

1 **Automated Residual Plot Assessment with the R Package**
2 **autovi and the Shiny App autovi.web**

3 Weihao Li¹, Dianne Cook¹, Emi Tanaka², Susan VanderPlas³ and Klaus
4 Ackermann¹

5 *Monash University, The Australian National University and University of*
6 *Nebraska*

Summary

7 Visual assessment of residual plots is a common approach for diagnosing linear
models, but it relies on manual evaluation, which does not scale well and can lead
to inconsistent decisions across analysts. The lineup protocol, which embeds the
observed plot among null plots, can reduce subjectivity but requires even more
human effort. In today's data-driven world, such tasks are well-suited for automation.
We present a new R package that uses a computer vision model to automate the
evaluation of residual plots. An accompanying Shiny app is provided for ease of use.
Given a sample of residuals, the model predicts a visual signal strength (VSS) and
offers supporting information to help analysts assess the adequacy of their model
fit.

8 *Key words:* initial data analysis; statistical graphics; data visualization; visual inference;
computer vision; machine learning; hypothesis testing; regression analysis;
model diagnostics

9 1. Introduction

10 Regression analysis is a widely used statistical modeling technique for data in many
11 fields. There is a vast array of software for conducting regression modeling and
12 generating diagnostics. The package `lmtest` (Zeileis & Hothorn 2002) provides a suite
13 of conventional tests, while the `stats` package (R Core Team 2022) offers standard
14 diagnostic plots such as residuals vs. fitted values, quantile-quantile (Q-Q) plots,
15 and residuals vs. leverage plots. Additional packages like `jtools` (Long 2022), `olsrr`

¹ Department of Econometrics and Business Statistics, Monash University, Wellington Road, VIC 3800, Australia

² Biological Data Science Institute, The Australian National University, 46 Sullivan's Creek Road, ACT 2600, Australia

³ Department of Statistics, University of Nebraska, Hardin Hall, 3310 Holdrege St Suite 340, Lincoln, NE 68583, United States

Email: `weihao.li@monash.edu`

(Hebbali 2024), `rockchalk` (Johnson 2022), and `ggResidpanel` (Goode & Rey 2019) deliver similar graphical diagnostics, often with enhanced aesthetics or interactive features. These tools collectively produce the core diagnostic plots outlined in the classical text by Cook & Weisberg (1982). The `ecostats` package (Warton 2023) extends these diagnostics by incorporating simulation envelopes into residual plots. Meanwhile, `DHARMa` (Hartig 2022) compares empirical quantiles (0.25, 0.5, and 0.75) of scaled residuals to their theoretical counterparts, with a strong focus on identifying model violations such as heteroscedasticity, misspecified functional forms, and issues specific to generalized linear and mixed-effect models, like over/under-dispersion. It also provides conventional test annotations to reduce the risk of misinterpretation.

However, relying solely on subjective assessments of these plots can lead to issues such as over-interpreting random patterns as model violations. Li et al. (2024a) demonstrated that visual inference methods, particularly those using the lineup protocol (Buja et al. 2009), offer more practical and reliable assessments of residual patterns than conventional tests, as they are less sensitive to minor departures. Packages such as `nullabor` (Wickham et al. 2020), `HLMdiag` (Loy & Hofmann 2014), and `regressinator` (Reinhart 2024) support this approach by enabling users to compare observed residual plots with plots generated under null hypothesis, thereby helping to quantify the significance of any detected patterns.

As noted in Li et al. (2024b), the lineup protocol has significant limitations in large-scale applications due to its reliance on human labor. To overcome this constraint, a computer vision model was developed alongside a corresponding statistical testing procedure to automate the assessment of residual plots. The model takes as input a residual plot and a set of auxiliary variables (such as the number of observations) and outputs a predicted visual signal strength (VSS). This VSS estimates the degree of deviation between the residual distribution of the fitted model and the reference distribution expected under correct model specification.

To make the statistical testing procedure and trained computer vision model widely accessible, we developed the R package `autovi` along with a companion web interface, `autovi.web`, which allows users to automatically assess their residual plots using the trained computer vision model.

The remainder of this paper is structured as follows: Section 2 introduces the definition and computation of visual signal strength. Section 3 provides a detailed documentation of the `autovi` package, including its usage and infrastructure. Section 4 focuses on the

50 `autovi.web` interface, describing its design and usage, along with illustrative examples.
 51 Finally, Section 5 presents the main conclusions of this work.

52 2. Definition and computation of visual signal strength

53 To train a computer vision model, a measure of the visible pattern in a plot is needed.
 54 We call this the **visual signal strength** (VSS), which measures how prominently a
 55 specific set of visual patterns appears in an image. This can be computed for a training
 56 set of data, and plots, where the generating distributions are specified.
 57 In the context of regression model diagnostics, VSS describes the clarity of visual
 58 patterns on a diagnostic plot that may indicate model violations. Violations can be
 59 categorized as weak, moderate, or strong, but here we treat it as a continuous positive
 60 real variable. Importantly, its interpretation depends on how it is linked to a function
 61 of the data or the underlying data generating process. Consequently, the calculation
 62 of VSS can vary across different model classes or within the same model, depending
 63 on the generating function.
 64 VSS is an estimate of the distance between the residual distribution of a fitted classical
 65 normal linear regression model and a reference distribution; more details can be found
 66 in Li et al. (2024b). The distance measure is based on the Kullback-Leibler (KL)
 67 divergence:

$$D = \log(1 + D_{KL}),$$

68 where D_{KL} is given by:

$$D_{KL} = \int_{\mathbb{R}^n} \log \frac{p(\mathbf{e})}{q(\mathbf{e})} p(\mathbf{e}) d\mathbf{e}, \quad (1)$$

69 here, $p(\cdot)$ and $q(\cdot)$ are the probability density functions of the reference residual
 70 distribution P and the true residual distribution Q , respectively.
 71 This distance measure depends on knowledge of the true residual distribution, which
 72 is unknown in practice. To compute D_{KL} for the training samples, Equation 1 takes
 73 different forms depending on the specific model violations. For instance, where necessary
 74 higher-order predictors, \mathbf{Z} , and their corresponding parameter, β_Z , are omitted from
 75 the fitted linear model, the distance measure can be expanded as follows:

$$D_{KL} = \frac{1}{2} (\boldsymbol{\mu}_z^\top (\text{diag}(\mathbf{R}\sigma^2))^{-1} \boldsymbol{\mu}_z),$$

76 where $\boldsymbol{\mu}_z = \mathbf{R}\mathbf{Z}\boldsymbol{\beta}_z$, $\mathbf{R} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ and \mathbf{X} is the design matrix of the
77 regression model.

