A Strategy for the Selection of Regression Models with Two Qualitative Regressors

Dr. Lobna Eid AL-Tayeb

Faculty of Commerce, Al-Azhar University (Girls' Branch), Egypt
lobnaalatyeb@azhar.edu.eg

Scientific Journal for Financial and Commercial Studies and Researches (SJFCSR)

Faculty of Commerce – Damietta University

Vol.3, No.2, Part 1., July 2022

APA Citation:

AL-Tayeb, L. E. (2022) A Strategy for the Selection of Regression Models with Two Qualitative Regressors, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 3(2)1, 657 - 672.

Website: https://cfdj.journals.ekb.eg/

A Strategy for the Selection of Regression Models with Two Qualitative Regressors

Dr. Lobna Eid AL-Tayeb

Abstract

This article proposes a strategy for the selection of regression models with quantitative response and two qualitative regressors. The proposal is based on the minimization of an estimator of the Mean Square of the Prediction Error ($M\widehat{S\,E\,P}$)within a predefined class of models. Some structures for the interactions are considered, among which are those proposed. Some graphs are proposed to diagnose the form of the interaction.

Keywords: qualitative variables, double classification, model selection

Introduction

Sir Francis Galton coined the word "regression" in 1885 [3] after demonstrating that children's height does not appear to represent that of their parents, but rather regresses to the average population. Currently, the term "regression analysis" refers to a broad range of statistical techniques for modelling variable relationships and predicting the value of one or more dependent (or response) variables from a collection of independent variables (or predictors).

The traditional way of approaching the regression problem with two qualitative regressors is that of the Double Classification [1], where from the very beginning a model is assumed to be adequate and afterwards, certain hypotheses for the parameters are tested in order to compare the effects for the different levels. In this way of approaching the problem the idea of model selection is not present.

Let x_1 and x_2 be two qualitative variables that can take $n_1$and n_2 different categorical values respectively. These categories will be denoted by i_1, i_1, ..., $i'_{\overline{n}1}$ and i'_1, i'_2, ..., $i'_{\overline{n}2}$. Suppose that observations have been made on a quantitative random variable Y at $m = n_1 \times n_2$ points in the set, $\{(i_j, i'_j): 1 \leq j \leq \overline{n}1, 1 \leq j' \leq \overline{n}2\}$.

Without loss of generality, we will assume $n_1 = \bar{n}_1$ and $n_2 = \bar{n}_2$, furthermore at the point (i_j, i'_j) and n_{i_j,i'_j}observations are made.

Traditionally, a model such as the following is assumed:

$$E\left(Y_{i_j,i'_{j,k}}\right) = \mu_{i_j,i'_j} = \mu + \alpha_{i_j} + \beta_{i'_j}, + \gamma_{i_{j'i'_j}} \qquad (1)$$

Observations are assumed to be uncorrelated. A model like the previous one is over parameterized (it contains the maximum number of parameters, that is, $n_1 + n_2 + (n_1 \times n_2) + 1$ parameters), which implies a reduction in the precision of the estimates, since the variance of the estimators grows as the number of parameters in the model grows. It is known that the least squares estimate in Equation-1 is,

$$\hat{\mu} = \overline{\overline{Y \ldots} = 1/n} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \sum_{k=1}^{n_{i_j,i'_{j'}}} Y_{i_j,i'_{j,k}} \qquad (2)$$

These estimators have interesting properties as they are least squares [9]. Considering different regression equations is equivalent to considering different partitions on the value space of the regressors. In the very simple case of n_1 = 3 and n_2 = 2 and the interactions equal to zero (additive model) [5]. The equations that define the possible models are:

1- Modify to firstly; Model with one parameter.
2- Modify with two parameters.
3- Modify with three parameters
4- Modify with four parameters.

The number of parameters in the model grows as the number of classes grows in the partition that the model induces in the value space $x = x_1 \times x_2$. Counting the number of models with p parameters would be

equivalent to counting the partitions where the sum of the class numbers of the partitions in the value spaces of x_1 and x_2 equals $p + 1$.

Proposition 1.1: Let x_1 and x_2 be two qualitative variables with n_1 and n_2 different values ($n_1 \leq n_2$). So there are:

$$\sum_{a=p+1-Min(n_2,p)}^{Min(n_1,p)} \{[\sum_{j=1}^{a} \frac{(-1)^{a-j}}{j!(a-j)!} j^{n_1}][\sum_{j=1}^{p+1-a} \frac{(-1)^{p+1-a-j}}{j!(p+1-a-j)!} j^{n_2}]\} \quad (3)$$

Different additive models with p parameters.

