Teaching Mathematics and Statistics in Sciences - IPA HU-SRB/0901/221/088
Interesting Mathematical Problems in Sciences and Everyday Life - 2011

Practical application of biostatistical methods in medical and biological research

Krisztina Boda

Department of Medical Physics and Informatics, University of Szeged, Hungary

Korányi fasor 9, Szeged, Hungary, 6720

boda@dmi.u- szeged.hu

Biostatistics is the application of mathematical statistics to medical and biological data. The methods are based on hard mathematics but their principle can be easily understood. The paper gives a short overview of the generalized linear models and describes the possibility of their application in medicine. The methods are illustrated by two practical examples. The first medical problem is the effect of intravenous lactate infusion on cerebral blood flow in Alzheimer's disease. Here a mixed model repeated-measurement ANOVA was used to examine the effect of Na-lactate infusion in time. Using mixed model, the variance-covariance structure of repeated measures can modelled, and missing values can be taken into consideration. The SAS software was applied for calculations. The other medical problem is the investigation of risk factors of respiratory complications in paediatric anaesthesia using relative risk regression. Here, strong correlation was found between several independent variables. When the independent variables are correlated, there are problems in the estimation of the regression coefficients. To avoid multicollinearity, the structure of the correlation of the candidate variables used in the multivariate model was first examined by factor analysis, later new artificial variables were formed. The final multivariate model gave us the most important risks factors. Based on the model, children at high risk for perioperative respiratory adverse events could be systematically identified at the preanaesthetic assessment and thus can benefit from a specifically targeted anaesthesia management.

1. Introduction

Statistics may be defined as a body of methods for learning from experience – usually in the form of numbers from many separate measurements displaying individual variations. Due to the fact that many non-numeric concepts, such as male or female, improved or worse, etc. can be described as counts, rates or proportions. The scope of statistical reasoning and methods is surprisingly broad. Nearly all scientific investigators find that their work sometimes presents statistical problems that demand solutions; similarly, nearly all readers of research reports find that the understanding of the reported results of a study requires a knowledge of statistical issues and of the way in which the investigators have addressed those issues.

One characteristic of medical and biological research is that the examinations result in data generally described by numbers. Biostatistics provides methods that permit a description and summary of such so that consequences may be drawn from them. Biostatistics is an application of mathematical statistics to the evaluation of biological and medical experimental data. It is based on probability theory and mathematical statistics.

Biostatistical methods are widely used in medical research. A scientific paper without such an evaluation is currently almost inconceivable. Moreover, the number of medical papers is increasing very rapidly year by year, while the evaluation of the experiments reported requires increasingly more sensitive methods. Meanwhile, the spreading of up-to-date knowledge is rendered more difficult by the specialisation at present going on throughout the medical profession.

The aim of the present work is to give a short overview of the generalized linear models and describes the possibility of their application in medicine. The methods will be illustrated by two practical examples.

2. The theory of generalized linear models

The general linear model

Notation

We denote random variables by upper case italic letters and observed values by the corresponding lower case letters. For example, the observations $y_1, y_2, \ldots y_n$ are regarded as realizations of the random variables $Y_1, Y_2, \ldots Y_n$. We use greek letters to denote parameters and the corresponding lower case roman letters are used to denote estimators and estimates. Vectors and matrices, are denoted by bold face lower and upper case letters, respectively. For example, y represents a vector of observations $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, or a vector of random variables $\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$. $\boldsymbol{\beta}$ denotes a vector of parameters and $\boldsymbol{X}$ is a matrix.

The general form of the linear model

The general form of the linear model is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where

$\boldsymbol{y}$ is an n x1 response vector,

$\boldsymbol{X}$ is an n x p matrix of constants ("design" matrix), columns are mainly values of 0 or 1 and values of independent variables,

$\boldsymbol{\beta}$ is a p x 1 vector of parameters, and

$\boldsymbol{\varepsilon}$ is an n x 1 random vector whose elements are independent and all have normal distribution N(0, σ^2).

