Accepted: 5 May 2017

DOI: 10.1111/jac.12220

MISCELLANEOUS

WILEY Journal of Agronomy and Crop Science

What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions

M. Kozak[1,2] | H.-P. Piepho[3]

[1]Department of Quantitative Methods in Economics, University of Information Technology and Management in Rzeszów, Rzeszów, Poland

[2]Department of Botany, Warsaw University of Life Sciences—SGGW, Warsaw, Poland

[3]Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

Correspondence
M. Kozak, Department of Quantitative Methods in Economics, University of Information Technology and Management in Rzeszow, Rzeszow, Poland
Email: nyggus@gmail.com

Abstract

We consider two questions important for applying analysis of variance (ANOVA): Should normality be checked on the raw data or on the residuals (or is it immaterial which of the two approaches we take)? Should normality and homogeneity of variance be checked using significance tests or diagnostic plots (or both)? Based on two examples, we show that residuals should be used for model checking and that residual plots are better for checking ANOVA assumptions than statistical tests. We also discuss why one should be very cautious when using statistical tests to check the assumptions.

KEYWORDS
ANOVA, diagnostic plots, linear models

1 | INTRODUCTION

Analysis of variance (ANOVA) is a procedure developed at the beginning of the 20th century by Sir Ronald Fisher to analyse treatment mean differences. Fisher worked with agricultural experiments, and now, around 100 years later, ANOVA is crucial for most agronomic research (McIntosh, 2015). Of course, it has also been used in numerous other scientific disciplines and non-scientific applications outside of agriculture.

It is because of this phenomenal popularity that ANOVA is applied in various ways. According to McIntosh (2015), there is nothing wrong with this as "the researcher and frequently an applied statistician play important roles in ANOVA because individual choices can shape the analysis and interpretation of results." However, statisticians and researchers take multiple—sometimes opposite—points of view on how ANOVA should be applied, which leads to methodological discussions and development of the method. Far too often such methodological discussions and recommendations go unnoticed by those who apply ANOVA, likely because of their statistical difficulty and the outlets in which they are published.

Most statistical methods have assumptions; ANOVA has them too. Thus, when applying ANOVA, one should check their validity. Different approaches have been proposed for this purpose, so users need to make some choices. Two of the most important choices can be phrased as the following questions:

Question 1 Should normality be checked on the raw data or on the residuals (or does it matter at all which of the two approaches we take)?

Question 2 Should normality and homogeneity of variance be checked using significance tests or diagnostic plots (or both)?

Our experience is that the way the user answers these questions affects the way he or she decides whether or not the assumptions are met. In this study, we argue that normality is best checked based on residuals (our answer to Question 1). We will demonstrate that using the raw data, a very common practice, can lead to an overly optimistic assessment of normality, or, conversely, may lead to rejection of normality when, in fact, the assumption is met. We further hope to convince readers that diagnostic plots are more helpful than significance tests in checking assumptions and identifying the most suitable remedy in case violations are identified (our answer to Question 2). If you agree with us, this is the time to put the paper down and get on with what's next on your list. Otherwise, do read on!

In what follows, we assume that the observations are independent, which is one of the most important ANOVA assumptions; this assumption is justified with proper randomization (Casler, 2015; Piepho, Möhring, & Williams, 2013). Thus, we will limit our discussion to classical ANOVA scenarios, in which there is one dependent variable, one or more treatment factors, and all observations are indeed

independent of each other (which must follow from the experimental design). Hence, the discussion will not cover repeated measures and other designs with correlated observations, although our main statements essentially carry over to these settings as well. Moreover, we assume for simplicity that heterogeneity of variance is not a problem (but see, e.g. Carroll and Ruppert (1988) or Piepho (2009) for possible remedies in case of heterogeneity).

The paper is organized as follows. First, we will present two motivating examples in which we show common ways of checking ANOVA assumptions. Then, we will reply to Questions 1 and 2 posed above. This will lead us to revise the analyses presented in these two examples. In the last section, we will summarize the whole discussion. All analyses are conducted and all graphs are created in R (R Core Team 2016).

Data analysis should always be a sequence of attempts—with the help of various tools—to understand the data. Thus, do not jump to conclusions after our first attempts to analyse the data in the examples; they can later turn out to be wrong! In fact, be aware that we will start by presenting initial analyses and conclusions that suggest themselves, but later will revise the initial conclusions in the light of further scrutiny. These initial conclusions are opposite to our final conclusions, which will reflect our final point of view.

2 | EXAMPLE 1

Let us consider an artificial data set presented in Table 1 and assume that the data come from a fully randomized experiment in a one-way layout with three treatments (A1, A2, A3), with 15 replications per treatment.

Before applying ANOVA, we may want to check whether the dependent variable is approximately normally distributed; thus, the null hypothesis is that the dependent variable follows a normal distribution. To test this hypothesis, we here use the Shapiro–Wilk test for the raw data comprising all 45 observations. The result is $W = 0.784$ with p-value $< .001$. This is worrying because the test rejects the null hypothesis of normality. So it seems we are led to conclude that the dependent variable is not normally distributed, and therefore, we should not proceed with ANOVA for these data. Now let's look at the normal quantile–quantile (QQ) plot of the dependent variable (Figure 1). A QQ plot is a scatterplot of two sets of quantiles plotted against one another (Atkinson, 1987); in the normal QQ plot, one of the sets (on the x-axis in Figure 1) comes from the theoretical normal distribution whereas the other (on the y-axis) represents the data. If the points form a roughly straight line, then both sets of quantiles came from the same distribution (or very similar distributions). In Figure 1, we do not see this pattern, and so the distribution of the dependent variable does not look normal at all, which agrees with the result of the Shapiro–Wilk test.

Conclusion up to here: The data do not follow a normal distribution, so we can either use nonparametric methods or look for a transformation to normalize the data.

TABLE 1 Artificial data used for Example 1

A1	A2	A3
23	20	43
21	19	41
26	21	44
20	23	48
22	22	43
28	22	41
23	22	41
22	22	43
22	21	42
19	23	43
23	24	39
24	23	41
25	25	45
21	26	46
22	28	41

3 | EXAMPLE 2

Now consider another artificial data set presented in Table 2. These data also are assumed to come from a fully randomized experiment in a one-way layout with three treatments (A1, A2, A3), but here with four replications per treatment.

Like in Example 1, we now check the assumptions. The Shapiro–Wilk test for all 12 observations gives the result of $W = 0.892$ with a p-value of $p = .127$. The normal QQ plot presented in Figure 2 suggests that the distribution of the data is close to normal. Bartlett's test for homogeneity of variance yields $K^2 = 5.77$ on two degrees of freedom, which gives $p = .056$. The p-value is close to the significance boundary of $\alpha = 0.05$, but still it seems reasonable not to reject the hypothesis that the variances are identical for the three treatments.

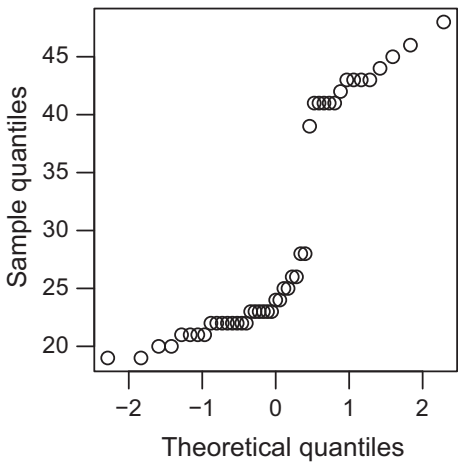

FIGURE 1 Normal QQ plot of raw data of the dependent variable from Example 1

Conclusion up to here: It seems that the assumptions of normality of the dependent variable and equality of variance for the three treatments are met: the *p*-values for the tests were close to the significance level, but did not reach it; the normal QQ plot showed that the residuals were close to being normally distributed. Thus, it appears that we can proceed with ANOVA.

