AUSTRIAN JOURNAL OF STATISTICS
Volume 39 (2010), Number 3, 182–202

Some Properties of a Recently Introduced Approach to Ordinal Regression

Florian Frommlet
Department of Statistics and Decision Support Systems, University Vienna

Abstract: The statistical properties of a novel approach to ordinal regression which was only recently introduced in the literature are discussed. It is shown that for ordinal explanatory variables the approach is equivalent to isotonic regression, with some advantages when dealing with two-sided alternatives. For ordinal response variables the procedure behaves very differently and is asymptotically equivalent to a two-sample t-test between the extreme categories. A penalized version is introduced to improve power, and the procedure is evaluated using Monte Carlo simulations. Finally the method is applied to microarray gene expression data on prostate cancer.

Zusammenfassung: Die statistischen Eigenschaften einer Methode zur Behandlung von Regression mit ordinalen Variablen wird untersucht, die erst vor kurzem in der Literatur vorgeschlagen wurde. Es wird gezeigt, dass sich dieser Zugang für erklärende ordinale Variablen nicht von der bekannten isotonen Regression unterscheidet. Ist hingegen die abhängige Variable ordinal, so verhält sich die neue Prozedur asymptotisch so wie ein Zweistichproben t-Test zwischen den beiden jeweils extremsten Kategorien. Eine pönalisierte Version des Verfahrens liefert im allgemeinen etwas bessere Power, was in einer Simulationsstudie dokumentiert wird. Schlussendlich wird die Methode auf Microarray – Genexpressionsdaten einer Prostatakrebsstudie angewendet.

Keywords: Isotonic Regression, Ordinal Regression, Least Squares, Asymptotics, Microarrays.

1 Introduction

The study of regression models for ordered categorical response variables dates back to the 1950's, for a comprehensive recent review article see Liu and Agresti (2005). Incited by a seminal article of McCullagh (1980) the topic of ordinal regression became a fruitful field of research. In a recent paper Torra, Domingo-Ferrer, Mateo-Sanz, and Ng (2006) proposed a new approach to regression with ordinal variables without latent variables. They suggest to map the levels of the ordinal variable into the interval $[0, 1]$ and take this mapping into account when calculating least squares estimates. Up to our knowledge this approach has not been investigated before, and the purpose of this article is to analyze the statistical properties of such a procedure, which is not really dealt with in Torra et al. (2006).

To this end we will restrict ourselves to simple univariate models, and we will study separately the cases where only the explicatory variable or only the dependent variable

is ordinal. In Section 1 we will show that in the first case the procedure is equivalent to isotonic regression, though it has some advantages when dealing with two-sided alternative hypothesis. The second case will be dealt with in Section 2, where the approach suggested by Torra et al. (2006) actually leads to a novel procedure, which is however asymptotically equivalent to a two-sample t-test between the values of the explicatory variable which correspond to the lowest and the highest level of the dependent variable. We will introduce penalties to obtain a more interesting procedure, and compare its power with simple regression where the levels of the dependent variable are assumed to be equidistantly fixed, and with logistic regression based on the proportional odds assumption (McCullagh, 1980). Finally the method is applied to publicly available gene expression data from a study on clinical prostate cancer behavior Singh et al. (2002). The performance of the new method is compared with the original data analysis and with an approach to ordinal regression based on Gaussian processes Chu, Ghahramani, Falciani, and Wild (2005).

2 Independent Variable Ordinal

Let X be an ordinal variable without an underlying latent variable, and Y a metric response variable. Without loss of generality we will code throughout this section the $l+1$ levels of X as $\Omega_X := \{0, 1, \ldots, l\}$. We study the following type of model

$$Y = b_0 + b_1 f(X) + \varepsilon\,, \tag{1}$$

where $f \in \mathcal{F} := \{f : \Omega_X \to [0,1],\ \text{nondecreasing, not constant}\}$ is the set of all mappings of the levels of X into the interval $[0,1]$ with $f(j) \geq f(i)$ if $j \geq i$. We do not allow constant mappings to avoid trivial degenerations. The error term ε is some random variable with $\mathrm{E}(\varepsilon) = 0$, which we will not further specify.

The goal is to perform least squares estimation of b_0 and b_1, which also minimizes over $\mathcal{F}$. This means that for a random sample of observations (y_j, x_j), $j = 1, \ldots, n$, we want to solve

$$\inf_{b_0, b_1; f} \left\{ \sum_{j=1}^{n} (y_j - b_0 - b_1 f(x_j))^2 :\ f \in \mathcal{F} \right\}. \tag{2}$$

We can actually identify any function $f \in \mathcal{F}$ with a vector $\mathbf{f}$ of dimension $l+1$, where $\mathbf{f} = (c_0, c_0 + c_1, \ldots, c_0 + \cdots + c_l)^\top$, with $c_i \geq 0$ and $\sum c_i \leq 1$. Now due to linearity the infimum of (2) is clearly shift invariant with respect to $\mathbf{f}$, and without loss of generality we can assume that $c_0 = 0$. Having done so the infimum of (2) is also scale invariant and we can further assume that $c_1 + \cdots + c_l = 1$. By denoting $\mathbf{c} = (c_1, \ldots, c_l)^\top$, we can thus formulate a parameterized version of (2):

$$\inf_{b_0, b_1; \mathbf{c}} \left\{ \sum_{j=1}^{n} (y_j - b_0 - b_1 g_j(\mathbf{c}))^2 : \mathbf{c} \in \Delta^l \right\}, \tag{3}$$

where $\Delta^l := \{\mathbf{c} \in \mathbb{R}^l : c_i \geq 0, c_1 + \cdots + c_l = 1\}$ is the standard simplex of dimension l, and $g_j(\mathbf{c}) := f(x_j) = c_1 + \cdots + c_{x_j}$ corresponds to the state of the $j-$th observation

of X. If we only want to consider nonincreasing regression lines, i.e. in the situation of a one-sided test, we ask for $b_1 \in \mathbb{R}^+$, otherwise we have $b_1 \in \mathbb{R}$.

The optimization problem (3) has a vary particular structure, it is a biquadratic program over $\mathbb{R}^2 \otimes \Delta^l$. Keeping $\mathbf{c}$ fixed we have the usual convex quadratic program of ordinary linear regression, whereas keeping b_0 and b_1 fixed we obtain a so called standard quadratic program (Bomze, 1998). Due to this special structure it is easy to prove that in (3) the minimum is actually attained:

Proposition 2.1 *The sum of squared errors* $R(b_0, b_1, \mathbf{c}) := \sum_{j=1}^{n}(y_j - b_0 - b_1 g_j(\mathbf{c}))^2$ *attains its minimum within* $\mathbb{R}^2 \otimes \Delta^l$*, i.e.*

$$\inf_{b_0,b_1;\mathbf{c}} \left\{ R(b_0, b_1, \mathbf{c}) : \mathbf{c} \in \Delta^l \right\} = \min_{b_0,b_1;\mathbf{c}} \left\{ R(b_0, b_1, \mathbf{c}) : \mathbf{c} \in \Delta^l \right\} .$$

Proof. For each fixed $\mathbf{c} \in \Delta^l$ the minimum of $R(b_0, b_1, \mathbf{c})$ is obtained by ordinary least squares regression. The solution $b_0(\mathbf{c})$, $b_1(\mathbf{c})$ thus obtained is continuous in $\mathbf{c}$, and Δ^l is a compact set. □

The same result holds for the one-sided situation where the minimum is attained in $\mathbb{R} \otimes \mathbb{R}^+ \otimes \Delta^l$. We will denote by $b_0^*, b_1^*; \mathbf{c}^*$ a vector that minimizes (3). Note that uniqueness of the solution is not always guaranteed, e.g. in case of $b_1^* = 0$ there are no conditions on $\mathbf{c}^*$. In general it is fairly easy to solve (3), because we can explicitly calculate the critical points. To this end we will introduce some notation. For an event $J \subset \Omega_X = \{0, \ldots, l\}$, we denote the mean of all observations y_j with $x_j \in J$ as $\bar{y}^J := \frac{1}{n_J} \sum_{x_j \in J} y_j$, where $n_J = \#(x_j \in J)$ is the number of observations with value in J. Without loss of generality we will assume that $\#(x_j = i) > 0$ for all $i \in \Omega_X$, i.e. we will neglect categories without any observations for our analysis.

