The Journal of Open Source Software

BetterReg: An R package for Useful Regression Statistics

Christopher L. Aberson[1]

1 Cal Poly Humboldt

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Mehmet Hakan Satman

Reviewers:

- @brunomontezano
- @62442katieb

Submitted: 14 March 2022
Published: unpublished

In partnership with

AMERICAN ASTRONOMICAL SOCIETY

This article and software are linked with research article DOI 10.3847/xxxxx <- update this with the DOI from AAS once you know it., published in the Astrophysical Journal <- The name of the AAS journal..

Summary

Statistics such as squared semi partial correlation, tolerance, and Mahalanobis Distances are useful for reporting the results of OLS Regression (e.g., Tabachnick et al., 2019). Similarly, values such as the Likelihood Ratio χ^2 (e.g. Cohen et al., 2015) and Likelihood R^2 (Menard, 2010).

Squared semipartial correlations sr^2 provide a measure of uniquely explained variances that is on the same scale as R^2 values. $R^2 change$ values quantify how much a set of predictors improves prediction. Comparisons of both dependent and independent regression coefficients provide a significance test addressing whether one coefficient is stronger than another. Tolerance values address multicollinearity by addressing variance unexplained in a predictor. Mahalanabis Distances are a popular measure of multivariate outliers presented on an χ^2 scale. The Likelihood Ratio χ^2 provides a significance test that is more stable than the commonly presented Wald Test for logistic regression. The Likelihood Ratio χ^2 is the most widely recommended Pseudo R^2 statistic for logistic. `BetterReg` is available on CRAN and GitHub (for developmental versions).

Statement of Need

The statistics provided in this package are not part of base R (R Core Team, 2021) or popular packages such `car`(Fox & Weisberg, 2019). To fill these gaps, I developed `BetterReg`.

Useage

`BetterReg` functions require existing regression models (either OLS or Logistic for most statistics), dataset (for some approaches), number of predictors (some functions), and desired amount of output (`Mahal` function).

part function for squared semipartial correlations

The part function requires an existing LM model and indication of number of predictors.

```
library(BetterReg)
mymodel<-lm(y~x1+x2+x3+x4+x5, data=testreg)
parts(model=mymodel, pred=5)

## Predictor 1: semi partial = 0.032; squared semipartial = 0.001
## Predictor 2: semi partial = 0.307; squared semipartial = 0.094
## Predictor 3: semi partial = 0.268; squared semipartial = 0.072
## Predictor 4: semi partial = 0.134; squared semipartial = 0.018
## Predictor 5: semi partial = 0.241; squared semipartial = 0.058
```

`R2change` function for addressing improvement in R^2 between models

The R2change function requires two models. Each model must have the same number of rows.

```
R2change(model1=mymodel1, model2=mymodel2)

## R-square change = 0.09
## F(2,995) = 54.764, p = 2.73174803699611e-23
```

`depbcomp` function for comparing dependent regression coefficients

The `depbcomp` function takes requires data and variable names. Dependent coefficients are coefficients from the same regression model.

```
depbcomp(data=testreg,y=y,x1=x1,x2=x2,x3=x3,x4=x4,x5=x5, numpred=5,comps="abs")

## Pred 1 vs. Pred 2  : t = 7.004, p = 4.57522908448027e-12
## Pred 1 vs. Pred 3  : t = 6.21, p = 7.79647457704868e-10
## Pred 1 vs. Pred 4  : t = 2.751, p = 0.00604702058333784
## Pred 1 vs. Pred 5  : t = 5.31, p = 1.3508334650858e-07
## Pred 2 vs. Pred 3  : t = 0.681, p = 0.495955077475793
## Pred 2 vs. Pred 4  : t = 4.189, p = 3.05299716290008e-05
## Pred 2 vs. Pred 5  : t = 1.612, p = 0.107363700946729
## Pred 3 vs. Pred 4  : t = 3.444, p = 0.000596991746199649
## Pred 3 vs. Pred 5  : t = 0.891, p = 0.373356929374812
## Pred 4 vs. Pred 5  : t = 2.553, p = 0.0108146623166698
```

`indbcomp` function for comparing independent regression coefficients

The `indbcomp` function requires data and variable names from two different samples. Independent coefficients are coefficients from different samples using the same regression model.

indbcomp(model1 = model1_2, model2 = model2_2, comps="abs")

```
## Predictor 1:  t = 0.362, p = 0.718
## Predictor 2:  t = 0.265, p = 0.792
```

`tolerance` function for multicollinearity assumptions

The `tolerance` function requires only a model.

```
mymodel<-lm(y~x1+x2+x3+x4+x5, data=testreg)
tolerance(model=mymodel)

##        x1        x2        x3        x4        x5
## 0.9976977 0.9990479 0.9931082 0.9953317 0.9980628
```

`Mahal` function for detecting multivariate outliers

The `Mahal` function requires model, predictors, and desired number of values to output.

```
mymodel<-lm(y~x1+x2+x3+x4+x5, data=testreg)
Mahal(model=mymodel, pred=5, values=10)

##      537      770      342      760      299      982      446      174
## 14.56342 15.03188 15.56224 15.60986 16.52869 16.80958 17.38597 18.11072
##      458      530
## 20.02762 25.09934
```

LRchi function for Logistic Regression Coefficients

The `LRchi` function takes input for the dependent variable name (y), up to 10 predictors (x1, x2, etc.), and the number of predictors.

```
LRchi(data=testlog, y="dv", x1="iv1", x2="iv2",numpred=2)

## Predictor: iv1; LR squared 34.09, p= 0

## Predictor: iv2; LR squared 0.19, p= 0.67
```

Pseudo function for Logistic Regression Effect Size

The `Psuedo` function takes an existing model as input

```
mymodel<-glm(dv~iv1+iv2+iv3+iv4, testlog,family = binomial())
pseudo(model=mymodel)

## Likelihood Ratio R-squared (McFadden, Recommended) = 0.26
## Cox-Snell R-squared) = 0.301
## Nagelkerk R-squared  = 0.402
```

References

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2015). *Applied multiple regression/correlation analysis for the behavioral sciences*. ISBN: 978-0-203-77444-1 978-1-134-80101-5 978-1-4106-0626-6 978-1-134-80094-0

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Menard, S. W. (2010). *Logistic regression: From introductory to advanced concepts and applications*. SAGE. ISBN: 978-1-4129-7483-7 978-1-4129-7119-5

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (Seventh edition). Pearson. ISBN: 978-0-13-479054-1