the influential cases really affect the prediction. To see this, we compute what Lesaffre and Albert call $\sum_{k \neq j} \Delta_k d_j^2$, using the approximation that they proposed (Lesaffre and Alber, 1989). Table 5.8 shows this diagnostic for outliers.

Points $128, 137, 245, 257, 265, 277$ and 347 all have a negative values on

Subject number	$\sum_{k \neq j} \Delta_k d_j^2$
128	-0.0578
137	-0.0662
161	0.0365
207	0.1260
245	-0.4200
257	-0.6985
265	-0.0417
277	-0.5907
279	0.0578
347	-0.0004

Table 5.8: $\Delta_k d_j^2$ diagnostic.

this diagnostic, indicating that these points affect the classification boundaries. All other points affect the estimates of the coefficients, but they do not affect the prediction process. In fact, if we drop points from the analysis, the misclassification rate remains the same. Looking at the deviance of the model with points $128, 137, 245, 257, 265, 277$ and 347 nd the model without them, for the former the deviance divided by the number of subjects is 2.69 while for the latter it is 2.68. This explains well that a statistician should be cautious about outliers and always check the change in results if points are dropped and, last but not least, compare the benefits with the costs. For the NESDA subsample data, we can conclude that there are some outliers, but they do not largely affect the model.



(a) (b) (c)

Figure 5.6: NESDA Data. Multinomial Distance Model:Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of studentized residuals.

5.2.2 *Baseline Category Logit Model Analysis*

To verify if the diagnostics extended to multinomial distance model work well, a baseline category logit model was fitted on the same data and the same diagnostics were applied. Also in this case we use R and the $BFGS$ Quasi-newton method. Table 5.9 contains the estimated the estimated coefficients.

Coefficient	0	1	2	3	4	5
α	-7.6730608	-4.9037020	-4.2339597	-3.7442133	-2.4267208	0
β_{gender}	-0.6617311	-0.4900898	0.3231788	-0.1866782	-0.1316032	0
$\beta_{extraversion}$	0.3509189	0.2670167	0.2156655	0.1953989	0.1327075	0

Table 5.9: Baseline category logit model: estimated coefficients.

Table 5.10 shows the classification table of the observed and fitted values. As we can see, the classification is very closed to the one given by the Multinomial Distance Model, with the difference that here there are some points classified in the last category. The misclassification rate is 0.57, 0.01 greater than the misclassification rate of the multinomial distance model.

		Fitted					
		0	1	2	3	4	5
	0	123	29	8	0	0	0
	1	64	45	7	0	0	0
	2	32	29	6	0	0	1
Observed	3	16	15	6	0	0	1
	4	4	10	5	0	0	1
	5	0	2	3	0	0	1

Table 5.10: Number of disorders for Baseline Category Logit Model: observed versus fitted values.

Diagnostics are applied for this model. Table 5.11) shows that the same subjects are detected as outlier as in the multinomial distance model.

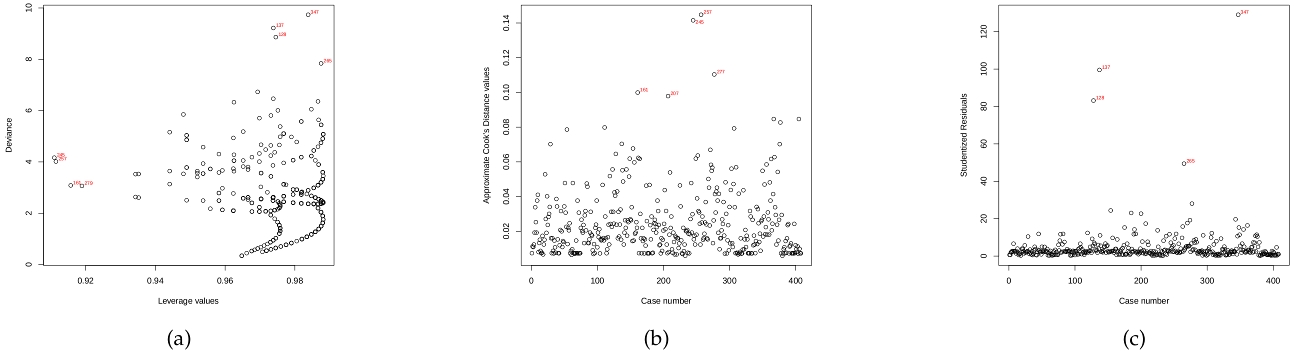
Figure 5.7 panel (a) shows the indivisual deviance values versus the leverage values. Here, again the same points are detected. Figure 5.7 panel (b) is the index plot of approximate Cook's distance values. Again, we detect the

Subject number	Gender	Extraversion	Numb.of disorders
161	0	17	4
245	0	15	2
257	1	15	3
279	0	18	5

Table 5.11: Baseline Category Logit Model: Subjects with high leverage.

same outliers as in multinomial distance model. Finally, figure 5.7 panel (c) is the index plot of studentized residuals for the baseline category logit model.

Since the diagnostics detect the same outliers and influecial cases, we can draw the same conclusions for the NESDA subsample data, which are that there are some outliers, but they do not heavily affect the results. In both cases, the misclassification rate is high, and further analysis to obtain the reasons are needed.



(a)

(b)

(c)

Figure 5.7: NESDA Data. Baseline Category Logit Model:Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of studentized residuals.