Furthermore, we define the index set $I \subseteq \{1, \ldots, l\}$, where $i \in I$ signifies that the corresponding boundary condition for $\mathbf{c}$ is not active, i.e. $c_i > 0$, whereas for $i \in \{1, \ldots, l\} \backslash I$ we have $c_i = 0$. Because of $\mathbf{c} \in \Delta^l$ it is not possible that $c_i = 0$ for all $i \in \{1, \ldots, l\}$. For any such $I = \{s_1, \ldots, s_\nu\}$, $1 \leq \nu \leq l$, we define $J_0 := \{0, \ldots, s_1 - 1\}$, $J_1 := \{s_1, \ldots, s_2 - 1\}, \ldots, J_\nu := \{s_\nu, \ldots, l\}$. Each J_k contains the indices such that if $x_j \in J_k$ then $f(x_j) = c_{s_1} + \cdots + c_{s_k}$. We adopt the common convention that $\sum_{r=1}^{0} c_{s_r} = 0$, so specifically J_0 contains all categories of X which are mapped to 0 under f.

Let $\mathcal{I} := \mathcal{P}(\{1, \ldots, l\}) \backslash \emptyset$ be the set of all nonempty index sets, $\mathcal{I}_1 := \{I \in \mathcal{I} : |I| = 1\}$ and $\mathcal{I}_2 := \mathcal{I} \backslash \mathcal{I}_1$. The index sets in $\mathcal{I}_1$ with one index correspond to the edges of Δ^l, whereas the index sets of $\mathcal{I}_2 \backslash \Omega_X$ correspond to the boundary manifolds of Δ^l. The minimum of (3) can occur either at the critical points of the manifolds indexed by $\mathcal{I}_2$, or at the points given by $\mathcal{I}_1$. The following lemma describes all potential critical points:

Lemma 2.1 *Let* $I = \{s_1, \ldots, s_\nu\} \in \mathcal{I}_2$ *and assume that* $\bar{y}^{J_0} \neq \bar{y}^{J_\nu}$*. If the vector*

$$\begin{aligned} b_0^I &= \bar{y}^{J_0} , \\ b_1^I &= \bar{y}^{J_\nu} - b_0^I , \\ c_{s_i}^I &= (\bar{y}^{J_i} - b_0^I)/b_1^I - \sum_{r=1}^{i-1} c_{s_r}^I , \qquad \textit{for } 1 \leq r < \nu , \end{aligned} \tag{4}$$

is strictly (3)-feasible (i.e. $c^I_{s_i} > 0$, for all $i \in \{1, \dots, \nu\}$), then it is a critical point of 3. The corresponding value of the objective function equals

$$R(b^I_0, b^I_1, \mathbf{c}^I) = \sum_{i=0}^{\nu} \sum_{j:x_j \in J_i} (y_j - \bar{y}^{J_i})^2 \,. \tag{5}$$

Remark: Note that for the critical point in the interior of $\mathbb{R}^2 \otimes \Delta^l$ (i.e. for $I = \{1, \dots, l\}$) we obtain just the usual residual sum of squares of a one way ANOVA. This will be the solution of (3) if the means over the different categories $\bar{y}^i$ are monotonic in i. Otherwise we will have a solution on the boundary of $\mathbb{R}^2 \otimes \Delta^l$, which means that some categories are joint together. In the one-sided situation strict feasibility requires additionally that $b_1 > 0$.
Proof. For any given I we have $c_i > 0$, for all $i \in I$ and $c_i = 0$, for all $i \in \Omega_X \backslash I$. The residual some of squares becomes $\sum_{i=0}^{\nu} \sum_{j:x_j \in J_i} (b_0 + b_1 \sum_{r=1}^{i} c_{s_r} - y_j)^2 = 0$. Partial derivation with respect to b_0, b_1, and c_i for $i \in I$ leads to the following system of linear equations:

$$\begin{aligned}
&\sum_{i=0}^{\nu} \sum_{j:x_j \in J_i} T_{ij} = 0 \\
&\sum_{i=1}^{\nu} \sum_{j:x_j \in J_i} T_{ij} \sum_{r=1}^{i} c_{s_r} = 0 \\
&b_1 \sum_{i=s}^{\nu-1} \sum_{j:x_j \in J_i} T_{ij} = 0 \,, \quad \text{for } 1 \le s < \nu \,,
\end{aligned} \tag{6}$$

where we have used the abbreviation $T_{ij} := b_0 + b_1 \sum_{r=1}^{i} c_{s_r} - y_j$. Now for $b_1 \neq 0$ we can conclude that $\sum_{j:x_j \in J_i} T_{ij} = 0$ for each $i \in I$, and it follows that (4) is a solution of (6). □

Obviously this problem is intimately related with isotonic regression. To clarify the connection we introduce the following definition:

Definition 2.1 *For a finite set $X = \{x_0, \dots, x_l\}$ with simple order $x_0 \le x_1 \le \dots \le x_l$ a real valued function g is called* isotonic *on X if $x, y \in X$ and $x \le y$ imply $g(x) \le g(y)$.*

Let us write $\bar{y}(i) := \bar{y}^{\{i\}}$ for the sample mean over each category and $\bar{\mathbf{y}} := (\bar{y}(0), \dots, \bar{y}(l))^\top$. As stated in the remark after Lemma 2.1 this natural estimate $\bar{\mathbf{y}}$ gives the optimal solution of (3) when the order restrictions are fulfilled. Otherwise we have:

Theorem 2.1 *For any sample $(y_j, x_j), j \in \{1, \dots, n\}$, a solution of (3) in the one-sided situation ($b_1 \in \mathbb{R}^+$) is given by the isotonic regression $\bar{y}^*$ of $\bar{\mathbf{y}}$ with weights $(n_0, \dots, n_l)$, which is defined as*

$$\bar{y}^* = \min_g \left\{ \sum_{i=0}^{l} (\bar{y}(i) - g(i))^2 n_i \right\} \tag{7}$$

in the class of isotonic functions g on Ω_X.

Proof. For any arbitrary isotonic function $g(i)$, $i \in \Omega_X$, define $b_0 = g(0)$, $b_1 = g(l) - g(0)$, and $f(i) = (g(i) - g(0))/(g(l) - g(0))$. This defines an isomorphism between functions $b_0 + b_1 f(i)$ with $b_1 \geq 0$, $f \in \mathcal{F}$, and the class of isotonic functions on Ω_X. Then $\sum_{j=1}^{n}(y_j - b_0 - b_1 f(x_j))^2 = \sum_{i=0}^{l}\sum_{j:x_j=i}(y_j - g(i))^2$ and from the usual decomposition

$$\sum_{j:x_j=i}(y_j - g(i))^2 = (\bar{y}(i) - g(i))^2 n_i + \sum_{j:x_j=i}(y_j - \bar{y}(i))^2$$

we conclude that the two problems (7) and (1) are equivalent. □

A general introduction to isotonic regression can be found either in Robertson, Wright, and Dykstra (1988), or Barlow, Bartholomew, Bremner, and Brunk (1972) where in Chapter 2 several algorithms are presented how to find the solution of (7). Specifically the pool-adjacent-violator algorithm (PAVA) finds the isotonic regression $\bar{y}^*$ by starting with $\bar{\mathbf{y}}$ and applying a process of *amalgamation of means* over categories till the pooled means fulfill the order restriction. These pooled means are just of the form $\bar{y}^{J_i}$ as in Lemma 2.1, which provides another possibility to establish equivalence of (7) and (1).

Isotonic regression is more naturally formulated for the one-sided case, which corresponds to test the null hypothesis

$$H_0 : \mu_0 = \mu_1 = \cdots = \mu_l$$

against the one-sided alternative

$$H_1 : \mu_0 \leq \mu_1 \leq \cdots \leq \mu_l \,.$$

This test was first introduced by Bartholomew (1959a), where one can find among others the distribution of the test statistic $\bar{\chi}^2_{l+1} := \sigma^{-2}\sum_{i=0}^{l}(\bar{y}(i) - \bar{y}^*(i))^2 n_i$ under H_0. In a subsequent paper Bartholomew (1959b) considered the two-sided alternative

$$H_2 : \mu_0 \leq \mu_1 \leq \cdots \leq \mu_l \qquad \text{or} \qquad \mu_0 \geq \mu_1 \geq \cdots \geq \mu_l \,,$$

which he dealt with by applying isotonic regression separately under the increasing and under the decreasing alternative. This situation can be easier handled within this framework by minimizing (1) with $b_1 \in \mathbb{R}$. Solutions can be found e.g. via Lemma 2.1 and applying a PAVA-like algorithm. The following example illustrates a peculiarity of the two-sided case.

