Latent Class Regression Model in IRLS Approach

S. LIPOVETSKY AND W. M. CONKLIN
GfK Custom Research Inc.
8401 Golden Valley Rd., Minneapolis, MN 55427, U.S.A.
<lipovetsky><mconklin>@GfKcustomresearch.com

(Received March 2004; revised and accepted January 2005)

Abstract—We consider latent class regressions for the simultaneous construction of several regression models by the data clusters. Maximum likelihood objective of observations belonging to at least one data segment is developed. Solution is reduced to the iteratively reweighted least squares (IRLS) procedure that defines coefficients of all models and the characteristics of fitting. Together with the regression models, this approach yields probabilities of each observation belonging to each of the classes. This technique can also be used for finding parameters of mixed distributions. The suggested approach enriches results of the regression modeling and clustering in practical applications.

Keywords—Regression, Latent classes, Iteratively reweighted least squares.

I. INTRODUCTION

We consider simultaneous constructing of several regressions by subsets of a given data set. Such an approach corresponds to so-called latent class models known in various statistical applications [1–3]. Latent class techniques are applied in factor and scaling analyses [4–8], structural equations and latent structure analysis [9–11], and social modeling [12–14], where the term latent has a meaning of an unobserved variable. The concept of latent class is widely used in the marketing research field where the term latent denotes the segments or subsets of data [15–20]. In this paper, we consider latent class regression modeling that has been studied in various works [21–26]. We suggest a new formulation of the maximum likelihood (ML) objective of the probability that each observation belongs to at least one class. This formulation produces a solution expressed as an iteratively reweighted least squares (IRLS) procedure for solving nonlinear statistical problems [27–33]. We apply this technique to latent class regression models and to the estimation of the parameters of mixed distributions. In addition to regressions, we obtain the probabilities of belonging to each class for each observation.

The paper is arranged as follows. Section 2 describes the regular maximum likelihood for linear regression and for parameters of normal distribution. Section 3 introduces maximum likelihood for probability that each observation belongs to at least one of several possible classes. Section 4 presents numerical examples for latent class modeling, and Section 5 summarizes.
2. MAXIMUM LIKELIHOOD FOR LINEAR REGRESSION

Let us at first describe a regular maximum likelihood approach known in regression analysis for the linear model

\[ y_i = a_0 + a_1 x_{i1} + \cdots + a_n x_{in} + \varepsilon_i, \]  

where observations \((i = 1, 2, \ldots, N)\) by the dependent variable \(y_i\) are fitted with the linear combination of the observed predictor variables \(x_{i1}, x_{i2}, \ldots, x_{in}\) \((n - \text{number of the independent variables})\), and \(\varepsilon_i\) denotes deviations from the linear form. Suppose the deviations \(\varepsilon_i\) correspond to a random noise defined by a normal distribution with the probability density function (pdf)

\[ f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right), \]  

where \(\sigma\) is the standard deviation. Probability of the random event corresponded to an \(i^{th}\) observation is

\[ p_i = \Delta\varepsilon f(\varepsilon_i) = \frac{\Delta\varepsilon}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right), \]  

where \(\Delta\varepsilon\) denotes a constant infinitesimal interval around the random term. Likelihood of the event that all independent observations occurred is the product of all probabilities (3) that is the objective of maximum likelihood

\[ \text{ML} = \prod_{i=1}^{N} p_i = \prod_{i=1}^{N} \frac{\Delta\varepsilon}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) = \left(\frac{\Delta\varepsilon}{\sqrt{2\pi\sigma}}\right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \varepsilon_i^2\right). \]  

Logarithm of the ML objective equals

\[ \ln \text{ML} = N \ln \frac{\Delta\varepsilon}{\sqrt{2\pi\sigma}} - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \varepsilon_i^2 \rightarrow \max. \]  

The first item at the right-hand side (4) is a constant, the other items contain parameters \(\sigma\) and coefficients of regression in the error term that can be expressed from (1) as

\[ \varepsilon_i = y_i - a_0 x_{i0} - a_1 x_{i1} - \cdots - a_n x_{in}, \]  

where we include an identical dummy variable \(x_0\) with the intercept \(a_0\). The first order conditions for maximizing (5) by \(a_k\) \((k = 0, 1, 2, \ldots, n)\) parameters are

\[ \frac{\partial \ln \text{ML}}{\partial a_k} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \varepsilon_i x_{ik} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - a_0 x_{i0} - a_1 x_{i1} - \cdots - a_n x_{in}) x_{ik} = 0, \]  

and by the standard deviation is

\[ \frac{\partial \ln \text{ML}}{\partial \sigma} = - \frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} \varepsilon_i^2 = 0. \]  

