A Method for Evaluating Validity of Piecewise-linear Models

Oleg V. Senko[1], Dmitriy S. Dzyba[2], Ekaterina A. Pigarova[3], Liudmila Ya. Rozhinskaya[3] and Anna V. Kuznetsova[4]

[1]*Dorodnicyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, 119991 Moscow, Russia*

[2]*Lomonosov Moscow State University, Leninskie Gory, Moscow, Russia*

[3]*Department of Neuroendocrinology and Bone Diseases, Endocrinology Research Centre, 11 Dmitry Ulyanov st., 117036 Moscow, Russia*

[4]*Emanuel Institute of Biochemical Physics, ul. Kosygina 4, 117997 Moscow, Russia*

Keywords: Regression Model, Optimal Complexity, Permutation Test.

Abstract: A method for evaluating optimal complexity of regression models is discussed. It is supposed that complicated model must be used only when any simple model fails describe exhaustively regularity that exists in data. At that null hypothesis about exhaustive explanation of data by simple regularity is tested with the help of complicated model. Validity of null hypothesis is evaluated with the help p-value that is calculated with the help of special version of permutation test. An application is discussed where developed technique is used to evaluate if more complicated piecewise-linear regressions must be used instead of simple regressions to describe correctly dependence of parathyroid hormone on vitamin D status.

1 INTRODUCTION

Standard task of statistical modelling is discussed. It is necessary to find statistical model that forecasts response Y by variables $X_1,\dots,X_n$:

$$Y = F(X_1,\dots,X_n) + \varepsilon,$$

where $F(X_1,\dots,X_n)$ is predicting function and ε is error term. Function F with minimal mathematical mean $\mathbb{E}\varepsilon^2$ is chosen from family $\widetilde{M}$ by data set $\widetilde{S}_0 = \{(y_1^0,\mathbf{x}_1^0),\dots,(y_m^0,\mathbf{x}_m^0)\}$, where $y_1^0,\dots,y_m^0$ are values of response variable Y and $\mathbf{x}_1^0,\dots,\mathbf{x}_m^0$ are vectors of predicting variables $X_1,\dots,X_n$. It is supposed that observations corresponding different objects from $\widetilde{S}_0$ are independent and are taken from the same probability space. Success of modelling depends on correct choice of predicting function F complexity or more exactly on complexity of family $\widetilde{M}$. Today there are several approaches for complexity optimization that allow to discourage overfitting effect. Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978), Hannan-Quinn information criterion (Hannan and Quinn, 1979), Risannen principle (Rissanen, 1978) may be mentioned there above. These techniques often allow to find out complexity level with best generalization ability. But in many application tasks it is important not only to find model of optimal complexity but also to estimate validity of choice. Let suppose that models may be searched inside simple family $\widetilde{M}_s$ and more complicated family $\widetilde{M}_c$. At that $\widetilde{M}_s \subseteq \widetilde{M}_c$. It is not sufficient to find out if optimal model must be searched inside family $\widetilde{M}_s$ or inside family $\widetilde{M}_c \backslash \widetilde{M}_s$. It is also necessary to evaluate our confidence that model found inside $\widetilde{M}_c \backslash \widetilde{M}_s$ really better describes data than model found inside family $\widetilde{M}_s$. It must noted that choice between two families sometimes corresponds to choice between two suppositions about type of process that generates studied data. It may be physical, chemical or biological process for example. Usually in statistics validity of choice between two hypotheses is evaluated with the help of p-values. The same way of validity evaluating is used in this paper. It is considered that complicated family must be used then and only then when any simple model fails to describe exhaustively regularity that exists in data. At that null hypothesis about exhaustive explanation of existing regularity by simple predictive function from $\widetilde{M}_s$ is tested with the help of complicated family $\widetilde{M}_c$. Such approach correspond to well known principle of of Occam's razor that is attributed William of Occam living in the 14th century. The most popular version of razor is formulated as "Entities should not be multiplied beyond necessity." Later razor principle

