Better Diagnostics for Linear Mixed-Effects Models Using Visual Inference

Adam Loy, Heike Hofmann, Dianne Cook*

February 24, 2015

Abstract

Linear mixed-effects (LME) models are versatile models that account for dependence structures when data are composed of groups. The additional flexibility of random effects models comes at the cost of complicating model exploration and validation due to the breakdown of asymptotic results, boundary issues, and patterns in diagnostic plots inherent to model structures. While these issues are well known and adjustments have been proposed, they require the analyst to keep track of all the special circumstances that may arise. In this paper we illustrate the use of visual inference for diagnosing LME model fits based on graphical tests that that can be broadly used in situations where the assumptions of conventional inferential procedures are violated. This approach provides a unified framework for diagnosing LME model fits and for model selection.

Keywords: Statistical graphics; Lineup protocol; Visual tests; Model diagnostics; Model selection

*Adam Loy is an Assistant Professor in the Department of Mathematics, Lawrence University, Appleton, WI 54911 (e-mail: adam.m.loy@lawrence.edu). Heike Hofmann is a Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011 (e-mail: hofmann@iastate.edu). Dianne Cook is a Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011 (e-mail: dicook@iastate.edu).

1 Introduction

Linear mixed-effects models are versatile models that account for dependence structures when data are composed of groups. Such structures occur, for example, when individuals are naturally grouped by organization (e.g., students within schools), geography (e.g., voters within states), or design (e.g., respondents assigned to interviewers). The additional flexibility offered by these models allows data to be incorporated at both the observation-level (level 1) and the group-level (level 2, or higher) while also accommodating dependencies between individuals within the same group.

For data organized in g groups, consider a continuous response linear mixed-effects model (LME model) for each group i, $i = 1, \ldots, g$:

$$\underset{(n_i \times 1)}{\boldsymbol{y}_i} = \underset{(n_i \times p)}{\boldsymbol{X}_i} \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n_i \times q)}{\boldsymbol{Z}_i} \underset{(q \times 1)}{\boldsymbol{b}_i} + \underset{(n_i \times 1)}{\boldsymbol{\varepsilon}_i} \tag{1}$$

where $\boldsymbol{y}_i$ is the vector of outcomes for the n_i level-1 units in group i, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices for the fixed and random effects, respectively, $\boldsymbol{\beta}$ is a vector of p fixed effects governing the global mean structure, $\boldsymbol{b}_i$ is a vector of q random effects describing the between-group covariance structure, and $\boldsymbol{\varepsilon}_i$ is a vector of level-1 error terms accounting for the within-group covariance structure. The random effects, $\boldsymbol{b}_i$, are assumed to be a random sample from $\mathcal{N}(\boldsymbol{0},\ \boldsymbol{D})$ and independent from the level-1 error terms, $\boldsymbol{\varepsilon}_i$, which are assumed to follow a distribution of $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{R}_i)$. Here, $\boldsymbol{D}$ is a positive-definite $q \times q$ covariance matrix and $\boldsymbol{R}_i$ is a positive-definite $n_i \times n_i$ covariance matrix. Finally, it is assumed that all between group effects have a covariance of zero.

Inference typically centers around either the marginal or conditional distribution of $\boldsymbol{y}_i$, depending on whether global or group-specific questions are of interest. Based on model (1) the marginal distribution of $\boldsymbol{y}_i$ for all $i = 1, \ldots, g$ is given by

$$\boldsymbol{y}_i \sim \mathcal{N}\left(\boldsymbol{X}_i\boldsymbol{\beta},\ \boldsymbol{V}_i\right), \tag{2}$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \sigma^2\boldsymbol{R}_i$, and the conditional distribution of $\boldsymbol{y}_i$ given $\boldsymbol{b}_i$ is defined as

$$\boldsymbol{y}_i | \boldsymbol{b}_i \sim \mathcal{N}\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,\ \sigma^2\boldsymbol{R}_i\right). \tag{3}$$

Residuals are central to the exploration of a linear mixed-effects model, similar to the classical linear model with uncorrelated error structure. For linear mixed-effects models, residual analysis is complicated by the fact that there are numerous quantities that can be defined as *residuals*, with each residual quantity being associated with different aspects of the model. The two fundamental residuals for model checking considered here are:

- the *level-1 (observation-level) residuals*, the conditional residuals or error terms: $\widehat{\boldsymbol{\varepsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i$,
- and the *level-2 (group-level) residuals*, the predicted random effects $\widehat{\boldsymbol{b}}_i$

where $\widehat{\boldsymbol{\beta}}$ is an estimate of the fixed effects,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{g} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{g} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{y}_i, \tag{4}$$

and $\widehat{\boldsymbol{b}}_i$ are predictions of the random effects, given as

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}\right), \quad \forall\ i = 1, \ldots, g. \tag{5}$$

When $\boldsymbol{V}_i$ is unknown, estimates for the covariance matrices are used in the above equations. These estimates are commonly found through maximum likelihood (ML) or restricted maximum likelihood (REML).

The additional flexibility of random effects models comes at the cost of complicating model exploration and validation. Problems addressed in this paper include:

1. *Model selection:*
   (a) *Breakdown of asymptotic results:* Test statistics used for model selection and validation rely on asymptotic reference distributions which often perform poorly in finite sample situations. For example, the empirical distribution of the predicted random effects does not resemble the theoretical distribution unless strong assumptions are met (Jiang, 1998, Theorem 3.2 and Lemma 3.1). In many finite sample situations these assumptions do not hold, making conventional use of Q-Q plots and tests of the empirical distribution function ineffective for distributional assessment (see the supplement of Loy and Hofmann, 2015, for supporting simulation results).

(b) *Boundary issues in the comparison of nested models:* When evaluating the significance of terms in the random effects structure, the likelihood ratio test statistic does not have the usual χ^2 reference distribution if we are testing whether a variance component lies on the boundary of the parameter space. This results in tests for the random effects that tend to be conservative.

2. *Model checking:* Residual plots might display noticeable patterns that are artifacts of both the data structure and the model estimation procedure rather than indications of lack of fit. This problem is especially pronounced for scatterplots comparing predicted random effects. Such plots can display points falling along a line that is oriented in conflict with the estimated correlation between the random effects (see Morrell and Brant, 2000, for an example). Additionally, plots of the error terms often exhibit patterns that appear to be indicative of heteroscedasticity, but are merely consequences of data imbalances or sparsity.

The above issues are well-known, and in special circumstances adjustments to the methodology have been proposed. For example, Stram and Lee (1994) suggest using a 50:50 mixture of χ^2_q and χ^2_{q+1} when testing q versus $q+1$ random effects; however, this adjustment is not successful in all cases. Lange and Ryan (1989) suggest using weighted Q-Q plots to assess the distributional assumptions made on the random effects. This approach is effective in settings where the residual variance, σ^2, is small relative to the variance components for the random effects, but breaks down when this is not the case (Loy and Hofmann, 2015).

The purpose of this paper is to illustrate the use of visual inference for diagnosing LME model fits, and propose graphical tests that can be broadly used in situations where the assumptions of conventional inferential procedures are violated. This approach harnesses the power of graphical diagnostics and, with the rigor of the visual inference framework, construct tests that address some of these problems without requiring additional assumptions.

The remainder of this paper is organized as follows. Section 2 provides an introduction to the framework of visual inference and the lineup protocol. Sections 3 and 4 expand on the problems encountered in model selection and model checking, respectively, and solutions utilizing visual inference. Throughout both sections multiple examples are used, and comparisons between conclusions on model building and fit from the new visual inference approach and existing diagnostic tools are provided. All example data sets are readily available for public use. A short description of each data set can be found in Appendix A.

2 Visual inference

Classical statistical inference consists of (i) formulating null and alternative hypotheses, (ii) calculating a test statistic from the observed data, (iii) comparing the test statistic to a reference (null) distribution, and (iv) deriving a p-value on which a conclusion is based. Each of these steps has a direct analog in visual inference, as outlined by Buja et al. (2009). This section highlights the parallels between conventional hypothesis tests and visual inference in the setting of linear mixed-effects models.

