© Journal for Modeling in Ophthalmology 2016;2:89-96
Original article

Role of ocular perfusion pressure in glaucoma: the issue of multicollinearity in statistical regression models

Alessandra Guglielmi[1], Giovanna Guidoboni[2], Alon Harris[3]

[1]*Dipartimento di Matematica, Politecnico di Milano, Italy*, [2]*Department of Mathematical Sciences, Indiana University Purdue University Indianapolis, Indianapolis, USA*, [3]*Eugene and Marilyn Glick Eye Institute, Indianapolis, IN, USA*

Abstract

Purpose: Intraocular pressure (IOP), mean arterial pressure (MAP), systolic blood pressure (SYS), diastolic blood pressure (DIA), ocular perfusion pressure (OPP) are important factors for clinical considerations in glaucoma. The existence of linear relationships among these factors, referred to as multicollinearity in statistics, makes it difficult to determine the contribution of each factor to the overall glaucoma risk. The aim of this work is to describe how to account for multicollinearity when applying statistical methods to quantify glaucoma risk.
Methods: Logistic regression models including multicollinear covariates are reviewed, and statistical techniques for the selection of non-redundant covariates are discussed. A meaningful statistical model including IOP, OPP and SYS as non-redundant covariates is obtained from a clinical dataset including 84 glaucoma patients and 73 healthy subjects, and is used to predict the probability that new individuals joining the study may have glaucoma, based on the values of their covariates.
Results: Logistic models with satisfactory goodness-of-fit to the clinical dataset include age, gender, heart rate and either one of the following triplets as covariates: (*i)*(SYS, DIA, OPP); (*ii)* (IOP, SYS, OPP); (*iii)* (IOP, SYS, DIA); or (*iv)* (IOP, SYS, MAP). Choosing triplet (*ii)*, higher disease probabilities are predicted for higher IOP levels. Similar predictions in terms of disease probability can be obtained for different combinations of OPP, SYS and IOP.

Correspondence: Alessandra Guglielmi, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy.
E-mail: alessandra.guglielmi@polimi.it

Conclusion: Multicollinearity does not allow to clearly estimate the single effect of an individual covariate on the overall glaucoma risk. Instead, statistically assessing the combined effects of IOP, OPP, and blood pressure provide useful predictions of disease probability.

Keywords: glaucoma, generalized linear models, logistic regression, multicollinearity, statistical methods, disease probability

1. Introduction

It is well known that elevated intraocular pressure (IOP) is a recognized risk factor for glaucoma. Several other glaucoma risk factors have been suggested, among which low blood pressure deserves particular mention. In order to combine the effects of elevated IOP and low blood pressure, a synthetic index called ocular perfusion pressure (OPP) has been proposed. The index is defined as $\mathrm{OPP} = (2/3)\mathrm{MAP} - \mathrm{IOP}$, where the mean arterial pressure (MAP) is defined as a linear convex combination of the systolic pressure (SYS) and diastolic pressure (DIA), namely $\mathrm{MAP} = (1/3)\mathrm{SYS} + (2/3)\mathrm{DIA}$. Thus, low values of the index OPP may be due to low MAP, elevated IOP or a combination of the two. Whether and to what extent IOP, OPP, MAP, SYS and DIA should be considered as risk factors for glaucoma is still a matter of debate in glaucoma research[1–3]. The present article considers this question from the statistical viewpoint and provides directions to its answer. The existence of formulas relating IOP, OPP, MAP, SYS and DIA is indicative of an issue that in statistics is known as *multicollinearity*, occurring when one or more covariates are defined as a function of the remaining variables. In this paper, we consider this issue from the theoretical viewpoint and provide examples from a real clinical dataset. Our analysis shows that it is the joint effect of all the covariates in the selected logistic model that determines the glaucoma risk, rather than the value of an individual covariate.

2. Methods

2.1 Description of the dataset

Our dataset contains $n = 157$ individuals, including 84 glaucoma patients and 73 healthy subjects. The data were collected within the Indianapolis Glaucoma Progression study and other clinical studies at *Eugene and Marilyn Glick Eye Institute*, Indianapolis (USA), directed by Prof. Alon Harris. The final goal of our statistical analysis is to identify a meaningful set of covariates, i.e. clinical parameters, that provide a good estimate of the probability that a new individual joining the study is a healthy subject or is suffering from glaucoma.

