Checking the Assumptions of Your Statistical Model Without Getting Paranoid

Jan Vanhove

University of Fribourg

Word count: 3965

Last changed: September 12, 2018

Author Note

Acknowledgements: Isabelle Udry, Alistair Cullum, Audrey Bonvin, Guillaume Rousselet.

Jan Vanhove, Department of Multilingualism, University of Fribourg.

Correspondence concerning this article should be addressed to Jan Vanhove, Department of Multilingualism, University of Fribourg, Rue de Rome 1, 1700 Fribourg, Switzerland.

Contact: jan.vanhove@unifr.ch

Abstract

This tutorial shows how you can graphically check the assumptions behind the general linear model (which includes *t*-tests, ANOVA, and ANCOVA as well as simple and multiple regression) without succumbing to paranoia. The basic principle is that you draw diagnostic plots for the actual model as well as for models fitted to simulated data so that the model assumptions are known to be accurate. If you can tell the actual model's diagnostic plot apart from the others, then perhaps a robustness check or a more complex model is in order; if you cannot, then the diagnostic plot in and of itself is not a cause for concern. An online appendix shows how to implement this procedure with the help of user-friendly R functions.

Keywords: ANOVA, general linear model, line-up protocol, model assumptions, model diagnostics, regression, t-test, visual inference

Checking the Assumptions of Your Statistical Model Without Getting Paranoid

To make the jump from data to inferences, statistical tests and models rely on assumptions. For example, Student's *t*-test and (ordinary least squares) linear regression models assume that the data are 'homoskedastic'. This means that the differences between the mean trends and the actual observations are all assumed to have been sampled from a distribution with the same width (variance). If the data do not conform to the test's or model's assumptions, this may render invalid or irrelevant the inferences derived from them. For instance, Student's *t*-test may produce *p*-values that tend to be either too low or too high in the presence of non-constant variance or 'heteroskedasticity' (Glass et al., 1972), whereas ordinary least squares regression coefficient estimates tend to be more variable from sample to sample than they would have been if the non-constant variance had been accounted for.

Sundry tests exist to verify model assumptions, but these come with assumptions of their own. Moreover, they do not inform the analyst as to the nature of any assumption violations (e.g., in what respect are the residuals non-constant?) nor directly as to their degree. As a result, statisticians recommend graphical checks instead (e.g., Gelman & Hill, 2007; Zuur et al., 2009). To check for non-constant variance, for instance, analysts can plot the model's RESIDUALS against the FITTED VALUES or against the predictor values and gauge whether there are any patterns in the spread of the residuals.[1] As the left panels of Figure 1 show, such plots can be helpful for diagnosing flagrant violations of model assumptions.

1 An observation's fitted value is the outcome value you obtain if you pass its predictor values through the regression equation. For instance, if a regression model estimates the relationship between a second-language learner's age of acquisition and their performance on a grammar test to be such that that the score is 190 − 1.22 × (age of acquisition), then the fitted value for a learner whose age of acquisition is 14 is 190 − 1.22 × 14 = 173. This is also commonly referred to as an observation's "predicted value." An observation's residual is the difference between the actually observed outcome value and the corresponding fitted value according to the model. If an observed value is 156 and its fitted value is 173, then the residual is 156 − 173 = −17.

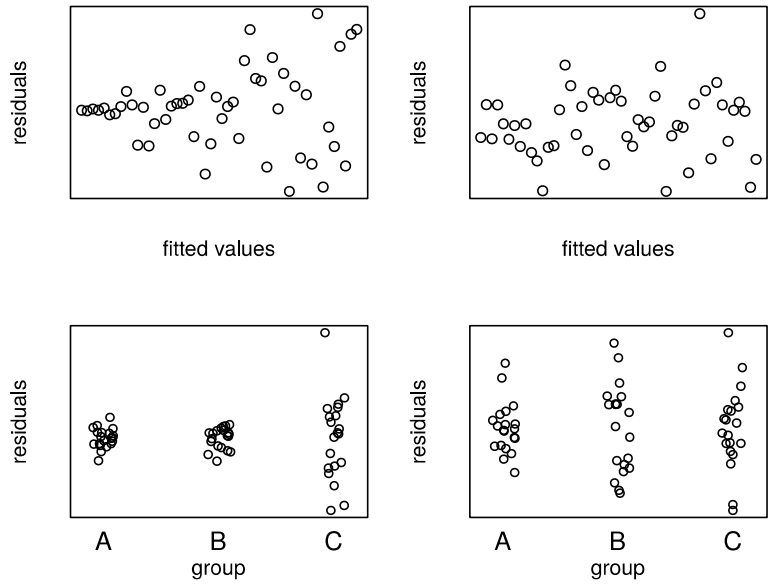

Figure 1: To check for non-constant variance, you can plot the model's residuals against their fitted values (top row; fictitious data) or against their corresponding predictor values (the 'group' variable in the bottom row; different fictitious data) and gauge if there is a pattern to the spread of the residuals. The funnel-shaped pattern in the top left plot and the visibly greater scatter for group C in the bottom left plot suggest that the residuals have non-constant variance. But what about the plots on the right?

