2014 Spanish Stata Users Group meeting

Development of the nomolog program and its evolution

Towards the implementation of a nomogram generator for the Cox regression

Alexander Zlotnik, Telecom.Eng.
Víctor Abraira Santos, PhD

Ramón y Cajal University Hospital

instituto ramón y cajal de investigación sanitaria irycis

NOTICE

- Nomogram generators for logistic and Cox regression models have been updated since this presentation.
- Download links to the latest program versions (nomolog & nomocox), examples, tutorials and methodological notes are available on this webpage:
  http://www.zlotnik.net/stata/nomograms/

Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work

Introduction

- What is a nomogram?

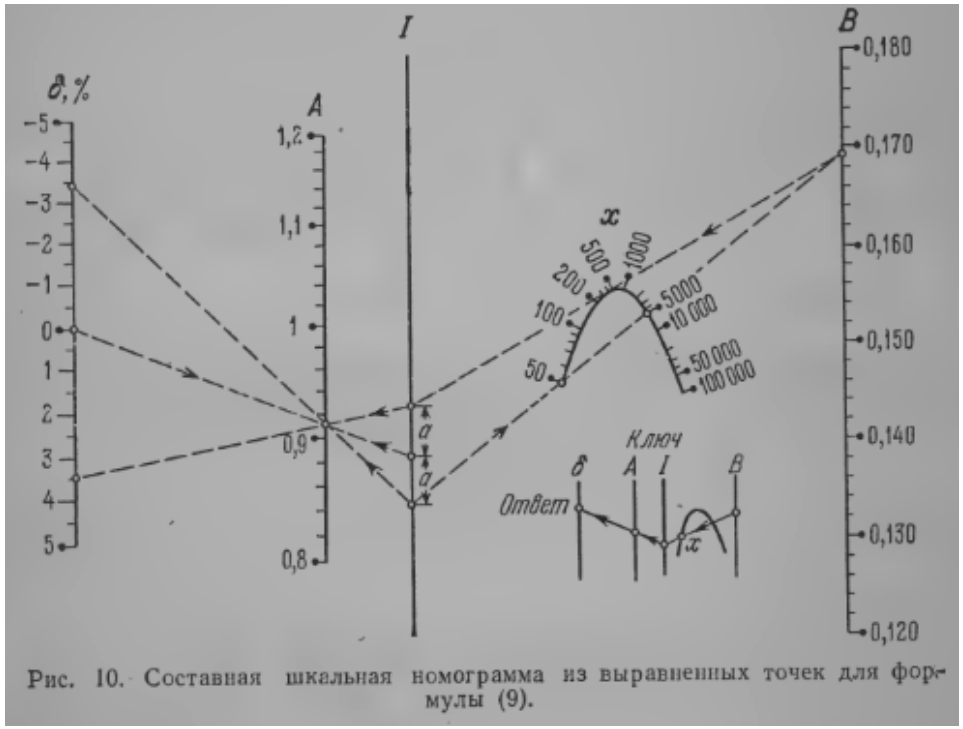
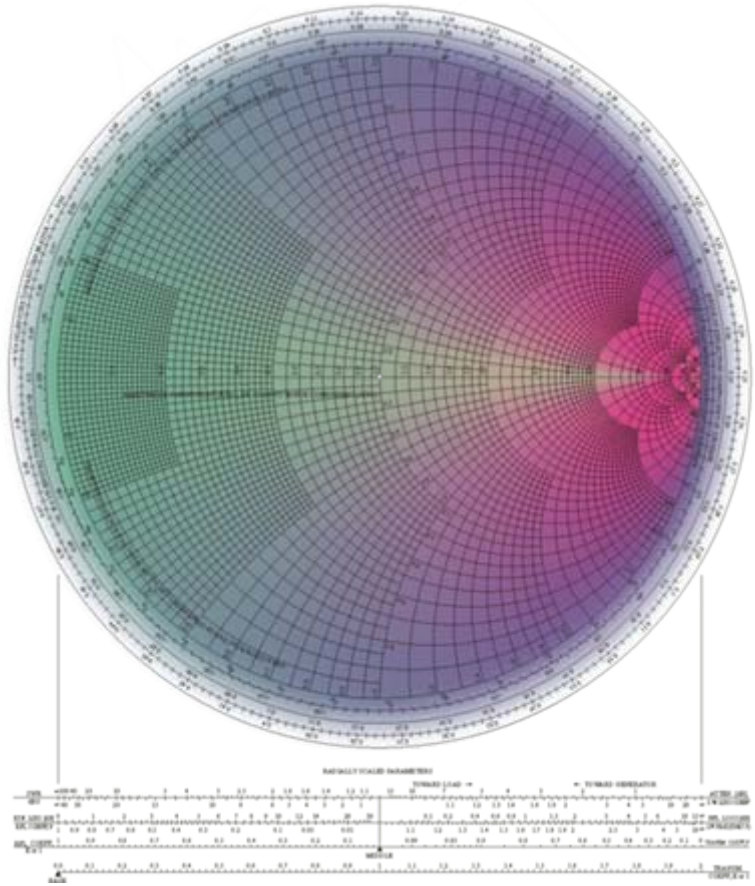


Рис. 10. Составная шкальная номограмма из выравненных точек для формулы (9).

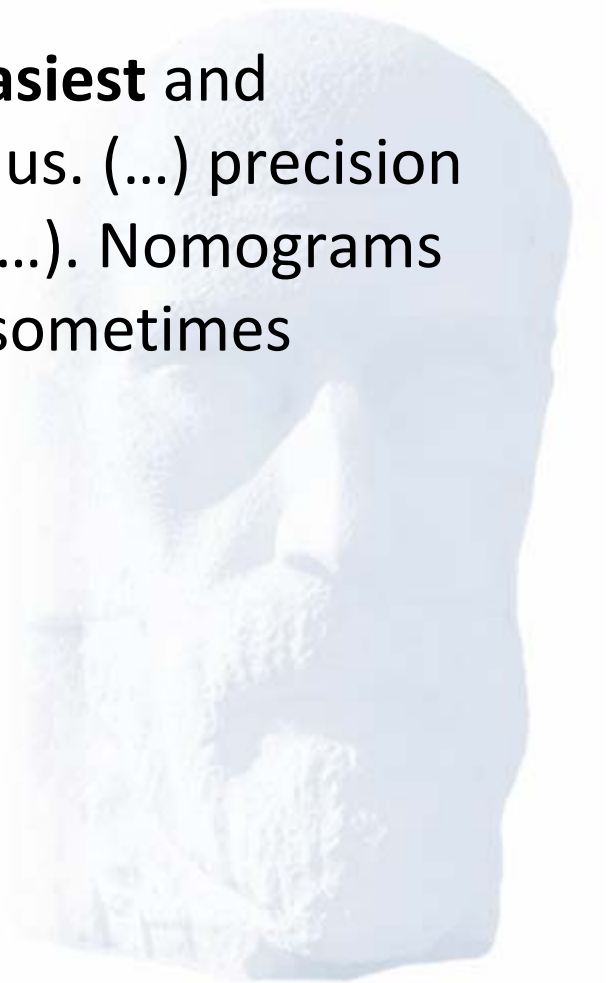
$$\delta = 100\,\frac{\lg x - Ax^B}{\lg x},$$



Introduction

- Nomograms are one of the **simplest, easiest** and **cheapest** methods of mechanical calculus. (…) precision is similar to that of a logarithmic ruler (…). Nomograms can be used for research purposes (…) sometimes leading to new scientific results.

Source: “Nomography and its applications”
G.S. Jovanovsky, Ed. Nauka, 1977



Introduction

- Sometimes complex calculations…

$$\Gamma_1^2 = \Gamma_0^2\Big[1 - \frac{\Gamma_0^4}{\Gamma_1^2} - \frac{\Gamma_0^4}{\Gamma_3^2} - 2\frac{\Gamma_0^4}{\Gamma_1\Gamma_3}\cos(2\theta_{32} - \psi_3 - 2\theta_{12} + \psi_1)\Big](1 + \Gamma_1^2 + 2\Gamma_1\cos\psi_1)(1 - \Gamma_1^2)^{-1} \times$$
$$\times\Big\{1 + \frac{\Gamma_0^4}{\Gamma_1^2} + \frac{\Gamma_0^4}{\Gamma_3^2} + 2\frac{\Gamma_0^4}{\Gamma_1\Gamma_3}\cos(2\theta_{32} - \psi_3 - 2\theta_{12} + \psi_1) + 2\Gamma_0^2\Big[\frac{\cos(2\theta_{12} - \psi_1)}{\Gamma_1} + \frac{\cos(2\theta_{32} - \psi_3)}{\Gamma_3}\Big]\Big\}^{-1}, \quad (70)$$

$$\Gamma_3^2 = \Gamma_0^2\Big[1 - \frac{\Gamma_0^4}{\Gamma_1^2} - \frac{\Gamma_0^4}{\Gamma_3^2} - 2\frac{\Gamma_0^4}{\Gamma_1\Gamma_3}\cos(2\theta_{12} - \psi_1 - 2\theta_{32} + \psi_3)\Big](1 + \Gamma_3^2 + 2\Gamma_3\cos\psi_3)(1 - \Gamma_3^2)^{-1} \times$$
$$\times\Big\{1 + \frac{\Gamma_0^4}{\Gamma_1^2} + \frac{\Gamma_0^4}{\Gamma_3^2} + 2\frac{\Gamma_0^4}{\Gamma_1\Gamma_3}\cos(2\theta_{12} - \psi_1 - 2\theta_{32} + \psi_3) + 2\Gamma_0^2\Big[\frac{\cos(2\theta_{12} - \psi_1)}{\Gamma_1} + \frac{\cos(2\theta_{32} - \psi_3)}{\Gamma_3}\Big]\Big\}^{-1} \quad (71)$$

$$\frac{\partial f_1}{\partial\psi_1} - \frac{\partial f_2}{\partial\psi_1} + \frac{\partial f_3}{\partial\psi_3} - \frac{\partial f_2}{\partial\psi_3} < 0,$$
$$\Big(\frac{\partial f_1}{\partial\psi_1} - \frac{\partial f_2}{\partial\psi_1}\Big)\Big(\frac{\partial f_3}{\partial\psi_3} - \frac{\partial f_2}{\partial\psi_3}\Big) - \frac{\partial f_2}{\partial\psi_1}\frac{\partial f_2}{\partial\psi_3} > 0, \quad (72)$$

где

$$f_1 = f_{01} - \frac{6}{5}F_0\frac{\sin\psi_1}{\frac{1+\Gamma_1^2}{2\Gamma_1} + \cos\psi_1},$$

$$f_2 = f_{02} - \frac{6}{5}F_0\Big[\frac{\sin(2\theta_{12} - \psi_1)}{\Gamma_1} + \frac{\sin(2\theta_{32} - \psi_3)}{\Gamma_3}\Big]\Big[\frac{1}{2\Gamma_0^2} + \frac{\Gamma_0^2}{2}\Big(\frac{1}{\Gamma_1^2} + \frac{1}{\Gamma_3^2}\Big) + \frac{\Gamma_0^2}{\Gamma_1\Gamma_3}\cos(2\theta_{12} - \psi_1 - 2\theta_{32} + \psi_3) + \frac{\cos(2\theta_{12} - \psi_1)}{\Gamma_1} + \frac{\cos(2\theta_{32} - \psi_3)}{\Gamma_3}\Big]^{-1},$$

$$f_3 = f_{03} - \frac{6}{5}F_0\frac{\sin\psi_3}{\frac{1+\Gamma_3^2}{2\Gamma_3} + \cos\psi_3}.$$

Stability conditions

Introduction

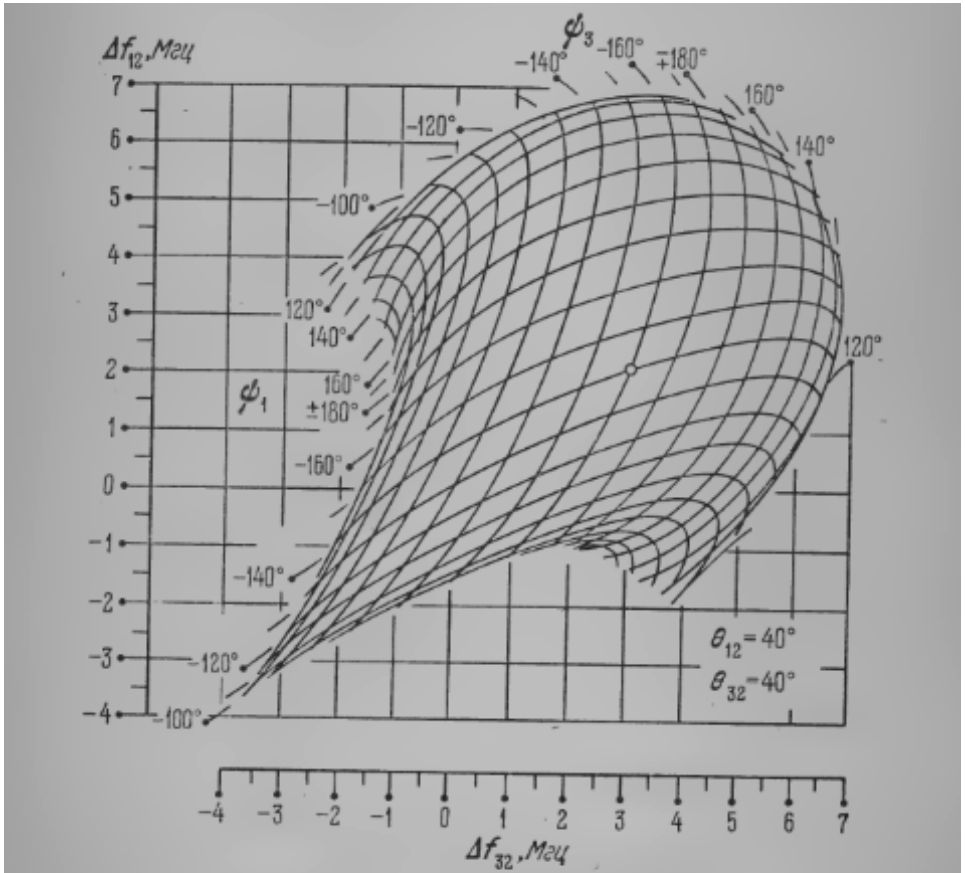


Рис. 56. Сетчатая номограмма для решения системы уравнений (68) — (71) при условиях (72) относительно ψ_1 и ψ_3, построенная при $\Gamma_0 = 0{,}1$, $F_0 = 10$ Мгц, $\theta_{12} = 40°$, $\theta_{32} = 40°$.

