Minimizing the effects of multicollinearity in the polynomial regression of age relationships and sex differences in serum levels of pregnenolone sulfate in healthy subjects

Milan Meloun¹,* , Martin Hill² and Helena Včeláková-Havlíková²
¹ Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, Pardubice, Czech Republic
² Institute of Endocrinology, Prague, Czech Republic

Abstract

Background: Pregnenolone sulfate (PregS) is known as a steroid conjugate positively modulating N-methyl-D-aspartate receptors on neuronal membranes. These receptors are responsible for permeability of calcium channels and activation of neuronal function. Neuroactivating effect of PregS is also exerted via non-competitive negative modulation of GABA_A receptors regulating the chloride influx. Recently, a penetrability of blood-brain barrier for PregS was found in rat, but some experiments in agreement with this finding were reported even earlier. It is known that circulating levels of PregS in human are relatively high depending primarily on age and adrenal activity.

Methods: Concerning the neuromodulating effect of PregS, we recently evaluated age relationships of PregS in both sexes using polynomial regression models known to bring about the problems of multicollinearity, i.e., strong correlations among independent variables. Several criteria for the selection of suitable bias are demonstrated. Biased estimators based on the generalized principal component regression (GPCR) method avoiding multicollinearity problems are described.

Results: Significant differences were found between men and women in the course of the age dependence of PregS. In women, a significant maximum was found around the 30th year followed by a rapid decline, while the maximum in men was achieved almost 10 years earlier and changes were minor up to the 60th year. The investigation of gender differences and age dependencies in PregS could be of interest given its well-known neurostimulating effect, relatively high serum concentration, and the probable partial permeability of the blood-brain barrier for the steroid conjugate.

Conclusions: GPCR in combination with the MEP (mean quadric error of prediction) criterion is extremely useful and appealing for constructing biased models. It can also be used for achieving such estimates with regard to keeping the model course corresponding to the data trend, especially in polynomial type regression models.


Keywords: age relationship of pregnenolone sulfate; multicollinearity; principal component regression; steroids.

Introduction

Pregnenolone sulfate (PregS) is known as a steroid conjugate positively modulating N-methyl-D-aspartate receptors on neuronal membranes. These receptors are responsible for permeability of calcium channels and activation of neuronal function (1–6). Neuroactivating effect of PregS is also exerted via non-competitive negative modulation of GABA_A receptors regulating the chloride influx (1, 7). Recently, a penetrability of blood-brain barrier for PregS was found in rat (8), but the experiments in agreement with this finding were reported even earlier (9–11). It is known that circulating levels of PregS in human are relatively high depending primarily on age and adrenal activity (12, 13). The use of PregS (with significant responsiveness to ACTH stimulation) was also suggested as a marker of adrenal dysfunction in children (13). Taken together, 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 the diagnostics of adrenal disorders but also in the diagnostics of some disturbances of the central nervous system as anxiety-depressive syndrome (14), or the diseases connected with aging, such as Alzheimer’s disease. de Peretti and Mappus described in detail the age dependence of PregS during childhood and adolescence, but the changes during adulthood were not investigated (12). We recently concentrated on this problem evaluating the age relationships of PregS in both sexes in the age range from 4 to 70 years using a polynomial regression model (15).

In this paper, we attempt to demonstrate a correct approach for the analysis of age dependencies of the hormones, which are commonly represented by the data with a non-Gaussian distribution and non-constant variance. In addition, multicollinearity (strong correlations between independent variables) is characteristic in the commonly used ordinary polynomial regression models. The difficulties as mentioned above could result in a complete misinterpretation of...
the data. Accordingly, we suggest an approach based on one of the biased regression models. A general-
ized principal component polynomial regression (GPCR) of the data transformed by power transfor-
mation was chosen as a suitable method capable of coping with the aforementioned problems. The aims
of this paper are to emphasize the importance of understanding the nature of any near-singularities in
the data that might cause problems with ordinary least squares regression, and to describe the algo-
rumth, the principal component regression. Several

criteria for the selection of suitable bias are demon-
strated and biased estimators based on the general-
ized principal component regression (GPCR) method
are described.

