Miminizing the Effects of Multicollinearity in Polynomial Regression of Age Relationships and Sex Differences in Serum Levels of Pregnelonone Sulfate in Healthy Subjects

Milan Meloun[2], Martin Hill[1], Helena Havlíková[1], Vladimír Pouzar[3]

[1]Institute of Endocrinology, Prague, Czech Republic

[2]Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic

[3]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Pregnenolon-sulfate (PregS) positively modulate the membrane N-methyl-D-aspartate (NMDA) receptors in CNS and periphery that are responsible for the permeability of calcium channels and the activation of neuronal function. In addition, these conjugates act as negative noncompetitive modulators of GABAA receptors that are responsible for the attenuation of neuronal excitability. Recently, the permeability of blood-brain barrier for PregS was found in rat, but the facts supporting this finding were known even before. Serum levels of PregS depend chiefly on age and on adrenal activity. The biased estimators based on the principal component regression PCR method avoiding multicollinearity problems are described. The purposes of this paper were to emphasize the importance of understanding the nature of any near-singularities in the data that might cause problems with the ordinary least squares regression, to described the algorithm of one biased regression method, the principal component regression. Several criteria for the selection of suitable bias are demonstrated.

Keyword(s): Biased linear regression; Mean quadratic error of prediction; Multicollinearity; Principal component regression PCR;

Introduction

Pregnenolon-sulfate (PregS) and dehydroepiandrosterone-sulfate (DHEAS) positively modulate the membrane N-methyl-D-aspartate (NMDA) receptors in CNS and periphery that are responsible for the permeability of calcium channels and the activation of neuronal function [1-3]. In addition, these conjugates act as negative noncompetitive modulators of GABAA receptors that are responsible for the attenuation of neuronal excitability [4-7]. Recently, the permeability of blood-brain barrier for PregS was found in rat [8], but the facts supporting this finding were known even before [9-12]. Serum levels of PregS depend chiefly on age and on adrenal activity [13-14]. Some authors [14] recommended the use of PregS with significant responsibility to ACTH stimulation as a marker of adrenal dysfunction in children. Taking together the complete information about the age dependence of PregS including the determination of reference limits in the age groups should be of interest not only in diagnostics of adrenal disorders but also in the diagnostics of some disturbances of central nervous system as anxiety-depressive syndrome [15] or the diseases connected with aging as Alzheimer disease. DePerretti [13] described the detail scan of PregS age dependence during childhood and adolescence, however the changes during adulthood were not investigated. A detailed study covering the changes of the steroid conjugate during all life span in the both sexes was not published so far. We have attempted to supplement this lack evaluating the age and sex relationships of PregS in the both sexes that covers the age span from 4 to 70 years of age using principal component regression. Besides the exact evaluation of the age and sex relationships of PregS we have tried to introduce a correct approach for the analysis of age dependencies of the hormones, which are mostly represented by the data with non-Gaussian distribution, non-constant variance. In addition, multicollinearity is currently present in ordinary polynomial regression models that are currently used for the description of the age dependence. The difficulties mentioned above could lead to complete misinterpretation of the data when using the ordinary polynomial regression. Accordingly, we have proposed a principal component polynomial regression on the data transformed by power transformation as an appropriate method enabling to cope with the problems mentioned above.

In this paper the estimators based on generalized principal components of transformed data by power transformation in comparison with the Box-Cox transformation are adopted. For suitable bias selection the criterion based on the *MEP,* R^2_P and *AIC* are preferred.

Methodology

The polynomial linear regression model with *n* observations of *m*-th order polynomial variable and for an additive model of measurements errors is assumed, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Vector ***y*** has dimensions ($n \times 1$) and matrix ***X*** has dimensions ($n \times m$). Random errors ε_i in dependent variable *y* should have a normal distribution $N(0, \sigma^2)$. When the least squares assumptions are valid, the parameter estimates ***b*** found by minimization of the sum of squared residuals *RSS*

$$RSS = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{m} x_{ij}\, b_j \right]^2 \approx \text{minimum} \qquad (1)$$

are the best linear unbiased estimators (BLUE), [16-17]. The convential least-squares estimator ***b*** has the form $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ with the corresponding variance $D(\boldsymbol{b}) = \sigma^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}$. However, some difficulties arise when the matrix $\boldsymbol{X}^T\boldsymbol{X}$ appears to be singular. In some cases, especially with polynomial models, the parameter estimates may be without physical meaning. The multicollinearity problem in regression refers to the set of problems created when there are near-singularities occurs among the columns of the ***X*** matrix and certain linear combinations of the columns of ***X*** are nearly zero [18].

(a) The condition number $K = \lambda_{\max}/\lambda_{\min}$ contains $\lambda_{\max}$ and $\lambda_{\min}$ the largest and the smallest eigenvalues of a matrix ***R*** [Belsey]. The condition number provides a measure of the sensitivity of the solution to the normal equations to small changes in ***X*** or ***y***. A large condition number indicates that a near-singularity is causing the matrix to be poorly conditioned. If $K > 1000$, very strong multicollinearity is detected.

(c) The *variance inflation factor* for the j-th regression parameter VIF_j being defined as the ratio of the variance of the jth regression coefficient to the same variance for orthogonal variables when $\boldsymbol{R}$ is the unit matrix. If $VIF_j > 10$, strong multicollinearity is detected. If there is a near-singularity involving X_j and the other independent variables, R_j^2 will be near 1.0 and VIF_j will be large. If X_j is orthogonal to the other independent variables, R_j^2 will be 0 and VIF_j will be 1.0.

Biased regression refers to this class of regression methods in which unbiasedness is no longer required. Generalized principal component regression solves the collinearity problem by elimination of those dimensions of the X-space that induce the problem and was described previously [18].

One of the main properties of regression models is a good predictive ability. This predictive ability can also be adopted for the selection of an, in some sense optimum, criterion parameter P. Various criteria for testing prediction ability may be used[1]; one of the most efficient seems to be the *mean quadratic error of prediction MEP* (in literature it is also known as the mean squared error of prediction *MSEP*) defined by the relationship

$$MEP = \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{(i)})^2}{n} \qquad (2)$$

where $\boldsymbol{b}_{(i)}$ is the estimate of regression parameters when all points except the ith one were used and $\boldsymbol{x}_i$ is the ith row of matrix $\boldsymbol{X}$. The statistic *MEP* uses a prediction $\hat{y}_i$ from an estimate constructed without including the ith point. The most suitable model is that which gives the lowest value (minimum) of the mean quadratic error of prediction *MEP*. Beyond the *MEP*, the predicted coefficient of determination R^2_p (maximum) and the Akaike information criterion *AIC* (minimum) can also be used. The *MEP* can be used to express the *predicted determination coefficient*,

$$\hat{R}_P^2 = 1 - \frac{n \times MEP}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2} \quad (3) \quad .$$

The *Akaike information criterion* is defined

$$AIC = n \ln\left(\frac{U(\boldsymbol{b})}{n}\right) + 2m \quad (4)$$

The most suitable model gives the lowest value of the mean quadratic error of prediction *MEP,* Akaike information criterion *AIC* and the highest value of the predicted determination coefficient, R_p^2. The calculated *P* do not correspond generally to a global minimum but parameter estimates and the statistical characteristics are greatly improved.

Transformation in case of non-normality of variable distributions

There are two basic reasons for transforming variables in regression: transformation of the dependent variable is indicated as possible remedies for non-normality and for heterogenous variances of the random errors **ε**. Transformations to improve normality have generally lower priority that transformation to simplify relationship or stabilize variance. Fortunately, transformations to stabilize variance often have the effect of improving normality as well.

Transformation for symmetry is carried out by a simple *power transformation*

$$y_i^{(\lambda)} = \left\{ \begin{array}{ll} y^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln y & \text{for parameter } \lambda = 0 \\ -y^{-\lambda} & \text{for parameter } \lambda < 0 \end{array} \right\} \quad (5)$$

which does not retain the scale, is not always continuous and is suitable only for positive *y*. Optimal estimates of parameter $\hat{\lambda}$ are sought by minimizing the absolute values of particular characteristics of an asymmetry. The robust estimate of an asymmetry $\hat{g}_P(y)$ may be expressed with the use of a relative distance between the arithmetic mean $\bar{y}$ and the median $\tilde{y}_{0.50}$ by

$$\hat{g}_P(y) = \frac{\bar{y} - \tilde{y}_{0.50}}{\sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}}} \tag{6}$$

as for symmetric distributions it is equal to zero, $\hat{g}_P(y) \approx 0$.

Transformation leading to the approximate normality may be carried out by the use of family of *Box-Cox transformation* [19] defined as

$$y_i^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\ \ln y & \text{for parameter } \lambda = 0 \end{cases} \tag{7}$$

where *x* is a positive variable and λ is real number or in the form with a variable standardization

$$y_i^{(\lambda)} = \begin{pmatrix} \dfrac{y_i^{\lambda} - 1}{\lambda (\bar{y})^{\lambda - 1}} & \text{for parameter } \lambda \neq 0 \\ \bar{y} \ln(y_i) & \text{for parameter } \lambda = 0 \end{pmatrix} \tag{8}$$

where $\bar{y} = \exp \sum_{i=1}^{n} [\ln(y_i)]/n$ is the geometric mean of the original dependent variable. The maximum likelihood solution is obtained by the least squares analysis on the transformed data for several choices of λ from, say $\lambda = -1$ to $+1$. Let $RSS(\lambda)$ be the residual sum of squares from fitting the model to transformed dependent variable $y^{(\lambda)}$ for the given choice of λ and let $\sigma^2(\lambda) = RSS(\lambda)/n$.

Box-Cox transformation lead to $(y - 1)$ when λ is equal to 1 and $\log y$ being the limiting form of the function as λ tends to 0. There is no reason to suppose that either of these values of λ is optimal, and hence it makes sense to try a range of values and see which yields the minimum of *RSS*. If it is to try ten values of λ, ten new dependent variables is generated within the regression application using the functional form and the different values of λ. Resulting y is regressed separately on the explanatory variables.

Box-Cox transformation has the following properties: a) The curves of transformation $g(x)$ are monotonic and continuous with respect to parameter λ because $lim_{\lambda \to 0} \frac{(x^{\lambda} - 1)}{\lambda} = \ln x$. b) All transformation curves share one point for all values of λ. The curves nearly coincide at points close to [0, 1]; *i. e.*, they share a common tangent line at that point. c) The power transformations of exponent -2; -3/2; -1; -1/2; 0; 1/2; 1; 3/2; 2 have equal spacing between curves in the family of Box-Cox transformation graph.

The Box-Cox transformation can be applied on the positive data only. To extend this transformation means to make a substitution of y values by $(y - y_0)$ values which are always positive. Here y_0 is the threshold value $y_0 < y_{(1)}$.

An excellent diagnostic tool enabling estimation of parameter λ may be done by the logarithm of the maximum likelihood function as

$$lnL(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1)\sum_{i=1}^{n} \ln y_i \qquad (9)$$

where $s^2(y)$ is the sample variance of transformed data y. The function $\ln L = f(\lambda)$ is expressed graphically for a suitable interval, for example, $-3 \le \lambda \le 3$. The maximum on this curve represents the maximum likelihood estimate $\hat{\lambda}$. The asymptotic $100(1 - \alpha)$ % confidence interval of parameter λ is expressed by □□□□ $L(\hat{\lambda}) - \ln L(\lambda)] \le \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This interval contains all values λ for which it is true that

$$\ln L(\lambda) \ge \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1) \qquad (10)$$

This Box-Cox transformation is less suitable if the confidence interval for λ is too wide. When the value $\lambda = 1$ is also covered by this confidence interval, the transformation is not efficient and is not recommended.

Experimental

Subjects and plasma samples

The blood was taken from 230 healthy women within 10 and 70 years of age, and from 179 healthy men within 4 and 69 years of age who had been invited by random selection for a survey of iodine deficiency in the district of Cheb in West Bohemia in the frame of the study on iodine deficiency in the Czech Republic. The blood was withdrawn from the cubital vein within 8 - 10 a.m. into heparin-coated vials. Not later than 2 h serum was separated and stored in a freezer at -20°C until processed.

Devices

The HPLC system was from Gilson (Villiers le Bel, France) and consisted of a pump 305 with manometric module 805, slave pump 306, dynamic mixer 811C, autoinjector 234 and fraction collector FC 203B. The UV detector LCD 2082 and column oven LCO 100 were from ECOM (Czech Republic). The reverse phase column ET 250/4 NUCLEOSIL® 100-5 C18 was from Macherey-Nägel (FRG). The CSW APEX system DataApex (Czech Republic) was used for the collecting and working up of chromatographic data.

Determination of pregnenolone sulfate

Pregnenolone sulfate was determined using the modified radioimmunoassay for determination of pregnenolone [20]. Briefly, tritiated pregnanolone as a tracer and antiserum raised against pregnenolone-19-O-carboxymethyoxime conjugated with bovine serum albumin exhibiting 42% cross-reactivity with pregnenolone and 6% cross-reactivity with progesterone were used for the determination. Progesterone and pregnenolone, the concentrations of which are about two orders of magnitude lower than in conjugated steroid were separated using ether extraction (25 ml of sample, 225 ml of distilled water and 1.25 ml of ether). The efficiency of separation was 79.6±3.1% and 87.8±3.3%, in progesterone and pregnenolone, respectively. The organic layer was discarded, while 100 ml of the polar one was used for radioimmunoassay. The calibration curve was constructed using pregnenolone sulfate standard. The sensitivity of analysis was 32 pg per tube, inter and intra-assay coefficient of variation was 10.9% and 4.3%, respectively. The necessity of separation of free steroids from the sample was confirmed by the analysis of interference of cross-reacting substances. The interference was evaluated by HPLC fractionation of the pooled sera from women in luteal phase of menstrual cycle and from umbilical cord of neonates followed by determination of immunoreactivity in dry residues of fractions. The HPLC separation was carried out using the reverse phase system with the column ET 100-5 C18 from Macherey-Nägel (Düren, Germany). The high-pressure gradient with eluent A consisting of 15% acetonitrile in water with addition 100 mg of ammonium bicarbonate per liter and methanol as eluent B was programmed as follows: 0-3 min, 0% B - 6 min, 40% B (linear gradient) - 20 min 80% B (linear gradient) 23 min 100% B (linear gradient) then drop to 0% B up to 30 min. The temperature of the column was 40˚C and constant flow rate was 1 ml/min. The retention times of standards in detectable concentrations (2.5 mg in 25 ml of methanol solution) were measured at 205 nm.

Procedure of statistical data treatment

The procedure for construction of a polynomial regression model consists of following steps:

Step 1. *Proposal of a model.*

Step 2. *Examination of multicollinearity and statistical significance of parameters estimates.*

Step 3. *Construction of a more accurate model using PCR:* on a base of *MEP* or *AIC* the most convenient regression model of transformed data is determined. If some parameters are statistically insignificant the most suitable parameter *P* is searched with the use of *MEP* and *AIC*.

Step 4. *Examination of a variables normality and their transformation, recalculation of results.*

Software used

For creation of computation of the principal component regression PCR the algorithm in *S-Plus* was written and also module Linear Regression of the *ADSTAT* package were used, *cf.* [21].

Results and Discussion

Statistical evaluation of data

We have attempted to describe age relationships and sex differences in serum levels of pregnenolone sulfate using principal component regression on polynomial model. Many problems in chemometrics concern an approximation of instrumental data of convex (or concave) increasing (or decreasing) values by a polynomial so that their course fulfils the condition of the shape of the data. For solution of these types of problems the principal component regression PCR with an optimum value of a criterion *P* minimizing characteristics *MEP, AIC* and maximizing R_P^2 can be used. Main aim is to find a degree of polynomial regression model *m* which describes the content of pregnenolone sulphate for male and female patients, respectively, in dependence on the age and also to estimate all polynomial parameters, $E(\varepsilon/\lambda) = \beta_0 + \beta_1 \lambda + ... + \beta_m \lambda^m$. The purpose of the least squares analysis will influence the manner in which the model is constructed. There are potential uses of regression equations given as providing a good description of the behavior of the responce variable; prediction of future responses and estimation of mean responses; extrapolation, or prediction of responses outside the range of the data; estimation of parameters β and developing realistic models of the process. Each objective has different implications on how much emphasis is placed on eliminating variables from the model, on how important it is that the retained variables be causally related to the response variable, and on the amount of effort devoted to making the model realistic.

Fig. 1

In the step 1 the proposal of a regression model for male patients data is used: Fig. 1 presents the statistics *MEP* and R_P^2 for increasing degree of polynomial *m* and the ordinary least squares OLS used. The lowest *MEP* value and the highest R_P^2 were achieved for a polynomial of the 6th degree. Even that the polynomial of the 5rd degree differs only slightly, the sixth degree polynomial was preferred. All parameters estimates from β_0 through β_6 are not significantly different from zero what is here a result of strong multicollinearity.

Fig. 2, 3, 4, 5

In step 2 the exploratory data analysis in regression is applied and the scatter plot of the regression curve of pregnenolone sulphate data in dependence on age of male patients (Fig. 2) shows a skewed asymmetric distribution of random errors in variable *y*. The rankit *Q-Q* plot of jackknife residuals (Fig. 3) proves the non-normal distribution. Ordinary residuals exhibit strong heteroscedasticity (Fig. 4) and Williams plot (Fig. 5) indicates some influential points among which there are several outliers (1, 74, 136, 142, 147, 60, 65, 150).

In step 3 an examination of multicolinearity concerns an estimation of the maximum condition number $K = 3.63 \times 10^8$ which is higher than 1000 and the largest value of the variance inflation factor $VIF = 3.05 \times 10^7$ is higher than 10 and both criteria indicate a strong multicollinearity. Since the test criterion $F_R = 6.75$ is greater than the corresponding quantile of the Fisher-Snedecor *F*-distribution $F_{0.95}(5, 179\text{-}6) = 2.15$, the proposed regression model is statistically significant. In contrast, the quantile of the Student *t*-distribution, $t_{0.975}(179\text{-}6) = 1.974$ is greater than all $t_0 = 1.435$, $t_1 = -1.790$, $t_5 = -1.953$,

$t_6 = 1.788$ but not then $t_2 = 2.141$, $t_3 = -2.186$ and $t_4 = 2.097$, therefore parameters β_0, β_1, β_5 and β_6 are statistically insignificant while β_2, β_3, β_4 significant. It may be concluded that the method of the ordinary least-squares (OLS) is not convenient for parameter estimation in case of strong multicollinearity in data and the method of the principal component regression PCR should be used instead.

Fig. 6

In step 4 a trial-and-error search of the most suitable value of the criterion parameter *P* with the use of the mean quadratic error of prediction *MEP*, the Akaike information criterion *AIC* and the predicted determination coefficient R_P^2 was applied, Fig. 6. The lowest *MEP* and *AIC* value and the highest value for R_P^2 is for $P = 2.0 \times 10^{-4}$ what means that for $P > 2.0 \times 10^{-4}$ all parameters estimates are statistically significant and therefore different from zero. While the ordinary least-squares method OLS, with $P = 10^{-34}$ found the polynomial

$$y = 810.2(564.8, \mathrm{N}) - 247.9(138.5, \mathrm{N})\, x + 27.54(12.86, \mathrm{S})\, x^2 -1.26(0.58, \mathrm{S})\, x^3 + 2.81(1.34, \mathrm{S}) \times 10^{-2} x^4 - 2.99(1.53, \mathrm{N}) \times 10^{-4} x^5 + 1.23(0.68, \mathrm{N}) \times 10^{-6} x^6$$

(in brackets is the parameter standard deviation and N means that the parameter estimate is statistically non-significant while S means significant) with *MEP* = 51192, *AIC* = 1921.1 and $R_P^2 = 14.91$, the method of the principal component regression PCR, with $P = 2.0 \times 10^{-4}$ found

$$y = -395.8(155.1, \mathrm{S}) + 64.93(17.65, \mathrm{S})\, x - 1.50(0.50, \mathrm{S})\, x^2 - 1.39(0.91, \mathrm{N}) \times 10^{-3} x^3 + 2.27(0.85, \mathrm{S}) \times 10^{-4} x^4 + 2.41(0.85, \mathrm{S}) \times 10^{-6} x^5 - 4.92(0.19, \mathrm{S}) \times 10^{-8} x^6$$

with lower value of the criterion *MEP* = 46032, *AIC* = 1926.9 and higher value of $R_P^2 = 34.75$. All parameters estimated by the method of the principal component regression PCR are statistically significant and therefore are acceptable even that an excellent curve fitting was achieved in both cases.

Table 1

Fig. 7

To examine normality of random errors distribution in dependent variable *y* and to find the most convenient variable transformation, the $RSS(\lambda)$ for different values of power λ were searched and the best power estimated (Table1, Fig. 7). Several resolution criteria ware applied to find the optimal power λ but the most important one was such λ for which the normality of residual distribution was achieved. It means for which the skewness g_1 is nearly zero and the kurtosis g_2 is nearly equal to 3. Resulting power was $\lambda = -0.15$. Fig. 8 shows the scatter plot of found polynomial ($m = 6$) through transformed data and criterion parameter $P = 0.0002$ and the rankit *Q-Q* plot of jackknife residuals then proves a normal distribution and a homoscedasticity of residuals. The method of the generalized principal component regression GPCR, with $P = 2.0 \times 10^{-4}$ and using transformed data found

$$y = 539.5(60.9, \mathrm{S}) + 36.58(6.93, \mathrm{S})\, x - 0.81(0.20, \mathrm{S})\, x^2 - 1.03(0.36, \mathrm{S}) \times 10^{-3}\, x^3 + 1.19(0.34, \mathrm{S}) \times 10^{-4}\, x^4 + 1.29(0.33, \mathrm{S}) \times 10^{-6}\, x^5 - 2.54(0.75, \mathrm{S}) \times 10^{-8}\, x^6$$

with value of the criterion $MEP = 7219.0$, $AIC = 1592.2$ and value of $R_P^2 = 51.12$. All parameters estimated by the method of the principal component regression PCR are statistically significant and therefore are acceptable even that an excellent curve fitting was achieved in both cases.

Fig. 8, 9

Analogically, the analysis of female patients data was provided and the optimum degree of polynom was found $m = 4$ (Fig. 9). Analogically as for the male data also for females a non-normality of random errors in dependent variable y and heteroscedasticity of ordinary residuals may be proven. The best power for power transformation was found $\lambda = 0.2$, Fig. 9. When the PCR method on transformed data was applied the optimum polynomial ($m = 4$) with parameter $P = 0.0004$ results. While the ordinary least-squares method (OLS with $P = 10^{-35}$) found the polynomial

$$y = -254.3(211.7, \mathrm{N}) + 27.0(29.5, \mathrm{N})\, x + 0.225(1.333, \mathrm{N})\, x^2 - 2.17(2.40, \mathrm{N}) \times 10^{-2}\, x^3 + 2.00(1.50, \mathrm{N}) \times 10^{-4}\, x^4$$

(in brackets is the parameter standard deviation and N means that the parameter estimate is statistically non-significant while S means significant) with $MEP = 24335$, $AIC = 2324.3$ and $R_P^2 = 54.71$, the method of the principal component regression PCR, with $P = 4.0 \times 10^{-4}$ and $\lambda = +0.20$ found

$$y = 0.661(0.217, \mathrm{S}) + 0.148(0.017, \mathrm{S})\, x - 1.940(0.287, \mathrm{S}) \times 10^{-3}\, x^2 - 1.903(0.239, \mathrm{S}) \times 10^{-5}\, x^3 + 3.007(0.538, \mathrm{S}) \times 10^{-7}\, x^4$$

with lower value of the criterion *MEP* = 0.189, *AIC* = -380.0 and higher value of R_P^2 = 62.54. All parameters estimated by the method of the principal component regression PCR are statistically significant and therefore are acceptable even that an excellent curve fitting was achieved with the use of both method. The ordinary residuals exhibit a normal distribution and an obvious homoscedasticity.

Interpretation of the results

The course of the age dependence in male differs from that in female (Fig. 8 and Fig. 9). While in women, a pronounced maximum after 30th year of age was followed by relatively rapid decline up to senescence , in men, the maximum after 20th year of age was succeeded by minor decline up to 40th year of age and plateau up to 60th year of age followed by more rapid decrease. Both age and sex differences were evaluated using two-way ANOVA with the sex as the first and the age group as the second factor. The data were transformed using power transformation of the original data to minimum skewness of residuals. Significant difference was found between sexes with tendency to lower levels of PregS in older women when compared with age-matched men.

Conclusion

The method of the principal component regression PCR in combination with the *MEP* criterion is very useful and attractive for constructing biased models. It can be also used for achieving such estimates which keep the model course corresponding to the data trend especially in polynomial-type regression models. In the search for the best degree of polynomial, several statistical characteristics of regression quality should be considered together. Significant differences were found between men and women in the course of age dependence of pregnenolone sulfate. In women, a significant maximum was found around 30th year of age followed by rapid decline, the maximum in men was achieved almost 10 years earlier and the changes were inconsiderable up to 60th year of age. The investigation of sex differences and age dependencies of pregnenolone sulfate could be of interest taking together its well-known neurostimulating effect, relatively high serum concentration and probable partial permeability of blood-brain barrier for the steroid conjugate which reflects for instance in correlation of complaints of the patients suffering with premenstrual syndrome with serum levels of the conjugate. As concerns the method of data analysis, the principal component regression is very useful tool for investigation of curvilinear dependencies especially in polynomial regression models.

Acknowledgements

The financial support of the Ministry of Education (Grant No MSM253100002) and of the Grant Agency of the Ministry of Health of the Czech Republic (Grant NB/7391-3) is gratefully acknowledged.

References

[1] Majewska, 1990 #810;

[2] Flood, 1992 #816;

[3] Flood, 1995 #813

[4] [Monnet, 1995 #818;

[5] Murray, 1997 #821;

[6] Mathis, 1996 #823;

[7] Wakerley, 1997 #827]

[8] [Wang, 1997 #458]

[9] [Rajkowski, 1997 #459;

[10] Corpechot, 1997 #1272;

[11] Young, 1996 #1370;

[12] Wang, 1996 #1274]

[13] de Peretti, 1983 #1277]

[14] de Peretti, 1986 #904

[15] Bicikova, 2001 #1838

[16] M. Meloun, J. Militký and M. Forina: Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods, Horwood, Chichester, 1994.

[17] M. Meloun, J. Militký, M. Hill, R. G. Brereton, Crucial problems in regression modelling and their solution, The Analyst, 127 (2002) 433-450.

[18] M. Meloun, J. Militký, CCLM (previous paper of this series, in press, dopíšu já, Milan).

[19] J. O. Rawlings, S. G. Pantula, D. A. Dickey, Applied Regression Analysis, Springer, New York 1998, page 411.

[20] Martin doplni

[21] ADSTAT, TriloByte Statistical Software, Pardubice 1999; http://www.trilobyte.cz.

Corresponding author:

Prof. RNDr. Milan Meloun, DrSc.,

Department of Analytical Chemistry,

University Pardubice,

532 10 Pardubice, Czech Republic,

Phone: +42466037026, **Fax:** +42466037068,

Email: milan.meloun@upce.cz

FIGURES:

Fig. 1 A search for the optimum polynom degree *m* obtained for the lowest value of the mean error of prediction *MEP* and for the highest value of the predicted determination coefficient R_P^2 concerning the ordinary least squares polynomial regression OLS of the age dependencies of pregnenolone sulphate in the serum of 179 men aged 4-69 years.

Fig. 2 The scatter plot of polynomial regression of the age dependencies of pregnenolone sulphate in the serum of 179 men aged 4-69 years from Fig. 1 calculated with the use of the ordinary least squares method OLS and original data. The curves of the mean prediction (the full line), the

95% Working-Hotteling interval bands of prediction (the dashed lines) are symmetrical and rather broad at ends of data interval and therefore do not permit prediction of *y* outside the data interval. All of the parameters of the polynomial were statisticaly insignificant.

Fig. 3 The rankit *Q-Q* plot of jackknife residuals for the polynomial dependence of pregnenolone sulphate in serum for 179 men calculated with the OLS analysis from Fig. 1 proves a non-normality of residual distribution and therefore indicates necessity of data transformation.

