Warsaw School of Economics
Institute of Econometrics
Department of Applied Econometrics

Department of Applied Econometrics Working Papers
Warsaw School of Economics
Al. Niepodleglosci 164
02-554 Warszawa, Poland

Working Paper No. 9-06

Segmentation model with respect to the difference in means

Marcin Owczarczuk
Warsaw School of Economics

This paper is available at the Warsaw School of Economics
Department of Applied Econometrics website at: http://www.sgh.waw.pl/instytuty/zes/wp/

Segmentation model with respect to the difference in means

Marcin Owczarczuk
Warsaw School of Economics
mo23628@sgh.waw.pl

Abstract

The aim of the paper is to formulate and solve the following segmentation problem. Given is a population described by independent variables: X_1, ... , X_n, (both continuous and categorical), the continuous dependent variable Y and the two-level categorical variable α with levels $\alpha = 1$ and $\alpha = 0$. $\overline{Y}_{\alpha=1}$ and $\overline{Y}_{\alpha=0}$ are the means of Y for observations at levels $\alpha = 1$ and $\alpha = 0$, respectively.

The goal is to create the segments of the population, described by the independent variables, that the difference in means $\overline{Y}_{\alpha=1} - \overline{Y}_{\alpha=0}$ is the feature that distinguishes the segments. I. e. the means should be as different as possible between segments and should be similar within the segment. The solution is based on regression trees approach.

Keywords: ANOVA, regression trees, segmentation

JEL codes: C44, M31, C21

1 Introduction

The problem of grouping observations, the clustering, can be described as dividing the set of observations into disjunctive subsets, so that the observations from the same subset are as near to each other as possible and the observations from different subsets are as far to each other as possible (Ćwik, Koronacki [2005]). In the majority of known methods there is a measure of distance or dissimilarity between two observations from the set. For example, as far as the k-means method is concerned, it can be a euclidean distance between the vectors of observation.

We formulate the problem of grouping in a different manner. The feature that distinguishes the segments, that is the conditional difference in means $\bar{Y}_{\alpha=1} - \bar{Y}_{\alpha=0}$ can be calculated only for a segment as a whole and it cannot be calculated for a single observation. Besides, there is no measure of distance or dissimilarity between two observations.

2 The formulation of the problem

Given are two samples from the same population. The first sample is exposed to some factor on the first level $\alpha = 1$, and the second sample is exposed to the same factor on the second level $\alpha = 0$. Both populations are characterised by values of a continuous variable Y (explained variable, dependent variable), values of both discrete and continuous variables $X_1, \ldots, X_n$ (explanatory variables, independent variables) and the value of a factor α. We know that the factor α has influence on variable Y, but this influence dependends on the values of the explanatory variables.

The general population should be divided into segments (by imposing conditions on the explanatory variables $X_1, \ldots, X_n$) so that the following conditions are fulfilled

1. *The condition of homogenity within the segment* The difference in means $\bar{Y}_{\alpha=1} - \bar{Y}_{\alpha=0}$ calculated for all the observations from the particular segment should be equal to the difference calculated for observations from any subsegment of this segment.

2. *The condition of heterogenity between the segments* The difference in means $\bar{Y}_{\alpha=1} - \bar{Y}_{\alpha=0}$ calculated for the observations from the particular segment should be significantly different than the difference calculated for observations from any other segment.

We assume that the mean and variance of variable Y exist.

The scheme of data generation is shown on Figure 1 and the scheme of segmentation on Figure 2.