… can be greatly simplified with a nomogram

Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work

Logistic regression nomograms

- Logistic regression-based predictive models are used in many fields, clinical research being one of them.

- Problems:
  - Variable importance is not obvious (coefficients may be small, but variable ranges may be large).
  - Calculating an output probability with a set of input variable values can be laborious for these models, **which hinders their adoption.**

Logistic regression nomograms

- Logistic regression nomogram generation
  - Plot all possible scores/points ($\alpha_1 x_i$) for each variable ($X_{1..N}$).
  - Get constant (α_0).
  - Transform into **probability of event** given the formula

$$p = \frac{1}{1 + e^{-(\alpha_0 + TP)}}$$

$$Total\ points = TP = \alpha_1 X_1 + \alpha_2 X_2 + \ldots$$

Logistic regression nomograms

- Example:

logit complications gender transfusions age

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Age	.0652398	.0069921	9.33	0.000	.0515356	.078944
transfusions	.0362445	.0115255	3.14	0.002	.0136549	.0588342
gender	.5388903	.1747807	3.08	0.002	.1963265	.8814542
_cons	-5.783012	.4558551	-12.69	0.000	-6.676472	-4.889553

$$\ln(p/1-p) = Y = -5.783012 + 0.0652398 * Age + 0.0362445 * transfusions + + gender * 0.5388903$$

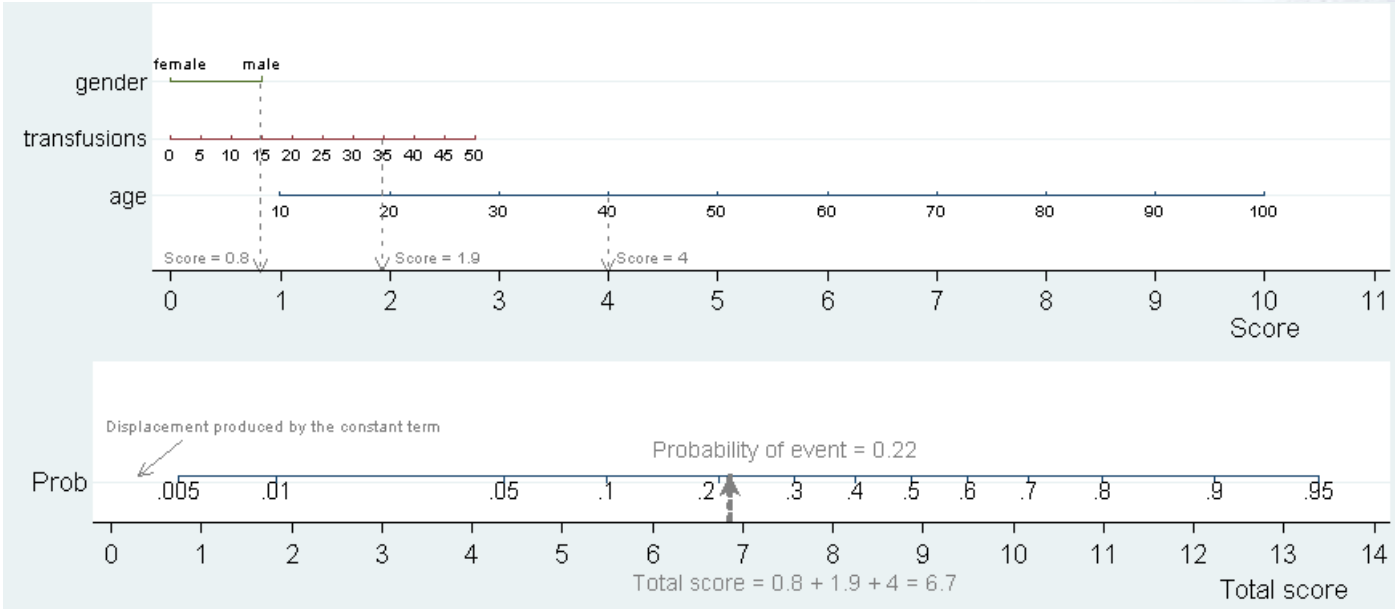
$$p = (e\wedge Y)/(1+e\wedge Y)$$

level = 0 => Female
level = 1 => Male

instituto ramón y cajal de investigación sanitaria irycis

Logistic regression predictive models

- Example:

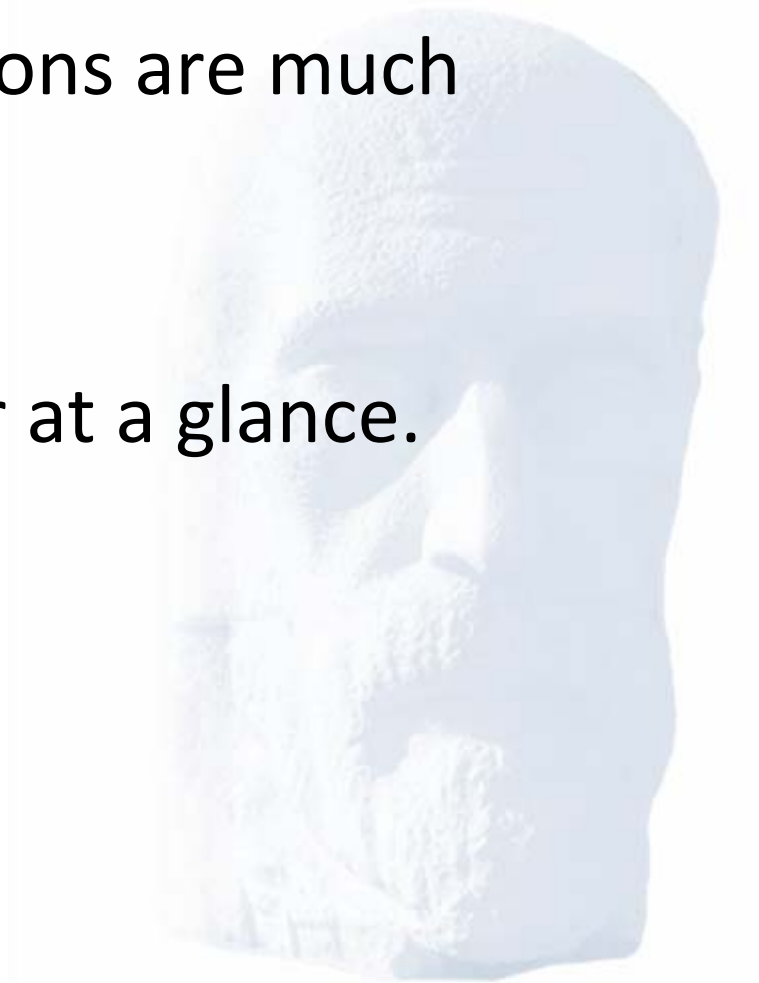


For a **40 year** old **male** who had **35 transfusions**,
Score(Male) ≈ 0.8; Score(35 transfusions) ≈ 1.9; Score(40 years old) ≈ 4.
The total score would be approximately 6.7,
which is equivalent to a probability of event of approximately 0.22.

Logistic regression nomograms

- Output probability calculations are much easier.

- Variable importance is clear at a glance.

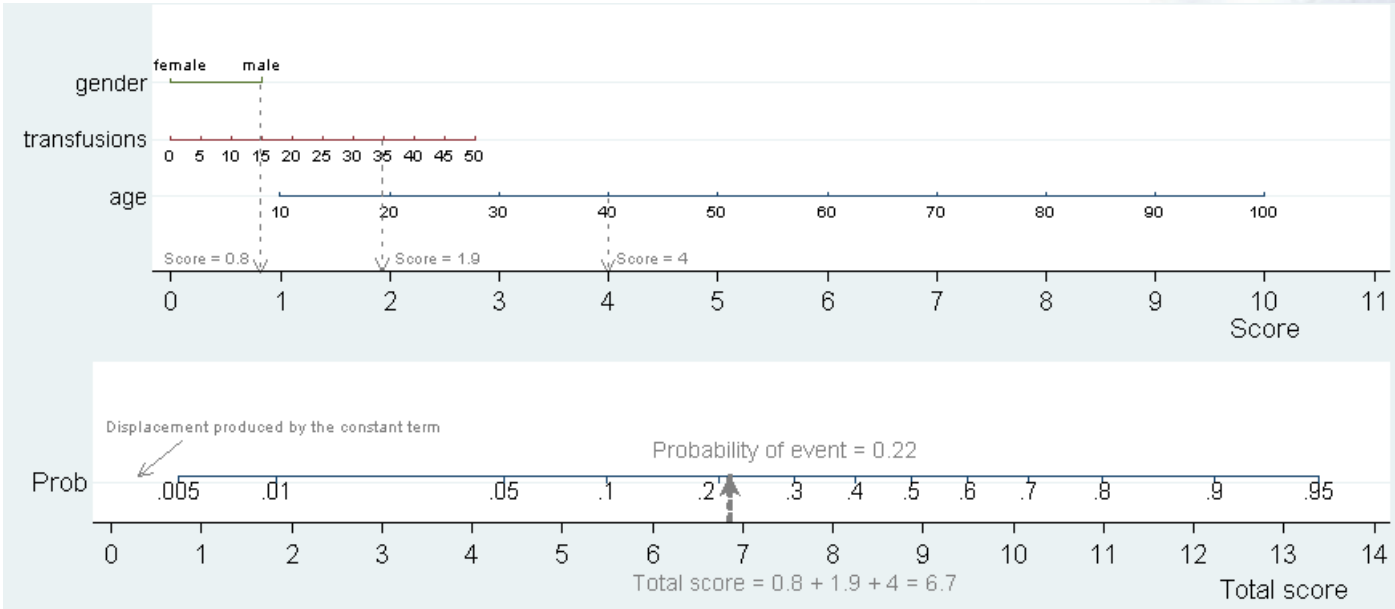


A (gentle) word of warning…

- Nomograms are a representation of a model, not a model validation tool.
- Nomograms don't have confidence intervals. It is *sort of* possible to graph nomograms with CIs, but it makes little sense.
- Nomograms should be used in models with good calibration, if possible with (extensive) external validation.

Score rescaling

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
age	.0652398	.0069921	9.33	0.000	.0515356	.078944
transfusions	.0362445	.0115255	3.14	0.002	.0136549	.0588342
gender	.5388903	.1747807	3.08	0.002	.1963265	.8814542
_cons	-5.783012	.4558551	-12.69	0.000	-6.676472	-4.889553



Score rescaling

- Scores are not equal to coefficient values because we rescale the scores…

$\varepsilon_i = \alpha_i \cdot F$

where

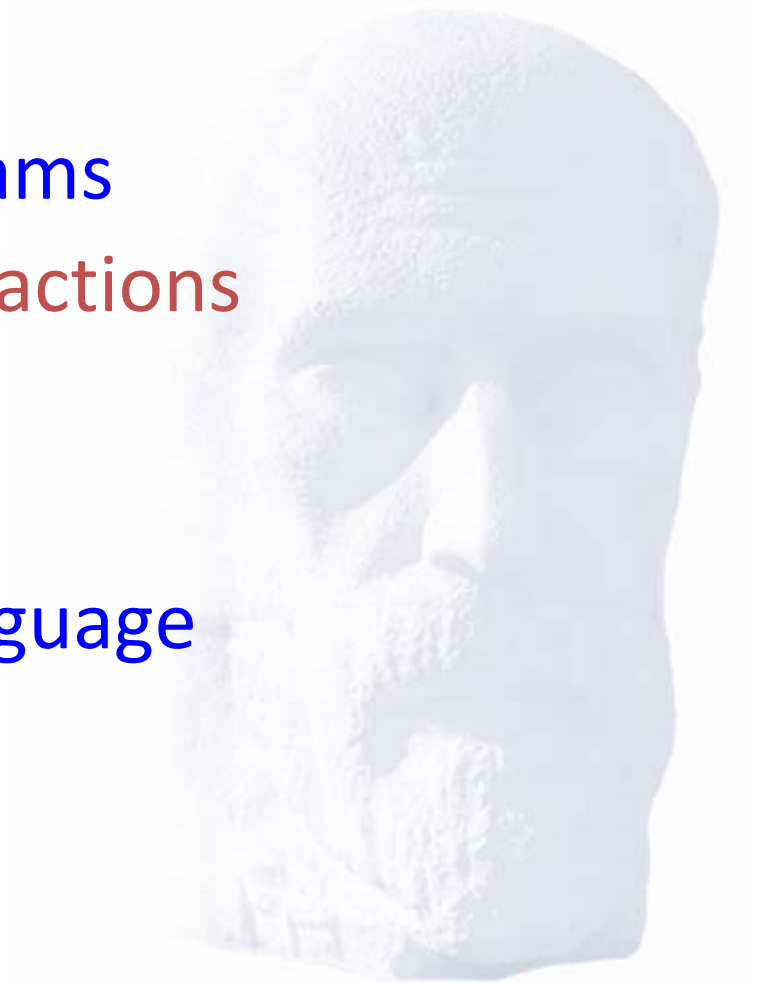
$F = 10 / max(\alpha_{i\ i=1..N}) \ \forall\ \alpha_i$

