Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Interplot: Plot the Effects of Variables in Interaction Terms

Achim Zeileis
Journal of Statistical Software

Second Author
Plus Affiliation

Abstract

This short article illustrates how to write a manuscript for the *Journal of Statistical Software* (JSS) using its LaTeX style files. Generally, we ask to follow JSS's style guide and FAQs precisely. Also, it is recommended to keep the LaTeX code as simple as possible, i.e., avoid inclusion of packages/commands that are not necessary. For outlining the typical structure of a JSS article some brief text snippets are employed that have been inspired by **?**, discussing count data regression in R. Editorial comments and instructions are marked by vertical bars.

Keywords: JSS, style guide, comma-separated, not capitalized, R.

1. Introduction: Count data regression in R

Interaction is a powerful tool to test conditional effects of one variable on the contribution of another variable to the dependent variable and has been extensively applied in the empirical research of social science since the 1970s (Wright). Unfortunately, the nonlinear nature determines that the statistical estimate of an interactive effect cannot be interpreted as straightforward as the coefficient of a regular regression parameter. Let's use a simple example to illustrate this point: The following model use an interaction term to test the conditional effect of Z on X's contribution (or the conditional effect of X on Z's contribution) to the variance of Y.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \times Z + \varepsilon.$$

The contribution of X on Y is typically represented by the marginal effect:

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z.$$

The standard error of this estimate is thus:

$$\hat{\sigma}_{\frac{\partial Y}{\partial X}} = \sqrt{var(\hat{\beta}_1) + Z^2 var(\hat{\beta}_3) + 2Z cov(\hat{\beta}_1, \hat{\beta}_3)}.$$

As Brambor, Clark, and Golder (2006, 70) indicates, the above equation suggests that its perfectly possible for the contribution of X on Y to be statistically significant for certain values of Z "even if all of the model parameters are insignificant." In other words, one cannot infer whether X has a meaningful conditional effect on Y simply from the magnitude and significance of either 1 or 3 (Ibid., 74). Instead, the conditional effect should be examined based on the marginal effect at every observed value of Z (Berry, Golder, and Milton 2012; Brambor, Clark, and Golder 2006; Braumoeller).

Furthermore, scholars have noticed that the point estimations of some interaction models, especially those depending on link functions (e.g., logit and probit) may not be as reliable as the estimates of the confidence intervals (Berry, DeMeritt, and Esarey 2016). Therefore, it is even more important to accurately present and examine the boundaries established by the confidence intervals than the point estimates in the test of conditional effect. (Berry *et al.* 2012) also pointed out that the substantive significance of the conditional effect highly relates to the distribution of the conditioning variable (viz., Z in the above example). They thus recommend researchers to present the frequency distribution of the conditioning variable together with the marginal effects (especially when the effect trend goes across the zero point).

More recently, (Esarey and Sumner) uncovered that the estimation of the marginal effects suggested by (Brambor *et al.* 2006) might cause a "multiple comparison problem" and result over- or underconfidence of the confidential intervals. They thus recommended to adjust the CIs with a critical t-statistics following the Benjamini1995 procedure.

The `interplot` package provides a convenient way to operate and visualize above points with one or a series of plots produced by a single function. The function visualizes the changes in the coefficient of one variable in a two-way interaction term conditional on the value of the other included variable. The plot also includes simulated 95% confidential intervals of these coefficients. In the current version, the function works with ordinary linear regression models, generalized linear models (e.g., logit, probit, ordered logit, etc.), and multilevel (mixed-effects) regressions, all with complete or multiply imputed data.

Comparing to established alternatives such as `effects::plot` and `sjplot::sjp.int`, `interplot` provides a more user-friendly way to quickly produce plots that are easy to interpret. `interplot` plots the changes in the *conditional coefficient* of one variable in the interaction, rather than changes in the dependent variable itself as in the aforementioned functions. This approach avoids displaying interaction effects across multiple panels or multiple lines in favor of a single plot containing all the relevant information. Moreover, by outputting `ggplot` objects, `interplot` allows users to easily further customize their graphs.

2. Installation

To install:

- the latest released version: `install.packages("interplot")`.
- the latest developing version: `devtools::install_github("sammo3182/interplot")`.