TABLE 2 Artificial data used for Example 2

A1	A1	A3
40	20	50
20	30	60
30	40	70
120	50	80

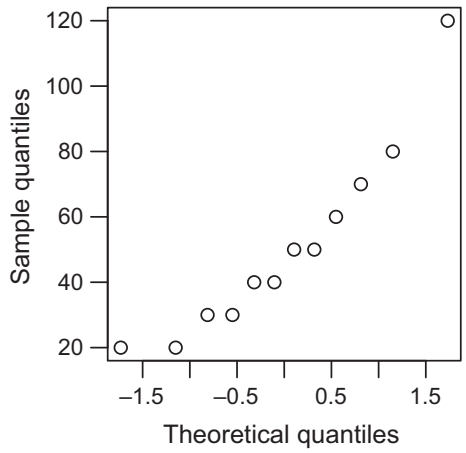

FIGURE 2 Normal QQ plot of raw data of the dependent variable from Example 2

4 | ANSWERING QUESTION 1

Question 1 Should normality be checked on the raw data or on residuals (or does it matter at all which of the two approaches we take)?

Let's recall the linear model for one-way ANOVA:

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j} \quad (1)$$

Here, $y_{i,j}$ is the *j*th observation ($j = 1,\ldots, J$ for balanced designs) for the *i*th treatment ($i = 1,\ldots,I$), μ is the intercept, τ_i is the effect of the *i*th treatment, and $\varepsilon_{i,j}$ is the error of the *j*th observation for the *i*th treatment. The distribution of the errors is normal, that is $\varepsilon \sim N(0, \sigma^2)$, σ^2 being the variance of ε. In a sample, these errors are estimated by so-called residuals:

$$\hat{\varepsilon}_{i,j} = y_{i,j} - \hat{\mu} - \hat{\tau}_i = y_{i,j} - \bar{y}_i, \quad (2)$$

where $\bar{y}_i$ is a sample mean of $y_{i,j}$ for the *i*th treatment. In many papers, we see sentences like "Normality of the variables was checked with the Shapiro-Wilk test." Now the question in such cases is: did the authors of such and similar sentences mean what they wrote? That is, did they use the Shapiro–Wilk test on the raw data (the original $y_{i,j}$ values of the variables to be analysed by ANOVA)?

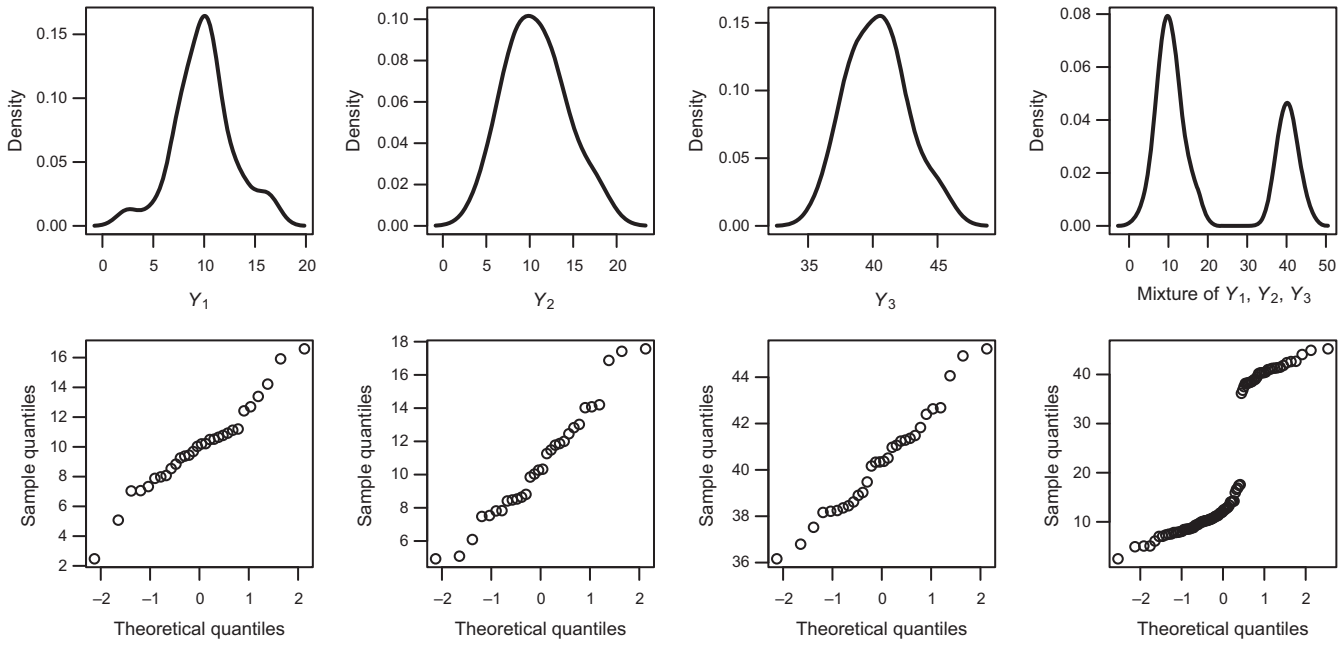

FIGURE 3 Density plots (top row) and normal QQ plots (bottom row) for three normal variables $Y_1 \sim N(10, 9)$, $Y_2 \sim N(10, 9)$, $Y_3 \sim N(40, 9)$ and a mixture of these three distributions (i.e. for the pooled data). Each of the three distributions considered independently is normal, but when we pool the data and consider it as one distribution (the mixture of distributions), it appears to be far from normal

Note that this is exactly what we did in the above analysis in Example 1.

Such an approach to checking data normality with the Shapiro–Wilk test actually tests the assumption that *all observations come from the same normal distribution*; if this is the case, then the treatments *do not affect the dependent variable*! If the treatments do affect the dependent variable, however, then under model (1) each treatment has its own normal distribution with a separate mean $\mu_i = \mu + \tau_i$.

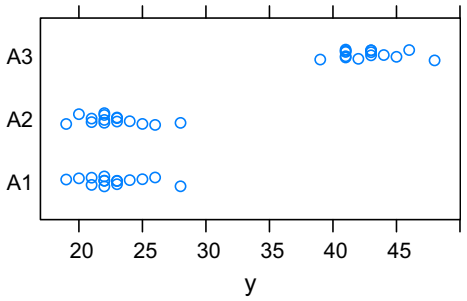

FIGURE 4 Data from Example 1 (see Table 1). Slight vertical jitter was added to visualize overlapping observations

The actual assumption is that the dependent variable should be normal *within each group* (each treatment); hence, in Example 1, we should assume that the dependent variable is normally distributed for each of the treatments A1, A2 and A3. This assumption is equivalent to the assumption of normality of errors from the linear model on which ANOVA is based. The key point here is that errors $\varepsilon_{i,j}$ of all treatments are assumed to come from the same normal distribution with zero mean and variance σ^2, whereas the observed data $y_{i,j}$ may come from different normal distributions when there are mean differences between treatments. Thus, if the means are different and we pool the data from different groups, the normal QQ plot will perceive the data set as a mixture of three normal distributions. Even though the data for each treatment have a perfect normal distribution, such a mixture may actually look very non-normal (Figure 3)! So, obviously a QQ plot *for the raw data* is not appropriate to check whether there is a problem with the normality assumption in ANOVA. Rather, *we should check normality of the errors, not of the observed data*. If the errors are normally distributed, then the

