Novel graphical diagnostics for regression models

Peter Solymos (solymos@ualberta.ca) and Subhash Lele (slele@ualberta.ca)

September 15, 2017

The Problem

Regression models are at the core of any classical statistical or machine learning analysis and are widely used in diverse disciplines, including public health, business analytics, and environmetrics. "*R is a free software environment for statistical computing and graphics*" – thus ideally suited for regression modeling and visual model diagnostics. Although visualization is one of the most powerful interpretation tools, this capacity of R has not been fully realized for regression models.

For a continuous response and a single continuous covariate, readily implemented visualizations through the *scatter plots* with the regression fit overlaid are routinely obtained in different graphical frameworks (i.e. **lattice**, **ggplot2**). Such tools are effective for diagnosing the possible inadequacies in the form of the model: Is linear model adequate? Do we need non-linear terms? The scatter plot is also useful to convey the results of the analysis. It visually shows how the response changes with changes in the single covariate.

Multiple regression, however, poses substantial problems. Although it is relatively easy to fit a multiple regression model and conduct the covariate selection, the results are generally difficult to convey. This is because the regression coefficients can no longer be interpreted as 'change in the response as the covariate is changed by one unit'. The interpretation depends on 'keeping all other covariates fixed'. This interpretation becomes even more difficult to convey if interaction terms are present.

In practice, many times, users are interested in manipulating one of the covariates and understanding the 'average change in the response variable', averaged over the distribution of the other covariates. Such effects are sometimes termed as 'marginal effects'. Most statistical textbooks on regression analysis do not discuss these plots. The **visreg** package includes *conditional effects*, the **dismo** package implements the very similar *partial dependence* plots for **gbm**'s but not marginal effects plots.

In the case of multiple regression, the *added variable plot*, which is a single variable scatter plot using the *residuals*, helps us diagnose if any additional covariates, or 'lurking variables', might be useful. The **spatstat** packages has exploratory techniques like this for spatial point process data (function `lurking`). Although it is relatively easy to implement, such visual exploratory tools are largely neglected.

For binary and count data, the available tool-set for visual regression diagnostic is more limited even for the single variable case. Scatter plots, for example, are nearly useless for most practical situations. These problems hinder interactive visual data exploration and model diagnostics, and encourages over-reliance on automated algorithms without true insight into the mechanisms responsible for the patterns.

The Plan

- We will develop an R package that generalizes the marginal effects plots for continuous response variable to binary and count data regression models (e.g. GLMs, GAMs, CARTs), including zero inflated, multinomial, and ordinal versions.
- We will implement (1) exploratory functions to be used prior to modelling, (2) functions to test the adequacy of the functional form of fitted models, and (3) functions to judge if some important covariates are missing from the model.
- We will design the package with easy extensibility in mind facilitating maintenance with dependencies and also allowing other package maintainers to write their own methods.
- We will take care that the new code will be tidyverse-aware, i.e. will allow easy interfacing with pipe-based workflows, using `data.frame`s, and **dplyr**-style verbs.
- We will also support high level plotting (i.e. **lattice**, **ggplot2**, **plotly**) ready for interactive web applications.
- We will document the software and provide tutorials, reproducible data analysis examples, and interactive web application demos using Shiny and OpenCPU.
- We will publish the resulting products in an open access form, in the *R Journal* or *JSS* (added as vignette to the package), and also in a computational/statistical journal.

Timeline and Project Milestones

We have implemented a preliminary version of marginal effect plots in the **ResourceSelection** package (`mep`) that we have used in teaching a spatial data analysis course. We will use this code as prototype for the package.

- *Month 1–3*: working out design, programming the R package, writing documentation, testing with smaller data sets. – *Outcome*: Working R package in GitHub repository.
- *Month 4–6*: testing on larger data sets, develop test cases, write tutorials, develop teaching material, reproducible examples, and web applications. – *Outcome*: Additional material added to GitHub project, package submitted to CRAN.

Failure Modes

- We can't integrate the package's workflow with mainstream packages. – *Solution path*: ask the tidyverse community and package developers for help.
- We can't get advanced interactive web applications to work. – *Solution path*: ask the Shiny and OpenCPU community and package developers for help.

The Team

Dr. Peter Solymos is a statistical ecologist at the Alberta Biodiversity Monitoring Institute and Boreal Avian Modelling Project (Edmonton, Alberta, Canada), with a research focus on developing and applying computational techniques for big data sets to better inform biodiversity conservation and natural resource management over large spatial scales. He is developing new algorithms and methodologies for more efficient data-information-knowledge pipelines (see e.g. species.abmi.ca). He has authored >50 peer reviewed publications including ones in the *R Journal* or *JSS* and many R packages among them **mefa4** for sparse matrix based data wrangling, **dclone** for MCMC based hierarchical modeling, **pbapply**, **ResourceSelection**, and **vegan**. Peter will bring R programming, data analytics, and statistical skills to the project.

Dr. Subhash Lele is a professor of statistics at the University of Alberta in Canada. His research interests include theoretical and applied statistics. He has been working in statistics for almost 30 years. He is author of the **ResourceSelection** package, he has published >80 papers in major journals. He has co-authored a book on Morphometrics, co-edited a book on the topic of quantification of statistical evidence. He has been on the editorial board of the *Journal of the American Statistical Association*, *Ecology* and *Ecological Monographs*, *Ecological and Environmental Statistics*, *Journal of Wildlife Management* and *Journal of Animal Ecology*. He has also served on various committees of the US National Academy of Sciences. Subhash will bring theoretical statistical, writing, and programming skills to the project.

How The ISC Can Help

We will use most funding to develop the R package and related demo applications. Total costs will be **7,000 USD**:

- Programming, project communication, *Months 1–3*: USD 3,000.
- Programming, project communication, *Months 4–6*: USD 2,000.
- Travel costs to *UseR! 2018* to promote the package: USD 2,000.

Dissemination

We will regularly post blogs about the project on our personal website aggregated by R-Bloggers, use Twitter, post to relevant mailing lists, and use GitHub issues and pull requests for feedback. The project will live on GitHub. We will work under a permissive open source license, probably MIT. We will propote the project at the *UseR! 2018* conference. We will write R Consortium blogs at start and end of the project. We also plan publications geared towards the R user community and the general statistical audiences.