Assume that the question of interest involves some assumption about a model (such as a null hypothesis of homogeneity of residual variance) while the alternative hypothesis encompasses any violation of this model assumption. For visual inference, the test statistic corresponds to a plot that displays an aspect of the model assumption and allows the observer to distinguish between scenarios under the null hypothesis from scenarios under alternative hypotheses. Plots drawn from data generated consistently with the null hypothesis are called *null plots*. The set of all null plots constitutes the reference distribution; thus, the plot of the observed data is indistinguishable from the null plots if the model assumption holds. In the lineup protocol the plot of the observed data is randomly embedded among a sample of, usually 19, null plots drawn from the reference distribution. These *lineups* are then presented to independent observers for evaluation.

Evaluation by independent observers allows for the estimation of a p-value associated with the lineup. Let X be the random variable describing the number of observers, out of N, identifying the data plot. If $X = x$ is the number of observers who chose the data plot from the lineup, then the p-value is the probability that at least x observers choose the data plot, given that the null hypothesis is true (i.e., the data plot is not any different from the other plots in the lineup). Under the null hypothesis the probability of choosing the true plot is $1/m$ (for a lineup of size m), and X is distributed according to a distribution similar

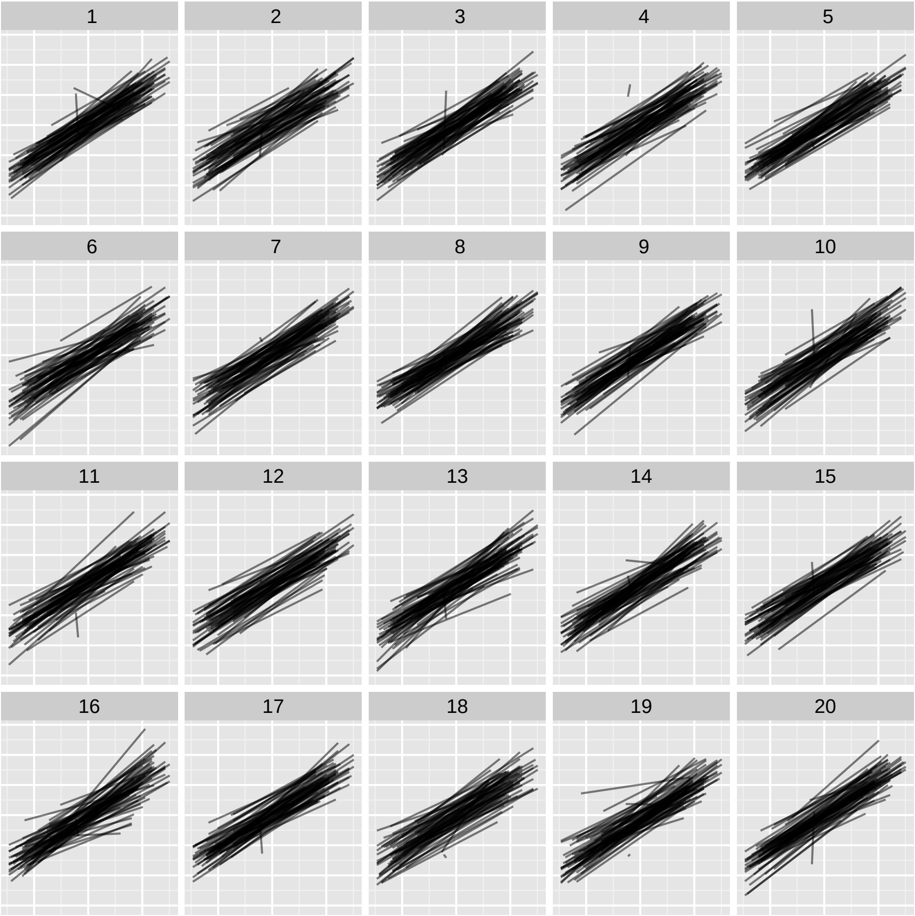


Figure 1: A lineup of size 20. Which plot is the most different from the other plots? What makes that plot different from the other plots?

to a Binomial distribution $B_{N,1/m}$, but adjusted for the dependencies between plots in a given lineup (Majumder et al., 2013, introduced visual p-values. Details of their calculation for this LME model application are in Appendix 2.1.).

Figure 1 shows a lineup. Each panel shows line segments of different lengths with varying slopes. Observers were asked the question 'Which plot is the most different?'. Any information revealing the context of the data, such as axis labels, units, titles, and legends, was carefully removed to avoid subjective bias (Meilgaard et al., 2006), ensuring observers make decisions that were based purely on the data display.

Based on a human subject study (described in more detail in Appendix B) run through the Amazon MTurk service (Amazon.com, Inc, 2014), 11 out of 73 observers chose the data plot, shown in panel #$(\sqrt{144}+4)$[1] of the lineup in Figure 1, resulting in a visual p-value of 0.0171. This leads us to reject the null hypothesis that the data plot is consistent with the model that generated the data, on which all of the other plots are based. The model and its corresponding null hypothesis are explained in detail in Section 3.

Unlike classical hypothesis tests, visual inference allows us to collect additional information on what aspect of the display led each observer to their choice. This information makes it possible to assess which part of the null hypothesis is violated, something not feasible in classical hypothesis tests. For example, 'Spread' as the reason for an observer's choice (over 'Outlier', 'Trend', 'Asymmetry', or 'Other') in Figure 1 was associated with the highest probability of picking the data plot over a null plot (see Table 3).

3 Model selection

Model selection for linear mixed-effects models relies on the comparison of nested models for the selection of both the fixed and random components. It is standard practice to use a t-test, F-test, or likelihood ratio test to determine whether a fixed effect describes a significant portion of the unexplained variability. When selecting random effects, likelihood ratio tests are most commonly used. However, situations often arise that complicate such tests. Some of these are outlined below:

Fixed effects: Likelihood ratio tests based on REML estimation cannot be used to test different fixed effects structures. Maximum likelihood estimation allows for such comparisons, but is anti-conservative. Defining the appropriate degrees of freedom for t- or F-tests provides another complication in testing scenarios of LME models. Various approximate F-tests propose solutions for estimating the degrees of freedom for these tests, but these typically lead to different results (Verbeke and Molenberghs, 2000). Inflated Type I error rates and low power become a problem in the approximate F-test when there are a few groups (Catellier and Muller, 2000). Kenward and Roger

[1] We encode the panel number as a mathematical expression to pose a cognitive obstacle, allowing the reader to evaluate the lineup before being biased by knowing the answer.

(1997) propose the use of a scaled Wald statistic that has a sampling distribution that is well approximated by an F distribution; however, this approach has inflated type I error rates in some small sample cases (Gomez et al., 2005). Skene and Kenward (2010) propose an alternative to the Kenward-Roger approximation that achieves nominal type I error rates, but this procedure suffers from low power.

Random components: When testing for the inclusion of a variance component, the parameter being tested lies on the boundary of the parameter space, and the asymptotic distribution of the likelihood ratio is no longer a χ^2 distribution. Approximations have been suggested and shown to be useful in many situations (Stram and Lee, 1994; Morrell, 1998), but no single approximation holds for all situations. This leads to a need for simulation studies to determine the proper adjustment to the reference distribution in every situation. Alternatively, the rule of thumb suggested by Stram and Lee can be utilized with the knowledge that the results may be sub-optimal, leading to either conservative or anti-conservative decisions.

Visual inference provides an alternative to conventional hypothesis tests that does not require different rules based on the method of estimation or location of a parameter in the parameter space, and avoids the tricky business of defining degrees of freedom. Rather, visual inference depends on the choice of an appropriate plot highlighting the aspect of the model in question, the number of null plots, and the number of independent observers. Below we discuss how visual inference can be utilized to test the significance of fixed and random effects.

3.1 Fixed effects

To test the significance of a fixed effect, we suggest using a plot comparing a residual quantity from the model without the variable of interest with the values of that variable. The residual used depends on the level at which the variable of interest enters the model: if the variable enters at the observation-level (level-1), then the level-1 residuals are used; if the variable enters at the group level, then both the level-1 and level-2 residuals are explored as the variable has the potential to explain additional variation at either level of the model.