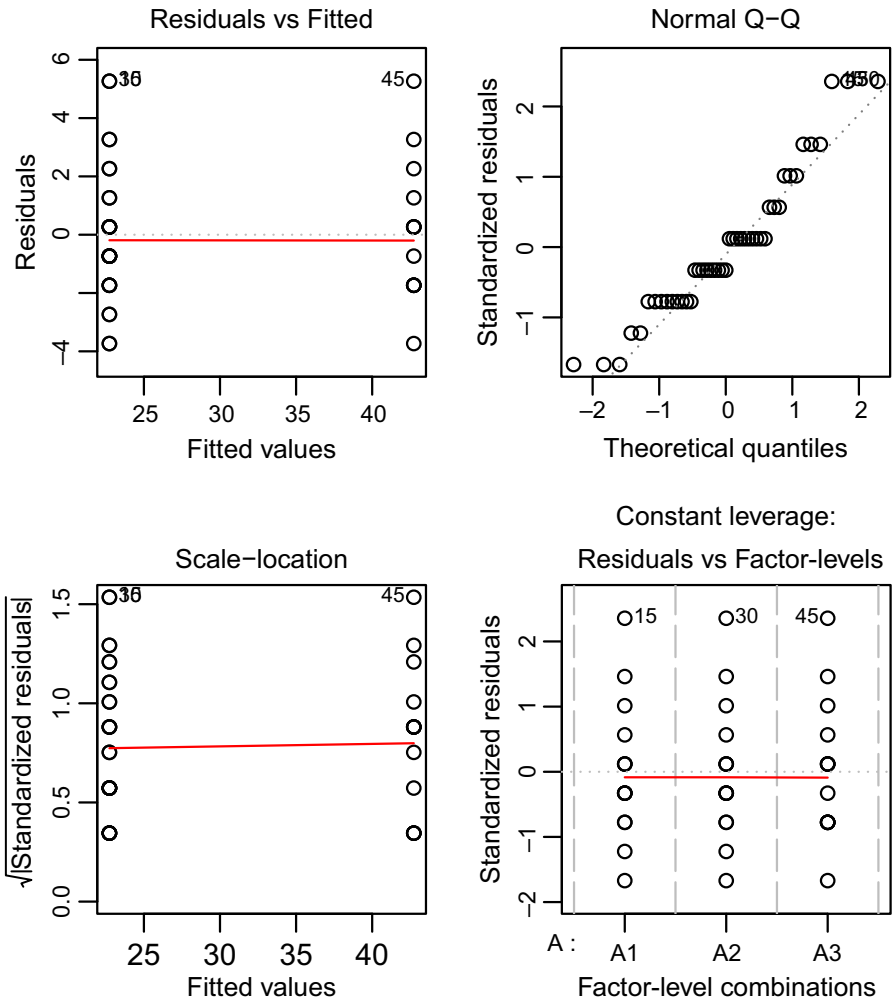

FIGURE 5 Residual plots for one-way ANOVA of data in Example 1

dependent variable within each treatment is normally distributed, too. The errors cannot be directly observed, but are estimated by the residuals $\hat{\varepsilon}_{i,j}$; these residuals can be used for checking normality. This is the best way to check this assumption, also better than checking the normality of the observed data within each group (in Example 1, this would mean independently checking of normality of the dependent variable for each treatment). Checking residuals will be more powerful because of a greater sample size compared to checking the observed data within each group. An additional benefit of analysing residuals is that such an analysis offers information on much more than just their normality: one can also check their homogeneity of variance (we will refer to this when answering the Question 2) and look for untypical observations.

5 | EXAMPLE 1 REVISITED

What was wrong with the way the assumptions were checked in Example 1? From the section "Answering Question 1," it is clear that we should not have ignored the treatment structure of the design, as we did when checking normality of the distribution of the raw data for the dependent variable. Instead, we should have checked normality of the distribution of the residuals from the fitted model.

A first stage of any analysis, however, should be exploring the data. Our first mistake was not to do so. Figure 4 shows the data, with the response plotted on the *x*-axis and the treatment labels on the *y*-axis (which has no quantitative scale here). It is immediately apparent (which was not so easy to see from Table 1) that the data for treatments A1 and A2 are the same or very similar (actually, here they indeed are exactly the same). We also immediately see that the data from treatment A3 are much higher in value. The within-treatment variances are very similar for all three treatments.

Next, we fit the model (1) and check the residuals. Figure 5 represents a standard set of diagnostic plots for checking the fit of a linear model; these plots are based on residuals from our model. Clearly, everything is fine with the residuals: the QQ plot (top right in Figure 4) does not show any deviations from the normal distribution! When it comes to checking normality, we need to be clear that we will never be able to check whether indeed the distribution is fully normal; instead, we should check whether it is *approximately normal*; DeVeaux, Velleman, and Bock (2008) call this the "Nearly Normal Condition," in which "the shape of the data's distribution is unimodal and symmetric."

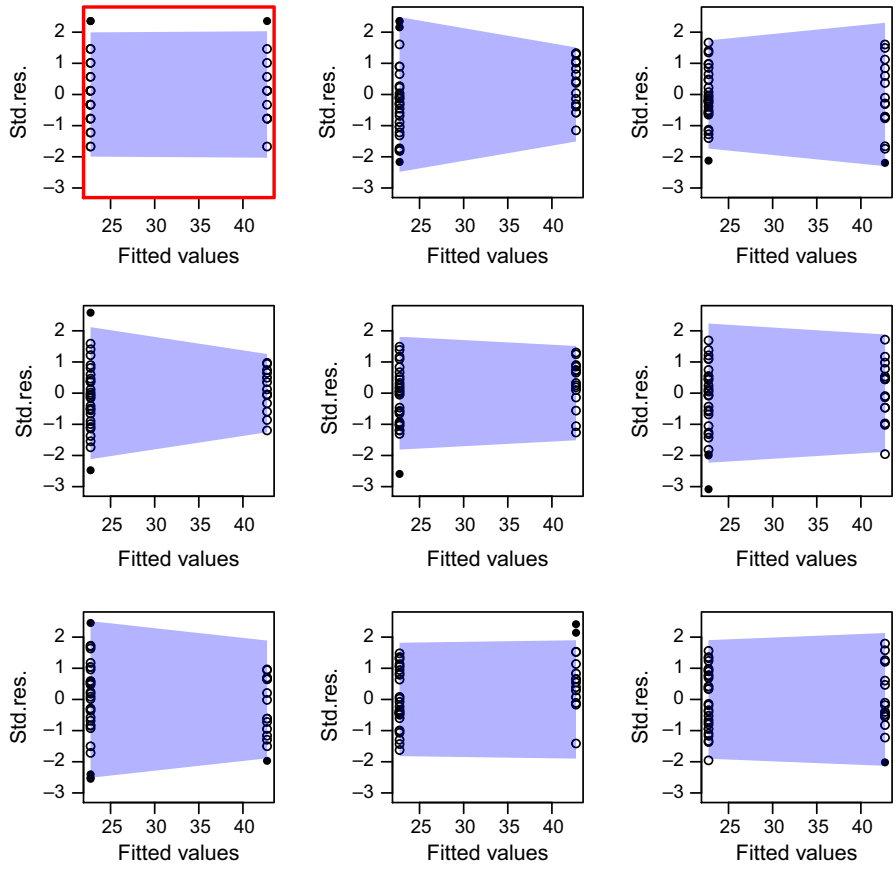

FIGURE 6 Wally plot based on the plot of standardized residuals versus values fitted by the model in Example 1. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript). The highlighted area indicates a local estimate of 1.96 of the standard deviation of the standardized residuals. The filled points are the outside± 1.96 of standard deviation of the residuals

In all these plots, we are looking for any deviations; here, there is nothing irregular or untypical. The residuals-versus-fitted value plot shows constant variances of residuals. In particular, the variance of the residuals is not related to treatments; the same is seen in the scale-location plot. The residuals-versus-treatments plot does not show anything atypical either. We might even say that these residuals are an example of the perfect residuals. Our initial conclusion was that one should not proceed with ANOVA because of lack of normality of the dependent variable was unwarranted. This conclusion was unwarranted.

