Paper 264-2013

Build Prognostic Nomograms for Risk Assessment Using SAS®

Dongsheng Yang, Cleveland Clinic Foundation, Cleveland, OH

ABSTRACT

Nomograms from multivariable logistic models or Cox proportional-hazards regression are a popular visual plot to display the predicted probabilities of an event for decision support. In this paper, we show how to build a prognostic nomograms after fitting a multivariable model, including how to assign points for each predictor under different situations such as main effect, interaction, piecewise linear effects. Furthermore, we also show how to use a power tool, graphic template language to construct a nogogram Finally, a SAS macro is developed to generate a nomogram.

INTRODUCTION

After fitting a regression model, it is important to interpret the estimated parameters (i.e., beta coefficients). It is easy to interpret coefficients of a simple regression but not for a multivariable regression model with interactions or non-linear items. For example, for a simple logistic regression (i.e., a single predictor), an odd ratio and a plot of predicted probability vs. predictor can be used to describe the relationship. However, when we describe a relationship between a predictor and response in a multivariable model, we need to hold other predictors to be constant (e.g., mean). Further, it is difficult to directly interpret the association between a predictor and the probability of response when considering other predictors in the model, especially for interaction, transformed predictor, and piecewise linear and non-linear predictors.

Frank Harrell (2000) pointed out that a good way is to construct a nomogram to display the associations between predictors and response in a multivariable model. In details, he said that " a nomogram not only sheds light on how the effect of one predictor on the probability of response depends on the levels of other factors, but it allows to quickly estimate the probability of response for individual subjects" [1]. In other words, it directly obtains the estimated probability of the event when predictor's values are provided, especially useful for users who do not have statistics background. So, nomograms are more and more popular visual plots in the predictive modeling.

In this paper, we show how to build a prognostic nomogram for a multivariable logistic model, including how to assign points for each predictor under different situations such as main effect, interaction, and piecewise linear effect. Furthermore, we also show how to use the power tool, PROC SGPLOT / graphic template language, to construct a nomogram. Finally, a SAS macro for generating nomogram is provided.

NOMOGRAM

In this section, we illustrate each key step of building a nomogram using results from a multivariable logistic model.

1. Scoring each predictor

In general, a point system of a nomogram is used to assign each predictor with a point ranges from 0 to 100 in a graphic interface. Based on the estimated regression coefficients, we rank the estimated effects disregarding statistical significance as well as direction (absolute beta values). The most important thing is to determine which predictor has the biggest impact in the model, then sequentially to assign other predictors based on their proportions to the point assigned to the biggest impact predictor. It means that each predictor is influenced by the presence of other predictors. We use estimated beta coefficients from a main effect logistic model as an example to illustrate how it works. **Table 1** shows the factor gender (male vs. female), age (10 to 90 years), and categorized blood pressure group (low, normal and high), and the corresponding estimated beta coefficients.

1) **Age (continuous variable)**: since an estimated beta coefficient of a continuous predictor depends on its unit, we choose the minimum and maximum values that will be displayed on the nomogram plot to obtain the absolute maximum beta coefficient value. For example, after we choose10 and 90 as the range of age, the calculated absolute maximum beta value = 0.033*(90-10) = 2.64. As compared with gender (0.48) and BP group (0.18), age has the biggest impact on the probability of event, so we assign 100 points to it. Specifically, we assign age of 90 to 100 points, age of 10 to 0 point, and assign points to other age values (i.e., age = 20, 40, 60 etc) based on the linear interpolation.

2) **Gender (categorical variable with two levels)**: gender is the second biggest impact predictor. The total point of the factor gender is assigned based on its proportion to the total point of the age=90 (point = 100). The formula is the ratio of:

   Total Point of male = 100* (absolute maximum beta male / beta age=90)
   = 100*(0.4769/2.64)
   = 18 points
   So a male patient is assigned 18 points, whereas a female is assigned 0 point.

3) **Categorized blood pressure with three levels:** It is easier to understand the point assignment, if we consider the BP group as 2 dummy variables. The points of the first dummy variable (BP group = 1 vs. 2) and the second dummy variable (BP group = 3 vs. 2) are assigned with 7 and 1 points based on their proportions to the point assigned to age, respectively.