But what about panels on the right-hand side? The assumption is that the data were *sampled* from distributions with the same variance. But due to sampling error, the residuals in any given sample may look as though they have non-constant variance even if they were in fact sampled from the same distribution. Here it seems that a judgement call is required, which can induce anxiety in beginning analysts. This tutorial is intended to alleviate some of this anxiety. It does so by giving a short introduction to a fairly recently developed technique (Buja et al., 2009; Loy et al., 2017; Majumder et al., 2013) that helps analysts to gauge whether any patterns in plots (diagnostic or other) can plausibly be accounted for by sampling error.

Assumptions of the general linear model

At their core, Student's t-test, simple and multiple regression, and ANOVA and ANCOVA are mathematically equivalent, the overarching label for them being the 'general

linear model' (not to be confused with the 'generalised linear model'). Correspondingly, these tools rely on the same assumptions, which Gelman and Hill (2007, Section 3.6) summarise in two pages. Here I will focus on the linearity, homoskedasticity (constant variance), and normality assumptions.[2] Figure 2 shows how these assumptions translate to a linear regression model with a single continuous predictor.

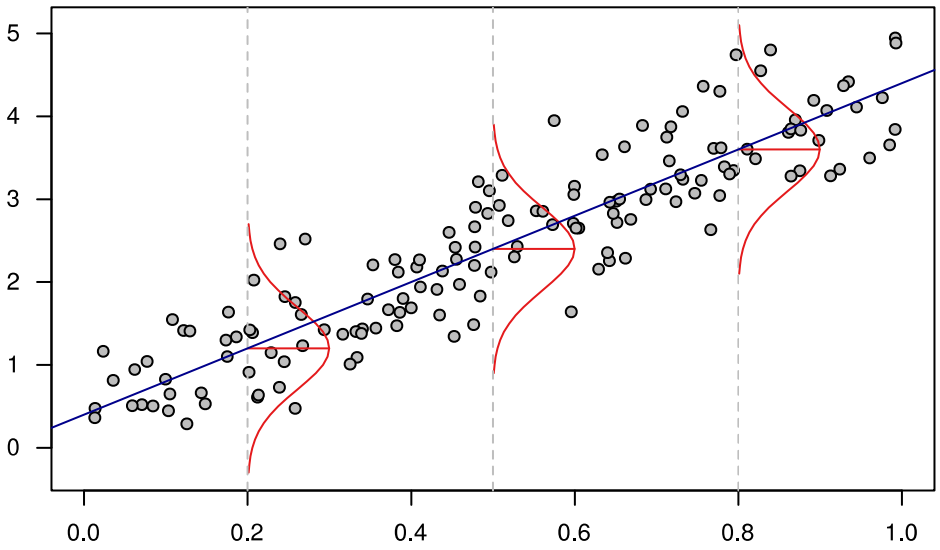

Figure 2: Illustration of the linearity, constant-variance and normality assumptions for a linear regression with one continuous predictor (x-axis) and one outcome (y-axis). The blue regression line connects the fitted y-values at each x-value and is assumed to be straight. The red density curves visualise that the model assumes the y-data to be normally distributed with the same variance at different x-values.

First, the general linear model assumes that the relationships between the outcome and any continuous predictors are linear[3]; in Figure 2, the straight regression line adequately captures the trend in the data. If a predictor is strongly nonlinearly related to the outcome, the model will still output estimated coefficients, standard errors, and *p*-values related to whatever linear trend best fits the predictor–outcome relationship. But these numbers may be quite irrelevant. Confronted with strong nonlinearities, analysts may want to include

2 A more important assumption than constant variance and normality is the assumption that the residuals were sampled independently from one another. This assumption is violated in, for instance, cluster-randomised designs (e.g., when entire classes of students are assigned to the control or intervention condition) or studies in which multiple data points were obtained per participant. Ignoring violations of the independence assumption can result in dramatic inferential errors, but these violations are difficult to diagnose without knowledge of the study design (but see Loy et al., 2017). A useful tool for dealing with such violations is the linear mixed-effects model. For an accessible introduction, see Winter (2013).

3 You can add non-linear transformations of the predictors, too, but then the model will assume that the transformed predictors are linearly related to the outcome.

additional variables that may account for them, transform the outcome and/or some predictors, or consider making use of tools that are able to capture nonlinear relationships (e.g., generalised additive models; for introductions, see Clark, 2018, and Zuur et al., 2009, Chapter 3).