Methodology

The polynomial linear regression model with n obser-
vations of the m-th order polynomial variable and for
an additive model of measurements errors is assumed.

\[ y = X \beta + \varepsilon \]

Vector \( y \) has dimensions \( (n \times 1) \) and matrix \( X \) has dimensions \( (n \times m) \). Random errors
\( \varepsilon \) in dependent variable \( y \) should have a normal dis-
tribution \( N(0, \sigma^2) \). When the least squares assump-
tions are valid, the parameter estimates \( b \) found by
minimization of the sum of squared residuals \( RSS \)

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

are the best linear unbiased estimators (16, 17). The
conventional least-squares estimator \( b \) has the form

\[ b = (X'X)^{-1}X'y \]

with corresponding variance \( \text{Var}(b) = \sigma^2 (X'X)^{-1} \). However, some difficulties arise when the matrix \( X'X \) appears to be singular. In some cases, especially with polynomial models, the parameter
estimates may be without physical meaning. The mul-
ticollinearity problem in regression refers to the set of
problems created when there are near-singularities
among the columns of the \( X \) matrix and certain linear
combinations of the columns of \( X \) are nearly zero (18).

a) The condition number \( K = \lambda_{\text{max}} / \lambda_{\text{min}} \) contains \( \lambda_{\text{max}} \)
and \( \lambda_{\text{min}} \), the largest and the smallest eigenvalues
of a matrix \( R \) (19). 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-sin-


ularity is causing the matrix to be poorly condi-
tioned. If \( K > 1000 \), very strong multicollinearity is
detected.

b) The variance inflation factor for the \( j \)-th regression
parameter \( \text{VIF}_j \) is defined as the ratio of the vary
ance of the \( j \)-th regression coefficient to the same
variance for orthogonal variables when \( R \) is the
unit matrix. If \( \text{VIF}_j > 10 \), 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 \( \text{VIF}_j \) will be large. If \( X_j \) is orthogonal
to the other independent variables, \( R_j^2 \) will be 0
and \( \text{VIF}_j \) will be 1.0.

Biased regression refers to that class of regression
methods in which unbiasedness is no longer
required. GPCR solves the collinearity problem by elimin-
ation of those dimensions of the \( X \)-space that
induce the problem and has been described previ-
ously (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; one of the
most efficient seems to be the mean quadratic error
of prediction \( MEP \) (in the literature it is also referred
to as the mean squared error of prediction \( MSEP \))
defined by the relationship

\[ MEP = \frac{\sum_{i=1}^{n} (y_i - x_i' \hat{b})^2}{n} \]  

where \( \hat{b} \) is the estimate of regression parameters
when all points except the \( \text{th} \) one were used and \( x_i \)
is the \( \text{th} \) row of matrix \( X \). The statistic \( MSEP \) uses a
prediction \( \hat{y} \) from an estimate constructed without
including the \( \text{th} \) 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 \( R_p^2 \)
(maximum) and the Akaike information criterion \( AIC \)
(minimum) can also be used. The \( MEP \) can be used
to express the predicted determination coefficient:

\[ R_p^2 \quad (100\%) = 1 - \frac{n \times MSEP}{\sum_{i=1}^{n} y_i^2 - n \times \bar{y}^2} \]  

The Akaike information criterion \( AIC \) is thus defined:

\[ AIC = n \ln \left( \frac{U(b)}{n} \right) + 2m \]  

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

Transformation in cases of non-normality
of variable distributions

There are two basic reasons for transforming
variables in regression: transformation of the depend-
ent variable is a possible remedy for non-normality
and for heterogeneous variances of the random
errors \( \varepsilon \). Transformations to improve normality gen-
erally have lower priority than transformation to sim-
plify relationship or stabilize variance. Fortunately, transformations to stabilize variance often have the effect of improving normality as well.

Transformation for symmetry of error distribution is carried out by a simple power transformation:

\[
y^{(l)} = \begin{cases} 
y^{l} & \text{for parameter } \lambda > 0 \\
\ln y & \text{for parameter } \lambda = 0 \\
\frac{\ln y}{\lambda} & \text{for parameter } \lambda < 0
\end{cases}
\]  

(5)

which does not retain the scale, is not always continuous, and is suitable only for positive \(y\). Optimal estimates of parameter \(\lambda\) are sought by minimizing the absolute values of particular characteristics of an asymmetry of residuals \(\hat{e}\).

