Behavior Research Methods (2024) 56:826–845
https://doi.org/10.3758/s13428-023-02072-x

Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics

Itamar Shatz[1]

Accepted: 17 January 2023 / Published online: 3 March 2023

Abstract

Statistical methods generally have assumptions (e.g., normality in linear regression models). Violations of these assumptions can cause various issues, like statistical errors and biased estimates, whose impact can range from inconsequential to critical. Accordingly, it is important to check these assumptions, but this is often done in a flawed way. Here, I first present a prevalent but problematic approach to diagnostics—testing assumptions using null hypothesis significance tests (e.g., the Shapiro–Wilk test of normality). Then, I consolidate and illustrate the issues with this approach, primarily using simulations. These issues include statistical errors (i.e., false positives, especially with large samples, and false negatives, especially with small samples), false binarity, limited descriptiveness, misinterpretation (e.g., of *p*-value as an effect size), and potential testing failure due to unmet test assumptions. Finally, I synthesize the implications of these issues for statistical diagnostics, and provide practical recommendations for improving such diagnostics. Key recommendations include maintaining awareness of the issues with assumption tests (while recognizing they can be useful), using appropriate combinations of diagnostic methods (including visualization and effect sizes) while recognizing their limitations, and distinguishing between *testing* and *checking* assumptions. Additional recommendations include judging assumption violations as a complex spectrum (rather than a simplistic binary), using programmatic tools that increase replicability and decrease researcher degrees of freedom, and sharing the material and rationale involved in the diagnostics.

Keywords Statistical assumptions · Assumption checks · Statistical diagnostics · Null hypothesis significance testing · Graphical methods · Visualization

Introduction

Statistical assumptions

Statistical methods, like hypothesis tests and regression models that are often used in the behavioral sciences, generally involve various assumptions. For example, linear models generally involve the assumption that their *residuals*[1] (i.e., the differences between observed and predicted values) are normally distributed (i.e., have a *Gaussian* distribution). This assumption also applies to common statistical tests that are special cases of linear

✉ Itamar Shatz
is442@cam.ac.uk

[1] University of Cambridge, Cambridge, UK

[1] This assumption actually pertains to the models' *errors*, which cannot be directly observed, and which are therefore estimated using the residuals (Barker & Shaw, 2015; Bilon, 2021; Cook & Weisberg, 1999; Knief & Forstmeier, 2021; Pek et al., 2018). Accordingly, throughout the paper, we will discuss this and related assumptions in the context of residuals, rather than errors. However, this approach has some flaws, like the problem of *supernormality*, where residuals appear to be normal even if the errors are not normal, particularly in small samples (Gnanadesikan, 1997; Weisberg, 2005). Such issues led to the development of various types of residuals, like *recursive residuals*, which may sometimes be better suited for diagnostic purposes than raw residuals (Hawkins, 1991; Kianifard & Swallow, 1996).

models, like the *t*-test and ANOVA, as well as to methods that extend these models, like linear mixed-effects models (Barker & Shaw, 2015; Casson & Farmer, 2014; Hox et al., 2018; Knief & Forstmeier, 2021; Pole & Bondy, 2012; Poncet et al., 2016; Rochon et al., 2012; Vallejo et al., 2021; Winter, 2019). Furthermore, it can apply to other quantitative methods, including inferential statistics, like confidence intervals (Alf & Lohr, 2007), and descriptive statistics, like mean and standard deviation (Al-Hoorie & Vitta, 2019). Additional information about this and other assumptions, particularly in the context of linear regression, appears in Appendix 1.

Violations of these assumptions can cause various issues, like statistical errors and biased estimates, whose impact can range from inconsequential to critical (Barker & Shaw, 2015; Ernst & Albers, 2017; Gel et al., 2005; Hayes & Cai, 2007; Hu & Plonsky, 2021; Knief & Forstmeier, 2021; Poncet et al., 2016; Rosopa et al., 2013; Schmidt & Finan, 2018; Troncoso Skidmore & Thompson, 2013; Vallejo et al., 2021; Zuur et al., 2010). Accordingly, it is recommended to consider the assumptions of statistical methods when using those methods, and to use statistical diagnostics to determine whether any assumptions are violated, and if so then how they are violated and to what degree (Barker & Shaw, 2015; Casson & Farmer, 2014; Gel et al., 2005; Hox et al., 2018; Osborne & Waters, 2003; Poncet et al., 2016; Schmidt & Finan, 2018; Tay et al., 2016; Zuur et al., 2010). When violations are detected, the diagnostics can also drive the decision of what to do; common options include switching methods (e.g., to robust non-parametric ones), transforming the data (e.g., by taking its logarithm), or sticking to the original analysis (Casson & Farmer, 2014; Pek et al., 2018; Pole & Bondy, 2012; Vallejo et al., 2021).

Motivation for this paper

As discussed above, checking assumptions is crucial to ensuring the validity of statistical analyses.

However, the way assumptions are currently checked is often flawed, due to issues like the use of statistical tests in a way that is likely to involve false positives, and these issues persist despite having been mentioned in various previous works (Anderson et al., 2001; Bilon, 2021; Cumming, 2014; Di Leo & Sardanelli, 2020; Ernst & Albers, 2017; Gelman & Stern, 2006; Knief & Forstmeier, 2021; Kozak & Piepho, 2018; Lakens, 2021; Rosnow & Rosenthal, 1989; Tijmstra, 2018; Wasserstein & Lazar, 2016; Winter, 2019; Zuur et al., 2010). Furthermore, lack of awareness and understanding of these issues contributes to the currently insufficient use and reporting of assumption checks in the scientific literature (Hoekstra et al., 2012; Hu & Plonsky, 2021; Nielsen et al., 2019; Nimon, 2012).[2] For example, a review and empirical analysis by Hu and Plonsky (2021) suggest that assumption checks are likely reported in under 25% of studies in social-science fields like linguistics, psychology, and education, and that many of these reports are lacking (e.g., because they mention only some of the relevant checks). This is supported by other research in the social sciences, such as a study by Ernst and Albers (2017), who found that in psychology research involving linear regression, only 2% of studies were both transparent and correct in reporting assumption checking, and a further 6% were transparent but incorrect. This is also supported by research in other fields, like medicine (e.g., Nielsen et al., 2019), though more research is needed in order to determine the exact rate of reporting, especially to understand how it varies across fields and whether it is increasing over time (Hu & Plonsky, 2021).

One reason for the persistence of the issues with assumption checks is insufficient statistical literacy among researchers, so a possible partial solution is to develop relevant resources on proper assumption-checking, which researchers can learn from (Hu & Plonsky, 2021; Loewen et al., 2014). There are already, as noted above, many works that mentioned these issues. However, they are generally limited, in the sense that they either do so only briefly and in passing (e.g., Winter, 2019), focus on only one or some of these issues (e.g., Tijmstra, 2018), and/or discuss these issues outside the context of statistical diagnostics (e.g., Gelman & Stern, 2006). Furthermore, some works (e.g., Bilon, 2021) present these issues from a technical perspective (e.g., using equations), which readers may struggle to understand and translate into practice, especially if they lack a strong quantitative background, as is often the case (Hu & Plonsky, 2021). All this is *not* meant to criticize these works, which simply had a different focus (e.g., exploring a single issue), but is rather meant to point out an existing and important gap in the literature.

The goal of the present paper is to address this gap, and to consequently improve the way assumption checking is conducted. Specifically, the paper expands on previous work in several ways. First, it aggregates the key common issues with assumption checking, to discuss all of them in one place. Furthermore, it illustrates these issues in a

[2] It is unclear to what extent the issue is that assumptions are not checked, as opposed to being checked but not reported (Hu & Plonsky, 2021). Nevertheless, the low rates of reporting—and especially of *correct* reporting—suggest there is a lack of understanding of the importance of assumption checking, and of how to do it properly (Hoekstra et al., 2012; Hu & Plonsky, 2021; Nielsen et al., 2019; Nimon, 2012). In addition, even when assumptions are checked correctly, a lack transparency regarding the checks (i.e., due to lack of reporting) can cause issues for researchers who try to interpret, compare, or replicate a study's analyses (Ernst & Albers, 2017; Hu & Plonsky, 2021).

manner that is meant to be intuitive and non-technical, in order to make the material accessible to diverse audiences, including those who have only limited statistics expertise but nevertheless use statistical methods in their work (Hu & Plonsky, 2021). Finally, it takes advantage of the aforementioned aggregation of these issues, in order to synthesize generalizable practical recommendations for improving assumption checking, which again should be accessible to diverse audiences.

The present paper therefore aims to serve as a resource that can be used in several key ways. First, it can be used by researchers to learn how to conduct better statistical diagnostics, and also to explain the rationale behind their diagnostic approach to readers and reviewers, by serving as a comprehensive reference. In addition, this paper may also be used by reviewers and editors, who can use it to guide the statistical diagnostics of authors, by mentioning it during the review process, and potentially also listing it as a methodological resource in the submission guidelines of journals (Hu & Plonsky, 2021; Loewen et al., 2014). Finally, it can also be used for pedagogical purposes, for example by educators who wish to direct their students to an accessible paper that explains how to conduct statistical diagnostics. This aligns with calls to improve the current state of statistical diagnostics in research (Hu & Plonsky, 2021; Nielsen et al., 2019; Nimon, 2012). This also aligns with the goal of *Behavior Research Methods* (BRM) to publish, among other things, "tutorials alerting readers to avoidable mistakes that are made time and time again" (Brysbaert et al., 2020, p. 1), in order to make research "more effective, less error-prone, and easier to run" (ibid.).

Brief overview of assumption testing

It is often recommended to *test* the assumptions of statistical methods before using them, using *null hypothesis significance tests* (NHST, sometimes referred to in this context as *numerical tests*). For example, when assessing the normality of residuals, a common recommendation is to use the *Shapiro–Wilk test* (Gel et al., 2005; Ghasemi & Zahediasl, 2012; Knief & Forstmeier, 2021; Mishra et al., 2019; Rochon et al., 2012). Generally, when this approach is used, if the resulting p-value of the test is $< .05$, then the residuals are considered significantly non-normal (i.e., the null hypothesis that the data is normally distributed is rejected), meaning that the assumption of normality is considered to be violated.

This approach to checking assumptions can be appealing for various reasons. For example, it involves a single well-established threshold (generally $p < .05$), which reduces some of the arbitrariness and researcher degrees of freedom when using such checks (Wicherts et al., 2016).[3] Second, it relies on the NHST framework, which many researchers are familiar with and are already using extensively in other parts of their work (Tijmstra, 2018; Troncoso Skidmore & Thompson, 2013; Veldkamp, 2017). Finally, it involves tests that are generally easy to implement from a programmatic perspective, and that are often reported automatically by certain software, so researchers may have the results of these tests available to them by default (Hoekstra et al., 2012).

However, as will be shown in the next section, there are various issues with this approach, which can cause serious problems for those who use it.

Issues with testing assumptions

The following subsections illustrate the key issues with using assumption tests for diagnostics of statistical methods. These issues include statistical errors, false binarity, limited descriptiveness, misinterpretation, and potential testing failure due to unmet test assumptions.

Statistical errors

Testing assumptions can cause both *false positives* (i.e., *type I errors*) and *false negatives* (i.e., *type II errors*), as shown below.

False positives occur when a test incorrectly leads to the conclusion that there is an assumption violation, in cases where there is no such violation. For example, this can happen if the test incorrectly leads to the conclusion that a certain distribution is non-normal, in a situation where it is actually normal (and should be normal). This issue is especially common with large samples, where even tiny, random, and inconsequential deviations from an expected distribution often lead to statistically significant differences from that distribution (Bilon, 2021; Bishara et al., 2021; Kozak & Piepho, 2018; Mishra et al., 2019).