V. Senko O., S. Dzyba ., A. Pigarova E., Ya. Rozhinskaya L. and V. Kuznetsova A..
A Method for Evaluating Validity of Piecewise-linear Models.
DOI: 10.5220/0005156904370443
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 437-443
ISBN: 978-989-758-048-2

was adopted by many scientists and another variants were invented. Principle was stated by Isaac Newton in form "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." Such form as it may be seen is most close to approach that is represented in the paper. Problems that are associated with Occam's razor are discussed in modern scientific literature on machine learning or knowledge discovery. Usually it is considered that razor is a way to improve forecasting ability. Arguments for and against such razors are represented in details in (Domingos, 1999). Approach that is discussed in this paper is based on testing of null hypotheses with the help of random permutation test. Let note that random permutation test now is rather popular technique allowing to evaluate statistical validity without any additional suppositions (Ernst, 2004; Good, 2005). Permutation tests also are used to study regression or recognition models (Kim et al., 2000; Ojala and Garriga, 2010; Golland et al., 2000).

2 EVALUATING VALIDITY OF COMPLICATED MODELS

2.1 Main Suppositions

It is supposed that optimal predicting function $F_0(\mathbf{x})$ is searched inside family *widetildeM* by some training set $\widetilde{S} = \{(y_1,\mathbf{x}_1),\dots,(y_m,\mathbf{x}_m)\}$ with the help of least squares technique:

$$F_0(\mathbf{x}) = arg\min_{F\in\widetilde{M}} Q[\widetilde{S},F(\mathbf{x})],$$

where $Q[\widetilde{S},F] = \sum_{j=1}^{m}[y_j - F(\mathbf{x_j})]^2$. Minimal value of $Q[\widetilde{S},F(\mathbf{x})]$ at set $\widetilde{M}$ will be referred to as $Q_{min}(\widetilde{S},\widetilde{M})$. The represented approach is based on several simple suppositions.

Supposition 1. More complicated function from $\widetilde{M}_c$ must be used only when there is no function inside family $\widetilde{M}_s$ that exhaustively describes data.

Supposition 2. It is considered that some function F exhaustively describes dependence of Y from $X_1,\dots,X_n$ if residuals $\{r_1 = y_1 - F(\mathbf{x}_1^0),\dots,r_m = y_m - F(\mathbf{x}_m^0)\}$ are realizations of mutually independent identically distributed random values $\xi_1,\dots,\xi_m$ that are independent on vector descriptions $\mathbf{x}$. It is supposed also that $\mathbb{E}(\xi_i) = 0,\ i = 1,\dots,m\}$.

Supposition 3. It is possible to reject (or verify) null hypothesis that function F exhaustively describes dependence on X variables with the help of complicated family $\widetilde{M}_c$.

2.2 Permutation Test Technique

Let $\widetilde{f}$ is set of all possible permutations of integers $\{1,\dots,m\}$. Let $\widetilde{S}_p(f,F)$ be data set that is received from initial data set $\widetilde{S}_0$ by random permutation of residuals $(r_1,\dots,r_m)$:

$$\widetilde{S}_p(f,F) = \{[r_{f(1)} + F(\mathbf{x}_1^0),\mathbf{x}_1^0],\dots,[r_{f(m)} + F(\mathbf{x}_m^0),\mathbf{x}_m^0]\}.$$

Definition Two permutations f' and f'' from $\widetilde{f}$ will be called equivalent if data sets $\widetilde{S}_p(f',F)$ and $\widetilde{S}_p(f'',F)$ are equal.

Let $\widetilde{f}_b = \{f_1^b,\dots,f_{\mathcal{N}}^b\}$ is such set of permutations that

- any two permutation from $\widetilde{f}_b$ are not equivalent,
- any permutation is equivalent to one of permutations from $\widetilde{f}_b$.

Let note that due to transitiveness of equivalence any permutation may be equivalent only one element from $\widetilde{f}_b$. Equivalence class $c(f)$ may be defined for each permutation from $\widetilde{f}_b$ that consists of all permutation that are equivalent to f. Equality

$$\widetilde{f} = \bigcup_{i=1}^{\mathcal{N}} c(f_j^b)$$

is true by definition of $\widetilde{f}_b$. Two statement are true.