78 The computer vision model approximates this mapping from a set of residuals to its
79 corresponding distance measure. It is trained on a large number of synthetic regression
80 models, where the data-generating process is known, allowing the distance measure
81 to be explicitly calculated. The model takes a residual plot as input and outputs the
82 corresponding distance measure. Additional details are provided in [Li et al. \(2024b\)](#).

83 3. R package: autovi

84 The main purpose of **autovi** is to provide rejection decisions and p -values for testing
85 the null hypothesis (H_0) that the regression model is correctly specified. The package
86 provides automated interpretation of residual plots using computer vision. The name
87 **autovi** stands for **a**utomated **v**isual **i**nference. This functionality can be accessed
88 through the R package **autovi**, or through a web interface, **autovi.web**, which enables
89 users to perform analyses without installing R, Python, or their associated dependencies
90 locally.

91 3.1. Motivation

92 Figure 1 shows three sets of plots of residuals against fitted values. The simulated
93 example in (a) might be interpreted as a heteroscedastic pattern, however the
94 automated reading would predict this to have a visual signal strength (VSS) of
95 1.53, with a corresponding p -value of 0.25. This means it would be interpreted as
96 a good residual plot, that there is nothing in the data to indicate a violation of
97 model assumptions. Skewness in the predictor variables is generating the apparent
98 heteroscedasticity, where the smaller variance in residuals at larger fitted values is
99 due to smaller sample size only. The Breusch-Pagan test ([Breusch & Pagan 1979](#)) for
100 heteroscedasticity would also not reject this as good residual plot.

101 The data in (b) is generated by fitting a linear model predicting `mpg` based on `hp`
102 using the `datasets::mtcars`. It is a small data set, and there is a hint of nonlinear
103 structure not captured by the model. The automated plot reading would predict a
104 VSS of 3.57, which has a p -value less than 0.05. That is, the nonlinear structure is
105 most likely real, and indicates a problem with the model. The conventional test, a

106 Ramsey Regression Equation Specification Error Test (RESET) (Ramsey 1969) would
 107 also strongly detect the nonlinearity.

108 The third example is generated using the `surreal` package (Balamuta 2024) where
 109 structured residuals are hidden in data, to be revealed if the correct model is specified.
 110 Here a quote based on Tukey is used as the residual structure “visual summaries focus
 111 on unexpected values”. The automated plot reading predicts the VSS to be 5.87, with
 112 a p -value less than 0.05. This structure is blindingly obvious visually, but a RESET
 113 test for nonlinear structure would not report a problem. (It would be detected by
 114 a Breusch-Pagan for heteroscedasticity and also Shapiro-Wilk test (Shapiro & Wilk
 115 1965) for non-normality.)

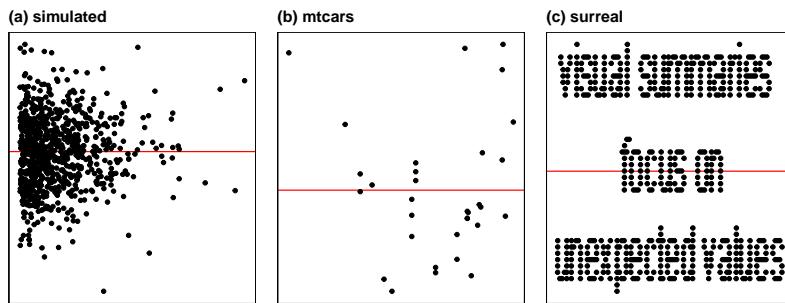


Figure 1. Reading residual plots can be a difficult task, particularly for students new to statistical modeling. The `autovi` package makes it easier. Here are three examples of residual plots, which may appear to have structure. According to `autovi`, the visual signal strengths (VSS) of these three examples are approximately (a) 1.53, (b) 3.57, (c) 5.87, resulting in (b), (c) being significant violations of good residuals, but (a) is consistent with a good residual plot.

116 3.2. Implementation

117 The `autovi` package is built on the `bandicoot` object-oriented programming (OOP)
 118 system (Li 2024), marking a departure from R’s traditional S3 generic system. This
 119 OOP architecture enhances flexibility and modularity, allowing users to redefine key
 120 functions through method overriding.

121 The `autovi` infrastructure effectively integrates multiple programming languages and
 122 libraries into a comprehensive analytical tool. It relies on five core libraries from
 123 Python and R, each playing a critical role in the analysis pipeline. In Python, `pillow`
 124 (Clark et al. 2015) handles image processing tasks such as reading and resizing PNG
 125 files of residual plots, then converting them into input tensors for further analysis.

126 TensorFlow (Abadi et al. 2016), a key component of modern machine learning, is used
 127 to predict the VSS of these plots using a pre-trained convolutional neural network.

128 In the R environment, `autovi` utilizes several libraries. `ggplot2` (Wickham 2016)
 129 generates the initial residual plots, saved as PNG files for visual input. `cassowaryr`
 130 (Mason et al. 2022) computes scagnostics (scatter plot diagnostics), providing numerical
 131 features that capture statistical properties of the plots. These scagnostics complement
 132 the visual analysis by offering quantitative metrics as secondary input to the
 133 computer vision model. `reticulate` (Ushey, Allaire & Tang 2024) enables seamless
 134 communication between R and Python.

135 3.3. Installation

136 The `autovi` package is available on CRAN. It is actively developed and maintained,
 137 with the latest updates accessible on GitHub. This paper uses `autovi` version 0.4.2.
 138 The package includes internal functions to check the current Python environment used
 139 by the `reticulate` package. If the necessary Python packages are not installed in the
 140 Python interpreter, an error will be raised. If you want to select a specific Python
 141 environment, you can do so by calling the `reticulate::use_python()` function before
 142 using the `autovi` package.