For example, a linear regression equation containing three independent variables can be written as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \boldsymbol{\varepsilon}$, or

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X}=\begin{bmatrix} 1\ x_{11}\ x_{12}\ x_{13} \\ 1\ x_{21}\ x_{22}\ x_{23} \\ \vdots\ \vdots\ \vdots\ \vdots \\ 1\ x_{n1}\ x_{n2}\ x_{n3} \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimations of the regression coefficients β_j can differ when several models are fitted to data. Moreover, the test of the hypotheses $\beta_j = 0$ depends on which terms were included in the model. Estimates, confidence intervals and hypothesis tests usually depend on which variables are included in the model. There is an exception when matrix $\boldsymbol{X}$ is orthogonal. In that case hypotheses H_{01}: $\beta_1=0,...,H_{0p}$: $\beta_p=0$ can be tested independently.

Orthogonality is perfect non-association between variables. Independence of variables is desired so that each addition of an independent variable adds to the prediction of the independent variable. If the relationship between independent variables is orthogonal, the overall effect of an independent variable may be partitioned into effects on the dependent variable in an additive fashion.

Models of ANOVA

ANOVA can be modelled by the general linear model.

Model of one-way ANOVA

The model of one-way ANOVA can be written in the following form:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1,...,t, \quad j = 1,...,n_t$$

where

y_{ij} denotes the i-th element of the j-th sample,

μ denotes the "overall population mean",
α_i denotes the effect of the i^{th} treatment, and
ε_{ij} denotes the random error, which is assumed to have $N(0, \sigma^2)$ distribution.
μ_i denotes the of the i^{th} population mean (treatment)

The null hypothesis H_0: $\mu_1 = \mu_2 = ... = \mu_t$ that all population means are equal now corresponds to the null hypothesis that H_0: $\alpha_1 = \alpha_2 = ... = \alpha_t$. This is a linear model and it can be rewritten in a form of a linear regression:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \beta_0 + \beta_1 D_{i2} + \beta_2 D_{i3} + ... + \beta_{t-1} D_t + \varepsilon_{ij}, \quad i=1,2,..,t,\ j=1,2,\ldots, n_t,$$

where the D_i-s are "dummy" variables formed from the independent variables, for example, in the following way:
Let the first group be a "reference" group. Then, let

D_{i2}=1 if an observation belongs in group 2; otherwise let D_{i2}=0.
D_{i3}=1 if an observation belongs in group 3; otherwise let D_{i3}=0.
...
D_{it}=1 if an observation belongs in group t; otherwise let D_{it} =0.

Then, if an observation belongs in group 1:

$$\mu_1 = \beta_0 + \beta_1(0) + \beta_2(0) + ... + \beta_{t-1}(0), \text{ i.e. } \mu_1 = \beta_0.$$

If an observation belongs in group 2, then

$$\mu_2 = \beta_0 + \beta_1(1) + \beta_2(0) + ... + \beta_{t-1}(0) = \beta_0 + \beta_1;$$

hence $\mu_2 = \mu_1 + \beta_1$, and $\beta_1 = \mu_2 - \mu_1$.
Similarly, the other coefficients are $\beta_1 = \mu_2 - \mu_1$, …, $\beta_{t-1} = \mu_t - \mu_1$; i.e. regression coefficients are estimates of the differences between group means. The test of the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ is equivalent to the test of the hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_{t-1} = 0$.

▪ Two-way analysis of variance

In two-way analysis, we wish to assess the effects of two qualitative factors (independent variables) on a dependent variable. We call the groups of a factor the levels of that factor. The goal of two-factor analysis is to estimate and compare the effects of the different factors on the dependent variable. Depending on the particular situation, we may wish to learn whether there are statistically significant differences

a) between the effects of the different levels of factor 1,
b) between the effects of the different levels of factor 2, or
c) between the effects of the different combinations of a level of factor 1 and a level of factor 2. Factors 1 and 2 interact if the relationship between the mean response and the different levels of one factor depends upon the level of the other factor.