Residuals can be checked with another interesting tool, the Wally plot (Ekstrøm, 2014). This type of plot can use any of the diagnostic plots (e.g., any of those in Figure 4). The idea is simple: first, create a sequence of diagnostic residual plots in which one of the plots is obtained from the model fitted to our data while the remaining ones are based on new data simulated from the fitted model under valid assumptions; randomize the order of plots, so you do not know which one of them is the actual plot for your data. In R, this can be done with the wallyplot function in the MESS package (Ekstrøm, 2016). Now guess which of the plots shows your data. While guessing, try to find a plot that seems the most untypical. If something is wrong with the fitted model, then one of the residual plots from the Wally plot will show this: this will be the plot for your data. If not, then probably everything is fine. Ekstrøm (2014) presented this technique for statistics teachers to use with students, but we believe it can be used with benefit by anyone fitting linear models.

Note that Wally plots can be produced for any diagnostic plot. Figure 6 shows the Wally plot based on the plot of standardized residuals versus fitted values while Figure 7 is based on the normal QQ plot. In both figures, no plot draws our attention with anything grossly untypical (although in the normal QQ plot from our data set, we see the specific pattern that is due to the discreteness of the data). The fitted model thus seems to meet the assumptions.

Conclusion revisited: The assumptions seem to be met, and we can proceed with ANOVA.

6 | ANSWERING QUESTION 2

Question 2 Should normality and homogeneity of variance be checked using significance tests or diagnostic plots (or both)?

Employing statistical tests is perhaps the most common way of checking the assumptions of normality and homogeneity of within-group variances. We followed this approach in our initial analysis for both Examples 1 and 2. For normality, this is often done for the observed data, as was discussed in Question 1 and shown in the examples; sometimes normality of the dependent variable within

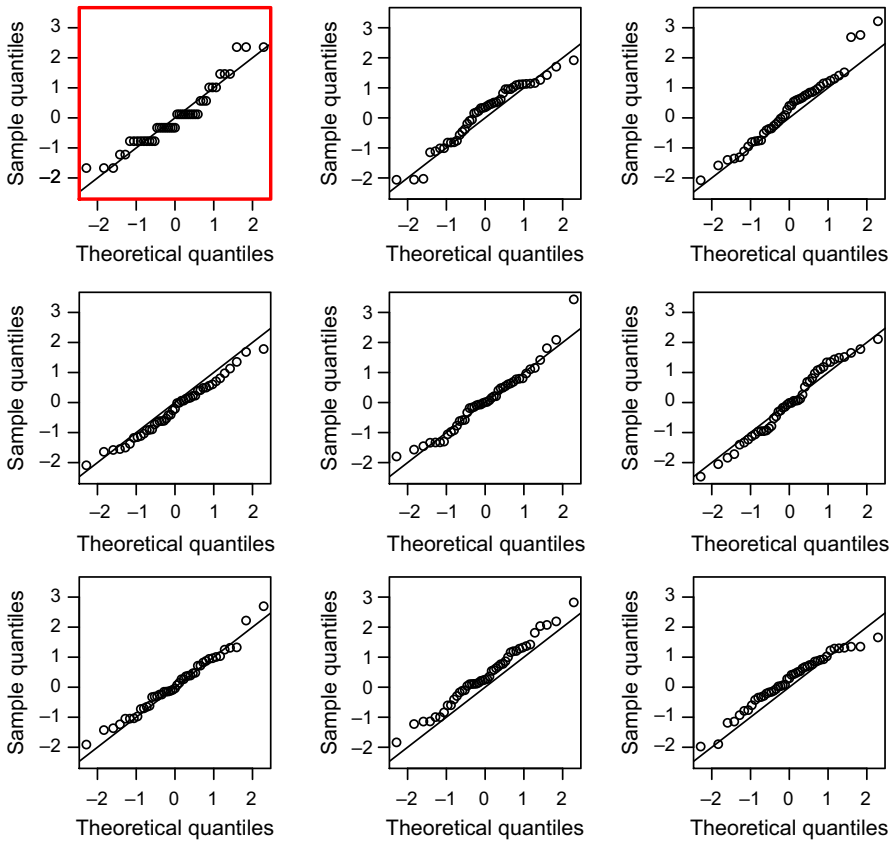

FIGURE 7 Normal QQ Wally plot for the model fitted in Example 1. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript)

groups or normality of residuals is tested. The Shapiro–Wilk test is commonly used for checking normality while for homogeneity of variances, Bartlett's and Levene's tests are probably the most commonly used, whereas the Brown–Forsythe test is a less popular alternative. For two groups, the F test is often used.

In our opinion, there are better ways of checking these assumptions than by statistical tests. This is because of the way statistical tests work. Here, the null hypothesis is as follows:

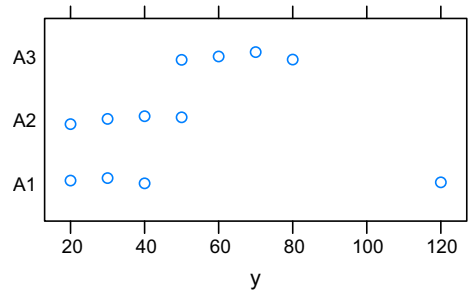

FIGURE 8 Data from Example 2 (see Table 2)

H_0: the assumption is met

Note this is a general formulation of H_0 that works for both assumptions (normality, homogeneity of variance). When this hypothesis is rejected, people often conclude that the assumption is violated. If the hypothesis is not rejected, the frequent conclusion is that the assumption is met, and one can follow with ANOVA. First of all, recall that not rejecting the null does not prove the null is true. Or as Altman and Bland (1995) put it, "Absence of evidence is not evidence of absence." Moreover, there is something untypical about this null hypothesis: what we actually prefer here is *not* rejecting the null hypothesis. This is untypical because we are used to testing null hypotheses that we prefer to be untrue.