Proof: It is known that if x is a qualitative variable with m different categories, the number of models with p parameters (number of partitions with p classes available in literature and cited by Bunke and Castell (1998)[6].

Therefore, the quantities, $H_1; H_2$ *where*

$$H_1 = \sum_{j=1}^{a} \frac{(-1)^{a-j}}{j!\,(a-j)!} j^{n_1}$$

$$H_2 = \sum_{j=1}^{p+1-a} \frac{(-1)^{p+1-a-j}}{j!\,(p+1-a-j)!} j^{n_2}$$

Represent the number of partitions with a class in the value space of x_1 and with $p + 1 - a$ classes in the value space of x_2; being its sum equal to $p + 1$. If M is an additive model defined by the equation that induces partitions such as those considered, it will have p parameters. For a fixed, each partition with a class in the value space of x_1 is combined with each of the H_2partitions with $p + 1 - a$ classes in the value space of x_2, that is, for a fixed the number of models with p parameters it will be $H = H_1 . H_2$. But the number of classes of the partitions in the value space of x_1 can be at least equal to $p + 1 - Min\,(n_2, p)$ and at most $Min\,(n_1, p)$. now adding for all possible values of the formula given in Eq.3 is obtained.

The number of additive models will serve as a reference to assess how large the number of possible models is.

2. MODELING OF INTERACTIONS

When working with two qualitative regressors and wanting to perform a complete analysis, it is necessary to determine the presence or not of the interaction term and estimate the variance. The usual model with interactions is over-parameterized and as already mentioned this is a big drawback. The consideration of certain forms of interactions proposed by many scientists including Huet (1991)[2] produces a decrease in the number of parameters. This consideration is tremendously important when there is only one observation for combinations of treatments, because then the classical theory of linear models cannot be used, since the estimation of the variance of the error has to be obtained from the sum of squares of the interactions.

The different models that are proposed to be considered others are, Additive model; Concurrent model; Regression model per column; Regression model per row; Fifth model; Mandel's model; Seventh, eighth, ninth and tenth models [8].

Subject to restrictions,

$$\sum_{j=1}^{n_1} \alpha_{i_j} = \sum_{j\prime=1}^{n_2} \beta_{i\prime_{j\prime}} = \sum_{j=1}^{n_1} \theta_{i_j} = \sum_{j\prime=1}^{n_2} \vartheta_{i\prime_{j\prime}} = \sum_{j\prime=1}^{n_1} \gamma^2{}_{i_j,i'_{j\prime}} = \sum_{j\prime=1}^{n_2} \gamma_{i_j,i'_{j\prime}} = 0$$

$$\sum_{j=1}^{n_1} \theta_{i_j}{}^2 = \sum_{j\prime=1}^{n_2} \vartheta^2{}_{i\prime_{j\prime}} = \sum_{l=1} U^2_{li_j} = \sum_{l=1} V^2_{li\prime_{j\prime}} = 1$$

The number of parameters in any model listed above will be equal to the number of parameters of the additive model contained in it, plus the number of parameters provided by the interactions.

3. A STRATEGY FOR THE SELECTION OF THE MODEL

It is considered that observations $Y_{i_j,i'_{j,k}}$ are made on a random variable$Y$, which satisfy the regression equation:

$$Y_{i_j,i'_{j,k}} = f\left(i_j, i'_{j'}\right) + \varepsilon_{i_j,i'_{j'}k} \tag{4}$$

They are assumed to be $\varepsilon_{i_j,i'_{j'}k}$ random (unobservable) errors with zero expectation and variance σ^2 and that they are unrelated.

The function $f\left(i_j, i'_{j'}\right)$belongs to a set M defined by:

$$M = \{g\ (x_1, x_2, \beta): \beta\ \in\ \beta\} \tag{5}$$

with

$$g\ (x_1, x_2, \beta) = \sum_r 1_{C_r^g}(x_1, x_2)\mu_{i_j,i'_{j'}} \tag{6}$$

$$\beta^t = (\mu, \alpha_{i_1}, \alpha_{i_2 \ldots\ldots}\alpha_{ip_1}, \beta_{i_1,1}, \beta_{i'_2 \ldots\ldots}\beta_{i'_2} \vdots \gamma_{1,}\gamma_{2,} \ldots . \gamma_{p3,})$$

where

$\forall\ |\gamma_l = \gamma_{li_j,i'_{j'}}$ for some $(i_j, i'_{j'})$ and is any of the types of interactions described and is given by one of the expressions from the different models given above.

$$N = \bigcup_r C_r^g$$

So, the selection of the model is given by the selection of a function g in M to approximate$f$. The set of observations will be used to calculate an

estimator β of the vector of parameters β and select a function $g\ (x_1, x_2, \hat{\beta})$.