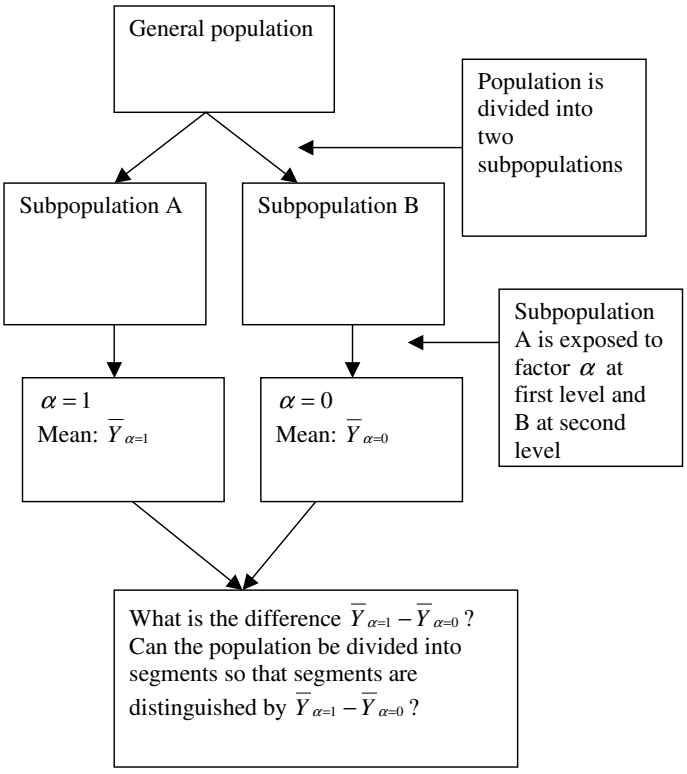

Figure 1: The scheme of data generation.

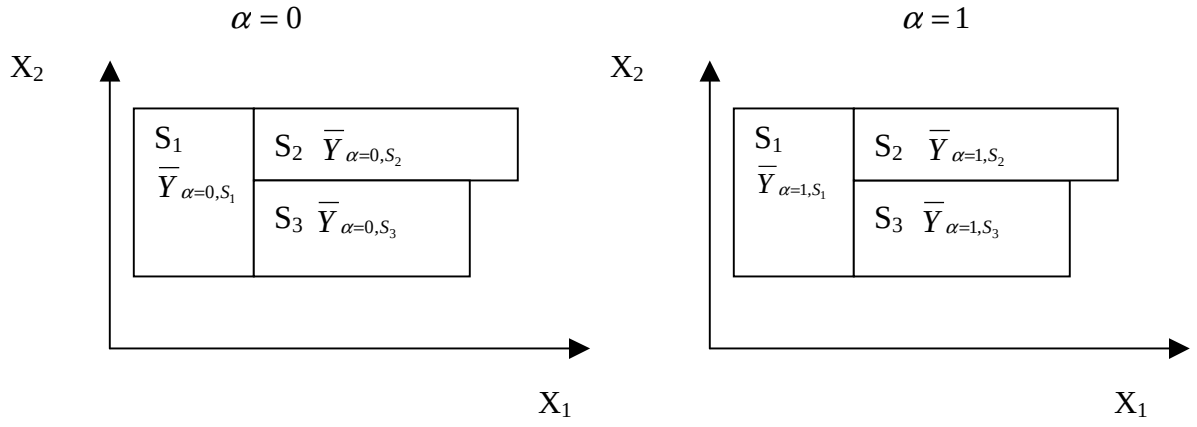

Figure 2: The scheme of segmentation. For simplicity we assume that there are two independent variables X_1, X_2 and three segments. The segmentation should ensure that the quantities $\left(\bar{Y}_{\alpha=1,(X_1,X_2)\in S_i} - \bar{Y}_{\alpha=0,(X_1,X_2)\in S_i}\right)$ for $i = 1, 2, 3$ are significantly different from each other. The area S_1 in both cases, that is for $\alpha = 0$ and $\alpha = 1$ defines the same subset in space of independent variables, analogously for S_2 and S_3.

3 The example of model application - optimization of marketing offers and campaign of banks

Banks offer their clients some services as promotion. The action of a bank is based on a fact that it offers to chosen group of its clients favorable conditions of purchasing particular services and the client may accept these conditions or not. For example the bank may offer purchasing the credit card without charges for maintenance or without charges for remittances.

Bank gains in that case because client must pay incentives when making debit. Besides this offer may discourage the client from resigning from account and changing to another bank. In connection to this offer the bank has particular costs. For example the client may create a debit and become insolvent. It may turn out that the client would purchase the credit card on ordinary conditions in the nearest future anyway and the bank does not collect the charges for maintance of an account. In that case it is reasonable to construct a segmentation of clients into three separate groups

1. group to which it is best to offer these services,
2. groups immune to this marketing campaign
3. groups which bring losses.

We may define the following variables:

Y - profit generated by a client in a particular time, for example in one quarter,

α - the fact that the client purchased the service ($\alpha = 1$) or did not ($\alpha = 0$),

$X_1, \ldots, X_n$ - variables describing the personal characteristics of client and the history of his or her account.