3. Basic Use

3.1. Run a Model

This example is based on the `mtcars` dataset, which is drawn from the 1974 volume of the US magazine *Motor Trend*. The dependent variable is mileage in miles per (US) gallon (`mpg`), and the independent variables are the number of engine cylinders (`cyl`) and automobile weight in thousands of pounds (`wt`).

```
data(mtcars)  #load the data
```

Suppose we are interested in how automobile weight affects the relationship between of the number of engine cylinders on mileage and how the number of cylinders affects the relationship between the car's weight and its mileage. Such conditional effects are modeled using a two-way multiplicative interaction term:

```
m_cyl <- lm(mpg ~ wt * cyl, data = mtcars)
summary(m_cyl)
```

```
Call:
lm(formula = mpg ~ wt * cyl, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2288 -1.3495 -0.5042  1.4647  5.2344 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  54.3068     6.1275   8.863 1.29e-09 ***
wt           -8.6556     2.3201  -3.731 0.000861 ***
cyl          -3.8032     1.0050  -3.784 0.000747 ***
wt:cyl        0.8084     0.3273   2.470 0.019882 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.368 on 28 degrees of freedom
Multiple R-squared:  0.8606,    Adjusted R-squared:  0.8457 
F-statistic: 57.62 on 3 and 28 DF,  p-value: 4.231e-12
```

The basic Poisson regression model for count data is a special case of the GLM framework **?**. It describes the dependence of a count response variable y_i ($i = 1, \dots, n$) by assuming a Poisson distribution $y_i \sim \mathrm{Pois}(\mu_i)$. The dependence of the conditional mean $E[y_i \,|\, x_i] = \mu_i$ on the regressors x_i is then specified via a log link and a linear predictor

$$\log(\mu_i) \quad = \quad x_i^\top \beta, \tag{1}$$

where the regression coefficients β are estimated by maximum likelihood (ML) using the iterative weighted least squares (IWLS) algorithm.

> TODO: Note that around the equation above there should be no spaces (avoided in the LaTeX code by `%` lines) so that "normal" spacing is used and not a new paragraph started.

R provides a very flexible implementation of the general GLM framework in the function `glm()` **?** in the **stats** package. Its most important arguments are

```
glm(formula, data, subset, na.action, weights, offset,
  family = gaussian, start = NULL, control = glm.control(…),
  model = TRUE, y = TRUE, x = FALSE, …)
```

where `formula` plus `data` is the now standard way of specifying regression relationships in R/S introduced in **?**. The remaining arguments in the first line (`subset`, `na.action`, `weights`, and `offset`) are also standard for setting up formula-based regression models in R/S. The arguments in the second line control aspects specific to GLMs while the arguments in the last line specify which components are returned in the fitted model object (of class '`glm`' which inherits from '`lm`'). For further arguments to `glm()` (including alternative specifications of starting values) see `?glm`. For estimating a Poisson model `family = poisson` has to be specified.

> As the synopsis above is a code listing that is not meant to be executed, one can use either the dedicated `{Code}` environment or a simple `{verbatim}` environment for this. Again, spaces before and after should be avoided.
> Finally, there might be a reference to a `{table}` such as Table 1. Usually, these are placed at the top of the page (`[t!]`), centered (`\centering`), with a caption below the table, column headers and captions in sentence style, and if possible avoiding vertical lines.

Type	Distribution	Method	Description
GLM	Poisson	ML	Poisson regression: classical GLM, estimated by maximum likelihood (ML)
		Quasi	"Quasi-Poisson regression'': same mean function, estimated by quasi-ML (QML) or equivalently generalized estimating equations (GEE), inference adjustment via estimated dispersion parameter

Type	Distribution	Method	Description
		Adjusted	"Adjusted Poisson regression": same mean function, estimated by QML/GEE, inference adjustment via sandwich covariances
	NB	ML	NB regression: extended GLM, estimated by ML including additional shape parameter
Zero-augmented	Poisson	ML	Zero-inflated Poisson (ZIP), hurdle Poisson
	NB	ML	Zero-inflated NB (ZINB), hurdle NB

Table 1: Overview of various count regression models. The table is usually placed at the top of the page (`[t!]`), centered (`centering`), has a caption below the table, column headers and captions are in sentence style, and if possible vertical lines should be avoided.