In the selection of the function $g\ (x_1, x_2, \hat{\beta})$. the minimization criterion of a $MSEP$ or MSE estimator will be used. The vector of parameters β will be estimated by the Least Squares Method.

In the first stage of the strategy, the models with a number of parameters less than or equal to a number p0 set by the user appropriately and within the permissible limits will be analyzed. Subsequently, a way must be found to reduce the number of models to be compared and the idea would be to make a reduction in such a way that between one step of the strategy and another the analyzed models do not change abruptly in terms of the number of parameters. In achieving this last objective, the concept of the neighbour model that is given below plays a fundamental role.

Definition 3.1. (Neighbouring model)

Let,

$$\delta \in \{\alpha;\ \beta;\ \gamma\ (x_1, x_2)\}$$

Let M_O be a model. Let π_0 be the partition determined by the function $g_0\ (x_1, x_2, \beta)$. that defines the model M_O. $s$ean C_{ok} the classes of this partition. It is said that it is a neighbour model of M_V^{δ}. M_O According to δ, if it is true:

1- There is one and only one δ and one and only one class C_{ok} such that:

$$C_{ok} = (C_{ok\prime})\ \delta \cup (C_{vk''})\ \delta$$

Where the superscript δ has been used to indicate that the class is affected only according to that parameter.

2- For all r ≠ k there exists r′, such that k′ ≠ r′ ≠ k′′, for which it is true:

$C_{or} = (C_{vr})^{\delta}$

3- Furthermore, the form of the interaction of the model is the same as that of the M_V^{δ}. M_O Model.

From the way the neighbour model has been defined, it is intuitively clear that it has one more parameter than M_O.

In a problem with two qualitative variables, there are three parameter names, which have been represented as follows:

α → levels of the first variable

β → levels of the second variable

γ → interactions

To build a selection strategy, based on neighbouring models, it is necessary to decide in what order the nominations are taken to increase the number of parameters in the model. Since there is no preference between one factor or another, this order is irrelevant. As you have to decide on an order, it is proposed: α → β → γ.

3.1. Strategy for model selection

Let γ_1 $(x_1, x_2,)$ be the form of interaction in the model M_l $(\cdot), l = 1,2,\ldots,10$. Let p (l) be the maximum number of allowable parameters for the model M_l $(\cdot)$.

let

$$M_l^0 = \{(M_l\ (p): p \le p_o; 1 \le p_o \le p(l)\}$$

Let

$\hat{r}\big(M_l\ (p)\big) = \hat{r}(l,\mathrm{p}(l))$

An estimator of the MSEP for the M_l (p) model.

1. Calculate the number of possible additive models for each $p = 1,2,\ldots,p$ (l). Being l a fixed number.

2. According to what is determined in the previous point, the computing facilities and the time and effort that one is willing to use, select a value p_0 and thus the set that M_l^0 we will call the set of basic models will be determined.

In what follows, the following notation will be used, for example $(M_l^0)_v^\alpha$is a neighbour model of according to the partitionsM_l^0 corresponding to the parameter α. That is, a parameter has been increased with respect to the number it contained M_l^0 but the increase is made in the denomination α. The class formed by the neighbours of according to the parameter α, will be denoted by $M_l^{v,\alpha}$

Perform the following steps:

