The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2015)
15, Number 2, pp. 537–546

A general-purpose nomogram generator for predictive logistic regression models

Alexander Zlotnik
Technical University of Madrid
Department of Electronic Engineering
Madrid, Spain
and Admissions, Clinical Documentation and
Clinical Information Systems Department
Ramón y Cajal University Hospital
Madrid, Spain
azlotnik@die.upm.es

Víctor Abraira
Clinical Biostatistics Unit
Ramón y Cajal University Hospital
IRYCIS
Madrid, Spain
and CIBERESP
Madrid, Spain

Abstract. Multivariate logistic regression is a statistical method commonly used in several fields to build predictive models. A nomogram is a tool that provides graphical depictions of all variables in the model and enables the user to easily compute output probabilities. Our objective was to build a flexible and easy-to-use nomogram generator in Stata. The script works after arbitrary `logit` or `logistic` commands.

Keywords: st0391, nomolog, graphics, logistic, logit, xtline, nomogram, Kattan nomogram, logistic nomogram, clinical nomogram

1 Introduction

Multivariate binary logistic regression is commonly used in several fields to build predictive models. These models help to fulfill one goal of a regression analysis: to predict the probability of the outcome by using a set of independent variables. Odds-ratio tables are the most common way to present predictive models in several disciplines, but they do not allow the direct calculation of output probabilities. Full logistic regression formulas are sometimes given in article appendixes, but as the number of variables in a model rises, these quickly become burdensome to apply. As a solution, Kattan et al. (1998) presented a graphical calculator for output probability score calculation.

Graphical calculators were given the generic name of nomograms in the late 19th century (Evesham 1986) and were widely used for a variety of engineering problems until the 1970s (Khovanskii 1972). Although some authors argue that Kattan nomograms do not fully comply with the definition of the term (Grimes 2008), from a practical point of view, they can be considered as such. Their precision, similar to that of a logarithmic ruler (Khovanskii 1972), is generally sufficient for biomedical problems.

In addition to enabling the user to obtain predicted values manually, nomograms provide excellent graphical depictions of all variables in the model (Harrell 2001) and provide a quick view of the weight of each variable.

Although predictive models are usually easy to build with current software tools, their applicability in medicine is often doubted (Perel et al. 2006). Clinicians reject published prognostic models primarily because of lack of evidence of accuracy, generality, and effectiveness of the model (Wyatt and Altman 1995) as well as user-friendly presentation (Perel et al. 2006). Although nomograms do provide an easy way of presenting the logistic regression models with a large number of variables, they should be used for output probability estimations only if the underlying model exhibits adequate calibration with both external and internal validation. Models with weaker calibration but a reasonable area under a receiver operating characteristic curve obtained on validation datasets may be used only for classification purposes, where a cutoff point is set to make a binary decision.

Although R (Harrell 2015) and, to a lesser extent, SAS (Yang 2013) can produce nomograms, to the best of our knowledge, nomogram generation has not been implemented in Stata. Our objective was to build a general-purpose nomogram generator entirely in Stata, without external software dependence, executable after `logistic` or `logit` commands. We found it important to allow automatic (or imposed) variable and data labeling as well as continuous variable ranges and division sizes.

Although the `marginsplot` command allows posterior probability calculations, as the number of variables increases, so does the complexity of the graph (superposed curved lines are displayed for each variable), which hinders its applicability for large models.

2 Logistic regression nomograms

As a reminder of nomogram structure, we can visualize a logistic regression model with variables age, gender, and number of transfusions as predictors of probability of hematological complications, built using the `logit` command.

```
. use nomolog_ex.dta
. logit outcome i.gender transfusions age
  (output omitted)
```

outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender						
Female	-.3592306	.1835125	-1.96	0.050	-.7189084	.0004472
transfusions	.029573	.0126694	2.33	0.020	.0047414	.0544046
age	.0688636	.0076487	9.00	0.000	.0538725	.0838547
_cons	-5.409792	.4634713	-11.67	0.000	-6.318179	-4.501405

The output probability using the nomogram of figure 1 is calculated as follows: obtain scores for all variable values, add all scores, and obtain the probability of event using the total score (TS) to probability graph. For example, for a 40-year-old male who had 35 transfusions, the scores for the respective variables are the following: Score(Male) $\approx$ 0.5; Score(35 transfusions) $\approx$ 1.5; and Score(40 years old) $\approx$ 4. The TS would be approximately 6, which is equivalent to a probability of event of approximately 0.16–0.17.

```
. nomolog, vli1(age,10,100,10,0)
Note: negative dummy ´Female´ in variable gender. Forced positive.
```