Transformation leading to approximate normality may be carried out by the use of the family of Box-Cox transformations (20), defined as:

\[
y^{(l)} = \begin{cases} 
(y^{l} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\
\ln y & \text{for parameter } \lambda = 0
\end{cases}
\]  

(6)

where \(y\) is a positive variable and \(\lambda\) is a real number or in a form with variable standardization. The maximum likelihood solution is obtained by the least squares analysis on the transformed data for several choices of \(\lambda\) from, say \(-3\) to \(+3\). Let \(\text{RSS}^{(l)}\) be the residual sum of squares from fitting the model to transformed dependent variable \(y^{(l)}\) for the given choice of \(\lambda\) and let 

\[
s^{2}(y) = \text{RSS}^{(l)}/n.
\]

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

Procedure of statistical data treatment

The procedure for the construction of a best polynomial regression subset \(\{1, x, x^2, x^3, \ldots, x^n\}\) comprises the following steps:

Step 1 Proposal of a model.

Step 2 Examination of the multicollinearity and statistical significance of the parameter estimates.

Step 3 Construction of a more accurate model using GPCR with an optimum value of a criterion \(P\) minimizing characteristics \(\text{MEP}, \text{AIC}\) and maximizing \(R^2\) can be used: on the base of \(\text{MEP}\) or \(\text{AIC}\) the most convenient regression model of the transformed data is determined. If some parameters are statistically non-significant, the most suitable parameter \(P\) is sought with the use of \(\text{MEP}\) and \(\text{AIC}\).

Step 4 Examination of variable normality and transformation, recalculation of results.

Software used

The algorithm in S-Plus was written for the creation of the computation of the GPCR and the linear regression module of the ADSTAT package was also used, cf. (21).

Results and discussion

Statistical evaluation of data

We attempt to describe age relationships and sex differences in serum levels of PregS using principal component regression on a polynomial model. For the solution of these types of problems, GPCR with an optimum value of a criterion \(P\) can be used. The main aim is to find a degree of polynomial regression model \(m\) which describes the content of PregS for male and female patients, respectively, in dependence on age, and also to estimate all the polynomial parameters, \(E(e|x) = \beta_0 + \beta_1 x + \ldots + \beta_m x^m\). The purpose of the least squares analysis will influence the manner in which the model is constructed. There are potential uses of the regression equations given in providing a good description of the behavior of the response variable, the prediction of future responses and the estimation of mean responses, the estimation of parameters \(\beta\) and development of realistic models of the process. Each objective has different implications on how much emphasis is placed on eliminating variables from the model, as to how important it is that the retained variables be causally related to the response variable, and on the amount of effort required to make the model realistic.

In step 1 the proposal of a regression model for male patients’ data is used: Figure 1 presents the MEP statistics for an increasing degree of polynomial \(m\) and the ordinary least squares (OLS) used. The lowest MEP value was achieved for a polynomial of the 7th degree. Although the polynomial of the 6th degree differs only slightly, the 6th degree polynomial was preferred here. The parameter estimates from \(\beta_0\) through \(\beta_6\) are not significantly different from zero, which here results from strong multicollinearity.

In step 2 the exploratory data analysis in regression is applied, and the scatter plot of the regression curve

Materials and methods

Subjects and plasma samples

Blood was taken from 230 healthy females aged between 10 and 70 years, and from 179 healthy males aged between 4 and 69 years, who had been invited by random selection for a survey of iodine deficiency in the district of Cheb in West Bohemia, within the framework of a study on iodine deficiency in the Czech Republic. Blood was withdrawn from the cubital vein between 08:00 h and 10:00 h into heparin-coated vials. After no more than 2 h, serum was separated and stored in a freezer at –20°C until processed.

Determination of pregnenolone sulfate

PregS was determined using the specific radioimmunoassay as described previously (15). The sensitivity of analysis was 32 pg/tube, and the inter- and intra-assay coefficients of variation were 10.9% and 4.3%, respectively.