1. Determine

$$M_l^0 = \underset{M \in M_l^0}{Avg\ Min\ \hat{r}(l,p)}$$

Let

$$M_l^0 = M_l^0(p')$$

where p' represents the number of parameters in M_l^0

2. Let

$$M_l^{v,\alpha} = \{(M_l^0)_v^\alpha\}$$

Determine

$$M_l^{1,\alpha} = \underset{M \in M_l^{v,\alpha}}{Avg\ Min\ \hat{r}(l,p'+1)}$$

Then,

$$M_l^{1,\alpha} = M_l^{1,\alpha}(p'+1)$$

3. Let

$$M_l^{v,\beta} = \{\left(M_l^{1,\alpha}\right)_v^\beta\}$$

Determine

$$M_l^{2,\beta} = \underset{M \in M_l^{v,\beta}}{Avg\ Min\ \hat{r}(l,p'+2)}$$

4. Let

$$M_l^{v,\gamma(x_1,x_2)} = \{\left(M_l^{2,\beta}\right)_v^{\gamma(x_1,x_2)}\}$$

Determine

$$M_l^{3,\gamma(x_1,x_2)} = \underset{M \in M_l^{v,\gamma(x_1,x_2)}}{Avg\ Min\ \hat{r}(l,p'+3)}$$

5. Let

$$M_l^{v,\alpha} = \{\left(M_l^{3,\gamma(x_1,x_2)}\right)_v^{\alpha}\}$$

Determine

$$M_l^{4,\alpha} = \underset{M \in M_l^{v,\alpha}}{Avg\ Min\ \hat{r}(l,p'+4)}$$

6. Let

$$M_l^{v,\beta} = \{\left(M_l^{4,\alpha}\right)_v^{\beta}\}$$

Determine

$$M_l^{5,\alpha} = \underset{M \in M_l^{v,\alpha}}{Avg\ Min\ \hat{r}(l,p'+5)}$$

7. Let

$$M_l^{v,\gamma(x_1,x_2)} = \{\left(M_l^{5,\alpha}\right)_v^{\gamma(x_1,x_2)}\}$$

Determine

$$M_l^{6,\gamma(x_1,x_2)} = \underset{M \in M_l^{v,\gamma(x_1,x_2)}}{Avg\ Min\ \hat{r}(l,p'+6)}$$

8. Repeat steps 5, 6 and 7. In this way, for $q \in \{z: 0 \leq z\}$, steps $2 + 3q, 3 + 3q$ and $4 + 3q$, would be determined by:

⋮

$2 + 3q.-$

$$M_l^{v,\alpha} = \{\left(M_l^{3q,\gamma(x_1,x_2)}\right)_v^{\alpha}\}$$

Determine

$$M_l^{1+3q,v} = \underset{M \in M_l^{v,\alpha}}{Avg\ Min\ \hat{r}(l, p' + 1 + 3q)}$$

Let, $3 + 3q$.-

$$M_l^{v,\beta} = \{\left(M_l^{1+3q,\alpha}\right)_v^{\beta}\}$$

Determine

$$M_l^{2+3q,v} = \underset{M \in M_l^{v,\beta}}{Avg\ Min\ \hat{r}(l, p' + 2 + 3q)}$$

Let, $4 + 3q$.-

$$M_l^{v,\gamma(x_1,x_2)} = \{\left(M_l^{3q,2,\beta}\right)_v^{\gamma\gamma(x_1,x_2)}\}$$

Determine

$$M_l^{3q,+3,v} = \underset{M \in M_l^{v,\gamma(x_1,x_2)}}{Avg\ Min\ \hat{r}(l, p' + 3 + 3q)}$$

When the corresponding denomination cannot be increased in a step, it is replaced by the next one in the order of affectation

4. SOME USEFUL GRAPHICS TO EXPLORE THE INTERACTION MODELING

Two models have been considered neighbours when the classes that they define remain invariant except for one of them that is divided into two new classes. However, according to the number of parameters provided by the interactions, quite homogeneous groups could be formed. A model that is selected with one type of interaction in group k can be examined by changing the form of its interaction for another in group $k + 1$.

$$
\begin{array}{ccccc}
 & & \vdots\, a)\alpha_i d_i & & \\
 & & \vdots\, b)c_i\beta_i & & \\
0 \rightarrow & \lambda\alpha_i\beta_i \rightarrow & \vdots\, c)\lambda\alpha_i\beta_i + c_i\beta_i & \rightarrow & \\
 & & \vdots\, d)\lambda\alpha_i\beta_i + \alpha_i d_i & & \\
1 & 2 & 3 & & \\
 & \vdots\, a)\lambda\, c_i d_i & & & \\
\vdots\, b)\lambda\alpha_i\beta_i + \alpha_i d_i + c_i\beta_i & \rightarrow & \vdots\, a)\lambda_1\, U_{1i}V_{1i} + \lambda_2\, U_{2i}V_{2i} & & \\
 & & \vdots\, b)\gamma_{ij} & & \\
 & \vdots & & & \\
 & \vdots & & & \\
 & 4 & & 5 &
\end{array}
\qquad (7)
$$