The right-hand side of equations (7) can be represented as a normal system

\[ a_0 \sum_{i=1}^{N} x_{i0} x_{ik} + a_1 \sum_{i=1}^{N} x_{i1} x_{ik} + \cdots + a_n \sum_{i=1}^{N} x_{in} x_{ik} = \sum_{i=1}^{N} y_i x_{ik}, \quad k = 0, 1, \ldots, n. \]  

This system in matrix form is

\[ X'X a = X' y, \]
where $X$ is the design matrix of the order $N$ by $1+n$ consisting of observations by independent $x_{i0}, x_{i1}, x_{i2}, \ldots, x_{in}$ variables, $X'$ is the transposed matrix, and $y$ is the vector-column of $N$ observations by the dependent variable. Thus, $X'X$ and $X'y$ are a matrix and a vector of cross-products of all $x$s among themselves and with $y$, respectively. Solution of system (10) can be presented as

$$a = (X'X)^{-1}X'y,$$

that is a well-known result for the coefficients of the linear regression (1). This solution can be obtained by the minimum least squares (LS) objective

$$\text{LS} = \sum_{i=1}^{N} \varepsilon_i^2 \rightarrow \text{min},$$

that corresponds to the least square (LS) objective. The objective (12) coincides with the last sum in (5) for the given standard error, when maximum likelihood (5) is reducing to the minimum least squares (12). Equation (8) yields the maximum likelihood estimate for the squared standard error

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^2,$$

that is the mean of LS residuals (12) for the estimated regression (11).

In the simplest case when a linear model (1) has just the intercept, so the design matrix $X$ consists solely of the identity variable $x_0$, solutions (11) and (13) degenerate to the simple known estimates for the parameters of the normal pdf (2)

$$a_0 = m = \frac{1}{N} \sum_{i=1}^{N} y_i, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - m)^2,$$

which are the mean and the variance of the $y$ variable.

3. MAXIMUM LIKELIHOOD FOR LATENT CLASS REGRESSION

Now we describe the maximum likelihood for the latent class modeling when instead of one model (1), we want to construct several models

$$y_i = a_{0j} + a_{1j} x_{i1} + \cdots + a_{nj} x_{in} + \varepsilon_{ij}, \quad j = 1, 2, \ldots, c,$$

where the index $j$ identifies a class, and there are $c$ classes. Suppose the deviations $\varepsilon_{ij}$ correspond to random noise defined by a mixture of normal distributions. Each distribution has pdf

$$f(\varepsilon_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left( -\frac{\varepsilon_{ij}^2}{2\sigma_j^2} \right),$$

where $\sigma_j$ is standard deviation for the $j^{th}$ pdf. We assume that probabilities (3) corresponded to densities (16) are independent. Then the probability that each $i^{th}$ observation belongs at least to one class can be defined as

$$p_i = 1 - \prod_{j=1}^{c} (1 - \Delta \varepsilon f(\varepsilon_{ij})) \approx \Delta \varepsilon \sum_{j=1}^{c} f(\varepsilon_{ij}),$$

where similarly to (3), $\Delta \varepsilon$ denotes a constant infinitesimal interval around the random terms. Approximation (17) can be considered as exact relation because of the infinitely small values
of $\Delta \varepsilon$. The maximum likelihood objective (4) constructed by the probabilities (17) across all observations is

$$\text{ML} = \prod_{i=1}^{N} p_i = \prod_{i=1}^{N} \left( \Delta \varepsilon \sum_{j=1}^{c} f(\varepsilon_{ij}) \right).$$  \hfill (18)

The logarithm of ML equals

$$\ln \text{ML} = N \ln \Delta \varepsilon + \sum_{i=1}^{N} \ln \left( \sum_{j=1}^{c} f(\varepsilon_{ij}) \right) \to \max.$$

Similarly to (5), the first item at the right-hand side (19) is a constant, but now the other items with latent classes pdf (16) contain parameters $\sigma_j$ and coefficients of latent class regressions in the error terms. The errors can be expressed from (15) as follows:

$$\varepsilon_{ij} = y_i - a_0 x_{i0} - a_1 x_{i1} - \cdots - a_n x_{in}, \quad j = 1, 2, \ldots, c,$$  \hfill (20)

where $x_0$ also means an identical dummy used with the intercept $a_0$.