Statement 1. In case supposition 2 is true for any $f_j \in \widetilde{f}_b$

$$\mathrm{P}\{\widetilde{S}_p[f_j,F] \mid \mathbf{x}_1 = \mathbf{x}_1^0,\dots,\mathbf{x}_m = \mathbf{x}_m^0\} = \prod_{i=1}^{m}\mathrm{P}(\xi_i = r_i)$$

Proof. Statement 1 may be easily received from independence of residuals r on vectors $\mathbf{x}$ and mutual independence of observations corresponding different objects from $\widetilde{S}_0$. It follows from supposition 2 that probabilities of data sets $\widetilde{S}_p(f_1,F),\dots,\widetilde{S}_p[f_{\mathcal{N}},F]$ are equal each other. Q.E.D.

Statement 2. All classes $c[f_1],\dots,c[f_N]$ are of the same size.

Proof. Really. Let $\{\widetilde{r}_1,\dots,\widetilde{r}_k\}$ be such partition of $\{r(1),\dots,r(m)\}$ that residuals r_* inside each element of partition are equal each other and residuals from different groups are different. Suppose that $\widetilde{J}_q = \{J_q(1),\dots,J_q[\mu(q)]\}$ is set of residuals numbers inside group $\widetilde{r}_q$ according some permutation $f_j \in \widetilde{f}_b$, where $\mu(q)$ is size of group $\widetilde{r}_q$ and $q = 1,\dots,k$. It is evident that for any permutation f_j' that is received from f_j by some permutations of numbers only inside sets $\widetilde{J}_1,\dots,\widetilde{J}_k$ equality of data sets $\widetilde{S}_p(f_j,F)$ and $\widetilde{S}_p(f_j',F)$ is preserved. At that for any permutation

f_j'' that is received from f_j by some permutation including exchanges between sets $\widetilde{J}_1,\ldots,\widetilde{J}_k$ data sets $\widetilde{S}_p(f_j,F)$ and $\widetilde{S}_p(f_j'',F)$ are not equal. So class $c(f_j)$ must include all permutations that are received from f_j by some permutations of numbers inside sets $\widetilde{J}_1,\ldots,\widetilde{J}_k$. Class $c(f_j)$ does not include any other permutations. Let note that amount of such permutations depends only on sizes of groups $\{\widetilde{r}_1,\ldots,\widetilde{r}_k\}$ and does not depend on specific permutation f_j. So size of class $c[f_j]$ does not depend on f_j. Q.E.D.

Set $\widetilde{\mathbf{S}}_b = \{\widetilde{S}_p(f_1,F),\ldots,\widetilde{S}_p(f_{\mathcal{N}},F)\}$ includes all possible data sets $\widetilde{S}$ satisfying conditions
a) empirical distribution of residuals r from forecasting function F in $\widetilde{S}$ coincides with empirical distribution of residuals r at initial data set $\widetilde{S}_0$ (condition $C_r(\widetilde{S}_0,F)$);
b) $\mathbf{x}$-descriptions in $\widetilde{S}$ completely coincide with $\mathbf{x}$-descriptions of $\widetilde{S}_0$(condition $C_{\mathbf{x}}(\widetilde{S}_0,F)$).
Let $\mathcal{P}$ is some predicate that is defined at set of all possible data sets of size m. Let predicate $\mathcal{P}$ be true at some subset $\widetilde{\mathbf{S}}_T(\mathcal{P})$ of set $\widetilde{\mathbf{S}}_b$. Probabilities of all data sets from $\widetilde{\mathbf{S}}_b$ are equal according statement 2. So equality $P\{\mathcal{P}(\widetilde{S}) = TRUE|\widetilde{S}\in\widetilde{\mathbf{S}}_b\}$ may be evaluated as ratio $\frac{|\widetilde{\mathbf{S}}_T(\mathcal{P})|}{|\widetilde{\mathbf{S}}_b|}$.