Let us denote the numbers of levels of factors 1 and 2 by t and l, respectively, and by N the total number of observations. The two-way ANOVA model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \Theta_{ij} + \varepsilon_{ijk}, \quad i = 1,...,t \quad j = 1,...,l, \quad k = 1,...,n_{ij}$$

where we use the following notations:

y_{ijk}= the k-th observed value of the dependent variable when we are using level i of factor 1 and level j of factor 2,
μ = an overall mean, (unknown constant),
α_i= the effect due to level i of factor 1 (an unknown constant),
β_j= the effect due to level j of factor 2, (an unknown constant),
Θij = the effect due to the interaction of level i of factor 1 and level j of factor 2 (an unknown constant),
ε_{ijk}= the k-th error term when we are using level i of factor 1 and level j of factor 2 (assumed to be distributed as $N(0, \sigma^2)$).

According to the above questions, the following null hypotheses can be tested:

a) $H_{01}: \quad \alpha_1 = \alpha_2 = ... = \alpha_t$

b) $H_{02}: \quad \beta_1 = \beta_2 = ... = \beta_l$

c) $H_{03}: \quad \Theta_{ij} = 0$

In two-way ANOVA, the total sum of squares is decomposed into four terms, according to the effects in the model. The results are generally written into an ANOVA table which contains rows for the effects of factors 1 and 2, the interaction and the error term.

The rows of this tables give the components for the effects of factor 1, factor 2, the interaction and the error term, while the columns contain the sum of squares, the number of degrees of freedom ((*t*-1), (*l*-1), (*t*-1)(*l*-1) and (*N*-*tl*)), the variances (i.e., the ratio of sum of squares and the degrees of freedom), the *F*-values (variance ratio: effect variances to the error variance), and the *p*-value of *F*.

There are three *F*-values in this table according to the three hypotheses.

Question c), i.e. the significance of interaction, H_{03} is tested first. In case of no significant interaction, the significance of each of factors 1 and 2 can be tested separately. If H_{01} is rejected, we can say that at least two of the factor 1 means differ. If *t*, the number of levels of factor 1, is more than two, we again have to use multiple comparisons to find pairwise differences.

In case of a significant interaction is significant, the relationship between the means of factor 1 depends on the level of factor 2. Multiple comparisons can be performed for each combination of one factor with a given level of the other factor. There are special methods against the increase of Type I error, because the use of *t*-tests independently is an incorrect solution.

▪ ANOVA with repeated measurements

The response to a drug treatment, for example, is often measured several times during or after administration of the drug, the intention being to compare treatments with respect to the trends in their effects over time and with respect to their mean levels of response. A widely used and general term is repeated measures data, which refers to data measured repeatedly on subjects either under different conditions, or at different times, or both. In ANOVA with repeated measurements, the repetition is expressed as a factor in the analysis, called the within-subject factor. Multivariate data refer to the case where the same subject is measured on more than one outcome variable. ANOVA with repeated measurements can be modelled using a univariate or multivariate approach. The results of the two approaches are not necessarily the same.

Suppose there are N study units or subjects with n_i measurements for subject i (e.g., n_i longitudinal observations for person i or n_i observations for cluster i). Let $\mathbf{y}_i$ denote the vector of responses for subject i and let $\mathbf{y}$ denote the vector of responses for all subjects

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}, \text{ so y has length } \sum_{i=1}^{N} n_i \ .$$

A normal linear model is

$$E(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\mu}, \quad \mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu},\mathbf{V}),$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$\mathbf{X}_i$ is the $n_i \times p$ design matrix for subject i and $\boldsymbol{\beta}$ is a parameter vector of length p. The variance-covariance matrix for measurements for subject i is

$$\mathbf{V}_i = \begin{bmatrix} \delta_{i11} & \delta_{i12} & \cdots & \delta_{i1_{n_i}} \\ \delta_{i21} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \delta_{in1} & & & \delta_{in_i n_i} \end{bmatrix}$$

and the overall variance-covariance matrix has the block diagonal form

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & 0 & & 0 \\ 0 & \mathbf{V}_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \mathbf{V}_N \end{bmatrix}$$

assuming that responses for different subjects are independent (where **O** denotes a matrix of zeros). Usually the matrices $\mathbf{V}_i$ are assumed to have the same form for all subjects.