Additionally, the type of plot depends on the variable type—if a continuous variable is targeted, a scatterplot with a smoother is suitable for testing; for a discrete covariate, we make use of side-by-side box plots. In this setting, the null plots are generated using the parametric bootstrap with a model that omits the variable of interest. The true plot is constructed from the same model but uses the observed data.

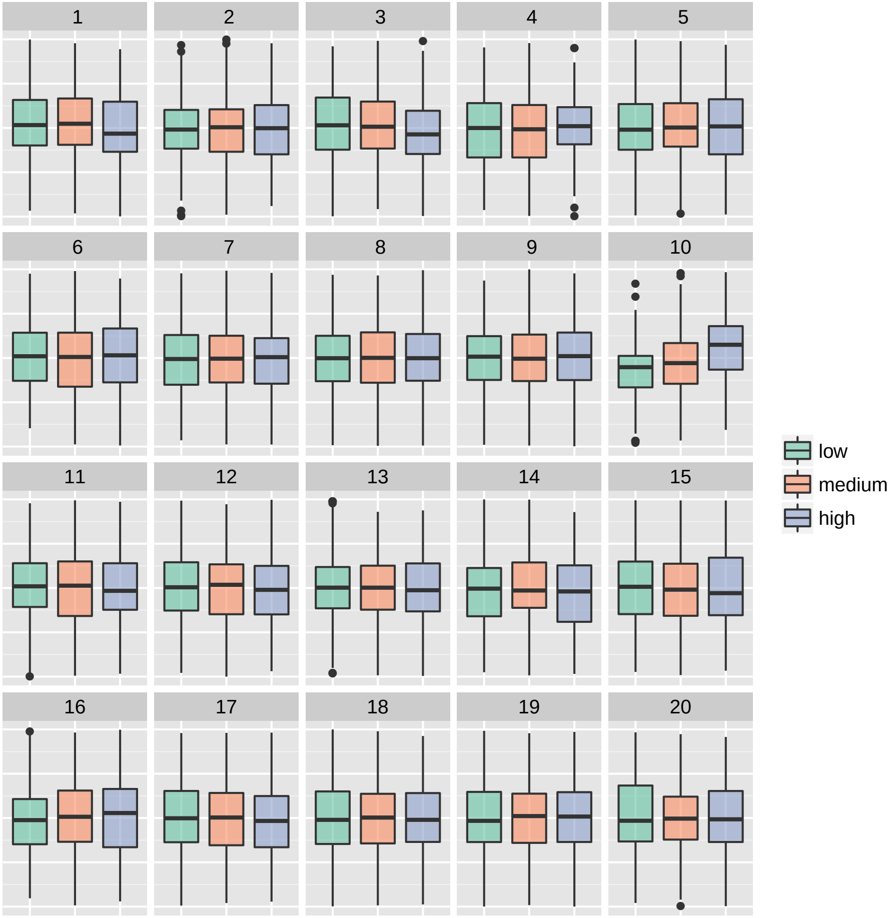


Figure 2: Lineup of level-1 residuals in box plots by groups to test significance of a discrete covariate. Which of the plots is the most different? Which feature led you to your choice?

Figure 2 illustrates the use of this type of lineup. In our study, 74 observers were asked to choose the plot that is the most different from the rest. Sixty five observers identified the data plot in panel #$(2^3 + 2)$, with 40.3% pointing to the trend as the distinguishing feature. This lineup is chosen to determine whether a child's language development (low, medium, or high) at age two is associated with the development of social skills for children diagnosed with autism spectrum disorder (see Appendix 1.2 for details on the data and analysis). Displayed on the y-axis are level-1 residuals from a longitudinal model, i.e., the grouping in model 1 is given by each individual. Clearly, language development at age two accounts for a significant amount of the remaining residual variability.

3.2 Random effects

Tests of the random part of a linear mixed-effects model focus on two questions: (1) whether a marginal random effect improves the model and (2) whether allowing the random effects to be correlated improves the model. Different plots must be used to answer each question. To answer the first question, we suggest using plots comparing the response and the explanatory variable of interest using appropriate (often linear) smoothers for each group. Scatterplots comparing the predicted random effects can be used to answer the second question.

The lineup in Figure 1 was chosen to test the relationship between scores from the General Certificate of Secondary Education Exam (GCSEE) and the standardized London Reading Test (LRT) (see Appendix 1.1 for details). Each line segment represents one of 65 inner-London schools. The slope of each line is determined by a linear regression relating the two test scores for all students at a school. The question of interest is whether random slopes for LRT scores are required to represent the relationship between GCSEE and LRT scores (H_1). Correspondingly, data for the null plots were created by simulating GCSEE scores from a model with the random intercept as its only random effect. The resulting scores for each school are regressed on LRT scores and model fits are shown as lines. If the model is appropriate, then the observed data should resemble the overall pattern of the lines in the null plots. In this example, we find that the true plot in panel $\#(\sqrt{144} + 4)$ is identifiable: 11 of the 73 observers pick the data plot, resulting in a visual p-value of 0.0171. The main comments participants gave to explain their choice were the spread and trend of the line segments in the plot. This is consistent with a larger variance of slopes than the null model allows; thus, we find evidence supporting the inclusion of a random slope for standardized LRT. This conclusion agrees with the results of the likelihood ratio

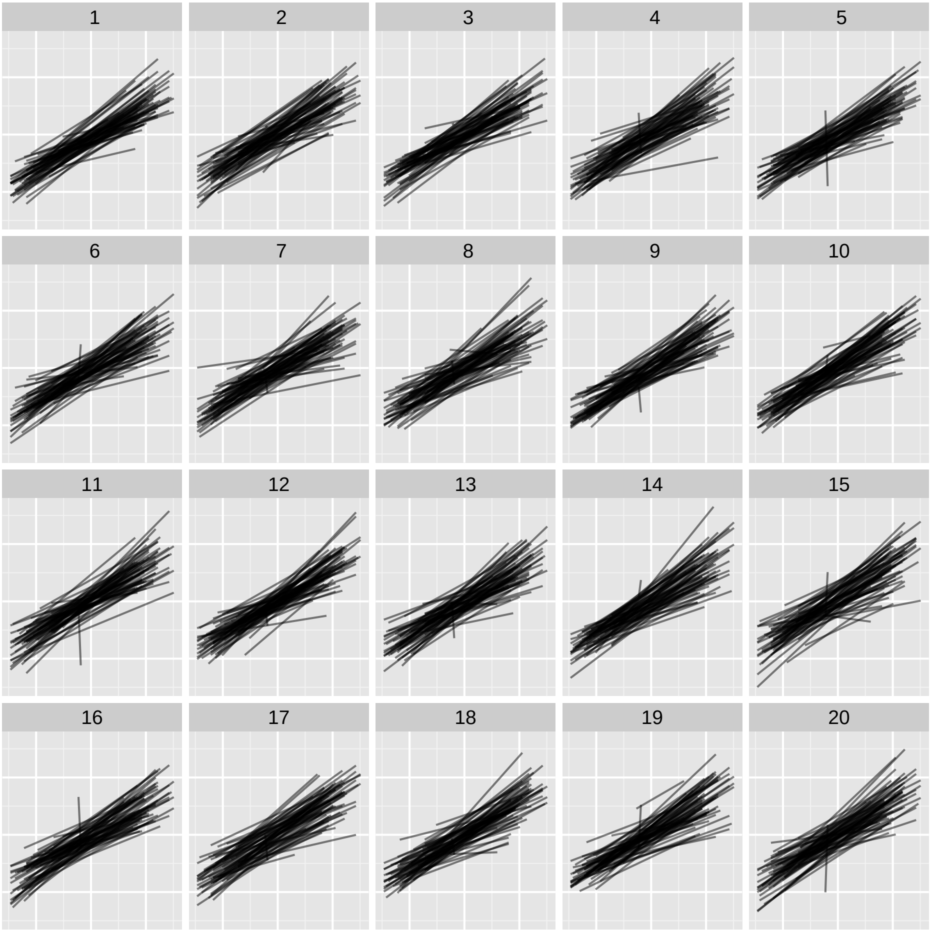


Figure 3: A follow-up to Figure 1 where the random slope has been included in the model. Which plot is the most different of the others?

test (which shows significance at a level of less than 0.0001), and did not require the use of an asymptotic distribution to calculate the p-value.