Example 2.1 *Consider a sample of three instances with* $\mathbf{x} = (0, 1, 2)$ *and corresponding* $\mathbf{y} = (a, b, a)$, $a, b \in \mathbb{R}$. *Note that for* f *there is only one degree of freedom, namely* $f(1) = c$. *For* $a \neq b$ *there are two solutions of (3):*

1. $c = 1$, $b_0 = a$, $b_1 = (b - a)/2$,
2. $c = 0$, $b_0 = (a + b)/2$, $b_1 = (a - b)/2$

with residual sum of squares $R(b_0, b_1, c) = (b - a)^2/2$. *In the trivial case* $a = b$ *we have the solution* $b_0 = a$, $b_1 = 0$ *and* c *might take any value within* $[0, 1]$.

Thus, in the non-trivial case we obtain two solutions, one with negative and one with positive slope. If the dependent variable Y itself is discrete this kind of behavior might be quite undesirable. Of course if Y is continuous, then such a situation would occur only with probability 0, but still there is some kind of instability with respect to small changes in the data of the dependent variable. This sort of problem becomes in the current context more transparent than in the formulation of two-sided isotonic regression.

3 Dependent Variable Ordinal

We next consider a metric variable X and an ordinal response with $l+1$ categories, i.e. $\Omega_Y := \{0, 1, \dots, l\}$. The corresponding model has the form

$$f(Y) = b_0 + b_1 X + \varepsilon\,, \tag{8}$$

where $f \in \mathcal{F} := \{f : \Omega_Y \to [0,1], f \text{ nondecreasing}, f(0) = 0, f(l) = 1\}$ and ε is some error function. We actually have to demand $f(l) = 1$, because otherwise the whole approach becomes entirely meaningless – for decreasing $f(l)$ naturally the sum of mean squared errors would also decrease. In the extreme case we would obtain the trivial solution $f \equiv 0$ and $b_0 = b_1 = 0$. As in the previous section we obtain a parameterized version by letting $f(i) = c_1 + \cdots + c_i$ for some vector $\mathbf{c} = (c_1, \dots, c_l)^\top \in \Delta^l$ and $g_j(\mathbf{c}) := f(y_j) = c_1 + \cdots + c_{y_j}$. In this section we will only consider the two-sided case, thus both b_0 and b_1 take values in $\mathbb{R}$.

Similar arguments as in Proposition 2.1 guarantee that a least squares estimate $b_0^*, b_1^*, \mathbf{c}^*$ can be obtained by solving

$$\min_{b_0, b_1; \mathbf{c}} \left\{ \sum_{j=1}^{n} (g_j(\mathbf{c}) - b_0 - b_1 x_j)^2 : \mathbf{c} \in \Delta^l \right\}. \tag{9}$$

Actually, (9) is in some sense simpler than (3), because we now have to deal with a quadratic optimization problem, and not with a biquadratic one. However, the critical points for this model do not have such a straight forward interpretation, and they are qualitatively very different from those of (3). We will use again the notation of Lemma 2.1, furthermore we define for an index set $J \subset \Omega_Y$ the variance of the x_j values with corresponding $y_j \in J$ as

$$\mathrm{var}(x)^J := \frac{1}{n_J} \sum_{y_j \in J} (x_j - \bar{x}^J)^2 = \frac{1}{n_J} \sum_{y_j \in J} x_j^2 - (\bar{x}^J)^2 = \overline{x^2}^J - (\bar{x}^J)^2\,.$$

To solve (9) we can use the following Lemma:

Lemma 3.1 *Let $I = \{s_1, \dots, s_\nu\} \in \mathcal{I}_2$. If the vector*

$$\begin{aligned}
b_0^I &= \frac{n_{J_\nu}}{n_{J_0} + n_{J_\nu}} - b_1^I \bar{x}^{J_0 \cup J_\nu}\,, \\
b_1^I &= \frac{n_{J_\nu}(\bar{x}^{J_\nu} - \bar{x}^{J_0 \cup J_\nu})}{(n_{J_0} + n_{J_\nu})\mathrm{var}(x)^{J_0 \cup J_\nu} + \sum_{i=1}^{\nu-1} n_{J_i} \mathrm{var}(x)^{J_i}}\,, \\
c_{s_i}^I &= b_0^I + b_1^I \bar{x}^{J_i} - \sum_{r=1}^{i-1} c_{s_r}^I\,, \qquad \textit{for } 1 \le i < \nu\,,
\end{aligned} \tag{10}$$

is (9)-feasible (i.e. $c_{s_i} > 0$, for all $i \in \{1, \ldots, \nu\}$), then it is a critical point of (9). The corresponding value of the objective function equals

$$R(b_0^I, b_1^I, \mathbf{c}^I) = \frac{n_{J_0} n_{J_\nu}}{n_{J_0} + n_{J_\nu}} \left(1 - b_1^I(\bar{x}^{J_\nu} - \bar{x}^{J_0})\right) . \tag{11}$$

Proof. Similar to the proof of Lemma 2.1 we can write the residual sum of squares for fixed index I as $\sum_{i=0}^{\nu} \sum_{j:y_j \in J_i} (b_0 + b_1 x_j - \sum_{r=1}^{i} c_{s_r})^2$. Writing $T_{ij} := b_0 + b_1 x_j - \sum_{r=1}^{i} c_{s_r}$ partial derivation with respect to b_0, b_1, and c_{s_r} leads to

$$\begin{aligned}
&\sum_{i=0}^{\nu} \sum_{j:y_j \in J_i} T_{ij} = 0 \\
&\sum_{i=0}^{\nu} \sum_{j:y_j \in J_i} T_{ij} x_j = 0 \\
&\sum_{i=s} \nu - 1 \sum_{j:y_j \in J_i} T_{ij} = 0 , \quad \text{for } 1 \leq s < \nu .
\end{aligned} \tag{12}$$

Now from the last set of $\nu - 1$ equalities we obtain $\sum_{j:y_j \in J_i} T_{ij} = 0$, for all $i \in \{1, \ldots, \nu - 1\}$, which leads to

$$b_0 + b_1 \bar{x}^{J_i} - \sum_{r=1}^{i} c_{s_r} = 0 , \quad \text{for } 1 \leq i < \nu . \tag{13}$$

The first equality of (12) gives

$$b_0 + b_1 \bar{x}^{J_0 \cup J_\nu} = \frac{n_{J_\nu}}{n_{J_0} + n_{J_\nu}} \tag{14}$$

and using (13) the second equality becomes

$$\sum_{j:y_j \in J_0} x_j(b_0 + b_1 x_j) + \sum_{j:y_j \in J_\nu} x_j(b_0 + b_1 x_j - 1) + b_1 \sum_{i=1}^{\nu-1} \sum_{j:y_j \in J_i} x_j(x_j - \bar{x}^{J_i}) = 0 . \tag{15}$$

Inserting (14) into (15) and some easy calculations finish the proof. □

We want to remark that just as in Section 2 the solution of (9) is given by the minimum of (11) over the finite number of critical points (with indices $I \in \mathcal{I}_2$) and over the edges of Δ^l (with indices $I \in \mathcal{I}_1$). In the latter case $\nu = 1$ and all categories of Y are either mapped to 0 or to 1. As we will see later on this is a rather untypical situation and will in practice hardly occur. The optimal values of b_0^I and b_1^I are still given by the corresponding formulas of (10).

The usual simple regression line has the property that the point $(\bar{x}, \bar{y})$ lies on it. Here this is typically not the case, but (10) shows that the regression line goes through $(\bar{x}^{J_0 \cup J_\nu}, \bar{y}^{J_0 \cup J_\nu})$. The formula of the slope b_1^I resembles simple regression considering only J_0 and J_ν, but the variances of $J_1, \ldots, J_{\nu-1}$ add to the denominator, and thus the regression line gets flatter. Note that for $b_1^I > 0$ the regression line defined by (10) will

be above the point $(\bar{x}^{J_0}, 0)$ and below $(\bar{x}^{J_\nu}, 1)$. In absolute terms its slope is smaller than a line running through those two points. The larger the variances $\mathrm{var}(x)^{J_i}$ the flatter the line.

Assume for a moment that we drop the order restriction implied by $\mathcal{F}$ and allow also for non-monotonous f. Clearly in the unrestricted case we obtain an optimal solution of (9) by $R(b_0^I, b_1^I, \mathbf{c}^I)$ as given in (11) for $I = \{1, 2, \ldots, l\}$. In the terminology of isotonic regression this is again a natural estimate and the solution under order restrictions is obtained by isotonization. This leads to the problem of finding I which minimizes $R(b_0^I, b_1^I, \mathbf{c}^I)$ while maintaining the order relation, which can be accomplished as in the previous section by applying a PAVA. For a more general discussion of algorithms of this kind we refer to Best and Chakravarti (1990).