There are several commonly used forms for the matrix $\mathbf{V}_i$. For example:

All the off-diagonal elements are equal:

$$\mathbf{V}_i = \delta^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

This is appropriate for clustered data where it is plausible that all measurements are equally correlated, for example, for elements within the same primary sampling unit such as people living in the same area. The term ρ is called the intra-class correlation coefficient. If the off-diagonal term ρ can be written in the form $\sigma_a^2/(\sigma_a^2 + \sigma_b^2)$, the matrix is said to have *compound symmetry* (CS).

First order autoregressive

$$\mathbf{V}_i = \delta^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \\ \vdots & & & \ddots & \\ \rho^{n-1} & \cdots & & \rho & 1 \end{bmatrix}$$

Unstructured correlation matrix: all the correlation terms may be different

$$\mathbf{V}_i = \delta^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & & \rho_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix}$$

The term repeated measures analysis of variance is often used when the data are assumed to be normally distributed. The calculations can be performed using most general purpose statistical software. Sometimes, the correlation structure is assumed to be either spherical or unstructured and correlations which are functions of the times between measurements cannot be modelled.

The generalized linear model

Exponential family of distributions

The distribution of a a single random variable Y belongs to the exponential family if it can be written in the form

$$f(y,\Theta)=\mathrm{s(y)}t(\Theta)e^{a(y)b(\Theta)}$$

where a, b, s and t are known functions and θ is a single parameter of the distribution. The if the above equation can be rewritten as

$$f(y,\Theta)=\exp[a(y)b(\Theta)+c(\Theta)+d(y)]$$

where $s(y) = \exp d(y)$ and $t(\theta) = \exp c(\theta)$. If $a(y) = y$, the distribution is said to be in **canonical** form and $b(\theta)$ is sometimes called the natural parameter of the distribution.

The exponential families include many of the most common distributions. For example, the Poisson, Normal and binomial distributions can all be written in the canonical form.

Generalized linear model

A generalized linear model has three components:

1. Response variables $Y_1, \ldots, Y_N$ which are assumed to share the same distribution from the exponential family;

2. A set of parameters $\boldsymbol{\beta}$ and explanatory variables

$$\boldsymbol{X}=\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} \cdots x_{1p} \\ \vdots \quad\quad \vdots \\ x_{n1} \cdots x_{np} \end{bmatrix}$$

3. A monotone, differentiable function g – called link function such that

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where

$$\mu_i = E(Y_i)\,.$$

- **Examples.**

Normal linear models

A special case of a generalized linear model is the model

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $Y_1, ..., Y_n$ are independent and distributed with $N(\mu_i,\sigma^2)$. The link function is the identity function, $g(\mu_i) = \mu_i$. This model is usually written in the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix}$, and the ε_i's are independent, identically distributed random variables with $N(0,\sigma^2)$ for $i= 1, ...,n$.

Logistic regression model

Consider n independent binary random variables $Y_1, ..., Y_n$ with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. The probability function of Y_i can be written as

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where $y_i = 0$ or 1.

The general linear model is

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}^T\boldsymbol{\beta}$$

where the link function is the logarithm of the **odds** $\pi/(1-\pi)$, called the **logit function**.

This is equivalent to modelling the probability π as

$$\pi = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}^T\boldsymbol{\beta}}}$$

If there is only one x explanatory variable which is also a binary variable, the model has the form

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

As

$$g(1) - g(0) = (\beta_0 + \beta_1 \cdot 1) - (\beta_0 + \beta_1 \cdot 0) = \beta_1$$

and

$$g(1) - g(0) = \ln\frac{\pi(1)}{1-\pi(1)} - \ln\frac{\pi(0)}{1-\pi(0)} = \ln\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \ln(OR)$$

we get, that $e^{\beta_1} = OR$. Here OR is the so called odds ratio. Odds ratio is used in retrospective studies as the approximation of the relative risk.

Relative risk regression model

Consider n independent binary random variables Y_1, ..., Y_n.

