Nonparametric Regression for Analyzing Correlation between Medical Parameters

Malgorzata Charytanowicz[1] and Piotr Kulczycki[2]

[1] Polish Academy of Sciences, Systems Research Institute, ul. Newelska 6, PL-01-447 Warsaw malgorzata.charytanowicz@ibspan.waw.pl
[2] Polish Academy of Sciences, Systems Research Institute, ul. Newelska 6, PL-01-447 Warsaw kulczycki@ibspan.waw.pl

Summary. In this paper, the use of nonparametric regression is applied for analyzing the correlation of hematologic parameters with creatinine. The sample population involved patients with renal insufficiency observed in the Stefan Kardynal Wyszynski Regional Specialists' Hospital in Lublin (Poland). The method presented here is based on the theory of statistical kernel estimators, which frees it of assumptions in regard to the form of regression function. This approach is universal, and can be used in medical parameter regression analysis, where arbitrary assumptions concerning the form of regression function are not recommended.

1 Introduction

Chronic renal failure (CRF) is characterized by a slow, progressive decline in glomerular filtration rate which increasingly effects other kidney functions. Progression of renal disease is associated with development of anemia. A negative correlation between creatinine and hematologic parameters (hemoglobin, hematocrit, red blood cells) exists. Inhibitors of erythropoiesis have been suggested to be important in the pathogenesis of anemia. Inadequate erythropoiesis occurs because the quantity of endogenous erythropoietin produced by the peritubular fibroblasts in the kidney, is insufficient in relation to the degree of anemia. Patients with CRF are unable to increase erythropoiesis sufficiently to compensate blood loss [6, 7]. Correction of anemia becomes possible due to the availability of recombinant human erythropoietin (rHuEPO) [9].

The goal of this paper is to provide a method allowing the analysis the relationship of the hematological parameters, performed separately, and creatinine. Creatinine was chosen as a marker of renal insuffiency. The data was derived from patients with renal insuffiency. The proposed method allows determine the stage of renal failure indicated the beginning of anemia. The mathematical apparatus relies on the theory of statistical kernel estimators [4, 5, 8], which frees the method from the types of regression functions.

2 Materials and methods

The sample population involved patients with renal insufficiency who had not received rHuEPO, observed in the Stefan Kardynal Wyszynski Regional Specialists' Hospital in Lublin (Poland), for periods up to five years (1994-1998). During this time, determinations of creatinine, as well as other biochemical and hematological parameters, were done routinely. The study is composed of 1915 determinations in the creatinine range from 1.1 to 5.0 mg/dL. The normal range for creatinine values equals 0.7-1.4 mg/dL. The number of observations with creatinine exceeded normal values equals 908. The mean, standard deviation and median values of standard blood parameters, including hemoglobin, hematocrit and red cells for this group are shown in Table 1.

Table 1. Mean (±S.D.) and median values of selected hematologic parameters for creatinine range ≤ 5.0 mg/dL

	Male ($n = 534$)		Female ($n = 374$)	
Parameters	Mean ± S.D.	Median	Mean ± S.D.	Median
Hemoglobin [g/dL]	13.28 ± 2.63	13.50	11.91 ± 1.93	11.81
Hematocrit [%]	38.37 ± 7.57	38.70	34.43 ± 5.43	34.25
Red Blood Cells [$\times 10^{12}$/L]	4.33 ± 0.87	4.37	3.84 ± 0.61	3.84

The mean ± S.D. of hemoglobin concentrations were 13.28 ± 2.63 g/dL with the median of 13.50 g/dL for males, and 11.91 ± 1.93g/dL with the median of 11.81 g/dL for females. The mean ± S.D. of hematocrit were 38.37 ± 7.57 % with the median of 38.70 % for males, and 34.43 ± 5.43% with the median of 34.25 % for females. The mean ± S.D. of red blood cells count were 4.33 ± 0.87 [$\times 10^{12}$/L] with the median of 4.37 [$\times 10^{12}$/L] for males, and 3.84 ± 0.61 [$\times 10^{12}$/L] with the median of 3.84 [$\times 10^{12}$/L] for females. In fact, all of the following parameters are lower than the normal ranges given in Table 2.