The following recalls and widens the discussion by Kozak (2009). A Type I error occurs when one incorrectly rejects a true null hypothesis. If we use the 0.05 significance level (standard in agricultural studies), then we accept a 5% possibility of deciding that the assumption is violated when in fact it is not. A Type II error occurs when one incorrectly retains a false null hypothesis, so deciding that the assumption is met while in fact it is not. The

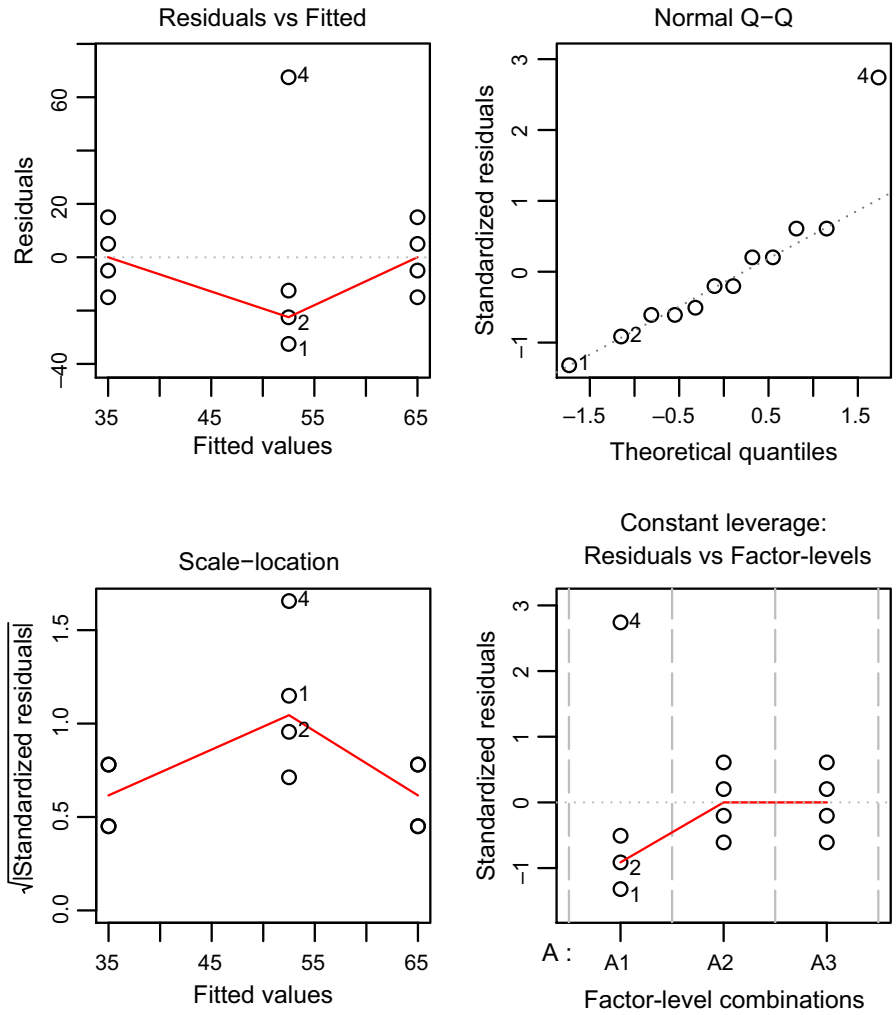

FIGURE 9 Diagnostic plots for one-way ANOVA of data in Example 1

power of a statistical test is the probability of a correct rejection of the null hypothesis. So the power is equal to (1—Type II error probability). Usually, we do not know the probability of a Type II error, but what we do know is that the smaller the sample size, the higher the Type II error rate and the smaller the power of the test.

And now comes the problem: not rejecting H_0 is a decision that most of us would be happy to see, but because of the above-mentioned characteristics of a Type II error, there are great chances that with small samples this decision is due to a Type II error. As Faraway (2002) puts it, "Even small differences from zero [departures from the null hypothesis] will be detected with a large sample. Now if we fail to reject the null hypothesis, we might simply conclude that the data were not enough to get a significant result. According to this view, the hypothesis test just becomes a test of sample size." Bottom line: just make the sample as small as possible if you want to convince (or mislead!) yourself that the assumptions are met. Zeileis and Hothorn (2002) add, "we will always be able to reject the null hypothesis provided we have enough data at hand. The question is not whether the model is wrong (it always is!) but if the irregularities are serious." Note also that if we reject a null hypothesis, we are in a much stronger position to draw a conclusion than if we do not. Strictly speaking, when the null hypothesis is not rejected, the analysis is inconclusive, and we had better not conclude anything at all (Hsu, 1996). Note, however, that when one does not reject the H_0 about the assumption, then this is exactly the situation in which the analysis is inconclusive and we should *not* be sure of our decision!

Campbell, Thompson, Guy, McIntosh, and Glaz (2015) discuss dangers of abusing the orthodox approach to hypotheses testing. One of their comment is that researchers pay too much attention to the yes–no decision made based on such testing, forgetting about all that is around this decision. What we discuss here about employing hypothesis testing to checking ANOVA assumptions is parallel to the discussion by Campbell et al. (2015), and adds an important application in which use of hypothesis testing can be dangerous and misleading.

According to many authors (e.g., Atkinson, 1987; Belsley, Kuh, & Welsch, 2005; Kozak, 2009; Moser & Stevens, 1992; Quinn & Keough, 2002; Rasch, Kubinger, & Moder, 2011; Schucany & Ng, 2006), significance tests should not be used for checking assumptions. Diagnostic residual plots are a better choice. On such diagnostic plots, we can identify untypical observations, heterogeneous within-treatment variances, and lack of normality of the residuals (and thus, of the dependent variable within treatments). Below we will show how they work.

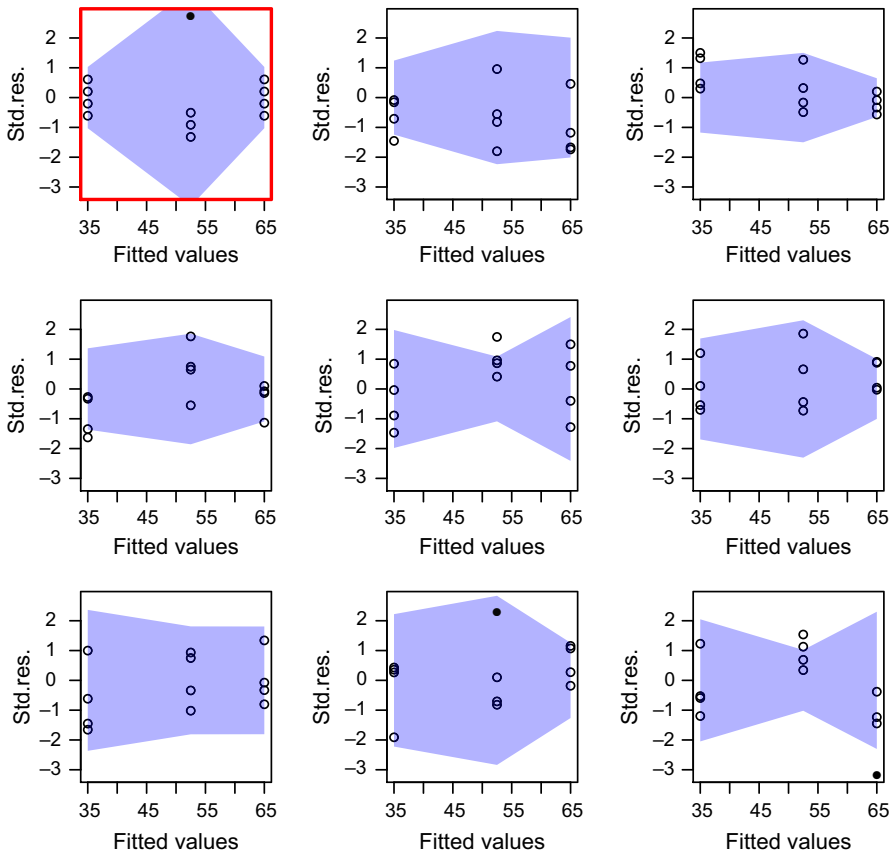

FIGURE 10 Wally plot based on the plot of standardized residuals versus the fitted values from the model in Example 2. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript)

7 | EXAMPLE 2 REVISITED

In Example 2, we checked the assumptions of normality of the dependent variable and variance homogeneity, and the initial decision was that the assumptions were met.