The adjustment must be then also made in the *Total Score* term

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

Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work



Dummy coefficient re-adjustment

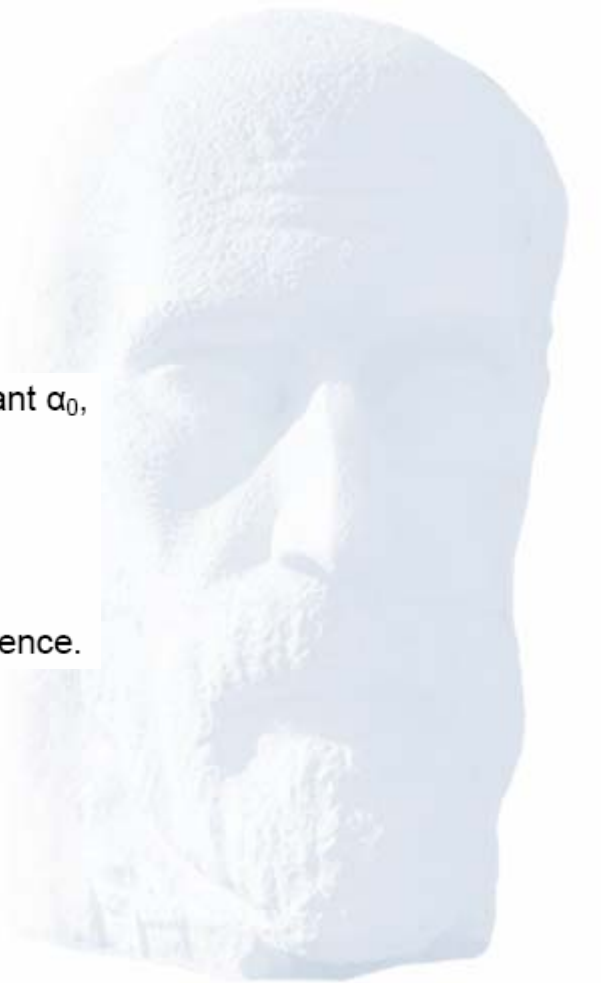
- Coefficients are forced positive

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

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

$TP = \alpha_0 + TS = \alpha_0 + \alpha_{A1} \cdot D_1 + \alpha_{A2} \cdot D_2 + \ldots + \alpha_{AN} \cdot D_N$

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



Dummy coefficient re-adjustment

and then

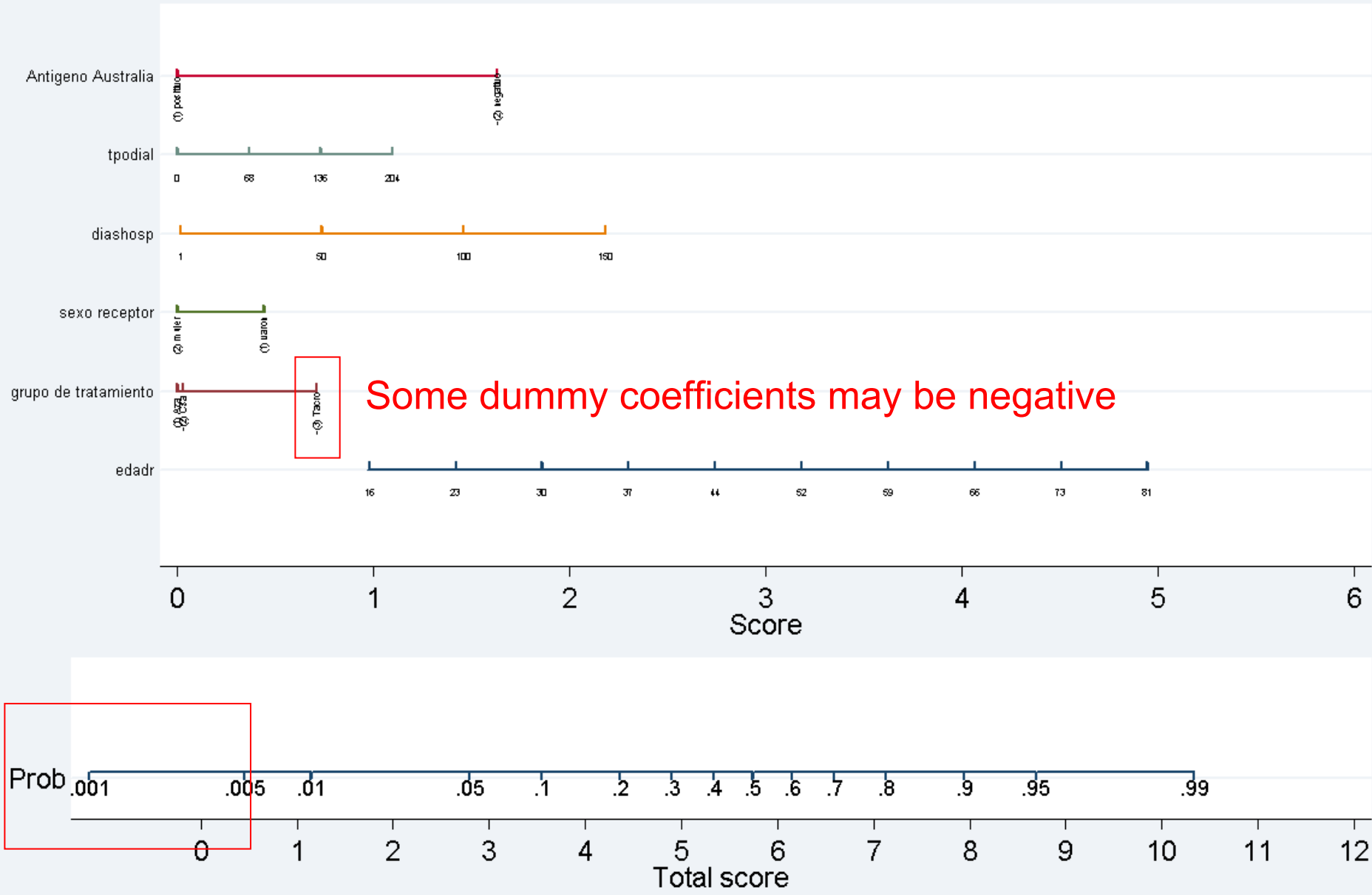
$$z = \beta_0 + \beta_{A1} \cdot D_1 + \beta_{A2} \cdot D_2 + \ldots + \beta_{AN} \cdot D_N$$

where

$$\beta_0 = \alpha_0 - min(\alpha_{Ai\ i=1..N})$$
$$\beta_1 = \alpha_1 - min(\alpha_{Ai\ i=1..N})$$
$$\ldots$$
$$\beta_N = \alpha_N - min(\alpha_{Ai\ i=1..N})$$

instituto ramón y cajal de investigación sanitaria irycis

Nomogram example

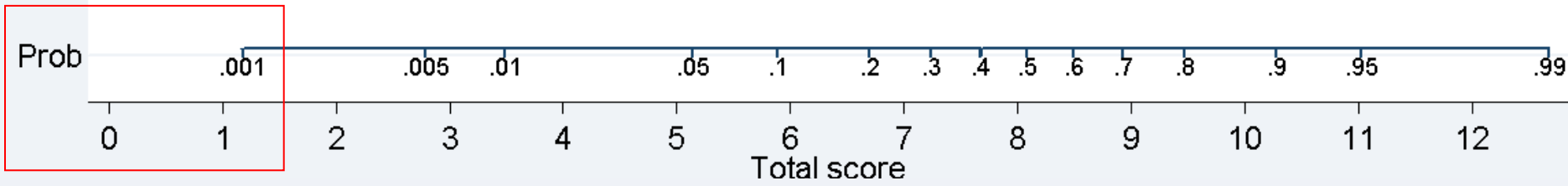


Ramon y Cajal University Hospital

Forced positive coefficients

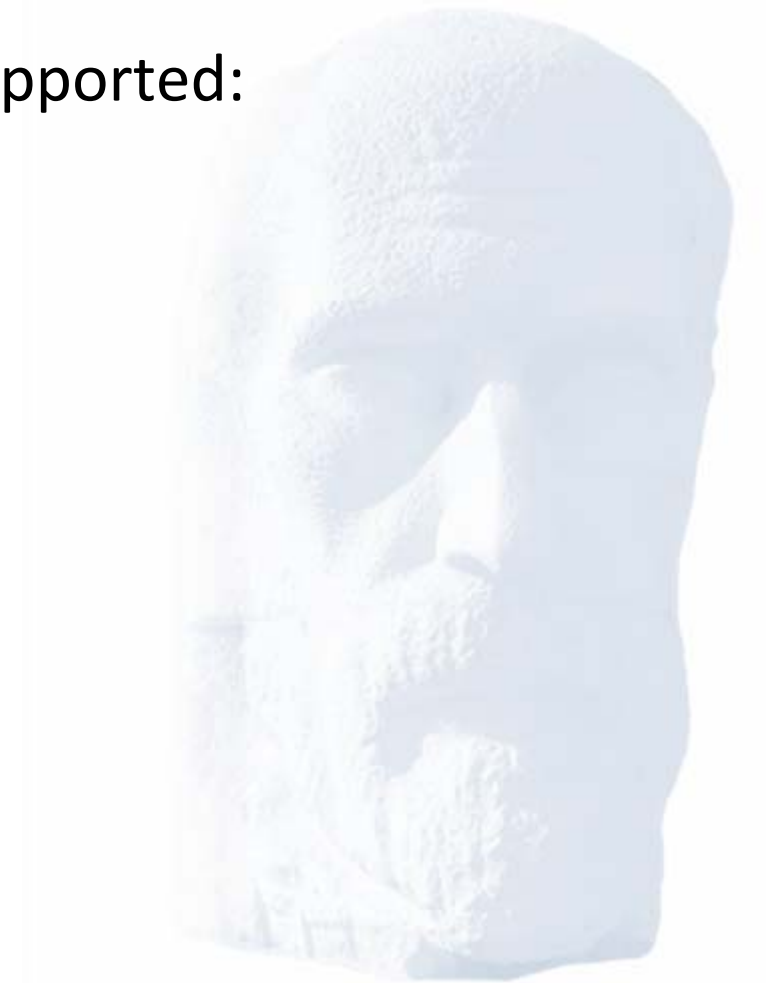
Nomogram example

This causes a displacement in the *Total score to Prob* conversion (due to α_0)



Interactions

- Three types of interactions are supported:
  - Continuous # Categorical
  - Categorical # Categorical
  - Continuous # Continuous



Interactions

In a model $z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$

if we make the transformation $y_1 = A - x_1$,
this will not only change coefficients α_1 and α_3,
but also coefficients α_0 and α_2 since

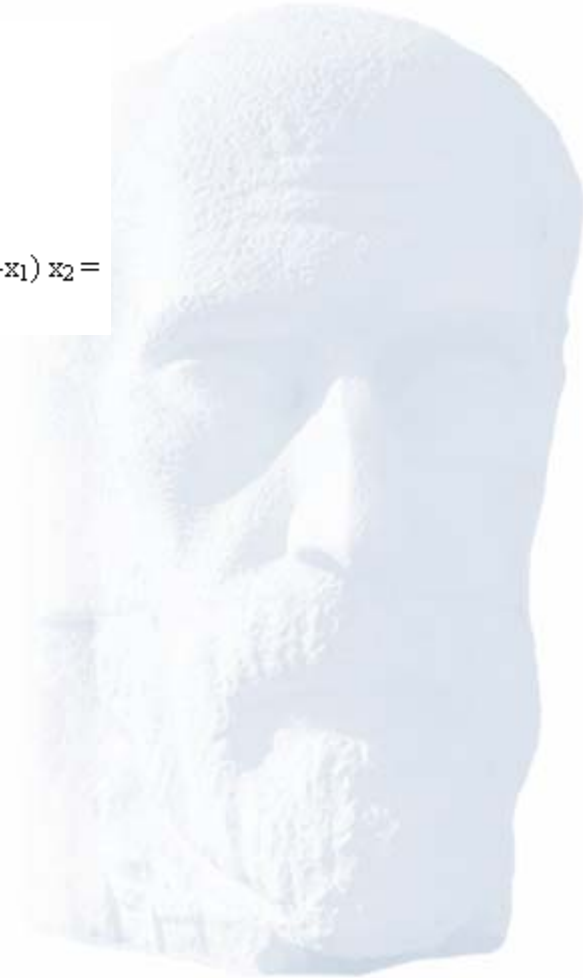
$$z = \beta_0 + \beta_1 y_1 + \beta_2 x_2 + \beta_3 y_1 x_2 = \beta_0 + \beta_1 y_1 + \beta_2 x_2 + \beta_3 y_1 x_2 = \beta_0 + \beta_1 (A - x_1) + \beta_2 x_2 + \beta_3 (A - x_1) x_2 =$$
$$\beta_0 + \beta_1 A - \beta_1 x_1 + \beta_2 x_2 + \beta_3 A x_2 - \beta_3 x_1 x_2 = \beta_0 + \beta_1 A - \beta_1 x_1 + (\beta_2 + \beta_3 A) x_2 - \beta_3 x_1 x_2$$

