IOB Institute of Molecular and Clinical Ophthalmology Basel

www.iob.ch

Post-selection inference for generalized linear models

Applying the usual statistical methods after variable selection is common in medical research, but can lead to severe bias. Novel statistical methods for correct post-selection inference have been implemented in the `R` package `selcorr`.

Marco Cattaneo, *Senior Statistician @ Department of Clinical Research, University of Basel & IOB*

Problem:

Variable selection is the process of (automatically) selecting a subset of relevant variables (predictors) for regression, mainly to simplify models and their interpretation. Its main issue is that ***naive*** statistical inference after selection is usually **biased**: p-values and confidence intervals should be corrected.

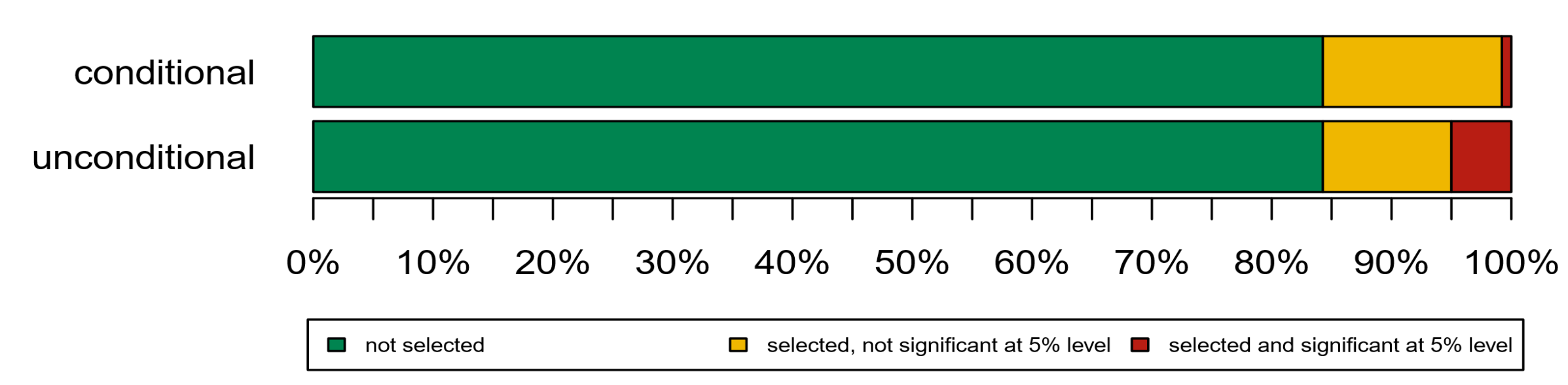
Naive inference after variable selection is common in medical research, mostly without any mention of its potential bias. In particular, **logistic regression** and selection via **AIC** (Akaike information criterion) are often used. In recent years, a few methods for correct statistical inference after variable selection have been published under the name *selective* or *post-selection* inference. However, only a couple of methods seems to have been implemented, and only for linear regression.

Solution:

The `R` package **`selcorr`** (Cattaneo, 2021, version 1.0) provides post-selection inference for generalized linear models and some standard selection procedures (in particular for logistic or linear regression and for AIC selection). It will be extended to other statistical models and selection procedures in future versions. Confidence intervals and p-values for the regression coefficients are corrected by parametric **bootstrap calibration**.

The package performs **unconditional** post-selection inference: if a variable has no effect, not selecting it is considered a correct inference (instead of no inference as for conditional approaches). For each selected variable, the p-value is based on the comparison of the optimal selected model with the one we would have obtained if the variable had not been available.

e.g. typical probabilities for a variable without effect:



Example:

Data from **ProgStar** report 15 (Schönbach et al., 2021, Am J Ophthalmol 230:123–133): 194 right eyes at baseline without missing values for the outcome and the 8 potential predictors. This is an extreme example: the correlation of total volume and mean sensitivity is so strong (r=0.998) that they are practically indistinguishable as predictors.