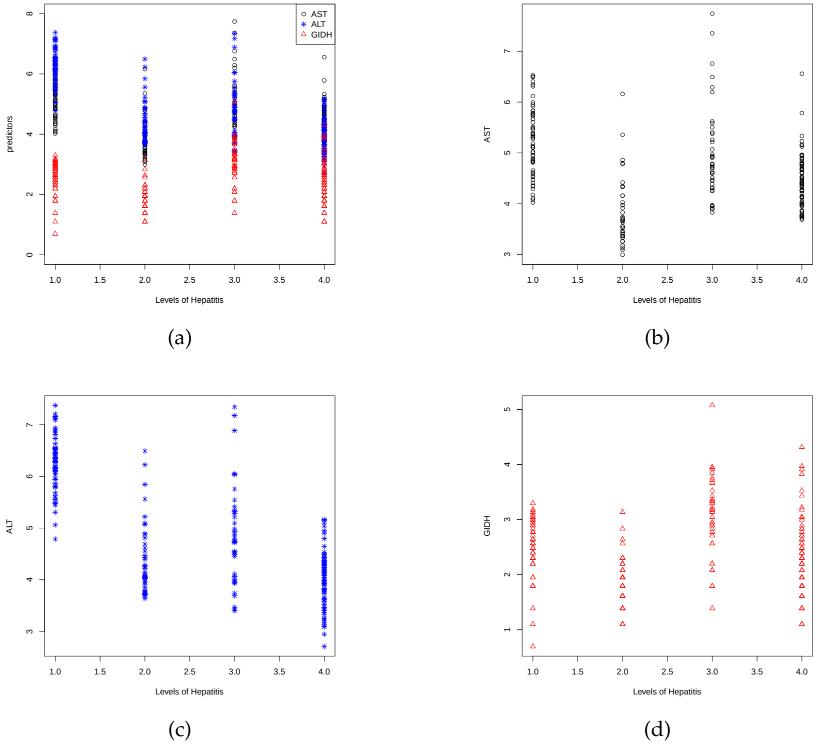
5.3 Hepatitis data

In an article of 1980 Plomteux (1980) showed that four level of hepatitis could be defined based on three liver function tests. This data set, then, is composed by a four cateogry response variable (1=acute viral hepatits, 2=persistent chronic hepatitis, 3=aggressive chronic hepatitis and 4=post-necrotic cirrhosis) and three predictor variables (aspartate aminotransferase (AST), alanine aminotransferase (ALT) and glutamate dehygrogenase (GIDH)). The total sample size is 218, from which 57 are in the response category 1, 44 are in category 2, 40 are in category 3 and 77 in category 4. For more details about the data and the experiment see Plomteux (1980) and Albert and Harris (1987). Because the predictor variables are largely skewed, we use their logarithm in the analysis. Figure 5.8 shows the distributions of each $\log$-predictor versus the response categories.

5.3.1 Multinomial Distance Model Analysis

In this case the response variable is not ordinal and therefore we fitted the Multinomial Distance Model in one, two and three dimensions. The main results are reported in Table 5.12. In one dimension the model does not work well since the information is more spread among the dimensions. The deviance is 291.09 and the misclassification rate is 0.28. In two dimensions, the deviance is 203 and the misclassification rate is 0.18. In three dimensions, which corresponds to baseline category logit model, the deviance is 192.63 and the misclassification rate is 0.16. The best model depends on a large number of factors. In fact, if you look at the Table 5.12, we see that Akaike Information Criterion and Bayesian Information Criterion indicate that the two dimensional model is the best. However, if you look at misclassification rate, the difference between the three models is not very large.

We chose the two dimensional multinomial distance model. The Multinomial Distance Model in more than one dimension is affected by translation and rotation problems. To solve these identification issues we fix some class point coordinates. For the translation problem we set $z_{1a} = 0$, for $a = 1, 2 ... A$,



(a)

(b)

(c)

(d)

Figure 5.8: Hepatitis data: Predictor Distributions versus response variable. Panel (a) plots all logarithmic predictors verus response categories. Panel (b) plots the distribution of $\log(AST)$ predictor. Plot (c) is the distribution of $\log(ALT)$ variable and panel (d) is the plot of $\log(GIDH)$ predictor versus response variable.

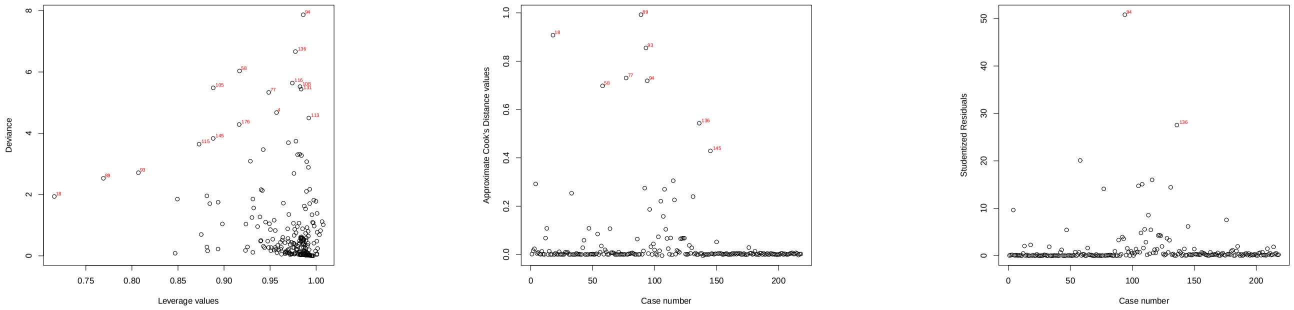
	1-dimension	2-dimensions	3-dimensions
Deviance	291.09	203.00	192.63
Misclassification rate	0.28	0.18	0.16
AIC	305.09	229.00	228.63
BIC	328.78	272.99	289.55
Number of parameters	7	13	18