143 We recommend using the Shiny app `autovi.web` if users encounter installation
 144 problems.

145 3.4. Usage

146 3.4.1. Numerical summary

147 Three steps are needed to get an automated assessment of a set of residuals and fitted
 148 values:

- 149 1. Load the `autovi` package using the `library()` function.
- 150 2. Create a checker object with a linear regression model.
- 151 3. Call the `check()` method of the checker, which, by default, predicts the VSS for
 152 the true residual plot, 100 null plots, and 100 bootstrapped plots. The method
 153 stores the predictions internally and prints a concise results report.

154 The code to do this is:

```
library(autovi)
checker <- residual_checker(lm(dist ~ speed, data = cars))
checker$check()
```

155 It produces the following summary:

156

```
157 -- <AUTO_VI object>
158 Status:
159 - Fitted model: lm
160 - Keras model: (None, 32, 32, 3) + (None, 5) -> (None, 1)
161     - Output node index: 1
162 - Result:
163     - Observed visual signal strength: 3.16 (p-value = 0.0396)
164     - Null visual signal strength: [100 draws]
165         - Mean: 1.274
166         - Quantiles:
167
168             25%   50%   75%   80%   90%   95%   99%
169             0.802 1.111 1.575 1.666 1.919 2.657 3.348
170
171     - Bootstrapped visual signal strength: [100 draws]
172         - Mean: 2.795 (p-value = 0.05941)
173         - Quantiles:
174
175             25%   50%   75%   80%   90%   95%   99%
176             2.455 2.941 3.177 3.300 3.474 3.537 3.668
177
178     - Likelihood ratio: 0.7333 (boot) / 0.06284 (null) = 11.67
```

179 The summary includes observed VSS of the true residual plot and associated *p*-value
 180 of the automated visual test. The *p*-value is the proportion of null plots (out of the
 181 total 100) that have VSS greater than or equal to that of the true residual plot. The
 182 report also provides sample quantiles of VSS for null samples and bootstrapped data
 183 plots, providing more information about the sampling variability and a likelihood of
 184 model violations. The likelihood is computed from the proportion of values greater

185 than the observed VSS in both the bootstrapped data values and the simulated null
 186 values.

187 **3.4.2. Visual summary**

188 Users can visually inspect the original residual plot alongside a sample null plot using
 189 `plot_pair()` or a lineup of null plot `plot_lineup()`. This visual comparison can
 190 clarify why H_0 is either rejected or not, and help identify potential remedies.

```
checker$plot_pair()
```

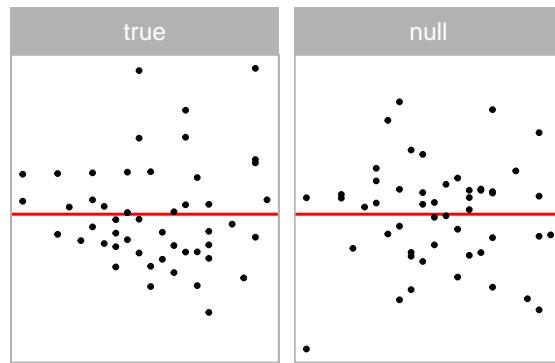


Figure 2. True plot alongside one null plot, for quick comparison.

191 The `plot_pair()` method (Figure 2) displays the true residual plot on the left and a
 192 single null plot on the right. If a full lineup was shown, the true residual plot would
 193 be embedded in a page of null plots. Users should look for any distinct visual patterns
 194 in the true residual plot that are absent in the null plot. Running these functions
 195 multiple times can help any visual suspicions, as each execution generates new random
 196 null plots for comparison.

197 The package offers a straightforward visualization of the assessment result through
 198 the `summary_plot()` function.

```
checker$summary_plot()
```

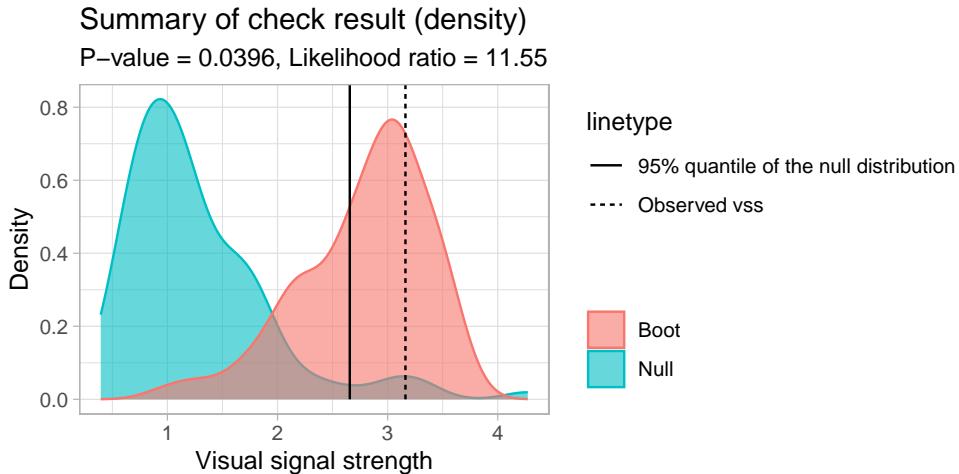


Figure 3. Summary plot comparing the densities of VSS for bootstrapped residual samples (red) relative to VSS for null plots (blue).

- 199 In the result, shown in Figure 3, the blue area represents the density of VSS for null
 200 residual plots, while the red area shows the density for bootstrapped residual plots.
 201 The dashed line indicates the VSS of the true residual plot, and the solid line marks
 202 the critical value at a 95% significance level. The p -value and the likelihood ratio are
 203 displayed in the subtitle. The likelihood ratio represents the ratio of the likelihood
 204 of observing the VSS of the true residual plot from the bootstrapped distribution
 205 compared to the null distribution.
- 206 Interpreting the plot involves several key aspects. If the dashed line falls to the right of
 207 the solid line, it suggests rejecting the null hypothesis. The degree of overlap between
 208 the red and blue areas indicates similarity between the true residual plot and null
 209 plots; greater overlap suggests more similarity. Lastly, the portion of the red area to
 210 the right of the solid line represents the percentage of bootstrapped models considered
 211 to have model violations.
- 212 This visual summary provides an intuitive way to assess the model's fit and potential
 213 violations, allowing users to quickly grasp the results of the automated analysis.