Like in Example 1 Revisited, we first examine the data with a simple plot (Figure 8). Evidently, something is wrong with the data: a large outlier can be seen for treatment A1, as high as 5-10 times higher than the other values for this treatment. How come, then, that we decided the ANOVA assumptions were met?

After answering Questions 1 and 2, it is clear what happened. First, we ignored the treatments when checking normality of the dependent variable using the raw data. And what we should have done is checking normality of the residuals from the model (1); Figure 9 shows the diagnostic plots for the model, based on the residuals. The normal QQ plot suggests that the residuals are not normally distributed. From each of these four diagnostic plots, we can see that one observation (labelled as "4," which means here it is a fourth row in the data table) is an outlier. Indeed, when we look at Table 2 and Figure 8, we see this outlying observation: the *y* values for treatment A1 are 20, 30 and 40, and this one is 120! Inspecting Figure 8, we might also ponder whether indeed the variances are equal. Let's have a look at them:

```
y variance for A1 is 2092
y variance for A2 is 167
y variance for A3 is 167
```

These variances do not look similar at all! What happened, then, with Bartlett's test, which did not reject the null hypothesis that the variances are the same? We know what happened: the sample size for each treatment is just 4, which does increase the type II error rate as compared to larger samples, and hence reduces the power to detect any departures. For larger samples, Bartlett's test probably would have rejected the null hypothesis that these variances are equal.

Now check the Wally plots in Figures 10 and 11. Here, we can detect the outlying observation in the observed data. But remember that in small samples, single observations are more likely than in larger samples to just appear like an outlier simply because the sample is too small to cover the whole distribution of the variable. However, the outlying observation for the fitted model in Figures 9-11 stands out so clearly that there will be little doubt it is a real outlier.

Now that we know we should not proceed with ANOVA, what to do next? The very first thing to do is inspect the outlying observation. Is it due to a mistake? It could be a data entry error (Kozak, Krzanowski, Cichocka, & Hartley, 2015 show how such mistakes can

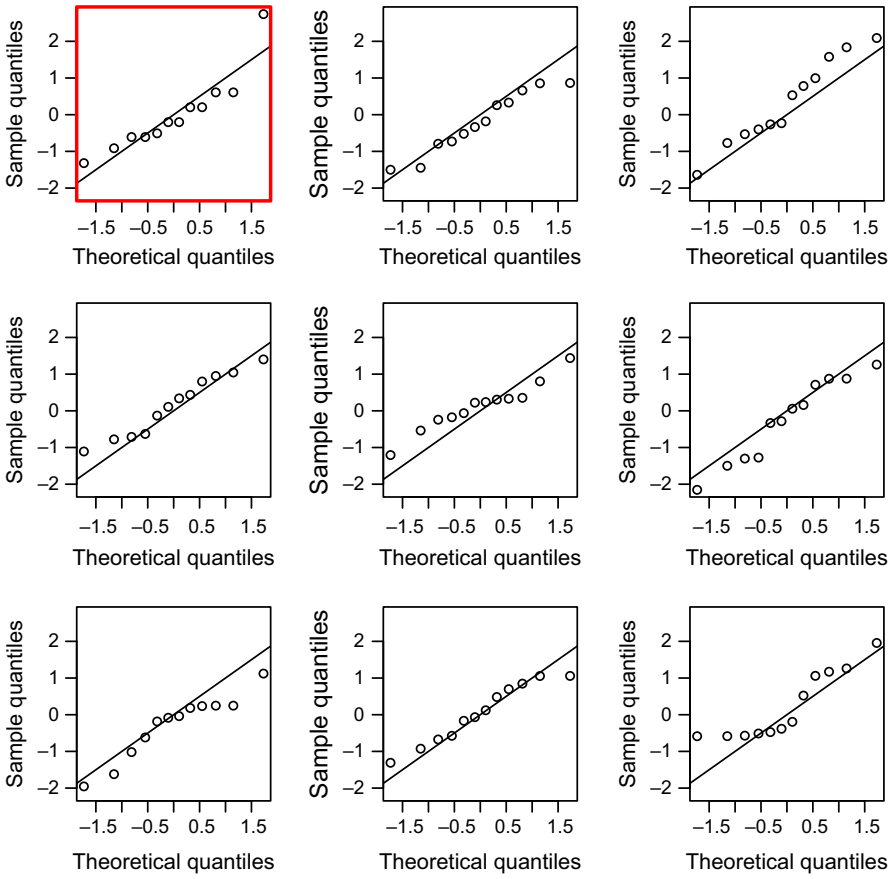

FIGURE 11 Normal QQ Wally plot for the model fitted in Example 2. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript)

affect statistical analysis) or a mistake in the experimental procedure. If we are sure, however, that such an outlier is *not* due to a mistake, then we should *not* merely remove it. Gotelli and Ellison (2004) stress this: "some researchers automatically delete outliers and extreme values prior to conducting their analyses. This is bad practice!" DeVeaux et al. (2008) claim that "Many outliers are not wrong: they're just different. Such cases often repay the effort to understand them. You can learn more from the extraordinary cases than from summaries of the overall data set."

Outliers deserve a separate discussion, and interested readers are referred to the cited references. The main point here is that when an outlier is detected and is not due to a mistake, something must be done. First, this outlier itself can be a source of information about the phenomenon (population) studied. However, sometimes a simple transformation will do the trick (i.e. the outlier will stop being an outlier), other times it should be removed from the analysis. When removing any outlier, one should explain why it was removed, why it could have occurred, what it might tell about the population and the phenomenon studied, and how our interpretation might change without this outlier.

Returning to the example, assume we find that it was a mistake in data entry, and instead of 120 it should be 20.

While previously the y variance in treatment A1 was much higher than in the other two treatments, now the situation is opposite. Will this be a problem? We will examine this from the diagnostic plots (Figures 12, 13, 14).

The model fits much better now. In difference to Figure 9, everything looks fine in these diagnostic plots. The same can be said about the Wally plots (Figures 13 and 14).

8 | CONCLUSION

At the beginning of the paper, we asked two questions about checking assumptions in ANOVA. The first one was, "Should normality be checked on the raw data or on residuals (or does it matter at all which of the two approaches we take)?" Our initial answer was that one should work with residuals and not with raw data; raw data are never a better choice. Both examples proved that based on raw data we can easily make incorrect decisions (e.g., deciding that the