The general linear model is

$$g(\pi) = \log(\pi) = \mathbf{x}^T\boldsymbol{\beta}$$

where the link function is the logarithm of the π.

If there is only one x explanatory variable which is also a binary variable, the model has the form

$$g(x) = \ln(\pi(x)) = \beta_0 + \beta_1 x$$

As

$$g(1) - g(0) = (\beta_0 + \beta_1 \cdot 1) - (\beta_0 + \beta_1 \cdot 0) = \beta_1$$

and

$$g(1) - g(0) = \ln \pi(1) - \ln \pi(0) = \ln \frac{\pi(1)}{\pi(0)} = \ln(RR)$$

we get, that $e^{\beta_1} = RR$. Here RR is the so called relative risk. Relative risk is used in prospective studies.

3. Application of generalized linear models to medical problems

The effect of intravenous lactate infusion on cerebral blood flow in Alzheimer's disease

The medical experiment

Intravenous Na-lactate could provoke increased CBF in normal subjects and adults with panic disorder, sometimes with concomitant panic attacks. A self-control design was used and the regional CBF was examined on 20 mild-moderate demented, late-onset, sporadic AD probands. Serum lactate level, blood pressure, venous blood pH, pCO2 and bicarbonate, and serum cortisol levels were measured at 0, 10 and 20 minutes after 0.9 % NaCl or 0.5 M Na-lactate infusion on two separate days.

Statistical model of two parameters are presented here: the venous blood pH and the systolic blood pressure (Figure 1). The other parameters can be analyzed in a similar way.

Statistical models and methods

Mixed models [1] are not especially new but most of the statistical textbooks do not yet include discussion of mixed models. The PROC MIXED of SAS fits a variety of mixed linear models and so they have became one of the most frequently used and cited programs [3,4]. SPSS [5] contains various GLM models and mixed models.

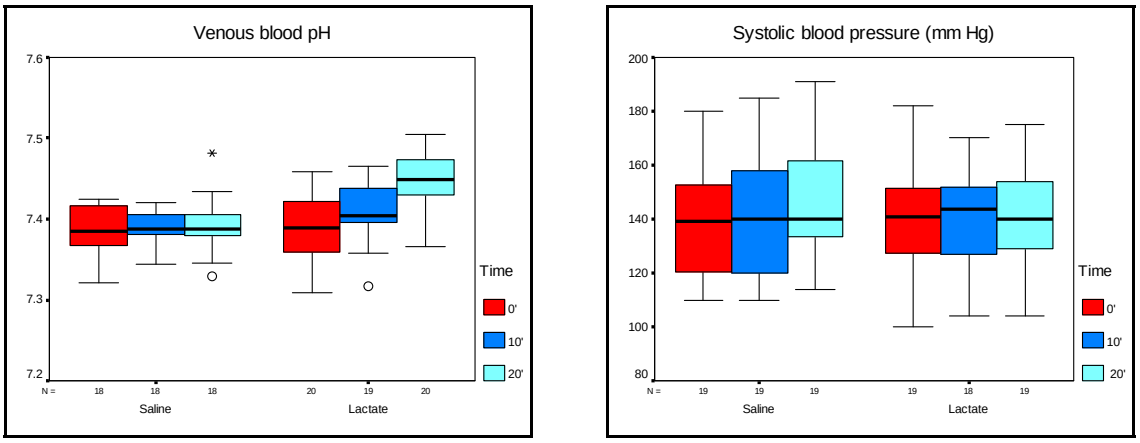

Figure 1. Box plot of the examined parameters for the two treatments in three time points. The boxes indicate the interquartile range of observed data, the line partitioning box corresponds to median observation. The whiskers are drawn to 1.5 times the interquartile range. Points beyond that range are drawn individually.

SAS and SPSS procedures were applied to the following models: univariate statistics, GLM univariate and multivariate tests and mixed models. In ANOVA models, there were two repeated measures factors: treatment (days) with 2 levels (NaCl or Na-lactate) and time with 3 levels (0, 10 and 20 minutes); both factors were fixed.