214 3.5. Modularized infrastructure

- 215 The initial motivation for developing `autovi` was to create a convenient interface for
 216 sharing the models described and trained in Li et al. (2024b). However, recognizing
 217 that the classical normal linear regression model represents a restricted class of

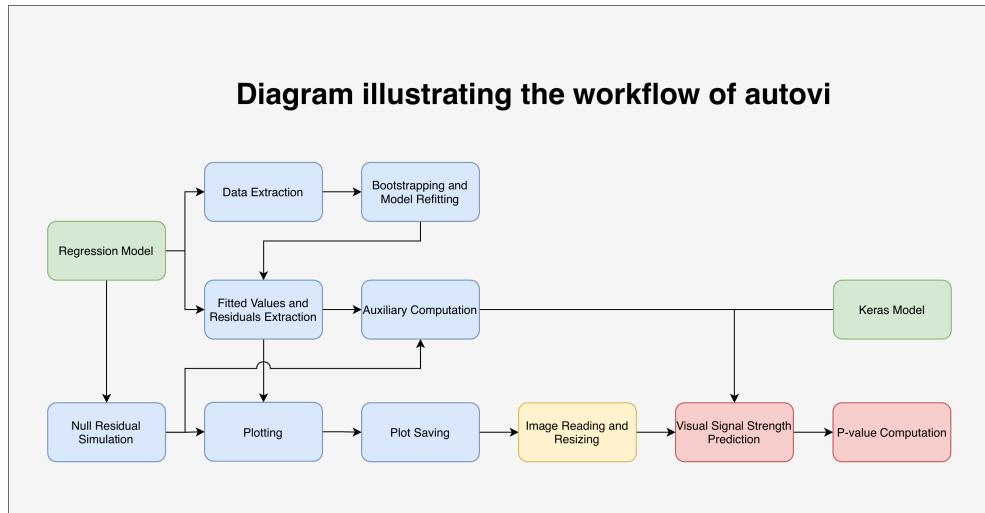


Figure 4. Diagram illustrating the infrastructure of the R package **autovi**. The modules in green are primary inputs provided by users. Modules in blue are overridable methods that can be modified to accommodate users' specific needs. The module in yellow is a pre-defined non-overridable method. The modules in red are primary outputs of the package.

models, we sought to avoid limiting the potential for future extensions, whether by the original developers or other developers. As a result, the package was designed to function seamlessly with linear regression models with minimal modification and few required arguments, while also accommodating other classes of models through partial infrastructure substitution. This modular and customizable design allows **autovi** to handle a wide range of residual diagnostics tasks.

The infrastructure of **autovi** consists of ten core modules: data extraction, bootstrapping and model refitting, fitted values and residuals extraction, auxiliary computation, null residual simulation, plotting, plot saving, image reading and resizing, VSS prediction, and *p*-value computation. Each module is designed with minimal dependency on the preceding modules, allowing users to customize parts of the infrastructure without affecting its overall integrity. An overview of this infrastructure is illustrated in Figure 4.

The modules for VSS prediction and *p*-value computation are predefined and cannot be overridden, although users can interact with them directly through function arguments. Similarly, the image reading and resizing module is fixed but will adapt to different Keras models by checking their input shapes. The remaining seven modules are designed to be overridable, enabling users to tailor the infrastructure to their specific needs. These modules are discussed in detail in the package documentation.

237 **3.6. Extension to Other Model Classes**

238 The `autovi` R package can be extended to accommodate other classes of models
 239 beyond linear regression, such as generalized linear models (`glm`). This is achieved by
 240 substituting the relevant overridable modules, as illustrated in Figure 4.

241 We provide an example of defining a new checker class tailored for Poisson regression
 242 using the `glm` framework:

- 243 1. Define a new class using `new_class()` with `AUTO_VI` as the parent class.
 244 2. Override the necessary methods using `register_method()`. In this example, we
 245 use Pearson residuals. To simulate null residuals, we assume the fitted model
 246 is correct and the estimated coefficients are accurate. New response values are
 247 generated accordingly, and a new model is fitted to this simulated response. Null
 248 residuals are then extracted from this refitted model.
 249 3. Create an alias for the `instantiate()` method of the new class.

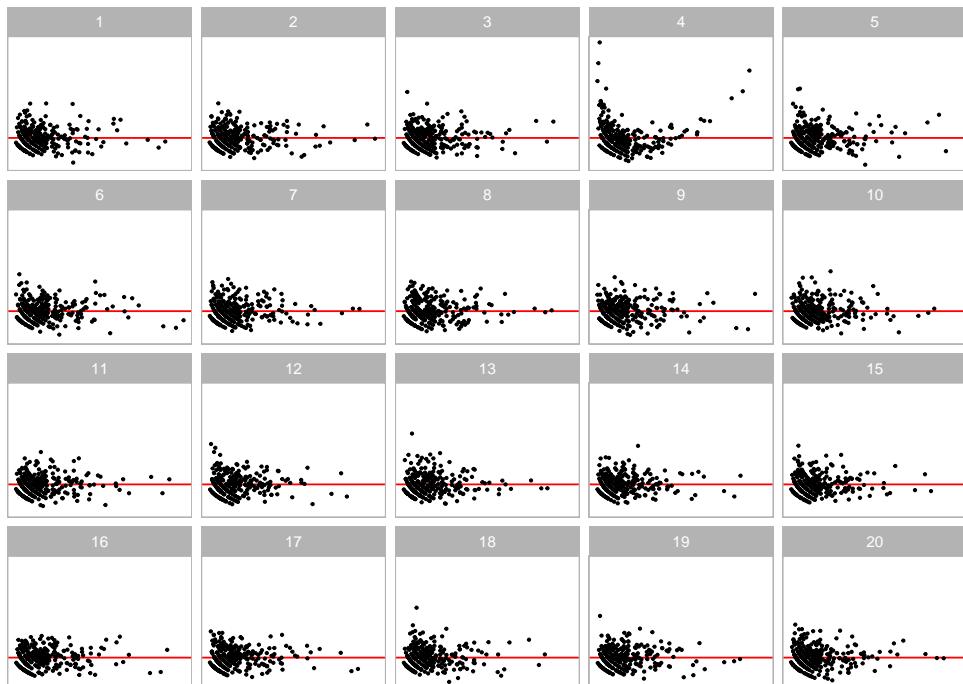
```
AUTO_POIS_VI <- new_class(AUTO_VI, class_name = "AUTO_POIS_VI")
register_method(
  AUTO_POIS_VI,
  get_fitted_and_resid = function(fitted_model = self$fitted_model) {
    tibble(.fitted = fitted(fitted_model),
           .resid = resid(fitted_model, type = "pearson"))
  },
  null_method = function(fitted_model = self$fitted_model) {
    dat <- model.frame(fitted_model)
    dat[[1]] <- rpois(nrow(dat), lambda = fitted(fitted_model))
    new_mod <- update(fitted_model, data = dat)
    return(self$get_fitted_and_resid(new_mod))
  }
)
auto_pois_vi <- AUTO_POIS_VI$instantiate
```

250 The resulting checker class can be employed analogously to the linear model case
 251 described in Section 3.4. For illustration, we fit a Poisson model in which the quadratic
 252 term of the predictor x is intentionally omitted. This misspecification manifests as a
 253 pronounced U-shaped pattern in the lineup display, which is also successfully identified
 254 by the computer vision model, yielding a p-value substantially below the conventional
 255 threshold of 0.05.

