

Scientific Notes

R-environment package for regression analysis

Emmanuel Arnhold⁽¹⁾

⁽¹⁾Universidade Federal de Goiás, Escola de Veterinária e Zootecnia, Campus Samambaia, Caixa Postal 131, CEP 74001-970 Goiânia, GO, Brazil. E-mail: earnhold@pq.cnpq.br

Abstract – The objective of this work was to develop a package in the R environment for automating and facilitating the regression analysis. Named *easyreg*, the package offers five functions. The *er1* function performs analyses in 13 models, including linear, nonlinear, and mixed models. The *er2* function considers the lack of fit in the analyses and in the following designs: completely randomized, randomized complete block, Latin squares, and repeated Latin squares. The *regplot* function generates graphics; the *bl* function estimates two-segment models; and the *regtest* function tests the equality of parameters and the identity of the models. These functions allow of a great number of analyses and confer practicality and versatility to the regression analysis.

Index terms: data analysis, equality of parameters, experimental designs, lack of fit, mixed models, model identity.

Pacote em ambiente R para análises de regressão

Resumo – O objetivo deste trabalho foi desenvolver um pacote em ambiente R para automatizar e facilitar análises de regressão. Denominado *easyreg*, o pacote disponibiliza cinco funções. A função *er1* realiza análises em 13 modelos, inclusive modelos lineares, não lineares e mistos. A função *er2* leva em conta a falta de ajuste nas análises e nos seguintes delineamentos: inteiramente casualizado, blocos ao acaso, quadrados latinos e quadrados latinos repetidos. A função *regplot* gera gráficos; a função *bl* estima modelos bissegmentados; e a função *regtest* testa a igualdade dos parâmetros e a identidade dos modelos. Estas funções permitem um grande número de análises e conferem praticidade e versatilidade à análise de regressão.

Termos para indexação: análise de dados, igualdade de parâmetros, delineamento experimental, falta de ajuste, modelos mistos, identidade de modelos.

The R environment (R Core Team, 2017) was created in 1996 by Ross Ihaka and Robert Gentleman, at the University of Auckland, New Zealand (Paternelli & Mello, 2011). Collaborators from different locations worldwide have further developed it. Among other advantages, its functions can be extended because of its easy programming, and its system of “packages” containing specific functions that considerably increase the capacity of analysis. R software is widely used in universities, research centers, and businesses. It is an important technological tool for the analysis and manipulation of data, competing with the best statistical software for this purpose, with the advantage of being free of charge and freely available for Mac, Windows, and Linux platforms.

The R base system displays many functions, as well as packages able to perform regression analyses. However, these functions generally perform separate analyses, and different functions are necessary to create a model, test parameters, or to analyze residues, in order to obtain a greater analysis control. However, for less experienced users, these many functions can turn the analyses a very difficult task.

Many packages have been developed that offer functions for automating analyses. These packages include “multcomp” (Hothorn et al., 2008), “pedigreemm” (Vazquez et al., 2010), “ExpDes” (Ferreira et al., 2013), “easynova and ds” (Arnhold, 2013, 2014), “GGEBiplotGUI” (Frutos et al., 2014), “ScottKnott” (Jelihovschi et al., 2014), “lsmeans” (Lenth, 2016), and “agricolae” (Mendiburu, 2016).

With these packages, analyses can be performed using R base functions, or creating new functions, thus offering users a more practical means of conducting regression analyses. These packages have been used by both less experienced users and users seeking practicality and versatility in their analyses.

With this approach, the present package, *easyreg* was developed, aiming at automating regression analyses in very common models and in agricultural sciences. The package's guide offers many examples of applications to agricultural data. The five functions (*er1*, *er2*, *regplot*, *regtest*, and *bl*) included in the package, in the R environment (Arnhold, 2016), are described as follows.

The *er1* function can perform regression analysis in 13 models (Table 1), including linear, nonlinear, and mixed models. This function extracts parameters from the models for analyses and other uses, and offers parameter testing and measures related to the quality of models, such as the coefficient and adjusted coefficient of determination, Akaike's information criterion (AIC), and Bayesian information criterion (BIC). Residuals, standard residues, discrepant data, and residual normality test are also provided. For some models, the maximum and minimal values, plateau, and line breaks are also estimated.