Therefore

$$\alpha_0 = \beta_0 + \beta_1 A \Rightarrow \alpha_0 - \beta_1 A = \beta_0 = \alpha_0 + \alpha_1 A$$
$$\alpha_1 = -\beta_1$$
$$\alpha_2 = \beta_2 + \beta_3 A \Rightarrow \alpha_2 - \beta_3 A = \beta_2 = \alpha_2 + \alpha_3 A$$
$$\alpha_3 = -\beta_3$$

Positive coefficients are not forced
in interaction terms

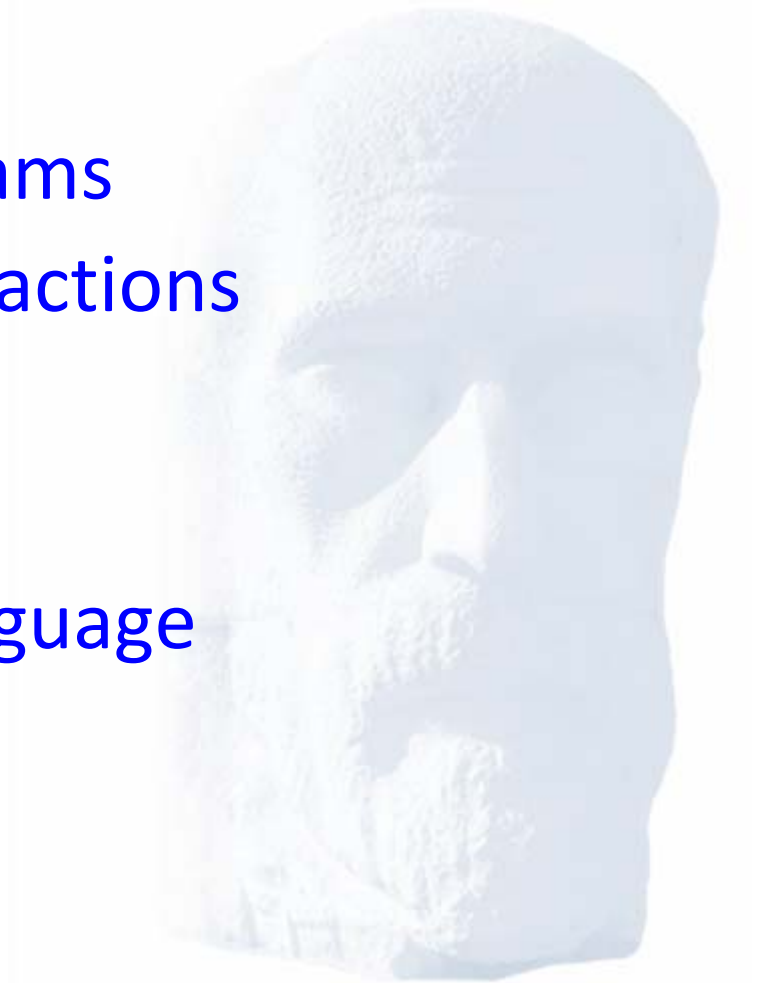
It is left to the user to find reference
terms which produce
positive interaction coefficients



instituto ramón y cajal de investigación sanitaria irycis

Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work



Installation

- Manual

Create **c:\ado\personal** (if it doesn't exist)

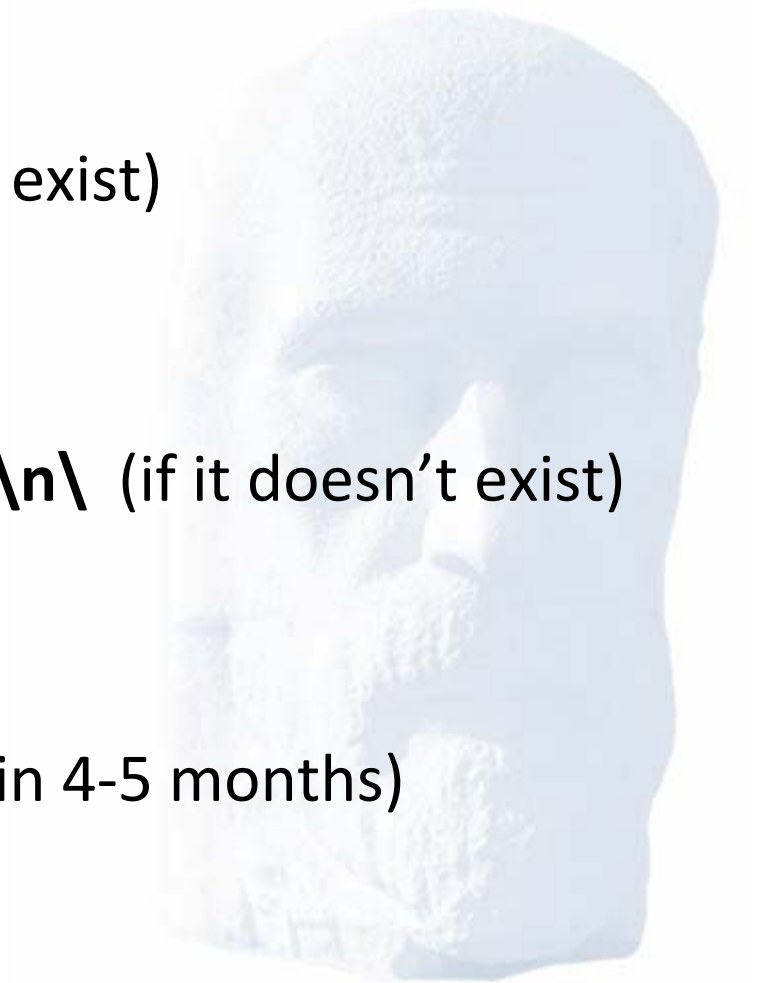
Copy the program files there.

Or, alternatively, create **c:\ado\plus\n** (if it doesn't exist)

Copy the program files there.

- Automatic (will become available in 4-5 months)

ssc install nomolog

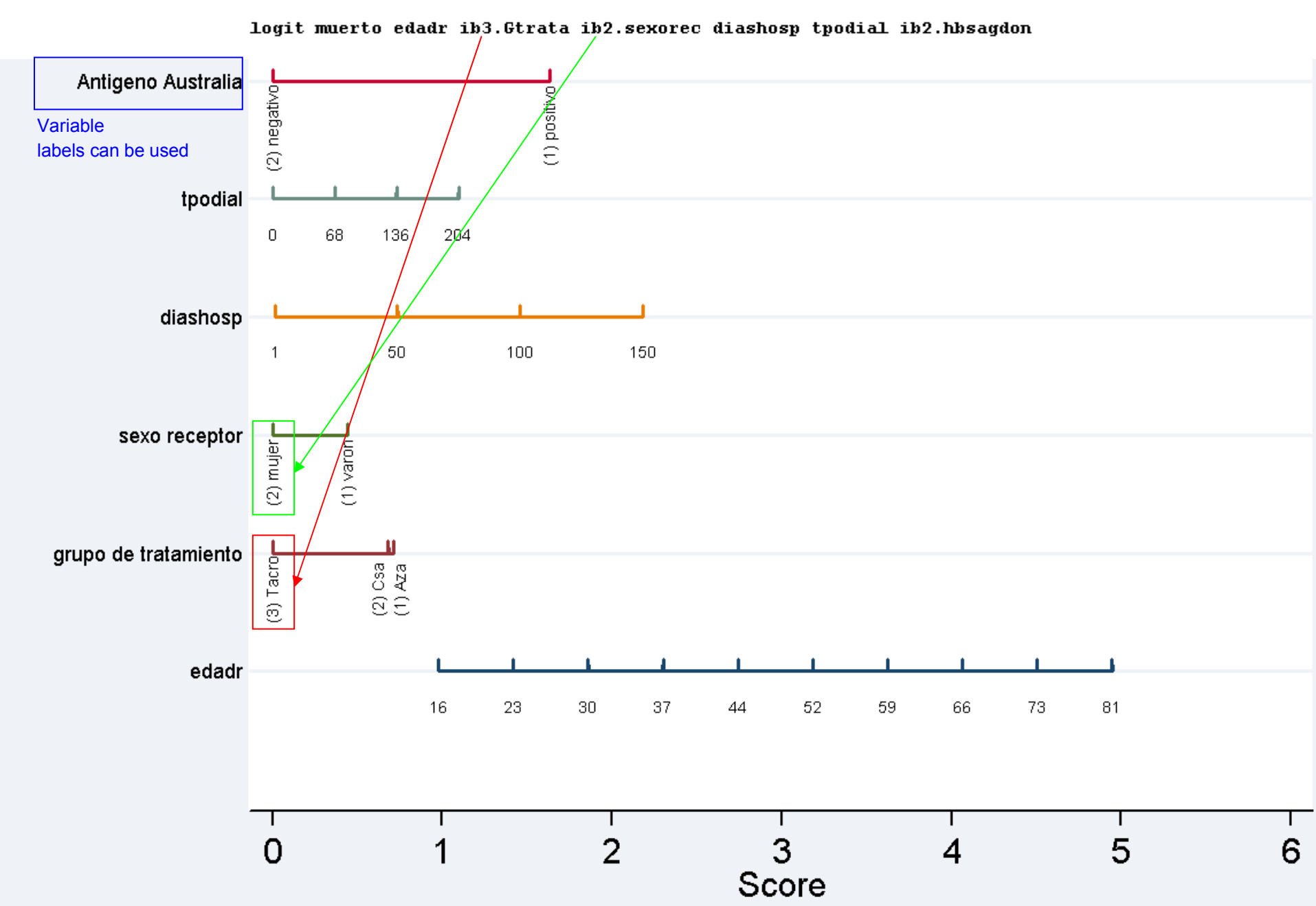


Usage

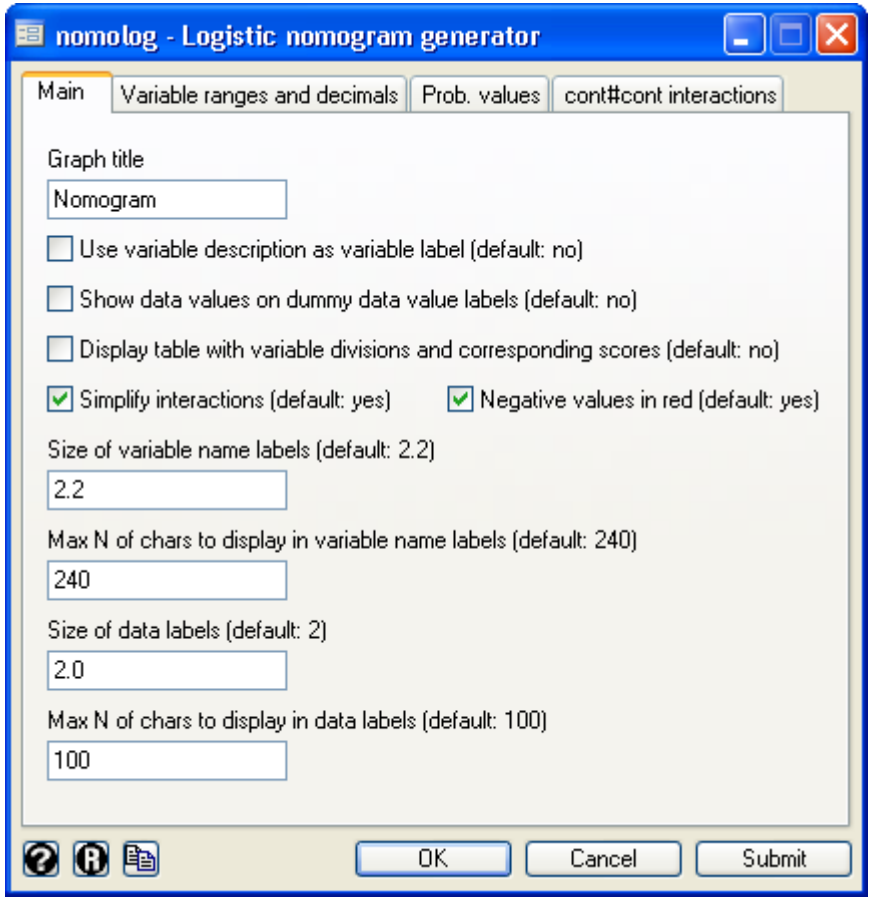
- **logistic … *anything* …** (usual syntax)
- **nomolog*, options***