Table 5.12: Main results of Multinomial Distance Model in different dimensions.

and for rotation problem we set $z_{12} = 0$. The total number of parameters to estimate are 13: $(3 + 1) \times 2$ coefficients, and $(4 - 1) + (4 - 2) = 5$ class point coordiantes. Table 5.13 shows the estimated values. To get fitted probabilities we have to compute the distances between subject points ($x_{iq}\beta_{qa}$) and class points (z_{ga}) (See equation 2.4). Let us analyze diagnostics for this model. Figure 5.9 panel (a) is the plot of the leverage values versus the individual deviances. In the bottomleft part of the graph, there are points with high leverage, while in the topright part there are points with low leverage but large individual deviance. Figure 5.9 panel (b) shows the index plot of the approximation to Cook's distance. Points with high leverage values or with low leverage but high deviance are detected. Finally, panel (c) is the index plot of the studentized residuals.

	Dimension	
	1	2
α	2.3879	-0.6666
β_{AST}	0.0059	0.0096
β_{ALT}	-0.0104	-0.0068
β_{GIDH}	-0.0084	0.1025
z_1	0.0000	0.0000
z_2	0.5860	0.0000
z_3	0.7080	1.5262
z_4	3.0010	1.3803

Table 5.13: Two dimensional Multinomial Distance Model: Estimated parameters.



(a) (b) (c)

Figure 5.9: Hepatits Data. 2-dimensional Multinomial Distance Model: Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of studentized residuals.

For subjects $18, 58, 77, 89, 93, 94$ and 136 we computed $\sum_{k\neq j}\Delta_k d_j^2$ to see wheter they affect the prediction process, Table 5.14 show the results.

Case number	$\sum_{k\neq j}\Delta_k d_j^2$
18	2.9060797
58	-2.9978030
77	2.4722745
89	3.3221793
93	-0.6618156
94	-0.4840309
136	-0.5389780

Table 5.14: Two dimensional Multinomial Distance Model: $\sum_{k\neq j}\Delta_k d_j^2$ diagnostic.

Deleting subjects with negative values of $\sum_{k\neq j}\Delta_k d_j^2$ improves the fit. In fact, fitting the model without those four subjects the misclassification rate becomes 0.16 and the deviance divided by the sample size is 0.79, compared to 0.93 which was the previous estimate for the completed data.

5.3.2 *Baseline Category Logit Model Analysis*

Baseline Category Logit Model is fitted on Hepatitis data as well. The number of parameters to estimate is $4 * 3 = 12$. Table 5.15 shows parameter estimates. Deviance is 192.63 and misclassification rate is 0.16.

Coefficients	1	2	3	4
α	-11.961094	5.991828	-6.240744	0
$\beta_{\log(AST)}$	-9.502010	-9.681154	-1.992770	0
$\beta_{\log(ALT)}$	13.492770	9.952411	2.722534	0
$\beta_{\log(GIDH)}$	-4.472537	-3.816352	1.139102	0

Table 5.15: Baseline category logit model: estimated coefficients.

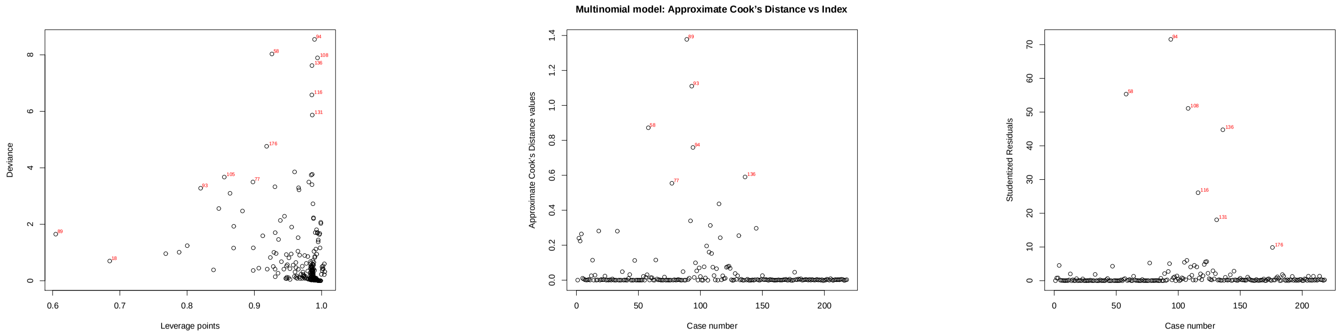
As before, we applied diagnostics on this model. Figure 5.10 panel (a) shows the leverage values versus the deviances. Detected points are the same of those detected in two dimensional multinomial distance model. Figure 5.10 panel (b) shows the index plot of the approximation to Cook's distance. Influential cases are the same of those detected in the previous model. Panel (c) shows the index plot of the studentized residuals. Again, detected points are the same of those detected by two dimensional multinomial distance model.