Responses measured on the same subject are usually correlated; also, variances of repeated measures often change with time. In univariate ANOVA models a special form of the covariance matrix is supposed, namely, the so called sphericity assumption or its special case, the compound symmetry (CS) covariance pattern – assuming equal correlations among all pairs of measures and equal variances of the repeated measurements. In multivariate ANOVA models repeated measures are considered as co-ordinates of a multidimensional vector, here, equal correlations are not required, the covariance pattern is called unstructured (UN). Both univariate and multivariate ANOVA can be performed by the GLM procedure of SAS and SPSS. However, subjects with missing values are ignored, their assumptions about correlation structure are not always realistic and rarely hold.

The method of mixed models can handle missing values, and a wide variety of covariance structures is available, for example, the first-order autoregressive structure {AR(1)}, where measures which are more proximate are more correlated than measures that are more distant. Covariance structures can be compared using several goodness of fit criteria. After selecting the appropriate covariance structure, inference on fixed effects is possible. If the examination the relationship of the response variable with time is in interest, a random coefficients model can be used. Here, regression curves are fitted for each patient and the regression coefficients are allowed to vary randomly between the patients.

For the first parameter a mixed covariance pattern model was found to be the most appropriate with a composite covariance structure, that takes into account the doubly repeated nature of our data. For the second parameter a random coefficients model was used.

Results and discussion

For the first parameter, a significant "treatment by time" interaction was found by univariate,multivariate and mixed models (Table 1 and 2). As a final model, mixed model was used with a composite covariance structure. It was constructed by taking the Kronecker product of an unstructured matrix with a first-order autoregressive type covariance matrix, so we assume equal correlation between treatments and a first-order autoregressive type covariance structure between the three time points.

Table 1. Results of GLM for venous blood pH data

	Univariate ANOVA				Multivariate ANOVA			
	df1	df2	F	p	df1	df2	F	p
Treatment	1	16	11.277	0.004	1	16	11.277	0.004
Time	2	32	20.718	0.000	2	15	19.651	0.000
Treatment*Time	2	32	14.171	0.000	2	15	8.702	0.003

Table 2. Results of mixed models using different covariance structures for venous blood pH data

Covariance structure	Unstructured				Composite UN@AR(1)			
-2 Log L (number of parameters)	-500.4 (21)				-454.2 (4)			
Fixed effects	df1	df2*	F	p	df1	df2*	F	p
Treatment	1	18.8	14.14	0.0013	1	21.8	8.77	0.0073
Time	2	18.6	22.21	<0.0001	2	38.6	15.86	<0.0001
Treatment*Time	2	18.6	10.35	0.0010	2	47.8	14.22	<0.0001

* Satterthwaite approximation for the denominator degrees of freedom

For the second parameter, the increase of mean systolic blood pressure was not obvious by GLM. Because of missing values, results are based on data of only 18 patients. Here, univariate ANOVA results seem to be acceptable – because covariance structure assumptions hold, showing a significant time-effect. However, assumptions of the multivariate approach are more realistic, showing a non-significant time-effect ($p>0.05$). Using mixed models with CS and UN covariance structures, the p-values are closer. A random coefficients model with random coefficients for patients and patients*time was also used to express the relationship of the systolic blood pressure with time. This model resulted in a significant linear time-trend ($p=0.028$).

Table 3. Results of GLM for systolic blood pressure data

	Univariate ANOVA				Multivariate ANOVA			
	df1	df2	F	p	df1	df2	F	p
Treatment	1	17	0.028	0.868	1	17	0.028	0.868
Time	2	34	3.492	0.042	2	16	2.736	0.095
Treatment*Time	2	34	1.433	0.253	2	16	1.424	0.270

Table 4. Results of mixed models using different covariance structures for systolic blood pressure data

Covariance structure	Compound Symmetry				Unstructured			
-2 Res Log L (number of parameters)	858.6 (2)				815.6 (21)			
Fixed effects	df1	df2*	F	p	df1	df2*	F	p
Treatment	1	89	0.02	0.653	1	18	0.14	0.717
Time	2	89	2.93	0.058	2	18	3.70	0.045
Treatment*Time	2	89	1.31	0.276	2	17	2.03	0.163

* Satterthwaite approximation for the denominator degrees of freedom

As a result for the other parameters, the serum lactate levels increased after the Na-lactate infusion and compensatory changes were found in the venous blood pH, pCO_2 and HCO_3 levels [6].