Let us introduce a *glaucoma indicator* for each individual in the dataset. Let $i = 1, \ldots, n$, with $n = 157$, be the index identifying each individual in the dataset, and let

y_i be the glaucoma indicator for the i-th individual, with $y_i = 1$ if the i-th individual suffers from glaucoma and $y_i = 0$ otherwise. The set of covariates considered for this analysis are: age in years (Age), Gender (1 if female, 0 if male), heart rate (HR), IOP, SYS, DIA, MAP and OPP. Empirical means of these variables and standard deviations for continuous covariates are reported in Table 1. There are 88 women and 69 men in the sample.

Table 1. Empirical means of all the covariates in the dataset; standard deviations for continuous variables are given between brackets.

Age	Gender	HR	IOP
59.95 (11.38)	88 F (69 M)	71.22 (12.60)	16.15 (4.00)

SYS	DIA	MAP	OPP
128.90 (18.32)	83.13 (11.21)	98.41 (12.46)	49.45 (9.39)

We recall that the following linear relationships exist among some covariates:

$$\mathrm{MAP} = \frac{1}{3}\mathrm{SYS} + \frac{2}{3}\mathrm{DIA}, \quad \mathrm{OPP} = \frac{2}{3}\mathrm{MAP} - \mathrm{IOP} = \frac{2}{9}\mathrm{SYS} + \frac{4}{9}\mathrm{DIA} - \mathrm{IOP}, \quad (1)$$

meaning that MAP, SYS, DIA, OPP and IOP are *multicollinear covariates*. Lack of awareness of multicollinearity may yield erroneous interpretation of statistical results[4]. A nice concise, but non-technical, overview of statistical problems that may be encountered when covariates are multicollinear can be found in Tu et al[5].

2.2 Logistic regression models and multicollinearity

The use of linear regression models to investigate the effect of IOP, MAP and OPP has recently been questioned. In this section, we aim at clarifying the main points of this debate. Let us assume that the glaucoma indicator y_i for each individual is the realization of a random variable Y_i, and that the individuals are independent. Let us denote by π_i the probability that $Y_i = 1$ and let us assume that π_i can be described by a logistic regression model, so that we can write

$$P(Y_i = 1) := \pi_i, \quad \log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \ldots, n, \quad (2)$$

where x_{ij} is the value assumed by the j-th covariate in the i-th individual. For example, the model discussed in Khawaja et al[6] considers two covariates, i.e. $p = 2$, with x_{i1} =IOP and x_{i2} =OPP for each patient i. In (2), the coefficients β_j represent the effect of the j-th covariate on the response (glaucoma indicator). Note that all the parameters β_j, for $j = 0, 1, \ldots, p$, are unknown and they are usually estimated from the dataset $\{(y_i, x_{i1}, , x_{i2}, \ldots, , x_{ip}), i = 1, \ldots, n\}$ via standard statistical techniques, such as maximum likelihood estimate (MLE). The estimated values of the parameters are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. Since individuals are randomly sampled from a larger population (i.e. from the world population), and a vector of covariate values identifies subpopulations of individuals, from (2), it follows that each exponential e^{β_j}, for

$j = 1, \ldots, p$, represents the conditional odds ratio of two subpopulations of individuals, one having the value of the j-th covariate fixed at some value $x^* + 1$, and the other having the value of the same j-th covariate equal to x^* namely

$$e^{\beta_j} = \frac{\text{odds if the covariate } x_{ij} \text{ is incremented by 1}}{\text{odds if the covariate } x_{ij} \text{ is not incremented}}, \tag{3}$$

while keeping all other covariates fixed (i.e. adjusting for the other covariates). It is common practice to drop the index i from the notation and simply write

$$x_j := x_{ij}, \quad Y := Y_i, \quad y := y_i \, .$$

Thus, in mathematical terms, we can rewrite (3) as

$$e^{\beta_j} = \frac{P(Y = 1 | x_j = x + 1, x_l = x_l^*) \,/\, (1 - P(Y = 1 | x_j = x + 1, x_l = x_l^*))}{P(Y = 1 | x_j = x, x_l = x_l^*) \,/\, (1 - P(Y = 1 | x_j = x, x_l = x_l^*))} \tag{4}$$

where x_l^* are fixed values, with $l = 1, \ldots, p$ and $l \neq j$. Thus, β_j quantifies the change in the response variable Y (whose realization corresponds to the glaucoma indicator y) for a unit change in the covariate x_j when the rest of the covariates are fixed (or no other covariates are present). In this perspective, β_j is usually interpreted as the effect of x_j on the response (glaucoma indicator), while adjusting for the other covariates, and $\hat{\beta}_j$ is its estimated value.

