Post-selection inference

`selcorr` (correction for selection)

Marco Cattaneo

Department of Clinical Research
University of Basel

22 Sempetmber 2022

selcorr

- The main issue of variable selection is that *naive* statistical inference after selection is usually biased: confidence intervals and p-values should be corrected.

selcorr

- The main issue of variable selection is that *naive* statistical inference after selection is usually biased: confidence intervals and p-values 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.

selcorr

- The main issue of variable selection is that *naive* statistical inference after selection is usually biased: confidence intervals and p-values 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 seem to have been implemented, and only for linear regression (but too slow or not directly applicable).

selcorr

- The main issue of variable selection is that *naive* statistical inference after selection is usually biased: confidence intervals and p-values 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 seem to have been implemented, and only for linear regression (but too slow or not directly applicable).
- The R package `selcorr` (version 1.0 on CRAN since 2021-12-21) 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.

selcorr

- The main issue of variable selection is that *naive* statistical inference after selection is usually biased: confidence intervals and p-values 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 seem to have been implemented, and only for linear regression (but too slow or not directly applicable).
- The R package `selcorr` (version 1.0 on CRAN since 2021-12-21) 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.

?selcorr

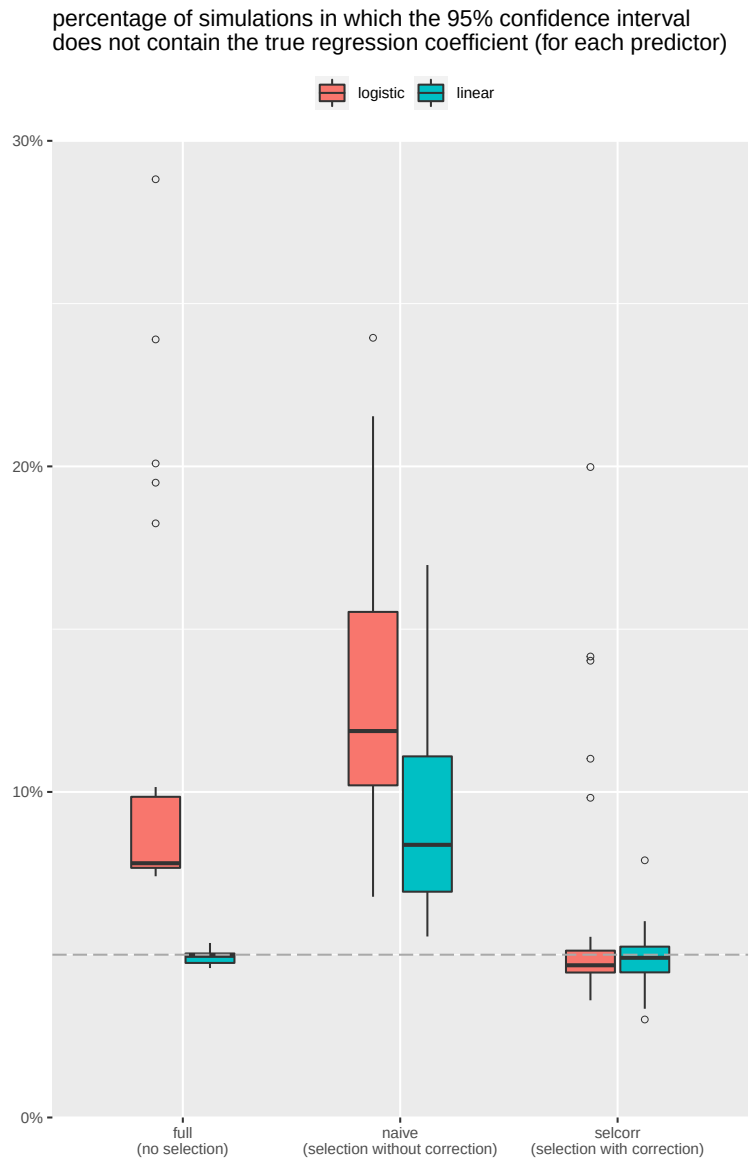
```
Post-Selection Inference for Generalized Linear Models

Description:

     'selcorr' returns (unconditional) post-selection confidence intervals and p-values for the coefficients of
     (generalized) linear models.

Usage:

     selcorr(
       object,
       fixed.vars = NULL,
       further.vars = NULL,
       boot.repl = 0,
       k = 2,
       conf.level = 0.95,
       quiet = FALSE
     )

Arguments:

  object: an object representing a model of an appropriate class. This is used as the initial model in a
          (bidirectional) stepwise model selection.

fixed.vars: the names of all independent variables that must be included in the selected model. The default is none.

further.vars: the names of all independent variables that can be included in the selected model, but are not part of
          'object'. The default is none.

boot.repl: a number or list of bootstrap replicates. The default is no bootstrapping. See Details and Examples for
          clarification.

       k: the multiple of the number of degrees of freedom used as penalty in the model selection. The default 'k =
          2' corresponds to the AIC.

conf.level: the level of the confidence intervals.

   quiet: if 'TRUE', then 'selcorr' does not generate an output.
```