3.1 Inference

If we want to construct a statistical test based for example on the absolute value of b_1^* we have to face the problem of how to specify the error function ε. This is not particularly straight forward, actually even the specification of H_0 and H_1 are not as obvious as in Section 2. Taking into account that Y is an ordinal variable we may simply consider the case where Y has a discrete distribution with $l+1$ levels and probabilities $p_0, \ldots, p_l$. We then can formulate the null hypothesis

$$H_0 : Y \text{ is independent of } X\,,$$

where X might be considered as a random variable or as given data. Alternative hypothesis will involve dependence of Y on X with some kind of inherent order. Denoting by $q_i(x) := p_0(x) + \cdots + p_i(x)$ the cumulative probabilities for given X one might consider e.g. the two-sided alternative

$$H_1 : \text{ All } q_i(x) \text{ either increase or decrease with } x\,.$$

The following theorem deals with consistency; under H_0 the regression line converges almost surely towards a constant.

Theorem 3.1 *Let X_i and Y_i be an i.i.d. sequences of independent random variables, X_i with finite mean μ, finite variance $\sigma^2 > 0$ and Y_i discrete with $l+1$ levels and probabilities $p_0, \ldots, p_l$. Furthermore let $(b_0^{(n)}; b_1^{(n)}; c_1^{(n)}, \ldots, c_l^{(n)})$ be the solution of (9) for the first n members of the random sequences. Then we have*

$$\left(b_0^{(n)}; b_1^{(n)}; c_1^{(n)}, \ldots, c_l^{(n)}\right) \xrightarrow{a.s.} \left(\frac{p_l}{p_0+p_l}; 0; \frac{p_l}{p_0+p_l}, 0, \ldots, 0, \frac{p_0}{p_0+p_l}\right)\,. \qquad (16)$$

Proof. Note that due to the strong law of large numbers $\bar{x}^{J_i} \xrightarrow{a.s.} \mu$ for any J_i. Because of $\sigma^2 > 0$ the variance terms in the denominator of b_1^I are bounded away from zero for any I, and we immediately obtain that $b_1^{(n)} \xrightarrow{a.s.} 0$.

Next consider that the scaled objective function of (9) formally converges to

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(g_j(\mathbf{c}) - b_0 - b_1 x_j)^2 &\to \sum_{i=0}^{l} \lim_{n\to\infty} \frac{1}{n} \sum_{j:y_j\in J_i} (b_0 - \sum_{r=1}^{i} c_r)^2 \\ &= p_0 b_0^2 + p_1(b_0 - c_1)^2 + p_2(b_0 - c_1 - c_2)^2 + \cdots + p_l(b_0 - 1)^2\,. \qquad (17)\end{aligned}$$

It is obvious that the minimum of (17) is obtained at $b_0^{\lim} := p_l(p_0 + p_l)^{-1}$ and $\mathbf{c}^{\lim} := (p_l(p_0+p_l)^{-1}, 0, \ldots, 0, p_0(p_0+p_l)^{-1})$, and it remains to argue that $b_0^{(n)}$ and $\mathbf{c}^{(n)}$ converge almost surely towards this limit.

From (13) it is evident that $\limsup b_0^{(n)} \leq 1$ and $\liminf b_0^{(n)} \geq 0$. Using

$$\left| \frac{1}{n} \sum_{j:y_j \in J_i} \left[b_1^2 x_j^2 + 2 b_1 x_j (b_0 - \sum_{r=1}^{i} c_r) \right] \right| \leq |b_1| \sum_{j:y_j \in J_i} \left(|b_1| \frac{x_j^2}{n} + 2 \left| b_0 - \sum_{r=1}^{i} c_r \right| \frac{|x_j|}{n} \right)$$

and the fact that by our assumptions the first two moments of X are bounded we conclude that the convergence in (17) is almost sure uniformly for the compact set $b_0 \in [0,1]$, $\mathbf{c} \in \Delta^l$. Now by standard arguments we conclude that $b_0^{(n)} \xrightarrow{a.s.} b_0^{\lim}$ and $\mathbf{c}^{(n)} \xrightarrow{a.s.} \mathbf{c}^{\lim}$. □

So asymptotically b_1^* converges towards 0, and it is easy to see that also $\mathrm{E}(b_1^*) = 0$ for finite n. Therefore one might consider to construct a statistical test based on b_1^*. For fixed I the numerator of b_1 in (10) can be rewritten as

$$n_{J_\nu} \left(\bar{x}^{J_\nu} - \bar{x}^{J_0 \cup J_\nu} \right) = \frac{n_{J_0} n_{J_\nu}}{n_{J_0} + n_{J_\nu}} \left(\bar{x}^{J_\nu} - \bar{x}^{J_0} \right) .$$

Due to the central limit theorem we have $\sqrt{n_{J_i}} \bar{x}^{J_i} \xrightarrow{d} \mathcal{N}(\mu; \sigma^2)$ for any J_i and under H_0 we have independence of the x_j corresponding to different J_i, because by definition all J_i are disjoint. Unfortunately independence between $\bar{x}^{J_0}$ and $\bar{x}^{J_\nu}$ is not given because n_{J_0} and n_{J_ν} are dependent random variables. However, it is still possible to obtain for fixed I that

$$\sqrt{n} \left(\bar{x}^{J_\nu} - \bar{x}^{J_0} \right) \longrightarrow \mathcal{N} \left(0; \sigma^2 (p_{J_0}^{-1} + p_{J_\nu}^{-1}) \right) . \tag{18}$$

A formal proof is provided in the appendix. Using the notation

$$\tilde{S}_I^2 := (n_{J_0} + n_{J_\nu}) \mathrm{var}(x)^{J_0 \cup J_\nu} + \sum_{i=1}^{\nu-1} n_{J_i} \mathrm{var}(x)^{J_i}$$

for the denominator of b_1 we define the test statistic

$$t^I = \frac{\tilde{S}_I \sqrt{n_{J_0} + n_{J_\nu}}}{\sqrt{n_{J_0} n_{J_\nu}}} b_1^I . \tag{19}$$

Note that $\tilde{S}_I^2 / \sigma^2$ is asymptotically χ^2-distributed with $n - \nu$ degrees of freedom, and therefore t^I will be approximately t-distributed with $n - \nu$ degrees of freedom.

Of course what we are actually interested in is t^{I^*}, the test statistic corresponding to b_1^*, the optimal solution of (9). By conditioning we immediately can see that the distribution of t^{I^*} will be a mixture of t-distributions. Like in the case of isotonic regression one could try to determine the level properties for each I, but we will not pursue this direction here. For testing one might consider the quantile of a t-distribution with $n - 1$ degrees of freedom as upper bound, and with $n - l$ degrees of freedom as lower bound for the true critical value.

We want to emphasize that in the given form the procedure roughly resembles a t-test between the X-values corresponding to the largest level and the smallest level of Y; as a

consequence of Theorem 3.1 we can be almost sure for large n under H_0 that $J_0 = \{0\}$ and $J_\nu = \{l\}$, even when we cannot be certain what I^* will exactly look like. We want to illustrate this behavior with a simple example, where we use step functions to simulate models under the alternative hypothesis H_1.

Example 3.1 *Let X be uniformly distributed on $[0,1]$ and Y categorical with five levels, i.e. $l = 4$. We will generate test examples where the cumulative probabilities $q_i(x)$ are non-increasing step functions with jumps at $x = 0.1, 0.2, 0.8, 0.9$. In other words we partition $[0,1]$ into five intervals*

$$M_0 = [0, 0.1], M_1 = (0.1, 0.2], M_2 = (0.2, 0.8], M_3 = (0.8, 0.9], M_4 = (0.9, 1]\,.$$

and define on each of those intervals Y a different discrete distribution, such that there is an order with respect to X. Note that non-increasing $q_i(x)$ lead to positive correlation between X and Y.