Conclusion

Medical experiments often result in repeated measures data. Using statistical software without knowing their main properties or using only their default parameters may lead to spurious results. Using only the default parameters simple models are supposed. Using carefully chosen statistical model may improve the quality of statistical evaluation of medical data.

Investigation of risk factors of respiratory complications in paediatric anaesthesia

The medical experiment

Perioperative respiratory adverse events (PRAE) remain one of the greatest concerns for the anaesthetist. Although some risk factors have been identified there is a lack of information about the relationship between the child's/family history, the anaesthesia management and the incidence of PRAE.

We prospectively included 9297 children over a 12-month-period having general anaesthesia. Data on the child's/family medical history of asthma, atopy, allergy, upper respiratory tract infection (URI) and passive smoking were collected. Anaesthesia management and all PRAEs were recorded.

Statistical models and methods

Univariate statistics were performed using Mann-Whitney U test and Chi-squared test for continuous and categorical variables, respectively. Multivariate models were developed for perioperative bronchospasm, laryngospasm and all other complications as dependent variables. Having many possible independent candidate variables, model development required variable selection to avoid problems of redundancy and overspecification. The choice of the independent variables in the multivariate models was based on uncorrected p-values of the univariate tests ($p<0.05$) and on medical considerations: some statistically significant variables were not included into the set of candidate independent variables. Also, categorical variables with several categories were transformed to binary variables along the highest relative risk (RR) following the univariate testing. For the different complications, relative risk, absolute risk reduction and 95%CIs were calculated.

It is well known that when the independent variables are correlated, there are problems in estimating model coefficients; the greater the multicollinearity, the greater the standard errors. To avoid multicollinearity, the structure of the correlation of the candidate variables used in the multivariate model was examined first by factor analysis and resulted in five factors

Instead of producing new artificial variables by factor analysis, we collapsed original variables belonging to the factors using the „or" logical operator. These collapsed variables were used in the multivariate analyses together with age and airway management. Multivariate analysis was performed by relative risk regression, since this method is appropriate for modelling the risk factors of prospective studies. It involves a generalized linear model with log link function and binomial dependent variable. Model fit was assessed via likelihood ratio test using stepwise elimination process variables, possible

interactions with age and some medically plausible interactions were also examined. Variables and their interactions were retained in the model if they significantly improved the model fit using the likelihood ratio test.

Results and discussion

Here we would like to show results of multivariate modelling. Other details can be read in the published paper [7].

In univariate models, possible risk factors were examined separately; variables and the univariate results are shown in Table 5. These variables were highly correlated. To avoid multicollinearity in multivariate modelling, the correlation structure was examined by factor analysis, which resulted in four factors. Instead of producing new artificial variables by factor analysis, we collapsed original variables belonging to the factors using the „or" logical operator. In multivariate models, age, gender, hayfever, airway management (TT, LMA or face mask) and the new collapsed variables (airway sensitivity, eczema, family history and anaesthesia) were examined. As a result of multivariate analyses (Table 5), some variables were not significant. The interactions with age and the following, medically plausible interactions were also not significant: airway sensitivity by anaesthesia and airway sensitivity by airway management (TT, LMA or face mask).

Table 5. Relative risk and 95% confidence interval (CI) for the risk factors of the occurrence for perioperative bronchospasm.