Reference
of PregS data in dependence on age of male patients (Figure 2) shows a skewed asymmetric distribution of random errors in variable y as shown in Table 1. The rankit Q-Q plot of jackknife residuals proves the non-normal distribution. The ordinary residuals exhibit strong heteroscedasticity (Figure 3) and some influential points.

In step 3 the examination of multicollinearity comprises an estimation of the maximum condition number $K=3.63 \times 10^6$ which is higher than 1000 and the largest value of the variance inflation factor $VIF=3.05 \times 10^7$ which is higher than 10; and both criteria indicate strong multicollinearity. Since the test criterion $F_{R}=6.75$ is greater than the corresponding quantile of the Fisher-Snedecor F-distribution $F_{0.975}(5,179)=2.15$, the proposed regression model is statistically significant. In contrast, the quantile of the Student t-distribution, $t_{0.975}(179-6)=1.974$ is greater than all $t_{q}$ values: $t_{q}=1.435$, $t_{q}=-1.790$, $t_{q}=-1.953$, $t_{q}=1.788$ but not than $t_{q}=2.141$, $t_{q}=-2.186$, and $t_{q}=2.097$; therefore, parameters $\beta_{3}$, $\beta_{1}$, $\beta_{5}$, and $\beta_{4}$ are statistically non-significant, while $\beta_{2}$, $\beta_{3}$, and $\beta_{4}$ are significant. It may be concluded that the OLS method is not convenient for parameter estimation in cases of strong multicollinearity in the data and that GPCR should be used instead.

In step 4 a trial-and-error search of the most suitable value of the criterion parameter $P$ using the mean quadratic error of prediction $MEP$, the Akaike information criterion AIC and the predicted determination coefficient $R_{p}^2$ was applied (Figure 4). The lowest $MEP$ and $AIC$ value and the highest value for $R_{p}^2$ is for $P=2.0 \times 10^{-4}$, which means that for $P>2.0 \times 10^{-4}$ all parameter estimates are statistically significant and therefore different to zero. While the OLS method, with $P=10^{-34}$, found the polynomial $y=810.2(564.8, N)–247.9(138.5, N) x + 27.54(12.86, S) x^2–1.26(0.58, S) x^3+2.81(1.34, S) x^4–2.99(1.53, N) x^5+1.23(0.68, N) x^6$ (where the parameter standard

Table 1 Search for the optimum power $\lambda$ in the power transformation of the dependent variable $y^\lambda$.

<table>
<thead>
<tr>
<th>Search for the optimum power $\lambda$ in the power transformation</th>
<th>$\lambda$</th>
<th>$\lambda$</th>
<th>$\lambda$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>0.5544</td>
<td>0.5562</td>
<td>0.5569</td>
<td>0.5567</td>
</tr>
<tr>
<td>$100R^2$ (%)</td>
<td>30.74</td>
<td>30.93</td>
<td>31.02</td>
<td>31.00</td>
</tr>
<tr>
<td>$100R^2_\gamma$ (%)</td>
<td>50.79</td>
<td>51.04</td>
<td>51.18</td>
<td>51.22</td>
</tr>
<tr>
<td>$MEP$</td>
<td>7285.7</td>
<td>7220.2</td>
<td>7219.0</td>
<td>7280.4</td>
</tr>
<tr>
<td>$AIC$</td>
<td>1593.8</td>
<td>1592.1</td>
<td>1592.2</td>
<td>1593.9</td>
</tr>
<tr>
<td>$RSS(\lambda) / \times 10^{-d}$</td>
<td>1.2170</td>
<td>1.2070</td>
<td>1.2080</td>
<td>1.2190</td>
</tr>
<tr>
<td>Skewness $g_1(\lambda)$</td>
<td>-1.43</td>
<td>-1.21</td>
<td>-1.00</td>
<td>-0.82</td>
</tr>
<tr>
<td>Kurtosis $g_2(\lambda)$</td>
<td>7.42</td>
<td>6.32</td>
<td>5.42</td>
<td>4.73</td>
</tr>
</tbody>
</table>