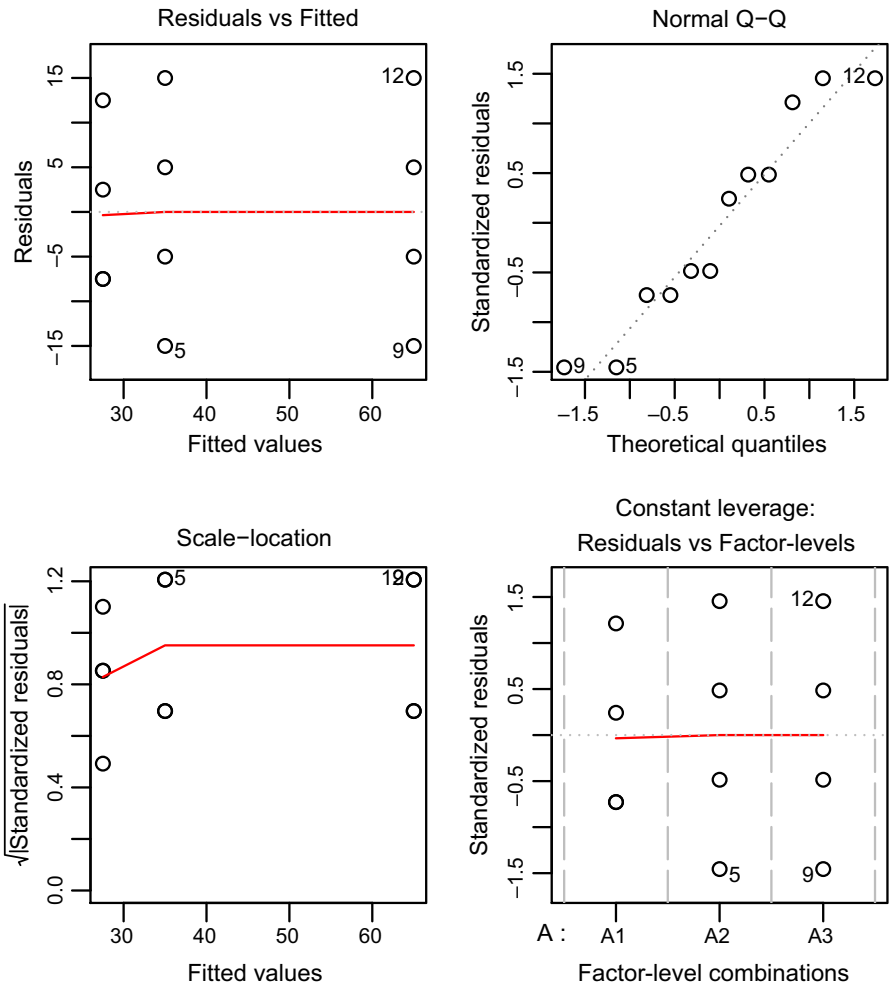

FIGURE 12 Diagnostic plots for one-way ANOVA of data in Example 1

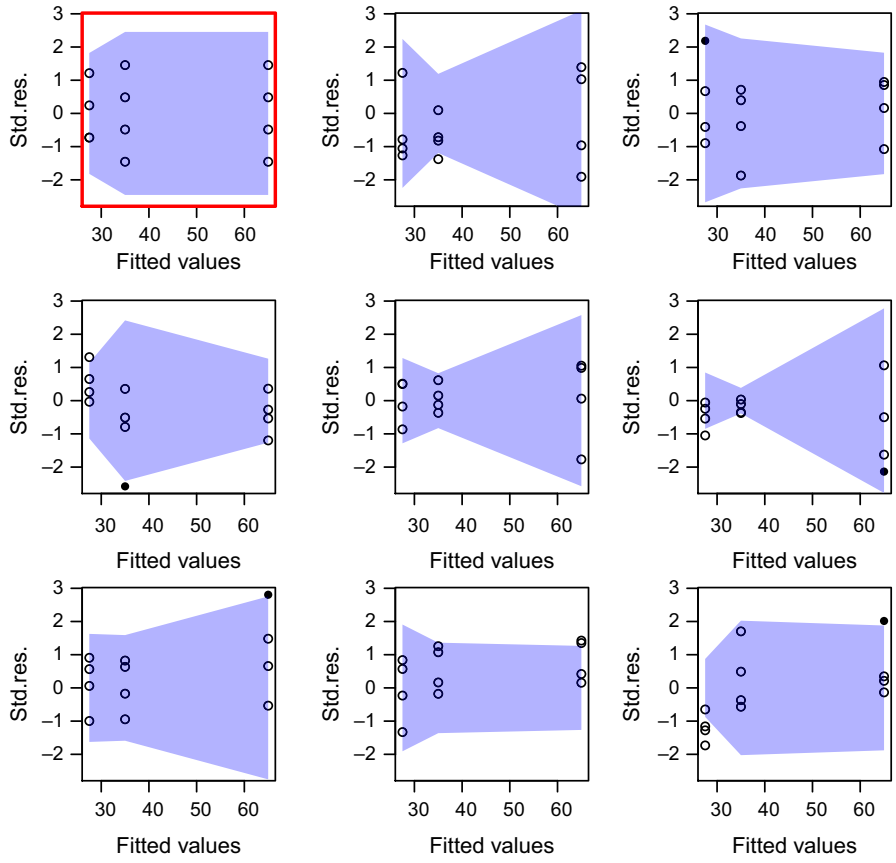

FIGURE 13 Wally plot based on the plot of standardized residuals versus the fitted values from the model in Example 2, after correcting the typing mistake for one of the observations. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript)

assumptions are not met even though they are, or vice versa) or fail to detect a serious outlier.

The second question was, "Should normality and homogeneity of variance be checked using significance tests or diagnostic plots (or both)?" Here, the answer might be seen not as clear as before, but we hope to have convinced you that diagnostic plots are a better choice. There are two possible reasons for the overuse of statistical tests to check assumptions. First, many researchers base their knowledge on books first published 40 years ago or earlier. Back then, using statistical tests was relatively simple while using diagnostic plots was difficult; thus, these books advised the former, often even not mentioning the latter. Second, most statistical software offers statistical tests for checking assumptions as a default. Using default tests is simple, so users use them. However, we explained why we think that significance tests are *not* a good way of checking assumptions (in general, not only for ANOVA). First of all, with large samples (a very desirable situation) we risk that even small (and irrelevant) departures from the null hypothesis (which states that the assumption is met) will be detected as significant, and so we would need to reject the hypothesis and state that the assumption is not met. With small samples, the situation is opposite: much larger (and important) departures would not be found significant. Thus, our advice is to use diagnostic plots instead of hypothesis testing to check ANOVA assumptions.

Using diagnostic plots, however, requires some expertise in reading and understanding them. Various textbooks on the topic (e.g. Atkinson, 1987; Atkinson & Riani, 2012; Fox, 1991) offer much help, but of course expertise mostly comes with practice and experience. As shown in the above examples, the Wally plot (Ekstrøm, 2014) is also an interesting graphical tool for checking assumptions of linear models. Another graphical tool to help interpret residual plots, not considered here, is confidence envelopes, which are tolerance bands around such plots (Atkinson, 1987; Schützenmeister, Jensen, & Piepho, 2012).

What also emerges from our examples is that one should *always* explore the data before analysing them. When we failed to do this initially in both examples, we missed important information about the data. When revisiting the examples, we checked the data graphically with simple plots (Figures 4, 8 and 15). A minute spent on making a simple plot can save a lot of effort and time later on: a first glance at Figure 8 hints that something might be wrong with the data. More difficult settings might require other graphical techniques, such as trellis displays (Cleveland & Fuentes, 1997) for factorial designs. We must remember, nonetheless, that such graphical techniques should not be used instead of diagnostic plots, but just as a means of exploring data before analysing them.

We worked with the simplest ANOVA situation, that is, a one-way balanced experiment in a completely randomized design. But

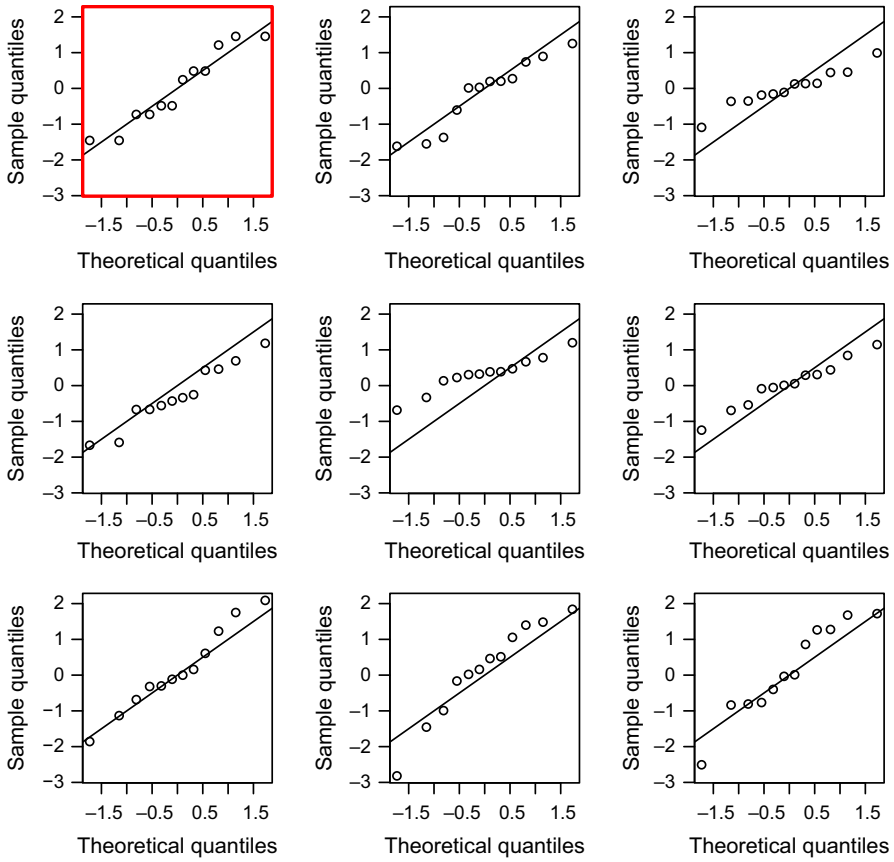

FIGURE 14 The normal QQ Wally plot for the model fitted in Example 2, after correcting the typing mistake for one of the observations. The plot for the observed data is highlighted with a thicker frame (red colour in the online version of the manuscript)

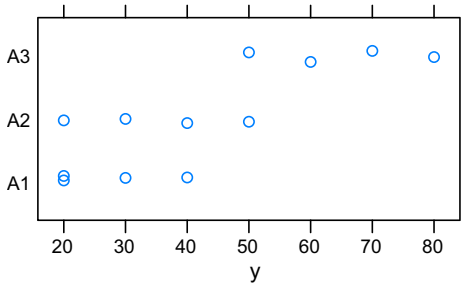

FIGURE 15 Data from Example 2 after correcting the outlier for treatment A1 (120 changed to 20). Slight vertical jitter was added to visualize overlapping observations

diagnostic plots are equally helpful for checking more complex designs and models, such as multifactor and blocked designs, also in unbalanced settings.

Various types of residuals and various graphs that use residuals exist (e.g. Atkinson, 1987; Fox, 1991). In ANOVA, we usually work with residuals to check normality. Generally, standardized or studentized residuals are better for checking assumptions than raw residuals, because the latter may display heterogeneous variance even when errors have constant variance (though for the special case of a balanced one-way both are equally suitable). The set of diagnostic plots used in Figures 5, 9 and 12 is a standard for balanced ANOVA. For unbalanced settings, however, sometimes the graph in the bottom right panel of these three figures is replaced with the leverage versus fitted values plot (e.g., in R it is a default graph for unbalanced designs, and one cannot graph the residuals versus factor levels graph). It is worth noting that diagnostic plots for linear regression are quite similar, but one should additionally analyse leverage plots.

To make valid decisions, we need to check assumptions. If the assumptions are not met, we should look for a remedy; if they are met, we can proceed with ANOVA. One should do whatever one can to ensure that the chosen statistical method can be used to analyse particular data. For ANOVA, diagnostic plots based on residuals from the fitted model are the best way of checking the assumptions.

ACKNOWLEDGEMENTS

We wish to thank two anonymous reviewers for helpful comments on the first version of the paper. In particular, one of the reviewers suggested two possible reasons for the overuse of statistical tests to check assumptions, reasons which we discuss in the last section.

REFERENCES

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485.

Atkinson, A. C. (1987). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Cambridge: Clarendon Press.

Atkinson, A. C., & Riani, M. (2012). *Robust diagnostic regression analysis*. New York: Springer Science & Business Media.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, Vol. *571*. John Wiley & Sons.

Campbell, K. G., Thompson, Y. M., Guy, S. O., McIntosh, M., & Glaz, B. (2015). "*Is*, or *is not*, the two great ends of Fate": Errors in agronomic research. *Agronomy Journal*, *107*, 718–729.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman & Hall.

Casler, M. D. (2015). Fundamentals of experiment design: Guidelines for designing successful experiments. *Agronomy Journal*, *107*, 692–705.

Cleveland, W. S., & Fuentes, M. (1997). Trellis Display: Modeling Data from Designed Experiments. Technical Report. BellLabs,Paris, France.

DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2008). *Intro Stats*, 3rd ed.. Addison Wesley.

Ekstrøm, C. T. (2014). Teaching 'instant experience' with graphical model validation techniques. *Teaching Statistics*, *36*, 23–26.

Ekstrøm, C.T. (2016). MESS: Miscellaneous esoteric statistical scripts. R package version 0.4-3. http://CRAN.R-project.org/package=MESS

Faraway, J.J. (2002). Practical regression and ANOVA using R. ftp://cran.r-project.org/pub/R/doc/contrib/Faraway-PRA.pdf

Fox, J. (1991). *Regression diagnostics: An introduction*, Vol. *79*. Newbury Park, CA: Sage.

Gotelli, N.J., Ellison, A.M. (2004). *A primer of ecological statistics*. Sinauer Associates, Inc. Publishers, Massachusetts, USA.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. London: Chapman & Hall.

Kozak, M. (2009). Analyzing one-way experiments: A piece of cake or a pain in the neck? *Scientia Agricola*, *66*, 556–562.

Kozak, M., Krzanowski, W., Cichocka, I., & Hartley, J. (2015). The effects of data input errors on subsequent statistical inference. *Journal of Applied Statistics*, *42*, 2030–2037.

McIntosh, M. S. (2015). Can analysis of variance be more significant? *Agronomy Journal*, *107*, 706–717.

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the 2-sample means test. *American Statistician*, *46*, 19–21.

Piepho, H. P. (2009). Data transformation in statistical analysis of field trials with changing treatment variance. *Agronomy Journal*, *101*, 865–869.

Piepho, H. P., Möhring, J., & Williams, E. R. (2013). Why randomize agricultural experiments? *Journal of Agronomy and Crop Science*, *199*, 374–383.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*, 219–231.

Schucany, W. R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics–Theory and Methods*, *35*, 2275–2286.

Schützenmeister, A., Jensen, U., & Piepho, H. P. (2012). Checking normality and homoscedasticity in the general linear model. *Communications in Statistics – Simulation and Computation*, *41*, 141–154.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. http://CRAN.R-project.org/doc/Rnews/

How to cite this article: Kozak M, Piepho H-P. What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *J Agro Crop Sci*. 2017;00:1–13. https://doi.org/10.1111/jac.12220