Second, the general linear model assumes that the residuals were all sampled from the same distribution. The three red curves in Figure 2 show the assumed distribution of the residuals at three predictor values; the ‘constant variance’ assumption entails that these distributions all have the same width as they do in Figure 2. Serious violations of this assumption may affect inferential results, and there may be better methods for estimating the model coefficients than in a general linear model. Alternatives to Student’s *t*-test that account for non-constant variance are Welch’ *t*-test and the Yuen–Welch test (see Wilcox, 2005, Section 5.3). For regression models or models corresponding to ANOVA or ANCOVA, generalised least squares models can be useful (for an introduction, see Zuur et al., 2009, Chapter 4). If only the inferential results (i.e., *p*-values or confidence intervals, as opposed to the estimates of the model coefficients) are of importance, bootstrapping can be used to verify if similar results are obtained without making the constant-variance assumption (for an introduction, see Hesterberg, 2015).

Finally, the general linear model assumes that the residuals were all sampled from *normal* distributions (hence the bell curves in Figure 2). Strictly speaking, *t*- and *F*-statistics – and the *p*-values derived from them – depend on this assumption, but with largish sample sizes, inferential results do not depend on this assumption as much. For this reason, Gelman & Hill (2007) consider the normality assumption the least important modelling assumption, and they advise against testing for it. When in doubt, the inferential statistics can be double-checked by bootstrapping. That said, the general linear model models *mean* trends: the regression line in Figure 2 connects the means of the modelled distribution of the y-value at each x-value. If the residuals have a skewed, bimodal or otherwise wonky distribution, then these mean trends may be correctly estimated, but they may not capture a salient aspect of the

data. For this latter reason, I think that it is a good idea to check for violations of the normality assumption nevertheless.

Incidentally, it is exceedingly unlikely that any of the general linear model's assumptions are literally true, i.e., that any relationship in the social sciences is *perfectly* linear or that any residual pattern shows perfectly constant variance and perfect normality. For instance, actual data is recorded to a limited number of decimal places and often has an upper or lower bound, whereas the theoretical normal distribution is infinitely smooth and extends from negative to positive infinity. But nevertheless, it can be *useful* to summarise a relationship as linear if the approximation is good enough. Similarly, the degree of non-constant variance and non-normality may be so benign that taking it into consideration adds to the technical complexity of the analysis without any tangible increase in insight. When checking for assumption violations, then, the question is less *"Is this assumption literally true?"* than *"Is this assumption a reasonable approximation?"*.

Model diagnostics and the line-up protocol

A useful technique for checking model assumptions is the line-up protocol developed by Baju et al. (2009). Applied to model diagnostics, it goes as follows; these steps are illustrated in the two examples below.

1. Fit the statistical model whose assumptions you want to check. Student's t-tests, ANOVAs, and ANCOVAs can be written as models, too.
2. Use the fitted model to generate *n* (say, 19) new sets with outcome values. Each of these new sets contains the same number of observations as the dataset on which the original model was fitted. These new data are generated by putting the dataset's predictor values through the model equation in order to obtain fitted values and then adding random observations from the estimated error distribution to them. Due to the random errors, the 19 new sets will differ somewhat from each other as well as from the real outcome values.

3. Refit the original model to each of the the newly generated sets of outcome values. Crucially, since these new values were simulated by means of the original statistical model, all of this model's assumptions are *literally true* for them. If their diagnostic plots reveal worrisome patterns, then you know for certain that these are due to randomness rather to assumption violations.
4. Create a line-up with diagnostic plots for the models fitted on the simulated outcome vectors and insert the diagnostic plot for the original model at an unknown random position in this line-up.
5. Scrutinise the diagnostic plots and see if you can work out which is the original model's. If you already happen to know what your actual model's diagnostic plot looks like, ask some colleagues to work out which plot contains your actual data.
6. Verify your or your colleagues' answer. If it is right, then this suggests that the original model's diagnostic plot stands out from diagnostic plots that show no evidence of violated assumptions. In this case, you may want to consider carrying out some robustness checks or changing the main analysis altogether. But if it is wrong, then this suggests that the original model's diagnostic plot shows no evidence of violated assumptions – i.e., the diagnostic plot (in itself) is not a reason to worry. All the while, bear in mind that there is an important difference between saying "this plot doesn't show evidence of assumption violations" and "all assumptions were met." Particularly in small samples, assumption violations are difficult to identify regardless of the method used.

The above is an informal use of the line-up protocol. For more formal uses, including the possibility of using it as a stringent statistical test, see Majumder et al. (2013) and Loy et al. (2017).