Table 2. Normal range of selected hematologic parameters

Parameter	Male	Female
Hemoglobin [g/dL]	14.00–18.00	12.00–16.00
Hematocrit [%]	42.00–52.00	37.00–47.00
Red Blood Cells [$\times 10^{12}$/L]	4.70–6.10	4.20–5.40

In treated patients, there was a significantly negative correlation between the hemoglobin concentration and creatinine ($r = -0.45, p < 0.0001$ for males, $r = -0.49, p < 0.0001$ for females). The correlation was significantly negative

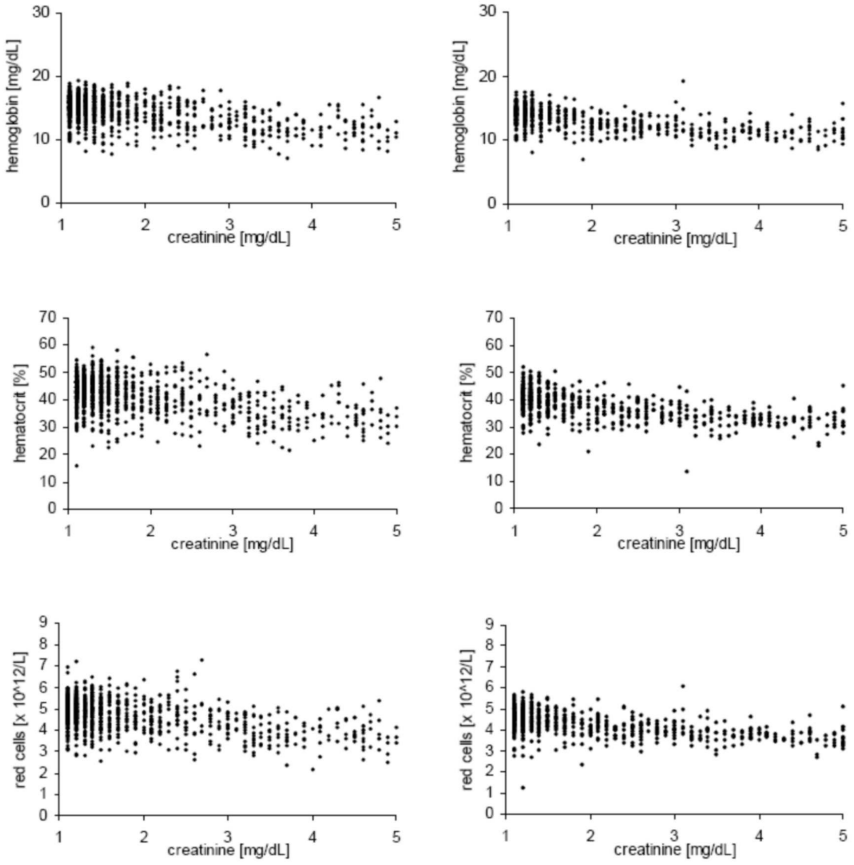

Fig. 1. Scatterplot of creatinine and selected hematologic parameters for males and females

between hematocrit and creatinine ($r = -0.46, p < 0.0001$ for males, $r = -0.49, p < 0.0001$ for females) and also between the red blood cells count and creatinine ($r = -0.49, p < 0.0001$ for males, $r = -0.50, p < 0.0001$ for females). The rate of hemoglobin, hematocrit and red blood cells decreases when creatinine increases. Scatter diagrams given on Figure 1 shows selected hematologic parameters as functions of creatinine, separately for males (first column) and females (second column).

During early renal failure (creatinine < 5 mg/dL) hematological parameters have often been represented as linearly related to creatinine [2]. The kernel-based nonparametric regression estimators give a much more flexible family of curves to choose from.

3 Nonparametric regression

3.1 Kernel regression

Classical parametric methods of determining an appropriate functional relationship between the two variables imposed arbitrary assumptions concerning the functional form of the regression function. The choice of parametric model depends very much on the situation. If a chosen parametric family is not of appropriate form, then there is a danger of reaching incorrect conclusions in the regression analysis. This also makes it difficult to take into account the whole accessible information. The rigidity of this regression can be overcome by removing the restriction that the model is parametric. This approach leads to nonparametric regression that let the data decide which function fits them best. In this study, a class of kernel-type regression estimators called *local polynomial kernel estimators* is presented.