The mixed models are performed using the *nlme* function. It is possible to estimate models with all random coefficients.

The *er2* function performs regression analysis based on the method of lack of fit. It considers completely randomized designs, randomized complete block

designs, Latin squares, and repeated Latin squares. The models considered are linear, quadratic, and cubic. This function estimates model parameters, and offers parameter testing (considering the design and the lack of fit), as well as the coefficient of determination and adjusted coefficient of determination.

The *regplot* function creates graphics and allows of the insertion of data and equations. An example of the *regplot* function is given in Figure 1. Here, a linear model was estimated using a plateau of the weight gain in the function of the methionine level in turkey chicks. In the *regplot* function, data should be inserted into a table, including explanatory and dependent variables in the first and second columns, respectively. The argument "design" describes the model. The model number can be found in the help function and description given in Table 1. In addition, defining the number of digits (*digits*), legend position (*position*), and the axes label (*xlab* and *ylob*) (Figure 1) is possible.

Like the *regplot* function, the *bl* function also creates figures. However, this function is specific to the analysis of models with two linear segments.

The *regtest* function performs tests to evaluate the equality of parameters and the identity of regression models based on the methodology of Regazzi (1993, 1999, 2003) and Regazzi et al. (2004). With this function, it is possible to apply tests in all models described by the *er1* function (Table 1).

Finally, similarly to packages such as *easyanova* (Arnhold, 2013) and *ExpDes* (Ferreira et al., 2013), and many others available for the R environment, the functions from the *easyreg* package provide results in

Table 1. Models available in the *er1*, *regplot*, and *regtest* functions.

Name	Mathematical description	Model number
Linear	$y \sim a + bx$	1
Quadratic	$y \sim a + bx + cx^2$	2
Linear plateau	$y \sim a + b \times (x - c) \times (x \leq c)$	3
Quadratic plateau	$y \sim (a + bx + c \times I(x^2)) \times (x \leq -0.5 \times b/c) + (a + I(-b^2 / (4c))) \times (x > -0.5 \times b/c)$	4
Two linear	if else ($x \geq d$, $(a - c \times d) + (b + c) \times x$, $a + b \times x$)	5
Exponential	$y \sim a \times \exp(bx)$	6
Logistic	$y \sim a \times (1 + b \times (\exp(-c \times x)))^{-1}$	7
van Bertalanffy	$y \sim a \times (1 + b \times (\exp(-c \times x)))^3$	8
Brody	$y \sim a \times (1 + b \times (\exp(-c \times x)))$	9
Gompertz	$y \sim a \times \exp(-b \times \exp(-c \times x))$	10
Lactation curve	$y \sim (a \times x^b) \times \exp(-c \times x)$	11
Ruminal degradation curve	$y \sim a \times (1 - \exp(-c \times x))$	12
Logistic bi-compartmental	$y \sim (a / (1 + \exp(2 - 4 \times c \times (x - e)))) + (b / (1 + \exp(2 - 4 \times d \times (x - e))))$	13

```

> require (easyreg)
> methionine=c (70, 75, 80, 85, 90, 95, 100, 105, 110)
> gain=c (98, 113, 123, 131, 138, 140, 141,139, 141)
> data=data.frame (methionine, gain)
> # linear plateau
> regplot (data, model=3, digits=2,
position=8, xlab="Methionine (% of NRC)",
ylab="Gain (g)")
$gain
$gain$`Linear plateau model`

              estimates
coefficient a      139.80
coefficient b         2.18
coefficient c       88.30
p-value t. test for a      0:00
p-value t. test for b      0.00
p-value t. test for c      0.00

```

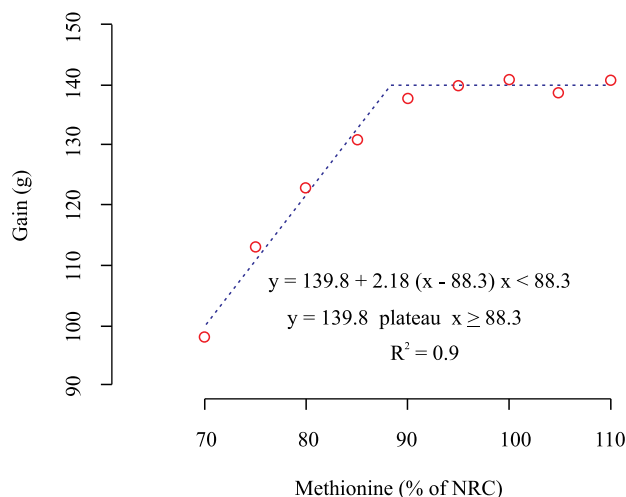


Figure 1. Example of an application of the regplot function with the programming in the console and the respective graph. This example considers a linear function with a plateau for daily weight gain (g) in the function of the methionine quantity (% of NRC) in turkey chicks.