We may formulate the decision problem in a following manner:
On which condition imposed on $X_1, \ldots, X_n$ (that is for which clients) we achieve the positive difference $\bar{Y}_{\alpha=1} - \bar{Y}_{\alpha=0}$? These are the segments of clients which generate profit because of the campaign.
Analogously:
On which condition imposed on $X_1, \ldots, X_n$ (that is for which clients) we achieve the negative difference $\bar{Y}_{\alpha=1} - \bar{Y}_{\alpha=0}$? These are the segments of clients which generate losses because of the campaign.

4 Preliminaries

In this section we briefly describe the one-way ANOVA (Faraway [2002]) and the regression trees (Nong Ye [2003]). These are the statistical tools used in segmentation algorithm described in Section 5.

4.1 ANOVA

Given is a factor α at $i = 1, \ldots, k$ levels and there are $j = 1, \ldots, J_i$ observations of a continuous dependent variable Y at each level of the factor. The ANOVA model can be formulated as follows

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, \ldots, k \qquad j = 1, \ldots, J_i, \tag{1}$$

where y_{ij} denotes the j-th observation of the variable Y at i-th level of the factor, μ denotes the global mean of the variable Y, $(\mu + \alpha_i)$ is the mean of variable Y at i-th level of a factor, and ε_{ij} are independent normally distributed random variables with mean zero and equal variances. We assume that the variances of the dependent variable are equal for all levels of factor.

$$\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2 \qquad i = 1, \ldots, k. \tag{2}$$

Since the parameters μ and α_i are not identifiable, some additional restrictions are necessary, for example

$$\sum_{i=1}^{k} \alpha_i = 0. \tag{3}$$

For these restriction we may write the global mean as

$$\mu = \bar{y}_{\cdot\cdot} = \frac{1}{\sum_{i=1}^{k} J_i} \sum_{i=1}^{k} \sum_{j=1}^{J_i} y_{ij}, \tag{4}$$

and the mean of variable Y at i-th level of factor as

$$\bar{y}_{i\cdot} = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}. \tag{5}$$

The following hypothesis is tested

$$H_0 : \ \alpha_1 = \cdots = \alpha_k = 0$$

$$H_1 : \ \exists \alpha_i \neq 0.$$

In other words we test wheter if the means of the response variable Y are equal for various levels of the factor.

The model of one-way analysis of variance may be formulated as a model of linear regression, that is

$$Y_i = \mu + \beta_1 X_1 + \cdots + \beta_1 X_k + \varepsilon_i, \tag{6}$$

where

$$\beta_1 + \cdots + \beta_k = 0, \tag{7}$$

and X_i are dummy variables:

$$X_i = \begin{cases} 1 & \text{for } i\text{-th level of the factor,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The following hypothesis is tested

$$H_0 : \ \beta_1 = \cdots = \beta_k = 0$$

$$H_1 : \ \exists \beta_i \neq 0.$$

4.2 Regression trees

In this subsection we describe only binary trees (Nong Ye [2003]).

Given are an explained continuous dependent variable Y, explanatory variables $X_1, \ldots, X_n$ (discrete and continuous) and a sample of observations described by these variables. The aim is to construct a model which would explain the variable Y by $X_1, \ldots, X_n$. This can be done by the regression tree which divides population into multidimensional cubes (leaves) by imposing restricting conditions on explanatory variables so that the variability of the variable Y in a leaf is as small as possible. By the variability we mean the variance of Y. In this model we predict the variable Y by its mean value in a leaf. It corresponds to estimating the unknown functional relation between variables Y and $X_1, \ldots, X_n$ by a step function.

For each leaf we define the mean of the explained variable Y

$$\hat{Y_k} = \frac{\sum_{i \in F_k} Y_i}{|F_k|} \tag{9}$$

and the sum of squared residuals

$$SSE_k = \sum_{i \in F_k} (Y_i - \hat{Y_k})^2, \tag{10}$$

where F_k denotes the set of observations which fall into leaf k.
As a prediction of the variable Y of a new observation $t = (X_{1,t}, \ldots, X_{n,t})$ we take the value $\hat{Y_k}$ of a leaf into which the observation t falls.