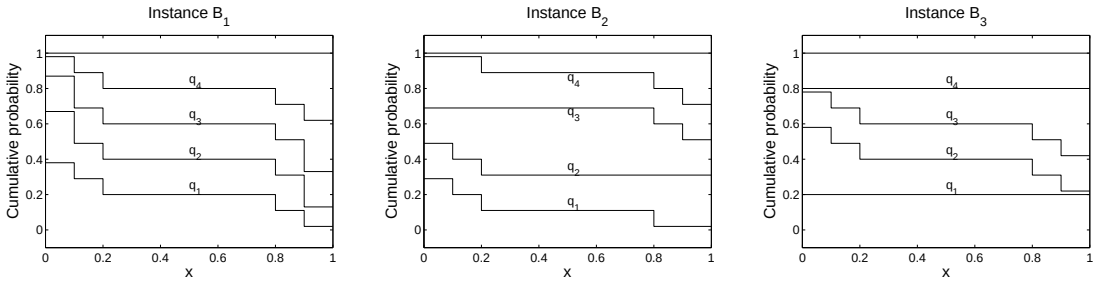
The joint distribution of X and Y can be specified by a 5×5 matrix P, where $P(i,j)$ is the probability that $Y = i - 1$ given that $X \in M_{j-1}$. We will consider matrices of the form $P = \frac{1}{5}E + \delta B$ where E is the all one matrix, B is some fixed matrix specifying the effect of a particular alternative and δ is a scalar parameter. For $\delta = 0$ we just obtain the null hypothesis H_0, where Y is uniformly discrete. Letting δ grow we can study the power of (19) which we compare with simple linear regression (keeping Y fixed) and with logistic regression using the proportional odds assumption (McCullagh, 1980).

We will consider three instances:

$$B_1 = \begin{pmatrix} 2 & 1 & 0 & -1 & -2 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 0 & -1 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ -2 & -2 & -1 & 0 & 1 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ -2 & -1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 1 shows the cumulative probabilities $q_i(x)$ for $\delta = 0.09$.

Figure 1: Cumulative probabilities $q_i(x)$ of example 3.1 for $\delta = 0.09$.



In Figure 2 we provide the corresponding power functions obtained by simulating 10000 replicates with $n = 100$ for various values of δ ranging from 0 to 0.1. Using the quantile of a t-distribution with 99 degrees of freedom as critical values leads to reasonable asymptotic behavior: At a significance level of $\alpha = 0.05$ and under the null hypothesis 5.16% of all test statistics lie above the threshold. For B_1 the new procedure is only

slightly less powerful than simple regression, for B_2 it is much more powerful, whereas for B_3 it cannot compete at all. In all three cases the multinomial logistic regression turned out to be strictly less powerful than simple regression.

The results of Example 3.1 are easy to understand when we consider that the new procedure is very similar to applying a two sample t-test between the data x_j with corresponding $y_j = 0$ and $y_j = l$ respectively. Basically all the information contained in the middle three rows of B_i does not enter into the decision process. This does not have a strong effect in case of instance B_1, though some power is lost by neglecting the information of the second and the fourth row. On the other hand in case of B_2 when only considering the three intermediate levels one observes an inverse trend, and discarding this information actually gives larger power. The situation for instance B_3 is the other extreme, where the dependence between X and Y is entirely due to the three intermediate levels. It is actually quite interesting to observe that the procedure has some very small power to detect large effects and is therefore not entirely equivalent with a two-sample t-test. But we conclude that a test based on t^{I^*} looses too much information to be of general interest, except in fairly particular situations.

3.2 Penalizing

One possibility of improvement consists in adding penalties for values of $c_2, \ldots, c_{l-1}$ getting too small, and for c_1 and c_l getting too large. To retain a quadratic optimization problem one might consider penalties of the form $K_i(c_i - \gamma_i)^2$, $i \in \{1, \ldots, l\}$. To decide how to choose the K_i we note that under H_0 the residual sum of squares (11) for the unpenalized model converges asymptotically towards $R^* = (n_0 n_l)(n_0 + n_l)^{-1}$. Thus if $K_i = o(n)$ the penalties will have for large n hardly any influence, whereas if K_i grows faster than n the contribution of the penalty terms for $c_i \neq \gamma_i$ will dominate R^*, which enforces $c_i \approx \gamma_i$ for large n. So the most interesting situation arises if the K_i are actually of order n, i.e. $K_i = \lambda_i n$.

We will present here a particularly simple penalization scheme, with all γ_i set to 0:

$$\min_{b_0, b_1; \mathbf{c} \in \Delta^l} \left\{ \sum_{j=1}^{n} (g_j(\mathbf{c}) - b_0 - b_1 x_j)^2 + n \sum i = 1^l \lambda_i c_i^2 \right\} . \tag{20}$$

For the discussion of a more general penalization scheme we refer to Frommlet (2008). The crucial idea is to choose the vector $\lambda = (\lambda_1, \ldots, \lambda_l)$ in such a way, that under the null hypothesis $c_i \to c_i^{\lim}$ for some prespecified values of $c_i^{\lim}$. These asymptotic values can be interpreted as a subjective expression of how close one believes different categories to be. Choosing for example $c_1^{\lim} = \cdots = c_l^{\lim}$ amounts to some prior believe that all categories are equidistant – which relates to a simple regression model. However, depending on the actual data the values of c_i are not fixed, which distinguishes the method from simple weighted regression. The original test statistic (19) based on Torra et al. (2006) corresponds to the rather unusual prior believe that all categories except for the two extreme ones are more or less identical.

By adding the penalty terms we loose the simple structure of (9) which allowed us to give explicit solutions of b_0, b_1 and a simple recursion formula to compute $\mathbf{c}$. However,

we can still follow the approach of Lemma 3.1 to obtain critical points. In the typical situation where no merging occurs for $Y = 0$ and $Y = l$, i.e. for $I = (1, s_2, \ldots, s_{\nu-1}, l)$ we obtain

$$b_0^I = \frac{n_l}{n_0 + n_l} - b_1^I \bar{x}^{0,l} + \frac{n}{n_0 + n_l}(\lambda_1 c_1^I - \lambda_l c_l^I)$$

and

$$b_1^I = \frac{\frac{n_0 n_l}{n_0+n_l}(\bar{x}^l - \bar{x}^0) + n\sum_{i=2}^{\nu-1}\Psi_i(\bar{x}^{J_i} - \bar{x}^{J_{i-1}}) + n\Psi_1(\bar{x}^{J_1} - \bar{x}^{0,l}) + n\Psi_\nu(\bar{x}^{0,l} - \bar{x}^{J_{\nu-1}})}{(n_0 + n_l)\mathrm{var}(x)^{\{0,l\}} + \sum_{i=1}^{\nu-1} n_{J_i}\mathrm{var}(x)^{J_i}},$$

with $\Psi_i = \lambda_{s_i} c_{s_i}^I$ for $i = 1, \ldots, \nu$. Therefore, both b_0^I and b_1^I depend on $\mathbf{c^I}$ which can in general not be expressed in a simple way. In any case all Ψ_i are bounded and therefore consistency still holds: $b_1^{(n)} \overset{a.s.}{\to} 0$ for $n \to \infty$ under H_0. However, a general result like (16) in Theorem 3.1 cannot be established, because everything depends now on the choice of the penalty parameters. Still we obtain the asymptotic distribution of the scaled numerator of b_1^I. For the sake of convenient notation we write $\bar{x}^{J_0} := \bar{x}^{0,l}$ as well as $\bar{x}^{J_\nu} := \bar{x}^{0,l}$ and thus obtain

$$\frac{1}{\sqrt{n}}\frac{n_0 n_l}{n_0 + n_l}(\bar{x}^l - \bar{x}^0) + \sqrt{n}\sum_{i=1}^{\nu}\Psi_i\left(\bar{x}^{J_i} - \bar{x}^{J_{i-1}}\right) \xrightarrow{d} \mathcal{N}\left(0; \sigma^2\left(\frac{p_0 p_l}{p_0 + p_l} + \Omega^{\lim}\right)\right),$$

where

$$\Omega = \sum_{i=1}^{\nu-1}\frac{n}{n_i}(\Psi_i - \Psi_{i+1})^2 + \frac{n}{n_0 + n_l}(\Psi_\nu - \Psi_1)^2$$

is the variance contribution from the penalty terms and $\Omega^{\lim}$ is its asymptotic limit. This leads immediately for given I to the test statistic

$$t^I = \frac{\tilde{S}_I}{\sqrt{\frac{n_0 n_l}{n_0+n_l} + \Omega}} b_1^I,$$

which is a generalization of (19) taking into account the penalties and which is again approximately t-distributed with $n - \nu$ degrees of freedom.