Variable	Univariate				Multivariate			
	p	RR	95%CI		p	RR	95%CI	
Age	0.325	0.985	0.956	1.015	-	-	-	-
Gender	0.004	0.667	0.505	0.882				
Hayfever	< 0.0001	2.915	2.153	3.947				
Upper respiratory tract infection (URI) <2 weeks	<0.0001	2.146	1.498	3.075				
Wheezing at exercise	<0.0001	7.730	5.870	10.178				
Wheezing >3 times in the last 12 months	<0.0001	7.168	5.307	9.680				
Nocturnal dry cough	<0.0001	10.510	7.932	13.927				
Airway sensitivity	**<0.0001**	**8.463**	**6.179**	**11.590**	**< 0.0001**	**5.653**	**4.089**	**7.816**
Eczema in the last 12 months	<0.0001	3.158	2.359	4.227				
Ever eczema	<0.0001	4.575	3.444	6.077				
Eczema	**<0.0001**	**4.533**	**3.416**	**6.016**	**<0.0001**	**2.601**	**1.950**	**3.470**
Asthma in the family, >2 persons	<0.0001	4.415	3.082	6.325				
Hayfever in the family, >2 persons	<0.0001	3.753	2.426	5.808				
Eczema in the family, >2 persons	0.028	2.190	1.089	4.401				
Smoking in the family, Mother and Father	<0.0001	2.603	1.894	3.576				
Family history	**<0.0001**	**2.932**	**2.212**	**3.887**	**<0.0001**	**1.863**	**1.413**	**2.458**
Airway managed by registrar vs. pediatric anesthesia consultant	<0.0001	3.847	2.473	5.984				
Inhalational induction of anesthesia	<0.0001	2.381	1.791	3.167				
Change of anesthesiologist during airway management	<0.0001	4.094	2.646	6.335				
Anesthesia	**<0.0001**	**3.872**	**2.163**	**6.929**	**<0.0001**	**3.078**	**1.727**	**5.484**
ENT surgery	0.043	1.458	1.012	2.101	-	-	-	-
Face mask vs. laryngeal mask (LMA)	0.118	1.933	0.846	4.418	0.304	1.538	0.677	3.493
Face mask vs. tracheal tube (TT)	<0.0001	5.105	2.252	11.574	0.002	3.523	1.564	7.937

Interpretation

This study identified from the child's/family medical history risk factors increasing the risk for PRAE. These children can be systematically identified at the preanaesthetic assessment and thus benefit from a specifically targeted anaesthesia management.

4. Conclusion

In this paper, we gave an introduction by examples to the theory, special properties, as well as some qualitative methods for generalized linear models. Our examples illustrated that the advanced techniques of biostatistics can help in developing the theory of medicine and might have an impact to the practice of curing.

5. Acknowledgement

Research supported by the Hungarian National Foundation for Scientific Research Grant No. T 049516, and cofinanced by the European Union through the Hungary-Serbia Cross-border Cooperation programme in the frame of the project IPA HU-SRB/0901/221/088.

References

[1] Brown H. and Prescott R.: Applied Mixed Models in Medicine. John Wiley & Sons, Chichester, 2001.

[2] Dobson A.J.: An Introduction to Generalized Linear Models. Chapman&Hall London · New York · Tokyo ·Melbourne · Madras, 1991.

[3] SAS Institute, Inc: The MIXED procedure in SAS/STAT Software: Changes and Enhancements through Release 6.11. Copyright © 1996 by SAS Institute Inc., Cary, NC 27513.

[4] Park T., and Lee Y.J.: Covariance models for nested repeated measures data: analysis of ovarian steroid secretion data. Statistics in Medicine 21: 134-164, 2002.

[5] SPSS 11.0. Copyright © SPSS Inc., Chicago IL, 1989-2001.

[6] Kálmán J., Palotás A., Kis G., Boda K., Túri P., Bari F., Domoki F., Dóda I., Árgyelán M., Vincze G., Séra T., Csernay L., Janka Z., Pávics L.: Regional cortical blood flow changes following sodium lactate infusion in Alzheimer's disease. European Journal of Neuroscience 21(6):1671-8, 2005.

[7] von Ungern-Sternberg BS., Boda K., Chambers NA., Rebmann C ., Johnson C., Sly PD, Habre W.:: Risk assessment for respiratory complications in paediatric anaesthesia: a prospective cohort study, The Lancet, 376 (9743): 773-783, 2010.