The Journal of Open Source Software

# gglm: An R package implementing the grammar of graphics for linear model diagnostic plots

**Grayson W. White**[1]

**1** Michigan State University, Department of Forestry

**DOI:**

**Software**

- Review
- Repository
- Archive

**Submitted:**
**Published:**

## Summary

gglm implements an interface to produce publication-ready model diagnostic plots that complies with the grammar of graphics (Wickham, 2010). Further, gglm utilizes the broom and broom.mixed R packages to provide support for diagnostic plots produced from a variety of model object classes across a wide variety of R packages (Bolker & Robinson, 2022; Robinson, Hayes, & Couch, 2023). A quartet of diagnostic plots can be quickly created using gglm's homonymous function, or plots can be created individually through instructive and intuitive layer functions added to a ggplot2 object (Wickham, 2016).

## Statement of Need

When scientists, statistical practitioners, students, and others implement statistical models, it is of the utmost importance that the modeling assumptions are verified through visual diagnostics in order to ensure valid statistical inference. The R statistical software language provides a method for producing diagnostic plots for linear model objects created with `stats::lm`, however these plots are visually unappealing, inconsistent with diagnostic plots across other R packages and model types, and out of place in modern statistics and data science courses focused on learning R with the `tidyverse` (Wickham et al., 2019).

`gglm` addresses the described issues with current diagnostic plots in R by providing a consistent interface for producing beautiful and publication-ready diagnostic plots across a large variety of R packages and model types (linear models, linear mixed models, generalized linear mixed models, etc.). `gglm` provides functionality to quickly produce four common diagnostic plots, similar to `stats::plot.lm`, but produced by `ggplot2`. Further, `gglm` provides a suite of layer functions adhering to the grammar of graphics which allow the user to create and fine-tune their diagnostic plots through `ggplot2`'s intuitive interface. The layer functions are particularly applicable in modern courses teaching linear regression where students have already learned `ggplot2`. For example, `gglm` and its layer functions are used in Harvard University's introductory statistics course (McConville, 2023). Outside of educational benefits, `gglm` has potential to allow researchers to more easily publish elegant diagnostic plots. `gglm` has been downloaded from CRAN over 23,000 times as of January 2024.

## Usage and Features

`gglm` achieves a balance in functionality by being both as easy to use as the built-in `stats::plot.lm` method, yet still highly intuitive and customizable for the curious user. `gglm` is designed with these traits in mind due to the understanding that an individual producing a diagnostic plot will most likely be in one of two camps: 1) the individual who wants an *easy* to use tool that allows them to quickly check their model diagnostics, or 2) the individual who wants an *intuitive and customizable* tool that allows them to look closely at their diagnostics for the purposes of education, fine-tuning graphics for publication, or other reasons. `gglm` satisfies the members of both camps.

The `gglm::gglm` function is made for folks in the first camp who are looking for a more aesthetically pleasing alternative to `stats::plot.lm`. In practice, the process of using `gglm::gglm` is as simple as and more general than using `stats::plot.lm`, with steps as follows:

- fit a model of any class listed in `gglm::list_model_classes`,
- call `gglm::gglm` on the saved model object.

The `gglm::stat_*` functions are thus for those in the second camp. `gglm` provides seven functions of this sort, including those that produce the following plots: Cook's distance by leverage, Cook's distance by observation number, fitted values by residual values, normal QQ, residual histogram, residual values by leverage, and scale by location. The steps to produce a diagnostic plot with these functions are more fluid than with `gglm::gglm`, but are easy to understand provided the user has an understanding of how to use `ggplot2`. One may use the workflow:

- fit a model of any class listed in `gglm::list_model_classes`,
- provide the saved model object as data to `ggplot2::ggplot`,
- add their intended diagnostic plot layer,
- add any more `ggplot2` layers such as themes, labels, annotations, and more to create their custom diagnostic plot.

## Comparison to Other Packages

Functionality similar to that of `gglm`'s is provided by a variety of R packages. As mentioned throughout, `stats` provides a `plot` method for producing diagnostic plots for `lm` objects with base R graphics (R Core Team, 2023). Further, `lindia` produces diagnostic plots for `lm` objects with `ggplot2` graphics, but does not include functions that adhere with the grammar of graphics (Lee & Ventura, 2023). Finally, many packages provide methods for plotting diagnostics based on their own model classes (see, e.g. `lme4::plot.merMod`), however these methods are do not have consistent usage across packages (Bates, Mächler, Bolker, & Walker, 2015). `gglm` hence addresses a significant gap in functionality by creating a consistent framework for producing diagnostic plots across R packages and model types while adhering to the grammar of graphics.

## References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bolker, B., & Robinson, D. (2022). *Broom.mixed: Tidying methods for mixed models*. Retrieved from https://CRAN.R-project.org/package=broom.mixed

Lee, Y. Y., & Ventura, S. (2023). *Lindia: Automated linear regression diagnostic*. Retrieved from https://CRAN.R-project.org/package=lindia

McConville, K. (2023). STAT 100: Introduction to statistics and data science. Harvard University Department of Statistics. Retrieved from https://mcconvil.github.io/stat100f23/

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robinson, D., Hayes, A., & Couch, S. (2023). *Broom: Convert statistical objects into tidy tibbles*. Retrieved from https://CRAN.R-project.org/package=broom

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3–28. doi:10.1198/jcgs.2009.07098

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686