The building of a tree is recursive. We begin from the whole set (parent node) of observations and try to divide it into two subsets (child nodes). For observations which are in a particular set, we choose the adequate split s, $s = \big(\{X_i \leq c\}, \{X_i > c\}\big)$, where c is properly choosen constant value and X_i is properly choosen explanatory variable. The variability of the Y in each obtained subset $\{X_i \leq c\}$ and $\{X_i > c\}$ (child nodes) should be as small as possible in comparison to the variability of the whole set (parent node). As a measure of the split quality one may use the difference between SSE of the child nodes and SSE of the parent node

$$q(s) = SSE_k - SSE_{kL} - SSE_{kR}, \tag{11}$$

where kL denotes the left child of the node k and kR - right child.
So we choose c and X_i to construct s that maximizes value $q(s)$.
Next we try to apply the just described procedure to each child nodes until the stoping criterion, for example the minimal number of observations in the child nodes, is met.

5 The construction of the model

In this section we propose the new algorithm for implementing the segmentation task. It is based on regression trees. The tree building algorithm may be formulated as follows:

BuildTree(node k, set D, split criterion SS)

1. Apply criterion SS to set D to find the optimal split
2. **if** you are allowed to make a split in the node k **then**:
   (a) use the optimal split to divide the set D into sets D_L and D_R
   (b) **BuildTree**(k_L,D_L,SS)
   (c) **BuildTree**(k_R,D_R,SS)
3. **endif**

As a condition stated in 2. we can use for example the restriction on the maximal tree depth or minimal number of observations in leafs.

The key difference between our algorithm and ordinary regression trees is the measure of variability in the node and the split criterion which is implied by this measure.

As a measure of variability of observations in the node k we take

$$SSE_k = SSE_{k,\alpha=1} + SSE_{k,\alpha=0} = \sum_{i \in F_k, \alpha=1} (Y_{i,\alpha} - \hat{Y}_{k,\alpha=1})^2 + \sum_{i \in F_k, \alpha=0} (Y_{i,\alpha} - \hat{Y}_{k,\alpha=0})^2, \quad (12)$$

where

$$\hat{Y}_{k,\alpha=1} = \frac{1}{|F_{k,\alpha=1|}} \sum_{i \in F_k, \alpha=1} Y_{i,\alpha}, \quad (13)$$

$$\hat{Y}_{k,\alpha=0} = \frac{1}{|F_{k,\alpha=0|}} \sum_{i \in F_k, \alpha=0} Y_{i,\alpha}. \quad (14)$$

The subscripts $\alpha = 1$ and $\alpha = 0$ denote that we consider only observations at level 1 and 0 of the factor α, respectively.

Comment. The proposed measure (12) is equal to the sum of squared residuals in the linear regression model with the variable Y as explained variable and factor α as the explanatory variable. It is also equal to the within variability in ANOVA.

By analogy to the regression trees we may define the decrease of SSE related to the split in a particular node

$$\begin{aligned} q(s) = SSE_k - SSE_{kL} - SSE_{kR} = \\ = \left(\sum_{i \in F_k, \alpha=1} (Y_{i,\alpha} - \hat{Y}_{k,\alpha=1})^2 + \sum_{i \in F_k, \alpha=0} (Y_{i,\alpha} - \hat{Y}_{k,\alpha=0})^2 \right) \\ - \left(\sum_{i \in F_{kL}, \alpha=1} (Y_{i,\alpha} - \hat{Y}_{kL,\alpha=1})^2 + \sum_{i \in F_{kL}, \alpha=0} (Y_{i,\alpha} - \hat{Y}_{kL,\alpha=0})^2 \right) \\ - \left(\sum_{i \in F_{kR}, \alpha=1} (Y_{i,\alpha} - \hat{Y}_{kR,\alpha=1})^2 + \sum_{i \in F_{kR}, \alpha=0} (Y_{i,\alpha} - \hat{Y}_{kR,\alpha=0})^2 \right). \end{aligned} \quad (15)$$

As a criterion SS in the tree building algorithm we take the split s that maximizes the value $q(s)$.

Comment. The presented construction may be considered as a special case of *model tree*. *Model tree* is a mixture of regression trees and linear regression. It consists of estimating in each node of a regression tree, instead of a constant function, the OLS line (separatelly for each leaf) with Y as a explained variable and properly chosen subset of explanatory variables $X_1, \ldots, X_n$. The aim of this construction is to improve the predictive power of the model (Wang, Witten [1997]). In our model we estimate in each leaf the OLS line with variable Y as explained variable and

factor α (after replacing it by the dummy variables) as explanatory variable. However, it should be noted that it is not our goal to predict the values of variable Y but the segmentation with respect to the difference in means of a continuous variable in groups determined by the categorical variable α. Besides, *model tree* does not guarantee that the variable α is used as a predictor in each leaf, because the selection of variables is based on their predictive properties. In our model due to the idea based on fitting the regression with factor α as the only one explanatory variable and due to the properly modified spliting criterion, we achieve the intended segmentation effect.