It is important to note, however, that the pre-trained computer vision model was developed specifically for diagnostic evaluation of linear regression. Its applicability to other model classes relies on the assumption that the null residual plots exhibit characteristics broadly consistent with those of well-behaved linear regression residuals, that is, residuals should be approximately randomly scattered around zero, display roughly constant variance across the range of fitted values, and exhibit no discernible structure or curvature. If these conditions are not met, or if model violations do not give rise to visually detectable patterns, the validity of the automated diagnostics may be compromised. For further discussion on extending the methodology to other model classes, refer to Li et al. (2024b).

```
x <- rnorm(300, sd = 0.5)
y <- rpois(300, lambda = exp(1 + x + x^2))
pois_checker <- auto_pois_viz(
  glm(y ~ x, family = "poisson"),
  keras_model = get_keras_model("vss_phn_32")
)
pois_checker$plot_lineup()
```

The true residual plot is at position 4.



```
pois_checker$check()
```

267

```
268 -- <AUTO_POIS_VI object>
269 Status:
270 - Fitted model: glm, lm
271 - Keras model: (None, 32, 32, 3) + (None, 5) -> (None, 1)
272   - Output node index: 1
273 - Result:
274   - Observed visual signal strength: 4.875 (p-value = 0.009901)
275   - Null visual signal strength: [100 draws]
276     - Mean: 1.331
277     - Quantiles:
278
279   25%    50%    75%    80%    90%    95%    99%
280   1.035  1.233  1.488  1.644  1.941  2.276  2.639
281
282   - Bootstrapped visual signal strength: [100 draws]
283     - Mean: 5.51 (p-value = 0.009901)
284     - Quantiles:
285
286   25%    50%    75%    80%    90%    95%    99%
287   5.330  5.505  5.698  5.735  5.830  5.903  6.013
288
289   - Likelihood ratio: 0.05096 (boot) / 0 (null) = Extremely large
```

290

4. Web interface: autovi.web

291 The `autovi.web` shiny application extends the functionality of `autovi` by offering a
 292 user-friendly web interface for automated residual plot assessment. This eliminates the
 293 common challenges associated with software installation, so users can avoid managing
 294 Python environments or handling version requirements for R libraries. The platform
 295 is cross-platform and accessible on various devices and operating systems, making it
 296 suitable even for users without R programming experience. Additionally, updates are
 297 managed centrally, ensuring that users always have access to the latest features. This
 298 section discusses the implementation based on `autovi.web` version 0.1.0.

299 **4.1. Implementation**

300 The interface `autovi.web` is built using the `shiny` (Chang et al. 2022) and
301 `shinydashboard` (Chang & Borges Ribeiro 2021) R packages. Hosted on the
302 `shinyapps.io` domain, the application is accessible through any modern web browser.
303 The R packages `htmltools` (Cheng et al. 2024) and `shinycssloaders` (Sali & Attali
304 2020) are used to render markdown documentation in shiny application, and for loading
305 animations for shiny widgets, respectively.

306 Determining the best way to implement the backend was difficult. In our initial
307 planning for `autovi.web`, we considered implementing the entire web application using
308 the `webr` framework (Moon 2020), which would have allowed the entire application to
309 run directly in the user’s browser. However, `webr` does not support packages which use
310 compiled fortran code, which is required by `splancs` (Rowlingson & Diggle 2023), a
311 dependency of `autovi`. In the future, it is possible that a working Emscripten (Zakai
312 2011) version of this package may allow full `webr` support.

313 We also explored the possibility of implementing the web interface using frameworks
314 built on other languages, such as Python. However, server hosting domains that
315 natively support Python servers typically do not have the latest version of R installed.
316 Additionally, calling R from Python is typically done using the `rpy2` Python library
317 (Gautier 2024), but this approach can be awkward when dealing with language syntax
318 related to non-standard evaluation. Another option we considered was renting a server
319 where we could have full control, such as those provided by cloud platforms like
320 Google Cloud Platform (GCP) or Amazon Web Services (AWS). However, deploying
321 and maintaining the server securely requires some expertise. Ultimately, the most
322 practical solution was to use the `shiny` and `shinydashboard` frameworks, which are
323 well-established in the R community and offer a solid foundation for web application
324 development.

325 The server-side configuration of `autovi.web` is carefully designed to support its
326 functionality. Most required Python libraries, including `pillow` and `numpy`, are pre-
327 installed on the server. These libraries are integrated into the Shiny application using
328 the `reticulate` package, which provides an interface between R and Python.

329 Due to the resource allocation policy of `shinyapps.io`, the server enters a sleep mode
330 during periods of inactivity, resulting in the clearing of the local Python virtual
331 environment. Consequently, when the application “wakes up” for a new user session,
332 these libraries need to be reinstalled. While this ensures a clean environment for each

333 session, it may lead to slightly longer loading times for the first user after a period of
334 inactivity.

335 In contrast to `autovi`, `autovi.web` leverages `TensorFlow.js`, a JavaScript library
336 that allows the execution of machine learning models directly in the browser. This
337 choice enables native browser execution, enhancing compatibility across different user
338 environments, and shifts the computational load from the server to the client-side.
339 `TensorFlow.js` also offers better scalability and performance, especially when dealing
340 with resource-intensive computer vision models on the web.

341 While `autovi` requires downloading the pre-trained computer vision models from
342 GitHub, these models in “.keras” file format are incompatible with `TensorFlow.js`.
343 Therefore, we extract and store the model weights in JSON files and include
344 them as extra resources in the Shiny application. When the application initializes,
345 `TensorFlow.js` rebuilds the computer vision model using these pre-stored weights.