Maximizing objective (19) by unknown parameters of regressions yields a system of the first-order partial derivatives

$$U(a_k^j) = \frac{\partial \ln \text{ML}}{\partial a_k^j} = \sum_{i=1}^{N} \frac{1}{\sum_{q=1}^{c} f(\varepsilon_{iq})} \frac{\partial f(\varepsilon_{ij})}{\partial a_k^j} = \sum_{i=1}^{N} \sum_{q=1}^{c} f(\varepsilon_{iq}) \frac{\varepsilon_{ij} \sigma_j^2 x_{ik}}{\sum_{q=1}^{c} f(\varepsilon_{iq})} = 0,$$  \hfill (21)

where the probability density functions are defined in (16). And maximizing objective (19) by unknown standard deviations yields the equations

$$U(\sigma_j) = \frac{\partial \ln \text{ML}}{\partial \sigma_j} = \sum_{i=1}^{N} \frac{1}{\sum_{q=1}^{c} f(\varepsilon_{iq})} \frac{\partial f(\varepsilon_{ij})}{\partial \sigma_j} = \sum_{i=1}^{N} \sum_{q=1}^{c} f(\varepsilon_{iq}) \frac{\varepsilon_{ij}^2 - \sigma_j^2}{\sigma_j^2} = 0.$$  \hfill (22)

Equations (21) correspond to a weighted normal system (7) or (9), and equations (22) to the weighted variance (8) or (13). However, the weights themselves are defined via distributions (16) and the deviations (20), and those in their turn depend on the estimated parameters.

Let us briefly describe how to solve the nonlinear system (21), (22) numerically by the Newton-Raphson procedure [34,35], and how to present the results in the so-called IRLS, or iteratively reweighted least squares algorithm known in solving nonlinear statistical problems. Consider a combined vector of all parameters $\gamma$ that in our case contains all $c(n + 1)$ coefficients of regressions (15) plus $c$ standard deviations of the classes in (16). Then the vector with the elements of scores (21), (22) can be approximated as

$$U = U^{(0)} + \frac{\partial U}{\partial \gamma} \left( \gamma^{(t+1)} - \gamma^{(t)} \right) = \frac{\partial \ln \text{ML}}{\partial \gamma} + \frac{\partial^2 \ln \text{ML}}{\partial \gamma \partial \gamma'} \left( \gamma^{(t+1)} - \gamma^{(t)} \right) = 0,$$  \hfill (23)

so the iteratively estimated vector of parameters can be expressed as

$$\gamma^{(t+1)} = \gamma^{(t)} - \left( \frac{\partial^2 \ln \text{ML}}{\partial \gamma \partial \gamma'} \right)^{-1} \frac{\partial \ln \text{ML}}{\partial \gamma} = \gamma^{(t)} - H^{-1} U \left( \gamma^{(t)} \right),$$  \hfill (24)

where $t$ is the iteration step. Hessian, or the matrix $H$ of the second derivatives, can be found by the derivatives of the scores (21), (22). But for statistical problems with a specific feature of
summing across observations, there is an easier way of using Fisher information matrix \( I \) defined as expectation of the outer product of the vector of the first derivatives

\[
I = -E(H) = E(UU'),
\]

where the elements of vector \( U \) are given in (21),(22). Let us construct Fisher matrix by its blocks. The block of the regressions coefficients can be obtained from (21) and presented as the matrix with elements

\[
I \left( a^j_{k1}, a^j_{k2} \right) = E \left( U \left( a^j_{k1} \right) U \left( a^j_{k2} \right) \right) = \sum_{i=1}^{N} \frac{\left( f(\varepsilon_{ij}) \right)^2}{\sum_{q=1}^{c} f(\varepsilon_{iq})} \frac{1}{\sigma_j^2 x_i x_i k1 x_i k2} \delta_{j r},
\]

where we assume that deviations (20) are independent by different regressions (when \( j \neq r \)), use approximation of \( \varepsilon_{ij}^2 = \sigma_j^2 \), and denote by \( \delta_{j r} \) the Kronecker delta function. Expression (26) defines a block-diagonal matrix where each \( j^{th} \) block of \( (n + 1) \) order can be represented as