?selcorr

```
Details:

     When 'boot.repl = 0', an approximate asymptotic distribution of the test statistic is used to calculate
     p-values and calibrate the profile-likelihood confidence intervals. This approach is faster, but p-values and
     confidence intervals can be more precisely calibrated by parametrically bootstrapping the test statistic (with
     'boot.repl' the number of replicates). Parallel computing can be used to speed up the bootstrapping: see
     Examples.

Value:

     the selected model is returned, without correction for model-selection, but with up to two additional
     components. There is an 'output' component corresponding to the post-selection inference, which is also printed
     unless 'quiet = TRUE'. When 'boot.repl' is not '0', there is also a 'boot.repl' component corresponding to the
     bootstrap replicates.

Examples:

     ## linear regression:
     selcorr(lm(Fertility ~ ., swiss))

     ## logistic regression:
     swiss.lr = within(swiss, Fertility <- (Fertility > 70))
     selcorr(glm(Fertility ~ ., binomial, swiss.lr))

     ## parallel bootstrapping:
     ## Not run:

     library(future.apply)
     plan(multisession)
     boot.repl = future_replicate(8, selcorr(lm(Fertility ~ ., swiss), boot.repl = 1000,
                                          quiet = TRUE)$boot.repl, simplify = FALSE)
     plan(sequential)
     selcorr(lm(Fertility ~ ., swiss), boot.repl = do.call("rbind", boot.repl))
     ## End(Not run)
```

simulation

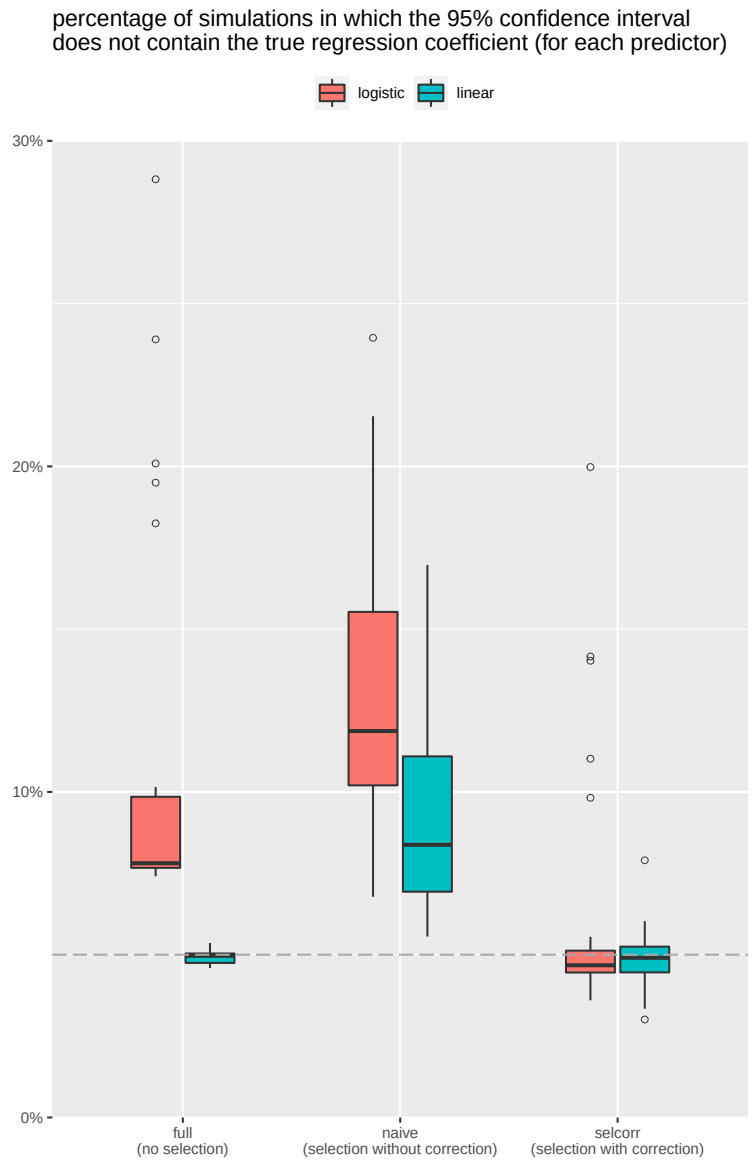
- Logistic and linear regressions with and without variable selection, as well as with and without correction were simulated 10 000 times. The linear predictor was the same for the logistic and linear models (only the outcome variable changed between simulations), consisting of 500 data points for 28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



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).



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.