Let's turn to two examples; the full R tutorial is available from http://taal.ch/RCode/walkthrough.html. [Note: I will put this link on osf.io.]

Example 1: A simple linear regression model

DeKeyser et al. (2010) recruited 62 Russian immigrants to Israel and subjected them to a 204-item Hebrew-language grammaticality judgement task (GJT). Their research question pertained to the relationship between the immigrants' age at immigration ('age of acquisition', AOA) and their GJT score, the theoretically relevant issue being whether there are nonlinearities in this relationship. Here we will check the assumptions of a reanalysis by Vanhove (2013), who fitted these data in a simple linear model.

The Line-Up In Action

Figure 3 shows steps 1 through 3 described above. To keep the plots legible, the data of only 15 out of 62 learners are plotted. In Step 1, the model coefficients are estimated (Panel A). Then, new outcome data are generated. This is done by generating fitted values from the model (Panel B), drawing random observations from the assumed error distribution (Panel C), and combining those (Panel D). The model is then re-estimated using the generated data (Panel E) and its new residuals are plotted (Panel F). This process is repeated 19 times.

In the full data set, the estimated regression equation for the AOA–GJT relationship is GJT = 187 − 1.2 × AOA, with the residuals being assumed to have been drawn from a normal distribution with a standard deviation of 16. The line-ups below pertain to this full-data model, not to the reduced dataset shown in Figure 3A.