Fig. 4 The sector pattern shape for the polynomial dependence of pregnenolone sulphate in serum for 179 men calculated with the OLS analysis from Fig. 1 proves heteroscedasticity and therefore indicates necessity of data transformation.

Fig. 5 Williams graph of jackknife residuals $\hat{e}_J$ on the diagonal elements of the hat matrix H_{ii} proves outliers (i. e. points above the horizontal boundary line $y = t_{0.95}(n\text{-}m\text{-}1)$) and high-leverages (i. e. points located right to the vertical boundary line $x = 2m/n$) concerning the age dependencies of pregnenolone sulphate in the serum of 179 men calculated with the OLS analysis from Fig. 1.

Fig. 6 A search for an optimum value of the criterion *P* for the age dependencies of pregnenolone sulphate in the serum of 179 men calculated with the generalized principal compoments regression GPCR from Fig. 1 according to which the terms corresponding to small eigenvalues are omitted. The optimum value concerns the minimum on curve of the mean error of prediction *MEP* in dependence on the GPCR criterion *P*.

Fig. 7 A search for an optimum power λ in the power transformation of the dependent variable y^λ is based on the maximum of the curve $R_P^2 = f(\lambda)$ for the age dependencies of pregnenolone sulphate in the serum of 179 men calculated with GPCR from Fig. 1

Fig. 8 The scatter plot of 6-th polynomial regression of the age dependencies of pregnenolone sulphate in the serum of 179 men from Fig. 1 calculated with the use of transformed dependent variable y^λ and GPCR analysis. The curves of the mean prediction (the full line), the 95% Working-Hotteling interval bands of prediction (the dashed lines) permit prediction of *y* outside the data

interval. All of the parameters of the polynomial were statisticaly significant (*t*-tests).

Fig. 9 The scatter plot of 4-th polynomial regression of the age dependencies of pregnenolone sulphate in the serum of 230 women in age 10-70 years calculated with the use of transformed dependent variable y^{λ} and OLS analysis. The curves of the mean prediction (the full line), the 95% Working-Hotteling interval bands of prediction (the dashed lines) permit prediction of *y* outside the data interval.

Tables:

Table 1. A search for an optimum power λ in the power transformation of the dependent variable y^{λ} is based on the maximum of the curve $R_P^2 = f(\lambda)$ for the age dependencies of pregnenolone sulphate in the serum of 179 men aged 4-69 years from Fig. 1 calculated with the use of the generalized principal component regression GPCR is based on an examination of following variables and test results: $n = 179$, $m = 6$, $P = 0.0002$ (GPCR), $R = \mathrm{f}(\lambda)$, $100R^2 = \mathrm{f}(\lambda)$, $R_P^2 = \mathrm{f}(\lambda)$, for normal distribution the skewness g_1 should zero and the kurtosis g_2 equal to 3, Cook-Weisberg test of homoscedasticity: homoscedastity can be accepted or rejected, Jarque-Berra normality test: normality can be accepted or rejected.

	Exponent λ is equal to						
	-0.27	-0.25	-0.20	-0.18	-0.15	-0.10	-0.01
R	0.5534	0.5544	0.5562	0.5566	0.5569	0.5567	0.5542
100R^2 *[%]*	30.63	30.74	30.93	30.98	31.02	31.00	30.31
R^2_P	50.66	50.79	51.04	51.11	51.18	51.22	51.03
MEP	7330.5	7285.7	7220.2	7212.1	7219.0	7280.4	7548.1
AIC	1594.7	1593.6	1592.1	1592.0	1592.2	1593.9	1600.6
$RSS(\lambda)$ $*10^{-6}$	1.2250	1.2170	1.2070	1.2060	1.2080	1.2190	1.2660
Skewness $g_1(e)$	-1.52	-1.43	-1.21	-1.13	-1.00	-0.82	-0.50
Kurtosis $g_2(e)$	7.93	7.42	6.32	5.94	5.42	4.73	3.89
Found λ can be	Not used	Not used	Not used	Used	Used	Not used	Not used