Let therefore, n elements $(x_i, y_i) \in \mathrm{R} \times \mathrm{R}$, $i = 1, 2, \ldots, n$ be given, where values x_i may designate some non-random numbers or realizations of the one-dimensional random variable X, whereas y_i designate realizations of the one-dimensional random variable Y. Assuming the existence of the function $f : \mathrm{R} \to \mathrm{R}$ having a continuous first derivative that:

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

where ε_i are independent random variables with zero mean and unit, finite variance. Let then $p \in \mathrm{N}$ be the degree of the polynomial being fit. The kernel regression estimator $\hat{f} : \mathrm{R} \to \mathrm{R}$, obtained by using weighted least squares with kernel weights, is given by the formula:

$$\hat{f}(x) = e_1^T (X^T W X)^{-1} X^T W y, \tag{2}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{3}$$

is the vector of responses,

$$X = \begin{bmatrix} 1 & x_1 - x & (x_1 - x)^2 & \ldots & (x_1 - x)^p \\ 1 & x_2 - x & (x_2 - x)^2 & \ldots & (x_2 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & (x_n - x)^2 & \ldots & (x_n - x)^p \end{bmatrix} \tag{4}$$

is an $n \times (p+1)$ design matrix,

$$W = diag\left(\frac{1}{h} K\left(\frac{x_1 - x}{h}\right), \frac{1}{h} K\left(\frac{x_2 - x}{h}\right), \ldots, \frac{1}{h} K\left(\frac{x_n - x}{h}\right)\right) \tag{5}$$

is an $n \times n$ diagonal matrix of kernel weights,

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{6}$$

is the $(p+1)\times 1$ vector having 1 in the first entry and zero elsewhere. The coefficient $h > 0$ is called a bandwidth, while the measurable function K :R$\rightarrow$[0,∞) of unit integral, symmetrical with respect to zero and having a weak global maximum in this place, takes the name of the kernel.

The choice of the kernel form has no practical meaning and thanks to this, it is possible to take into account the primarily properties of the estimator obtained. Most often the standard normal kernel expressed by a convenient analytical formula:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{7}$$

is used.

The practical implementation of the kernel regression estimators requires a good choice of bandwidth. If h is too small, a spiky rough kernel estimate is obtained, and if h is too large, it results a flat kernel estimate. A frequently used bandwidth selection technique is the cross-validation method [1, 3], which chooses h to minimize

$$\sum_{i=1}^{n} \left(y_i - \hat{f}_{-i}(x_i) \right)^2 \tag{8}$$

where $\hat{f}_{-i}(\cdot)$ denotes the regression estimator (2), without using the ith observation (x_i, y_i).

An important problem is the choice of the parameter p. For sufficiently smooth regression functions, the asymptotic performance of $\hat{f}$ improves for higher values of p. However, for higher p, the variance of the estimator becomes larger and in practice, a very large sample may be required. On the other hand, the even degree polynomial kernel estimator has a more complicated bias expression which does not lend itself to simple interpretation. These facts suggests the use of either $p = 1$ or $p = 3$. Moreover, for $p = 1$, the convenient explicit formulae exists:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{\left(\hat{s}_2(x) - \hat{s}_1(x) \left(x_i - x \right) \right) y_i K\left(\frac{x - x_i}{h} \right)}{\hat{s}_2(x)\hat{s}_0(x) - \left(\hat{s}_1(x) \right)^2} \tag{9}$$

where

$$\hat{s}_r(x) = \frac{1}{nh} \sum_{i=1i}^{n} \left(x_i - x \right)^r K\left(\frac{x_i - x}{h} \right) r = 0, 1, 2. \tag{10}$$

Therefore - except in more advanced statistical applications - $p = 1$ is preferred. The tasks concerning the choice of the kernel form, the bandwidth, as

well as additional procedures improving the quality of the estimator obtained, are found in [4, 5, 8]. The utility of local linear kernel estimators has been investigated in the context of some typical data derived from patients with renal insuffiency.

3.2 Results

In this study, the method of nonparametric regression based on a weighted local linear regression was used to analyze the relationship between creatinine and selected hematological parameters (hematocrit, hemoglobin and red cells). For ease of computation, the standard normal kernel (7) was used. The bandwidth was determined using the cross-validation method (8). The results were compared with results obtained for two parametric models:

- linear regression model [2]:

$$y = b_0 + b_1 x \tag{11}$$

- logarithmic regression model:

$$y = b_0 + b_1 log x \tag{12}$$

having the best fit to the data in the family of nonlinear functions.

The Mean Squared Error MSE

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13}$$

precisely, the Root of the Mean Squared Error RMSE was reported in comparison with the weighted local linear regression (9). Table 3 shows estimation errors RMSE obtained for these three considered regression models describing the relationship of hemoglobin, hematocrit and red cells to creatinine, separately for males and females.

Table 3. Comparing of the estimation errors for selected regression functions describing the relationship of hematological parameters (hemoglobin, hematocrit, red blood cells) to creatinine for males ($n = 1239$) and females ($n = 676$).

