Nomographic representation of logistic regression models: A case study using patient self-assessment data

Stephan Dreiseitl a,*, Alexandra Harbauer b, Michael Binder b, Harald Kittler b

a Department of Software Engineering, University of Applied Sciences Upper Austria at Hagenberg, Austria
b Department of Dermatology, Medical University of Vienna, Austria

Received 24 December 2004
Available online 17 March 2005

Abstract

Logistic regression models are widely used in medicine, but difficult to apply without the aid of electronic devices. In this paper, we present a novel approach to represent logistic regression models as nomograms that can be evaluated by simple line drawings. As a case study, we show how data obtained from a questionnaire-based patient self-assessment study on the risks of developing melanoma can be used to first identify a subset of significant covariates, build a logistic regression model, and finally transform the model to a graphical format. The advantage of the nomogram is that it can easily be mass-produced, distributed and evaluated, while providing the same information as the logistic regression model it represents.

1. Introduction

The use of nomograms as graphical representations of the calculations in simple models has a long history in engineering and medicine [1,2]. In the last decades they have been replaced, to a great extend, by the increased availability and use of computers. By being paper-based and not requiring an electronic device for calculations, nomograms have advantages in situations where the precision of electronic devices is not required, and when graphical evaluation is easier to perform than numerical computation.

Logistic regression models are well suited to be represented by nomograms. Several instances of replacing more complex logistic regression models with simpler nomograms have been reported in the literature. Some of the more recent examples are as follows: Ohori et al. [3] developed a nomogram for staging prostate cancer; a nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients was derived by Van Zee et al. [4]; and a nomogram was used to identify the decreased risk of positive pelvic lymph nodes in patients with prostate cancer [5]. Other clinical applications of nomograms derived from logistic regression models are available in the literature [6–13].

In general, nomographic depictions of logistic regression models are given as a series of straight lines with a common linear scale, with the scale factors of the individual lines given by the coefficients of the covariates in the model. The graphical evaluation of the model consists of locating the values of the covariates on the respective lines, and determining corresponding point values on the common point scale. As the last step, the total point sum is converted to a probability by graphical conversion to a probability scale. This process is straightforward and easy to apply for any number of covariates in the model. A disadvantage of this evaluation is the fact that the individual score contributions have to be summed up by hand.

* Corresponding author. Fax: +43 7236 3888 2099.
E-mail address: Stephan.Dreiseitl@fh-hagenberg.at (S. Dreiseitl).
The objective of this paper is to show how an alternative form of nomogram can be constructed to represent logistic regression models. In our construction, there is no need to add up probability scores, as all calculations are replaced by the drawing and intersection of straight lines. We illustrate the general process of nomogram construction by using data from a patient self-assessment study on the risks of developing melanoma. This data, which had been published previously [14], and the logistic regression model derived from it, are presented in Section 2. In Section 3, the process of converting a logistic regression model to a nomogram is explained in detail, and applied to the logistic regression model derived earlier. Advantages and disadvantages of the resulting nomograms, compared to the traditional approach, and possible applications in clinical medicine are discussed in Section 4.

2. Derivation of logistic regression model

The data that is the basis for the logistic regression model and nomogram derived in this paper were obtained in a study on the ability of patients to self-assess their risk of developing melanoma [14]. The motivation for that study was the observation that the early diagnosis of melanoma is crucial for successful treatment [15]. This early detection and diagnosis of melanoma is only possible if high-risk patients consult physicians regularly; reliable self-assessment could help to identify such high-risk patients.

Almost all studies on patient self-assessment use paper-based questionnaires to obtain answers. While this format provides a means for comprehensive information elicitation, it might in some instances be beneficial to replace lengthy questionnaires with a visual decision aid, such as a nomogram.

Patient self-assessment is more easily possible in dermatology than in other medical domains, because the values of contributing factors can be determined visually [16–19]. Nevertheless, some studies on patient self-assessment have been carried out in other areas, such as orthopedics [20] and cancer care [21]. The results of these studies vary, with some reporting agreement between patient and physician assessment [16,21], while others show no such agreement [17–20].