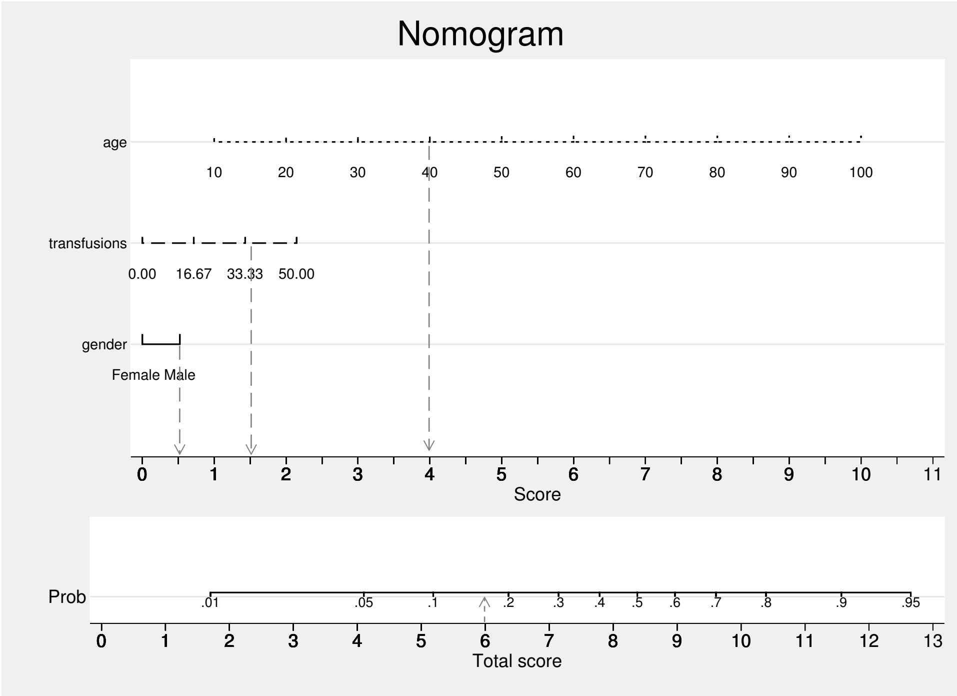

Figure 1. Nomogram example

To facilitate the comprehension of nomogram usage, we created the nomogram with the `nomolog` command and then enhanced the resulting graph using Stata's Graph Editor.

This example illustrates the following uses of logistic nomograms:

1. Logistic nomograms provide output probability calculation using a fast and simple graphical method. A nomogram, once generated, does not require a computer to estimate event probability. This allows its use in environments and situations where computing devices are unavailable or inconvenient. A classic example is that it can enable patients to make informed treatment decisions because, given that the underlying models have been validated on external datasets and exhibit adequate calibration, it allows an easy evaluation of what-if scenarios (Kattan and Marasco 2010). Ideally, models derived from meta-analysis should be used for this purpose.

2. Variable weight is clear at a glance because shorter variable scales indicate lesser relative importance. This eliminates the black-box effect (that is, the lack of understanding of underlying mechanisms of calculation), which is increasingly

frequent in complex models. As a by-product, the nomogram also becomes a useful tool for exploratory or descriptive analysis.

3 Program features[1]

3.1 Imposed variable ranges

The program allows specifying static ranges and division sizes for continuous variables. This can be useful to ease calculations on nomograms. An age variable might have a minimum of 16 and a maximum of 82 in the dataset. However, it is far more natural to set a minimum of 10 and a maximum of 100 with divisions of 10. This can be easily done using the syntax or the graphical dialog box that comes with the program. Because this might produce out-of-sample predictions, a notice is added to each graph specifying the actual variable ranges if these are modified.

3.2 Interactions

Two variable interactions may be used. Both `##` and `#` operators are supported. If an interaction between continuous variables is defined, one of them must be limited to a set of reference points so that the nomogram remains linear.

Given an interaction between two variables (`X1` and `X2`), the scale of the first one is shown (`X1`) and various scales for the other (`X2`), one for each value of the first one. If both variables are categorical, `X1` is the first variable in the command, and `X2` is shown for all values of `X1`. If both are continuous, the same criterion is used, and the user must specify a maximum of five values of `X1` for which `X2` is represented. If one variable is continuous and the other is categorical, the categorical value is considered `X1` and the other `X2`. For example, if the command includes `i.gender##i.treatment`, the `gender` scale would be shown as well as two `treatment` scales, one for `male` and another for `female`. If the command includes `c.age##c.dialisystime`, several values of `age` have to be specified. One scale of `dialisystime` would be displayed for each.