Table 1. Beta coefficients and assigned points

Predictor	Level	Beta	Values of variable	Absolute maximum beta value	Rank	Assigned points
Age	Unit=1	0.0330	10 to 90 by 10	0.033*(90-10) = 2.64	1	100 assigned to age = 90; 0 assigned to age = 10; 13, 25, 38, 50, 63, 75, 88 are assigned to age 20 to 80, respectively.
Gender	Male	0.4769	0, 1	0.4769	2	100*(0.4769/2.64) =18 assigned to male; 0 assigned to female
BP Group			1, 2, 3			
	1	0.1818		0.1818	3	100*(0.1818/2.64) = 7
	3	0.0136		0.0136	4	100*(0.0136/2.64) = 1
	2					0

* BP Group (blood pressure): 1 = low, 2 = normal, 3 = high.

2. Linear predictor method to assign points

However, sometimes models include interactions, piecewise linear or non-linear items, where directly assigning points is difficult. For example, using **Table 1** predictors we fit an interaction model:

$$\text{Logit}(P_{Y=1}) = -2.4117 + 0.0455*\text{age} + 1.7797*(\text{gender = male}) + 0.1852*(\text{BP group = 1}) + 0.015*(\text{BP group = 3}) - 0.0261*(\text{age} \times \text{gender = male})$$

Obviously, it is not easy to determine the biggest impact factor and assign points when the model is not a main effect model. Based on the estimated coefficients, the linear predictor (LP) is the best way to calculate points. The calculated results are showed in **Table 2** and **Table 3.**

Below are the LP calculations based on the estimated beta coefficients from the above multivariable logistic regression model:

$LP_{female\ and\ age} = 0.0455 * Age + 1.7797 * 0_{female} - 0.0261 * 0_{female}$
$LP_{male\ and\ age} = 0.0455 * Age + 1.7797 * 1_{male} - 0.0261 * 1_{male}$
$LP_{BP = Low} = 0.1852 * 1_{BP = Low}$
$LP_{BP = High} = 0.015 * 1_{BP = High}$

In calculated LP values, the age×gender is the biggest impact term (e.g., biggest LP value) in the model, and the BP group is the second biggest impact factor. So we assign 100 point to the biggest LP (age= 90 and gender = female: LP = 4.095) and 0 to the smallest LP (age = 10 and gender = female: LP = 0.455).

After that, all others (LP) are assigned using linear interpolation method, based on their proportions to the smallest and biggest LPs. Here is the formula to calculate points using linear interpolation (**Table 2 and 3)**.

Table 2. Linear Predictor (LP) and Assigned Points for Age and Gender

Age	Gender	linear predictor (LP)	Point
90	female	4.095	100
80	female	3.64	87.5
70	female	3.185	75
60	female	2.73	62.5
50	female	2.275	50
40	female	1.82	37.5
30	female	1.365	25
20	female	0.91	12.5
10	female	0.455	0
90	male	3.526	84.4
80	male	3.332	79.0
70	male	3.138	73.7
60	male	2.944	68.4
50	male	2.75	63.0
40	male	2.556	57.7
30	male	2.362	52.4
20	male	2.168	47.1
10	male	1.974	41.7
Point = $100*(LP_i$ – smallest LP) / (biggest LP – smallest LP).			
Point $_{age = 10 and gender = male}$ = 100*(1.974 – 0.455) / (4.095 – 0.0455) = 41.7			

Table 3. Linear Predictor (LP) and Assigned Points for Blood Pressure

BP group	linear predictor (LP)	Points
1 = low	0.1852	5.09
2 = Normal	0	0
3 = high	0.015	0.41
Point $_{BP = low}$ = 100*0.1852 / (4.095 – 0.0455) = 5.09		
Point $_{BP = high}$ = 100*0.015 / (4.095 – 0.0455) = 0.41		

3. Get total points and linear project to the probability scale (0 - 1)

To get total points and linear project to the probability scale, **Table 4**:

1) Points per unit of linear predictor: 100/(max LP – min LP)
2) Linear predictor units per point: 1/ Points per unit of linear predictor
3) $LP_{\text{for total point of 0}}$: $\beta_0 + X\beta$, hold each factor at reference level
4) $LP_{\text{for total point > 0}}$ = ln [Risk of Y = 1 / (1- Risk of Y = 1)]
5) Total Point = Points per unit of linear predictor * (LP_i – $LP_{\text{for total point of 0}}$)

Example:

Points per unit of linear predictor: 100 / (4.095 – 0.0455) = 27.5
Linear predictor units per point: 1/27.5 = 0.0364
$LP_{\text{for total point of 0}}$: -2.4117 +0.045*10+0+0-0 = -1.9567, where reference levels are age = 10, gender = female, and BP group = normal.

For risk = 0.3:

$LP_{\text{for total point > 0}}$ = ln [0.3 / (1-0.3)] = -0.85
Total Point = 27.5 * (-0.85 + 1.9567) = 30.5

Table 4. Total Points and Linear Project to the Probability Scale (0-1)

Risk of Y = 1	LP	Total points	Label
0.124	-1.9567	0	Min
0.2	-1.39	15.7	
0.3	-0.85	30.5	
0.4	-0.41	42.6	
0.5	0	53.8	
0.6	0.405	64.9	
0.7	0.847	77.0	
0.8	1.386	91.8	
0.865	1.696	105	Max

4. Construct a nomogram plot using PROC SGPLOT or graph template language

After calculated points of each factor, total points and their risk probabilities of event, we create a SAS data set that includes all elements needed to generate a nomogram plot. Here is a part of the SAS data set (**Table 5**)

We use the series plot and the highlow plot in the PROC SGPLOT to create a nomogram plot. We set up low = y-0.1 and high = y for locations of ticks and labels. We can also control tick location by setting up low = y and high = y + 0.1 to change the direction of levels of a categorical variable in a plot (e.g., **Fig 1**. BP group).

Below are the SAS codes:

```
proc format;
        value yfmt
                6 = 'Points'
                5 = 'Age at Gender = Female'
                4 = 'Age at Gender = Male'
                3 = 'Blood Pressure'
                2 = 'Total Points'
                1 = 'Risk of Event';
run;

proc sgplot data = fig_data noautolegend ;
        series x = x y = y/group = y_label
                                lineattrs = (pattern = 1 thickness = 2);
        highlow x=x high = high low=low /       lowlabel=x_label
                                                highlabel=x_label2;
        yaxis display= ( nolabel noline noticks ) tickvalueformat =yfmt. ;
        xaxis display= none  ;
run;
```

Table 5. SAS data set for a nomogram

y	y_label	x_label	x	low	high	x_label2
1	Risk of Event	0.2	20	0.9	1	
……						
1	Risk of Event	0.8	80	0.9	1	
2	Total Points	0	12.4	1.9	2	
…..						
2	Total Points	105	86.5	1.9	2	
3	Blood Pressure		0	3	3.1	Normal
3	Blood Pressure	High	0.41	2.9	3	
3	Blood Pressure	Low	5	2.9	3	
4	Age at Gender = Male	10	42	3.9	4	
…….						
4	Age at Gender = Male	90	84	3.9	4	
5	Age at Gender = Female	10	0	4.9	5	
……						
5	Age at Gender = Female	90	100	4.9	5	
6	Points	0	0	5.9	6	
6	Points		5	5.9	6	
6	Points	10	10	5.9	6	
6	Points		15	5.9	6	
6	Points	20	20	5.9	6	
……						

Below is the generated nomogram plot (**Fig 1**):