\[
I \left( a^j \right) = \frac{1}{\sigma_j^2} X' W_j X,
\]

where a diagonal matrix of the \( N^{th} \) order for the weights for observations within each \( j^{th} \) block is defined as follows:

\[
W_j = \text{diag} \left( \frac{f(\varepsilon_{1j})}{\sum_{q=1}^{c} f(\varepsilon_{iq})}, \ldots, \frac{f(\varepsilon_{ij})}{\sum_{q=1}^{c} f(\varepsilon_{iq})}, \ldots, \frac{f(\varepsilon_{Nj})}{\sum_{q=1}^{c} f(\varepsilon_{iq})} \right).
\]

Thus, for each \( i^{th} \) observation, a weight for \( j^{th} \) class is defined by \( j^{th} \) share of pdf (16) summed total across all the classes.

Considering other blocks of Fisher information matrix corresponded to cross-products of scores for regression coefficients (21) and for standard deviations (22), we assume that deviations \( \varepsilon_{ij} \) in regressions are independent from the deviations \( \varepsilon_{ij}^2 - \sigma_j^2 \) of squared errors from their variance. So, the corresponded nondiagonal blocks of Fisher information matrix are zero matrices. Finally, the block of cross-products of the scores (22) among themselves can be presented as a matrix with the elements

\[
I(\sigma_j, \sigma_r) = E(U(\sigma_j)U(\sigma_r)) = \sum_{i=1}^{N} W_{ij}^2 \left( \frac{\varepsilon_{ij}^2 - \sigma_j^2}{\sigma_j^2} \right)^2 \delta_{j r},
\]

where \( W_{ij} \) are the weights (28). We also assumed that the deviations \( \varepsilon_{ij}^2 - \sigma_j^2 \) of squared errors from their variance are independent across the different regressions (when \( j \neq r \)). Block (29) is a diagonal matrix of a “variance of variance” meaning.

Now we can write the expressions for Newton-Raphson procedure in explicit form. For this aim let us rewrite the scores (21) in matrix form

\[
U \left( a^j \right) = \frac{1}{\sigma_j^2} X' W_j \varepsilon_j = \frac{1}{\sigma_j^2} X' W_j (y - Xa^j),
\]

where \( a^j \) is a vector of coefficients of each \( j^{th} \) regression (15), the weight diagonal matrix is defined in (28), and the vector of residuals for each regression is defined in (20). Using Newton-Raphson
scheme (24) with Fisher information matrix (25) for the iterations on the vector of coefficients for each \( j \)th regression, and taking into account the expressions (27) and (30), we get

\[
(a^j)^{(t+1)} = (a^j)^{(t)} + I^{-1} (a^j) U (a^j) = (a^j)^{(t)} + (X'W_j^2 X)^{-1} X'W_j \left( y - X (a^j)^{(t)} \right). \tag{31}
\]

Similarly, for the estimation of the residual standard deviation for each regression we use the scheme (24),(25) with \( j \)th score (22) and \( j \)th element in (29), so the iteration procedure is

\[
(\sigma_j)^{(t+1)} = (\sigma_j)^{(t)} + I^{-1}(\sigma_j) U(\sigma_j) = (\sigma_j)^{(t)} + \frac{\sum_{i=1}^{N} W_{ij} \left( \frac{\varepsilon_{ij}^2 - (\sigma_j)^{2(t)}}{\sigma_j^{2(t)}} \right)}{\sum_{i=1}^{N} W_{ij} \left( \frac{\varepsilon_{ij}^2 - (\sigma_j)^{2(t)}}{\sigma_j^{2(t)}} \right)}, \tag{32}
\]

where \( W_{ij} \) are the elements of \( j \)th matrix (28), and the residuals \( \varepsilon_{ij} \) for each regression are calculated by (20). On each step of iterations for the regressions parameters (31), we can use equations (22) solved for a simple estimation of the residual variances, that in explicit form is

\[
\sigma_j^2 = \frac{\sum_{i=1}^{N} W_{ij} \varepsilon_{ij}^2}{\sum_{i=1}^{N} W_{ij}}, \tag{33}
\]

where weights are defined in (28). Relations (33) are weighted generalizations of the regular estimates (13) for residual standard errors squared.