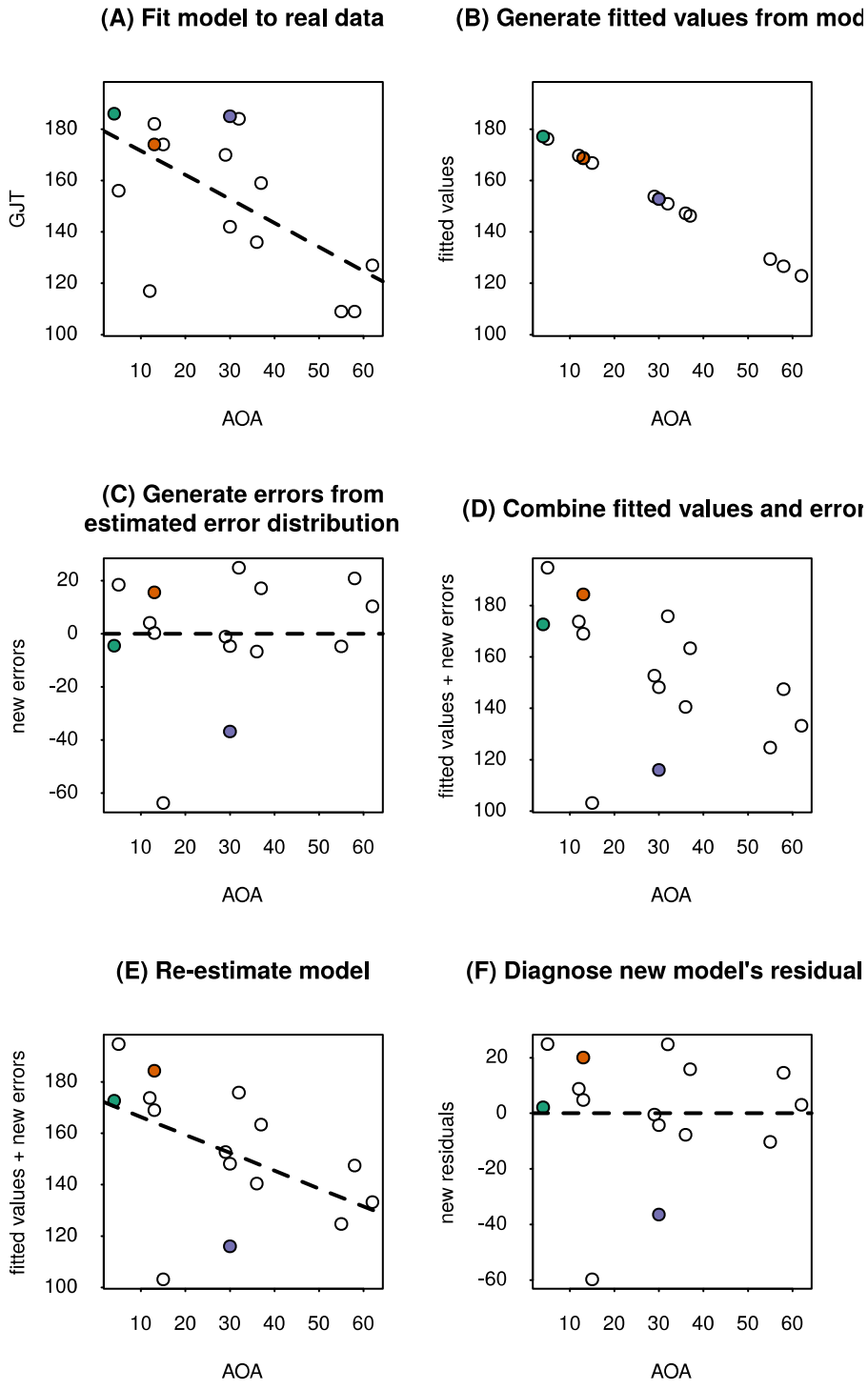

Figure 3: How the data for one new diagnostic plot are generated. This process is repeated 19 times in the line-up protocol. For expository purposes, a sample of only 15 observations is shown; the green, orange and purple points show the same AOA observations in all panels.

Linearity

For models with a single predictor, checking whether the predictor–outcome relationship is approximately linear can be done by drawing scatterplots of this relationship. But for more complex models, this does not always work, and a more generally useful approach is to plot the model's residuals against (a) its fitted values and (b) the predictor values. These plots should not show any remaining trends. To facilitate the diagnosis, a nonlinear smoother is often added to these plots; this smoother should essentially resemble a flat line but for sampling error. The resultant graph is shown as Figure 4. Which plot do you think shows the strongest residual pattern?

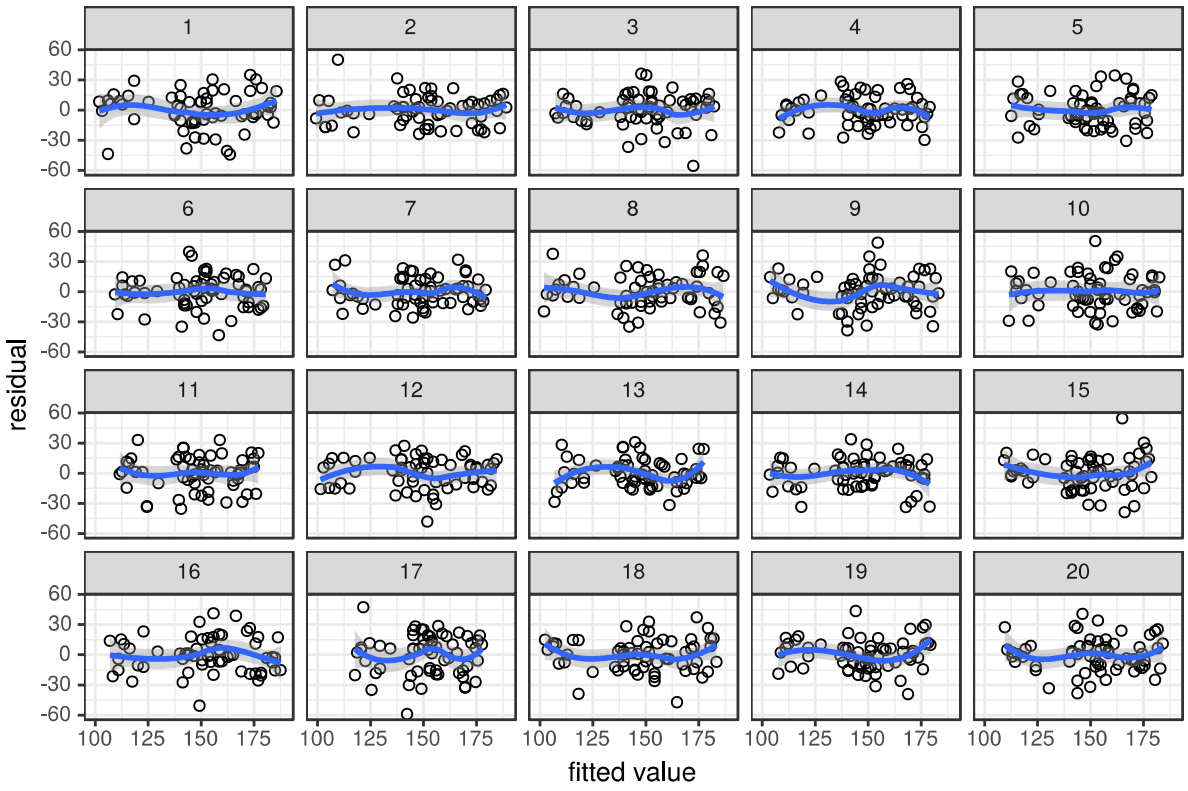

Figure 4: Fitted values and residuals. Which panel shows the strongest trend, suggestive of a nonlinear trend in the data?