346 To allow communication between `TensorFlow.js` and other components of the Shiny
347 application, the `shinyjs` R package ([Attali 2021](#)) is used. This package allows calling
348 custom JavaScript code within the Shiny framework. The specialized JavaScript
349 code for initializing `TensorFlow.js` and calling `TensorFlow.js` for VSS prediction is
350 deployed alongside the Shiny application as additional resources.

351 4.2. Usage

352 The workflow of `autovi.web` is designed to be straightforward, with numbered
353 steps displayed in each panel. There are two example datasets provided by the
354 web application. The single residual plot example uses the `dino` dataset from the
355 R package `datasauRus` ([Davies, Locke & D'Agostino McGowan 2022](#)). The lineup
356 example uses residuals from a simulated regression model that has a non-linearity
357 issue. We walk through the lineup example to further demonstrate the workflow of
358 the web application.

359 4.2.1. Reading data and setting parameters

360 The user can select to upload data as either a single set of residuals and fitted values
361 in a two (or more) column CSV file or a pre-computed lineup of residuals and null
362 datasets in a three (or more) column CSV file (i.e. multiple sets of residuals and fitted
363 values with a column indicating the set label). Here we illustrate use with lineup
364 example data sets (Figure 5). To use the lineup example data, click the “Use Lineup
365 Example” button. The data status will then update to show the number of rows and

366 columns in the dataset, and the CSV type will automatically be selected to the correct
 367 option. Since the example dataset follows the variable naming conventions assumed
 368 by the web application, the columns for fitted values, residuals, and labels of residual
 369 plots are automatically mapped such that the column named as `.fitted` is mapped
 370 to fitted values, `.resid` is mapped to residuals and if applicable, `.sample` to labels of
 371 the residual set (middle image). If the user is working with a custom dataset, these
 372 options must be set accordingly. Whenever a data containing a lineup, the user must
 373 manually select the label for the true residual plot, otherwise the web application does
 374 not provide all the results. The last step is to click the play button (right image) to
 375 start the assessment.

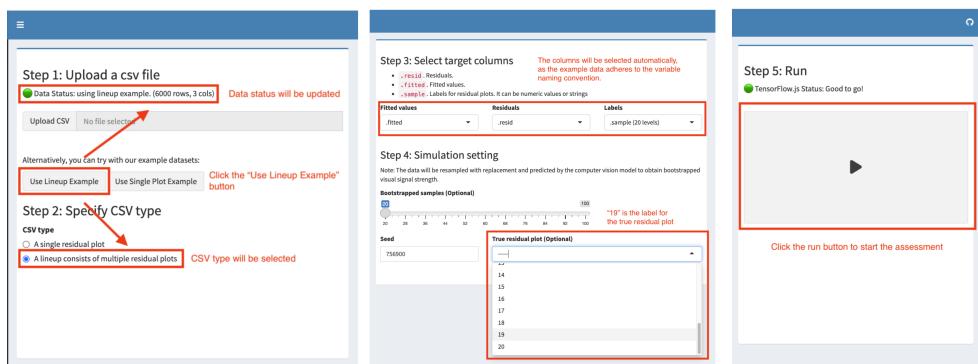


Figure 5. To begin the workflow for `autoovi` using the lineup example dataset, the user clicks the “Use Lineup Example” button (left) to load the example dataset, during which the data status and CSV type will be automatically updated. The user must manually select the label for the true residual plot (middle) to compute further results. The user initiates the assessment of the lineup example data by clicking the run button (right).

376 4.2.2. Results provided

377 Results are provided in multiple panels. The first row of the table Figure 6 is the most
 378 crucial to check, as it provides the VSS and the rank of the true residual plot among
 379 the other plots. The summary text beneath the table provides the *p*-value, which can
 380 be used for quick decision-making. The lineup is for manual inspection, and the user
 381 should see if the true residual plot is visually distinguishable from the other plots, to
 382 confirm if the model violation is serious.

383 The density plot in Figure 7 offers a more robust result, allowing the user to compare
 384 the distribution of bootstrapped VSS with the distribution of null VSS. Finally, the
 385 grayscale attention map (right image) can be used to check if the target visual features,

386 like the non-linearity present in the lineup example, are captured by the computer
 387 vision model, ensuring the quality of the assessment.

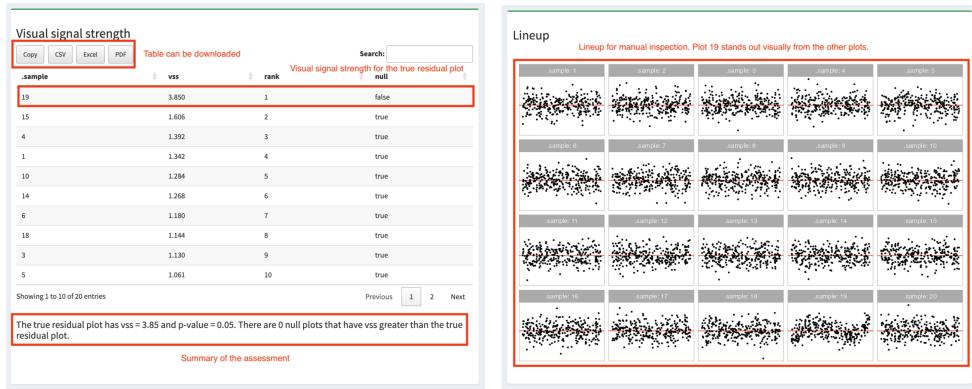


Figure 6. Results for the lineup. The VSS of the true residual plot is displayed in the first row of the table of VSS values for all the null plots (left image), with a summary text beneath the table providing the p -value to aid in decision-making. A lineup of residual plots allows for manual inspection (right image).