Regrouping terms in (31), we represent this expression as

\[
(a^j)^{(t+1)} = (X'W_j^2 X)^{-1} (X'W_j^2) \left( z^j \right)^{(t)}, \tag{34}
\]

where the so-called working variable is

\[
(z^j)^{(t)} = X (a^j)^{(t)} + W_j^{-1} \left( y - X (a^j)^{(t)} \right) = (\tilde{y}_j)^{(t)} + W_j^{-1}(\varepsilon_j)^{(t)}. \tag{35}
\]

In (35), \( \tilde{y}_j \) and \( \varepsilon_j \) are vectors of theoretical values of the dependent variable (15) and residuals (20) estimated by each latent class regression. We see that procedure (34),(35) corresponds to a weighted least squares solution for each \( j \)th latent class regression of the variable \( z^j \) (35) by the same set of the predictors as in (15). Relations (34),(35) define the IRLS, or iteratively reweighted least squares known in solving nonlinear statistical problems.

Thus, with some randomly distributed weights we construct weighted regressions (15), find residuals (20) and standard errors (33), reestimate the weights by (28), and continue the iteration process by (34),(35). IRLS procedure usually quickly converges without problems. Statistical significance for the coefficients of latent class regressions can be found very easily because diagonal elements of the matrix \((X'W_j^2 X)^{-1}\) in (34) are estimates of variance \( \sigma^2(a^j) \) for the coefficients \( a^j \). The \( t \)-statistics for the parameters of latent class regressions are \( t_k = \frac{a_k}{\sigma(a_k)} \). So-called deviance, a characteristic of fitting quality, is usually defined via the minimum value of the objective (19) as deviance = \(-2 \ln ML\). More complicated characteristics of quality, the questions of the convergence, speed, dependence on the initial values of the parameters of the IRLS procedures can be found in the literature on the nonlinear statistical modeling, for instance, in [1,30,31].

In the simple case when each linear model (15) has just the intercept, the problem of latent class regressions reduces to the estimation of the parameters in the mixture of normal distributions. Similarly to reducing of the normal LS system (11) to estimation of the mean and variance in (14),
the IRLS procedure (34),(35) can also be simplified. Using just identity vector in place of the \( x \) variables in (26), we reduce it to the total of squared elements of the weights matrix (28). The general expression (31) can be presented as follows:

\[
m_{ij}^{(t+1)} = m_{ij}^{(t)} + \frac{\sum_{i=1}^{N} W_{ij} (y_i - m_{ij}^{(t)})}{\sum_{i=1}^{N} W_{ij}^2},
\]

(36)

where weights are defined in (28), and the parameter of center for each distribution (16) is denoted as \( m_j = a^j_0 \). Together with estimation for standard deviations (33), expression (36) defines the IRLS procedure for finding parameters of the latent class distributions.

All relations (28)–(36) actually correspond to the weights \( W^{(t)} \) and deviations \( \varepsilon^{(t)} \) reestimated on each iteration step, although in many places we skip the iteration index to simplify the formulae. Together with the estimated parameters of the models and their standard errors, we obtain the sets of the weights \( W_{ij} \) (28) that identify probabilities of every \( i^{th} \) observation belonging to each \( j^{th} \) latent class. When the weights are finalized, we can use other regression techniques for estimation of net effects and adjusting the models to the presence of multicollinearity in the data [36,37].

We can generalize the regular ML (4) and the latent class ML (18) and present them as one combined multiobjective problem

\[
\text{ML} = \prod_{i=1}^{N} \prod_{j=1}^{c} \Delta \varepsilon f(\varepsilon_{ij}) \left( \prod_{i=1}^{N} \Delta \varepsilon \sum_{j=1}^{c} f(\varepsilon_{ij}) \right)^{1-\lambda},
\]

(37)

where two terms in parentheses corresponds to the objectives (4) and (18), respectively, and the regular part of ML corresponds to several latent classes (15) with probabilities defined by (16). Parameter \( \lambda \) of multiobjective function defines shares of both regular and latent class ML functions. Repeating derivation (19)–(35) with objective (37), we obtain a similar IRLS procedure, with the only difference in the weight definition that now becomes

\[
\hat{W}_{ij} = \lambda + (1 - \lambda) W_{ij},
\]

(38)

where \( W_{ij} \) are the weights (28). The parameter \( \lambda \) of multiobjective function can vary from 0 to 1. For example, \( \lambda = 1 \) reduces problem (37) to the regular ML (4) and the regular LS regression, whereas \( \lambda = 0 \) reduces (37) to the latent class ML (18) and to the IRLS procedure (34),(35). Intermediate values of \( \lambda \) correspond to the weighting between the identical weights for the regular regression and iteratively defined latent class weights. The extension of the weights (28) to the weights (38) can be useful if the iterative process encounters with converging difficulties.