This issue is illustrated in Fig. 1. The plots it contains show that, as the sample size increases (going from left to right), the distribution of the randomly generated samples approaches normality, as indicated by the observed distribution (the blue shaded area) aligning with an expected

[3] However, it does not eliminate these issues entirely. For example, researchers may need to choose between multiple available tests, such as the Shapiro–Wilk test and the *Kolmogorov-Smirnov (K-S) test*, where the K-S test generally has lower power to detect non-normality (Ghasemi & Zahediasl, 2012; Steinskog et al., 2007). Furthermore, researchers also be faced with many other alternatives to choose from, which can also lead to different results, such as the *Lilliefors test*, the *Anderson–Darling test*, and the *Jarque-Bera test* (Das & Imon, 2016; Ghasemi & Zahediasl, 2012; Pole & Bondy, 2012; Steinskog et al., 2007).

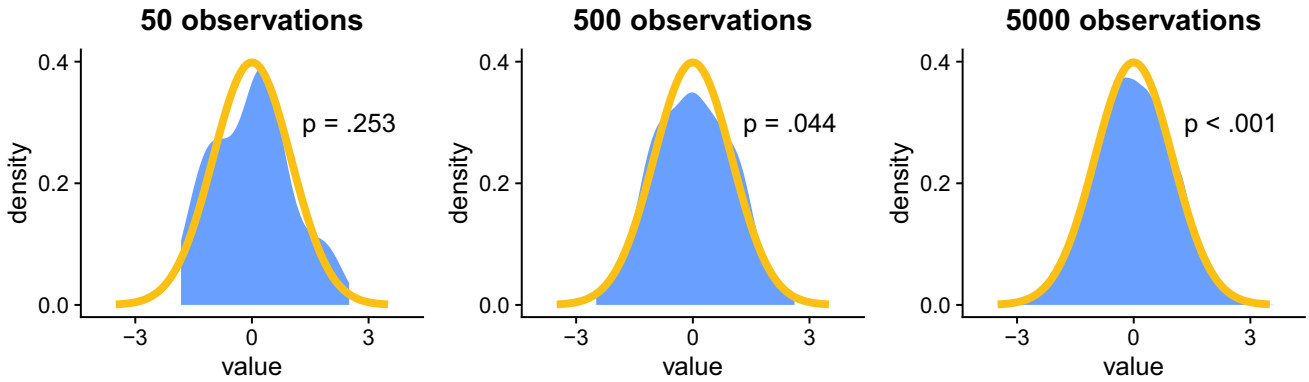


Fig. 1 **General background**: The *x*-axis indicates standardized values (*mean* = 0, *standard deviation* = 1); the *y*-axis indicates value density (e.g., 0.2 means 20% of observations have this value). The orange line indicates the expected density for a normal distribution; the blue shaded area indicates the observed density for the samples. *p*-values are from Shapiro–Wilk normality tests. **Specific background**: Each sample was randomly generated using identical settings to have a roughly normal distribution with random noise. Samples differ in size, which increases from 50 → 500 → 5000 (left-to-right). Note that the leftmost panel is not truncated, but because it has fewer observations, by chance none are more than 1.8 SD below the mean. **Takeaway**: As sample size increases, the observed distribution approaches normality but the *p*-value decreases, illustrating the risk of false positives in assumption testing, especially in large samples

normal distribution (the orange line). However, these plots also show that as the sample size increases—and the sample approaches normality—the *p*-value of the associated assumption test decreases. Paradoxically, this means that the smallest sample ($N = 50$), which is the least normal, might be interpreted as the only sample where the normality assumption is not violated (since $p > .05$). Conversely, the medium sample ($N = 500$), which is closer to normality, might be interpreted as non-normal (but somewhat borderline, since $p = .044$), and the largest sample ($N = 5000$), which is closest to normality, might be interpreted as entirely non-normal (since $p < .001$).

This figure demonstrates that, although increasing the sample size is generally beneficial to statistical analyses, it can cause issues when testing assumptions, since large samples are likely to appear to be significantly different from expected distributions according to NHST (Bilon, 2021; Bishara et al., 2021;Kozak & Piepho, 2018 ; Mishra et al., 2019). This *size-significance paradox* of assumption testing can lead to unwarranted lack of confidence in results from large and "overpowered" samples, where minor assumption violations may be incorrectly interpreted as worse than they are (Bishara et al., 2021; Kozak & Piepho, 2018).

Conversely, the second type of statistical errors that assumption tests can cause—false negatives—occur when a test incorrectly leads to the conclusion that there is no assumption violation, in cases where there is one. For example, this can happen if the test leads to the conclusion that a certain distribution is normal (or more accurately, not significantly non-normal), in a situation where it is not. This issue is especially common with small samples, where even substantial and systematic deviations from an expected distribution may not be statistically significant, due to insufficient statistical power (i.e., insufficient ability to detect such deviations at a statistically significant level) (Kozak & Piepho, 2018; Mishra et al., 2019).

This issue is illustrated in Fig. 2. The plots that it contains show three samples with substantial deviations from normality—due to noise, skewness, and bimodality—as indicated by the shape of the observed distributions (blue shaded area). However, in all these cases, the samples might be interpreted as normal based on the associated assumption test (since $p > .05$).

Accordingly, assumption testing can also lead to unwarranted confidence in small and underpowered samples, where the tests are sometimes unable to detect even strong assumption violations.[4] Together with the issue of false positives, this ironically means that, according to assumption tests, large and normal sample can sometimes be seen as non-normal, whereas small and non-normal samples can sometimes be seen as normal.

False binarity

Assumption tests generally involve a hard *threshold* (or *cut-off*), so assumption violations are determined only based on

[4] Note that a "strong" violation (e.g., substantial deviation from normality) does *not* necessarily entail "strong" consequences for analysis. For example, in a simulation study, Knief and Forstmeier (2021) found that "Gaussian models are robust to non-normality over a wide range of conditions, meaning that *p*-values remain fairly reliable except for data with influential outliers judged at strict alpha levels. Gaussian models also performed well in terms of power across all simulated scenarios. Parameter estimates were mostly unbiased and precise except if sample sizes were small or the distribution of the predictor was highly skewed." (p. 2576).

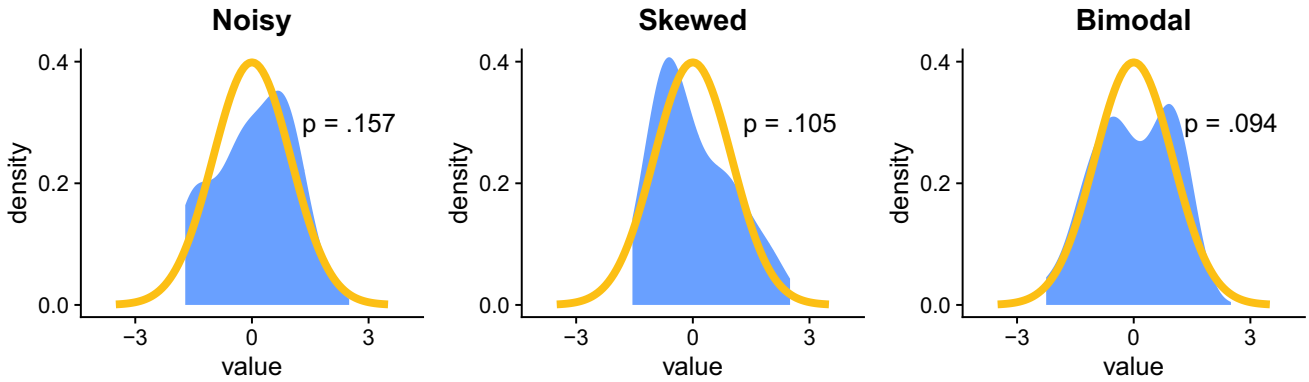


Fig. 2 General background: The *x*-axis indicates standardized values (*mean* = 0, *standard deviation* = 1); the *y*-axis indicates value density (e.g., 0.2 means 20% of observations have this value). The orange line indicates the expected density for a normal distribution; the blue shaded area indicates the observed density for the samples. *p*-values are from Shapiro–Wilk normality tests. **Specific background**: The first sample (left-to-right) is normally distributed but noisy, the second is skewed, and the third is bimodal ($N = 30$ in all samples). **Takeaway**: Despite the substantial deviations from normality, the tests do not detect non-normality, illustrating the risk of false negatives in assumption testing, especially in small samples

whether $p < .05$, in a *binary* (or *dichotomous*) way. This can lead to completely different interpretations of the data based on inconsequential differences in *p*-values (Gelman & Stern, 2006; Greenland et al., 2016; Halsey, 2019; Wasserstein & Lazar, 2016). For example, as shown in Fig. 3, if the result of a normality test is $p = .051$, then the sample might be considered "normal" (or more accurately, not "significantly" non-normal), whereas if the result is $p = .049$, then the sample might be considered "significantly" non-normal, even though the difference between these values is functionally meaningless.[5]

In addition, this binary thinking also compresses a diverse spectrum of possible assumption violations into a narrow false dichotomy. This simplistic view of assumption violations ignores potential nuances, such as that there are different *types* of violations (as was shown in Fig. 2 and will be shown in the next subsection), as well as different *magnitudes* of violations. This issue with magnitude is illustrated in Fig. 4, where, for example, the bottom-right plot appears substantially more non-normal than the bottom-left plot, but both may simply be considered as "non-normal" based on an assumption test (since $p < .05$ in both cases). Note that this plot also illustrates an associated issue with using a hard threshold in assumption tests, since the bottom-left and bottom-right plots are both categorized as non-normal, even though the distribution of the bottom-left plot is more similar to that of the top-right plot (which is not non-normal).

Limited descriptiveness

Assumption tests, particularly when used with a binary mindset, generally only indicate whether the distribution at hand is "significantly" different from some expected distribution (Greenland et al., 2016; Wasserstein & Lazar, 2016). However, this does not indicate much about how different the distribution is from expected (in terms of magnitude), as was shown in Fig. 4, or in what way the distribution is different, as was shown in Fig. 2. This latter issue is further illustrated in Fig. 5, which contains three plots, each representing a sample that deviates significantly from normality (due to noise, skewness, and bimodality, as indicated by the shape of the observed distributions). Here, the assumption tests detect the assumption violation (unlike in Fig. 2, where they failed to do so), but do not provide any further information about its nature, since all tests merely indicates that $p < .001$.

The informativeness of graphical methods compared to numerical methods is also illustrated in *Anscombe's quartet* (Anscombe, 1973) and the *Datasaurus dozen* (Matejka & Fitzmaurice, 2017), which appear in Appendix 2. These are collections which show how data with very different distributions can have the same summary statistics (e.g., mean, SD, and correlation).

Misinterpretation

The results of assumption tests can be misinterpreted due to issues that commonly occur when people interpret the results of NHST (Gelman & Stern, 2006; Greenland et al., 2016;

[5] As Rosnow & Rosenthal (1989, p. 1277) state: "…surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?" This reflects the continuous—rather than binary—nature of *p*-values, as well as the fact that although the associated threshold of .05 is widely adopted, it is also arbitrary (Gelman & Stern, 2006; Greenland et al., 2016). Essentially, this issue means that a tiny quantitative difference in *p*-value leads to a disproportionate qualitative difference in analyses, as going from $p = .051$ to $p = .049$ means that the distribution is suddenly considered "non-normal."