We analyze case $58, 77, 89, 93, 94, 108, 116, 131, 136$ and 176 further. Table 5.16 shows $\sum_{k \neq j} \Delta_k d_j^2$ diagnostic to investigate which points affect the prediction rule.

Case number	$\sum_{k \neq j} \Delta_k d_j^2$
58	0.260949579
77	0.873243297
89	0.162310564
93	-0.118970295
94	5.513176780
108	0.002561748
116	0.095021850
131	0.039422361
136	-0.005191135
176	0.022836977

Table 5.16: Baseline Category Logit Model: $\sum_{k \neq j} \Delta_k d_j^2$ diagnostic.

According to $\sum_{k \neq j} \Delta_k d_j^2$ diagnostic only subjects 93 and 136 affect the prediction. In two dimensional multinomial logit model also points 54 and 94 were detected. After deleting points 93 and 136, the fit is slightly better, the deviance divided by the sample size is 0.82 while before was 0.88. The misclassification rate does not change. This was predicted by the fact that the magnitude of $\sum_{k \neq j} \Delta_k d_j^2$ for deleted cases are not too large. Conclusions for this model is not the same of those for 2-dimensional multinomial distance model. It is important to note that the two models are different and probably they do not have exactly the same outliers.



(a) (b) (c)

Figure 5.10: Hepatitis Data. Baseline Category Logit Model:Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of studentized residuals.

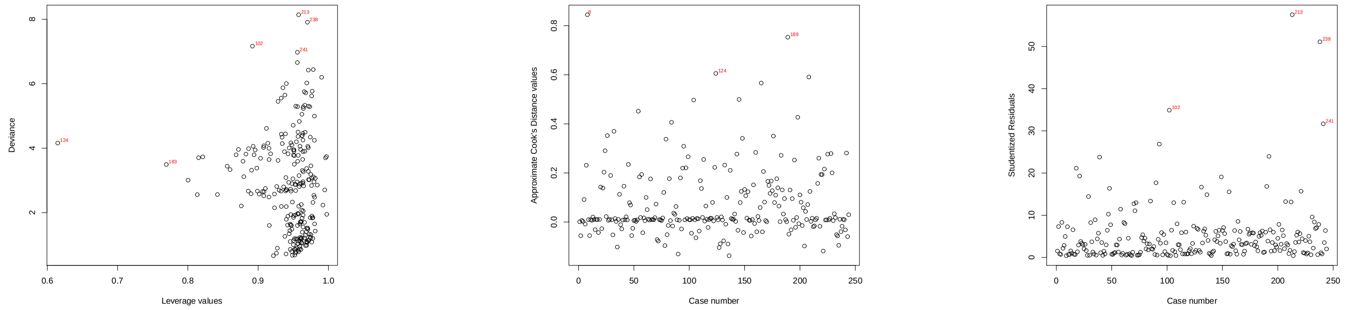
5.4 *Early Childhood Longitudinal Study - Kindergarten Cohort Analysis*

Data set comes from a cohort study performed in U.S.A.. The goal is to assess the proficiency in early reading at the end of the kindergarten year. The response variable is the proficiency level distinguished in 6 levels. Predictor variables are gender, minority status, whether the child attended half-day kindergarten, number of family risks, frequency with which parents read books to child, family socio-economic status and assessment age. For more details see O'Connell and Liu, 2011. We fit an one dimensional multinomial distance model and apply diagnostics to compare detected outliers with those detected from O'Connell and Liu in their article. They applied analysis on the full sample and two subsamples. We decided to use subsample I. Table 5.17 contains descriptive statistics for the chosen sample.

Model deviance is 708.58 and the deviance divided by the sample size is 2.9. Misclassification rate is 0.91, then the fit is very poor. Only one point, case 124, has very low leverage value. In figure 5.11 panel (a) we can see that there are some points with large individual deviance. Figure 5.11 panel (b) shows the index plot of the approximation to the Cook's distance. We can see that there are point $8, 124$ and 189 which have slightly different values of that diagnostics from the others. Finally, panel (c) is the index plot of studentized residuals. Detected points are $120, 213, 238$ and 241.

	Reading Proficiency Level						
	0	1	2	3	4	5	$Total$
	n=26	n= 32	n=54	n=108	n=16	n=8 N = 244	
$\%\ male$	12(46%)	21(66%)	32(59 %)	57(53 %)	6(38 %)	3(38 %)	131(54 %)
$\%\ minority$	19(73.08%)	16(50%)	30(55.56 %)	36(33.33 %)	6(37.5 %)	3(37.5 %)	110(45.08 %)
$risknummean$	1.38	0.75	0.69	0.5	0.25	0.25	0.64
$sesmean$	-0.8108	-0.2256	0.0057	0.1073	0.1019	0.2250	-0.0532
$\%\ halfday$	18(69.23 %)	15(46.88%)	23(42.59 %)	53(49.07 %)	9(56.25 %)	4(50 %)	122(50%)
$\%\ readbk$	18(69.23 %)	23(71.78%)	42(77.78 %)	88(81.48 %)	15(93.75 %)	7(87.5 %)	193(79.10 %)
$agemean$	75.74	75.44	75.81	74.57	74.44	77.95	75.18

Table 5.17: Summary Statistics for Kindergarten data.