Or, use the Graphical User Interface

- **db nomolog**



Usage

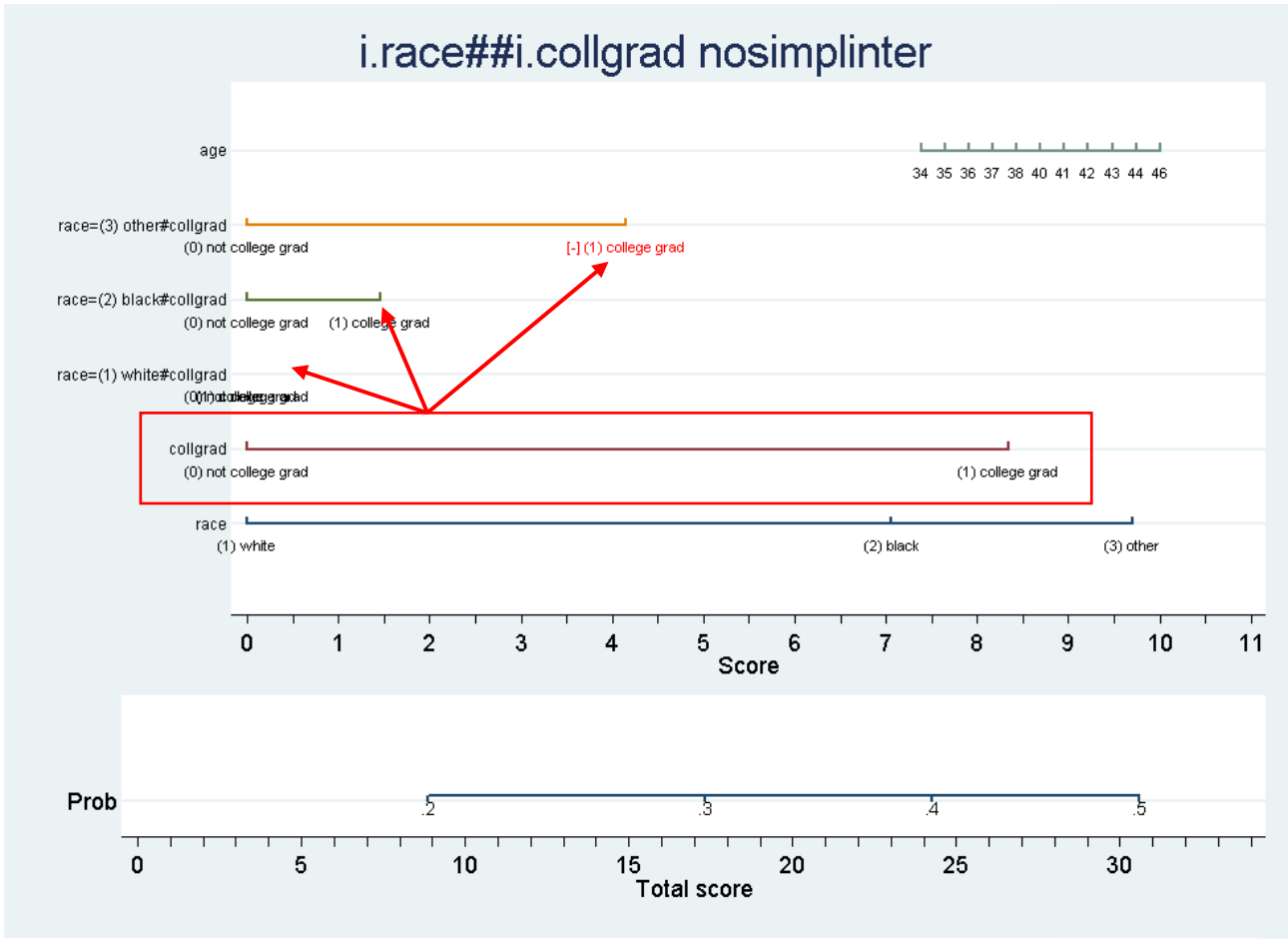


instituto ramón y cajal de investigación sanitaria irycis

Interaction simplification

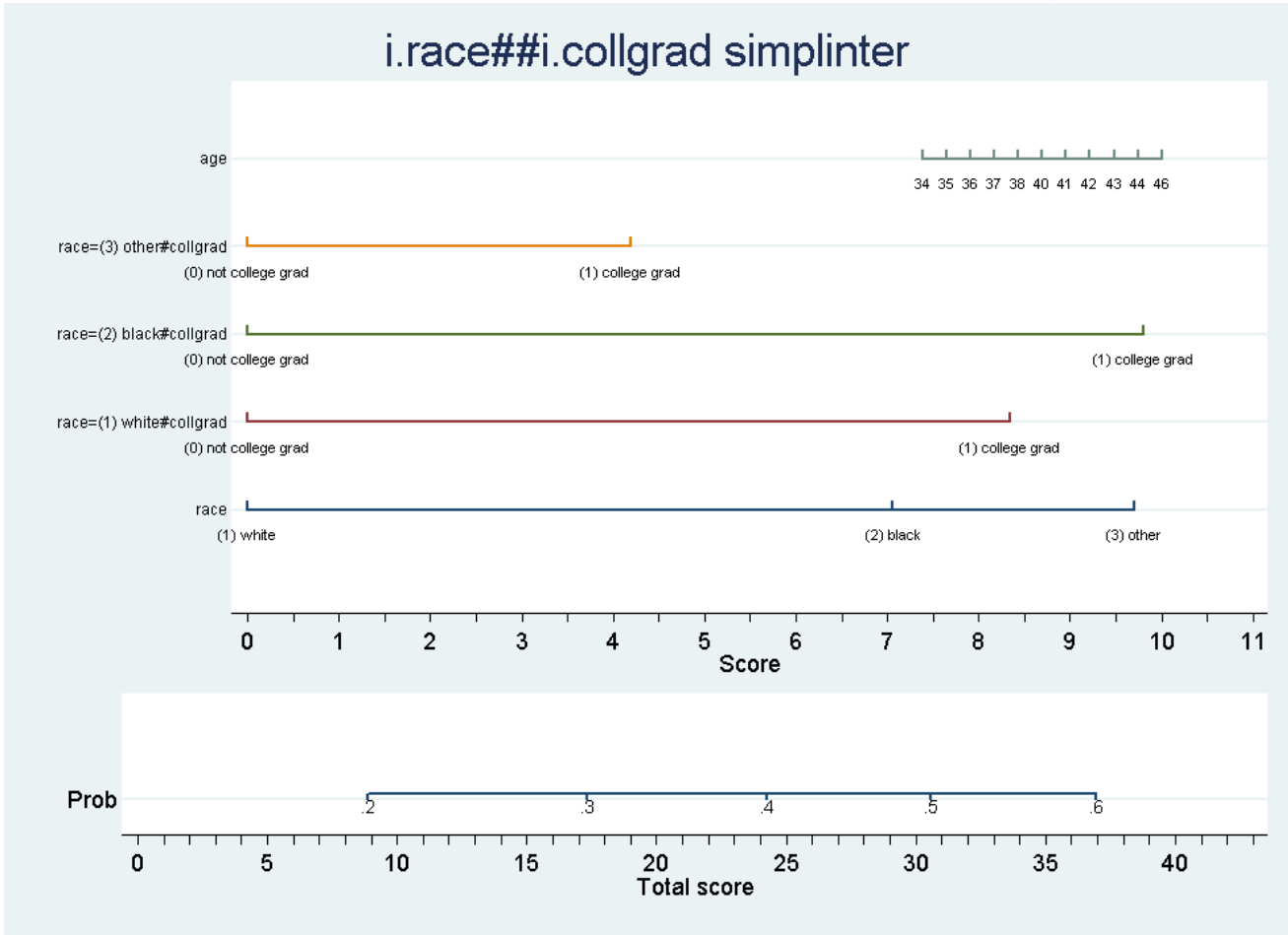
```
sysuse nlsw88, clear
logit union i.race##i.collgrad age
matrix list e(b)

e(b)[1,13]
          union:       union:       union:       union:       union:       union:       union:
             1b.           2.           3.           0b.           1.     1b.race#     1b.race#
           race         race         race     collgrad     collgrad  0b.collgrad  1o.collgrad
y1            0    .45104275     .619655            0     .5325678            0            0

          union:       union:       union:       union:       union:       union:
        2o.race#     2.race#     3o.race#     3.race#
    0b.collgrad   1.collgrad  0b.collgrad   1.collgrad          age        _cons
y1            0    .09356574            0   -.26507807     .0138832   -1.9534634
```

Interaction simplification



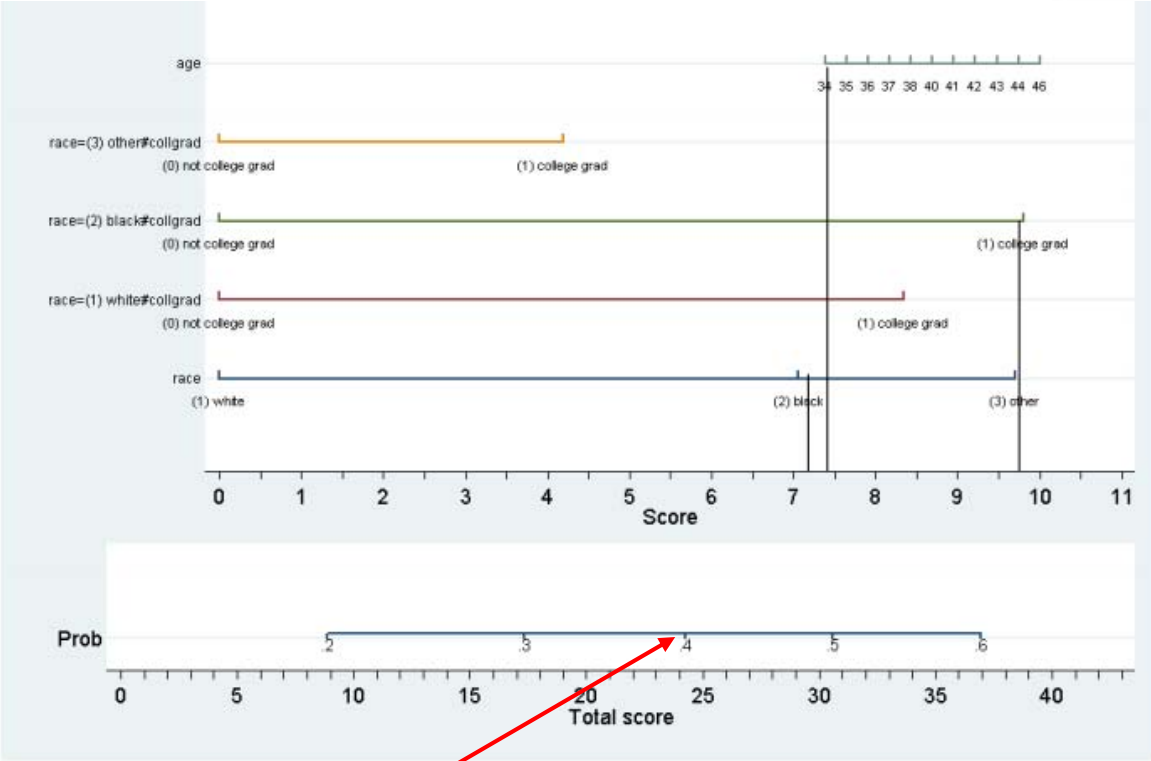
Not simplified interaction

Interaction simplification



Simplified interaction

Interaction simplification



Here we calculate the predicted probabilities with –predict– and compare them to the ones obtained with a nomogram.

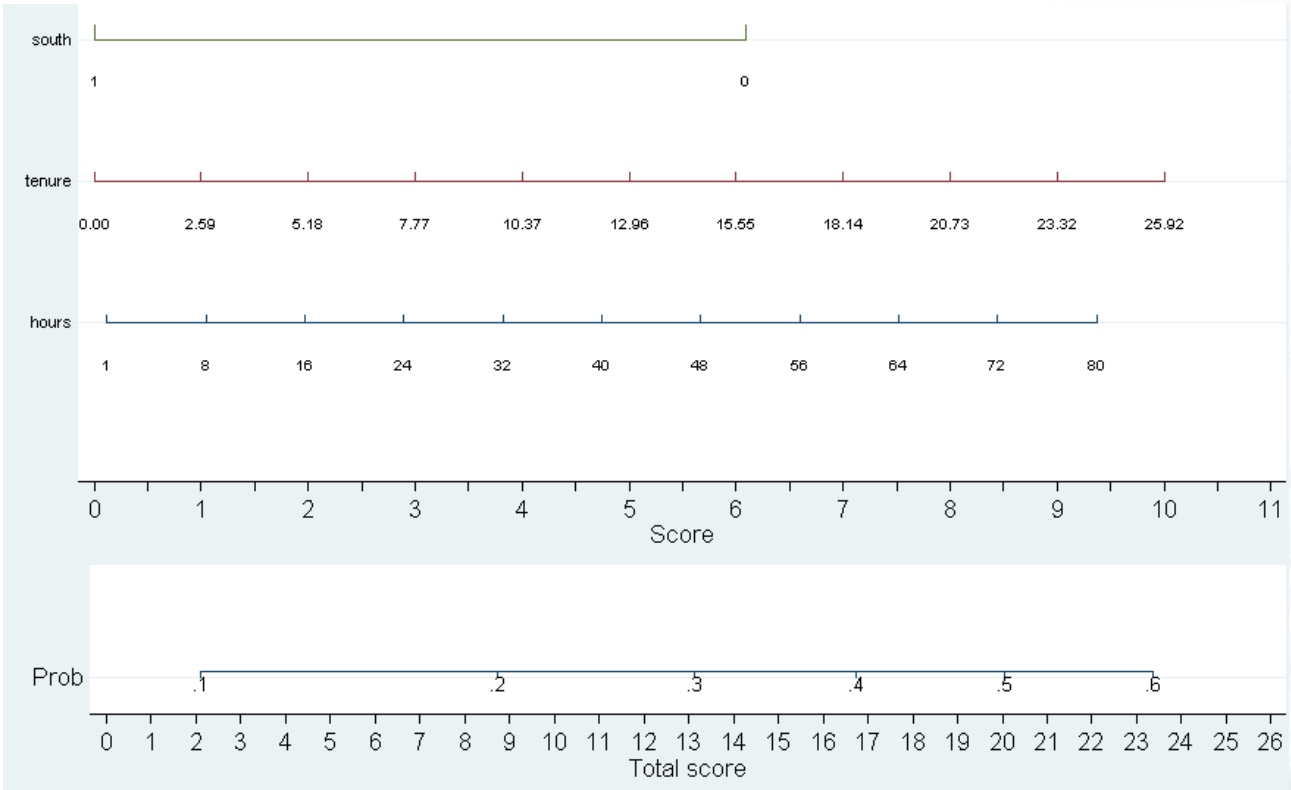
```
age=34 => ≈7.4
race="black" & collgrad="college grad" => ≈9.75
collgrad="college grad" (simplified)
race="black" => ≈7.1

Total score ≈24.25 => Prob ≈0.4
```

	A	B	C	D	E	F	G	R
1	idcode	age	race	married	never_married	grade	collgrad	prob
1007	2341	34	black	married	0	16	college grad	0.400289685
1431	3282	34	black	single	0	17	college grad	0.400289685
2098	4838	34	black	single	0	17	college grad	0.400289685

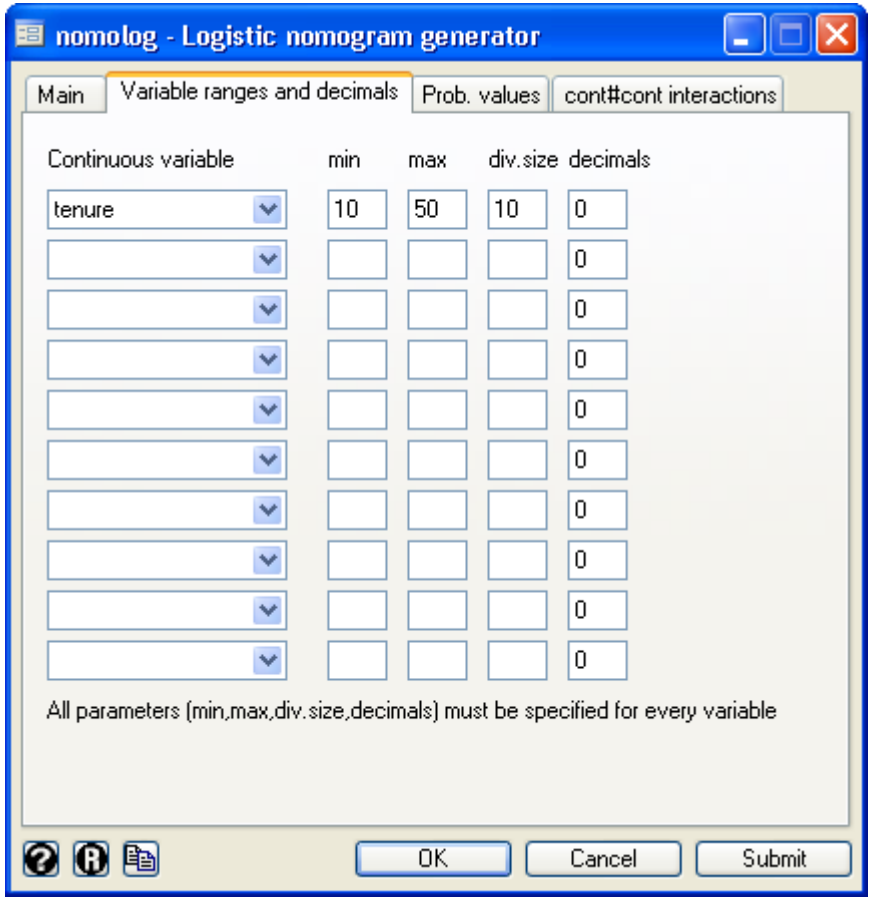
instituto ramón y cajal de investigación sanitaria irycis

Imposed variable ranges

```
sysuse nlsw88, clear
logit union tenure i.south
```



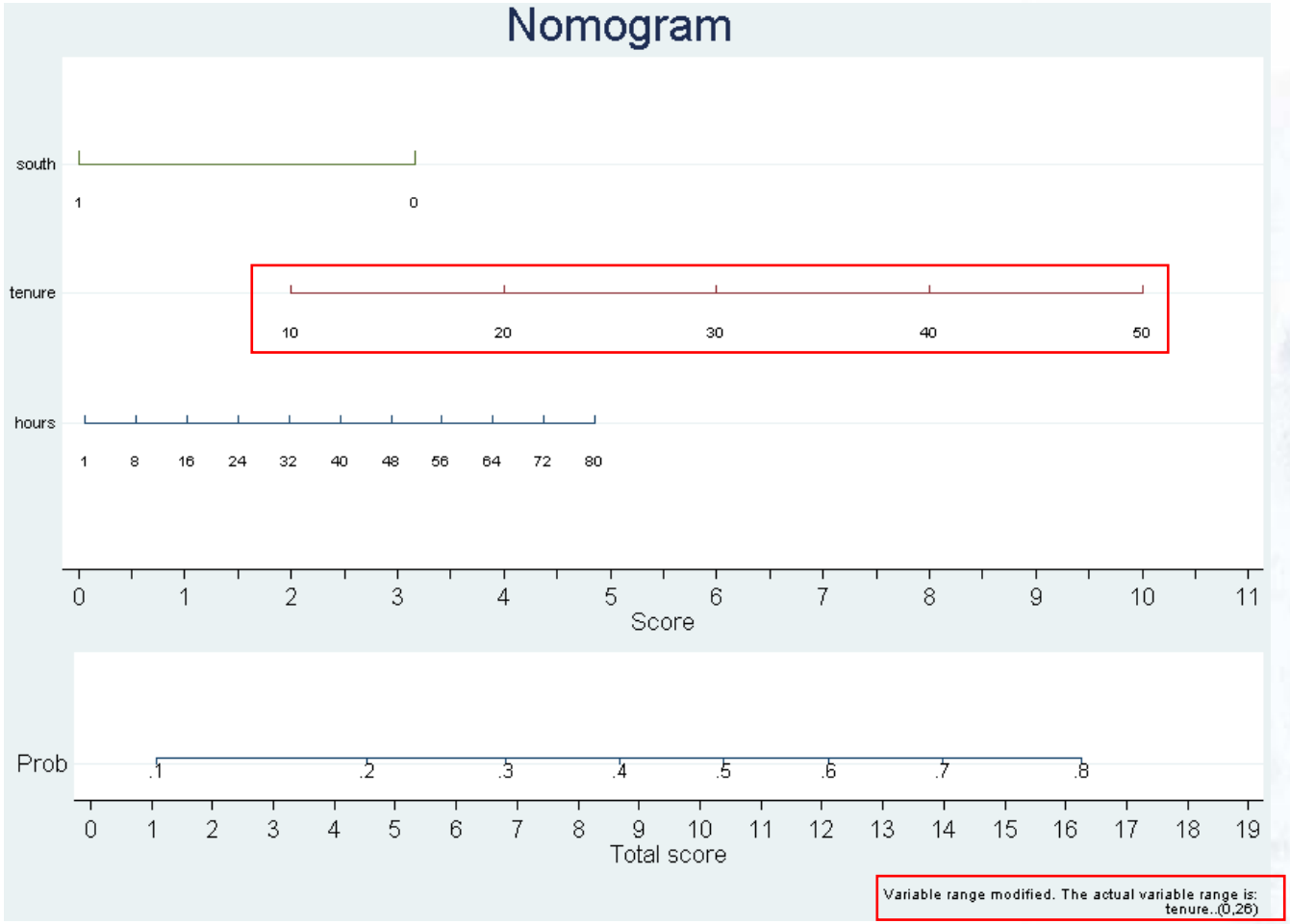
Imposed variable ranges



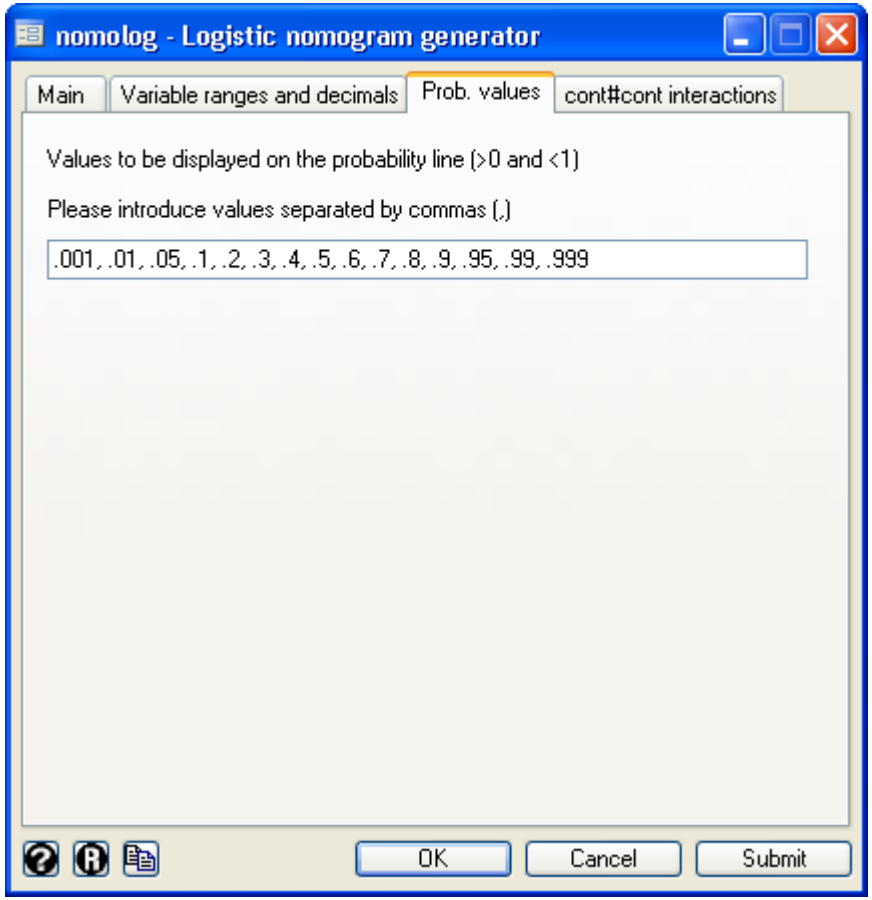
Warning:

Imposing variable ranges for non-existent values will produce out-of-sample predictions

Imposed variable ranges



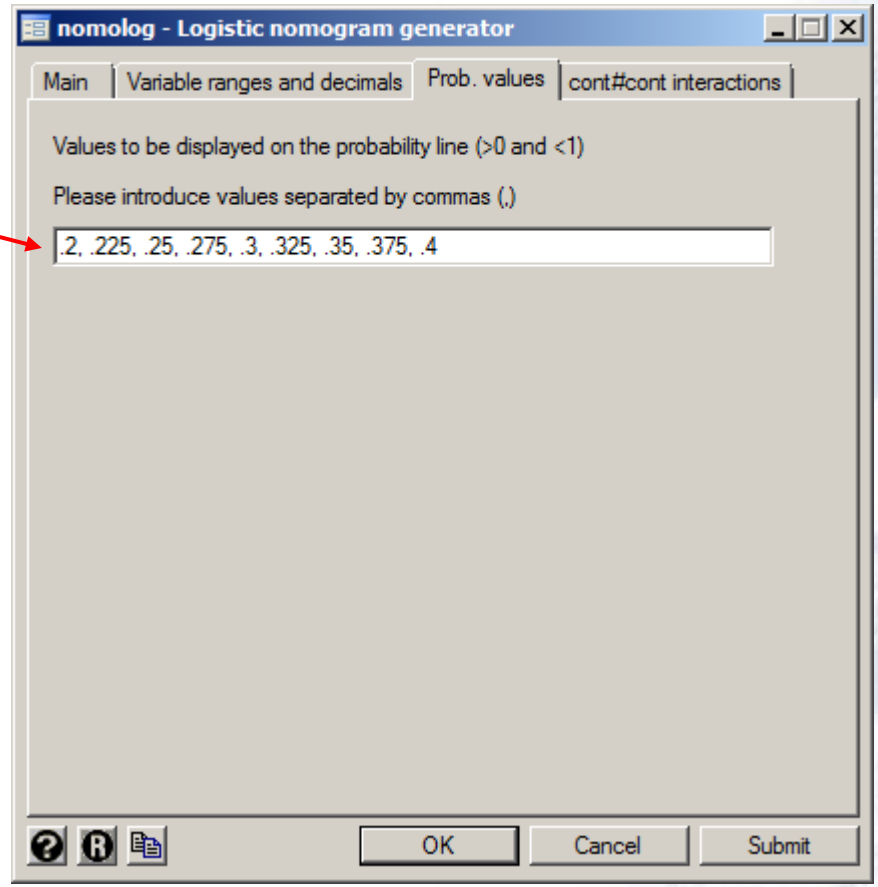
Total Score -> Probability values



Sometimes the default probability line values lack the sufficient resolution in the area of interest

