selcorr

correction for selection

Marco Cattaneo

Department of Clinical Research
University of Basel

30 October 2020

example

```
> summary(step(lm(DV ~ IV1 + IV2 + IV3 + IV4 + IV5 + IV6 + IV7 + IV8, DB)))

Call:
lm(formula = DV ~ IV2 + IV3 + IV4 + IV5, data = DB)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0055  -2.9442   0.0952   3.5790   9.6031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.221133   2.624620   4.275 3.11e-05 ***
IV2         -0.037505   0.025334  -1.480  0.14052
IV3         -0.018400   0.008418  -2.186  0.03013 *
IV4          0.056861   0.016229   3.504  0.00058 ***
IV5          0.015556   0.007696   2.021  0.04476 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.783 on 178 degrees of freedom
Multiple R-squared:  0.07484,        Adjusted R-squared:  0.05405
F-statistic:   3.6 on 4 and 178 DF,  p-value: 0.007543
```

example

```
> summary(step(lm(DV ~ IV1 + IV2 + IV3 + IV4 + IV5 + IV6 + IV7 + IV8, DB)))

Call:
lm(formula = DV ~ IV2 + IV3 + IV4 + IV5, data = DB)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0055  -2.9442   0.0952   3.5790   9.6031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.221133   2.624620   4.275 3.11e-05 ***
IV2         -0.037505   0.025334  -1.480  0.14052
IV3         -0.018400   0.008418  -2.186  0.03013 *
IV4          0.056861   0.016229   3.504  0.00058 ***
IV5          0.015556   0.007696   2.021  0.04476 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.783 on 178 degrees of freedom
Multiple R-squared:  0.07484,        Adjusted R-squared:  0.05405
F-statistic:   3.6 on 4 and 178 DF,  p-value: 0.007543
```

example

```
> summary(step(lm(DV ~ IV1 + IV2 + IV3 + IV4 + IV5 + IV6 + IV7 + IV8, DB)))

Call:
lm(formula = DV ~ IV2 + IV3 + IV4 + IV5, data = DB)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0055  -2.9442   0.0952   3.5790   9.6031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.221133   2.624620   4.275 3.11e-05 ***
IV2         -0.037505   0.025334  -1.480  0.14052      0.1442
IV3         -0.018400   0.008418  -2.186  0.03013 *    0.0430 *
IV4          0.056861   0.016229   3.504  0.00058 ***  0.0016 **
IV5          0.015556   0.007696   2.021  0.04476 *    0.0581 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.783 on 178 degrees of freedom
Multiple R-squared:  0.07484,        Adjusted R-squared:  0.05405
F-statistic:   3.6 on 4 and 178 DF,  p-value: 0.007543  0.0762
```

background

- after variable/model selection, p-values of goodness-of-fit tests for the selected model should usually be corrected upwards (since the model is selected on the basis of the data, it will fit these data relatively well)

background

- after variable/model selection, p-values of goodness-of-fit tests for the selected model should usually be corrected upwards (since the model is selected on the basis of the data, it will fit these data relatively well)

- a very large number of papers in medical research use variable/model selection without corrections for p-values and confidence intervals (often also for central results and without indication of the discarded variables, although in the highest-level medical literature usually not as primary analyses)

background

- after variable/model selection, p-values of goodness-of-fit tests for the selected model should usually be corrected upwards (since the model is selected on the basis of the data, it will fit these data relatively well)
- a very large number of papers in medical research use variable/model selection without corrections for p-values and confidence intervals (often also for central results and without indication of the discarded variables, although in the highest-level medical literature usually not as primary analyses)
- in the last few years, a few correction methods have been suggested under the name of selective/post-selection inference, although mostly with a focus on machine learning approaches (and very large numbers of covariates)

goal

- R package `selcorr` on CRAN providing a user-friendly implementation of post-selection inference with a focus on medical statistics

goal

- R package `selcorr` on CRAN providing a user-friendly implementation of post-selection inference with a focus on medical statistics
- code based on the asymptotic distribution of likelihood ratios (which depends only on the covariance matrix of the independent variables) is available, but should be further developed and documented

goal

- R package `selcorr` on CRAN providing a user-friendly implementation of post-selection inference with a focus on medical statistics
- code based on the asymptotic distribution of likelihood ratios (which depends only on the covariance matrix of the independent variables) is available, but should be further developed and documented
- in a first step, the package will complement the output of the functions `lm` and `glm` with corrected p-values after variable selection by AIC/BIC

goal

- R package `selcorr` on CRAN providing a user-friendly implementation of post-selection inference with a focus on medical statistics
- code based on the asymptotic distribution of likelihood ratios (which depends only on the covariance matrix of the independent variables) is available, but should be further developed and documented
- in a first step, the package will complement the output of the functions `lm` and `glm` with corrected p-values after variable selection by AIC/BIC
- in further steps, the package can then be extended to other statistical models and selection procedures