Supposition 4. Let $\mathcal{P}_{pv} = Q_{min}(\widetilde{S},\widetilde{M}_c) < Q_{min}(\widetilde{S}_0,\widetilde{M}_c)$. It is suggested to use conditional probability

$$P\{\mathcal{P}(\widetilde{S}) = TRUE|\widetilde{S}\in\widetilde{\mathbf{S}}_b\} = \frac{|\widetilde{\mathbf{S}}_T(\mathcal{P}_{pv})|}{|\widetilde{\mathbf{S}}_b|} \quad (1)$$

as p-value that evaluates validity of null hypothesis about exhaustiveness.

Statement 3. Equality is true

$$\frac{|\widetilde{\mathbf{S}}_T(\mathcal{P}_{pv})|}{|\widetilde{\mathbf{S}}_b|} =$$

$$= \frac{|f\in\widetilde{f}|Q_{min}[\widetilde{S}_p(f,F),\widetilde{M}_c] < Q_{min}(\widetilde{S}_0,\widetilde{M}_c)|}{|\widetilde{f}|} \quad (2)$$

Proof. Really, evidently

$$\mathcal{P}]\widetilde{S}_p(f,\widetilde{S}_0)) = \mathcal{P}[\widetilde{S}_p(f',\widetilde{S}_0)]$$

if f is equivalent f'. According statement 2 all equivalence classes are of the same size. Let n_c be number of permutations in each equivalence class. Then

$$\frac{|\widetilde{\mathbf{S}}_T(\mathcal{P}_{pv})|*n_c}{|\widetilde{\mathbf{S}}_b|*n_c} =$$

$$= \frac{|f\in\widetilde{f}|Q_{min}(\widetilde{S}_p(f,F),\widetilde{M}_c) < Q_{min}(\widetilde{S}_0,\widetilde{M}_c)\}|}{|\widetilde{f}|}.$$

Q.E.D.
Thus ratio

$$\frac{|\{f\in\widetilde{f}|Q_{min}[\widetilde{S}_p(f,F),\widetilde{M}_c] < Q_{min}(\widetilde{S}_0,\widetilde{M}_c)\}|}{|\widetilde{f}|} \quad (3)$$

theoretically allows to calculate exact p-value testing validity of null hypothesis about exhaustive description of existing regularity by simple regularity from $\widetilde{M}_s$. But practically it is impossible to calculate exact p-values because of huge amount of possible permutation. However it is easily to estimate 3 using relatively small number of random permutations that are generated by random numbers generator. Let

$$\widetilde{f}_g = \{f_j|j=1,\ldots,N_g\}$$

be set of permutations calculated by by random numbers generator. Then p-value may be estimated as ratio

$$\frac{|\{f_j\in\widetilde{f}_g|Q_{min}(\widetilde{S}_p(f,F),\widetilde{M}_c) < Q_{min}(\widetilde{S}_0,\widetilde{M}_c)\}|}{N_g} \quad (4)$$

2.3 Choice of Simple Model

Technique described in previous subsection may be used only if simple model from $\widetilde{M}_s$ has been previously chosen. Supposition 1 declares that complicated model must not be used when there is simple model that exhaustively describes data. Such model may be searched by evaluating all predicting functions from $\widetilde{M}_s$ with the help of described in previous section PT version. But it is practically impossible to implement such approach. In this paper only two simple predicting functions from $\widetilde{M}_s$ are evaluated. At first simple predicting function is studied that is searched with the help of standard least squares technique. It is naturally to hope that in many task LS regression is very close to a model that exhaustively describes data. However experiments with optimal valid partitioning method (Senko and Kuznetsova, 2006) demonstrated that really false complicated regularity $\mathcal{R}_c$ may be mistakenly evaluated as valid. Such mistakes take place when regularity are verified relatively simple regularity $\mathcal{R}_s$ that in the best way approximate data. But at that $\mathcal{R}_s$ significantly deviates from verified complicated regularity $\mathcal{R}_c$. So a method was developed in (Kuznetsova et al., 2013) verifying more complicated model $\mathcal{R}_c$ relatively simple model that minimally deviates from $\mathcal{R}_s$.
Let try to explain why such technique may be useful. Suppose that $F_s(\mathbf{x})$ is some predicting function from $\widetilde{M}_s$, $F_c^o(\mathbf{x})$ is $arg\min_{F(\mathbf{x})\in\widetilde{M}_c} Q[\widetilde{S}_0,F(\mathbf{x})]$,

$\delta(j) = F_s(\mathbf{x}_j) - F_c^o(\mathbf{x}_j)$.
Discussed approach is based on evaluating upper boundary of $Q_{min}[\widetilde{S}_p(f,F_s),\widetilde{M}_c]$ where $f \in \widetilde{f}$.