Total Score -> Probability values

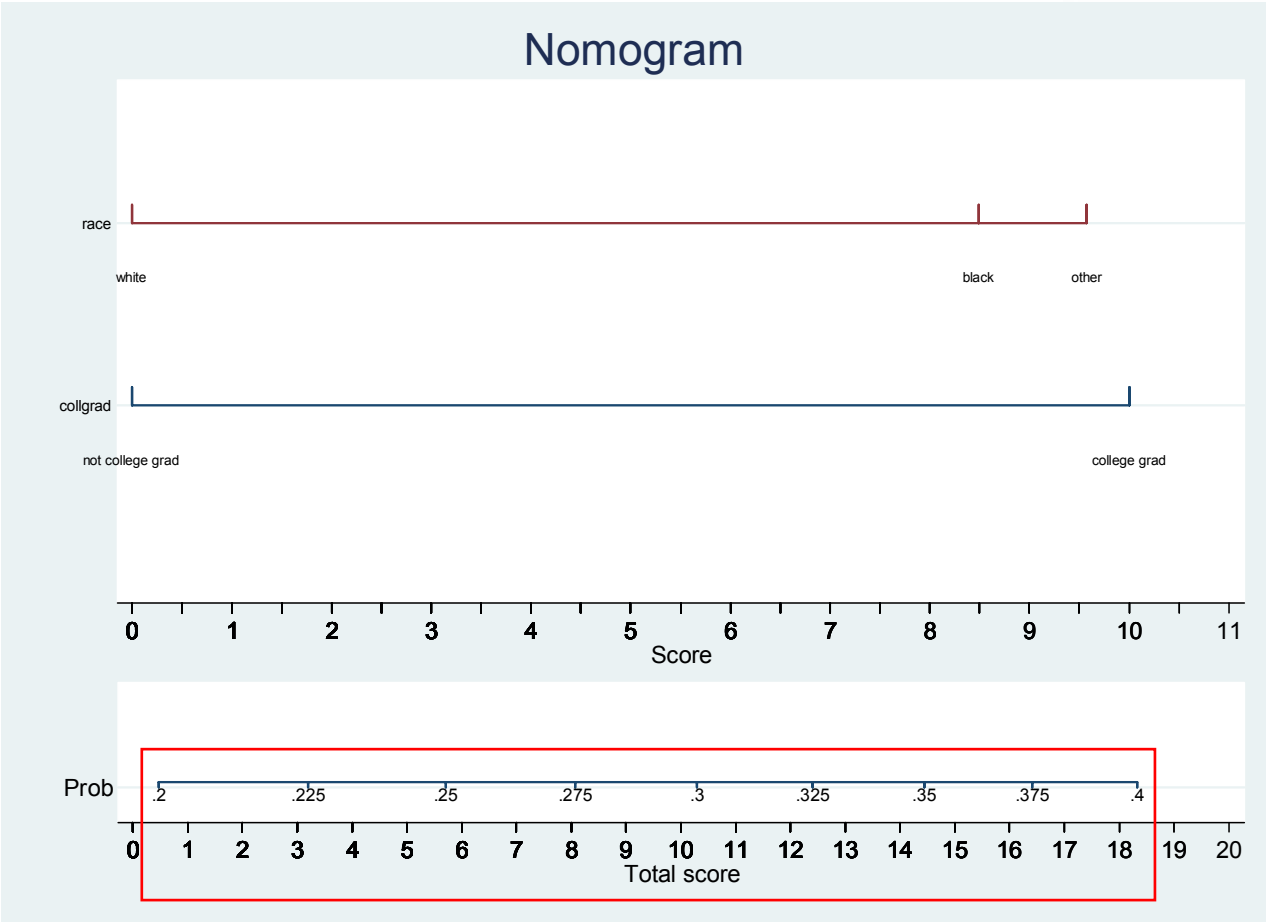
```
sysuse nlsw88
logit union i.collgrad i.race
db nomolog
```



If this is the case,

we can modify them

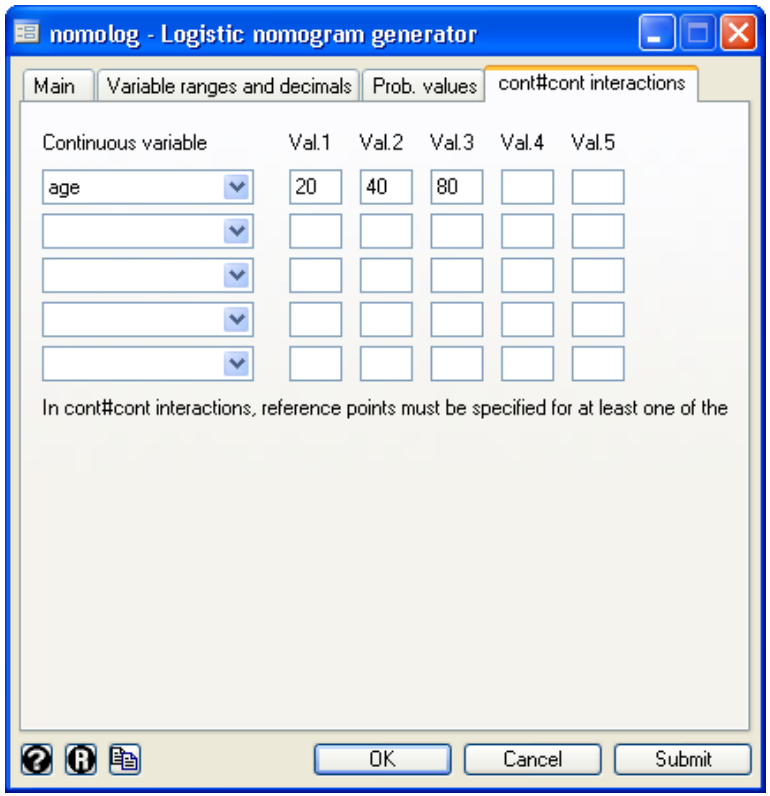
Total Score -> Probability values



instituto ramón y cajal de investigación sanitaria irycis

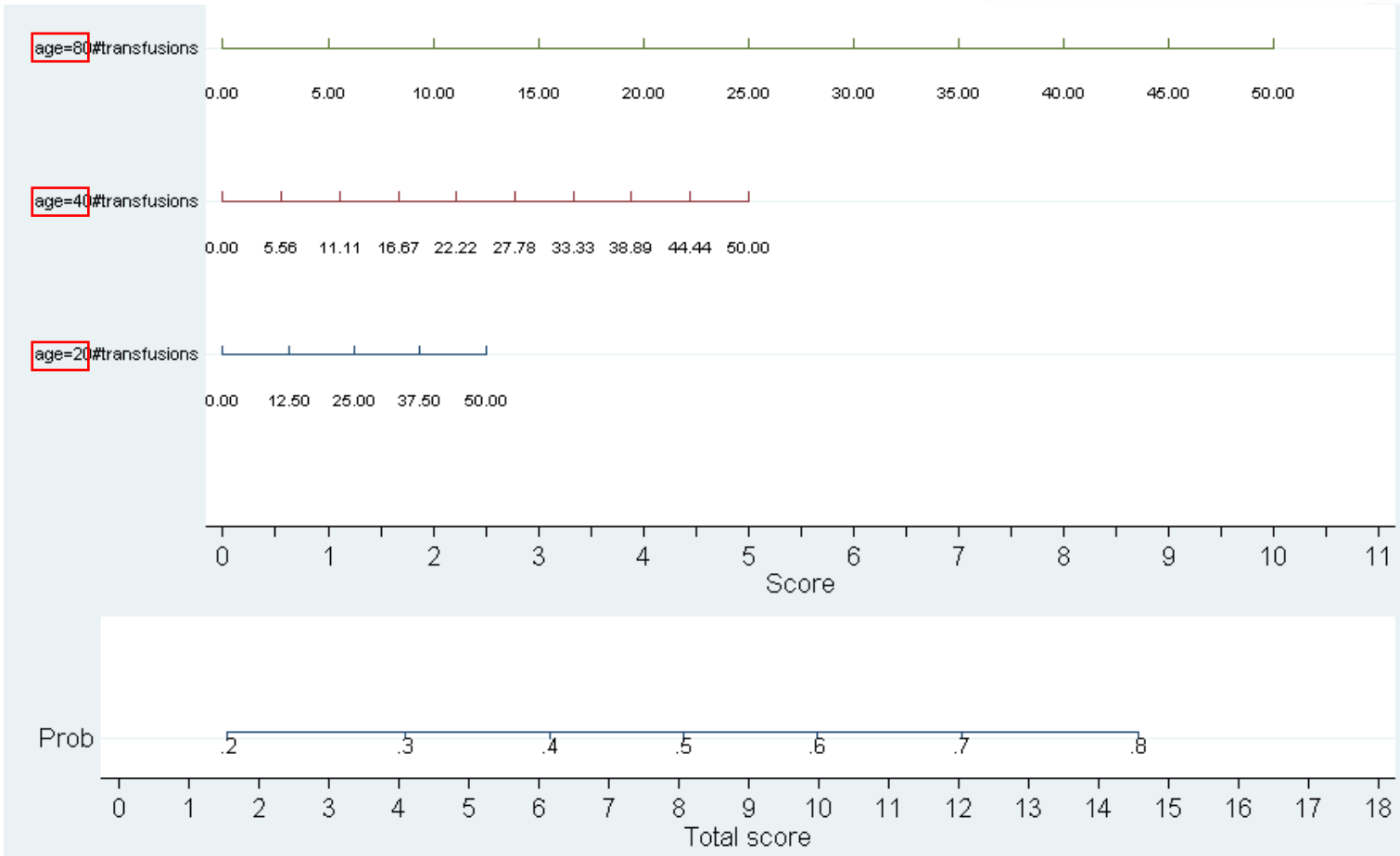
cont # cont interactions

```
sysuse nomolog_ex, clear
logit outcome c.age#c.transfusions
```



cont#cont interactions must be particularized to be represented on a linear nomogram

cont # cont interactions



Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work

Cox regression nomograms

- Logistic regression

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

- Cox regression

$$S(t) = S_0(t)^{\exp(\beta_1 X_1+\ldots+\beta_k X_k)}$$

instituto ramón y cajal de investigación sanitaria irycis

Cox regression nomograms

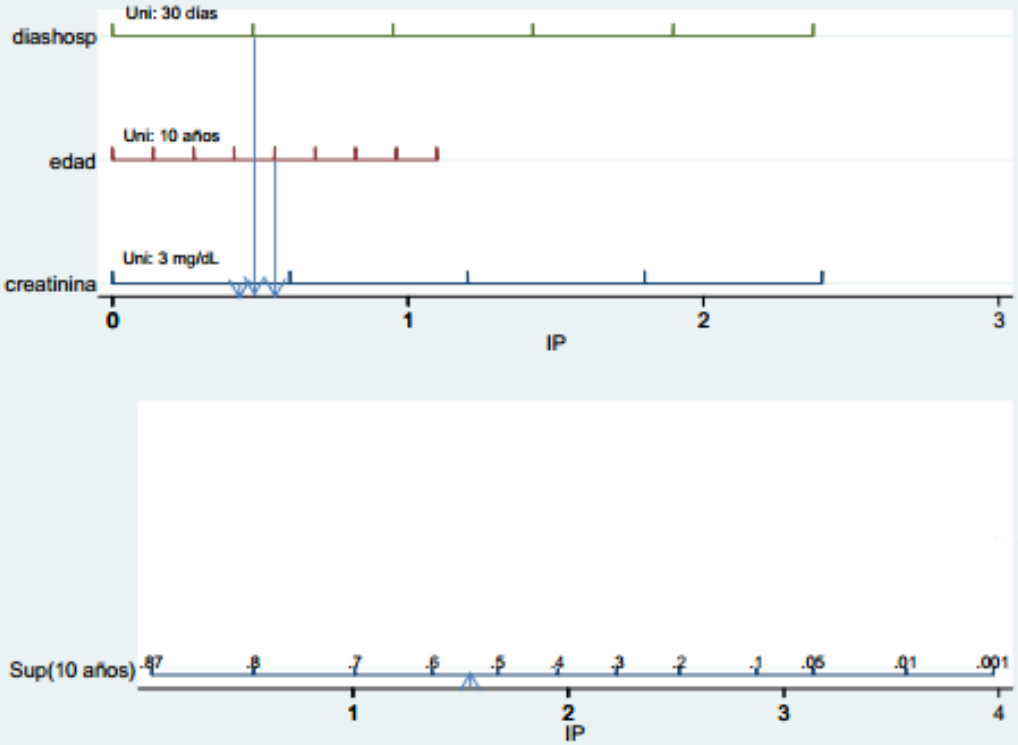
$$S(t) = S_0(t)^{\exp(\beta_1 X_1 + \ldots + \beta_k X_k)}$$

- Coefficients can be obtained from the e(b) matrix

- The base survival can be calculated as

```
use dataset.dta
stset studytime, failure(died==1)
stcox …
predict _s0, basesurv
egen  _sup10y=min(_s0)  if  _t<=10
```

Logistic vs Cox nomograms



The calculation can then be performed in a similar way to the one of the logistic regression nomogram

Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work

Large programs in Stata

Large programs in Stata

- This is a wild guess, but I'd say that in Stata more than 70% of an average programmer's time is spent debugging.

- In large programs it is very important to be able to test individual components and their interactions. If an error is detected, it may be located much faster this way.

- Stata doesn't make this easy.

Large programs in Stata

- Stata lacks a debugger. Using –di– and –set trace on– is very time-consuming.