As you notice, several of the plots seem to exhibit some nonlinearies. But most of these are entirely spurious. If you could not tell the true data from the simulated data, then the diagnostic plot in and of itself does not suggest that you need to worry about modelling nonlinearities. (The actual data's position equals the atomic number associated with a

Nirvana song.[4]) This does not mean that we have demonstrated that the true relationship between AOA and GJT is perfectly linear – but with these data it seems that next to nothing is lost by treating it as linear.

Constant Variance

To check for violations of the constant variance assumption, we can plot the absolute values of the residuals (or alternatively, the square roots of their absolute values) against the fitted values and see if there is any variation in the average residual scatter. Since we already know the position of the true data in the previous line-up, the line-up is generated anew. Which panel in Figure 5 shows the strongest pattern in the scatter of the residuals?

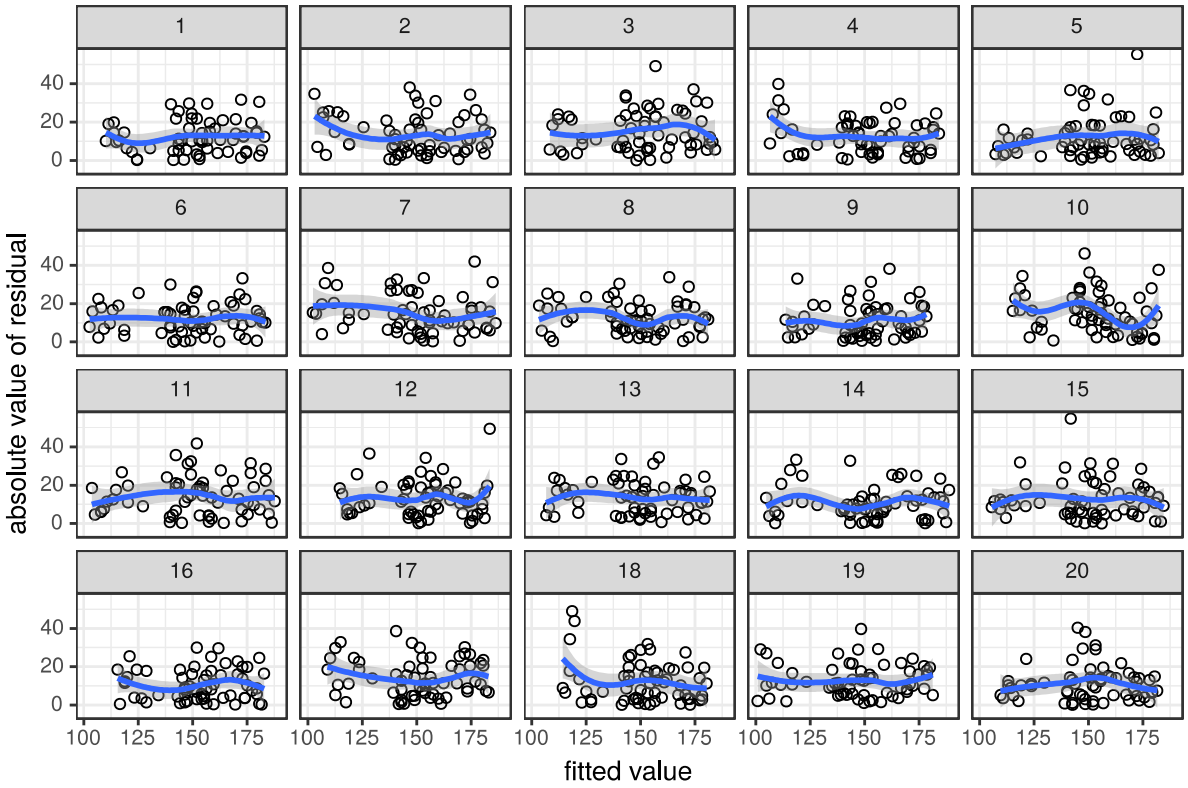

Figure 5: Fitted values (x-axis) vs. the absolute values of the residuals (y-axis). Which panel shows the strongest trend, suggestive of non-constant variance? (The actual data's position occurs in Paul Desmond's most famous tune.)

4 The solutions are put as quiz questions so that readers do not accidentally glance at them. The answers can also be found in the online tutorial.

Normality

To check for the normality of the residuals, we can follow the same procedure. Depending on your personal preferences, the residuals can be plotted using a histogram or a quantile−quantile plot; to save space, neither is shown here (see online tutorial).

Impossible Data

It can be informative to check if the model ever generates outcome values that are actually impossible given your subject-matter knowledge. Impossible simulated data highlight that you possess relevant knowledge that your statistical model has not picked up on and may help you to refine the model accordingly. For instance, the present model occasionally generates outcome data above the theoretically maximum value of 204. Follow-up steps to remedy this could include fitting the data in a binomial regression model or in a censored regression model, but these are beyond the scope of this tutorial.