Note that participants are unable to identify the true data from a follow-up lineup (Figure 3) where the random slope was included in the model. Data for the null plots were generated from a parametric bootstrap using a normal distribution for the error terms and random effects as specified in model 1. None of the 64 observers identified the data plot in panel #$(2 \cdot 3^2)$. The covariance structure for the LRT scores as given in the model of the null hypothesis can therefore not be rejected as improperly specified.

Participants should only see one of Figures 1 or 3 because both show the same data, even in the same design. After viewing the first of these figures we cannot, strictly speaking, assume that a participant is still an unbiased judge, because, theoretically, the data from the second lineup could be identified by recognizing it as the same panel that was previously shown. While the chance of this is slim, we only exposed participants to one of a set of dependent lineups.

Having considered the value of a random slope in the model, we next consider whether the model needs to allow the random effects to be correlated (H_1). While this is an example of a standard likelihood ratio test problem—a correlation of zero is not on the boundary of the parameter space—using a lineup keeps all tests of the random effects in a unified framework. The lineup in Figure 4 shows scatterplots of the predicted random effects with overlaid regression lines. The null plots in the lineup are created by simulation from the model that does not allow for correlation between the random effects, and the true plot is created using the predicted random effects from such a model fit to the observed data. The slopes of the regression lines are indicative of the amount of correlation. If the correlation between the random effects is not necessary, then the true plot will display little correlation and be indistinguishable from the null plots. The lineup allows us to gauge the amount of correlation between the random effects while accounting for the effect of shrinkage in the model, avoiding the over-interpretation of structure in such plots discussed by Morrell and Brant (2000). In Figure 4 the true plot in panel $\#(10 + \sqrt{25})$ was identified by 41 of the 69 observers, providing very strong evidence in support of the additional parameter for correlation between random effects, which agrees with a p-value of 0.0041 from the classical likelihood ratio test.

4 Model checking

In the formulation of model (1) we make a number of assumptions that must be satisfied. In this section we discuss how residual plots can be used with lineups to check the assumptions of homogeneous residual variance, linearity, and normality of the random effects. While we only focus on these assumptions, the discussion is general enough to reveal how visual inference can be extended to check other aspects of the model.

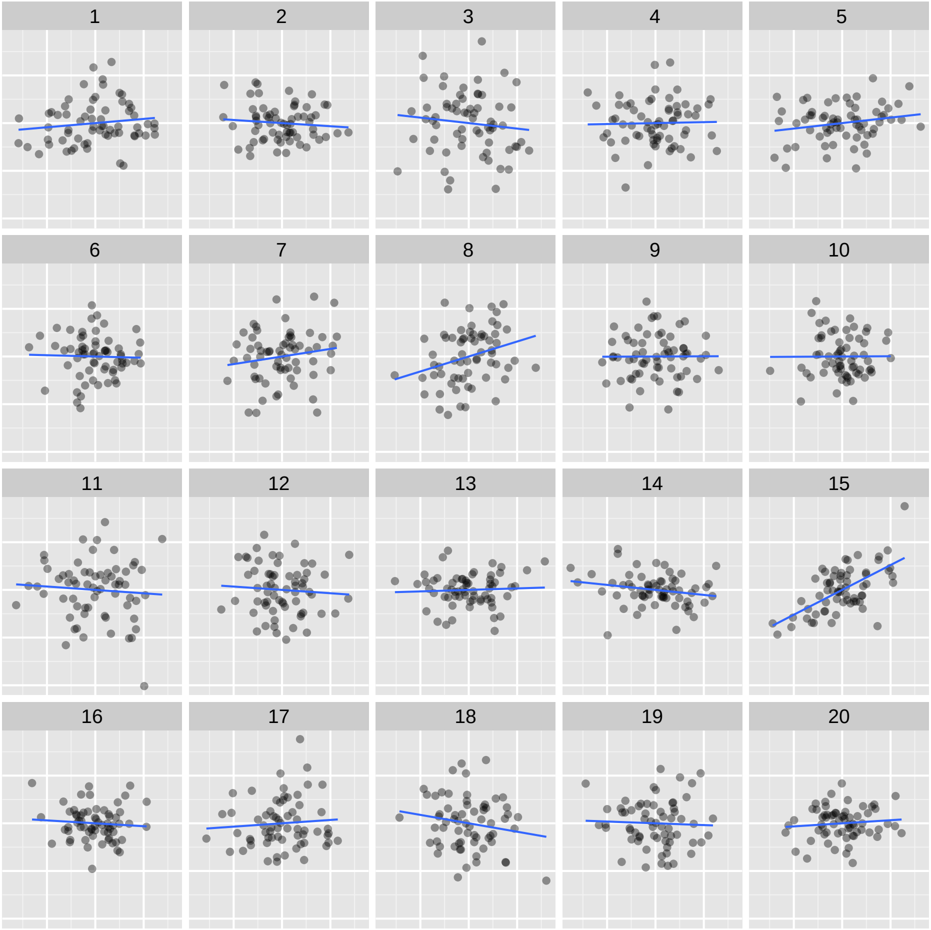


Figure 4: Lineup for testing the correlation between random effects. Which plot is the most different of the others? What feature in the panel led you to your choice?

4.1 Homogeneity of variance

Model (1) assumes homogeneity of the within-group variance. To check this assumption we must verify the homogeneity of the within-group residual variance across the levels of all explanatory variables and check that the within-group variance is also constant between groups. Such investigations are often carried out using plots of the level-1 residuals. In order to guard against mis- or over-interpretation of the residual plots, we can again employ lineups.

Figure 5 shows a lineup of 20 plots of the level-1 residuals against pressure in the dialyzer study (see Appendix 1.4). In this example 29 of 85 observers identified the true plot in

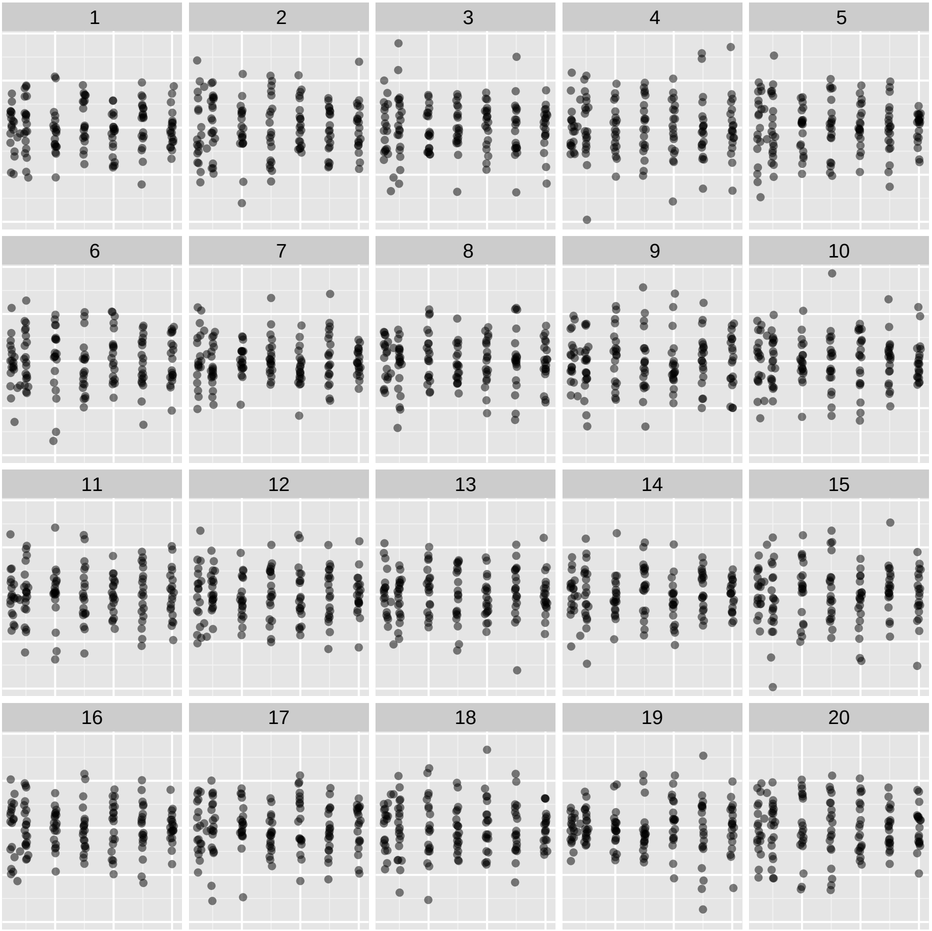


Figure 5: Lineup testing homogeneity of the level-1 residuals. Which of the plots is the most different? Which feature led you to your choice?

panel #$(2^4 + 3)$, providing evidence of heteroscedasticity.