3.3 Forced positive coefficients and score rescaling

Because positive coefficients are convenient in nomograms to avoid subtractions when calculating the TS, the program forces positive coefficients in all noninteraction variables. As a result, the constant of the model is also changed, and label order is inverted within continuous variables with negative coefficients. Forcing positive interaction terms is not supported unless these terms are linearly independent of the rest of the variables. Negative interaction coefficients are displayed in red by default and always preceded by a minus sign.

1. Additional explanations, answers to frequent questions, visual examples, and a modification of the program, which generates nomograms for Cox regression models is available on this webpage: http://www.zlotnik.net/stata/nomograms/.

Calculation details for forced positive coefficients for noninteraction coefficients and score rescaling are presented in appendix A. The rescaling of variable score at 10 points is explained in appendix B.

3.4 Dialog box design

All features are available through a graphical user interface (dialog box), which greatly simplifies the use of the nomogram generator (figure 2).

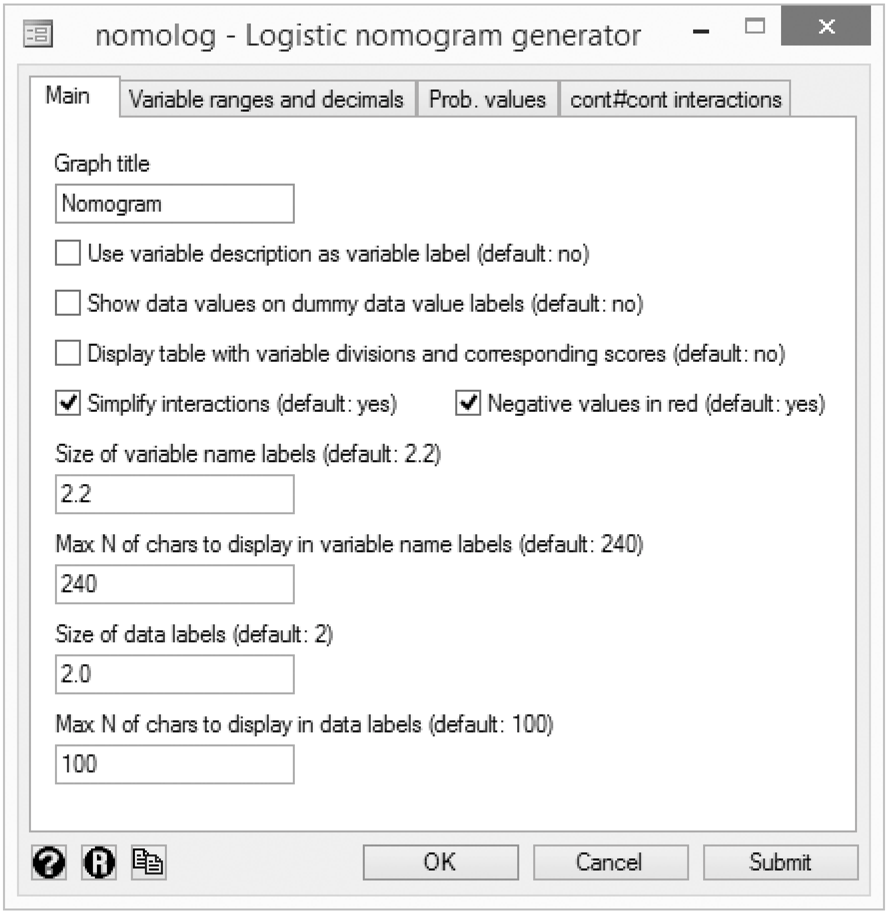

Figure 2. Nomogram dialog box

To generate graphics with publication-ready formatting, one might want to use variable descriptions instead of variable names. In some cases, especially in exploratory analyses, it may be helpful to display dummy category coding (that is, which code corresponds to which category) on the graph. Tabulation of nomogram divisions and corresponding scores can be used in complementary graphs and calculations. All of these options are disabled by default because they are not likely to be required in the most common usage scenarios.

The output of the program is a standard Stata graph, where all elements can be modified manually using Stata's Graph Editor. When many variables with several divisions are used, it becomes cumbersome to modify all of them manually; hence, variable and data label size can be defined for all of them at once.