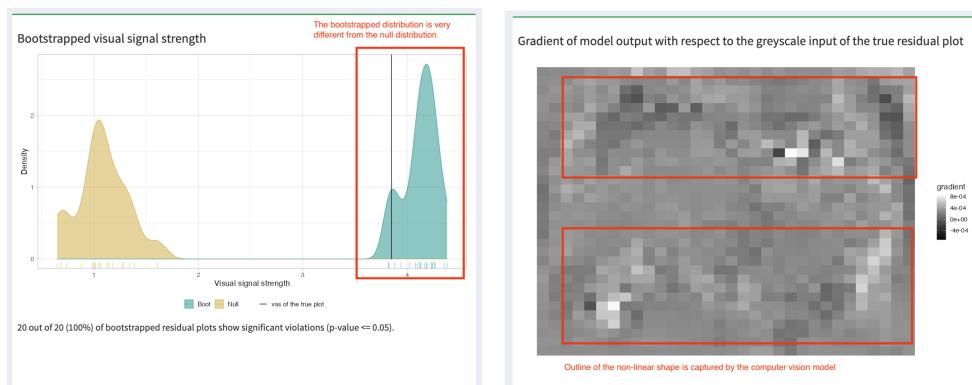


Figure 7. Summaries assessing the strength of the pattern and which elements of the plot contribute. The density plot helps verify if the bootstrapped distribution differs from the null distribution (left image). The attention map (right image) offers insights into whether the computer vision model has captured the intended visual features of the true residual plot.

388

5. Conclusions

389 This paper presents new regression diagnostics software, the R package **autovi** and
 390 its accompanying web interface, **autovi.web**. It addresses a critical gap in the current
 391 landscape of statistical software. While regression tools are widely available, effective
 392 and efficient diagnostic methods have lagged behind, particularly in the field of residual
 393 plot interpretation.

394 The `autovi` R package, introduced in this paper, automates the assessment of residual
395 plots by incorporating a computer vision model, reducing reliance on time-consuming
396 and potentially inconsistent human interpretation. This automation improves the
397 efficiency of the diagnostic process and promotes consistency in model evaluation
398 across different users and studies.

399 The development of the accompanying Shiny app, `autovi.web`, expands access to these
400 advanced diagnostic tools, by providing a user-friendly interface. It makes automated
401 residual plot assessment accessible to a broader audience, including those who may not
402 have extensive programming experience. This web-based solution effectively addresses
403 the potential barriers to adoption, such as complex dependencies and installation
404 requirements, that are often associated with advanced statistical software.

405 The combination of `autovi` and `autovi.web` offers a comprehensive solution to the
406 challenges of residual plot interpretation in regression analysis. These tools have the
407 potential to significantly improve the quality and consistency of model diagnostics
408 across various fields, from academic research to industry applications. By automating
409 a critical aspect of model evaluation, they allow researchers and analysts to focus more
410 on interpreting results and refining models, rather than grappling with the intricacies
411 of plot assessment.

412 The framework established by `autovi` and `autovi.web` opens up exciting possibilities
413 for further research and development. Future work could explore the extension of these
414 automated assessment techniques to other types of diagnostic plots and statistical
415 models, potentially revolutionizing how we approach statistical inference using visual
416 displays more broadly.

417 6. Resources and supplementary material

418 The current version of `autovi` can be installed from CRAN, and source
419 code for both packages are available at github.com/TengMCing/autovi and
420 github.com/TengMCing/autovi_web respectively. The web interface is available from
421 autoviweb.netlify.app.

422 This paper is reproducibly written using Quarto (Allaire et al. 2024) powered by
423 Pandoc (MacFarlane, Krewinkel & Rosenthal 2024) and pdfTeX. The full source code
424 to reproduce this paper is available at github.com/TengMCing/autovi_paper.

425 These R packages were used for the work: `tidyverse` (Wickham et al. 2019), `lmtest`
 426 (Zeileis & Hothorn 2002), `kableExtra` (Zhu 2021), `patchwork` (Pedersen 2022),
 427 `rcartocolor` (Nowosad 2018), `glue` (Hester & Bryan 2022), `here` (Müller 2020),
 428 `magick` (Ooms 2023), `yardstick` (Kuhn, Vaughan & Hvitfeldt 2024) and `reticulate`
 429 (Ushey, Allaire & Tang 2024).

430 **References**

- 431 ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G.S.,
 432 DAVIS, A., DEAN, J., DEVIN, M. et al. (2016). Tensorflow: Large-scale machine learning on
 433 heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- 434 ALLAIRE, J., TEAGUE, C., SCHEIDEGGER, C., XIE, Y. & DERVIEUX, C. (2024). Quarto. doi:
 435 10.5281/zenodo.5960048. URL <https://github.com/quarto-dev/quarto-cli>.
- 436 ATTALI, D. (2021). *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*.
 437 URL <https://CRAN.R-project.org/package=shinyjs>. R package version 2.1.0.
- 438 BALAMUTA, J.J. (2024). *surreal: Create Datasets with Hidden Images in Residual Plots*. URL
 439 <https://CRAN.R-project.org/package=surreal>. R package version 0.0.1.
- 440 BREUSCH, T.S. & PAGAN, A.R. (1979). A simple test for heteroscedasticity and random coefficient
 441 variation. *Econometrica: Journal of the Econometric Society* , 1287–1294.
- 442 BUJA, A., COOK, D., HOFMANN, H., LAWRENCE, M., LEE, E.K., SWAYNE, D.F. & WICKHAM, H.
 443 (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical
 444 Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**,
 445 4361–4383.
- 446 CHANG, W. & BORGES RIBEIRO, B. (2021). *shinydashboard: Create Dashboards with 'Shiny'*.
 447 URL <https://CRAN.R-project.org/package=shinydashboard>. R package version 0.7.2.
- 448 CHANG, W., CHENG, J., ALLAIRE, J., SIEVERT, C., SCHLOERKE, B., XIE, Y., ALLEN, J.,
 449 MCPHERSON, J., DIPERT, A. & BORGES, B. (2022). *shiny: Web Application Framework for
 450 R*. URL <https://CRAN.R-project.org/package=shiny>. R package version 1.7.3.
- 451 CHENG, J., SIEVERT, C., SCHLOERKE, B., CHANG, W., XIE, Y. & ALLEN, J. (2024). *htmltools*:
 452 Tools for HTML. URL <https://CRAN.R-project.org/package=htmltools>. R package version
 453 0.5.8.
- 454 CLARK, A. et al. (2015). Pillow (pil fork) documentation. *readthedocs*.
- 455 COOK, R.D. & WEISBERG, S. (1982). *Residuals and influence in regression*. New York: Chapman
 456 and Hall.
- 457 DAVIES, R., LOCKE, S. & D'AGOSTINO McGOWAN, L. (2022). *datasauRus: Datasets from the
 458 Datasaurus Dozen*. URL <https://CRAN.R-project.org/package=datasauRus>. R package
 459 version 0.1.6.
- 460 GAUTIER, L. (2024). *Python interface to the R language (embedded R)*. URL <https://pypi.org/project/rpy2/>. Version 3.5.16.
- 462 GOODE, K. & REY, K. (2019). *ggResidpanel: Panels and Interactive Versions of Diagnostic Plots
 463 using 'ggplot2'*. URL <https://CRAN.R-project.org/package=ggResidpanel>. R package version
 464 0.3.0.
- 465 HARTIG, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed)
 466 Regression Models*. URL <https://CRAN.R-project.org/package=DHARMa>. R package
 467 version 0.4.6.
- 468 HEBBALI, A. (2024). *olsrr: Tools for Building OLS Regression Models*. URL [https://CRAN.R-project.org/package=olsrr](https://CRAN.R-