But by definition of $\widetilde{S}_p(f,F_s)$

$$Q_{min}[\widetilde{S}_p(f,F_s),\widetilde{M}_c] < Q[\widetilde{S}_p(f,F_s),F_c^o] =$$

$$= \sum_{j=1}^{m} [r_{f(j)} + F_s(\mathbf{x}_j) - F_c^o(\mathbf{x}_j)]^2 = \sum_{j=1}^{m} [r_{f(j)} + \delta(j)]^2 =$$

$$= \sum_{j=1}^{m} r_{f(j)}^2 + 2\sum_{j=1}^{m} \delta(j) r_{f(j)} + \sum_{j=1}^{m} \delta^2(j).$$

On another hand

$$Q_{min}(\widetilde{S}_o,\widetilde{M}_c) = Q(\widetilde{S}_o,F_c^o) =$$

$$= \sum_{j=1}^{m} [y_j - F_c^o(\mathbf{x}_j)]^2 = \sum_{j=1}^{m} [y_j - F_s(\mathbf{x}_j) +$$

$$+ F_s(\mathbf{x}_j) - F_c^o(\mathbf{x}_j)]^2 = \sum_{j=1}^{m} [r_j + \delta(j)]^2 =$$

$$\sum_{j=1}^{m} r_j^2 + 2\sum_{j=1}^{m} \delta(j) r_j + \sum_{j=1}^{m} \delta^2(j).$$

Taking into account that

$$\sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} r_{f(j)}^2$$

we receive that

$$Q[\widetilde{S}_p(f,F_s),F_c^o] - Q_{min}(\widetilde{S}_o,\widetilde{M}_c) =$$

$$= 2\sum_{j=1}^{m} \delta(j)[r_{f(j)}) - r_j] \le 2\sum_{j=1}^{m} |\delta(j)| \cdot |[r_{f(j)}) - r_j]|.$$

Thus upper bound for $Q_{min}[\widetilde{S}_p(f,F_s),\widetilde{M}_c]$ tends to $Q_{min}[\widetilde{S}_o,\widetilde{M}_c]$ as $\max_{j=1,\ldots,m}|\delta_j|$ tends to 0. It is more probable that inequality

$$Q_{min}[\widetilde{S}_p(f,F_s),\widetilde{M}_c] < Q_{min}[\widetilde{S}_o,\widetilde{M}_c]$$

is true when $\max_{j=1,\ldots,m}|\delta_j|$ is small. So we may hope that p-value that is calculated by ratios 4 will be greater when $\max_{j=1,\ldots,m}|\delta_j|$ is small. Thus small p-value received when F_o^c is verified relatively closest simple model is strong argument for absence of simple model from $\widetilde{M}_s$ that cannot be rejected using complicated model. Existence of such argument corresponds to Supposition 1 correctness.

3 APPLICATION EXAMPLE

3.1 Objectives

Effect of vitamin D status(vitD) on parathyroid hormone (PTH) concentration was studied (Kim et al., 2012). Now serum 25 (OH) D is the best indicator of the (vitD), but target levels of vitamin D in the blood are still represent a matter of debate. So the priority arrears of the research are the development of a method-dependent reference values with the use of biomarkers for vitD sufficiency. One such widely recognized biomarker is the correlation of vitD with PTH. But supposition exists that vitD correlates with PTH only when vitD concentration is less than certain threshold level and there is correlation "loss" when vitD concentration is higher than threshold level. Goal of our research was statistical verification of last supposition and search of optimal model that describes dependence of PTH on vitD. It must be noted that discussed supposition corresponds to use of piecewise-linear model.

