


# BetterReg: An R package for Useful Regression Statistics

Christopher L. Aberson <sup>1</sup>

<sup>1</sup> Cal Poly Humboldt

DOI: [10.21105/joss.04280](https://doi.org/10.21105/joss.04280)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

---

Editor: [Mehmet Hakan Satman](#)

## Reviewers:

- [@brunomontezano](#)
- [@62442katieb](#)

Submitted: 14 March 2022

Published: 02 June 2022

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## Summary

Statistics such as squared semi partial correlations, tolerance, and Mahalanobis Distances are useful for reporting the results of OLS (Ordinary Least Squares) Regression ([Tabachnick et al., 2019](#)) as well as Likelihood Ratio Chi-square ([Cohen et al., 2002](#)) and Likelihood R-square ([Menard, 2010](#)). Such statistics are not part of base R ([R Core Team, 2022](#)) popular packages such as *car* ([Fox & Weisberg, 2019](#)). To fill these gaps, the BetterReg package is developed to provide these statistics and measures.

Squared semipartial correlations provide a measure of uniquely explained variances that is on the same scale as  $R^2$  values. Tolerance values address multi-collinearity by addressing variance unexplained in a predictor. Mahalanobis Distance is a popular measure of multi-variate outliers that are presented on a  $\chi^2$  scale. The Likelihood Ratio  $\chi^2$  provides a significance test that is more stable than the commonly presented Wald Test and the Likelihood Ratio  $\chi^2$  is the most widely recommended Pseudo  $R^2$  statistic for the Logistic Regression.

The target audience for this package is researchers using OLS and Logistic Regression. Presently, there is not any R package that provides those statistics, so the calculation requires researchers to write their own code. These statistics are widely available in commercial programs such as SAS, SPSS, and Stata.

## Usage

BetterReg functions require existing regression models (either OLS or Logistic for most statistics), dataset names (for some approaches), number of predictors (some functions), and desired amount of output (the 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 the required data and variable names as arguments. 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 the coefficients obtained 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 produce the 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 likelihood ratio chi square

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

```
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 Pseudo 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, G. S., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge. London, UK. <https://doi.org/10.4324/9780203774441>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression (third edition)*. Sage. Thousand Oaks, U.S.A. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Menard, S. W. (2010). *Logistic regression: From introductory to advanced concepts and applications*. Sage. Thousand Oaks, U.S.A. <https://us.sagepub.com/en-us/nam/logistic-regression/book227554>
- R Core Team. (2022). *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. London, UK. <https://www.pearson.com/us/higher-education/program/Tabachnick-Using-Multivariate-Statistics-7th-Edition/PGM2458367.html>