4. NUMERICAL EXAMPLES

For a numerical simulation we considered the points belonging to two planes with opposite coefficients:

- (a) \( y = 3 + 6x_1 - 10x_2 \), and
- (b) \( y = -3 - 6x_1 + 10x_2 \).

Normal random noise with a zero mean and a standard deviation of 200 was added to both data sets consisting of 20 points each, that significantly distorted both subsets of data. The results of regression fitting are presented in Table 1. In the last two columns the coefficients of equations (a) and (b) are shown, together with the standard error used for the data distortion. The first two columns contain the regular regressions constructed by the data-1 (a) and data-2 (b) subsets. We
see that the coefficients of the regressions that fitted data-1 or data-2 reproduce a general structure of the exact dependencies (with the exception of the intercepts $a_0$, more sensitive to the quality of data). In data-1 the coefficients for predictors $x_1$ and $x_2$ are about 4 and $-8$ (the original values are 6 and $-10$), whereas in data-2 the coefficients are about $-5$ and 11 (the original values $-6$ and 10). Below the coefficients of regressions, are $t$-statistics for these coefficients, then the coefficients of multiple determination (0.46 and 0.60), $F$-statistic and corresponded total $p$-value for each model. By these characteristics, the models are of a good quality. Residual standard errors are about 183 for each regression—close to the original random noise.

Table 1. Regressions by simulated data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data Set 1</th>
<th>Data Set 2</th>
<th>Total Data</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Model (a)</th>
<th>Model (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>4.23</td>
<td>-5.56</td>
<td>-0.67</td>
<td>2.82</td>
<td>2.82</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>$a_2$</td>
<td>-8.18</td>
<td>11.05</td>
<td>1.43</td>
<td>-8.48</td>
<td>15.06</td>
<td>-10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>$t_0$</td>
<td>-0.78</td>
<td>-1.79</td>
<td>1.27</td>
<td>0.75</td>
<td>-8.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_1$</td>
<td>1.96</td>
<td>-2.59</td>
<td>-0.31</td>
<td>1.98</td>
<td>-2.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_2$</td>
<td>-3.56</td>
<td>4.82</td>
<td>0.62</td>
<td>-6.21</td>
<td>11.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.46</td>
<td>0.60</td>
<td>0.01</td>
<td>0.51</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>7.13</td>
<td>12.98</td>
<td>0.21</td>
<td>19.38</td>
<td>62.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>0.01</td>
<td>0.00</td>
<td>0.81</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std.</td>
<td>183.2</td>
<td>182.6</td>
<td>262.2</td>
<td>115.2</td>
<td>98.3</td>
<td>200.0</td>
<td>200.0</td>
</tr>
</tbody>
</table>

Suppose we have just the combined data by both subsets. The next column in Table 1 contains the regular regression output obtained by the combined total data. This regression demonstrates very poor results—insignificant coefficients of regression (small $t$ and big $p$-values), negligible coefficient of multiple determination (0.01), and poor quality of the model in total (small $F$- and big $p$-values). It could be expected because of mutually opposite behavior of the sources (a) and (b) for the total data. The estimate of standard error (about 262) resembles the original standard error of the noise. By the results of regression modeling we would conclude that there is no dependence on the predictors. With the tool of the latent class regression modeling, we can try to reconstruct two regressions for fitting this data. The columns in Table 1 with latent class regressions show the result of the IRLS modeling. The resurrected models for two classes are pretty good, especially for small samples of 20 observations in each subset (a) and (b). The obtained coefficients reflect the structure of the original equations (a),(b), have good $t$-values (except one intercept), good multiple coefficients of determination (0.51 and 0.77), and the high quality of the models in total (big $F$ and small total $p$-value for each regression). Estimates for standard error are less than error used for the noise. These results were obtained in the process (33)-(35) that converged after about 40 iterations.