3.2 Data Set

The study included patients (n = 139, males 18%, mean age 48,5 $\pm$ 18 years) in which levels of total 25(OH)D (LIAISON, DiaSorin) and PTH (ELECSYS, Roche) were measured during autumn period (September-October). In selection of patients we used exclusion criteria: presence of primary hyperparathyroidism, secondary or tertiary hyperparathyroidism on the background terminal chronic renal failure, blood creatinine level of more than 100 mmol/l or GFR less than 60 ml/min/1,73m2, intake of active vitD metabolites within 1 month prior the blood test.

3.3 Search of Optimal Regression

It is supposed that response variable Y is predicted by variable X with the help of piecewise-linear model with 2 segments

$$Y = \beta_0^l + \beta_1^l X + \varepsilon_l, when X \le B$$
$$Y = \beta_0^r + \beta_1^r X + \varepsilon_r, when X \ge B \quad (5)$$

At that it is supposed that

$$\beta_0^l + \beta_1^l B = \beta_0^r + \beta_1^r B. \quad (6)$$

Let $\widetilde{M}_{pwl}^B$ be family of all piecewise-linear predicting functions with 2 segments and fixed B . For each B regression coefficients $\beta_0^l, \beta_1^l, \beta_0^r, \beta_1^r$ are calculated from observations

$$(y_1,x_1),\ldots,(y_m,x_m)$$

with the help of standard least squares technique. It is evident that search of coefficients may be reduced to of quadratic programming task:

$$Q(\widetilde{S}_0,\widetilde{M}^B_{pwl}) = \sum_{x_j<B}(y_j-\beta^l_0-\beta^l_1 x_j)^2 + \\ + \sum_{x_j<B}(y_j-\beta^r_0-\beta^r_1 x_j)^2 \to \min, \quad (7)$$

when constraint (6) is satisfied. Partial derivatives of Lagrange function

$$Q(\widetilde{S}_0,\widetilde{M}^B_{pwl}) + \lambda(\beta^l_0+\beta^l_1 B-\beta^r_0-\beta^r_1 B)$$

by coefficients $\beta^l_0,\beta^l_1,\beta^r_0,\beta^r_1$ must be equal 0 for the task (7). Let $\widetilde{X}_m = \{x_1,\ldots,x_m\}$. Using 4 equalities for partial derivatives and constraint (6) we receive system of 5 equations.

$$\begin{cases} m_l\beta^l_0+\overline{X}_l\beta^l_1-\frac{1}{2}\lambda=\overline{Y}_l \\ \overline{X}_l\beta^l_0+d^l_x\beta^l_1-\frac{B}{2}\lambda=c^l_{xy} \\ m_r\beta^r_0+d^r_x\beta^r_1+\frac{1}{2}\lambda=\overline{Y}_r \\ \overline{X}_r\beta^r_0+d^r_x\beta^r_1+\frac{B}{2}\lambda=c^r_{xy} \\ -\beta^l_0-\beta^l_1 B+\beta^r_0+B\beta^r_1=0 \end{cases} \quad (8)$$

where

- m_l is number of points in $\widetilde{X}_m$ satisfying inequality $x_j < B$,
- m_r is number of points in $\widetilde{X}_m$ satisfying inequality $x_j > B$,
- $d^l_x = \sum_{x_j<B} x_j$, $\overline{X}_r = \sum_{x_j>B} x_j$,
- $c^l_{xy} = \sum_{x_j<B} x_j y_j$, $c^r_{xy} = \sum_{x_j>B} x_j y_j$,
- $\overline{X}_l = \sum_{x_j<B}(x_j)^2$, $\overline{X}_r = \sum_{x_j>B}(x_j)^2$,

Optimal regression coefficients belongs to solution of system (8). Let

$$\widetilde{X}_c = \{x^c_{j'j''} = \frac{x_{j'}+x_{j''}}{2} | x_{j'} \neq x_{j''}, x_{j'} \in \widetilde{X}_m, x_{j''} \in \widetilde{X}_m\}$$

be a set of boundaries separating neighbour points from $\widetilde{X}_m$. To find LS piecewise-linear regression it is sufficient to calculate $Q(\widetilde{S}_0,\widetilde{M}^B_{pwl})$ for all boundary points from $\widetilde{X}_c$ and to select boundary corresponding to minimal $Q(\widetilde{S}_0,\widetilde{M}^B_{pwl})$.