Residual scatterplots are useful in checking that level-1 residuals are homoscedastic with respect to the explanatory variables, but they do not show the same power as box plots when they investigate potential differences in variability between groups, as can be seen in the example of the lineups in Figures 13 and 14 of Appendix 2.2. To visualize this assumption, we suggest using side-by-side box plots of the level-1 residuals. While box plots are more powerful, they still require the use of lineups. When the plots are considered individually, unbalanced group sizes will cause artificial structure in these plots. To overcome this difficulty, we create a lineup of side-by-side box plots for each group ordered by their interquartile range (IQR), which we have come to call a *cyclone plot*. Figure 6 shows a lineup of cyclone plots for 66 patients in a longitudinal study investigating the ability of methylprednisolone to treat patients with severe alcoholic hepatitis (see Section 1.3 for details). The true plot in panel #$(2^3 + 5)$ is easily identified from the field of null plots (by 50 of 75 observers) revealing heteroscedasticity across groups that might not be apparent in other residual plots.

While the data plot is easily identified, any panel from this lineup considered separately exhibits a structure that might, taken by itself, lead an analyst to the conclusion that within-group variance increases across the vertical axis. However, placing the true plot into the lineup forces the analyst to consider most of this structure as inherent to the data rather than evidence against the hypothesis of homogeneity. The fact that observers are able to still identify the data plot indicates that the data plot has additional structure inconsistent with homogeneity. The use of lineups incorporates the comparison of the data to what is expected, eliminating the subjective interpretations we encounter with the use of single plots.

An alternative approach to detect heteroscedasticity of the level-1 residuals across groups is to use a test based on the standardized measure of dispersion given by

$$d_i = \frac{\log\left(s_i^2\right) - \left[\sum_i (n_i - r_i) \log\left(s_i^2\right) / \sum_i (n_i - r_i)\right]}{\left(2/(n_i - r_i)\right)^{1/2}}, \tag{6}$$

where s_i^2 is the residual variance within each group based on separate ordinary least squares regressions and r_i is the rank of the corresponding model matrix (Raudenbush and Bryk, 2002). The test statistic is then

$$H = \sum_{i=1}^{g^*} d_i^2 \tag{7}$$

which has an approximate $\chi^2_{g^*-1}$ reference distribution when the data are normal and the group sizes are "large enough." Here we use g^* because "small" groups may be excluded from the calculation as small group sizes provide less reliable information about the residual variance (assuming that there are enough observations to fit the model), but this is a subjective choice. A common rule of thumb is to exclude groups with samples sizes smaller than 10. If the distributional assumptions are violated, or we do not have large enough group sizes, the approximation to the χ^2 distribution breaks down. In the methylpred-

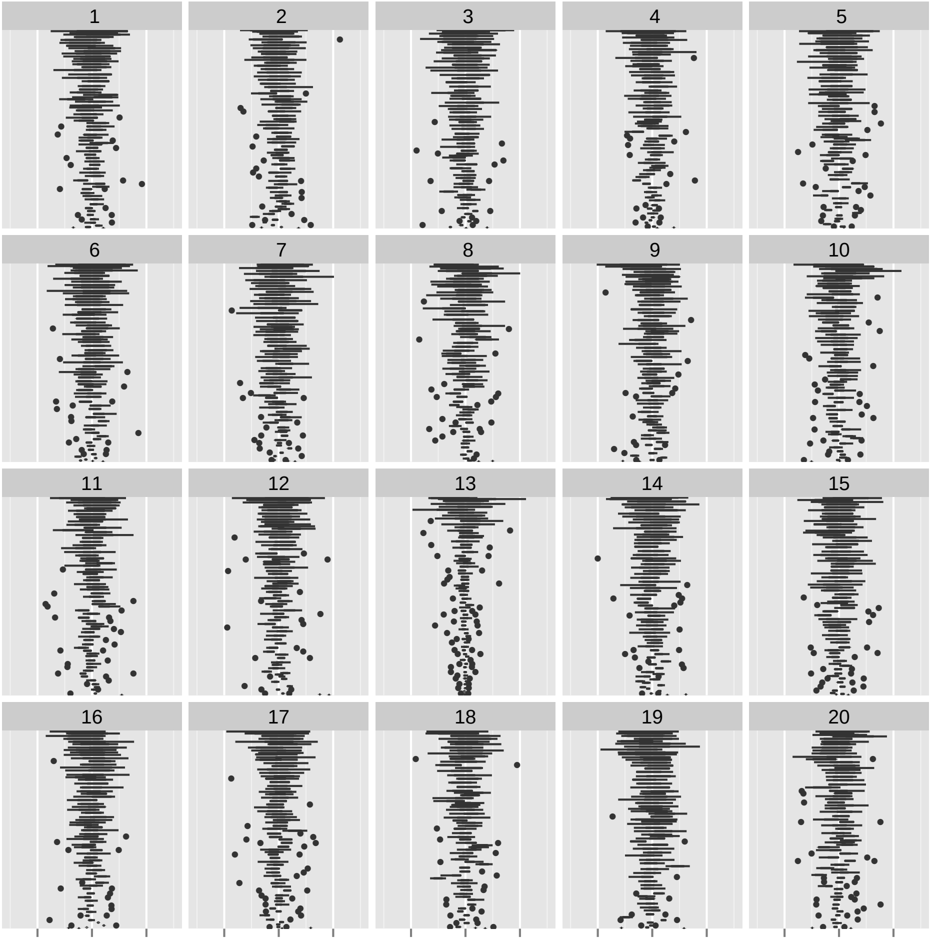


Figure 6: Lineup testing homogeneity of the level-1 residuals between groups. Which of the plots is the most different? Which feature led you to your choice?

nisolone study each subject was observed at most five times, with 19 subjects dropping out of the study early. Due to the small group sizes the χ^2 approximation is inappropriate, forcing the analyst to rely on simulation to construct the sampling distribution of the test statistic, which is not only computationally more demanding than the generation of 19 null plots, but lacks power with small sample sizes. Additionally, we have found it to be sensitive to the choice of the minimum group size (see Table 1 for an example using the radon data set). Performing the simulation-based version of the test in this example results in a p-value of 0.0886, providing much weaker evidence for heterogeneity than the visual test.