To choose the penalties λ_i we assume that in the asymptotic limit under H_0 no merging between categories occurs, i.e. $I^{\lim} = \{1, 2, \ldots, l\}$. For given $c_1^{\lim}, \ldots, c_l^{\lim}$ the penalties λ_i are computed by solving the following system of linear equations:

$$\lambda_{k+1} c_{k+1}^{\lim} - \lambda_k c_k^{\lim} = p_k\left(\sum_{r=1}^{k} c_r^{\lim} - b_0^{\lim}\right), \qquad k \in \{1, \ldots, l-1\}, \tag{21}$$

where $b_0^{\lim} = \frac{p_l}{p_0+p_l} - \frac{1}{p_0+p_l}(\lambda_1 c_1^{\lim} - \lambda_l c_l^{\lim})$. This system has $l - 1$ equations for l penalty parameters. To fully determine the penalties we add an equation for the overall weight of the penalty

$$\lambda_1 + \cdots + \lambda_l = K.$$

According to simulations the actual choice of K has no strong influence on the results and we choose $K = l$. Using this penalty scheme we study again Example 3.1. To assess the effect of various parameters on the statistical power of the given model we study three cases:

Model 1: $c_1^{\lim} = c_2^{\lim} = 0.25$, which corresponds to simple linear regression,

Model 2: $c_1^{\lim} = 0.4, c_2^{\lim} = 0.1$, which is closer to the model without penalty,

Model 3: $c_1^{\lim} = 0.1, c_2^{\lim} = 0.4$,

and according to symmetry $c_4^{\lim} = c_1^{\lim}$ and $c_3^{\lim} = c_2^{\lim}$.

Table 1 provides the power of the various models at a significance level $\alpha = 0.05$ and for $\delta = 0.1$, the largest effect parameter in our simulations. We used the same random instances generated in the previous simulations. In all simulations the asymptotic approximation worked quite well and one can rely upon the test statistic being t-distributed under the null hypothesis. For sample sizes much smaller than $n = 100$ one might prefer to use permutation tests or bootstrap sampling rather than threshold levels based on asymptotics.

Model	B_1	B_2	B_3
No penalty	0.88	0.93	0.06
Model 1	0.90	0.62	0.33
Model 2	0.90	0.86	0.11
Model 3	0.87	0.39	0.54
Simple Reg.	0.91	0.64	0.34
Logistic Reg.	0.84	0.43	0.24

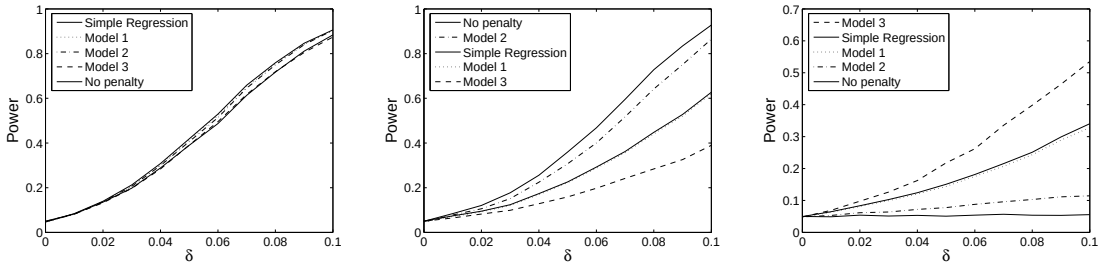
Table 1: Power of various models for three instances of Example 3.1 at effect size $\delta = 0.1$

As expected, for the first models with $c_i^{\lim} = 0.25$ the results are fairly close to simple regression for all three instances B_1, B_2, and B_3. Specifically the advantage of the new approach for instance B_2 is lost by introducing this kind of penalties. For model 2, where $c_1^{\lim} = 0.4$ and $c_2^{\lim} = 0.1$, the general behavior is similar to the model without penalties, though a little bit less extreme – the model without penalties corresponds to $c_1^{\lim} = 0.5$ and $c_2^{\lim} = 0$. Compared to simple regression model 2 performs slightly worse for instance B_1, it performs much better for instance B_2 and it has smaller power for instance B_3. Finally model 3 shows exactly the opposite behavior, it outperforms simple regression for instance B_3, but it does not work well for instance B_2. Figure 2 illustrates the general behavior of the power for the various models.

3.3 Gene expression analysis

We will use now our approach to analyze publicly available gene expression data from a study on prostate cancer Singh et al. (2002). Based on microarray experiments the expression levels of 12600 genes are compared with the Gleason score, which evaluates how effectively cancer cells are able to structure themselves. A Gleason score between 2 and 4 means well differentiated cells, a score between 5 and 6 describes intermediate differentiation, a score of 7 is intermediate to badly differentiated, and a Score between 8 and 10 means badly or undifferentiated tumors. The study included 52 patients, 26 with score 6 and 20 with score 7. The 6 patients with score larger than 7 were merged to form the top level.

Figure 2: Results for the three instances of example 3.1: The power of ordinal regression procedures with penalties (Model 1, 2, and 3) is compared with simple regression and with the ordinal regression procedure without penalties. The order of performance is reflected by the order in the legends.



In the original analysis of Singh et al. (2002) the Gleason score was treated as a metric variable. It is indicated that significant association was statistically determined by permutation tests with respect to Pearson correlation coefficients. However, 29 genes are reported with p-values smaller than 0.001, and it would appear that these are not p-values based on permutation tests but ordinary p-values. Furthermore these 29 genes are presented in such a way, that 8 of them are not at all identifiable in the original list of 26000 genes, and only 18 can be determined with certainty. For several other genes there are two possible interpretations of the given specification. We base our comparison with the original results on the 18 clearly identifiable genes and include if possible also the ambiguous ones in our discussion.

In Chu et al. (2005) it was pointed out that the Gleason score is clearly an ordinal variable, and thus the problem of predicting the score from gene expression data is a typical problem of ordinal regression. Chu et al. (2005) develop an ordinal regression procedure based on Gaussian processes. They assume there is an unobservable latent function $f(X_i) \in \mathbb{R}$ associated with the gene expression labels X_i, where the categories of the ordinal scale correspond to intervals on the real line, and the specific value of Y_i is determined by $f(X_i)$. They use a Bayesian approach for prediction of categories, and based on leave-one-out cross validation they determine a set of 21 genes to predict the Gleason score.

For the selected genes of Chu et al. (2005) the serial numbers of the original data are given, which makes a comparison of results much easier. It is remarkable that there is not a single gene which was selected both by Singh et al. (2002) and Chu et al. (2005), which puts the results of both manuscripts somewhat into perspective. Clearly the focus of Chu et al. (2005) was on prediction, whereas Singh et al. (2002) was applying a multiple testing approach, but the discrepancy between their results is quite disturbing.

We will reanalyze the data applying simple linear regression as well as our ordinal regression procedure without penalties, and with the following penalties:

Model 1: $\lambda_1 = 1.3734, \lambda_2 = 0.6266, \quad \Rightarrow \quad c_1^{\lim} = 0.3, c_2^{\lim} = 0.7,$

Model 2: $\lambda_1 = 0.9260, \lambda_2 = 1.0740, \quad \Rightarrow \quad c_1^{\lim} = 0.5, c_2^{\lim} = 0.5,$

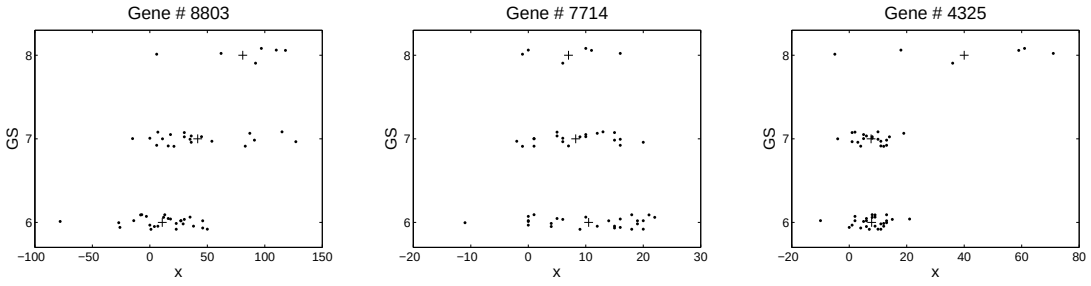
Model 3: $\lambda_1 = 0.4787, \lambda_2 = 1.5213, \quad \Rightarrow \quad c_1^{\lim} = 0.7, c_2^{\lim} = 0.3.$