Let the i-j-th remainder be

$$r_{ij} = \hat{Y}_{ij} - Y_{ij}$$

It is known that Zwanzig (1979); Humak (1983)

$$r_{ij,i'j'} \underset{\rightarrow}{L} N\left(f\left(i_j, i'_{j'}\right)\right) - g(i_j, i'_{j'}, \beta(.))$$

As the Normal distribution is a symmetric distribution, its mean coincides with the median and this fact can be used to think of some graphic situations that allow further exploration. If the model is correct then the median of the limiting distribution is zero.

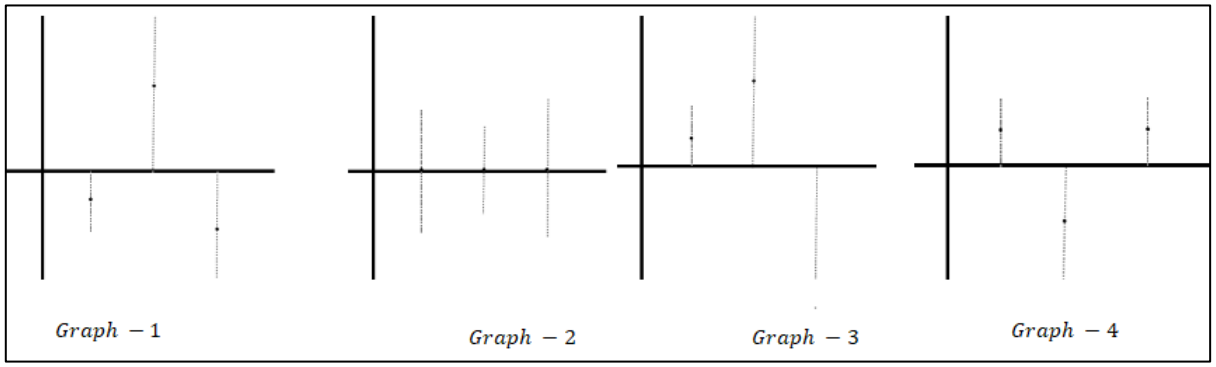
In order to gain clarity, the case where both qualitative variables can only take three values will be considered.

Case 1. Suppose that a model with a form of interaction such as the one described in column 2 of Eq.5 was selected, and that the correct model (interaction) is of the form in a) of column 3:

Med $r_{\text{lm}} = \alpha_l d_m - \lambda\alpha_l\beta_m$

$= \alpha_l\ (d_m - \lambda\beta_m) = \alpha_l k_m$ That is, the median of the l-m-th residue is a function of α_l. What $\sum_{l=1}^{3} \alpha_l = 0$.

It has to:



Med $r_{\text{lm}} = \alpha_2 k_m - \alpha_3 k_m$

= - Med $r_{2\text{m}}$ - Med $r_{3\text{m}}$

A graph of residuals showing these against the $\widehat{\alpha_i} o$ against its indices, it can provide invaluable information to guide subsequent explorations.

Some graphical illustrations are given in graph 1-4 a graph like number 2 does not suggest that relationship and, therefore, no changes in that sense. While the graphs numbers 1, 3 and 4, if they suggest changes, and then the model should be explored by making said change in the interaction modelling.

Case 2. Consider the situation of case 1 but now the true model is the one given in b) of column 3 in Eq.5. The result is similar but the dependency arises with respect to β_j:

Med $r_{\text{i,j}} = k_i \beta_j$.

The residuals should be plotted against or against their indices. $\hat{\beta}_j$

These graphs can be useful in almost all situations except the one where the model given in a) of column 5 in Equation 5 is involved.

Conclusions

All the models referred to above except 1 and 10; they are non-linear; therefore, it seems reasonable to reject the hypothesis of constant variance. This has been considered in a program made in Turbo Pascal 7.0. Twenty data sets were simulated with models No. 2 and No. 4 (ten with each) and it was obtained that the strategy selected the optimal model in 70% of the cases. The results are found in Tables 1 and 2.