(a) (b) (c)

Figure 5.11: Early Childhood Longitudinal Study Data. Multinomial Distance Model: Panel (a) plots leverage values versus individual deviances. Panel (b) is the index plot of the approximation to the Cook's distance. Plot (c) is the index plot of studentized residuals.

For these points we computed the $\sum_{k\neq j} \Delta_k d_j^2$ diagnostic to see if they affect prediction. Table 5.18 shows the results. Among the others, points $102, 213$ and 238 also affect the prediction.

Case number	$\sum_{k\neq j} \Delta_k d_j^2$
8	5.54991201
102	-1.64145230
124	3.08089830
183	4.92655205
189	1.13849317
213	-1.32184322
238	-0.07999275
241	0.75579429

Table 5.18: Multinomial Distance Model: $\sum_{k\neq j} \Delta_k d_j^2$ diagnostic.

O'Connell and Liu, in their article, fitted five models, one for each split based on six response variable. For each split, they applied diagnostics for simple logistic regression and they detected 8 cases. Table 5.19 shows these outliers.

Model	Case number
(0) vs others	70, 115, 124, 207
(0+1) vs others	124
(0+1+2+3) vs others	241, 136
(0+1+2+3+4) vs 5	149, 238

Table 5.19: Detected outliers in the five simple logistic regressions.

The main outlier is point 124. It has the largest leverage value in O'Connell and Liu's analysis as well as in ours. Cases $124, 238$ and 241 are outliers in both analyses. We also inspected prediction influence and concluded that cases $115, 213$ and 238 affect the prediction process. There are some differences between other outliers but it depends on the fact that O'Connel and Liu (2011) fitted 5 simple logistic regression models while we fitted a multinomial model.

DISCUSSION 6

The usefulness of a model is defined by its characteristics and the possibility to evaluate it after the fit. In fact, the most important part in a statistical analysis it is to assess wheter the model fit is good. Without this part, any model losts the practical verve and it is not still useful.

The Multinomial Distance Model is a good tool in classification problems. Its main weakness is the lack of diagnostics. This deficiency makes the model less appetible than others.

To evaluate outliers and influential cases in a model, we need to define its design matrix. In fact, starting from this, it is possible to compute diagnostic measures to assess cases which are far from the centroid of the space spanned by X. The Multinomial Distance model is a bilinear model, which means that it is multiplicative in the parameters. This feature makes the assessment process difficult.

In this work we showed that it is possible to extend the generalized linear model diagnostics to multinomial distance model. We started from the fact that it is possible to rewrite the one dimensional multinomial distance model as a baseline category logit model form. This means that we can find both coefficient and design matrices.

The main features are the $pseudo - \widetilde{\mathbf{S}}$ coefficient matrix and the *pseudo*-design matrix $\widetilde{\mathbf{X}}$. In fact, we have been able to define the same matrices of a baseline category logit models, based on the fact that also in multinomial distance model there is one *pseudo*-coefficient for each category. Thus, we have:

$$Pseudo - \widetilde{\mathbf{S}} = \begin{bmatrix} \widetilde{\alpha}_1 & \widetilde{\alpha}_2 & \dots & \widetilde{\alpha}_G \\ \widetilde{\beta}_{11} & \widetilde{\beta}_{12} & \dots & \widetilde{\beta}_{1G} \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{\beta}_{q1} & \widetilde{\beta}_{q2} & \dots & \widetilde{\beta}_{qG} \end{bmatrix}$$

where $\widetilde{\alpha}_g$ are the *pseudo*-intercepts and $\widetilde{\beta}_{gq}$ the *pseudo*-coefficients. As we can see, this matrix is the same as the coefficient matrix we estimate when fitting the baseline category logit model. In the same way, we have also defined the *pseudo*-design matrix $\widetilde{\mathbf{X}}$, which is formed by N stacked blocks:

$$\widetilde{\mathbf{X}}_k = \begin{bmatrix} \mathbf{x}_k & 0 & 0 & \dots & 0 \\ 0 & \mathbf{x}_k & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{x}_k \end{bmatrix}$$

and each $\mathbf{x}_k$ is the observed predictor vector of subject k. Thus, $\widetilde{\mathbf{X}}$ is a $NG \times G$ matrix. Once we obtained this matrices we can apply multiple group logistic regression diagnostics to multinomial distance model as it discussed in chapter 4.

6.1 Some Extentions

The obtained results can be generalized to all models for which it is possible to compute *pseudo*-$\widetilde{\mathbf{S}}$ matrix and *pseudo*-design matrix.

In chapter 2 we presented some other models for multicategorical response variable. We saw that the adjacent category logit model is a different parametrization of the baseline category logit model (see equation (2.15)). Thus, for this model it is possible to apply any multiple group logistic regression diagnostics without any effort.

Continuation-ratio logit model allows to different intercepts and coefficients for each response category. Therefore, we can compute the *pseudo* design matrix and apply multiple group logistic regression diagnostics. For this model it is also possible to extend one step approximation to the estimated coefficients. The coefficient matrix and the *pseudo*-$\widetilde{\mathbf{S}}$ matirx here are the same.