Let us consider an application of the latent class regression in a real research project on the prescriptions of drugs for patients. The data was elicited from 240 medical doctors who in 5-point scale (from totally agree to totally disagree) estimated the following variables: $y$—likelihood to prescribe a drug, $x_1$—the drug produces a quick relief, $x_2$—it is safe to use, $x_3$—it is safe in severe cases, $x_4$—few side effects, $x_5$—long lasting relief, $x_6$—effective for kids, $x_7$—provides the most symptom relief, $x_8$—provides high patient satisfaction, $x_9$—provides high quality of life, $x_{10}$—the drug of a low cost.

Table 2 presents the results of a regular regression modeling. In the first column we see the variables, with $x_0$ corresponded to the intercept, and $R^2$ is the coefficient of multiple determination. In the second column we see the pair correlations of $y$ with $x_5$; all the correlations are positive. The next four columns contain the regular regression results—coefficients of the model, their $t$-statistics, net effects of contribution of each regressor into the coefficient of multiple de-
termination, and below is its value $R^2 = 0.26$, and the percent shares of the net effects. Because of multicollinearity, the first variable enters into the model with the negative sign, and its net effect is negative. The last four columns contain the results of stepwise regression with the main variables that yield practically the same coefficient of multiple determination, so the quality of the model is not very high. The final set of predictors consists of $x_2$, $x_4$, $x_5$, $x_6$, $x_9$, and $x_{10}$, and the biggest share is contributed by $x_9$, “provides high quality of life" attribute.

Table 2. Regular regression modeling by total data (size $N = 240$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
<th>Regular Regression</th>
<th>Stepwise Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$ Coeff.</td>
<td>$t$ Value</td>
<td>Net Effect</td>
</tr>
<tr>
<td>$x_0$</td>
<td>0.43 1.39 0.53</td>
<td>-0.03 -0.46 -0.01 -2.06</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.23</td>
<td>-0.03 -0.46 -0.01 -2.06</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.31</td>
<td>0.09 1.69 0.03 10.02</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.24</td>
<td>0.05 1.02 0.01 4.45</td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.31</td>
<td>0.16 3.05 0.05 17.35</td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.34</td>
<td>0.13 1.85 0.03 13.03</td>
<td></td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.21</td>
<td>0.06 2.01 0.02 7.26</td>
<td></td>
</tr>
<tr>
<td>$x_7$</td>
<td>0.28</td>
<td>0.04 0.54 0.01 3.06</td>
<td></td>
</tr>
<tr>
<td>$x_8$</td>
<td>0.39</td>
<td>0.24 3.52 0.08 28.92</td>
<td></td>
</tr>
<tr>
<td>$x_9$</td>
<td>0.37</td>
<td>0.15 2.21 0.04 17.17</td>
<td></td>
</tr>
<tr>
<td>$x_{10}$</td>
<td>0.14</td>
<td>0.01 0.34 0.00 0.80</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.26 100.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Employing the IRLS procedure (34),(35) for two-class regression, we obtain two subsets of data identified by a bigger weight (28) for each observation, then we use the stepwise procedure for eliminating nonsignificant regressors by the partial $F$-test. The results of two-class regressions are presented in Table 3, where for each class there are the columns of regression output similar to those described for the previous table. The coefficients of multiple determination are 0.29 and 0.73, so both regressions make a better fitting than the regular regression for total data. In the class-1 model, the main contributors are $x_4$, $x_8$, and $x_9$, so this model is similar to the total regression, although the variables $x_4$, “few side effects”, and $x_8$, “provides high patient satisfaction”, now play an important role, as it could be expected by their high pair correlations with $y$ (see Table 2). In the class-2 model, the main contributors are $x_2$, $x_1$, and $x_3$ (the attributes of safety, and quick and long relief), while $x_9$ has a negative input in this model. By the sample sizes of two classes, $N_1 = 158$ and $N_2 = 262$, we see that class-2 model with the high $R^2 = 0.73$ corresponds to the larger subset of the doctors, whose decision on a drug prescription is rather defined by more conservative safety issue than by patients' high satisfaction or quality of life.