In our own study on patient self-assessment [14], 202 cases randomly selected from a group of patients with diagnosed melanoma were matched with 202 controls selected at random from patients consulting a general practitioner in Vienna. All 404 study participants were required to complete a 10-item questionnaire asking about factors that are known to be associated with higher risk of developing melanoma. The questions asked for the following information: hair color, eye color, skin type, skin reaction to sun exposure, history of sunburns, affinity for sun exposure, affinity for artificial tanning, personal assessment of skin damage due to sun exposure, estimate of number of nevi, and number of large congenital nevi. All responses were either nominal or ordinal, with the number of nevi discretized into five ordinal categories. The answers to the questions relating to sun exposure and artificial tanning were combined into one variable representing sun affinity. This left eight answers to consider.

After completing the questionnaire, the study participants were examined by a dermatologist. The physician determined the skin type, the amount of skin damage due to sun exposure, the number of nevi and large congenital nevi, and family history of skin cancer.

Based on the patients’ answers to the questionnaire, we built a multi-variate logistic regression model to predict the presence or absence of melanoma in the study group. The discriminatory power of this model, measured by the area under the ROC curve, was 0.76 (95% CI 0.74–0.78).

A main focus of the previous study was to identify which of the 10 questions asked are the most relevant, in the sense that a model build from the answers to these questions is as good as a model build from the whole questionnaire. Using logistic regression analysis, we identified three risk factors that are all independently associated with the melanoma group: skin damage, skin type, and nevi count. A reduced model using these three factors is not significantly different from the full model, with the area under the ROC curve being 0.73 (95% CI 0.68–0.77, \( p = 0.10 \)). This indicates that patients are quite proficient at determining their own risk state, when compared to the physician diagnosis. Unfortunately, as evidenced by the rather low areas under the ROC curves, neither physicians nor patients can provide a highly accurate assessment of the risk state.

In summary, the main findings of the previous study are that

- patient self-assessment is possible, as the discriminatory power of the model based on patient answers is not significantly different from the model based on physician evaluation; and
- a reduced model with three covariates is not significantly different from the full model with eight covariates.

The second observation means that a three-variable nomogram, as a graphical representation of the logistic regression model with three covariates, has the same discriminatory power as a 10-item questionnaire.
3. Nomogram construction

In this section, we will show how the three-covariate logistic regression model we had obtained previously can be represented graphically in a novel way. The nomogram construction presented here exploits the fact that a multi-variate logistic regression model represents a linear relationship between the covariates; the non-linearity of the model is the result of applying the non-linear logistic function to this linear relationship. For two covariates \( x_1 \) and \( x_2 \), this means that all combinations of \( x_1 \) and \( x_2 \) that result in a given (constant) probability estimate \( p(x_1, x_2) = c \) lie on a straight line. To make this process clear, consider, as an example, the logistic regression model

\[
p(x_1, x_2) = \frac{1}{1 + e^{-(2 - 0.4x_1 + x_2)}}.
\]

If the two covariates are represented by two parallel axes, all lines connecting \( x_1 \) and \( x_2 \) values with \( p(x_1, x_2) = c \) intersect in one point. Fig. 1A shows some points \((x_1, x_2)\) with \( p(x_1, x_2) = 0.5 \) joined by straight lines. Using the model in Eq. (1) with this constraint, we see that these are the lines joining points for which \( x_2 = 2 - 0.4x_1 \) holds. It can, for example, be observed that \( p(0, 2) = 0.5 \), and that \( p(1, 1.6) = 0.5 \). All these lines intersect in one point. For different values of \( c \), the intersection points of these imaginary lines lie on a straight line, forming a graphical probability scale. For the model above, this probability scale is shown in Fig. 1B. The output of the logistic regression model in Eq. (1) can be calculated graphically by drawing a straight line through the two values of the covariates. The intersection of this line with the probability line gives the model output. Note that because the nomogram is derived from a logistic regression model, which is non-linear in the covariates, the final output scale is not linear.