- Error messages are usually not very informative.

  invalid syntax

  r(198);

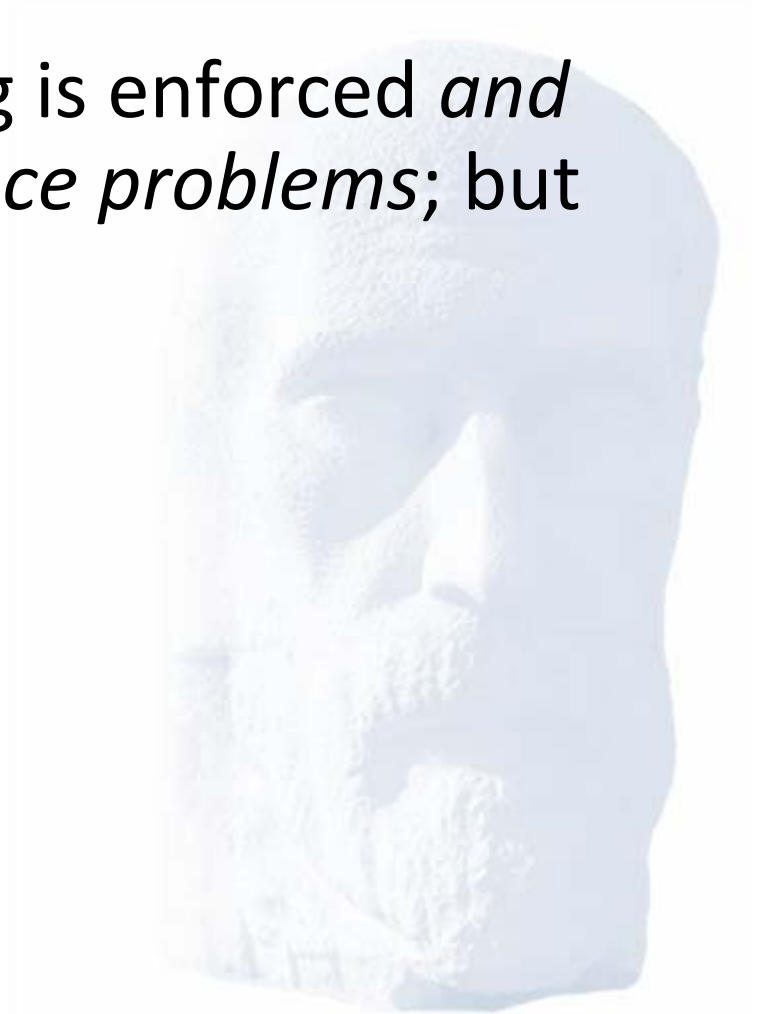
Large programs in Stata

- Curly braces ({ }) positioning is enforced *and can lead to very-hard-to-trace problems*; but indentation is not.
- This is valid syntax

```
if `a' > `b' {
  …
}
```

- This is not

```
if `a' > `b'
{
  …
}
```



Large programs in Stata

- Some recommendations for large ADO programs:
  - Use indentation
  - Use “START x” “ END x“-style comments

```
698     if `iDebug' > 1 {
699      noisily di "asCoef_var_`i'_lvalue_`sValue'_coef=" = `asCoef_var_`i'_lvalue_`
700      noisily di "asCoef_var_`i'_lvalue_`sValue'_refcoef=" = `asCoef_var_`i'_lvalu
701     }
702    }
703    // END else if (`bDxD') {
704    else if (`bDxC' | `bC') {
705     local i = `i' + 1 //real var counter
706    }
707    // END else if (`bDxC')
708   } //END if !strmatch("`rvar'","_cons") & !strmatch("`rvar'","*c.*#*c.*")
709
710   local k = `k' + 1 // e(b) coefficient counter
711  }
```

Large programs in Stata

- Some recommendations for large ADO programs:
  - Create a debug mode and produce some output as the program proceeds
  - Try to use meaningful variable names

```
if `iDebug' > 0 {
 display "sValue_dxd=" "`sValue'"
}

local asCoef_var_`i'_lvalue_`sValue'_coef = rcoefs[1,`k']

if `iDebug' > 1 {
 noisily di "asCoef_var_`i'_lvalue_`sValue'_coef=" = `asCoef_var_`i'_lvalue_`sValue'_coef'
 noisily di "asCoef_var_`i'_lvalue_`sValue'_refcoef=" = `asCoef_var_`i'_lvalue_`sValue'_refcoef'
}
```

Depending on the value of iDebug, we produce more or less output

Large programs in Stata

- The right way to create temporary files is

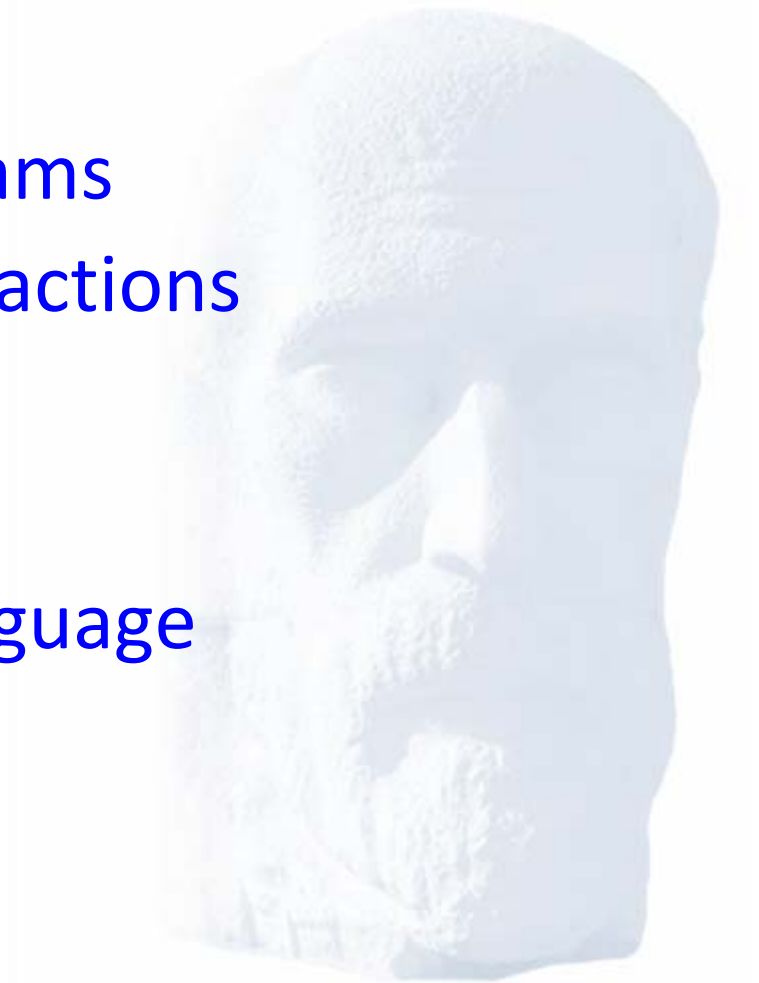
  tempfile *pt_filename*

  This guarantees that these files are unique and that they are deleted once the program ends.

- Try to test the program after each significant change, so that you know which change caused the error.

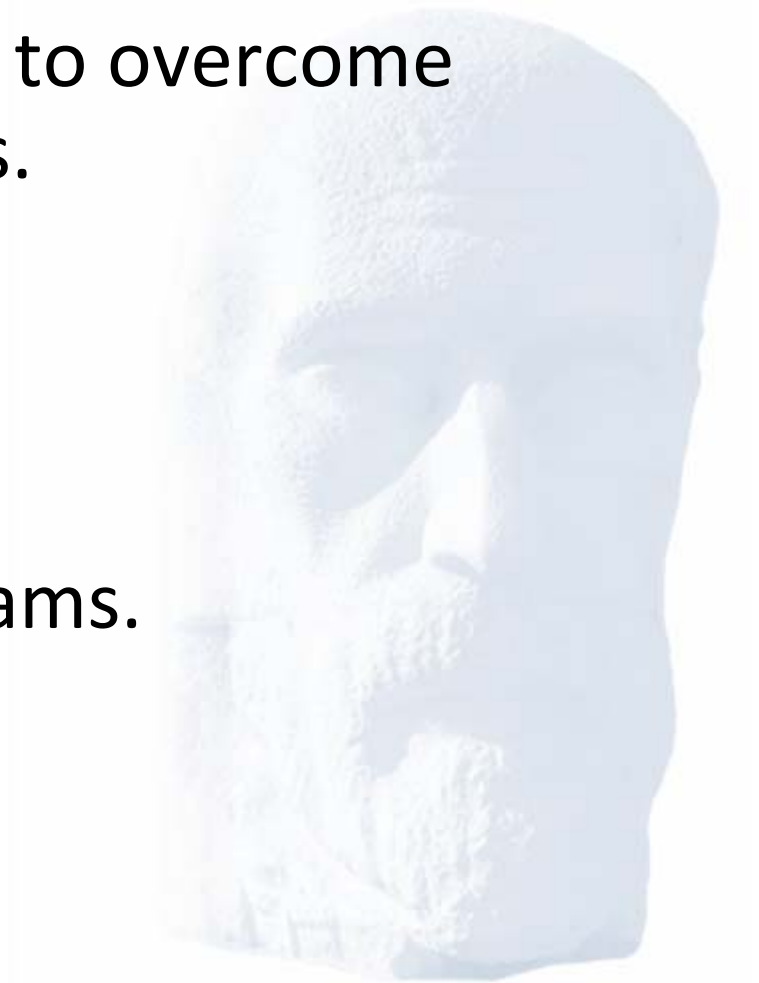
Structure of Presentation

- Introduction
- Logistic regression nomograms
- Positive coefficients & interactions
- The –nomolog– package
- Cox nomograms
- Large programs in Stata language
- Future work



Future work

- Explore **–addplot–** as a way to overcome some of **–xtline–** limitations.
- Cox regression nomograms.
- Poisson regression nomograms.

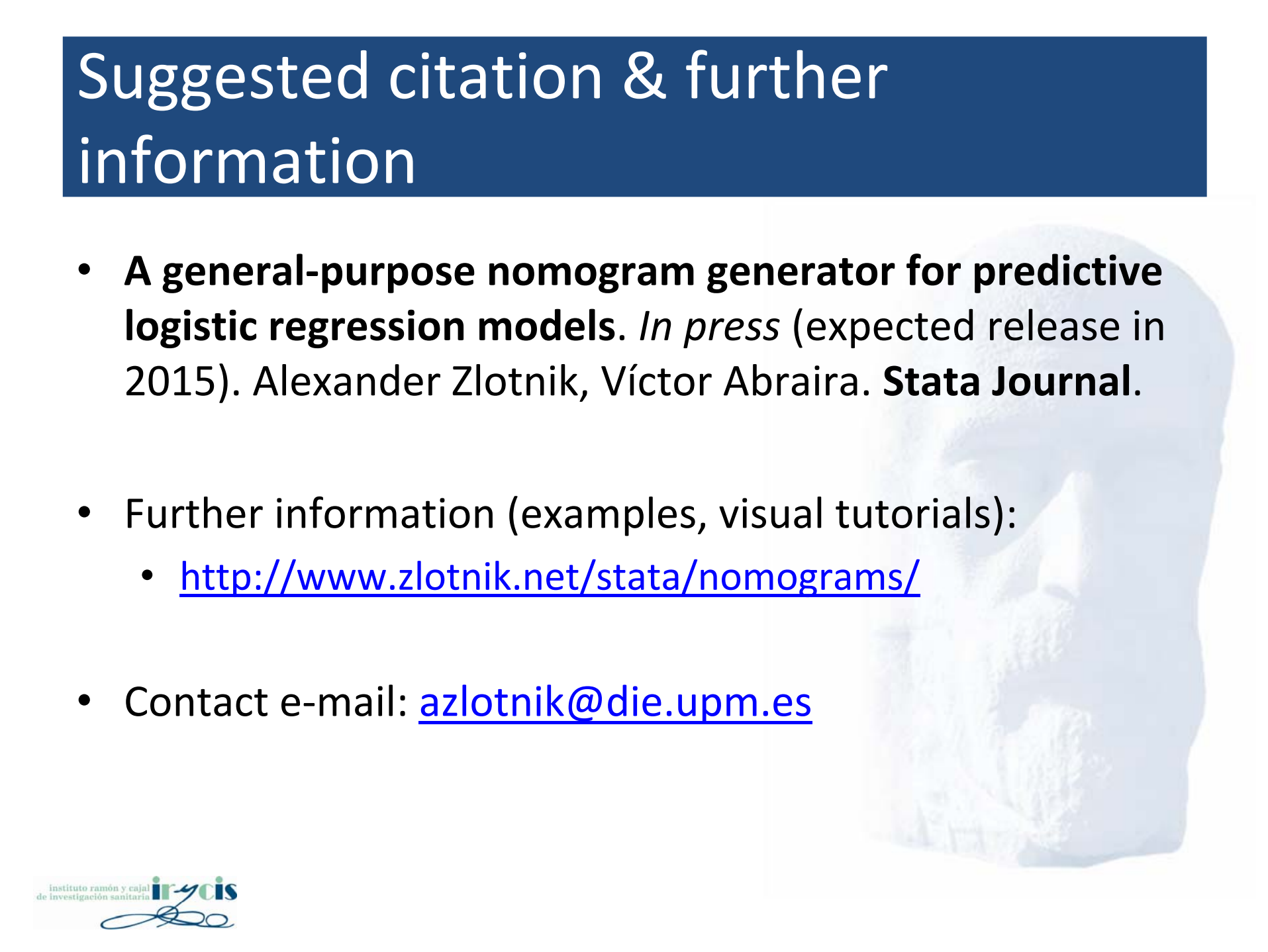


Suggested citation & further information

- **A general-purpose nomogram generator for predictive logistic regression models**. *In press* (expected release in 2015). Alexander Zlotnik, Víctor Abraira. **Stata Journal**.

- Further information (examples, visual tutorials):
  - http://www.zlotnik.net/stata/nomograms/

- Contact e-mail: azlotnik@die.upm.es



instituto ramón y cajal de investigación sanitaria irycis