Certain analyses may require displaying both the main effects and the interaction without a simplification of the nomogram. This may be done by disabling the interaction simplification check box. In interactions of two continuous variables, data points of `X1` for which `X2` is to be represented can be defined in the **cont#cont interactions** tab.

4 Example. Evolution of clinical back pain in routine clinical practice.

Here we illustrate the use of this command with an extract from a predictive model developed on a back pain registry of 17 centers of the Spanish National Health Service (Kovacs et al. 2012).

Clinically relevant improvement of low back pain was used as the dependent variable. Independent variables used in the model were chronicity of back pain (chronic, subacute, acute), baseline severity of referred pain, baseline severity of low back pain, baseline grade of disability in Roland–Morris questionnaire score, previous surgery (yes/no), and neuroreflexotherapy. A predictive model of improvement of low back pain is built using the `logit` command. A nomogram is generated (figure 3).

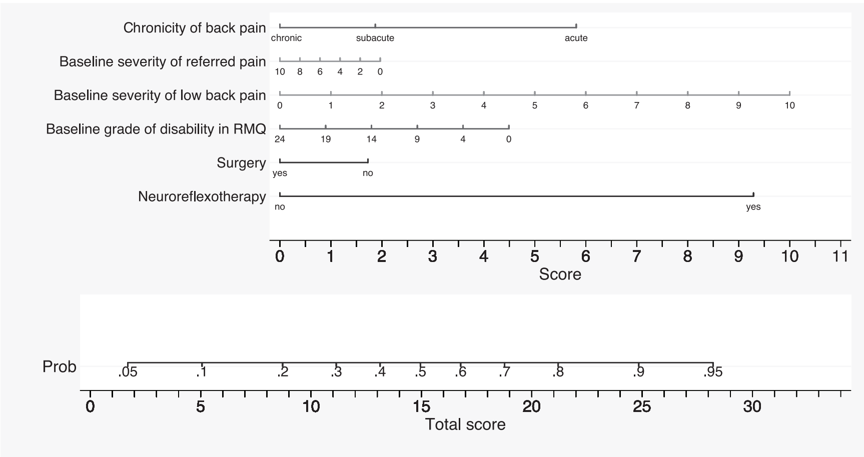

Figure 3. Nomogram for low back pain clinically relevant improvement

5 Conclusion

We believe that we have developed a flexible general-purpose and easy-to-use program to generate logistic regression nomograms. The more relevant features include the following:

- Continuous variable ranges and sizes of divisions may be defined by the user. For example, if the variable age in our dataset has a range of 16 to 81 and the nomogram uses 5 divisions by default for this variable, we might want to define the variable range to be 10 to 100 with divisions of size 10 without changing the coefficient.
- The nomogram is paginated in batches of 10 variables, which improves model readability in settings with many variables.
- Our program produces a vector graph, which is modifiable with Stata's Graph Editor for refinement of graph elements, such as labels and styles. This saves time when preparing graphs for publications.
- The program uses fully commented and orthogonally structured code, completely independent of the regression generator itself, which eases modification and extendability.

6 Limitations

This implementation produces nomograms in the way they were conceived by Kattan et al. (1998) and subsequently applied in most of the literature—without confidence intervals. Displaying these would make the figure cumbersome because the standard error of the TS depends on individual variable values.

Interactions with three or more variables are not supported, because they are relatively infrequent in biomedical research.

Although Stata possesses a rich graphical library, custom graphs are challenging to develop as limits of nonmodifiable functions belonging to the Stata core libraries are reached. In this case, time-series graphs are used with the `xtline` command, which imposes a hard-coded limitation of 70 labels. This is why exceeding labels are not displayed and categorical variables are limited to 10 dummies each. This can be mitigated by applying user-defined variable division numbers. If the `divtable` option is used, scores corresponding to all variable divisions will be displayed, even those exceeding these limits.

7 Future work

A calculator version of the program is also planned to enable nomogram generation directly from user-provided coefficients. Cox regression support will be added in subsequent development, most likely in a separate package.

8 Acknowledgments

We thank all StataCorp technical support personnel whom we had the opportunity to work with, especially statistician Joy Wang. We also thank all the members of the Clinical Biostatistics Unit of the Ramón y Cajal University Hospital who participated in the testing of this program.

9 References

Evesham, H. A. 1986. Origins and development of nomography. *Annals of the History of Computing* 8: 324–333.