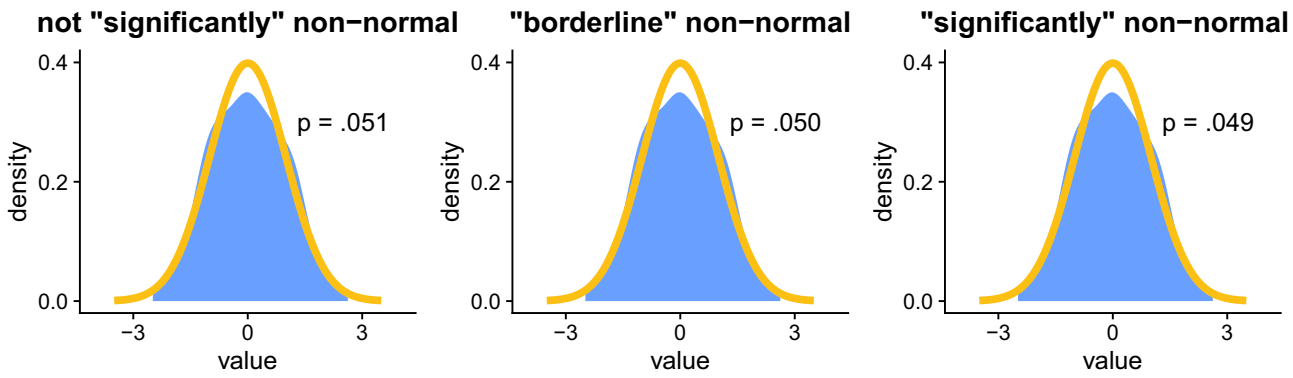


Fig. 3 General background: The *x*-axis indicates standardized values (*mean* = 0, *standard deviation* = 1); the *y*-axis indicates value density (e.g., 0.2 means 20% of observations have this value). The orange line indicates the expected density for a normal distribution; the blue shaded area indicates the observed density for the samples. *p*-values are from Shapiro–Wilk normality tests. **Specific background**: Each sample ($N = 500$) was randomly generated to have a normal distribution with slightly more noise than the previous (left-to-right). **Takeaway**: Based on assumption tests with a hard threshold ($p < .05$), the normality of these samples is classified differently, even though the differences in normality between the samples are tiny

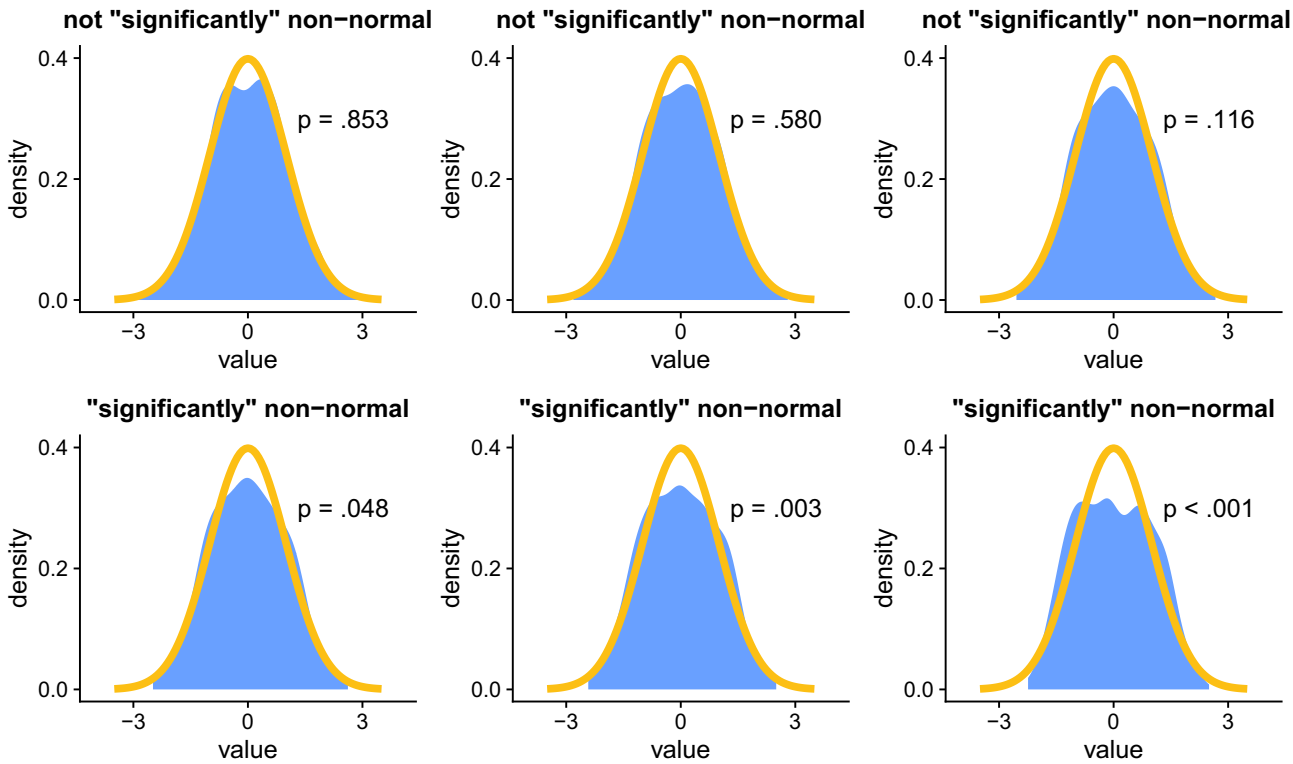


Fig. 4 General background: The *x*-axis indicates standardized values (*mean* = 0, *standard deviation* = 1); the *y*-axis indicates value density (e.g., 0.2 means 20% of observations have this value). The orange line indicates the expected density for a normal distribution; the blue shaded area indicates the observed density for the samples. *p*-values are from Shapiro–Wilk normality tests. **Specific background**: Each sample ($N = 500$) was randomly generated to have a normal distribution, with substantially more noise going from left-to-right and then top-to-bottom. **Takeaway**: Assumption tests with a hard threshold ($p < .05$) designate the top plots as "normal" (or more accurately, as not non-normal) and the bottom plots as non-normal, but do not capture substantial differences in distributions within each group (e.g., the increased non-normality in the bottom-right plot compared to the bottom-left one)

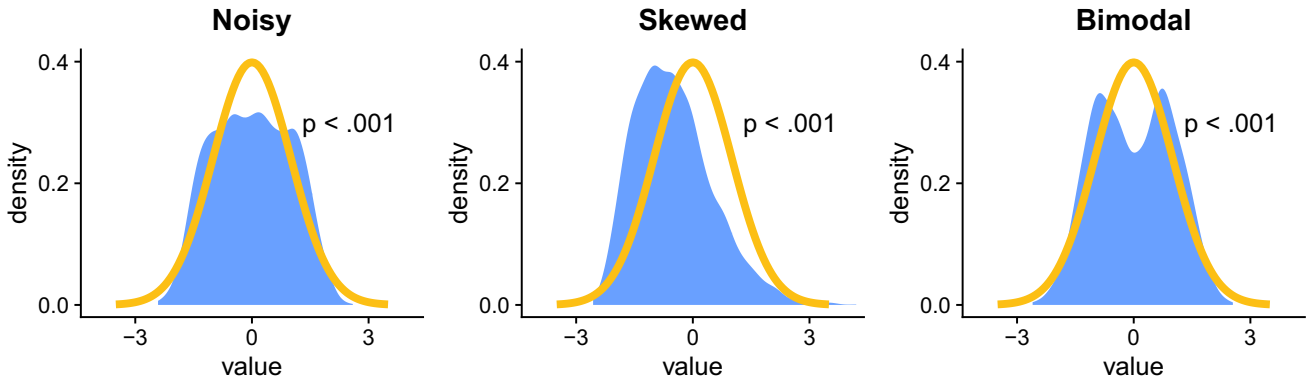


Fig. 5 General background: The *x*-axis indicates standardized values (*mean* = 0, *standard deviation* = 1); the *y*-axis indicates value density (e.g., 0.2 means 20% of observations have this value). The orange line indicates the expected density for a normal distribution; the blue shaded area indicates the observed density for the samples. *p*-values are from Shapiro–Wilk normality tests. **Specific background**: The first sample (left-to-right) is normally distributed but noisy, the second sample is skewed, and the third is bimodal ($N = 3000$ in all samples). **Takeaway**: The assumption tests indicate that there is non-normality in all three samples, but do not reveal the substantially different ways in which the samples are non-normal

Lakens, 2021; Wasserstein & Lazar, 2016). For example, the result of tests might be misinterpreted as suggesting that there is a substantial difference between $p = .051$ and $p = .049$ in the Shapiro–Wilk test, even though this difference is generally meaningless (Gelman & Stern, 2006; Wasserstein & Lazar, 2016), as was shown in the previous sub-section on false binarity. Similarly, the resulting *p*-value might be incorrectly interpreted as an effect size, for instance if people assume that a low *p*-value from the Shapiro–Wilk test (e.g., $p < .001$) indicates strong non-normality (Wasserstein & Lazar, 2016). In addition, non-significant results might be misinterpreted as indicating that a sample is "significantly" normal, even though the associated test only shows that we cannot reject the null hypothesis that the distribution is normal (Greenland et al., 2016; Wasserstein & Lazar, 2016).

Potential testing failure due to unmet test assumptions

In addition to the aforementioned issues with assumptions tests, which pertain primarily to statistical significance, another issue is that the tests themselves can have various assumptions. This adds further complexity to the diagnostics, as well as more room for error, since researchers can neglect to account for these added assumptions.

To illustrate this, we will look at another common assumption; that a model's residuals have *constant variance* (Casson & Farmer, 2014; Cook & Weisberg, 1983; Hayes & Cai, 2007; Rosopa et al., 2013; Schmidt & Finan, 2018; Zuur et al., 2010). As with the normality assumption, testing this assumption can also lead to various issues, like false negatives. This is illustrated by the right plot in Fig. 6, where the test fails to detect a clear pattern of non-constant variance, which is indicated by the curved shape of the line and systematic—rather than random—distribution of the residuals around it.

A potential reason for this issue, beyond sample size, is that the commonly used *Breusch–Pagan test* of constant variance itself has assumptions, including normality of residuals, violations of which can cause failure to detect non-constant variance (Barker & Shaw, 2015; Cribari-Neto & Zarkos, 1999; Halunga et al., 2017; Waldman, 1983). This is illustrated in Fig. 7, where, despite clear visual patterns of non-constant variance (as indicated by the regression line being curved and having the residuals distributed systematically around it), the associated Breusch–Pagan test fails to detect the violation, partially due to the violation of the normality assumption.

A path forward

In the previous section, we saw the key issues associated with assumption testing. In this section, we will see practical recommendations for improving statistical diagnostics, especially in light of the aforementioned issues.