Male		Female	
Regression Equation	*RMSE*	Regression Equation	*RMSE*
Hemoglobin *HB* [g/dL]			
$HB = 16.75 - 1.20 \cdot creatinine$	1.77	$HB = 15.08 - 0.95 \cdot creatinine$	1.41
$HB = 15.84 - 2.63 \cdot \log(creatinine)$	1.77	$HB = 14.45 - 2.27 \cdot \log(creatinine)$	1.44
kernel estimator (9)	1.73	kernel estimator (9)	1.35

Hematocrit HT [%]		
$HT = 47.38 - 3.14 \cdot creatinine$	5.23 $HT = 43.04 - 2.56 \cdot creatinine$	4.10
$HT = 44.98 - 6.82 \cdot \log(creatinine)$	5.24 $HT = 41.34 - 6.06 \cdot \log(creatinine)$	4.05
kernel estimator (9)	5.12 kernel estimator (9)	3.92
Red Blood Ceels RBC [$\times 10^9$/L]		
$RBC = 5334.48 - 347.01 \cdot creatinie$	6.15 $RBC = 4857.05 - 274.40 \cdot creatinine$	4.89
$RBC = 5067.61 - 750.08 \cdot \log(creatin)$	6.17 $RBC = 4671.56 - 641.49 \cdot \log(creatin)$	4.86
kernel estimator (9)	6.02 kernel estimator (9)	4.69

The RMSE for both the linear and logarithmic functions are comparable. The mean square error for males is a bit greater than for females. The local linear estimator gives the lowest mean squared errors and the difference is in the range from 2% to 4%.

Further analysis was performed using kernel estimators of regression. The functional form of the relationship of all dependent variables was determined as a function of creatinine. Using obtained forms, the values of creatinine corresponding to the values less than the lower limit of normal ranges of hemoglobin concentrations, hematocrit and red blood cells count were calculated. The results, obtained separately for males and females, are given in Table 4:

Table 4. Calculated values of creatinine corresponding to the values less than the lower limit of normal ranges of selected hematologic parameters

	Creatinine [mg/dL]	Creatinine [mg/dL]
Parameter	Male	Female
Hemoglobina [g/dL]	2.11	2.76
Hematocrit [%]	2.02	2.32
Red Blood Cells [$\times 10^{12}$/L]	2.15	2.63

The analysis of various hematological parameters has allowed the claim that anemia in chronic renal failure develops when creatinine exceed 2 mg/dL for males and 2.3 mg/dL for females. The dispersion of all these parameters is greater in the male's group. The first step in the correction of renal anemia is confirmation of the diagnosis. This diagnosis should be done before rHuEPO treatment. If renal anemia appears, other causes of anemia such as blood loss, iron deficiency or malnutrition should be sought and corrected. Treatment decisions for patients with renal anemia should be recommended at an early stage, before symptoms appear.

4 Summary

Anemia is a common complication of chronic renal failure (CRF). Over fifty percent of patients with CRF die because of anemia. It has been reported that dialysis have no significant meaning in treatment of anemia in patients with renal disease. In eighty years, much progress has been made in the management of anemia due to the availability of recombinant human erythropoietin (rHuEPO). The clearest benefit of rHuEPO in CRF is a substantial reduction in transfusion dependency, which reduces the need for hospital admission and the risk of viral transmission. Treatment of anemia with rHuEPO has also been shown to improve cognitive function, socialization and quality of life in dialysis patients and in consequence, is required and beneficial. To obtain the maximum benefits of early anemia correction, physical training and diet are necessary.

The aim of this study was to analyze the relationship between hematological parameters and stage of kidney disease. A negative correlation between creatinine and hematocrit, creatinine and hemoglobin and also between creatinine and red cells was observed. The proposed procedure, based on the methodology of kernel estimators, has allowed the claim that anemia in renal insuffiency develops when creatinine exceed 2 mg/dL. The kernel regression method enabled better use of the available data and more sophisticated analysis.

References

1. Fan J (1992) Journal of the American Statistical Association 87: 998–1004
2. Hakim RM, Lazarus JM (1988) American Journal of Kidney Diseases XI: 238–247
3. Hardle W, Marron JS (1985) The Annals of Statistics 13:1465–1481
4. Kulczycki P (2005) Estymatory jadrowe w analizie systemowej. WNT, Warszawa
5. Kulczycki P, Charytanowicz M (2005) Applied Mathematics and Computer Science 15:393–404
6. Orlowski T (1992) Choroby nerek. PZWL, Warszawa
7. Poplawski A (1995) Przeglad Lekarski 52:78–79
8. Wand MP, Jones MC (1994) Kernel Smoothing. Chapman and Hall, London
9. Winearles CG, Oliver DO, Pippard MJ, Reid C, Downing MR, Cotes PM (1986) Lancet 2:1175–1182