Index	No penalty	Model 1	Model 2	Model 3	Lin. Reg.	Chu #	Singh
583	9.7e-05	7.3e-05⋆	3.2e-05 ⋆⋆	3.9e-05 ⋆⋆	4.3e-06 ⋆⋆	1	
8803	1.4e-04	1.0e-04	5.3e-05 ⋆⋆	7.2e-05 ⋆⋆	1.1e-05 ⋆⋆		
11421	4.8e-04	3.4e-04	2.4e-04	3.6e-04	1.2e-04		
12022	1.2e-04	1.0e-04	1.5e-04	4.2e-04	5.3e-05		SPARC
12092	2.0e-04	1.4e-04	7.6e-05 ⋆⋆	1.0e-04 ⋆	2.0e-05 ⋆		
5837		3.8e-04	1.7e-04	1.8e-04	7.2e-05	13	
1667	1.2e-04	1.1e-04	2.3e-04		9.5e-05		Collagen
6599	6.1e-05	9.1e-05	3.6e-04		1.4e-04		
9335	3.1e-04	2.6e-04	3.7e-04		2.2e-04		
10787	4.3e-06 ⋆⋆	2.8e-05 ⋆⋆	3.8e-04		6.5e-05	16	
6118			2.8e-04	1.1e-04 ⋆	1.2e-04	4	
8795		4.3e-04	4.1e-04		2.6e-04		
4325	4.8e-06 ⋆⋆	4.0e-05 ⋆⋆			1.2e-04	12	
11355				2.9e-04	4.4e-04		
7750				3.4e-04			
1666	2.7e-04	3.0e-04					Collagen
6174	2.0e-04	3.9e-04					
6658	3.6e-04	4.7e-04					
7901	1.9e-04	2.6e-04					Follastin
9277	3.6e-04	4.5e-04					
10458	3.4e-04	4.8e-04					
4998	2.9e-04						
7954	2.1e-04						

Table 2: Results from micro array data analysis. The first column gives the serial numbers of the original data, bold print indicates genes that were significant under permutation test for at least one model. The next five columns give unadjusted p-values for ordinal regression (based on asymptotic approximation) as well as linear regression; only p-values with $p \leq 0.0005$ are listed. Genes are ordered according to which models had $p \leq 0.0005$. $\star\star$ indicates significance for permutation tests at level $\alpha = 0.05$, $\star$ at level $\alpha = 0.1$. The last two columns specify genes which were also detected by Chu et al. (2005) and Singh et al. (2002), respectively. The column 'Chu #' gives the serial number in Table 6 of Chu et al. (2005), the column 'Singh' gives the annotation provided in Figure 1 of Singh et al. (2002).

Table 2 provides the results of our analysis. Unadjusted p-values based on asymptotic theory are provided for all genes with $p \leq 0.0005$. This threshold value was chosen in such a way that comparison with the results of Singh et al. (2002) and Chu et al. (2005) becomes possible. For a proper treatment of multiple testing permutation tests were performed. Based on 10000 random permutations of the Gleason score variable test statistics for the Five models were computed and the maximum over the 12600 genes taken. Threshold values were than obtained for each of the 5 models based on the upper tail quantiles controlling at $\alpha = 0.05$, and $\alpha = 0.1$. Test statistics exceeding those threshold values are indicated with $\star\star$ and $\star$ respectively in Table 1. Based on permutation tests we would recommend to consider only five ($\alpha = 0.05$) or six ($\alpha = 0.1$) genes as

Figure 3: Gleason Score data for three genes: Gene expression levels on x-axis vs. Gleason score (GS) on y-axis. For better visibility of the data random noise was added to GS. The symbol + gives the mean values of x within groups.



(a) (b) (c)

significant. Interestingly four of them were also detected by Chu et al. (2005), though none of them by Singh et al. (2002).

Concerning the original results reported by Singh et al. (2002) only three of the 29 reported genes show up in Table 2, one of them twice because the description 'Collagen type I, alpha-2' fits both to gene #1666 and #1667. Taking into account that the statistical analysis described in Singh et al. (2002) is fairly similar to the linear regression approach it is difficult to understand how the results presented in their Figure 1 were actually obtained. From the reported genes which were identifiable in particular #2741 (HSU66684), #4756 (STE20-like kinase), #8935 (acetolactate synthase) and #9153 (cyclin H) showed no association with the Gleason score at all. It is not clear if the particular results presented in Figure 1 of Singh et al. (2002) are actually based on the data published at

`http://www.broadinstitute.org/cgi-bin/cancer/`
`publications/pub_paper.cgi?mode=view&paper_id=75`

It is certain however, that the work of Chu et al. (2005) is based on the very same data. In Figure 3(a) the measurements for gene #8803 are shown, one of the two genes which we consider as important that was not reported in Chu et al. (2005). There appears to be a rather strong positive trend, and the graph for gene #12092 looks very similar. One reason why those two genes might have been neglected by Chu et al. (2005) is that they take a model based approach, and as usual in micro array data there is a strong amount of multicollinearity. Specifically both genes #8803 and #12092 are negatively correlated with gene #583, which has the strongest association with the Gleason score. In this particular setting it would appear to be rather desirable to learn about all genes which are related with the Gleason score. Missing two important genes would speak against the algorithm from Chu et al. (2005).

On the other hand Figure 3(b) presents data for the second of the 21 genes reported by Chu et al. (2005). From visual inspection it is difficult to argue that there would be any influence of gene #7714 on the Gleason score at all. Among the 21 reported genes there are 5 which are entirely unrelated with GS (these are gene #7714, #9264, #7049, #9878 and #11233). All other reported genes had at least in one of our considered models unadjusted p-values smaller than 0.05, but except for those listed in Table 2 no reported gene had p-values smaller than 0.0001.

We mentioned already that the expression levels of the first three genes (#583, #8803, #12092) selected under permutation tests were strongly correlated. Their data suggests a trend, where larger expression levels of gene #583 (as well as smaller expression levels of #8803 and #12092) are associated with less cell differentiation. This is best captured by Model 2 and the linear regression model. It is striking that under permutation tests none of the three genes was detected by the model without penalties. On the other hand genes #4325 and #10787, whose expression levels were also strongly correlated, were only detected by Model 1 and the model without penalties. Figure 3(c) provides the explanation why this is the case. Expression levels for GS = 6 and GS = 7 appear practically identical, but 4 of the 6 individuals with extremely bad cell differentiation had particularly large expression levels of gene #4325 and #10787. Both of these genes were also found by Chu et al. (2005), but based on our analysis it becomes quite clear that for those two genes the influence on GS is qualitatively much different compared with the previous three genes.

4 Conclusion

In this article we have analyzed the statistical properties of an approach to ordinal regression introduced by Torra et al. (2006), where ordinal variables are mapped into the interval $[0, 1]$. The authors claim that fixing these mappings a priori would lead to a bias in the resulting models, and therefore they suggest to estimate them by least squares optimization. We have shown in Section 2 that in the case of an ordinal explicatory variable the approach actually works, though it coincides with the well known procedure of isotonic regression.

In Section 3 we have seen that if the dependent variable is ordinal then the least squares estimation of the maps f in (8) does not at all lead to an "unbiased" situation. Asymptotically the procedure is equivalent to choosing the map $f(i) = p_l(p_0+p_l)^{-1}$ for $0 < i < l$, a choice which in most situations leads to a significant loss of power compared e.g. to simple regression, which is equivalent to the choice $f(i) = i/l$. Torra et al. (2006) introduced their new approach to ordinal regression actually when both dependent and explicatory variables are ordinal. In Frommlet (2008) the shortcomings in this specific situation are discussed.

In summary the procedure of Torra et al. (2006) cannot really be recommended for dependent ordinal variables. An improvement based on penalties was suggested and evaluated in a small simulation study. In principle the choice of such penalties based on asymptotic considerations is not much different from choosing a-priori a specific map f. Different penalties are suitable for different alternative hypothesis, which can be described in terms of the relative proximity of adjacent ordinal scales. This was made transparent by Example 3.1 as well as by the analysis of micro array data.

Specifically in the gene expression data analysis the type of model for which a gene is significant has some direct interpretation: Genes selected only by model 1 are significantly different expressed only for Gleason score GS > 7, genes selected by model 2 tend to have a trend, etc. In general our analysis suggests that the number of important genes reported both in Singh et al. (2002) as well as in Chu et al. (2005) are rather optimistic. In particular the findings of the original study Singh et al. (2002) can hardly be confirmed.

Based on the data we would suggest the following slightly more modest conclusion: Three genes (#583, #8803, #12092) have a general influence on cell differentiation. Gene #583 (RET finger protein-like 3) is known to be related with oncogenic activity Chu et al. (2005) and its role in the regulation of growth or differentiation of different cell types has been discussed for a long time Tezel et al. (1999). Gene #8803 is known to express insulin-like growth factor-binding protein-3, and gene #12092 expresses apolipoprotein E. For two genes (#4325, #10787), both known to be related with the development of drug resistance, expression in patients with GS > 7 was significantly enhanced. Finally for gene #6118, whose functional role is not yet known, expression was somewhat larger for better differentiated cells. Although one might expect that more genes have an influence on differentiation of tumor cells, evidence from the data appears to be not particularly strong, which is largely due to the rather small sample size. We would therefore suggest that all other genes reported both in Singh et al. (2002) and Chu et al. (2005) should be considered only with great precaution.