The construction becomes more complex with each additional covariate. To include a third covariate \( x_3 \) in the calculations, we observe that an intersection point between \( x_1 \) and \( x_2 \) can be seen as a new auxiliary covariate that encodes the linear relationship between \( x_1 \) and \( x_2 \). Using this new covariate, one can repeat the construction process outlined in Fig. 1, with the new covariate and \( x_3 \) taking the places of \( x_1 \) and \( x_2 \). A graphical representation of this process is shown in Fig. 2.

To evaluate this model, one has to first intersect the line joining two covariate values \( x_1 \) and \( x_2 \) with the auxiliary axis (call this point \( A \)). Then, one has to find the intersection of the line joining \( A \) and the \( x_3 \) value with the final probability axis and read off the model output. Note that this evaluation process uses only straight lines; thus, the nomogram can be stretched arbitrarily in \( x \)- and \( y \)-direction without effecting the evaluation.

In the following, we will give a more detailed example using the logistic regression model based on the three significant covariates obtained from the self-assessment questionnaire. These covariates are

- the nevi count with five ordinal values (0–5/5–10/10–20/20–50/>50),
- the skin type with four ordinal skin phototypes according to Fitzpatrick [22], and
- the skin damage due to sun exposure with three ordinal values (no/minor/major damage).

The same three factors were identified as significant in a logistic regression model based on the dermatologist’s examination of the study subjects. From the patients’ data, we obtained the model

\[
p(x_1, x_2, x_3) = \frac{1}{1 + e^{-(2.9 + 0.38x_1 + 0.485x_2 + 0.782x_3)}},
\]

391

Fig. 1. Nomogram construction for two covariates using the model in Eq. (1). (A) Five sets of points \((x_1, x_2)\) with the linear relationship \( x_2 = 2 - 0.4x_1 \) are joined by straight lines. Observe how the lines intersect in exactly one point: this is the point for which \( p(x_1, x_2) = 0.5 \). (B) All imaginary intersection points for various probability levels are joined by a straight line. This line represents the output of the logistic regression model. Evaluating the nomogram for \( x_1 = 3 \) and \( x_2 = 1 \) yields a value of \( p(3, 1) \approx 0.55 \).
with $x_1$ being the nevi count, $x_2$ the skin type, and $x_3$ the skin damage. The covariate encoding started at 0, so that $x_1$ (nevi count) takes values in the set \{0, \ldots, 4\}, $x_2$ (skin type) in the set \{0, \ldots, 3\}, and $x_3$ (skin damage) in the set \{0, 1, 2\}. Note that the skin phototypes are labeled in reverse order of melanoma risk; this means that skin type IV is represented as $x_2 = 0$, down to skin type I with an encoding of $x_2 = 3$. Following the construction process outlined above, an auxiliary axis between the first two covariates (nevi count and skin type) was first calculated, followed by the inclusion of the skin damage covariate in the model.

The nomogram obtained from the model in Eq. (2) is shown in Fig. 3. It must be emphasized that Fig. 3 is a direct representation of Eq. (2); this means that the discriminatory power of the graphical model is exactly the same as the power of the logistic regression model.

Because of the values of the covariates in the logistic regression model, the final probability estimate axis is squashed, with the axis tick marks representing 0.1 increments in probability estimates being closer in the middle of the axis. Although this limits the resolution of the nomogram around $p = 0.5$, the graphical representation is sufficient for separating patients into low-risk vs. high-risk groups. ROC analysis reveals how “low-risk” and “high-risk” groups could be defined: at a specificity of 90%, the logistic regression model (and thus the nomogram) has a sensitivity of 39%; in our model the threshold that achieves these values is at $p = 0.69$. The low sensitivity value limits the usefulness of the nomogram for mass melanoma screening; however, the general nomogram construction process presented here can be useful as a graphical representation of other models with higher discriminatory power.

It has to be cautioned that the probability value calculated in Eq. (2) and Fig. 3 is derived from a data set with melanoma prevalence of 50% (202 cases and 202 controls). To obtain an accurate probability of belonging to the melanoma group in a general population with much lower prevalence, the intercept of the logistic regression model has to be adjusted [23]. Since this adjustment only affects the intercept value, the nomogram remains the same, with the exception of the labeling on the output axis. In our case, the resulting model output range will then be much lower, depending on the population and hence prevalence in question.