This interpretation of β_j does not extend to the case of multicollinear covariates, which is indeed the case for IOP, MAP and OPP, as shown by the relationships in (1). Let us clarify this issue by means of a simple example. Let us suppose that the glaucoma indicator depends only on two covariates, say IOP and OPP. Thus, in this case $p = 2$, $x_1 =$IOP and $x_2 =$OPP. According to (2), dropping the index i, the logarithm of the odds of having glaucoma, i.e. $\log(\pi/(1-\pi))$, is given by:

$$\log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 \times \mathsf{IOP} + \beta_2 \times \mathsf{OPP} \tag{5}$$

thereby suggesting that β_1 represents the effect of IOP (in the logit scale) on the glaucoma indicator for OPP fixed and that β_2 represents the effect of OPP on the glaucoma indicator for IOP fixed. However, IOP and OPP are related to each other via (1); thus, for IOP fixed, OPP can vary only if MAP varies, yielding:

$$\begin{aligned} \log \frac{\pi}{1 - \pi} &= \beta_0 + \beta_1 \times \mathsf{IOP} + \beta_2 \times OPP \\ &= \beta_0 + \beta_1 \times \mathsf{IOP} + \beta_2 \times \left(\frac{1}{3} \mathsf{MAP} - \mathsf{IOP} \right) \\ &= \beta_0 + \frac{\beta_2}{3} \times \mathsf{MAP} + (\beta_1 - \beta_2) \times \mathsf{IOP}. \end{aligned}$$

The last line suggests that the effect of IOP on the glaucoma indicator is represented by the difference $\beta_1 - \beta_2$ for MAP fixed, which is a different conclusion than that suggested by (5). The controversial aspect related to the interpretation of logistic regression parameters for multicollinear covariates in glaucoma has been discussed in recent works[2,6]. However, Khawaja et al[6] erroneously argue that the intrinsic relationship between IOP and OPP precludes any useful interpretation of OPP as glaucoma risk factor, whereas the issue is just that a different statistical approach should be used to properly account for the relationship between IOP and OPP when analizing clinical data, as discussed in the next section.

2.3 Accounting for multicollinearity in statistical analysis

Multicollinearity does not allow, in general, to interpret the regression parameters e^{β_j} and their estimates $e^{\hat{\beta}_j}$ as the effects of variations of single covariates, while keeping all the others fixed. However, when analyzing a dataset, the main statistical question should not be: "what is the meaning of the regression parameters in the logistic model?", rather "which covariates should be included in the logistic regression in order to obtain a statistical model capable of predicting the response with good accuracy?". The latter question does indeed make sense also in the case of multicollinear variables, as discussed in classical Statistics textbooks (see, for istance, Section 4.6 of the book by Agresti[4]).

For the specific example involving IOP and OPP discussed above, the statistical question should not be whether or not β_2 describes the effect of OPP on the glaucoma indicator for fixed IOP (the answer is obviously *no* since OPP and IOP are intrinsically related); rather, the *real statistical question* is whether OPP, IOP and blood pressure should all be considered as risk factors in glaucoma. On the ground of statistical tools, we prove that the answer to the last question is positive. In order to show this, we need to: identify redundant covariates for the determination of the glaucoma indicator *(Step1)*; obtain statistical models that include only non-redundant covariates and provide good estimates of the probability of having glaucoma for a new individual joining the study *(Step 2)*.

Step 1. The dependency of x_j on the other covariates can be quantified using the variance inflation factor (VIF). For the covariate x_j, this factor is defined as $\text{VIF}_j = 1/(1-R_j^2)$, where R_j^2 denotes the value of the index $R^2 \in (0,1)$ in a linear regression model where the value of x_j is determined by the other covariates (see, for istance, Section 4.6.5 of the book by Agresti[4]). If x_j is predicted very well by the other covariates in the linear model, then $R_j^2 \approx 1$ (the higher R^2, the best is the corresponding linear model in predicting the output); as a consequence, VIF_j in this case will be large. As a rule of thumb, the covariate x_j is considered to be redundant if $\text{VIF}_j > 10$. By applying this simple rule to the covariates in the dataset described in Section 2.1, we found that the VIF values were larger than 10 for IOP, SYS, DIA, MAP and OPP, as expected.

Step 2. The outcomes of Step 1 imply that logistic models for the glaucoma in-

dicator can include Age, Gender, HR, and only some covariates among IOP, SYS, DIA, MAP and OPP. Actually, we could select any three covariates among the five above, and would obtain a statistically significant model for any of these choices. The software R is able to select the covariates through a stepwise backward eliminating procedure that starts from a complex model fitted to the dateset and sequentially removes terms, such as the largest p-value in a test of significance, or the least deterioration in the Aikake Information Criterior (AIC), which is a statistical tool to measure goodness-of-fit[4]. For the dataset described in Section 2.1, the software R found that, using the stepwise backward eliminating procedure, the best model includes Age, Gender, HR, SYS, DIA and OPP as covariates. The AIC in this case is optimal, but the same optimal value of AIC is obtained when, in addition to Age, Gender, HR, we select either (IOP, SYS, OPP) or (IOP, SYS, DIA) or (IOP, SYS, MAP). For all these four models, the VIF values for the covariates included in the model are similar (and all smaller than 10). In conclusion, these four models are equivalent in terms of goodness-of-fit measures and measure of dependency among covariates. Therefore, we can choose any of these four models in order to predict the probability of having glaucoma for a new individual entering the study with given values of the selected covariates.