a practical manner. Therefore, the package can aid less experienced users, or users who have some difficulty in using the R software for data analyses. It can also help users who seek agility in the process of data analysis.

References

- ARNHOLD, E. **easyreg**: Easy Regression. R package version 1.0. 2016. Available at: <<http://CRAN.R-project.org/package=easyreg>>. Accessed on: Nov. 18 2016.
- ARNHOLD, E. Package in the R-environment for analysis of variance and complementary analyses. **Brazilian Journal of Veterinary Research and Animal Science**, v.50, p.488-492, 2013. DOI: 10.11606/issn.1678-4456.v50i6p488-492.
- ARNHOLD, E. Pacote em ambiente R para automatizar estatísticas descritivas. **Sigmae**, v.3, p.36-42, 2014.
- FERREIRA, E.B.; CAVALCANTI, P.P.; NOGUEIRA, D.A. **ExpDes**: experimental designs package. R package version 1.1.2. 2013. Available at: <<http://CRAN.R-project.org/package=ExpDes>>. Accessed on: Nov. 8 2016.
- FRUTOS, E.; PURIFICACIÓN GALINDO, M.; LEIVA, V. An interactive biplot implementation in R for modeling genotype-by-environment interaction. **Stochastic Environmental Research and Risk Assessment**, v.28, p.1629-1641, 2014. DOI: 10.1007/s00477-013-0821-z.
- HOTHORN, T.; BRETZ, F.; WESTFALL, P. Simultaneous inference in general parametric models. **Biometrical Journal**, v.50, p.346-363, 2008. DOI: 10.1002/bimj.200810425.
- JELIHOVSCHI, E.G.; FARIA, J.C.; ALLAMAN, I.B. ScottKnott: A package for performing the Scott-Knott clustering algorithm in R. **Trends in Applied and Computational Mathematics**, v.15, p.3-17, 2014.
- LENTH, R.V. Least-squares means: The R package lsmeans. **Journal of Statistical Software**, v.69, p.1-33, 2016. DOI: 10.18637/jss.v069.i01.
- MENDIBURU, F. de. **agricolae**: Statistical procedures for agricultural research. R package version 1.2-4. 2016. Available at: <<http://CRAN.R-project.org/package=agricolae>>. Accessed on: Nov. 5 2016.
- PETERNELLI, L.A.; MELLO, M.P. **Conhecendo o R**: uma visão estatística. Viçosa: Ed. da UFV, 2011. 185p.
- R CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <<http://www.R-project.org/>>. Accessed on: May 27 2017.
- REGAZZI, A.J. Teste para verificar a identidade de modelos de regressão e a igualdade de parâmetros no caso de dados de delineamentos experimentais. **Revista Ceres**, v.46, p.383-409, 1999.
- REGAZZI, A.J. Teste para verificar a identidade de modelos de regressão e a igualdade de alguns parâmetros num modelo polinomial ortogonal. **Revista Ceres**, v.40, p.176-195, 1993.
- REGAZZI, A.J. Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear. **Revista Ceres**, v.50, p.9-26, 2003.

REGAZZI, A.J.; SILVA, C.H.O. Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear. I. Dados no delineamento inteiramente casualizado. **Revista de Matemática e Estatística**, v.22, p.33-45, 2004.

VAZQUEZ, A.I.; BATES, D.; ROSA, G.J.M.; GIANOLA, D.; WEIGEL, K.A. Technical note: an R package for fitting generalized linear mixed models in animal breeding. **Journal of Animal Science**, v.88, p.497-504, 2010. DOI: 10.2527/jas.2009-1952.

Received on December 12, 2016 and accepted on October 16, 2017