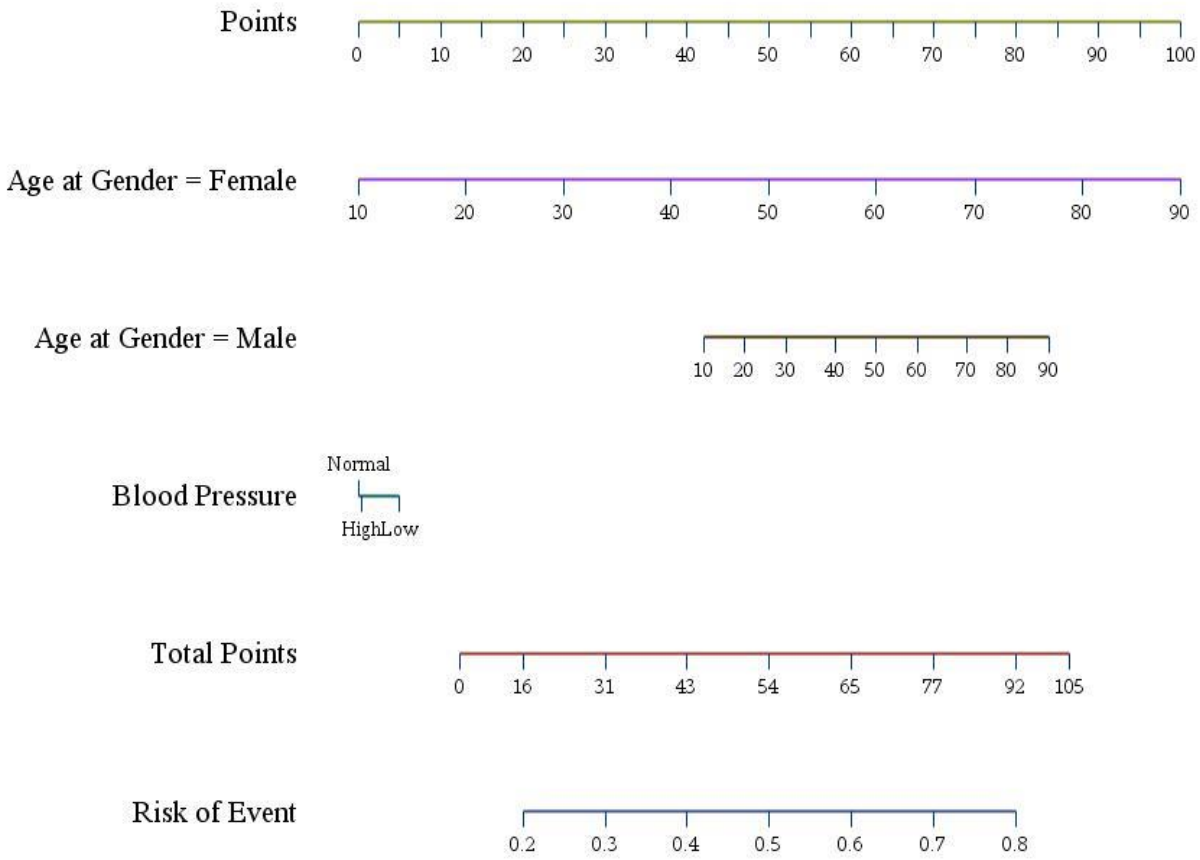

Fig 1. Nomogram plot

INTERPRETATION

To interpret a nomogram is very easy and straightforward. First, we add points from point axis for each predictor to obtain a total point. And then project the total point to the axis of risk of event. For example, given a patient with age = 50, female, and Blood pressure = low, the total point is 55 read from the above nomogram, and the corresponding probability of having the event is about 0.5.

CONCLUSION

Nomogram provides a user-friendly graphic interface to interpret effect sizes of predictors in a multivariable model, and more important it provides predicted probability of an event. Furthermore, it is very easy to construct a nomogram using PROC SGPLOT or graphic template language.

ACKNOWLEDGEMENTS

I would like to acknowledge my colleague, Chenghong Yu, a senior statistical programmer. He helped me in learning how to use the linear predictive method to assign points.

REFERENCES

1. F. Harrell. Regression Modeling Strategies. Springer, New York, NY, 2001.

2. Frank E Harrell Jr (2013). rms: Regression Modeling Strategies. R package version 3.6-3. http://CRAN.R-project.org/package=rms.

3. Lasonos A, Schrag D et al: How To Build and Interpret a Nomogram for Cancer Prognosis. Journal of Clinical Oncology, volume 26 , 2008.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Dongsheng Yang Enterprise: Cleveland Clinic
City, State ZIP: 44195
Work Phone: (216) – 636-5418
E-mail: yangd@ccf.org

Web: http://www.lerner.ccf.org/qhs/