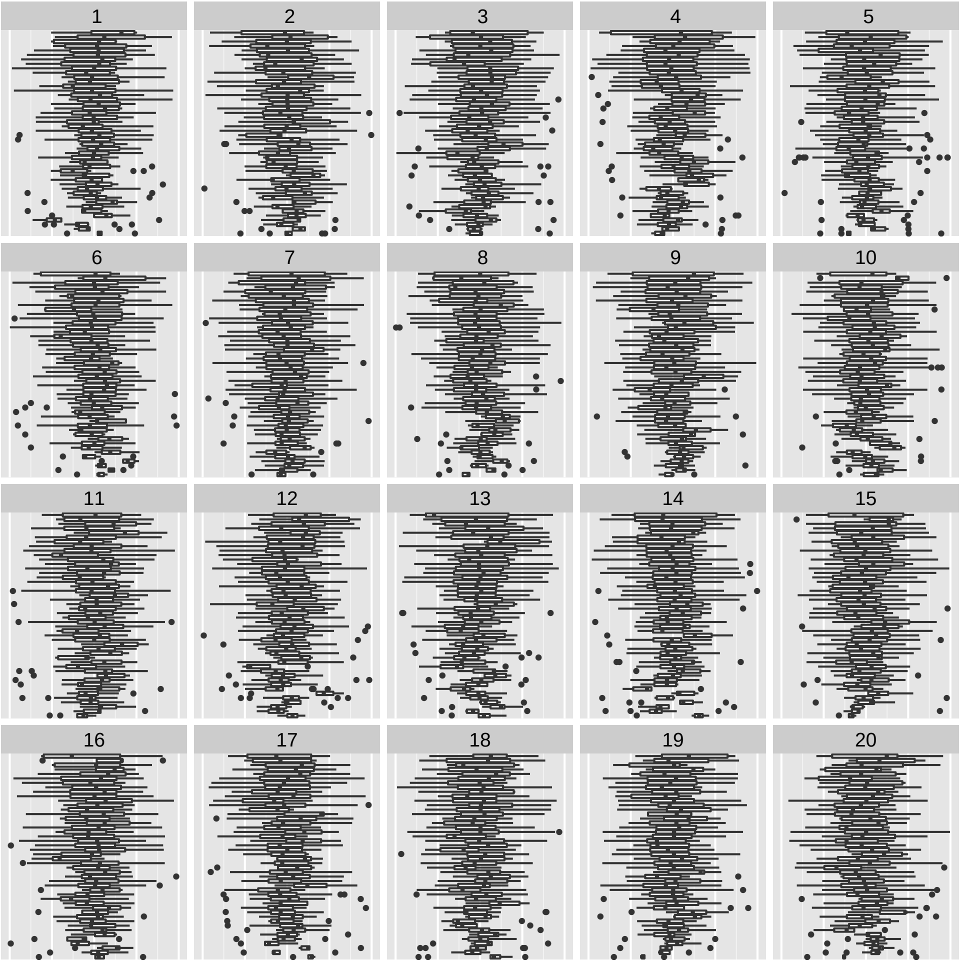


Figure 7: Lineup of 20 box plots (ordered by IQR) of level-1 residuals used to test the assumption of homogeneous level-1 residual variance. Which is the real plot?

Figure 7 shows another lineup of cyclone plots. The data underlying this example are radon measurements across counties in Minnesota (see Section 1.5 for details). Here, level-1 residuals by county are plotted from a model that only includes counties with at least five observations. Only one out of 59 participants identified the data plot shown in panel #$(4^2 - 6)$ from the lineup, providing no evidence against homogeneity. In this situation the conventional test yields a p-value of 0.149, if counties with fewer than 10 observations are excluded, thereby agreeing with the visual inference result. However, if only counties with fewer than 5 observations are excluded, the conventional test indicates strong evidence of heterogeneity based on a p-value of 0.0017. This sensitivity to the group size is a clear

Table 1: A comparison of the naive and the simulation-based version (based on 10,000 simulated statistics) of the conventional test of homoscedastic error terms across groups for the radon data. There is a clear discrepancy between the two p-values, indicating a need for simulation-based methods. Additionally, the sensitivity of the simulation-based test to the minimum group size is apparent.

Minimum group size	H	d.f.	Naive p-value	Simulated p-value
3	116.6	73	0.0009	0.9178
4	96.8	62	0.0031	0.7980
5	77.9	45	0.0017	0.6066
6	75.8	38	0.0003	0.5313
7	59.0	33	0.0036	0.4697
8	51.2	29	0.0066	0.4119
9	39.6	26	0.0426	0.3509
10	27.7	21	0.1490	0.2595
11	26.6	19	0.1145	0.2260
12	23.7	17	0.1281	0.1952
13	23.7	16	0.0966	0.1873
14	8.2	11	0.6940	0.1360
15	5.1	7	0.6429	0.0764

weakness of the conventional test, and casts doubt on its usefulness in situations with small group sizes. Table 1 provides further exploration of this sensitivity. In contrast, the visual test is not overly sensitive to group sizes. Counties with less than 5 observations were eliminated because box plots are not appropriate for such small group sizes, but could be included in the representation as dot plots. While we are still slightly constrained by group size, we are far less constrained than with the conventional test.

4.2 Linearity

Scatterplots with smoothers can also be used to check that the relationship between the explanatory variables and response variable is in fact linear. Figure 8 shows such a lineup testing the linearity of an observation-level explanatory variable. Out of 63 observers, 60 identified the true plot in panel #$(2^3 + 2)$, providing evidence that the mean structure is misspecified. This example comes from the dialyzer study and considers a model with only linear and quadratic terms for transmembrane pressure (see Appendix 1.4 for details on the data). Based on the data panel, it is clear that a higher-order polynomial is required. Once a polynomial of degree four is included in the mean structure of the fixed effects, the nonlinear pattern in the residuals is removed—a lineup of this situation is shown in Figure 5. In this lineup, the data plot is identified due to heteroscedasticity in the residuals. This highlights the flexibility of lineups due to the general phrasing of the alternative hypothesis. By tracking observers' reasons for the choice of plot in a lineup, we can distinguish between different alternatives. Using the same setup we will get test results based on where we are in the modeling process: as long as the mean structure is not correctly specified, it is most likely the distinguishing feature. Once the mean structure is properly specified, the lineup changes to test for homogeneity of variance.

To extend checks of linearity to group-level variables we suggest the use of the level-2 residuals.

4.3 Normality

Recall that in model (1) we assume that the random effects, $\boldsymbol{b}_i$, are a random sample from $\mathcal{N}(\mathbf{0},\ \boldsymbol{D})$ and are independent from the error terms, $\boldsymbol{\varepsilon}_i$, which are assumed to be a random sample from $\mathcal{N}(\mathbf{0},\ \sigma^2 \boldsymbol{R}_{n_i})$. In many situations, however, the predicted random effects are highly influenced (i.e., confounded) by the error terms. Consequently, traditional checks for normality such as Q-Q plots and the Anderson-Darling test are not appropriate. Formally this is seen by the strong assumptions that are required for the empirical distributions of the residuals in the linear mixed-effects model to converge in probability to their true distributions (Jiang, 1998, Theorem 3.2 and Lemma 3.1). If these assumptions are not satisfied, then the empirical distribution of the residuals may not resemble the hypothesized distribution under a properly specified model. When this occurs, individual Q-Q plots will often lead to erroneous conclusions about the distributional assumptions which can be overcome, at least partially, using lineups.

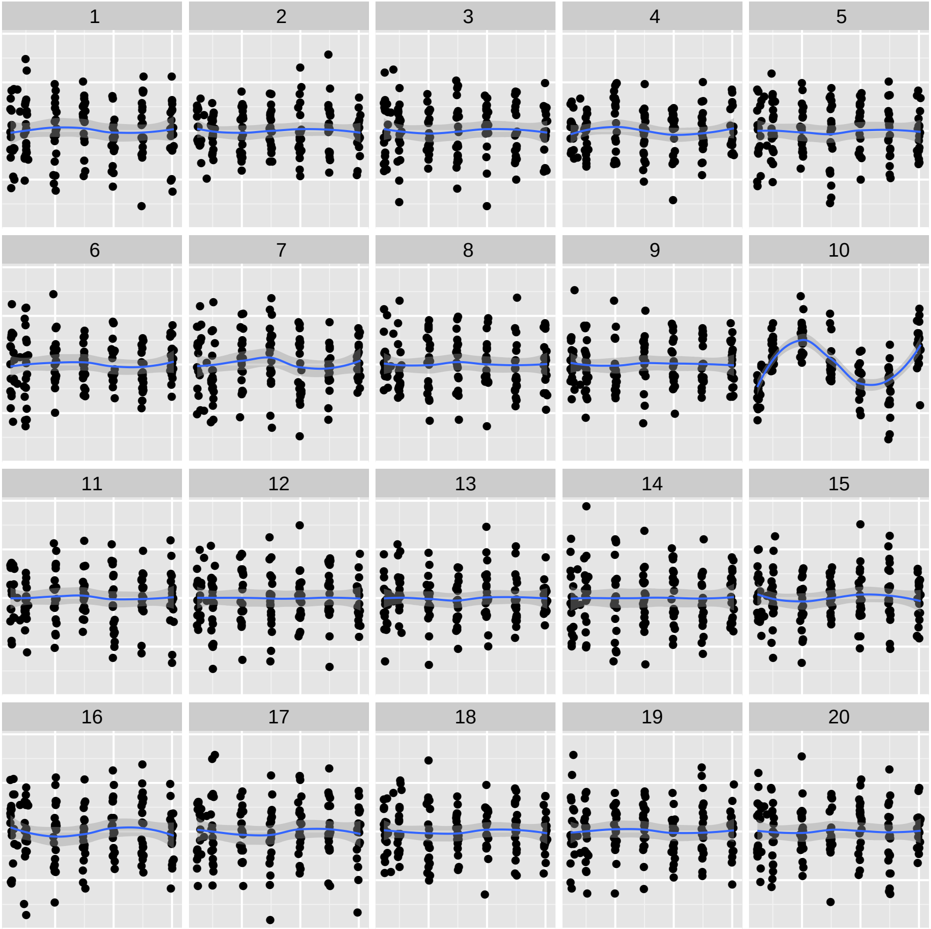


Figure 8: Lineup testing for nonlinearity of a covariate. Which of the plots is the most different? Which feature led you to your choice?