Beware the issues with assumption testing

One way to minimize the issues that we saw with assumption tests is to account for these issues when using such tests. For example, to avoid false binarity when interpreting the results of a Shapiro–Wilk test, remember that the *p*-value is measured on a continuous spectrum (e.g., there is little difference between $p = .051$ and $p = .049$). In addition, to minimize these issues, you should consider them when deciding whether to use assumption tests in the first place, as there are

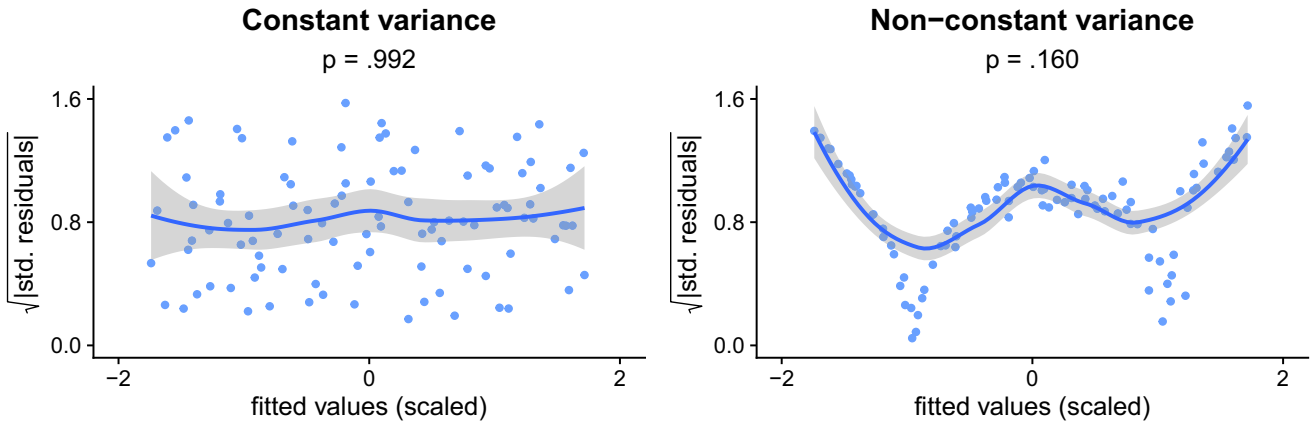


Fig. 6 **Background**: Each scatterplot shows the square root of the absolute values of standardized residuals in a simple regression model ($N = 100$), as a function of their scaled fitted values. Each plot also contains a corresponding smoothed regression line, with a grey band for the 95% CI. Constant variance is indicated by a flat and horizontal line, with residuals spread randomly around it. *p*-values are from *Breusch–Pagan tests*. **Takeaway**: There is clear non-constant variance in the right plot, since the line is heavily curved and the residuals are systematically distributed around it, but the associated assumption test fails to detect this

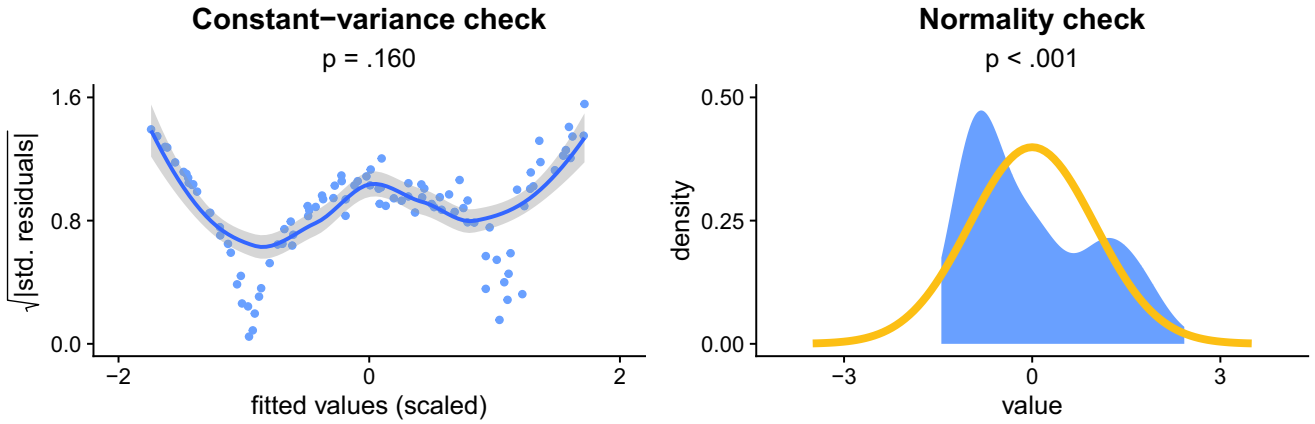


Fig. 7 **Background**: The left plot shows a constant-variance check, for the model with non-constant variance shown in Fig. 6. The right plot shows a normality check for this model (based on the paradigm described in Figs. 1, 2, 3, 4 and 5). **Takeaway**: The violation of the normality assumption hinders the ability of the Breusch–Pagan test to detect the non-constant variance

cases where it is preferable to replace or supplement them with alternative checks, as discussed next.

Use visualizations in statistical diagnostics

When seeing the issues with assumption tests, we also saw how graphical methods, like residual plots, can help reduce or avoid these issues, and consequently improve statistical diagnostics. This means that you will often benefit from using graphical methods in your diagnostics, as has also been noted by others (Bilon, 2021; Hox et al., 2018; Knief & Forstmeier, 2021; Kozak & Piepho, 2018; Winter, 2019; Zuur et al., 2010).

In addition, using such diagnostics is becoming easier than ever, and in many cases is as easy as using assumption tests. This is illustrated in Fig. 8, which shows how a range of relevant visual assumption checks can be generated with a simple function in R.[6]

When using graphical methods, there is the question of how to conduct *graphical inference* (or *visual statistical inference*), which involves drawing conclusions about statistical properties using visualizations (Hullman & Gelman, 2021; Loy, 2021; Majumder et al., 2013; Wickham et al., 2010). In the present context, a key question is how to judge the severity of assumption violations, in order to determine whether a certain visual pattern deviates enough from expected to merit a response. Though the question of

[6] The ease of use of these *turnkey* tools—where you only need to provide the data and potentially choose some basic settings—might help increase the rate of assumption checking over time.

graphical inference is under active consideration (Hullman & Gelman, 2021), and though there is no perfect method for this, there are nevertheless some methods that can help.

One such method is to use visual aids, like confidence intervals that are overlain on plots, to add information regarding the certainty associated with the visual patterns. Examples of this were shown in Fig. 8 (e.g., in the homogeneity of variance plot), where they were generated automatically by the *performance* package in R. A similar approach is utilized in the R *qqtest* package, which generates *self-calibrating* QQ-plots that visually incorporate sampling variation into the display (Oldford, 2016). Furthermore, it is sometimes beneficial to supplement visualizations with numerical aids, including effect sizes and statistical significance (as will be discussed in the next two sub-sections), since they may provide complementary information that aids the judgment and decision-making process (Flatt & Jacobs, 2019; Hartig, 2021).

Another method you can use is the *lineup protocol*, (Buja et al., 2009; Loy, 2021; Majumder et al., 2013; Wickham et al., 2010). This involves generating (e.g., using simulation) a random set of similar distributions that do *not* contain the assumption violation of interest (e.g., residual non-normality), and then checking if you and others can identify the plot containing the original data out of the ones containing the data without a violation. The less able people are to identify the original plot, the less likely it is that it involves a substantial violation (Buja et al., 2009; Loy, 2021; Majumder et al., 2013; Wickham et al., 2010). This protocol—as well as the Rorschach protocol which is mentioned next—can be implemented using programmatic tools like the *nullabor* package in R (Wickham et al., 2010).

In addition, to better understand how to judge visual patterns, it could help to look at relevant examples that are used to illustrate the presence/absence of violations, for example in package documentation (e.g., of the R *DHARMa* package, which is discussed in the next sub-section), or in other instructional sources (e.g., Winter, 2019). When doing this, you can also use the *Rorschach protocol*, by examining randomly generated plots in which assumptions are not violated (e.g., without non-normality). This can help calibrate your expectations regarding the variability that such plots can involve, to reduce the tendency to view random patterns in the data as assumption violations (Buja et al., 2009; Wickham et al., 2010).

Furthermore, when assessing visual patterns, you can ask for input from other individuals (Majumder et al., 2013). In doing so, you should prioritize input from those with expertise in interpreting such plots, who may be able to judge them better, and try to not reveal the goals of the research or any previous judgments of the visual patterns until after these individuals provided their input, to minimize potential bias (Veldkamp, 2017; Wicherts et al., 2016).

Finally, once you make a decision regarding a visual assumption check, you should describe your rationale, for example by explaining which visual patterns you found concerning and why. You should also share the relevant plots (as supplementary material if necessary), in order to be transparent and ensure that other researchers can see these visualizations and apply their own judgment to them.

Use effect sizes in statistical diagnostics

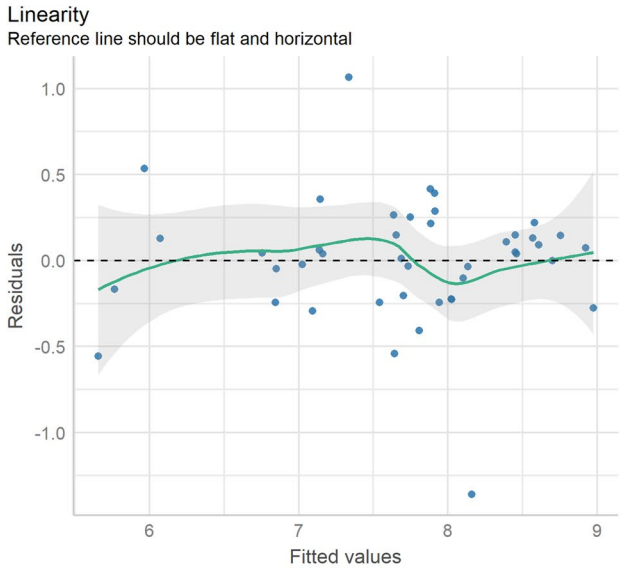
There are situations where you can benefit from using numerical measures of effect size in your diagnostics, to help identify assumption violations and quantify their magnitude. Such effect sizes are not currently commonly used for the two main assumptions that were discussed so far in the paper (normality and constant variance), but can be used for other important assumptions.

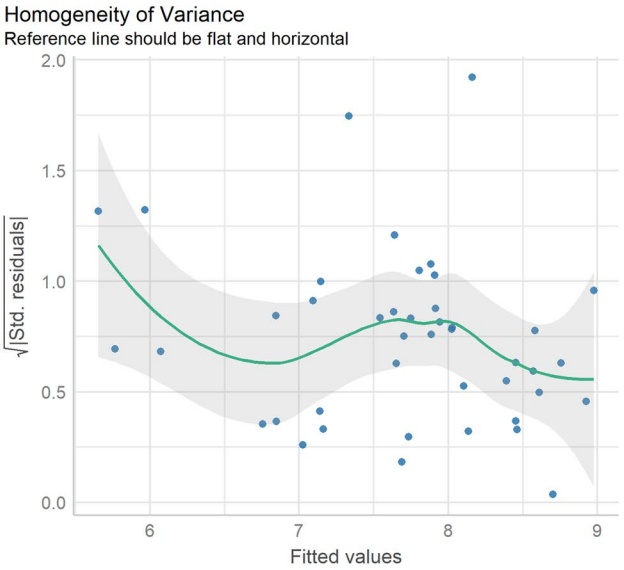
A key example of this appears in the context of Poisson regression models, a type of *generalized linear model* (GLM), used for working with count data (Forthmann & Doebler, 2021; Green, 2021; Winter, 2019).[7] The Poisson distribution assumes that the mean and the variance of the data are equal (i.e., that there is *equidispersion*), but this assumption is often violated (Coxe et al., 2009). This can be either due to *overdispersion*, when the variance is bigger than expected, or *underdispersion*, when the variance is smaller than expected (Brooks et al., 2017; Forthmann & Doebler, 2021). Overdispersion leads to underestimated (i.e., *liberal*) standard errors, *p*-values, and confidence intervals, while underdispersion leads to overestimated (i.e., *conservative*) standard errors, *p*-values, and confidence intervals (Brooks et al., 2017; Forthmann & Doebler, 2021). Accordingly, it is important to check dispersion when using Poisson models, and based on the results of these checks, researchers may, for example, choose to use alternative methods, like negative binomial models (Brooks et al., 2017, 2019; Winter, 2019).