Grimes, D. A. 2008. The nomogram epidemic: Resurgence of a medical relic. *Annals of Internal Medicine* 149: 273–275.

Harrell, F. E., Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer.

———. 2015. *rms: Regression Modeling Strategies.* R package version 4.3-0. http://cran.r-project.org/web/packages/rms/index.html.

Kattan, M. W., J. A. Eastham, A. M. F. Stapleton, T. M. Wheeler, and P. T. Scardino. 1998. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *Journal of the National Cancer Institute* 90: 766–771.

Kattan, M. W., and J. Marasco. 2010. What is a real nomogram? *Seminars in Oncology* 37: 23–26.

Khovanskii, G. S. 1972. *Nomography and its Possibilities.* Moscow, Russia: Nauka.

Kovacs, F. M., J. Seco, A. Royuela, J. Corcoll Reixach, and V. Abraira. 2012. Predicting the evolution of low back pain patients in routine clinical practice: Results from a registry within the Spanish National Health Service. *Spine Journal* 12: 1008–1020.

Perel, P., P. Edwards, R. Wentz, and I. Roberts. 2006. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making* 6: 38.

Wyatt, J. C., and D. G. Altman. 1995. Commentary: Prognostic models: Clinically useful or quickly forgotten? *British Medical Journal* 311: 1539.

Yang, D. 2013. Build prognostic nomograms for risk assessment using SAS®. Paper 264-2013. SAS Global Forum 2013. http://support.sas.com/resources/papers/proceedings13/264-2013.pdf.

About the authors

Alexander Zlotnik is a PhD student in the Department of Electronic Engineering at the Technical University of Madrid. He is also an internal consultant in the Admissions, Clinical Documentation, and Clinical Information Systems Department at the Ramón y Cajal University Hospital.

Víctor Abraira is the head of the Clinical Biostatistics Unit at Ramón y Cajal University Hospital, the head of the Research Group at CIBERESP, and the director of priority area in the Instituto Ramón y Cajal de Investigación Sanitaria.

A Forced positive coefficients

The function to transform the TS in probability is the logistic function.

$$p = \frac{1}{1+e^{-(\alpha_0+\mathrm{TS})}} = \frac{1}{1+e^{-(\alpha_0+\alpha_1 x_1+\cdots+\alpha_N x_N)}}$$

Given a categorical variable A with N categories and a regression constant α_0,

$$\mathrm{TP} = \alpha_0 + \mathrm{TS} = \alpha_0 + \alpha_{A1} \times D_1 + \alpha_{A2} \times D_2 + \cdots + \alpha_{AN} \times D_N$$

If $\exists \alpha_{Ai \;\; i=1\ldots N} < 0$, the most negative coefficient $\min(\alpha_{Ai \;\; i=1\ldots N})$ is set as reference.

Then

$$\mathrm{TP} = \beta_0 + \beta_{A1} \times D_1 + \beta_{A2} \times D_2 + \cdots + \beta_{AN} \times D_N$$

where

$$\begin{aligned} \beta_0 &= \alpha_0 - \min(\alpha_{Ai \;\; i=1\ldots N}) \\ \beta_1 &= \alpha_1 - \min(\alpha_{Ai \;\; i=1\ldots N}) \\ &\ldots \\ \beta_N &= \alpha_N - \min(\alpha_{Ai \;\; i=1\ldots N}) \end{aligned}$$

These changes can be linearly combined for different variables adjusting the constant each time.

B Score rescaling

The maximum score for any variable is normalized at 10 to ease calculations.

$$\epsilon_i = \alpha_i \times F$$

where

$$F = \frac{10}{\max(\alpha_{i \;\; i=1\ldots N})} \forall \alpha_i$$

The adjustment must be then also made in the TS term,

$$\mathrm{TS} \times F = \left(\frac{p}{1-p} - \alpha_0\right) \times F$$

Maximum and minimum probabilities are

$$p_{\min} = \frac{1}{1 + e^{-(\alpha_0)}}$$
$$p_{\max} = \frac{1}{1 + e^{-\{\alpha_0 + \max(\mathrm{TS})\}}}$$

The TS has a restriction to improve readability,

$$\mathrm{TS}|_{p\max} > \mathrm{TS}|_{p=0.999} => \mathrm{TS}_{\max} = \mathrm{TS}|_{p=0.999}$$
$$\mathrm{TS}|_{p\max} \leq \mathrm{TS}|_{p=0.999} => \mathrm{TS}_{\max} = \mathrm{TS}|_{p\max}$$