Example 2: ANOVA on Likert-scale data

To investigate if people belonging to different social classes differ in their appreciation for experiental vs. material purchases, Lee et al. (2018, Study 2) recruited 469 participants categorised as belonging to a higher or lower social classes and asked them to recall either an experiential or material purchase they had made. They were then asked to rate how happy this purchase had made them on a 7-point Likert scale. Such data are commonly analysed in an ANOVA, but if you fit them explicitly as a linear model, you obtain the same results, as shown in the online materials. (The online materials also show how you can write a Student's *t*-test as a linear model.)

Since both purchase type (experiential vs. material) and social class (high vs. low) are categorical predictors, the linearity assumption is satisfied by default, so we skip right to the constant variance assumption.

Constant Variance

The principle is the same as in the previous example, but there is a problem. If you plot the residuals' absolute values (or the square roots thereof), you will immediately be able to spot the actual data plot – but not because of a blatant violation of the constant variance assumption: since the outcome data were measured on a (discrete) Likert scale, the actual residuals are considerably more discrete than the residuals for the simulated data (see online materials). While this demonstrates that the statistical model does not pick up on a salient aspect of the data, the discrete nature of the Likert data has little bearing on the constant variance assumption per se. To this end, I suggest to instead compute the standard deviation or variance of the residuals per unique combination of predictor values ('cell') and embed these in a line-up (see Figure 6).

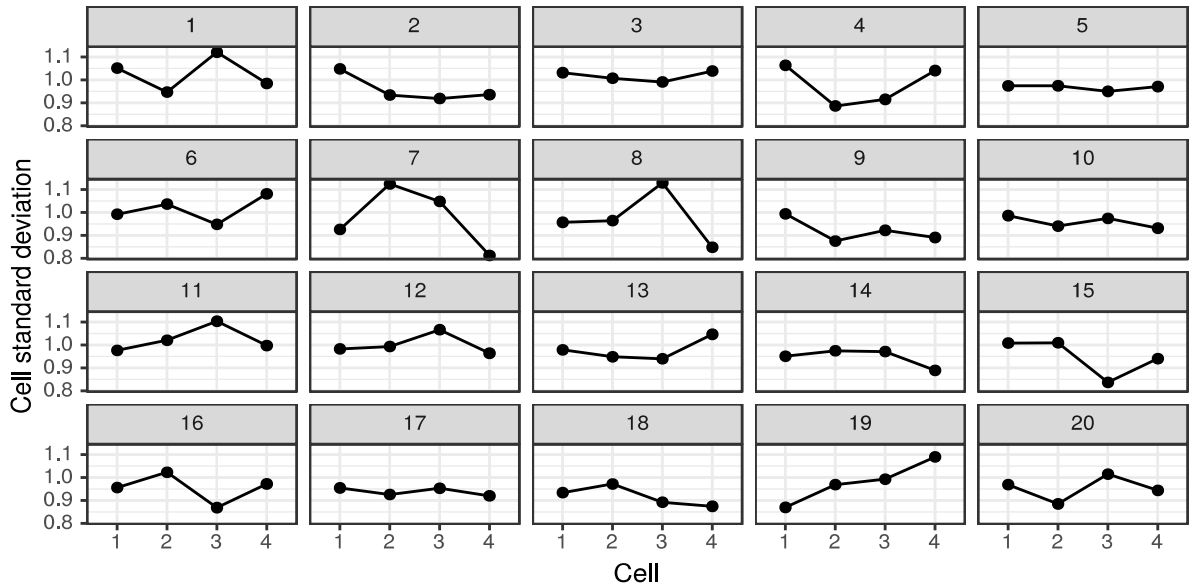

Figure 6: The fitted values for the four cells (x-axis) vs. standard deviation of the residuals per cell (y-axis). The cells are ordered arbitrarily. Which panel shows the largest cell-to-cell differences along the y-axis, suggestive of heteroskedasticity? (A song title from the Steely Dan album *Gaucho* contains the actual data's position.)

If you did not pick the actual data plot, this suggests that, while the model does not pick up on the discrete nature of the outcome data, it is not the constant variance assumption per se that seems to be clearly violated.

Normality

The line-up clearly shows that the residuals are not normally distributed (Figure 7). With this many data points, though, the inferential results hardly depend on the normality assumption. In the online materials, this is confirmed by constructing confidence intervals for the estimated model coefficients using the bootstrap.