4. Illustrations

For a simple illustration of basic Poisson and NB count regression the `quine` data from the **MASS** package is used. This provides the number of `Days` that children were absent from school in Australia in a particular year, along with several covariates that can be employed as regressors. The data can be loaded by

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure 1.

> For code input and output, the style files provide dedicated environments. Either the "agnostic" `{CodeInput}` and `{CodeOutput}` can be used or, equivalently, the environments `{Sinput}` and `{Soutput}` as produced by `Sweave()` or **knitr** when using the `render_sweave()` hook. Please make sure that all code is properly spaced, e.g., using `y = a + b * x` and *not* `y=a+b*x`. Moreover, code input should use "the usual" command prompt in the respective software system. For R code, the prompt `R>` should be used with `+` as the continuation prompt. Generally, comments within the code chunks should be avoided – and made in the regular LaTeX text instead. Finally, empty lines before and after code input/output should be avoided (see above).

As a first model for the `quine` data, we fit the basic Poisson regression model. (Note that JSS prefers when the second line of code is indented by two spaces.)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine, family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

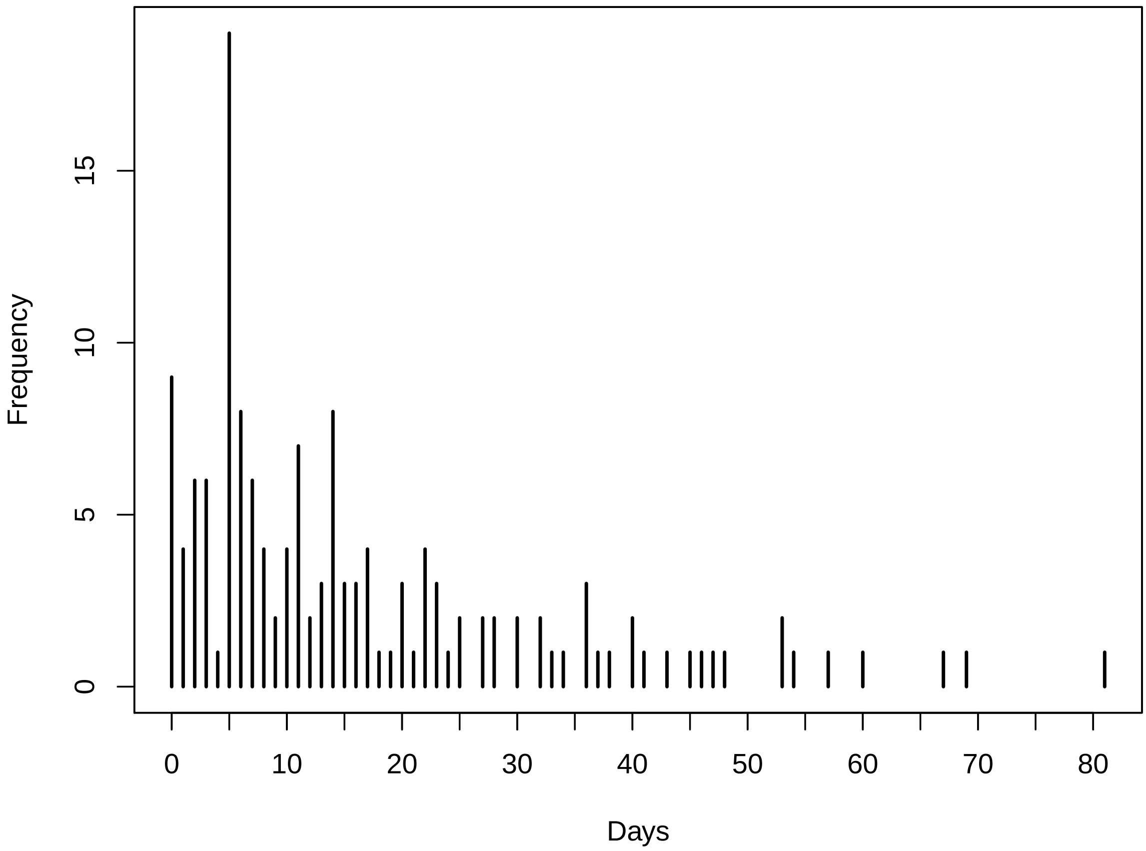

Figure 1: Frequency distribution for number of days absent from school.