 469 project.org/package=olsrr). R package version 0.6.0.

- 470 HESTER, J. & BRYAN, J. (2022). *glue: Interpreted String Literals*. URL <https://CRAN.R-project.org/package=glue>. R package version 1.6.2.
- 471 JOHNSON, P.E. (2022). *rockchalk: Regression Estimation and Presentation*. URL <https://CRAN.R-project.org/package=rockchalk>. R package version 1.8.157.
- 472 KUHN, M., VAUGHAN, D. & HVITFELDT, E. (2024). *yardstick: Tidy Characterizations of Model Performance*. URL <https://CRAN.R-project.org/package=yardstick>. R package version 1.3.1.
- 473 LI, W. (2024). *bandicoot: Light-weight python-like object-oriented system*. URL <https://CRAN.R-project.org/package=bandicoot>.
- 474 LI, W., COOK, D., TANAKA, E. & VANDERPLAS, S. (2024a). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *Journal of Computational and Graphical Statistics* **33**, 1497–1511. doi:10.1080/10618600.2024.2344612.
- 475 LI, W., COOK, D., TANAKA, E., VANDERPLAS, S. & ACKERMANN, K. (2024b). Automated assessment of residual plots with computer vision models. *arXiv preprint arXiv:2411.01001*.
- 476 LONG, J.A. (2022). *jtools: Analysis and Presentation of Social Scientific Data*. URL <https://cran.r-project.org/package=jtools>. R package version 2.2.0.
- 477 LOY, A. & HOFMANN, H. (2014). *Hlmdiag: A suite of diagnostics for hierarchical linear models in R*. *Journal of Statistical Software* **56**, 1–28.
- 478 MACFARLANE, J., KREWINKEL, A. & ROSENTHAL, J. (2024). Pandoc. URL <https://github.com/jgm/pandoc>.
- 479 MASON, H., LEE, S., LAA, U. & COOK, D. (2022). *cassowaryr: Compute Scagnostics on Pairs of Numeric Variables in a Data Set*. URL <https://CRAN.R-project.org/package=cassowaryr>. R package version 2.0.0.
- 480 MOON, K.W. (2020). *webr: Data and Functions for Web-Based Analysis*. URL <https://CRAN.R-project.org/package=webr>. R package version 0.1.5.
- 481 MÜLLER, K. (2020). *here: A simpler way to find your files*. URL <https://CRAN.R-project.org/package=here>. R package version 1.0.1.
- 482 NOWOSAD, J. (2018). 'CARTOCOLORs' palettes. URL <https://nowosad.github.io/rcartocolor>. R package version 1.0.
- 483 OOMS, J. (2023). *magick: Advanced Graphics and Image-Processing in R*. URL <https://CRAN.R-project.org/package=magick>. R package version 2.7.4.
- 484 PEDERSEN, T.L. (2022). *patchwork: The composer of plots*. URL <https://CRAN.R-project.org/package=patchwork>. R package version 1.1.2.
- 485 R CORE TEAM (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- 486 RAMSEY, J.B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* **31**, 350–371.
- 487 REINHART, A. (2024). *regressinator: Simulate and Diagnose (Generalized) Linear Models*. URL <https://CRAN.R-project.org/package=regressinator>. R package version 0.2.0.
- 488 ROWLINGSON, B. & DIGGLE, P. (2023). *splancs: Spatial and Space-Time Point Pattern Analysis*. URL <https://CRAN.R-project.org/package=splancs>. R package version 2.01-44.
- 489 SALI, A. & ATTALI, D. (2020). *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*. URL <https://CRAN.R-project.org/package=shinycssloaders>. R package version 1.0.0.
- 490 SHAPIRO, S.S. & WILK, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611.
- 491 USHEY, K., ALLAIRE, J. & TANG, Y. (2024). *reticulate: Interface to 'Python'*. URL <https://CRAN.R-project.org/package=reticulate>. R package version 1.35.0.
- 492 WARTON, D.I. (2023). Global simulation envelopes for diagnostic plots in regression models. *The American Statistician* **77**, 425–431.

- 519 WICKHAM, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
520 URL <https://ggplot2.tidyverse.org>.
- 521 WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., McGOWAN, L.D., FRANÇOIS, R.,
522 GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., PEDERSEN, T.L., MILLER,
523 E., BACHE, S.M., MÜLLER, K., OOMS, J., ROBINSON, D., SEIDEL, D.P., SPINU, V.,
524 TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. (2019). Welcome to
525 the tidyverse. *Journal of Open Source Software* **4**, 1686. doi:10.21105/joss.01686.
- 526 WICKHAM, H., CHOWDHURY, N.R., COOK, D. & HOFMANN, H. (2020). *nullabor: Tools for*
527 *Graphical Inference*. URL <https://CRAN.R-project.org/package=nullabor>. R package version
528 0.3.9.
- 529 ZAKAI, A. (2011). Emscripten: an llvm-to-javascript compiler. In *Proceedings of the ACM*
530 *international conference companion on Object oriented programming systems languages and*
531 *applications companion*. pp. 301–312.
- 532 ZEILEIS, A. & HOTHORN, T. (2002). Diagnostic checking in regression relationships. *R News* **2**,
533 7–10.
- 534 ZHU, H. (2021). *kableExtra: Construct complex table with kable and pipe syntax*. URL
535 <https://CRAN.R-project.org/package=kableExtra>. R package version 1.3.4.