6 Computer simulations

In this section we illustrate, by examples, the algorithm performance. The main goal is to show that our construction fulfills some natural conditions of a segmentation algorithm. All the calculations were done using R package.

6.1 Analysis of the splitting criterion

The criterion of splitting the observations of the node of a tree should fulfill the following conditions

1. In case when the population is not homogenous that is $\exists_{c \in dom(X_i)}$ that the difference in means of variable Y in the subpopulation $\{Y_i : X_i \leq c\}$ and subpopulation $\{Y_i : X_i > c\}$ are significantly different, the split $s = \big(\{X_i \leq c\}, \{X_i > c\}\big)$ should result in significantly higher than zero value of $q(s)$.

2. In case when the population is homogenous that is $\forall_{c \in dom(X_i)}$ the subpopulation $\{Y_i : X_i \leq c\}$ and the subpopulation $\{Y_i : X_i > c\}$ are characterised by the same difference in means of the variable Y, each split should result in the same value of a criterion $q(s)$.

In order to ilustrate the above conditions we generated two samples, each with 400 observations, using the following schemes:

1.

$$X \sim U[0, 1]$$

$$\varepsilon_i \sim N(0, 0.04)$$

$$Y_i = \begin{cases} 3 + \varepsilon_i & \text{for } X \in [0, 0.5) \text{ and } \alpha \in \{0, 1\} \\ 5 + \varepsilon_i & \text{for } X \in [0.5, 1] \text{ and } \alpha = 1 \\ 1 + \varepsilon_i & \text{for } X \in [0.5, 1] \text{ and } \alpha = 0 \end{cases} \tag{16}$$

2.

$$X \sim U[0, 1]$$

$$\varepsilon_i \sim N(0, 0.04)$$

$$Y_i = 3 + \varepsilon_i \quad \text{for } X \in [0, 1] \text{ and } \alpha \in \{0, 1\}$$

We set the minimal number of observations at each level of the factor in each child node equal 10.

Figures 3 and 4 illustrate these cases.

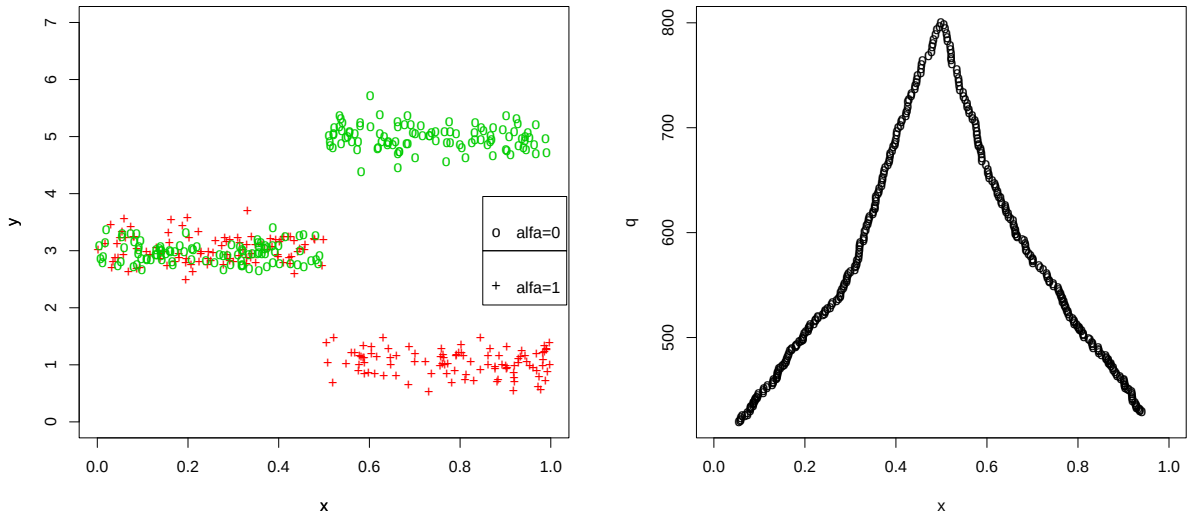

Figure 3: The left-hand side graph shows variable Y as a function of the factor α and values of the explanatory variable X. The right-hand side graph shows the splitting criterion q as a function of a splitting point X. The level of X has influence on the difference in means.