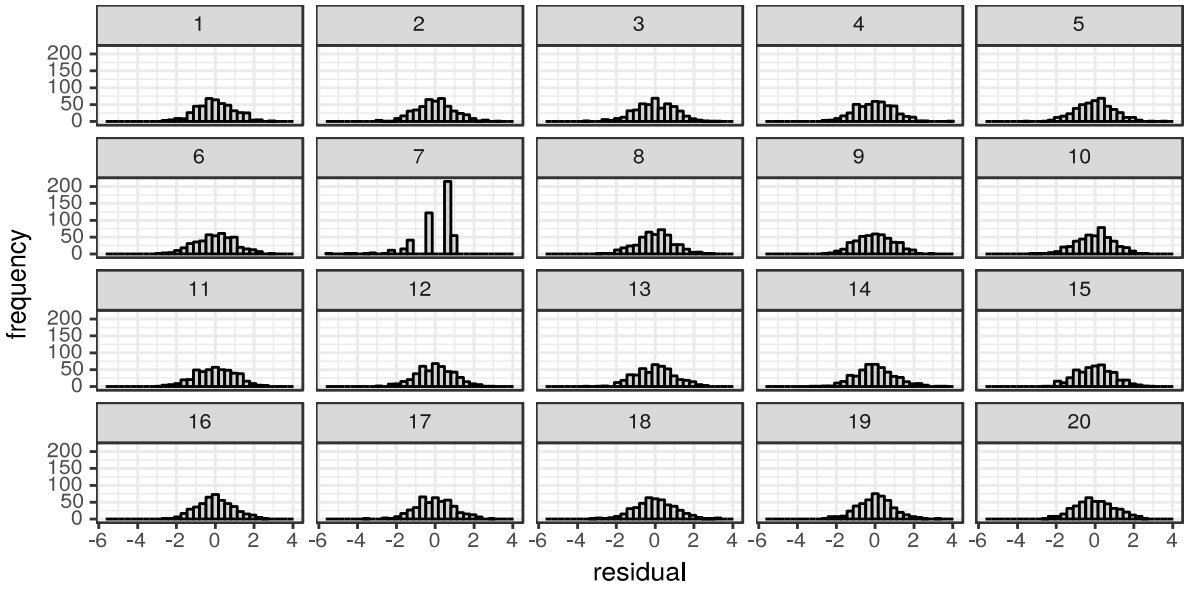

Figure 7: Histograms of the residuals. The true data panel sticks out like a sore thumb.

More reasonable concerns would be (a) that the residuals are negatively skewed due to the presence of a strong ceiling effect (nearly 60% of the respondents gave the maximum response) and (b) that the model ignores the discrete nature of the outcome data. Indeed, a linear model fitted on these data will happily generate decimal outcome data and outcome data above 7, both of which we know to be impossible. One way to address these concerns would be to refit the data using an ordinal model (see Bürkner & Vuorre, 2018, for a tutorial). In this case, doing so would not affect Lee et al.'s conclusion about the presence of an interaction between purchase type and social class on happiness ratings.

Summary

Suspected patterns in diagnostic plots may be due to randomness (sampling error) alone and do not have to indicate assumption violations. Outcome data simulated from the model being diagnosed conform literally to these assumptions. Refitting the model on such

simulated data and plotting the refitted models' diagnostic plots and the original model's side-by-side makes it easier to gauge whether the original model's diagnostic plot shows evidence of violated assumptions without falling victim to paranoia.

References

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, 367*(1906), 4361-4383. doi:10.1098/rsta.2009.0120

Bürkner, P.-C., & Vuorre, M. (2018). Ordinal regression models in psychology: A tutorial. doi:10.31234/osf.io/x8swp

Clark, M. (2018). *Generalized additive models*. Available from https://m-clark.github.io/generalized-additive-models/.

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*(3), 413-438. doi:10.1017/S0142716410000056

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet the assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi:10.2307/1169991

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician, 69*(4), 371-386. doi:10.1080/00031305.2015.1089789

Lee, J. C., Hall, D. L., & Wood, W. (2018). Experiential or material purchases? Social class determines purchase happiness. *Psychological Science, 29*(7), 1031-1039. doi:10.1177/0956797617736386

Loy, Adam, Hofmann, H., & Cook, D. (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics, 26*(3), 478-492. doi:10.1080/10618600.2017.1330207

Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference,

applied to linear models. *Journal of the American Statistial Association, 108*(503), 942-956. doi:10.1080/01621459.2013.808157

Vanhove, J. (2013). The critical period hypothesis in second language acquisition: A statistical critique and a reanalysis. *PLOS ONE, 8*(7), e69172. doi:10.1371/journal.pone.0069172

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*, 2nd edn. Amsterdam, The Netherlands: Elsevier.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. Retrieved from https://arxiv.org/abs/1308.5499v1.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer. doi:10.1007/978-0-387-87458-6