In this article we have only considered the least squares approach suggested by Torra et al. (2006). As an alternative to the presented penalizing scheme one might like to consider a likelihood ratio test. To this end it is necessary to specify a probabilistic model. The simplest scenario is to require that X conditional on each level of Y is normal distributed with fixed variance σ^2. It is then easy to show that under these assumptions the ML-estimate and the least squares estimate with regard to (8) coincide. However, to construct a likelihood ratio test we also need the likelihood function under the null hypothesis, but for $b_1 \to 0$ the likelihood flattens out and converges towards 0. We are confronted with the situation that under the null hypothesis the parameters $\mathbf{c}$ are not identifiable and have to be treated as nuisance parameters. It might be an interesting topic for further research to address this particular problem of inference where some nuisance parameters are not identified under the null hypothesis.

Acknowledgment

I would like to thank Małgorzata Bogdan for reading the original manuscript and giving many helpful remarks, as well as an anonymous referee for constructive suggestions.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions; The Theory and Application of Isotonic Regression*. New York: Wiley.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. I. *Biometrika*, *46*, 36-48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives. II. *Biometrika*, *46*, 328-335.

Best, M. J., and Chakravarti, N. (1990). Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, *47*, 425-439.

Bomze, I. (1998). On standard quadratic optimization problems. *Journal of Global Optimization*, *13*, 369-387.

Chu, W., Ghahramani, Z., Falciani, F., and Wild, D. L. (2005). Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, *21*, 3385-3393.

Frommlet, F. (2008). *Critical remarks on a novel approach to ordinal regression without latent variables* (Tech. Rep. No. 208-06). ISDS.

Liu, I., and Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Sociedad de Estadistica e Investigacion Operativa*, *14*, 1-73.

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

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*, 203-209.

Tezel, G., Nagasaka, T., Iwahashi, N., N., N. A., Iwashita, T., Sakata, K., et al. (1999). *Different nuclear/cytoplasmic distributions of RET finger protein in different cell types*, *49*, 881-886.

Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., and Ng, M. (2006). Regression for ordinal variables without underlying continuous variables. *Information Sciences*, *176*, 465-474.

Appendix (Proof of (18) in Section 3.1)

For the sake of notational convenience we want to prove

$$\sqrt{n}(\bar{x}^{J_\nu} - \bar{x}^{J_0}) \xrightarrow{d} \mathcal{N}(0; \sigma^2(p_{J_0}^{-1} + p_{J_\nu}^{-1}))$$

only for the situation where $J_0 = \{0\}$ and $J_\nu = \{l\}$. The generalization should be obvious.

Lemma 4.1 *Let X_i and Y_i be i.i.d. sequences of independent random variables, X_i with finite mean μ, finite variance $\sigma^2 > 0$ and Y_i discrete with probabilities $(p_0, p_1, \ldots, p_l)$. We then have $\sqrt{n}(\bar{X}^l - \bar{X}^0) \xrightarrow{d} \mathcal{N}(0; \sigma^2(p_0^{-1} + p_l^{-1}))$.*

Proof: We can write

$$\sqrt{n}\left(\frac{1}{n_l}\sum_{i:Y_i=l} X_i - \frac{1}{n_0}\sum_{i:Y_i=0} X_i\right) = \sqrt{n}\left(\frac{1}{E(n_l)}\sum_{j=1}^{E(n_l)} X_{i_j} - \frac{1}{E(n_0)}\sum_{k=1}^{E(n_0)} X_{i_k}\right)$$
$$+\sqrt{n}\left(\frac{1}{n_l}\sum_{i:Y_i=l} X_i - \frac{1}{E(n_l)}\sum_{j=1}^{E(n_l)} X_{i_j}\right) + \sqrt{n}\left(\frac{1}{E(n_0)}\sum_{k=1}^{E(n_0)} X_{i_k} - \frac{1}{n_0}\sum_{i:Y_i=0} X_i\right),$$

where we still have to specify how exactly we choose the subindices i_j and i_k. If we request that $i_j \neq i_k$ for all j and k then $\sum_{j=1}^{E(n_l)} X_{i_j}$ and $\sum_{k=1}^{E(n_0)} X_{i_k}$ are independent and

we immediately conclude from the central limit theorem that the first term on the right hand side converges as desired:

$$\sqrt{n}\left(\frac{1}{E(n_l)}\sum_{j=1}^{E(n_l)} X_{i_j} - \frac{1}{E(n_0)}\sum_{k=1}^{E(n_0)} X_{i_k}\right) \xrightarrow{d} \mathcal{N}(0;\sigma^2(p_0^{-1}+p_l^{-1}))\,.$$

Due to Cramer's (or Slutsky's) theorem it remains to show that the other two terms on the right hand side converge in probability towards 0. It is evident that the second term can be rewritten as

$$T_2 := \sqrt{n}\left(\frac{1}{n_l}\sum_{i:Y_i=l}(X_i-\mu) - \frac{1}{E(n_l)}\sum_{j=1}^{E(n_l)}(X_{i_j}-\mu)\right)$$

and similar for the third term, so without loss of generality we can assume that $\mu = 0$. Concerning the choices of the subindices we have basically four situations:

- $n_0 \geq E(n_0)$, $n_l \geq E(n_l)$: The choice is obvious, i_j has to be a subset of $\{i : Y_i = l\}$ and i_k has to be a subset of $\{i : Y_i = 0\}$.
- $n_0 < E(n_0)$, $n_l < E(n_l)$: For i_j choose first all indices $\{i : Y_i = l\}$ and then choose arbitrary indices from $\{i : Y_i \neq 0\}$. For i_k choose again first all indices $\{i : Y_i = 0\}$ and then arbitrary from the remaining unused indices.
- $n_0 < E(n_0)$, $n_l \geq E(n_l)$: First choose the i_j as a subset of $\{i : Y_i = l\}$, then choose for i_k first all $i : Y_i = 0$, and then fill up arbitrarily.
- $n_0 \geq E(n_0)$, $n_l < E(n_l)$: Like above, only n_0 and n_l exchange roles.

Having made those choices we can now look at the behavior of the second term. Let us first look at the situation where $n_l \leq E(n_l)$. We can then rewrite

$$\begin{aligned} T_2 &= \sqrt{n}\left(\frac{1}{n_l} - \frac{1}{E(n_l)}\right)\sum_{i=1}^{n_l} X_i - \sqrt{n}\frac{1}{E(n_l)}\sum_{i=n_l+1}^{E(n_l)} X_i \\ &= \frac{np_l - n_l}{p_l\sqrt{n}}\frac{1}{n_l}\sum_{i=1}^{n_l} X_i - \frac{1}{p_l\sqrt{n}}\sum_{i=n_l+1}^{E(n_l)} X_i\,. \end{aligned}$$

Due to the central limit theorem the term $\frac{n_l - np_l}{p_l\sqrt{n}}$ converges in distribution towards a normally distributed random variable and due to the law of large numbers for randomly indexed sequences $\frac{1}{n_l}\sum_{i=1}^{n_l} X_i$ converges to 0 in probability (we need that n_l converges almost surely towards infinity, which is clear). Now according to Cramer's theorem the product of both converges in distribution towards 0, which furthermore implies convergence in probability towards 0.

For the second term we first apply Chebyshev inequality with respect to $\sum_{i=n_l+1}^{np_l} X_i$ for given n_l:

$$\Pr\left(\left|\frac{1}{p_l\sqrt{n}}\sum_{i=n_l+1}^{np_l} X_i\right| > \epsilon \middle| n_l\right) \leq \frac{(np_l - n_l)\sigma^2}{np_l^2\epsilon^2}\,.$$

Taking expectation with respect to n_l and applying Lyapunov's inequality leads to

$$\Pr\left(\left|\frac{1}{p_l\sqrt{n}}\sum_{i=n_l+1}^{np_l} X_i\right| > \epsilon\right) \leq \frac{\sigma^2}{np_l^2\epsilon^2}\left(E(n_l - np_l)^2\right)^{1/2} = O(n^{-1/2})\,.$$

The case $n_l > E(n_l)$ can be dealt similarly.

Author's address:

Florian Frommlet
Department of Statistics and Decision Support Systems
University Vienna
Brünner Straße 72
1210 Wien

E-Mail: florian.frommlet@univie.ac.at