As an example of the application of the nomogram, consider a patient using the nomogram to evaluate his or her risk state. Since the disease prevalence is not considered here, the risk state is not an absolute number, but can only be seen as a relative risk in a population. To obtain a risk estimate, the patient connects the values for covariates $x_1$ and $x_2$ with a straight line, and connects the intersection point of that line with the gray auxiliary axis with the value for covariate $x_3$. Using this line, the patient then reads off the probability estimate from the output axis. As shown in Fig. 4, for a person...
with 20–50 nevi, skin type III and minor skin damage due to sun exposure, the risk assessment is \( p \approx 0.34 \).

4. Discussion

Although programmable calculators and personal digital assistants (PDAs) have begun to allow the storage of models in digital form, nomographic representations of logistic regression models are still popular as decision support tools in many areas of clinical medicine [3–13].

The new format presented in this paper was derived to be used by physicians and patients alike. We believe that one of the advantages of this new format is that it is easier to evaluate for people who did not have specific instructions in its use. The evaluation process, the drawing of straight lines, could be explained in a text accompanying the nomogram.

When used as a decision-support tool by physicians, the new format may also be advantageous, because the relationship between covariates is visible at a glance; this is not possible with the traditional nomogram format. This advantage may be especially interesting in application areas where the nomogram user gets to choose the values of covariates (e.g., a physician making treatment decisions that involve a number of factors). Keeping all other covariates constant, the contribution of single factors can then easily be gauged by visual inspection. As an example, consider the nomogram in Fig. 4: one can see that, for nevi count 20–50 and skin type II, choosing all possible values for the skin damage covariate changes the model output between about 0.2 and 0.5.

As with the traditional format, the nomogram presented here is an accurate representation of the logistic regression model from which it was derived (up to the limits of graphical precision). No information is lost when converting the model to graphical format. The major drawback of nomograms, that they do not provide as much significant digits in the result as electronic devices, is not really a disadvantage in application areas where approximations or even dichotomous yes/no answers are sufficient.

A disadvantage of our proposed format is that it is only applicable to models with a small number of covariates. While the construction process given in Section 3 can be generalized to an arbitrary number of covariates, the two-dimensional layout means that the nomogram may become more and more cluttered with each additional covariate axis. Certainly, being able to assess the contribution of individual covariates would not be possible as easily as for the case of three covariates.

As for patient self-assessment, easy-to-use nomograms may be an alternative to questionnaires. There are some studies investigating logistic regression analysis of questionnaire results [24–27]. In these studies, logistic regression models are used to identify significant covariates, and sometimes also to determine whether patients are able to accurately assess their own health state [27]. There is, however, no study employing the process presented in this paper: to use logistic regression analysis for dimensionality reduction (keeping only significant covariates), converting the resulting model to a nomogram, and using this nomogram as a patient self-assessment aid.

The results of our own study on patient self-assessment show that there is good agreement between patient and physician assessment [14]. It turns out, however, that the diagnostic task of discriminating between high-risk and low-risk patients is hard even for physicians. This can be seen from the fact that the area under the ROC curve is not large, for both patient and physician models. So although patients are able to assess their risk state about as well as physicians, models derived from this assessment are not accurate enough to be used as a mass-screening tools. This, however, is not a drawback of the nomogram construction process itself. Future work will involve the study of patient and physician acceptance of the new proposed nomogram format for models with higher discriminatory power.

5. Conclusion

In this paper, we developed a novel method of nomogram construction based on a logistic regression model. The process of converting the logistic regression model to its nomographic representation is presented in detail using, as an example, data from a patient self-assessment study.

The construction process presented this paper is sufficiently general to be applied to any three-covariate logistic regression model; extensions to larger models are possible. We discuss advantages and disadvantages of the nomogram developed here compared with the format that is currently most widely used in medical applications. Mathematica code for the automatic generation of nomograms from logistic regression models is available from the authors.

References