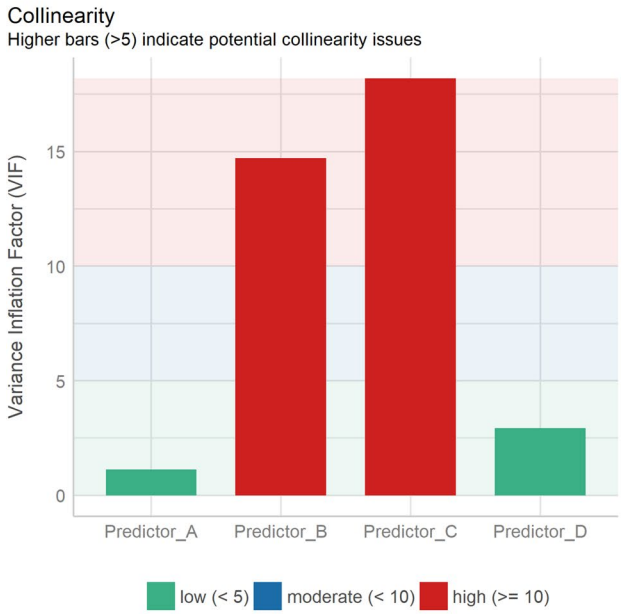
One way to check dispersion is to use the *testDispersion* function in the *DHARMa* package in R (Hartig, 2021), which compares the variance of a model's observed residuals against the variance of its expected residuals (as determined based on simulations).[8] This outputs a *p*-value for the

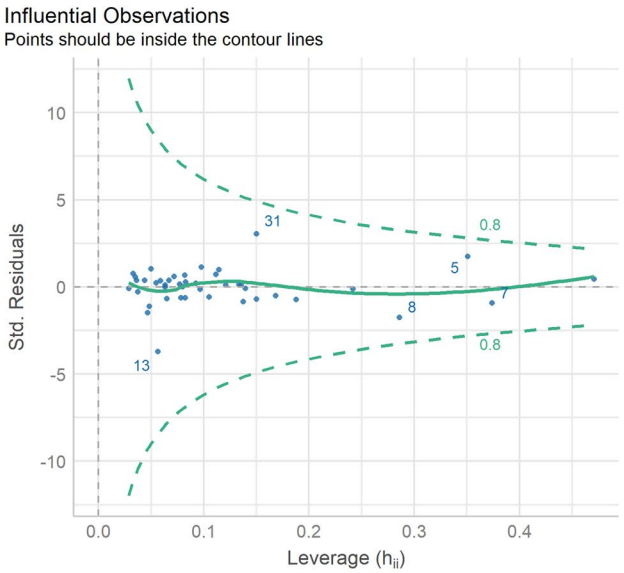
[7] Count data can be equal only to zero or positive integers (i.e., it can take values like 0, 1, 2, 3…) (Green, 2021; Winter, 2019). It is used to model things like the number of speech errors in a text (Winter, 2019), the number of publications of a researcher (Forthmann & Doebler, 2021), or the number of times someone exercises per week (Green, 2021).

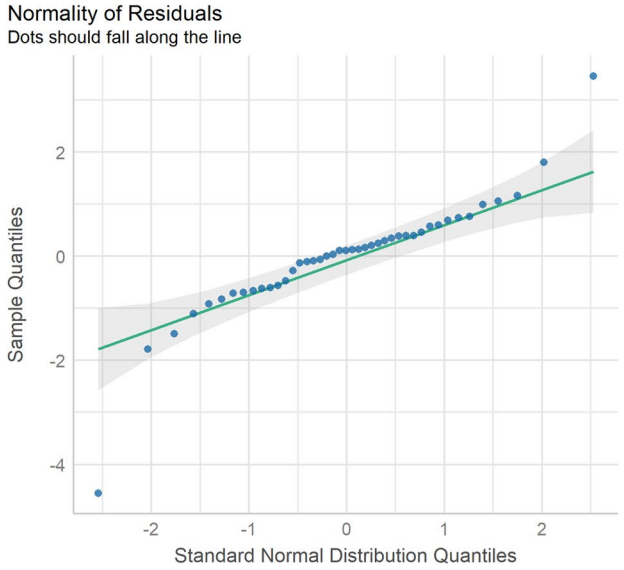
[8] For more information on this approach to diagnostics of GLMs and GLMMs (*generalized linear mixed models*), see the DHARMa documentation (Hartig, 2021). The documentation also demonstrates how visualizations can complement the dispersion test. One example of this is by showing a histogram of expected variance values based on simulations, which can help to see how different the observed variance is from the expected. Another example of this is by showing a QQ plot of the residuals, which can help to visually assess dispersion and other potential issues (like deviation from normality). The DHARMa documentation also shows how these diagnostics can be generated through a few simple functions in R, in a similar way to how checks for LMs can be generated using the *performance* package, as was shown in the previous sub-section.

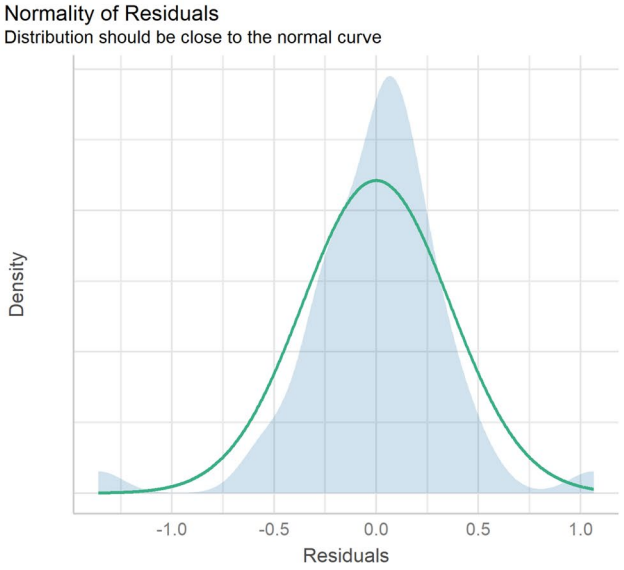

Linearity
Reference line should be flat and horizontal
Residuals
Fitted values


Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values


Collinearity
Higher bars (>5) indicate potential collinearity issues
Variance Inflation Factor (VIF)
Predictor_A
Predictor_B
Predictor_C
Predictor_D
low (< 5)
moderate (< 10)
high (>= 10)


Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (hii)
31
5
13
0.8
0.8


Normality of Residuals
Dots should fall along the line
Sample Quantiles
Standard Normal Distribution Quantiles


Normality of Residuals
Distribution should be close to the normal curve
Density
Residuals

◂**Fig. 8** Diagnostics for a regression model, generated using the *performance* package in R (Lüdecke et al., 2021), by simply running *check_model*(model_name). These checks are for linearity, homogeneity of variance (i.e., constant variance), collinearity, influential observations, and normality of residuals; for more details, see the package's documentation

dispersion test, together with a dispersion ratio as an effect size, where a ratio > 1 indicates overdispersion, while a ratio < 1 indicates underdispersion.

In a large-scale simulation (N = 150,000) with a Poisson model, this function outputted the following results: $ratio_{dispersion} = 0.98$, $p < .001$. Relying only on the p-value would suggest that there is a statistically significant deviation from the expected dispersion, but will not reveal whether the problem is overdispersion or underdispersion, or what is the magnitude of the deviation from the expected dispersion. However, looking at the associated effect size (the dispersion ratio) reveals that there is underdispersion (since the ratio < 1), and more importantly, that this deviation from the expected dispersion is very small (since the dispersion ratio is very close to 1) (Hartig, 2021).

A related issue with Poisson models is *zero-inflation*, which occurs when count data contains more zeros than expected (Brooks et al., 2017, 2019). This issue can cause biased parameter estimates, and can be addressed using solutions like zero-inflated models (Brooks et al., 2017, 2019; Green, 2021; Harrison, 2014).

One way to check for this issue is to use the *testZeroInflation* function in the DHARMa package, which is similar to the *testDispersion* function (Hartig, 2021). This function compares the observed number of zeros with expected number of zeros (based on simulations). It outputs a p-value for the associated test, together with a ratio of observed to expected zeros as an effect size, where a ratio < 1 indicates that the observed data has fewer zeros than expected, while a ratio > 1 indicates that it has more zeros than expected (i.e., has zero-inflation).

In a large-scale simulation (N = 150,000) with a Poisson model, this function outputted the following results: $ratio_{observed\_to\_expected\_zeros} = 1.01$, $p < .001$. Again, relying only on the p-value would suggest that there is a statistically significant deviation from the expected number of zeros. However, this will not reveal whether there are more or fewer zeros than expected (despite the name of this function, it tests both possibilities by default), or how big the deviation is. Looking at the associated effect size (the ratio of observed to expected zeros) reveals that there are more zeros than expected in the model (since the ratio > 1), but that this deviation from the expected ratio is again very small (since the ratio is very close to 1) (Hartig, 2021).

Finally, another example of a numerical effect size that can be used in statistical diagnostics is the *variance inflation factor* (VIF). It quantifies the severity of *collinearity* (or *multicollinearity*) in regression models, where a higher VIF indicates greater collinearity, and consequently greater inflation in the standard errors of the coefficients (Alin, 2010; Dormann et al., 2013). VIF is often shown using plots like the one in Fig. 9, which provide a visual representation of this numeric measure.

Remember that assumption tests can be useful

As we saw previously, assumption tests have certain limitations, such as that they do not provide information about the way in which observed distributions differ from expected. Nevertheless, assumption tests can sometimes be useful, primarily when they complement other types of assumption checks (Flatt & Jacobs, 2019; Hartig, 2021).

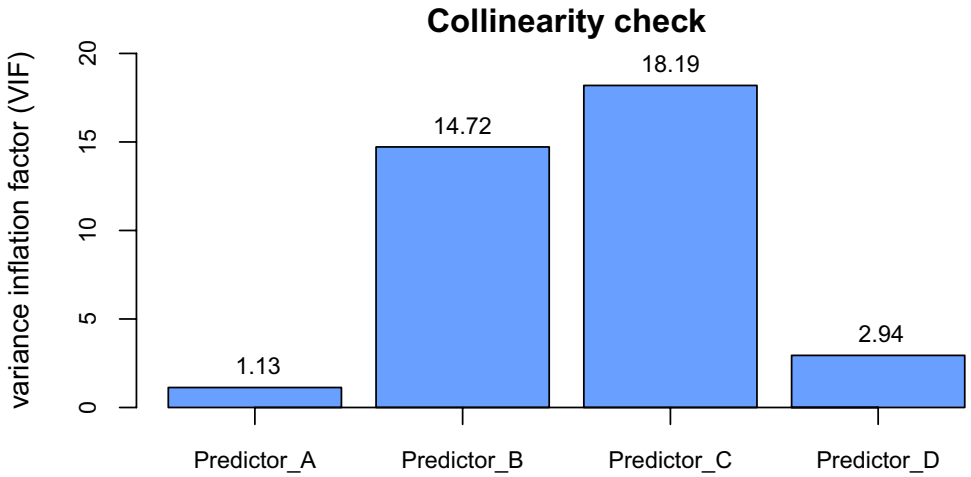


Fig. 9 An example collinearity check for a multiple regression model, where a higher VIF indicates greater collinearity

A key example of this appears in statistical diagnostics of Poisson models, which, as discussed in the previous sub-section, should generally be checked for overdispersion and underdispersion by calculating the relevant effect size (dispersion ratio) (Hartig, 2021). When doing this, it can sometimes be beneficial to run an associated significance test for the difference between the observed and expected dispersion—while considering the sample's size—as this can help assess whether a certain deviation from expected might be due to chance (Hartig, 2021).

In addition, assumption tests can also be beneficial when used for initial diagnostics. For example, tests can be used in this manner when generating a large number of models (e.g., hundreds or thousands), which may be infeasible to assess visually; in such situations, the assumption tests may be used to identify a subset of models that are more likely to have issues, and these models can then be inspected visually. As with the assessment of dispersion, here too it is possible to use further statistics when assessing the results from the assumption tests, and particularly effect size (where available) and sample size.

When using assumption tests in this and other capacities, it is important to remember the issues that they can involve (e.g., false negatives), and to make sure that you are minimizing the risks of those issues (e.g., by considering whether your samples are large enough for tests to detect non-normality). This also involves considering the assumptions that these tests have, and choosing tests that are most appropriate for your situation.

In addition, when deciding whether and how to use assumption tests, you should consider other relevant factors, including how concerned you are over false positives/negatives, and how you want to balance the validity of your analyses with the time spent checking their assumptions. For example, if you need to run diagnostics on a large number of models (e.g., 100) for a preliminary analysis where you are not concerned about false positive/negatives, then you might decide that for your specific purposes it is better to use automated assumption tests rather than visual checks. Alternatively, if you are conducting diagnostics and are more concerned over false negatives than false positives (e.g., since you can visually assess any models of concern), then you may change your *p*-value threshold to reflect this (e.g., from .05 to .10), while again also considering factors such as the size of your samples (Bishara et al., 2021).

Finally, note that it may be beneficial to use *equivalence tests* in this context, rather than traditional NHST. Such tests "examine whether the hypothesis that there are effects extreme enough to be considered meaningful can be rejected" (Lakens et al., 2018, p. 260). They add flexibility to the diagnostic process, by enabling researchers to set the quantitative bounds involved in rejecting the hypothesis, based on relevant evidence. In the context of normality, for example, using these tests can lead people to move from asking "can we reject the hypothesis that the data are normally distributed?", to asking "is this distribution consistent enough with a normal distribution for our present purposes?"

Remember that visualizations and effect sizes are imperfect

Despite the potential benefits of using visual checks and effect sizes in statistical diagnostics, it is important to remember that they have limitations.

One limitation of visual checks is that they may not detect certain issues. For example, the panel of assumption checks for linear models that was shown in Fig. 8 is not expected to detect dependence between observations (e.g., due to the presence of multiple data points per participant). Furthermore, people sometimes fail to notice issues that are evident in visualizations (e.g., non-normality), due to errors in judgment (Bishara et al., 2021; Fisch, 1998). This issue may be worse in small samples (e.g., just 10 observations), where it can be harder to interpret visual patterns with confidence (Cook & Weisberg, 1983; Weissgerber et al., 2016). The opposite issue can also occur, when people *overinterpret* visual patterns. For example, this can involve believing that a certain residual pattern is indicative of substantial non-normality, when in reality it is merely the result of some trivial noise. Psychologically, the tendency to see such patterns in random noise can be considered a form of *apophenia* (Wickham et al., 2010), and can be attributed to causes like Gestalt principles (Dixon, 2012).

Another limitation of visual checks is that they can have misleading results in some cases. For example, as Cook and Weisberg (1983) note, using residual plots to check for constant variance can wrongly seem to indicate that there is non-constant variance if the density of the points is uneven along the *x*-axis, since areas with higher density generally also have a greater spread on the *y*-axis, due to the increased number of observations.

Another issue with visual checks is the subjectivity involved in the interpretation of visual patterns, which can cause issues like biased interpretation of results (Bishara et al., 2021). Furthermore, there is often arbitrariness in the choice of which graphical methods to use. For example, visual normality checks can be performed using many methods other than the density plots that were used in this paper, including *histograms*, *boxplots* (also called *box-and-whisker plots*), and *normal quartile plots* (also called *Q-Q plots*) (Das & Imon, 2016; Mishra et al., 2019; Pole & Bondy, 2012). Moreover, many of these methods have various settings that can influence their interpretation, like bin size in

histograms and the opacity of points in dot plots (Correll et al., 2019). Similarly to the choice of which assumption tests to use, the choice of which graphical methods to use increases the researcher degrees of freedom (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016). This, in turn, can lead to issues like selecting a method because it supports one's hypotheses rather than because it is the most appropriate method to use, for instance due to an unconscious confirmation bias (Veldkamp, 2017; Wicherts et al., 2016).

Finally, visual checks may also be infeasible in some cases. For example, this can be the case if you need to deal with a very large number of models (e.g., 5000), which may take too long for you to assess visually.

Effect sizes also have various limitations when used in statistical diagnostics. For example, consider the VIF, which was mentioned previously as a measure of effect size for collinearity. One issue with VIF is that it is often interpreted using arbitrary rules of thumb, for instance when a VIF of 4 or 10 is considered to indicate the presence of "severe" collinearity, which merits a change to analyses (O'Brien, 2007). This can lead to a similar issue with false binarity as the p threshold of .05 (as discussed in §3.2), for instance if a VIF of 4.01 is viewed as indicating severe collinearity, while a VIF of 3.99 does not.

Furthermore, VIF values are often considered in isolation, based only on their magnitude, but other factors, like sample size, also play a key role in determining how substantial the impact of collinearity is on associated statistical inferences (O'Brien, 2007).[9] In addition, the influence of collinearity should be considered in the context of the inferential goals of the analysis. As Belsley et al. (2004, p. 116) note:

> … for example, if an investigator is only interested in whether a given coefficient is significantly positive, and is able, even in the presence of collinearity, to accept that hypothesis on the basis of the relevant t-test, then collinearity has caused no problem. Of course, the resulting forecasts or point estimates may have wider confidence intervals than would be needed to satisfy a more ambitious researcher, but for the limited purpose of the test of significance [initially] proposed, collinearity has caused no practical harm… These cases serve to exemplify the pleasantly pragmatic philosophy that collinearity doesn't hurt so long as it doesn't bite.

Finally, another issue with VIF is that it merely quantifies the degree of collinearity present in the data, but does not say anything about what type of collinearity exists (Alin, 2010), which is reminiscent of the issue of limited descriptiveness of assumption tests discussed in §3.3. This is problematic, since different types of collinearity may necessitate different responses. For example, Iacobucci et al. (2016) show that using mean centering reduces what they refer to as "micro" collinearity, but not "macro" collinearity.

All this does *not* mean that visual checks and effect sizes should be avoided in statistical diagnostics, but rather that they should be used with appropriate caution, similarly to other statistical methods, like assumption tests.

Use proper terminology

Because of the issues associated with assumption testing, it is important to draw a clear terminological distinction between *testing* and *checking* assumptions. Specifically, the term "assumption testing" should only be used to refer to the testing of assumptions using statistical tests (e.g., the Shapiro–Wilk test). Conversely, the term "assumption checking" can be used to refer to all forms of assumption checks, including statistical tests, as well as visualizations and numerical assessments of effect sizes.

In addition to enabling a more nuanced discussion of statistical diagnostics, drawing this distinction will help emphasize the importance of considering more than just statistical significance when checking assumptions. Doing this may, in turn, help prevent certain cases where people are told to "test" their assumptions, and interpret this as meaning that they should only use statistical tests in their diagnostics. This is particularly relevant for researchers with a limited statistical background, who are less likely to understand the issues with assumption tests, and who comprise a substantial portion of those who use statistical methods in practice (Hu & Plonsky, 2021).

Key practical recommendations

Based on the material discussed in the paper so far, the following are key practical recommendations for conducting statistical diagnostics:

1. Remember the potential issues with assumption checks when deciding whether/how to use them, and when interpreting their results.

[9] As O'Brien (2007, p. 675) notes: "…unless the collinearity is perfect, increasing the sample size (using more cases of the same sort) will reduce the variance of the regression coefficients" and "When the focus is on the variance of a regression coefficient and the stability of result: the sample size, the proportion of the variance in the dependent variable associated with the independent variables, the variance of the independent variable whose coefficient is of concern, and the multi-collinearity of the independent variable of concern with the other independent variables are all important."

2. Prefer the most appropriate type of assumption check to use in your particular situation (e.g., visualization over a test), even if other methods are more common.
3. Use a combination of diagnostic methods where appropriate (e.g., significance testing initially, followed by visualization).
4. Draw a terminological distinction between assumption *testing* (which involves only statistical tests) and assumption *checking* (which can involve statistical tests, as well as other methods, including visualization and numerical effect sizes).
5. Prefer using existing programmatic tools for diagnostics, and using their default settings, unless you have a compelling reason to do otherwise. This can make the diagnostic process easier to implement, more replicable, and more comparable across studies, while also reducing researcher degrees of freedom (Gelman & Loken, 2014; Simmons et al., 2011; Wicherts et al., 2016). Examples of relevant tools include the *performance* and *DHARMa* R packages, whose use was demonstrated earlier.
6. Explain the rationale behind your diagnostic process, including what you checked, how, and why, and how you interpreted the results. When doing this, acknowledge any important limitations and arbitrariness in your process, for example if there were other reasonable methods you could have used.
7. Share all the material that you used in the diagnostics, like the code that you used and the resulting plots. It may be best to do this as part of online supplementary material, particularly if space constraints would otherwise prohibit you from sharing important information (Hu & Plonsky, 2021).
8. Judge assumption violations as a complex spectrum, rather than a simplistic binary. This means that you should consider not only whether there is a violation, but also what the violation is, what caused it, how severe it is, and how it affects your particular analyses, while also considering factors like the robustness of your methods to this type of violation, the size of your sample, and your inferential goals. You may realize that your method is robust enough or the violation is minor enough that nothing needs to be done, especially when analyzing large samples, which are usually—but not always—more robust to violations (Casson & Farmer, 2014; Ernst & Albers, 2017; Fagerland, 2012; Ghasemi & Zahediasl, 2012; Knief & Forstmeier, 2021; Kozak & Piepho, 2018; Lumley et al., 2002; Pole & Bondy, 2012; Poncet et al., 2016; Schmidt & Finan, 2018; Tijmstra, 2018). This is often the case, for example, with violations of normality (Knief & Forstmeier, 2021) or collinearity (O'Brien, 2007). You may also realize that even if the violation is substantial, it does not affect the goals of your particular analyses. For example, if the goal of an analysis is mainly inference of the regression line, rather than prediction of individual data points, then the normality assumption might not be important (Gelman et al., 2022).
9. Remember that statistical diagnostics cannot detect all the potential issues in statistical methods. For example, such diagnostics cannot, in many cases, detect key issues with the validity of the model specification (e.g., omitted-variable bias) or with the representativeness of the sample (e.g., systematic biases in the sampling process). However, these issues can be far more important than issues with normality and constant variance (Gelman et al., 2022), as mentioned in Appendix 1. Accordingly, it is important to consider such issues before running your analyses, even if you cannot check them using formal methods. Another example of such an issue are inappropriate causal interpretations of regression results, which might only be identified by doing things like assessing the language used to present results (Bordacconi & Larsen, 2014).
10. Remember that assumption checks cannot guarantee that your analyses are correct. Rather, they provide evidence—which you then assess—regarding the presence and severity of certain assumption violations. As such, they can only increase your confidence that your analyses are not grossly wrong, or provide you with information regarding what kind of issues they suffer from (Faraway, 2016; Hartig, 2021; Winter, 2019). In this regard, it helps to remember Box's aphorism that "All models are wrong but some are useful" (Box, 1979, p. 202), and his recommendation to worry selectively: "Since all models are wrong the scientist must be alert to what is importantly wrong" (Box, 1976, p. 972).

Conclusions

When using statistical methods like linear models, you should generally check if and how their assumptions are violated. This is because assumption violations can have various consequences, so assessing assumptions is crucial to deciding whether and how to proceed with analyses. A common way to do this is to use statistical tests, like the Shapiro–Wilk test of normality, but as shown in the present paper, this approach involves various potential issues, including statistical errors (false positives and false negatives), false binarity, limited descriptiveness, misinterpretation (e.g., of p-values as effect sizes), and potential testing failure due to unmet test assumptions.

Despite this, assumption tests can sometimes be beneficial. However, the aforementioned issues mean that if assumption tests are used, then this should be done with caution, and generally to supplement visualization and/or numerical effect sizes, though these types of assumption checks also have limitations. In addition, assumption checks can also be improved

by following other practical guidelines which were outlined in the paper, including explaining the rationale behind the diagnostic process, sharing all the relevant material, and judging assumption violations as a complex spectrum.

Appendices

Appendix 1: Assumptions of linear regression models

The following is a brief summary of the key assumptions of simple linear regression models. It is based primarily on the description of Gelman et al. (2022, pp. 153–155), who also ranked the assumptions in the following (decreasing) order of importance:

1. ***Validity.*** The data should map to your research question, by (i) having an outcome that accurately reflects the phenomenon of interest, (ii) including all relevant predictors in the model, and (iii) the model generalizing to all cases to which it will be applied.
2. ***Representativeness.*** The sample should be representative of its parent population, primarily in terms of the distribution of the response variable given the predictors in the model.
3. ***Additivity and linearity.*** The deterministic component of the model should be a linear function of its separate predictors. Addressing violations of this can involve things like adding interactions or transforming the data.
4. ***Independence of errors.*** The model's errors should be independent of one another. Addressing violations of this can involve things like using mixed-effects models. Note that this and the other assumptions involving errors are checked by examining the model's *residuals*, which approximate the errors, since they cannot be observed directly (Cook & Weisberg, 1999; Knief & Forstmeier, 2021).
5. ***Equal variance of errors.*** The variance in the errors should be constant across all levels of the independent variables (i.e., along the regression line) (Barker & Shaw, 2015; Winter, 2019). This means that the conditional variance of the response variable is the same for all observations (Cribari-Neto & Zarkos, 1999; Fox, 2022). This assumption is also called *constant variance*, *homogeneity of variance*, and *homoscedasticity*. It is contrasted with *unequal variance*, which is also called *non-constant variance*, *heterogeneity of variance*, and *heteroscedasticity*.
6. ***Normality of errors.*** The errors should be normally distributed. Note that this applies only to the errors, so while the conditional distribution of the response as a function of the predictors should be normal, the raw response and predictors can be non-normal.

An important caveat is that there is variability in the literature in terms of which assumptions are discussed and how, due to things like varied conventions across fields. For example, a list of regression assumptions in Fox (2022) explicitly mentions most of the mathematical assumptions listed above (linearity, constant variance, normality, and independence), but not additivity, and not the conceptual assumptions (validity and representativeness). Alternatively, in econometrics, the focus and framing of the discussion often revolves around the *Gauss-Markov assumptions*, and includes concepts like exogeneity, auto-correlation, and omitted-variable bias (Belsley et al., 2004; Verbeek, 2008).

In addition, there are two other important caveats about these assumptions. First, certain issues that are important to consider in statistical diagnostics are not always considered assumptions. For example, while Fox (2022) mentions lack of (perfect) collinearity as an assumption, Gelman et al. do not, although they do discuss collinearity as a potential issue elsewhere in their text. Similarly, although Fox does not mention lack of outliers as an assumption, he does discuss outliers as a potential issue elsewhere in his text. The second caveat is that other types of models can have other assumptions, like the *equidispersion* assumption of Poisson GLMs, which was discussed in the paper (Coxe et al., 2009).

Appendix 2: Informativeness of graphical methods

In addition to the simulations that were shown in the body of the paper, another way to illustrate the potential benefits visualization in statistical analyses is through datasets that demonstrate how data with very different distributions can have the same summary statistics. Included below are two such datasets— *Anscombe's quartet* and the *Datasaurus dozen*—which are often used for this purpose.

Anscombe's quartet

Appendix Fig. 10 contains plots of *Anscombe's quartet* (Anscombe, 1973), showing four datasets with very different distributions of observations, but functionally identical summary statistics. These datasets demonstrate how visualization can be useful for detecting many common types of assumption violations (Cook & Weisberg, 1999; Hox et al., 2018; Winter, 2019); for example, dataset 2 reveals a pattern of non-linearity.

Data for the plots came from the *anscombe* dataset in base R (R Core Team, 2021), which is based on Anscombe (1973).

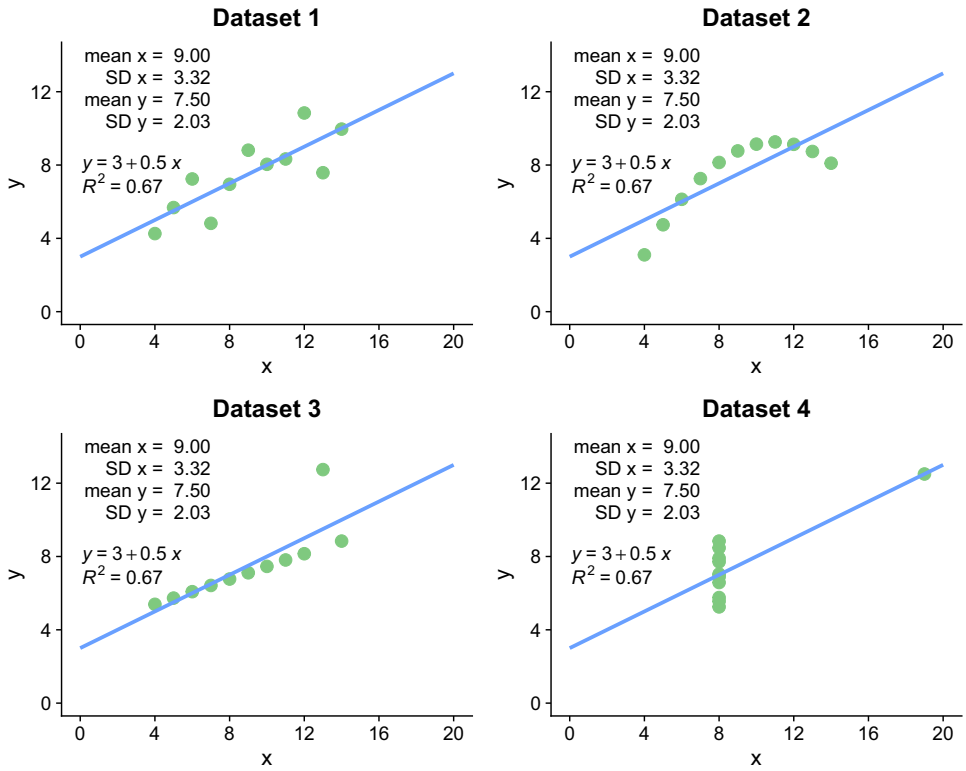


Fig. 10 Plots of *Anscombe's quartet* (Anscombe, 1973). The observations (green points) are distributed very differently across each dataset. However, despite this, the mean and SD of each variable are the same across the datasets, as are the regression equations (represented by the blue line) and corresponding R^2

Datasaurus dozen

Appendix Fig. 11 contains plots of the *Datasaurus dozen* (Matejka & Fitzmaurice, 2017), showing 13 (i.e., a baker's dozen) datasets with very different distributions of observations but functionally the same summary statistics.

Data for the plots came from the *datasauRus* package in R (Davies et al., 2022), which is based on the datasets from Matejka and Fitzmaurice (2017) and the original Datasaurus by Cairo (2016).

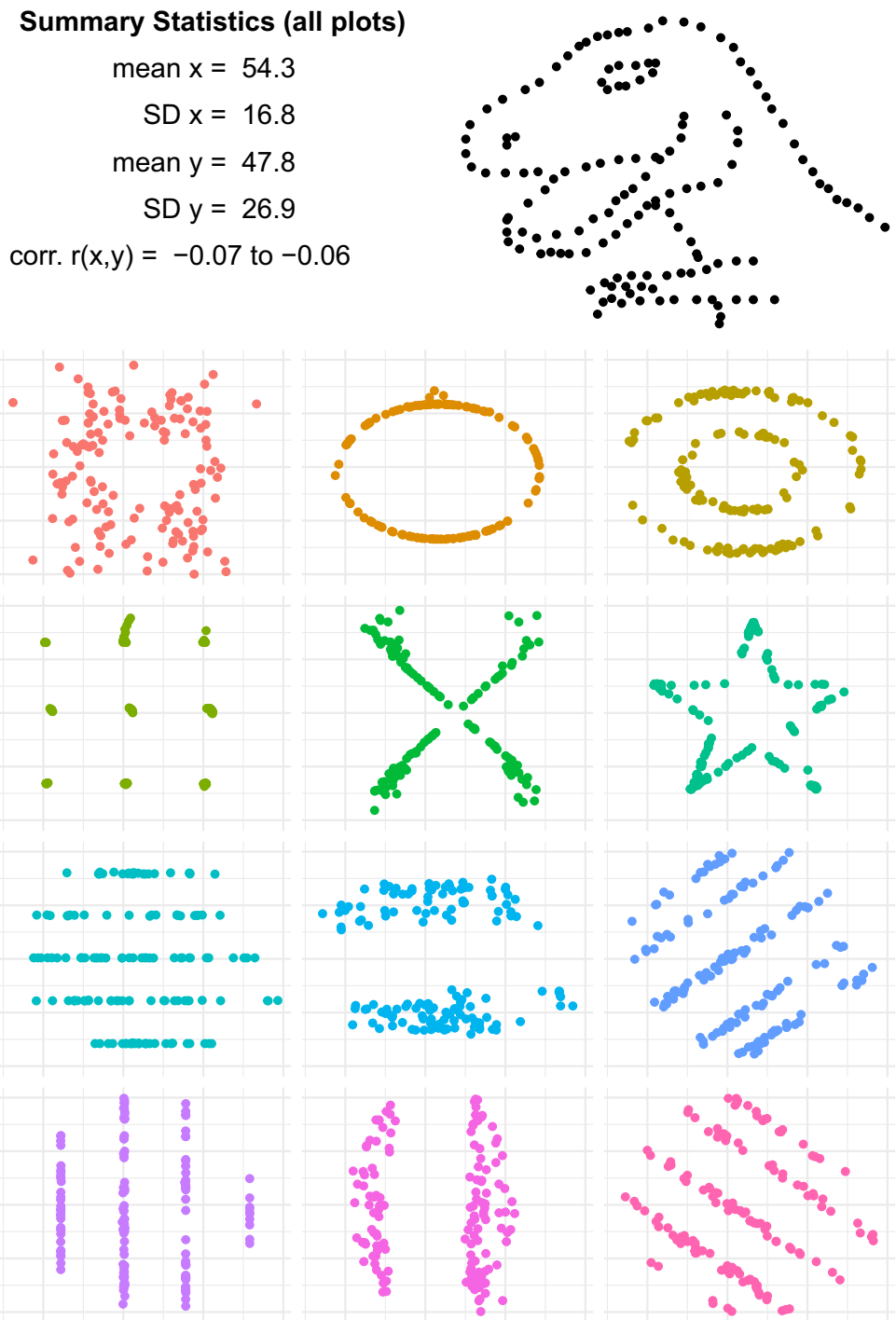


Fig. 11 The *Datasaurus dozen* (Matejka & Fitzmaurice, 2017). The datasets in all the plots—including the dinosaur—have functionally identical summary statistics.

Declaration

Competing interests The author reports that there are no competing interests to declare.

References

Alf, C., & Lohr, S. (2007). Sampling assumptions in introductory statistics classes. *American Statistician, 61*(1), 71–77. https://doi.org/10.1198/000313007X171098

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research, 23*(6), 727–744. https://doi.org/10.1177/1362168818767191

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 370–374. https://doi.org/10.1002/wics.84

Anderson, D. R., Link, W. A., Johnson, D. H., & Burnham, K. P. (2001). Suggestions for presenting the results of data analysis. *The Journal of Wildlife Management, 65*(3), 373–378.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21. https://doi.org/10.1007/978-3-540-71915-1_35

Barker, L. E., & Shaw, K. M. (2015). Best (but oft-forgotten) practices: Checking assumptions concerning regression residuals. *American Journal of Clinical Nutrition, 102*(3), 533–539. https://doi.org/10.3945/ajcn.115.113498

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Bilon, X. J. (2021). Normality and significance testing in simple linear regression model for large sample sizes: A simulation study. *Communications in Statistics: Simulation and Computation*, Advance online publication. https://doi.org/10.1080/03610918.2021.1916824

Bishara, A. J., Li, J., & Conley, C. (2021). Informal versus formal judgment of statistical models: The case of normality assumptions. *Psychonomic Bulletin and Review, 28*(4), 1164–1182. https://doi.org/10.3758/s13423-021-01879-z

Bordacconi, M. J., & Larsen, M. V. (2014). Regression to causality: Regression-style presentation influences causal attribution. *Research and Politics, 1*(2), 1–6. https://doi.org/10.1177/2053168014548092

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.28920/dhm51.2.230

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brooks, M. E., Kristensen, K., Benthem, K. J. Van, Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brooks, M. E., Kristensen, K., Darrigo, M. R., Rubim, P., Uriarte, M., Bruna, E., & Bolker, B. M. (2019). Statistical modeling of patterns in annual reproductive rates. *Ecology, 100*(7), 1–7. https://doi.org/10.1002/ecy.2706

Brysbaert, M., Bakk, Z., Buchanan, E. M., Drieghe, D., Frey, A., Kim, E., Kuperman, V., Madan, C. R., Marelli, M., Mathôt, S., Svetina Valdivia, D., & Yap, M. (2020). Into a new decade. *Behavior Research Methods, 53*, 1–3. https://doi.org/10.3758/s13428-020-01497-y

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367*(1906), 4361–4383. https://doi.org/10.1098/rsta.2009.0120

Cairo, A. (2016). *Download the Datasaurus: Never trust summary statistics alone; always visualize your data*. https://web.archive.org/web/20220728213556/http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

Casson, R. J., & Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clinical and Experimental Ophthalmology, 42*(6), 590–596. https://doi.org/10.1111/ceo.12358

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika, 70*(1), 1–10. https://doi.org/10.1093/biomet/70.1.1

Cook, R. D., & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. John Wiley & Sons. https://doi.org/10.1002/9780470316948.ch14

Correll, M., Li, M., Kindlmann, G., & Scheidegger, C. (2019). Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 830–839. https://doi.org/10.1109/TVCG.2018.2864907

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Cribari-Neto, F., & Zarkos, S. G. (1999). Bootstrap methods for heteroskedastic regression models: Evidence on estimation and testing. *Econometric Reviews, 18*(2), 211–228. https://doi.org/10.1080/07474939908800440

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5–12. https://doi.org/10.11648/j.ajtas.20160501.12

Davies, R., Locke, S., & McGowan, L. D. (2022). *datasauRus: Datasets from the Datasaurus Dozen (0.1.6)*. R package.

Di Leo, G., & Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental, 4*, Article 18. https://doi.org/10.1186/s41747-020-0145-y

Dixon, D. (2012). Analysis tool or research methodology: Is there an epistemology for patterns? In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 191–209). Palgrave Macmillan. https://doi.org/10.1057/9780230371934_11

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice-a systematic review of common misconceptions. *PeerJ, 5*, e3323. https://doi.org/10.7717/peerj.3323

Fagerland, M. W. (2012). T-tests, non-parametric tests, and large studies—A paradox of statistical practice? *BMC Medical Research Methodology, 12*(1), 78.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models* (2nd ed.). CRC Press (Taylor & Francis Group).

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21*, 111–123. https://doi.org/10.4018/978-1-7998-8409-5.ch001

Flatt, C., & Jacobs, R. L. (2019). Principle assumptions of regression analysis: Testing, techniques, and statistical reporting of imperfect data sets. *Advances in Developing Human Resources, 21*(4), 484–502. https://doi.org/10.1177/1523422319869915

Forthmann, B., & Doebler, P. (2021). Reliability of researcher capacity estimates and count data dispersion: A comparison of Poisson, negative binomial, and Conway-Maxwell-Poisson models. *Scientometrics, 126*(4), 3337–3354. https://doi.org/10.1007/s11192-021-03864-8

Fox, J. D. (2022). *Regression diagnostics*. Sage. https://doi.org/10.4135/9781071878651

Gel, Y., Miao, W., & Gastwirth, J. L. (2005). The importance of checking the assumptions underlying statistical analysis: Graphical methods for assessing normality. *Jurimetrics, 46*, 3–29.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gelman, A., Hill, J., & Vehtari, A. (2022). *Regression and other stories*. Cambridge University Press.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Gnanadesikan, R. (1997). *Methods for statistical analysis of multivariate data* (2nd ed.). Wiley.

Green, J. A. (2021). Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychology and Behavioral Medicine, 9*(1), 436–455. https://doi.org/10.1080/21642850.2021.1920416

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters, 15*(5), 20190174. https://doi.org/10.1098/rsbl.2019.0174

Halunga, A. G., Orme, C. D., & Yamagata, T. (2017). A heteroskedasticity robust Breusch–Pagan test for Contemporaneous correlation in dynamic panel data models. *Journal of Econometrics, 198*(2), 209–230. https://doi.org/10.1016/j.jeconom.2016.12.005

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ, 2*, e616. https://doi.org/10.7717/peerj.616

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models*. R package https://cran.r-project.org/package=DHARMa

Hawkins, D. M. (1991). Diagnostics for use with regression recursive residuals. *Technometrics, 33*(2), 221–234. https://doi.org/10.1080/00401706.1991.10484809

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*(4), 709–722. https://doi.org/10.3758/BF03192961

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*, Article 137. https://doi.org/10.3389/fpsyg.2012.00137

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.1198/jasa.2003.s281

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research, 37*(1), 171–184. https://doi.org/10.1177/0267658319877433

Hullman, J., & Gelman, A. (2021). Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review, 3*(3). https://doi.org/10.1162/99608f92.3ab8a587

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods, 48*(4), 1308–1317. https://doi.org/10.3758/s13428-015-0624-x

Kianifard, F., & Swallow, W. H. (1996). A review of the development and application of recursive residuals in linear models. *Journal of the American Statistical Association, 91*(433), 391–400. https://doi.org/10.1080/01621459.1996.10476700

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods, 53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *Journal of Agronomy and Crop Science, 204*(1), 86–98. https://doi.org/10.1111/jac.12220

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science, 16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48*(2), 360–388. https://doi.org/10.1002/tesq.128

Loy, A. (2021). Bringing visual inference to the classroom. *Journal of Statistics and Data Science Education, 29*(2), 171–182. https://doi.org/10.1080/26939169.2021.1920866

Lüdecke, D., Ben-shachar, M. S., Patil, I., Makowski, D., Waggoner, P., Patil, I., Ben-shachar, M. S., Patil, I., & Makowski, D. (2021). Assessment of regression models performance. *The Journal of Open Source Software, 6*(59), 1–8. https://doi.org/10.21105/joss.03132

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*(1), 151–169. https://doi.org/10.1146/annurev.publheath.23.100901.140546

Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association, 108*(503), 942–956. https://doi.org/10.1080/01621459.2013.808157

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–1294. https://doi.org/10.1145/3025453.3025912

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Nielsen, E. E., Nørskov, A. K., Lange, T., Thabane, L., Wetterslev, J., Beyersmann, J., De Unã-Álvarez, J., Torri, V., Billot, L., Putter, H., Winkel, P., Gluud, C., & Jakobsen, J. C. (2019). Assessing assumptions for statistical analyses in randomised clinical trials. *BMJ Evidence-Based Medicine, 24*(5), 185–189. https://doi.org/10.1136/bmjebm-2019-111174

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology, 3*, 1–5. https://doi.org/10.3389/fpsyg.2012.00322

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Oldford, R. W. (2016). Self-calibrating quantile–quantile plots. *The American Statistician, 70*(1), 74–90. https://doi.org/10.1080/00031305.2015.1090338

Osborne, J. W., & Waters, E. (2003). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation, 8*(2), 1–5. https://doi.org/10.7275/r222-hv23

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*, 1–17. https://doi.org/10.3389/fpsyg.2018.02104

Pole, B. J. D., & Bondy, S. J. (2012). Normality assumption. In *Encyclopedia of research design* (pp. 932–934). SAGE. https://doi.org/10.4135/9781412961288

Poncet, A., Courvoisier, D. S., Combescure, C., & Perneger, T. V. (2016). Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. *Methodology, 12*(2), 61–71. https://doi.org/10.1027/1614-2241/a000110

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.r-project.org/

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology, 12*, Article 81. https://doi.org/10.1186/1471-2288-12-81

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(10), 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Rosopa, P. J., Schaffer, M. M., & Schroeder, A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods, 18*(3), 335–351. https://doi.org/10.1037/a0032553

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology, 98*, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steinskog, D. J., Tjøtheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Monthly Weather Review, 135*(3), 1151–1157. https://doi.org/10.1175/MWR3326.1

Tay, L., Parrigon, S., Huang, Q., & LeBreton, J. M. (2016). Graphical descriptives: A way to improve data transparency and methodological rigor in psychology. *Perspectives on Psychological Science, 11*(5), 692–701. https://doi.org/10.1177/1745691616663875

Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests does not suffice: A plea for plausibility. *Psychonomic Bulletin and Review, 25*(2), 548–559. https://doi.org/10.3758/s13423-018-1447-4

Troncoso Skidmore, S., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods, 45*(2), 536–546. https://doi.org/10.3758/s13428-012-0257-2

Vallejo, G., Fernández, M. P., & Rosário, P. (2021). Combination rules for homoscedastic and heteroscedastic MANOVA models from multiply imputed datasets. *Behavior Research Methods, 53*(2), 669–685. https://doi.org/10.3758/s13428-020-01429-w

Veldkamp, C. L. S. (2017). *The human fallibility of scientists* [Tilburg University]. https://psyarxiv.com/g8cjq/

Verbeek, M. (2008). *A guide to modern econometrics* (2nd ed.). John Wiley & Sons.

Waldman, D. M. (1983). A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. *Economics Letters, 13*(2–3), 197–200. https://doi.org/10.1016/0165-1765(83)90085-X

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). John Wiley & Sons.

Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., & Milic, N. M. (2016). From static to interactive: Transforming data visualization to improve transparency. *PLoS Biology, 14*(6), 1–8. https://doi.org/10.1371/journal.pbio.1002484

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology, 7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 973–979. https://doi.org/10.1109/TVCG.2010.161

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge. https://doi.org/10.4324/9781315165547

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x

Open practices statement All the paper's materials (i.e., code and data) are available in the following Open Science Framework (OSF) repository: https://doi.org/10.17605/OSF.IO/59B68

None of the experiments were preregistered.