Table 1: Model→ $E\left(Y_{i_j,i'_{j'}}\right) = \mu + \alpha_{ij} + \beta_{i',j'} + \lambda\alpha_{ij}\beta_{i',j'}$

MSEP	Selected the optimal model	MSEP of the optimal model
2.42	x	
2.29	x	
4.33	-	4.012
1.25	x	
3.25	x	
2.30	-	2.025
3.55	x	
1.64	x	
5.52	-	4.293
2.89	x	

Table 2: Model→ $E\left(Y_{i_j,i'_{j'}}\right) = \mu + \alpha_{ij} + \beta_{i',j'} + c_i\beta_{i',j'}$

MSEP	Selected the optimal model	MSEP of the optimal model
15.52	x	
7.21	x	
6.07	x	
10.24	x	
23.44	-	15.697
21.29	-	19.567
10.10	-	9.804
4.41	x	
9.81	x	
35.39	x	

References

[1]. Belitz, Christiane & Lang, Stefan. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics & Data Analysis. 53. 61-81. 10.1016/j.csda.2008.05.032.

[2]. Ballabio, Davide & Todeschini, Roberto. (2009). Multivariate Classification for Qualitative Analysis. 10.1016/B978-0-12-374136-3.00004-3.

[3]. Browne, Janet. (2016). Inspiration to Perspiration: Francis Galton's Hereditary Genius in Victorian Context. 10.1057/9781137497673_6.

[4]. Johnson, D.E. and F. Graybill (1972): "An analysis of two-way model with interaction and no replication", The Journal of the Acoustical Society of America, 67, 862-868.

[5]. Kim, Byungchan. (2012). On the number of partitions of n into k different parts. Journal of Number Theory. 132. 1306–1313. 10.1016/j.jnt.2012.01.003.

[6]. Kaya Uyanık, Gülden & Güler, Neşe. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences. 106. 234–240. 10.1016/j.sbspro.2013.12.027.

[7]. Kuo , Martin & Yi-Chi Chen (2021). Variable selection in finite mixture of regression models with an unknown number of components. Computational Statistics & Data Analysis Volume 158, June 2021, 107180.

[8]. Rodrigues, Paulo Canas, Pereira, Dulce Gamito Santinhos, & Mexia, João Tiago. (2011). A comparison between Joint Regression Analysis of and the Additive Main and Multiplicative Interaction model: the robustness with increasing amounts missing data. Scientia Agricola, 68(6), 679-686.

[9]. Snee, Ron & Rayner, Arthur. (1982). Assessing the Accuracy of Mixture Model Regression Calculations. Journal of Quality Technology. 14. 67-79. 10.1080/00224065.1982.11978792.

الملخص العربي

A STRATEGY FOR THE SELECTION OF REGRESSION MODELS WITH TWO QUALITATIVE REGRESSORS

د. لبني عيد الطيب

استاذ مساعد بكلية التجارة جامعة الازهر بنات القاهرة

ظهرت كلمة "الانحدار" في عام 1885علي يد (Francis Galton) بعد أن أظهر أن طول الأطفال لا يبدو أنه يمثل ارتفاع والديهم، بل يرجع إلى متوسط عدد السكان. حاليًا، يشير مصطلح "تحليل الانحدار" إلى مجموعة واسعة من الأساليب الإحصائية لنمذجة العلاقات المتغيرة والتنبؤ بقيمة واحد أو أكثر من المتغيرات التابعة (أو المتغيرات) من مجموعة من المتغيرات المستقلة (أو المتنبئين).

اقترحت هذه المقالة استراتيجية لاختيار نماذج الانحدار ذات الاستجابة الكمية واثنين من عوامل الانحدار النوعي. يعتمد الاقتراح على تصغير مقدر المربع المتوسط لخطأ التنبؤ ((M S E P)) ضمن فئة محددة مسبقًا من النماذج. يتم النظر في بعض الهياكل للتفاعلات، من بينها تلك المقترحة. تم اقتراح بعض الرسوم البيانية لتشخيص شكل التفاعل

جميع النماذج المشار باستثناء نموذجين؛ هم غير خطيين، لذلك يبدو من المعقول رفض فرضية التباين المستمر. تم أخذ ذلك في الاعتبار في برنامج تم إنشاؤه في .Turbo Pascal 7.0 تمت محاكاة عشرين مجموعة بيانات مع النموذجين رقم ٢ ورقم ٤ (عشرة مع كل مجموعة) وتم الحصول على أن الاستراتيجية اختارت النموذج الأمثل في ٧٠٪ من الحالات. تم العثور على النتائج في الجدولين ١ و٢.