We can also compute the same matrices for Stereotype model. This model is multiplicative in the parameters. We have to estimate $G-1$ intercepts, one β and $G-2$ ϕ parameters. Then, to obtain probabilities, we have to multiply the ϕ parameters by β, getting $G-1$ *pseudo*-coefficients

$$\widetilde{\alpha}_g = \alpha_g \qquad \widetilde{\beta}_g = \phi_g \beta.$$

Once we computed the *pseudo*-$\widetilde{\mathbf{S}}$ matrix and the *pseudo* design matrix, multiple group logistic regression diagnostics cab be easily applied to Stereotype model, too.

Finally, we can generalize the *pseudo* matrices to the Ideal Point Discriminant Analysis. After fitting the model, we obtain two matrices and one vector of coefficients, that are, **Z** for the group coordinates, **B** for the predictors and **w** for the bias parameters. We can compute the *pseudo*-$\widetilde{\mathbf{S}}$ matrix where

$$\widetilde{\alpha}_g = (2\alpha(z_g) - z_g^2) + \log(w_g) \qquad \widetilde{\beta}_g = 2\beta(m_g) + \log(w_g).$$

Thus, multiple group logistic regression diagnositics can be applied to the Ideal Point Discriminant Analysis, too.

6.2 Conclusions

This monograph proposed a way to compensate for the lack of model evaluating tools for the Multinomial Distance Model. It is shown that the *pseudo*-$\widetilde{\mathbf{S}}$ matrix and the *pseudo* design matrix can also be computed for all models that allow different intercepts and coefficients for each response category. Therefore, the obtained results are also extended to other models.

The multiple group logistic regression diagnostics can also be applied to Multinomial Distance Model in more than one dimension. In fact, in chapter 4 we showed that it is possible to compute the *pseudo* coefficients for more than one dimensional multinomial distance model.

As we explained in chapter 4, some problems arise when we try to use one step approximation to the case deletion method to obtain estimated coefficients. The main problem is that the *pseudo* design matrix is not the same design matrix as the one used in the estimating process and the *pseudo*-$\widetilde{\mathbf{S}}$ matrix is not the matrix of the coefficient estimated.

Furthermore, in this monograph we considered only this approximation to assess outliers and leverage points. Further analysis and comparisons could be done. For example, finding a way to extend one step approximation of $dfbeta$ or comparing with other models like categorical regression model $(CATREG)$. Moreover, a bayesian approach to detect outliers could be applied.

7 Appendix

New Algorithm to compute the Multinomial Distance Model in one dimension:

```
deviance.MDM <- function(pars, Y, X){
# pars is the vector of initializing values for parameters
# Y is the response matrix (N x G)
# X is the predictor matrix (N x Q) with the first column
# equals to 1 for the intercept
# extract matrix B from pars
n = nrow(Y)
  J = ncol(Y)
  p = ncol(X)
  B = matrix(pars[1:p], p, 1)
# create Z - matrix with coordinates of class points
  Z = matrix(0, J, 1)
# only one constrain
  Z[1:(nrow(Z)-1)] = pars[(p+1):(p+(J-1))]
# make the matrix of coefficients
  Betas = 2*B%*%t(Z)
  Betas[1,] = Betas[1,]-t(Z^2)
# Make the linear predictors
  U = X%*%Betas
# Compute probabilities and deviance
  P = exp(U)
  sp = rowSums(P)
  P = (1 / sp) * P
  dev = -2 * sum(Y * log(P))
}
```