Table 4, arranged similarly to the previous table, contains the results of three-class regression modeling. The coefficients of multiple determination are in three classes: 0.81, 0.46, and 0.86, so the regressions are even better than those of two-classes. In Table 4, the Class 1 regression is mostly defined by the predictors $x_3$ and $x_2$ of long relief and safety to use, with additional contribution of similar by their sense variables $x_1$, $x_3$, $x_7$, and $x_8$. The Class 2 regression in Table 4 consists of the regressors $x_2$, $x_4$, and $x_9$, of safety, few side effects, and high quality of life. The last Class 3 model in Table 4 has main contributors of $x_2$, $x_9$, and $x_6$, of safety, few side effects, and effectiveness for kids, respectively. It is interesting to note that the last regressor does not arise at all in the models in Table 3, but it yields the main input for the third model in Table 4.

Comparison of the two-class and three-class solutions where the subsets sizes are identified by maximum weight for each respondent is presented in Table 5. The first one in three-class subset (3LC-1) consists of 30 and 102 observations from the first and second ones of two-class subsets (2LC-1 and 2LC-2, respectively), etc. By this contingency table we see that the first class 3LC-1
Table 3. Stepwise latent two classes regressions (sizes $N_1 = 158$ and $N_2 = 262$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Class 1 Regression</th>
<th>Class 2 Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>$t$ Value</td>
</tr>
<tr>
<td>$x_0$</td>
<td>1.76</td>
<td>3.10</td>
</tr>
<tr>
<td>$x_1$</td>
<td>-0.22</td>
<td>-1.80</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-0.25</td>
<td>-2.53</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.30</td>
<td>3.73</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.39</td>
<td>3.10</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.40</td>
<td>3.21</td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.29</td>
<td>100.00</td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Stepwise latent three classes regressions (sizes $N_1 = 132$, $N_2 = 88$, and $N_3 = 200$).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Class 1 Regression</th>
<th>Class 2 Regression</th>
<th>Class 3 Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>$t$ Value</td>
<td>Net Effect</td>
</tr>
<tr>
<td>$x_0$</td>
<td>-2.69</td>
<td>-8.20</td>
<td>0.10</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.27</td>
<td>4.65</td>
<td>0.09</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.43</td>
<td>7.84</td>
<td>0.08</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.22</td>
<td>4.51</td>
<td>0.01</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-0.08</td>
<td>-1.64</td>
<td>0.09</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.60</td>
<td>9.66</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 5. Contingency table for two- and three-latent class models.

<table>
<thead>
<tr>
<th>Model</th>
<th>3LC-1</th>
<th>3LC-2</th>
<th>3LC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2LC-1</td>
<td>30</td>
<td>73</td>
<td>55</td>
</tr>
<tr>
<td>2LC-2</td>
<td>102</td>
<td>15</td>
<td>145</td>
</tr>
</tbody>
</table>

and the third class 3LC-3 in three latent classes mostly correspond to the second class 2LC-2 solution from Table 3. Similarly, the second class 3LC-2 in three latent classes is mostly defined by the first class 2LC-1 solution from Table 3. Comparison of Table 4 with Table 3 shows such kind of correspondence, although with some mixture of the variables and with a new player $x_6$ of effectiveness for kids. Further attempts with four or more classes yield models with only one variable in a class, with small subsample sizes, and correspondingly with statistically nonreliable results, so it does not make sense to increase number of latent classes for this data.

Generally, the results of latent class regression modeling can be considered similarly to the interpretation of the loadings used in factor analysis, where the choice of factor solution depends on the possibility to find an appropriate explanation for the variables grouping in the aggregates. As it is known in the statistical modeling, the IRLS procedure usually converges rather quickly. For instance, in a case of about several hundred observations (similar to the data used in our
example) the convergence is reached after 10-20 iterations, that takes several minutes with a help of such statistical tool as the S-plus package [38]. The time and number of iterations are approximately proportional to the size of data, so for several thousand observations the convergence can be reached in several dozen minutes. Although some values of the regression coefficients could weakly depend on the initial random distribution of the weights, the general structure of the regression models for the latent classes is usually consistent and stable.

5. SUMMARY

We considered a maximum likelihood evaluation of the latent class regressions, presented this approach in a convenient form of iteratively reweighted least squares, and showed how to obtain both coefficients of the models and the probabilities of belonging to each class for each observation. Some extensions for latent class estimation of mixed distributions, and for a multiobjective latent class modeling are also considered. The described technique is very promising, and we continue to investigate the features of the developed approach for its extension to other models, particularly, to the discrete response models. The latent class approach enriches the field of regression modeling and segmentation analysis, and is convenient and useful in practical applications.

REFERENCES