Figures 9 and 10 illustrate the use of lineups to test the distributional assumptions in a linear mixed-effects model. Figure 9 presents a lineup of the predicted random slopes from the radon study, in which group sizes are very unbalanced and there is a high degree of shrinkage. Confidence bands based on the normal distribution are applied to each lineup and reveal that the empirical distribution of the predicted random effects—both for the null plots and true plot—does not align with a normal distribution. While such confidence bands show the relationship of the predicted random effects to the hypothesized distribution, it is known that this is an ill conceived comparison in many cases; however, the lineups are not comparing the predicted random effects to the normal distribution, but

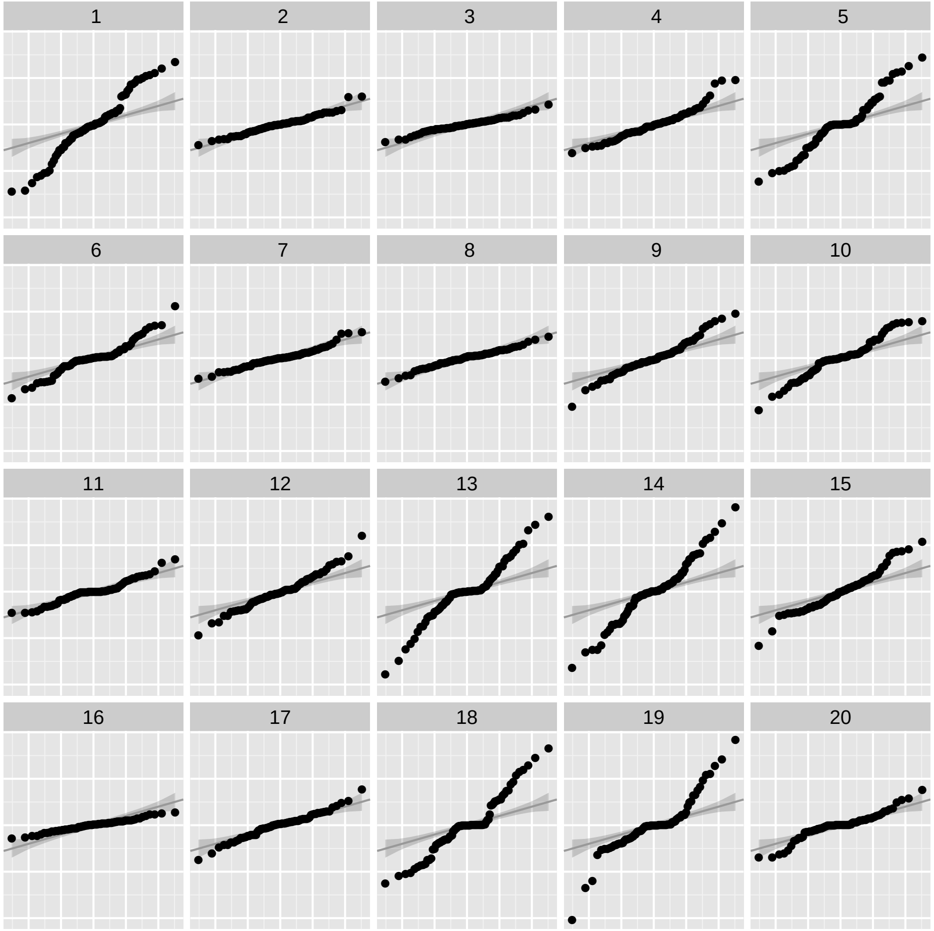


Figure 9: Lineup testing for normality of the random slope for the radon data. Which of the plots is the most different? Which feature led you to your choice?

rather are comparing the empirical distribution of the random effects between the null and observed plots. Consequently, the conclusions drawn from the lineups relate to evidence of consistency between the true plot and what is expected under a properly specified model. For example, the true plot in panel #$(2^4 - 6)$ in Figure 9 is indistinguishable from the null plots (none of 68 observers identified this plot), providing no evidence of a violation of normality; however, when compared only to the normal distribution, the observed Q-Q plot would be rejected by any standard test for normality (e.g., the p-value of the Anderson-Darling test is .0004 for the data panel). Panel #19 was identified most often from the lineup in Figure 9; it was picked by 29 out of 68 observers. Other panels selected at least

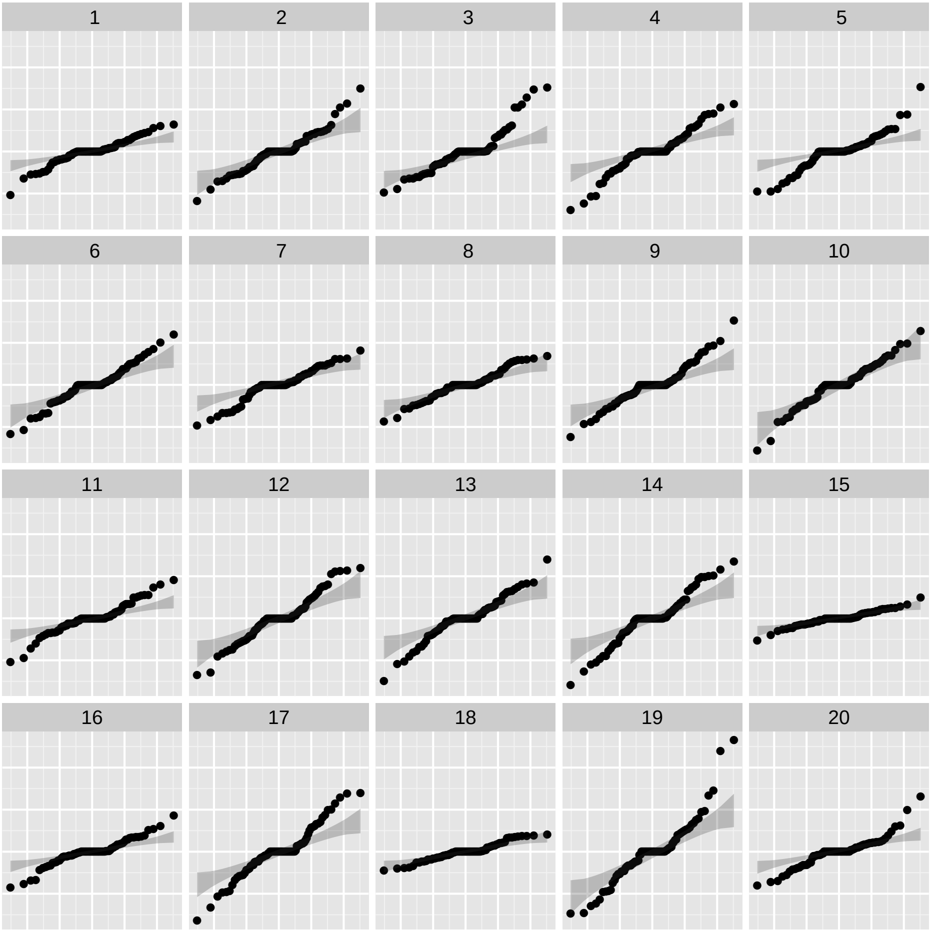


Figure 10: Lineup testing for normality of random slopes with error terms simulated from a t_3 distribution. Which of the plots is the most different? Which feature led you to your choice?

four times were #1, 13, 14, 16, and 18. Additionally, this example shows that even the null plots that were generated from a normal distribution no longer look normal after model estimation—in fact, 16 of the null plots fail the Anderson-Darling test of normality at a significance level of 0.05. This is due to confounding between the different levels of the mixed-effects model (Loy and Hofmann, 2015).