We may note that in case 1 the highest value of the splitting criterion function is in $X = 0.5$ just as it was expected. In case 2 the splitting criterion function is approximatelly constant and equal zero.

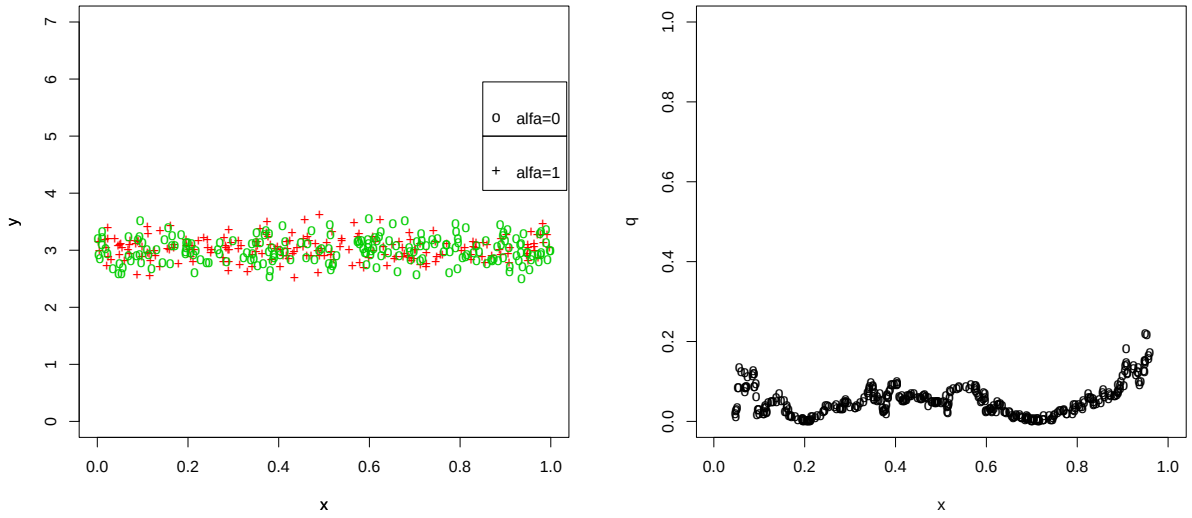

Figure 4: The left-hand side graph shows variable Y as a function of the factor α and values of the explanatory variable X. The right-hand side graph shows the splitting criterion q as a function of a splitting point X. The level of X has no influence on the difference in means.

6.2 The example of tree performance

In order to show the tree performance we generated 1000 observations according to the following scheme:

$$X \sim U[0,1]$$

$$\varepsilon_i \sim N(0, 0.0225)$$

$$Y_i = \begin{cases} 2+\varepsilon_i & \text{for } X \in [0,0.4) \\ 5+\varepsilon_i & \text{for } X \in [0.4,0.6) \\ 3+\varepsilon_i & \text{for } X \in [0.6,1] \end{cases} \text{ for } \alpha = 1, \qquad Y_i = \begin{cases} 4+\varepsilon_i & \text{for } X \in [0,0.2) \\ 1+\varepsilon_i & \text{for } X \in [0.2,0.8) \\ 6+\varepsilon_i & \text{for } X \in [0.8,1] \end{cases} \text{ for } \alpha = 0 \tag{17}$$

The graph of variables for this scheme is presented on Figure 5. During the building tree we set the condition of the possibility of the split application as the minimal number of observations at each level of factor in each node equal 50.

We may note that the model fits well to the data if the dependency between the explained variable Y and the explanatory variables is approximatelly piecewise constant, which is a typical characteristics of trees. One should expect that in case of a different functional dependency, for

example linear, the quality of tree is lower.

7 Conclusion

In this paper we formulated the problem of segmentation of population based on the criterion of the difference in means determined by levels of a categorical variable. Next we constructed the model, based on regression trees, which realises this task. On simulations, we showed the basic features of this solution.