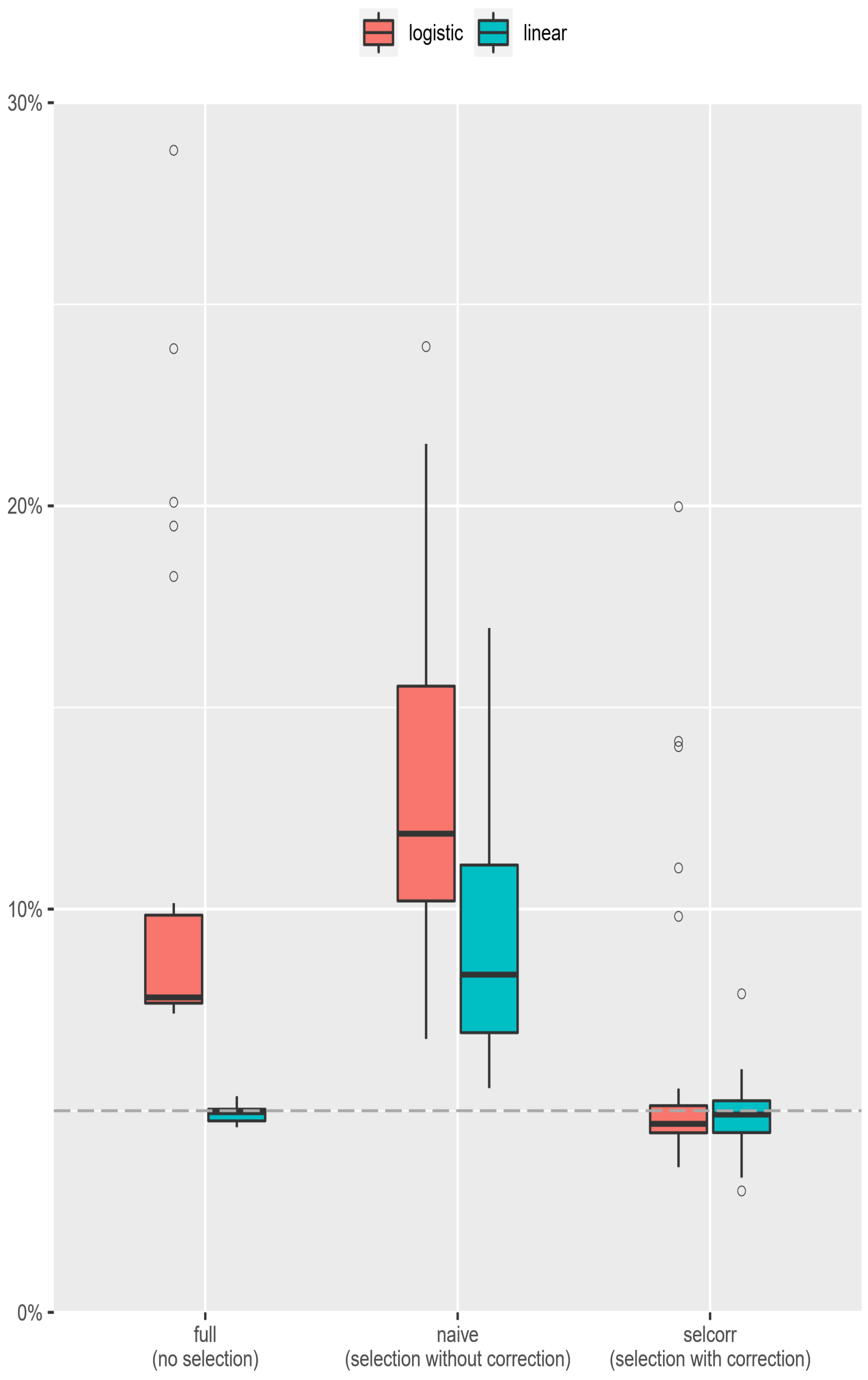
	full (no selection)	naive (selection without correction)	selcorr (selection with correction)
age (years)	0.96 [0.93, 0.99] (p=0.008)	0.95 [0.93, 0.98] (p<0.001)	0.95 [0.92, 0.98] (p=0.001)
BCVA (ETDRS letters)	0.99 [0.96, 1.02] (p=0.484)	----------------------------------	0.00 [-∞ , +∞] (p=1.000)
duration of disease (years)	1.00 [0.94, 1.06] (p=0.931)	----------------------------------	0.00 [-∞ , +∞] (p=1.000)
female sex	0.69 [0.34, 1.42] (p=0.315)	----------------------------------	0.00 [-∞ , +∞] (p=1.000)
mean sensitivity (dB)	0.41 [0.12, 1.31] (p=0.139)	0.70 [0.62, 0.77] (p<0.001)	0.70 [0.30, 1.11] (p=0.124)
RPE pigmentary abnormalities	0.49 [0.21, 1.08] (p=0.081)	0.47 [0.21, 1.04] (p=0.066)	0.47 [0.21, 1.06] (p=0.068)
total volume (0.1 dB-sr)	1.83 [0.49, 7.10] (p=0.371)	----------------------------------	0.00 [-∞ , +∞] (p=1.000)
white ethnicity	2.40 [0.84, 7.54] (p=0.114)	2.59 [0.92, 7.99] (p=0.081)	2.59 [0.88, 8.43] (p=0.084)

***Multivariable logistic regression models** for the effects of potential predictors on the odds of flecks outside of the vascular arcades: odds ratio estimates, with 95% confidence intervals and p-values (based on 10 000 bootstrap replicates for `selcorr`).*

Simulation:

Logistic and linear regressions with and without variable selection, as well as with and without correction were simulated 10 000 times. The linear predictor was the same for the logistic and linear models (only the outcome variable changed between simulations), consisting of **500 data points** for **28 predictor variables** (14 with an effect). Post-selection inference was based on 100 bootstrap replicates.

The logistic regression inferences are approximate even without variable selection, while the linear regression ones are exact (when the model assumptions are satisfied). For linear regression, variable selection (with correction) increases the **statistical power** (sign test: p=0.006), while for logistic regression comparing the power makes no sense, since the inferences without selection are not calibrated.



***Boxplots** of the percentage of simulations in which the 95% confidence interval does not contain the true regression coefficient, for each predictor variable (it should be 5%).*