Found $\lambda$ can be: Not used Not used Used Not used Not used

The search for the optimum power $\lambda$ in the power transformation of the dependent variable $y^\lambda$ is based on the maximum of the curve $R_{p}^2=f(\lambda)$ for the age dependencies of pregnenolone sulfate in the serum of 179 males aged 4–69 years from Figure 2, calculated with the use of generalized principal component regression (GPCR) based on an examination of the following variables and test results: $n=179$, $m=6$, $P=0.0002$ (GPCR), $R_f(\lambda)$, $100R^2_f(\lambda)$, $100R^2_\gamma_f(\lambda)$; for normal distribution, the skewness $g_1$ should be equal to zero and the kurtosis $g_2$ equal to 3.
The sector pattern shape for the polynomial dependence of pregnenolone sulfate in serum of 179 men calculated with the ordinary least squares (OLS) analysis from Figure 1 proves heteroscedasticity and therefore indicates a need for data transformation.

The search for the optimum value of the criterion $P$ for the age dependencies of pregnenolone sulfate in the serum of 179 men calculated with generalized principal component regression (GPCR) from Figure 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 $\text{MEP}$ in dependence on the GPCR criterion $P$.

A search for the optimum power $\lambda$ in the power transformation of the dependent variable $y^\lambda$ based on the maximum of the curve $R^2_s$ for the age dependencies of pregnenolone sulfate in the serum of 179 men calculated with generalized principal component regression (GPCR) from Figure 2.

To examine the normality of random error distribution in the dependent variable $y$ and to find the most convenient variable transformation, the $\text{RSS}(\lambda)$ for different values of power $\lambda$ were sought out and the best power estimated (Table 1, Figure 5). Several resolution criteria were applied to find the optimal power $\lambda$, but the most important was such $\lambda$ for which the normality of residual distribution was achieved, i.e., for which the skewness $g_1$ is nearly zero and the kurtosis $g_2$ is nearly equal to 3. The resulting power was $\lambda = -0.15$. Figure 6 shows the scatter plot of the polynomial found ($m=6$) through transformed data and the criterion parameter $P=0.0002$, and a rankit Q-Q plot of jackknife residuals then proves a normal distribution and a homoscedasticity of residuals. The GPCR method with $P=2.0 \times 10^{-4}$ and using transformed data found $y^{\lambda=0.15} = -395.8(155.1, S) + 46.93(17.65, S) x - 1.50(0.50, S) x^2 - 1.39(0.91, N) x^3 + 2.27(0.85, S) x^4 + 2.41(0.85, S) x^5 - 0.81(0.20, S) x^6 + 1.29(0.33, S) x^7 - 2.54(0.75, S) x^8$ with the values of $\text{MEP}=5192.0$, $\text{AIC}=1592.2$ and $R^2_s \times 100\% = 51.12\%$. All parameters estimated by GPCR are statistically significant and are
The course of the age dependence in males differs from that in females (Figures 6 and 7). In women, a pronounced maximum after the 30th year was followed by a relatively rapid decline up to senescence, while in men the maximum after the 20th year was succeeded by a minor decline up to the 40th year and a plateau up to the 60th year followed by a more rapid decrease. As confirmed using two-way analysis of variance (ANOVA) [the details of statistical analysis with ANOVA were published elsewhere (15)] with sex and age group as the first and second factors, respectively, both the sex and the age differences were highly significant (p<0.0001), as were the differences in the shapes of the age dependencies (age/sex interactions).

Conclusions

GPCR in combination with the MEP criterion is extremely useful and appealing for constructing biased models. It can also be used for achieving such estimates with regard to keeping 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 the age dependence of PregS. In women, a significant maximum was found around the 30th year followed by a rapid decline, while the maximum in men was achieved almost 10 years earlier and changes were minor up to the 60th year. The investigation of gender differences and age dependencies in PregS could be of interest given its well-known neurostimulating effect, relatively high serum concentration, and the probable partial permeability of the blood-brain barrier for the steroid conjugate, which is reflected, e.g., in the correlation of complaints by patients suffering from premenstrual syndrome with serum levels of the conjugate. With regard to the method of data analysis, principal component regression is an extremely useful tool for the investigation of curvilinear dependencies, especially in polynomial regression models.

Acknowledgements

Financial support from the Grant Agency of the Czech Republic (grant no. GAČR 303/06/1817) is gratefully acknowledged.

References