In a comparison with the BIC the latter model is clearly preferred.

```
R> library("MASS")
R> BIC(m_pois, m_nbin)

       df      BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)


Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
EthN        -0.24591    0.39135  -0.628  0.52977
SexM        -0.77181    0.38021  -2.030  0.04236 *
AgeF1       -0.02546    0.41615  -0.061  0.95121
AgeF2       -0.54884    0.54393  -1.009  0.31296
AgeF3       -0.25735    0.40558  -0.635  0.52574
LrnSL        0.38919    0.48421   0.804  0.42153
EthN:SexM    0.36240    0.29430   1.231  0.21818
EthN:AgeF1  -0.70000    0.43646  -1.604  0.10876
EthN:AgeF2  -1.23283    0.42962  -2.870  0.00411 **
EthN:AgeF3   0.04721    0.44883   0.105  0.91622
EthN:LrnSL   0.06847    0.34040   0.201  0.84059
SexM:AgeF1   0.02257    0.47360   0.048  0.96198
SexM:AgeF2   1.55330    0.51325   3.026  0.00247 **
SexM:AgeF3   1.25227    0.45539   2.750  0.00596 **
SexM:LrnSL   0.07187    0.40805   0.176  0.86019
AgeF1:LrnSL -0.43101    0.47948  -0.899  0.36870
AgeF2:LrnSL  0.52074    0.48567   1.072  0.28363
AgeF3:LrnSL       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
Residual deviance: 167.53  on 128  degrees of freedom
```

```
AIC: 1100.5

Number of Fisher Scoring iterations: 1

              Theta:  1.604
          Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

5. Summary and discussion

As usual...

Computational details

If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R~3.4.1 with the **MASS**~7.3.47 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at [https://CRAN.R-project.org/].

Acknowledgments

All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Berry WD, DeMeritt JHR, Esarey J (2016). "Bias and overconfidence in parametric models of interactive processes." *American Journal of Political Science*, **60**, 521–539. Citation Key: BerryEtAl2016.

Berry WD, Golder M, Milton D (2012). "Improving tests of theories positing interaction." *The Journal of Politics*, **74**(3), 653–671. `doi:10.1017/S0022381612000199`. URL https://www.journals.uchicago.edu/doi/10.1017/S0022381612000199.

Brambor T, Clark WR, Golder M (2006). "Understanding interaction models: improving empirical analyses." *Political Analysis*, **14**(1), 63–82. doi:10.1093/pan/mpi014. URL https://www.cambridge.org/core/product/identifier/S1047198700001297/type/journal_article.

Braumoeller BF (????). "Hypothesis Testing and Multiplicative Interaction Terms." **58**(04). ISSN 0020-8183, 1531-5088. doi:10.1017/S0020818304040251. URL http://www.journals.cambridge.org/abstract_S0020818304040251.

Esarey J, Sumner JL (????). "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." **51**(9), 1144–1176. ISSN 0010-4140, 1552-3829. doi:10.1177/0010414017730080. URL http://journals.sagepub.com/doi/10.1177/0010414017730080.

Wright GC (????). "Linear Models for Evaluating Conditional Relationships." **20**(2), 349. ISSN 00925853. doi:10.2307/2110649. 2110649, URL https://www.jstor.org/stable/2110649?origin=crossref.

More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).
For more technical style details, please check out JSS's style FAQ at [https://www.jstatsoft.org/pages/view/style#frequently-asked-questions] which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details…

Using BibTeX

References need to be provided in a BibTeX file (`.bib`). All references should be made with `@cite` syntax. This commands yield different formats of author-year citations and allow to include additional details (e.g.,pages, chapters, …) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.
Cleaning up BibTeX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- item JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- item Titles should be in title case.
- item Journal titles should not be abbreviated and in title case.
- item DOIs should be included where available.
- item Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

Achim Zeileis
E-mail: Achim.Zeileis@R-project.org
URL: https://www.zeileis.org/

Second Author

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd