The Journal of Open Source Software

rigr: Regression, Inference, and General Data Analysis Tools in R

Yiqun T. Chen[1,2], **Brian D. Williamson**[3], **Taylor Okonek**[1], **Charles J. Wolock**[1], **Andrew J. Spieker**[4], **Travis Y. Hee Wai**[5], **James P. Hughes**[1], **Scott S. Emerson**[1], **and Amy D. Willis**[1¶]

1 Department of Biostatistics, University of Washington, Seattle, USA **2** Data Science Institute & Department of Biomedical Data Science, Stanford University, Stanford, USA **3** Kaiser Permanente Washington Health Research Institute, Seattle, USA **4** Vanderbilt University Medical Center, Nashville, USA **5** VA Puget Sound Health Care System, Seattle, USA ¶ Corresponding author

DOI: 10.21105/joss.04847

Software

- Review
- Repository
- Archive

Editor: Øystein Sørensen

Reviewers:

- @stefanocoretta
- @tomsing1

Submitted: 22 September 2022
Published: 26 December 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

Regression models are an essential analytical tool in modern biomedical research, with applications ranging from disease outcomes prediction to vaccine efficacy estimation (Bland & Altman, 1986; Rothman et al., 2008). While R provides a set of comprehensive tools for fitting regression models (e.g., `lm`, `glm` in base R (R Core Team, 2021); `coxph` in the `survival` package (Therneau, 2022)), existing routines require users to navigate multiple R packages with different conventions and syntaxes, creating barriers for both practitioners and learners.

To alleviate these barriers, we developed an R package called `rigr` that facilitates common data analyses in R, with an emphasis on straightforward, modern regression modeling. On a high level, `rigr` compiles output from existing routines together in an intuitive format, and adds functionality to existing functions. For example, users can fit linear models, generalized linear models, and proportional hazards models using a single function `regress`. They can also easily perform complex inference and obtain robust standard errors in this single package. We also provide functions for descriptive statistics and one- and two-sample inference with improved clarity of output.

Statement of need

R has become one of the leading languages used for statistical analyses in the biomedical sciences, with tens of thousands of available open-source packages that can perform statistical tasks ranging from simple *z*-tests to complex deep learning models. However, navigating regression analyses in R can be challenging, especially for biomedical researchers with limited programming experience. In particular, fitting commonly-used regression models (e.g., linear, logistic, Poisson, and proportional hazards regressions) requires different functions in R, each possessing a function-specific syntax. This creates an unnecessary burden and hinders learners from making conceptual connections between these related regression models. Barriers of entry are even higher for obtaining customized, "modern" inferential results. For instance, heteroskedasticity-robust (so-called "sandwich") standard errors, which are robust to certain misspecified variance or covariance structures of outcome variables (King & Roberts, 2015; Mansournia et al., 2021), are not provided by default in the popular `lm`, `glm`, or `coxph` functions in R. Moreover, if researchers want to perform joint inference on a linear combination of regression coefficients, they need to manually extract robust standard errors and then call more specialized packages to perform the calculation. In summary, while combining existing packages in R can provide the necessary tools for both simple and advanced analyses, the process is laden with unnecessary complexity and details.

We designed the `rigr` package to lower the barriers to obtaining robust and interpretable regression results in R; to emphasise connections between commonly-used statistical models; and to facilitate the use of distribution-free inference tools for regression. A single regression function `regress` can fit linear models, frequently-used generalized linear models, and proportional hazards models, and the output of a `regress` call displays exponentiated coefficients and confidence intervals when appropriate. Together, this allows researchers to (i) make connections between different classes of models, and (ii) easily report estimated coefficients on an interpretable scale (e.g., odds ratio for logistic regressions and hazard ratio for proportional hazards regressions). Moreover, heteroskedasticity-robust standard errors are used in inference by default, and tests for linear contrasts and multiple partial *F*-tests (e.g., for testing the null hypothesis that multiple regression coefficients are equal to zero) are easy to specify. This means that `rigr` users can easily test multiple linear hypotheses (which includes ANOVA as a special case) with robust standard error estimates in a single function (`lincom` or anova), which presently requires pooling results from multiple packages in R (including `stats` (R Core Team, 2021), `survival` (Therneau, 2022), `sandwich` (Zeileis et al., 2020) and `car` (Fox & Weisberg, 2019)). Finally, the package also provides easy-to-use functions for descriptive statistics, and one- and two-sample inference with clearly formated outputs. When designing the outputs for functions in `rigr`, we deliberately include only key information on the inferential results while omitting additional details that might be confusing to newcomers to regression analyses.

Examples

regress: General Regression for an Arbitrary Functional

We first demonstrate the use of the `regress` function, which is a single function that can be used to fit linear, logistic, Poisson, and proportional hazards regressions depending on the specified `fnctl` argument.

In the following example, we use `regress` to fit a linear regression model of atrophy (a measure of loss of neurons estimated by the degree of ventricular enlargement relative to the predicted ventricular size; with 0 indicating no atrophy and 100 indicating the most severe degree of atrophy) on sex, weight, age, and race. We see that the robust standard error estimates are reported by default, as are the results of an *F*-test for the null hypothesis that there is no difference in mean atrophy messure across populations matched on sex, weight and age but who differ in their self-reported racial categories. The `mri` dataset represents a subset of the Cardiovascular Health Study dataset (Emerson, 2005; Kuller et al., 2007). In order to preserve patient anonymity and facilitate its use in the classroom, the data were modified by adding random noise and rescaling some variables. Such modifications preserved the general relationships among the variables in the dataset.

```
# Loading library and dataset
library(rigr)
library(dplyr)
data(mri)
# renaming category for display purposes
mri$race_display <- recode(mri$race,
 ”Subject did not identify as White, Black or Asian” = ”Others”)
regress(fnctl = ”mean”, atrophy ~ sex + weight + age + race_display, data = mri)

# Residuals:
#     Min       1Q  Median       3Q      Max
# -34.053   -8.469  -0.538    7.405   54.071

# Coefficients:
```

```
#                     Estimate  Naive SE  Robust SE   95%L      95%H
# [1] Intercept        -22.71     7.646     8.098       -38.61    -6.813
# [2] sexMale           5.441     0.9977    1.034        3.411     7.470
# [3] weight           0.01998   0.01672   0.01787     -0.01510   0.05505
# [4] age               0.7153   0.08478   0.09125       0.5361    0.8944
#     race_display
# [5]    Black         -2.387     2.109     2.055       -6.422     1.648
# [6]    Others        -3.135     3.885     4.148       -11.28     5.008
# [7]    White         -0.2016    1.822     1.790       -3.716     3.313
#                     F stat    df    Pr(>F)
# [1] Intercept        7.87     1     0.0052
# [2] sexMale          27.70    1     < 0.00005
# [3] weight           1.25     1     0.2639
# [4] age              61.45    1     < 0.00005
#     race_display     1.23     3     0.2992
# [5]    Black         1.35     1     0.2459
# [6]    Others        0.57     1     0.4500
# [7]    White         0.01     1     0.9104

# Residual standard error: 11.99 on 728 degrees of freedom
# Multiple R-squared:  0.1456,  Adjusted R-squared:  0.1385
# F-statistic: 18.17 on 6 and 728 DF,  p-value: < 2.2e-16
```

The next set of code demonstrates how regress can be used to run other regression models.

```
# Logistic regression of atrophy on sex height, weight, race, and
# a degree-2 polynomial of ldl.
regress(fnctl = ”odds”, diabetes ~ height+weight*sex +
 polynomial(ldl, 2), data = mri)

# Fitting a proportional hazards regression (Cox regression).
library(survival)
regress(fnctl = ”hazard”, Surv(obstime, death)~age+race, data=mri)
```

In both cases, the outputs will include the exponentiated coefficients (i.e., coefficients on the odds ratio and hazard ratio scales, respectively) by default.

Testing linear hypotheses

We now demonstrate how to produce point estimates, interval estimates, and p-values for linear combinations of regression coefficients using rigr.

```
# Linear regression of LDL on age, stroke, and race (with robust SE by default)
ldl_reg <- regress (”mean”, ldl~age+stroke, data = mri)

# Testing coefficient created by .5*age - stroke (the first 0 comes
# from excluding the intercept)
single_comb <- c(0, 0.5, -1)
lincom(ldl_reg, single_comb)

# H0: 0.5*age-1*stroke   =  0
# Ha: 0.5*age-1*stroke  !=  0
#      Estimate Std. Err.   95%L   95%H      T Pr(T > ¦t¦)
# [1,]   -1.367     2.152 -5.593  2.859 -0.635       0.526

# Test multiple combinations:
# .5*age - stroke = 0 and Intercept + 60*age = 125 (two marginal tests)
```

```
multiple_comb <- matrix(c(0, 0.5, -1, 1, 60, 0), byrow = TRUE, nrow = 2)
lincom(ldl_reg, multiple_comb, null.hypoth = c(0, 125))

# H0: 0.5*age-1*stroke   =  0
# Ha: 0.5*age-1*stroke  !=  0
#      Estimate Std. Err.   95%L   95%H      T Pr(T > ¦t¦)
# [1,]   -1.367     2.152 -5.593  2.859 -0.635       0.526

# H0: 1*(Intercept)+60*age   =  125
# Ha: 1*(Intercept)+60*age  !=  125
#      Estimate Std. Err.    95%L    95%H     T Pr(T > ¦t¦)
# [1,]  126.989     3.568 119.984 133.994 0.557       0.577

# Test joint null hypothesis:
# H0: .5*age - stroke = 0 AND Intercept + 60*age = 125 (one joint test)
lincom(ldl_reg, multiple_comb, null.hypoth = c(0, 125), joint.test = TRUE)
#      Chi2 stat df p value
# [1,]    0.6911  2   0.708
```

Acknowledgments

The authors gratefully acknowledge contributions from Thomas Lumley, Terry M. Therneau and Solomon Lim, and support from colleagues in the Department of Biostatistics at the University of Washington.

References

Bland, J. M., & Altman, D. G. (1986). Regression analysis. *The Lancet*, *327*(8486), 908–909. https://doi.org/10.1016/S0140-6736(86)92832-1

Emerson, S. S. (2005). *Documentation for MRI and Cerebral Atrophy*. http://www.emersonstatistics.com/datasets/mri.pdf

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

King, G., & Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, *23*(2), 159–179. https://doi.org/10.1093/pan/mpu015

Kuller, L. H., Arnold, A. M., Longstreth Jr, W., Manolio, T. A., O'Leary, D. H., Burke, G. L., Fried, L. P., & Newman, A. B. (2007). White matter grade and ventricular volume on brain MRI as markers of longevity in the cardiovascular health study. *Neurobiology of Aging*, *28*(9), 1307–1315. https://doi.org/10.1016/j.neurobiolaging.2006.06.010

Mansournia, M. A., Nazemipour, M., Naimi, A. I., Collins, G. S., & Campbell, M. J. (2021). Reflection on modern methods: Demystifying robust standard errors for epidemiologists. *International Journal of Epidemiology*, *50*(1), 346–351. https://doi.org/10.1093/ije/dyaa260

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

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (Vol. 3). Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Therneau, T. M. (2022). *A package for survival analysis in R*. https://CRAN.R-project.org/package=survival

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*(1), 1–36. https://doi.org/10.18637/jss.v095.i01