3. Results

The outcomes of Steps 1 and 2 confirm that it makes perfect sense to consider either (SYS, DIA, OPP), or (IOP, SYS, OPP), or (IOP, SYS, DIA), or (IOP, SYS, MAP) as sets of covariates in order to predict glaucoma probability, despite the existence of functional relationships between them. In this section, we consider the model in equation (2) for $p = 6$ and covariates x_{ij}, $j = 1, \ldots, 6$, given by Age, Gender, HR, IOP, SYS and OPP, respectively, measured for all patients i in the dataset; as usual, independence among patients is assumed. The fitted coefficients are $\hat{\beta}_0 = -16.905$, $\hat{\beta}_1 = 0.107$, $\hat{\beta}_2 = -1.160$, $\hat{\beta}_3 = -0.036$, $\hat{\beta}_4 = 0.177$, $\hat{\beta}_5 = 0.120$, $\hat{\beta}_6 = -0.089$. We use this model to predict the probability of having glaucoma for a new female patient, aged 60 and with HR= 71, joining the study. Table 2 shows that, for given values of OPP, SYS, Age and Gender, different disease probabilities are predicted depending on the level of IOP. In particular, higher IOP levels correspond to higher probabilities of having glaucoma. On the other hand, similar predictions in terms of disease probability can be obtained for different combinations of OPP, SYS, IOP (within the ranges of values in our dataset), suggesting that these covariates should all be considered as important risk factors in glaucoma.

We remark that including diastolic blood pressure in the model would also be an option, as indicated by the cases (SYS, DIA, OPP) and (IOP, SYS, DIA) at the beginning of Section 3. However, we cannot simultaneously include all the covariates in the same statistical model because of the relationship among them, i.e. multicollinearity. It is important to emphasize that removing some of the covariates from the statistical model does not mean that the model does not account for that covariate; rather,

Table 2. Predicted disease probabilities for new female patients, aged 60, HR= 71. For OPP and SYS fixed, higher values of IOP correspond to higher probabilities of having glaucoma (left side of the Table). However, similar disease probabilities are obtained for different values of the covariates (right side of the Table).

OPP	SYS	IOP	Disease prob	Disease prob	OPP	SYS	IOP
43	129	12	0.678	0.673	52	134	13
43	129	16	0.811	0.809	34	132	10
43	129	20	0.897	0.899	20	115	20

variations in that covariate are accounted for through variations in the other collinear variables in the model.

4. Conclusions

The main question motivating our work is whether IOP, OPP and blood pressure should all be interpreted as risk factors in glaucoma. Based on the statistical techniques and analysis reported in this article, our answer is that it is the *joint effect* of IOP, OPP and blood pressure, or, more precisely, of all the covariates in the selected logistic model, that determines the probability of disease, rather than the value of an individual covariate. Importantly, the main statistical interest should be the prediction of disease probabilities for new patients entering the study, presenting specific values of the covariates included in the model, rather than the estimated individual effect of a single predictor.

References

1. Costa V, Harris A, Anderson D, Stodtmeister R, Cremasco F, Kergoat H, et al. Ocular perfusion pressure in glaucoma. Acta Ophthalmol, 2014;92(4): e252–66.
2. Khawaja AP, Crabb DP, Jansonius NM. Time to abandon over-simplified surrogates of ocular perfusion pressure in glaucoma research. Acta ophthalmologica, 2015;93(1): e85–e86.
3. Costa V, Anderson D, Harris A. Surrogates for ocular perfusion pressure are not perfect. Acta Ophthalmol, 2015;93(1): e86–7.
4. Agresti A. Foundations of Linear and Generalized Linear Models. John Wiley & Sons, 2015;
5. Tu YK, Kellett M, Clerehugh V, Gilthorpe M. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. British dental journal, 2005;199(7): 457–461.
6. Khawaja AP, Crabb DP, Jansonius NM. The Role of Ocular Perfusion Pressure in Glaucoma Cannot Be Studied With Multivariable Regression Analysis Applied to SurrogatesLetters. Investigative ophthalmology & visual science, 2013;54(7): 4619–4620.