The function to compute the pseudo design matrix:

```
mat=function(m){
# m is the predictor matrix
m. = do.call(cbind, replicate((J-1), m, simplify=F))
ID = list((J-1))
for(i in 1:(J-1)){
ID[[i]] = matrix(rep(1,ncol(m)),1)
}
U = as.matrix(bdiag(ID))
return(m.*U)
}


##############################################################
#####################    DIAGNOSTICS     #####################
##############################################################

#following Lesaffre and Albert (1989):

# compute Sigma matrix (see pag. 433)
Sigma <- list(rep(0,n))
for(i in 1:n){
Sigma[[i]] <- sqrt(diag(1/P[i,]))
}
SI <- as.matrix(bdiag(Sigma))

# compute the Q.hat matrix  (see pag. 433)
Q <- list(rep(0,n))
```

```
for(i in 1:n){
Q[[i]] <- rbind((matrix(rep.int(P[i,1:(J-1)],(J-1)),(J-1),
(J-1),byrow=T)*(diag(1,(J-1))-matrix(rep.int(P[i,1:(J-1)],
(J-1)),(J-1),(J-1)))),-P[i,J]*P[i,1:(J-1)])
}
Q.hat <- bdiag(Q)

# compute the V.hat matrix  (see pag. 427)
V <- list(rep(0,n))
for(i in 1:n){
V[[i]] <- matrix(rep.int(P[i,],(J-1)),(J-1),(J-1),byrow=T)*
(diag(1,(J-1))-matrix(rep.int(P[i,],(J-1)),(J-1),(J-1)))
}
V.hat <- bdiag(V)

# compute the pseudo design matrix X
X.new <- list(rep(0,n))
for(i in 1:n){
X.new[[i]] <- mat(matrix(rep.int(X[i,],(J-1)),(J-1),ncol(X),
byrow=T))
}
X. <- do.call(rbind,X.new)

# finally compute the Generalized Hat Matrix (pag. 433)
W <- (solve(t(X.)%*%V.hat%*%X.))
H <- (SI%*%Q.hat%*%X.%*%W%*%t(X.)%*%t(Q.hat)%*%SI)

# and the generalized M matrix
M <- as.matrix(diag(nrow(H))-H)

# compute the det of the diagonal blocks of the M matrix
# to detect leverage points
```

```
m <- list(rep(0,n))
det. <- numeric(0)
index <- seq(1,(n*J), by=J)
for(i in 1:n){
m[[i]] <- as.matrix(M[index[i]:(index[i]+(J-1)),index[i]:
(index[i]+(J-1))])
det.[i] <- det(m[[i]])
}

# compute standardized residual vector (pag. 428)
r = Y-P
chi <- matrix(0,n,J)
for (i in 1:n){
chi[i,] <- (Sigma[[i]])%*%t(t(r[i,]))
}

# compute studentized residuals
# (Thanks to Cajo Ter Braak for pow.matrix function)
m.1 <- lapply(m,solve)
chi.star <- numeric(0)
for(i in 1:n){
chis <- pow.matrix(m.1[[i]], 0.5)%*%chi[i,]
chi.star[i] <- t(chis)%*%chis
}

# compute the individual deviance
d=numeric(0)
for(i in 1:n){
d[i] <- -2*(Y[i,]%*%log(P[i,]))
}

# compute one step approximation to Cook's distance
```

```
h. <- list(rep(0,n))
index <- seq(1,(n*J), by=J)
for(i in 1:n){
h.[[i]] <- as.matrix(H[index[i]:(index[i]+(J-1)),index[i]:
(index[i]+(J-1))])
}
m.1 <- lapply(m,solve)
cooks <- numeric(0)
for(i in 1:n){
cooks[i] <- round(chi[i,]%*%(m.1[[i]])%*%h.[[i]]%*%
(m.1[[i]])%*%t(t(chi[i,])), digits=5)
}

# compute one step approximation to neighboring effects
# (code for simulation study)
delta.d1 <- matrix(0,n,4)
index2 = index[c(301,302, 303,304)]
index3 = c(301,302, 303, 304)
for (i in 1:ncol(delta.d1)){
for(j in 1:n){
if(index3[i]!=j){
delta.d1[j,i] <- as.matrix(2*chi[j,]%*%H[index[j]:
(index[j]+(J-1)),index2[i]:(index2[i]+(J-1))]%*%
m.1[[index3[i]]]%*%t(t(chi[index3[i],])) +
chi[index3[i],]%*%m.1[[index3[i]]]%*%
H[index2[i]:(index2[i]+(J-1)),index[j]:
(index[j]+(J-1))]%*%H[index[j]:(index[j]+(J-1)),
index2[i]:(index2[i]+(J-1))]%*%
m.1[[index3[i]]]%*%t(t(chi[index3[i],])))
}
else{
delta.d1[j,i]=0
}
}
}
colSums(delta.d1)

###############################################################
################    GRAPHICAL DIAGNOSTICS      ###############
###############################################################

# plot individual deviances versus leverage values
x11()
names = factor(1:n)
plot(det.,d, ylab="Deviance", xlab="Leverage values")

# Index plot of cook’s Distance
x11()
plot(c(1:n), cooks, xlab="Case number",
ylab="Approximate Cook’s Distance values")

# Index plot of Studentized residuals
x11()
plot(c(1:n), chi.star, xlab="Case number",
ylab="Studentized residuals")


###############################################################
##################      LOO DIAGNOSTICS      #################
###############################################################

# case deletion diagnostics
p <- ncol(X)
res <- matrix(0,((J-1)+p),n)
```

```
SE <- list(rep(0,n))
DEVIANCES <- numeric(0)
for(i in 1:n){
stats <- optim(pars0, deviance.MDM, NULL, Y[-i,], X[-i,],
method ="BFGS", control = list(trace = 2, maxit = 100),
hessian = T);
res[,i] <- stats$par
SE[[i]] <- stats$hessian
DEVIANCES[i] <- stats$value
}

# plot deviances computed deleting each observation in turn
plot(c(1:n), DEVIANCES, xlab="Deleted Case Number",
ylab="Residual Deviance")

# compute dfbetas
delta.beta=matrix(0,nrow(res),ncol(res))
for(i in 1:ncol(res)){
delta.beta[,i]=coeff-res[,i]
# coeff is the vector of coefficient estimated on complete data
}

# compute standard errors SE(-k)
Stand.Err <- matrix(0,nrow(delta.beta), ncol(delta.beta))
for(i in 1:n){
Stand.Err[,i] <- sqrt(diag(solve(SE[[i]])))
}

# standardized dfbetas
influential.function <- delta.beta / Stand.Err

# plot dfbetas for each parameter (simulation study)
```

```
x11()
plot(c(1:n),delta.beta[1,], xlab="Deleted Case Number",
ylab="Coefficient Difference" )
x11()
plot(c(1:n),delta.beta[2,],xlab="Deleted Case Number",
ylab="Coefficient Difference"  )
x11()
plot(c(1:n),delta.beta[3,],xlab="Deleted Case Number",
ylab="Coefficient Difference" )
x11()
plot(c(1:n),delta.beta[4,],xlab="Deleted Case Number",
ylab="Coefficient Difference" )
x11()
plot(c(1:n),delta.beta[5,],xlab="Deleted Case Number",
ylab="Coefficient Difference")


#  compute LOO neighbouring effect
Prob.casedeleting <- list(rep(0,n))
for(i in 1:n){
p <- 2
B <- matrix(res[1:p,i], p, 1)
# create Z - matrix with coordinates of class points
Z <- matrix(0, J, 1)
l <- nrow(res)
Z[1:(nrow(Z)-1)] <- res[(p+1):l,i]
Betas = 2*B%*%t(Z)
Betas[1,] = Betas[1,]-t(Z)^2
# now define "linear predictors"
U <- X%*%Betas
# compute probabilities:
p <- exp(U)
sp <- rowSums(p)
Prob.casedeleting[[i]] <- (1 / sp) * p
}

Pr <- apply(P, 1, max)
# P is the probability matrix computed on complete data
Prob.final <- matrix(0,n,n)
for(i in 1:n){
Prob.final[,i] <- apply(Prob.casedeleting[[i]], 1, max)
}
# finally, compute a(-k)
a.minusk <- matrix(0, n , n)
for(i in 1:n){
a.minusk[,i] <- Pr-Prob.final[,i]
}
```