3.4 Data Analysis Results

Let x_{vd} be concentration of serum 25(OH)D, y_{ph} be concentration of PTH, $y_{lph} = \log y_{ph}$. Optimal piecewise-linear regressions calculating y_{ph} and y_{lph} were chosen in $\widetilde{M}_{pwl}$ with the help of technique described in previous section. Optimal boundary point B was equal $23.95\frac{ng}{ml}$ for model predicting y_{ph} from x_{vd} (task I) and $B = 24.7\frac{ng}{ml}$ for piecewise-linear regression predicting y_{lph} from x_{vd} (task II). Dependence of $Q(\widetilde{S}_0,\widetilde{M}^B_{pwl})$ on B in task I is given at figure 1.

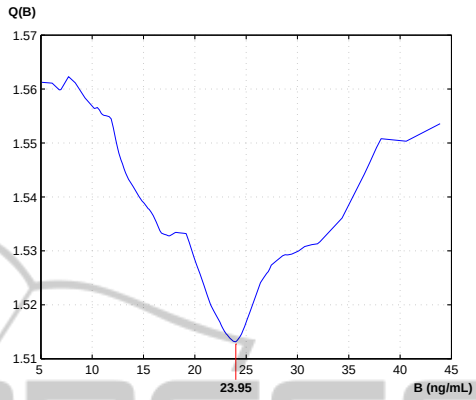

Figure 1: Dependence $Q(\widetilde{S}_0,\widetilde{M}^B_{pwl})$ on B in task I.

It is seen from figure 1 that point 23.95(ng/ml) corresponds unique expressed global minimum of $Q(\widetilde{S}_0,\widetilde{M}^B_{pwl})$. Graphic of piecewise-linear function from model I is represented at figure 2. It is seen

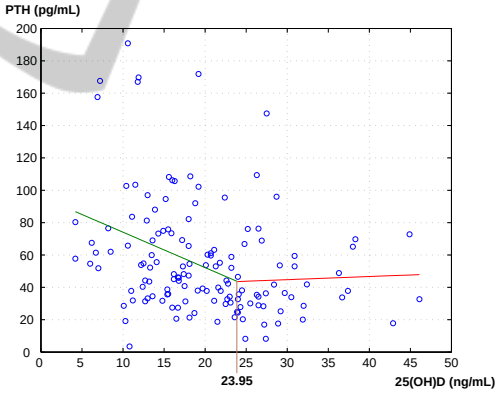

Figure 2: Optimal piecewise regression for task I.

that slope of linear predicting function inside left segment significantly exceeds slope of linear predicting function inside right segment. Correlation coefficient between y_{ph} and x_{vd} in group of patients with $x_{vd} <$ 23.95 is equal -0.2934 (significant at p¡0.01). Correlation coefficient in group of patients with $x_{vd} > 23.95$ is close to zero (equal 0.0351). Such results are in good agreement with supposition that vitD correlates with PTH only when vitD concentration is less than certain threshold level. However statistical significance of such correlation analysis is not too high because correlation coefficients are calculated in groups formed by boundary B that was previously found by the same data set. Let try to validate result with the

help of procedures verifying complicated models relatively simple models that were discussed in previous sections.

3.5 Verification

At the first stage null hypothesis about independence of y_{ph} on x_{vd} was tested with the help previously discussed in (Senko and Kuznetsova, 2006) permutation test version. Set of random permutations of integers $1,\ldots,m$ was formed with the help of random numbers generator. This set $\widetilde{f}_{rng}$ consisted of N_g elements. Data sets $\{\widetilde{S}_p(f_j)|f_j \in \widetilde{f}_{rng}\}$ was built from $\widetilde{S}_0$ by random permutation of y_{ph} positions relatively fixed positions of x_{vd}. Statistical validity of null hypothesis is evaluated with the help of p-value that is equal ratio

$$\frac{|\{f_j \in \widetilde{f}_{rng}|Q_{min}[\widetilde{S}_p(f_j),\widetilde{M}_{pwl}] < Q_{min}(\widetilde{S}_0,\widetilde{M}_{pwl})\}|}{N_g}.$$