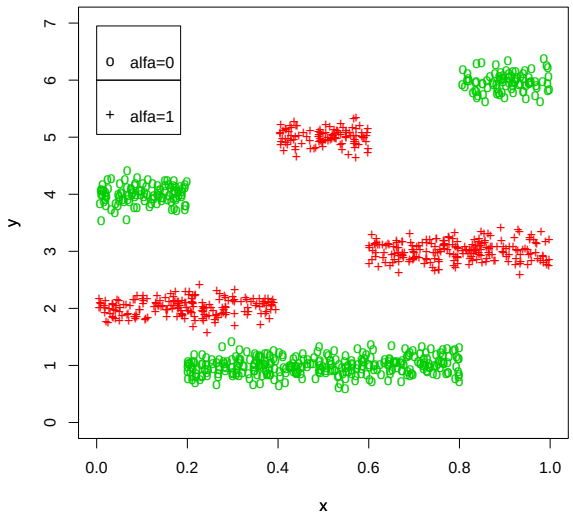

Figure 5: The graph of the explained variable Y as a function of explanatory variable X and the factor α

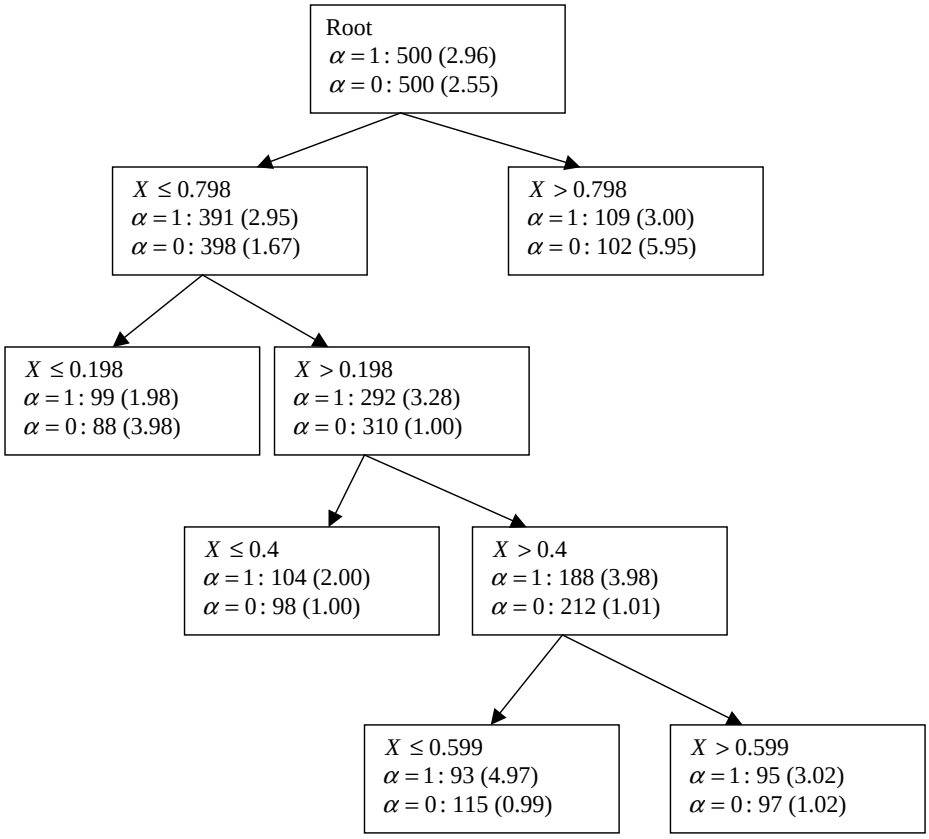

Figure 6: The tree achieved as a result of the algorithm for data from Figure 5. In each node we reported the number of observations and, in the parentheses, the mean of variable Y at each level of factor.

References

[1] Breiman L., Friedman J. H., Olshen R.A., Stone C. J., *Classification and regression trees*, Wadsworth, Belmont CA., 1984

[2] Ćwik J., Koronacki J. *Statystyczne systemy uczące się*, WNT, Warsaw 2005

[3] Faraway J., *Practical Regression and Anova using R*, 2002, http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

[4] Koronacki J., Mielniczuk J. *Statystyka dla studentów kierunków techniczych i przyrodniczych*, WNT, Warsaw 2001

[5] Nong Ye (red.), *The handbook of data mining*, Lawrence Erlbaum Associates, Mahwah 2003

[6] Wang Y., Witten I.H. Induction of model trees for predicting continuous classes, *Proc European Conference on Machine Learning Poster Papers*, s. 128-137, Prague, 1997.