BIBLIOGRAPHY

Agresti, A., (2002). *Categorical Data Ananlysis*. J. Wiley. New York.

Agresti, A., (2010). *Ananlysis of Ordinal Categorical Data*. J. Wiley. New York.

Anderson, J. A., (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19-35.

Anderson, J. A., (1974). Diagnosis by logistic discriminant function. *Applied Statistics*,23,397-404.

Anderson, J. A., (1982). Logistic Discrimination. In P. R. Krishnaiah and L. Kanal (Eds.), *Handbook of statistics 2*.Amsterdam: North Holland.

Anderson, J. A., (1984). Regression and Ordered Categorical Variables. *Journal of the Royal Statistical Society*, 46, 1-30.

Coombs, C. H., (1964). *A theory of Data*. J. Wiley. New York.

Day, G., Shocker, A., Srivastava, R., (1979). Customer-Oriented Approaches to Identifying Product Markets. *Journal of Marketing*, 43, 4.

De Rooij M., (2009). Ideal Point Discriminant Analysis with a special emphasis on visualization. *Psychometrika*, 74, 317-330.

Fahrmeir L., Tutz, G., (2001). *Multivariate Statistical Modelling Based on Generalized Linear Model*, Springer, New York.

Fox, J., (1991). *Regression Diagnostics: An Introduction*. Sage Publications.

Goodman, L. A., (1983). The analysis of dipendence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics*, 39, 149-160.

Lesaffre, E., Albert A., (1989). Multiple-Group Logistic Regression Diagnostics. *Journal of the Royal Statistical Society*, 38, 425-440.

Liu, I., Agresti, A., (2005). The Analysis of Ordered Categorical Data: An Overview and Survey of Recent Developments. *Sociedad de Estadistica e Investigacion Operativa TEST*, 14, 1-30.

Luce, R. D., (1959). *Individual choice behaviour: A theoretical analysis*. J. Wiley. New York.

McFadden, D., (1980). Econometric models for probabilistic choice among products. *Journal of Business*, 53, S13-S29.

McGullagh, P., (1980). Regression models for Ordina Data. *Journal of the Royal Statistical Society*, 42, 109-142.

Nelder, J., Wedderburn, R. W. M., (1972). Generalized linear models. *J. R. Statist. Soc.*, 135, 370-384.

O'Connell, A. A., Liu X., (2011). Model Diagnostics for Proportional and Partial Proportional Odds Models. *Journal of Modern Applied Statistical Methods*, 10, 139-175.

Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., ...Consortium, N. R. (2008). The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17,121140. doi:10.1002/mpr.256.

Peterson, Bercedis and Frank E. Harrell, Jr. (1990). Partial Proportional Odds Models for Ordinal Response Variables. *Applied Statistics*, 39, 205-217.

Pregibon, D., (1981). Logistic Regression Diagnostics. *Ann. Statist.*, 9, 705-724.

Simon, G., (1974). Alternative Analysis for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69, 971-976.

Spinhoven, P., de Rooij, M., Heiser, W., Smith, J.H., Penninx, B. W. J. H. (2012, May 7). Personality and Changes in Comorbidity Patterns Among Anxiety and Depressive Disorders. *Journal of Abnormal Psychology*, Advance online publication. doi: 10.1037/a0028234.

Takane, Y., Bozdogan, H., Shibayama, T., (1987). Ideal Point Discriminant Analysis. *Psychometrika*, 52, 371-392.

Takane, Y., (1987). Analysis of contingency tables by Ideal Point Discriminant Analysis. *Psychometrika*, 52, 493-513.

Takane, Y., (1989). Ideal Point Discriminant Analysis and ordered response categories. *Behaviormetrika*, 26, 31-46.

Takane, Y., (1998). Visualization in ideal point discriminant analysis. *Psychometrika*, 52, 371-392.

Tutz, G. (1991). Sequential models in categorical regression. *Comput. Statist. Data Anal.*, 11, 275-295.

Williams, D., (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36(2), 181-91.

Wolfe, P., (1969). Convergence Conditions for Ascent Methods. *SIAM Review*, 11, 226-235.