In other words p-value is calculated as fraction of random data sets where dependence of y_{ph} on x_{vd} is approximated better than at initial set $\widetilde{S}_0$. Values $Q_{min}(\widetilde{S}_0,\widetilde{M}_{pwl})$ and $Q_{min}[\widetilde{S}_p(f_j),\widetilde{M}_{pwl}]$ are calculated with the help of procedure that is describe in section 3.3. Piecewise-linear modeling of y_{ph} from x_{vd} allows to reject null hypothesis with p-value equal 0.000041. Piecewise-linear modeling of y_{lph} from x_{vd} allows to reject null hypothesis with p-value equal 0.000079. At that number of random permutations was equal 10^6. Then piecewise-linear regressions were verified relatively simple regression models. Optimal piecewise-linear regression $y_{ph} = F^o_{pwl}(x_{vd}) + \varepsilon_{pw}$ was verified by testing null hypothesis about exhaustive description of dependence by simple linear regression $y_{ph} = \alpha_0 + \alpha_1 x_{vd} + \varepsilon_1$. Piecewise-linear regression $y_{lph} = F^o_{pwl} x_{vd} + \varepsilon_{pw}$ was verified by testing null hypothesis about exhaustive description of dependence by simple linear regression $y_{lph} = \alpha^l_0 + \alpha^l_1 x_{vd} + \varepsilon_2$.

Two ways of regression coefficients $\alpha_0, \alpha^l_0, \alpha_1, \alpha^l_1$ calculating were considered:

- simple regression coefficients were searched with the help of standard LS procedure,
- such simple regression coefficients were chosen that distance between verified piecewise-linear regression and simple regression was minimal.

Let suppose that x values in $\widetilde{S}_0$ belong to some interval (a_l, a_h). Then distance between two predicting functions $F_1(x)$ and $F_2(x)$ is calculated by formula

$$D[F_1(x),F_2(x)] = \int_{a_l}^{a_h} [F_1(x) - F_2(x)]^2 dx.$$

Ratio (4) was used to estimate p-values. At that number of permutations was equal 10^6. Results of verification are represented in table.

Table 1: Results of verification.

target	type of symple model	p-value
y_{ph}	standard LS	0.022
y_{lph}	standard LS	0.026
y_{ph}	most close to F^o_{pw}	0.015
y_{lph}	most close to F^o_{pw}	0.0218

It is seen from table that p-values for null hypotheses about exhaustive description of data by simple regressions do not exceed 0.026. This result is strong argument that simple regressions are not sufficient to explain data and more complicated piecewise-linear regression models are really necessary. Thus supposition that vitD correlates with PTH only when vitD concentration is less than certain threshold level is statistically valid.

4 CONCLUSIONS

So method was proposed that allows to evaluate validity of choice between simple or complicated regression models in terms of p-values. Method is based on testing null hypothesis about independence of deviations from simple predicting function on X variables. Method was successfully used for evaluating correctness of biomedical supposition that vitamin D status correlates with parathyroid hormone levels. Method may be used in variety of tasks where a problem of choice between more complicated or simple models.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, vol.19, iss.6,:pp. 716–723.

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, vol. 3, iss. 4:pp. 409–425.

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685.

Golland, P. et al. (2000). Permutation test for classification. *Journal of Machine Learning Research*, 1.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Science+Business Media, Inc.

Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B (Methodological)*, vol.41:pp. 190–195.

Kim, G. et al. (2012). Relationship between vitamin d, parathyroid hormone, and bone mineral density in elderly koreans. *J Korean Med Sci*.

Kim, H.-J. et al. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statist. Medicine*, vol.19.

Kuznetsova, A. et al. (2013). Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. *Pattern Recognition and Image Analysis*, 22(4):10–25.

Ojala, M. and Garriga, G. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, vol. 14, iss. 5:pp. 465–658.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, vol. 6:pp. 461–464.

Senko, O. and Kuznetsova, A. (2006). The optimal valid partitioning procedures. *InterStat, Statistics in Internet,http://ip.statjournals.net*.