To determine whether lineups can detect distributional violations we constructed another lineup where the “true plot” was constructed from random effects simulated from a t_3 distribution. This lineup is shown in Figure 10. Twenty nine out of 70 observers identified the true plot in panel #($\sqrt{49} + 3 \cdot 4$), providing evidence against the null hypothesis of normality. The fact that we can distinguish the true plot from the null plots in Figure 10 indicates that lineups of Q-Q plots provide an avenue for distributional assessment where conventional methods fail. Further investigation is needed to explore limitations of this approach. The ability of a lineup to distinguish a t distribution for the random effects in the radon study shows that the approach has fewer limitations than conventional approaches, justifying our preference.

5 Discussion and Conclusion

We have presented a graphical approach to model selection and diagnosis, using lineups constructed by simulation from the model. Lineup tests provide us with the framework to test hypotheses and also allow a subsequent exploration of the plots in the lineup for additional insight into the data structure. This approach relies on the simulation process, the design of the graphics created, and observers, but avoids the reliance on asymptotic reference distributions; thereby circumventing the pitfalls of many commonly used tests.

The graphical approach is relatively new, and involves working with human observers recruited on the web. There is a vast experience in statistics in engaging subjects in a traditional lab setting, where researchers are closely engaged with the experiment. With MTurk, researchers are working with subjects in a long distance relationship, and there are some participants who will try to “game the system.” Nevertheless, results are promising, as MTurk experiments have replicated studies conducted in the traditional lab setting. Specifically, Heer and Bostock (2010) achieved matching results to those of Cleveland and McGill (1984) using MTurk, and Kosara and Ziemkiewicz (2010) found similar results between a lab study and one performed using MTurk.

As with conducting surveys, avoiding leading questions is very important. For this study, observers were generically asked to pick the plot with the most different features. Contextual information, such as labels, axis tick marks, and titles, were removed to avoid subjective bias. Care was taken, in constructing lineup sequences and different plot designs, so that observers saw the data only once. Because of the finite nature of possible comparison in a lineup, multiple replications of lineups (e.g., five, as in this study) are recommended using different sets of $m - 1$ null plots.

By asking observers for the distinguishing feature(s) of the plot they chose, graphical tests also provide us with information about the specific violation(s) of the null that is captured in a general alternative hypothesis. A fixed selection of reasons were provided for observers to choose from, but they could also enter a different reason in a text box. This allows us to examine responses by reasons for selection to assess individual sub-hypotheses, and investigate different types of errors. For example, in Figure 10 some participants chose plots because they showed the "most straight line" or "closest fit" to the line. These reasons would indicate that the person may not have had prior experience in reading Q-Q plots, and mistakenly looked for compliance with a theoretical distribution. This can be interpreted as a Type II error. Type III errors (Mosteller, 1948), where the null hypothesis is rejected for the wrong reason, can also be detected by participants' choices.

Plot design is important—some designs are better for revealing anomalies than others. At a population level, the results obtained from the same plot design are very stable. In this paper design choices were made using our best judgment based on experience and results from cognitive psychology. As more studies are conducted, more about the power of designs for model diagnostics will be learned, enabling more informed decisions about best practices (Hofmann et al., 2012). A further benefit of using the graphical framework for testing is that graphics adapt relatively well to big data situations (Unwin et al., 2006). This provides us with a viable approach to assess the practical relevance of a result versus results that show statistical significance purely based on the dimension of the problem. Barring bad design choices, graphical tests allow us to judge practical relevance of a result as "If we do not see it, it might be there but not be relevant."

The diagnostics in this paper draw from various sources. Some of the diagnostics presented here, are well known diagnostic tools, such as Q-Q plots or scatterplots of residuals with trend lines as suggested by Cook and Weisberg (1999) for ordinary least squares regression. Some diagnostics are suggestions from the literature specific to mixed effects models. An example of a new diagnostic addressing a practical need are the cyclone plots of Figure 6. The overarching purpose of these examples is to show graphical diagnostics in a wide variety of situations that all need special consideration in the conventional hypothesis testing setting, but that all fit within the same graphical inference framework. Many situations, such as outlier detection, were not discussed in this paper, but the results also extend to these problems. The reader is encouraged to examine more examples provided by Buja et al. (2009) and Majumder et al. (2013), and Roy Chowdhury et al. (2014) for biological applications.

6 Acknowledgments

We would like to thank Mahbubul Majumder and Eric Hare, who conducted the Amazon MTurk study (mahbub.stat.iastate.edu/feedback_turk11/homepage.html). R (R Core Team, 2013) was used to implement all analyses and methods discussed in this paper: lme4 (Bates et al., 2014a,b) and HLMdiag (Loy, 2013; Loy and Hofmann, 2014) were used for model fitting and calculation of diagnostics, respectively. ggplot2 (Wickham, 2009), nullabor (Wickham, 2012), and gridSVG (Murrell and Potter, 2013) provided the basis for the visualizations. This work was funded in part by National Science Foundation grant DMS 1007697. All studies were conducted with approval from the internal review board IRB 10-347.

References

Amazon.com, Inc (2005–2014). Mechanical Turk. https://www.mturk.com/mturk/welcome.

Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., Welch, K., and Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of Consulting and Clinical Psychology*, 75(4):594–604.

Anderson, D. K., Oti, R. S., Lord, C., and Welch, K. (2009). Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 37(7):1019–1034.

Bates, D., Maechler, M., and Bolker, B. (2011). *mlmRev: Examples from Multilevel Modelling Software Review*. R package version 1.0-1.

Bates, D., Maechler, M., and Bolker, B. (2012). *MEMSS: Data sets from Mixed-effects Models in S*. R package version 0.9-0.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014a). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014b). lme4: Linear mixed-effects models using Eigen and S4. *arXiv.org*.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., and 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.

Carithers, R. L., Herlong, H. F., Diehl, A. M., Shaw, E. W., Combes, B., Fallon, H. J., and Maddrey, W. C. (1989). Methylprednisolone therapy in patients with severe alcoholic hepatitis. *Annals of Internal Medicine*, 110(9):685–690.

Catellier, D. J. and Muller, K. E. (2000). Tests for gaussian repeated measures with missing data in small samples. *Statistics in Medicine*, 19(8):1101–1114.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Gelman, A. and Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., and Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford Review of Education*, 19(4):425–433.

Gomez, E. V., Schaalje, G. B., and Fellingham, G. W. (2005). Performance of the Kenward–Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics - Simulation and Computation*, 34(2):377–392.

Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 203–212, New York, NY, USA. ACM.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448.

Hofmann, H., Röttger, C. G., Cook, D., Buja, A., and Dixon, P. (2015). Distributions for visual inference under different lineup scenarios. *arXiv.org*.

Jiang, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statistica Sinica*, 8:861–886.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997.

Kosara, R. and Ziemkiewicz, C. (2010). Do mechanical turks dream of square pie charts? In *Proceedings BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 373–382. ACM Press.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17(2):624–642.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Institute, Cary.

Loy, A. (2013). *HLMdiag: Diagnostic tools for hierarchical (multilevel) linear models*. R package version 0.2.2.

Loy, A. and Hofmann, H. (2014). HLMdiag: A suite of diagnostics for hierarchical linear models in R. *Journal of Statistical Software*, 56(5):1–28.

Loy, A. and Hofmann, H. (2015). Are you normal? the problem of confounded residual structures in hierarchical linear models. *Journal of Computational and Graphical Statistics*, to appear(ja):xx–xx.

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

Meilgaard, M. C., Carr, B. T., and Civille, G. V. (2006). *Sensory Evaluation Techniques*. CRC Press, 4 edition.

Morrell, C. and Brant, L. (2000). Lines in random effects plots from the linear mixed-effects model. *The American Statistician*, pages 1–4.

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, 54(4):1560–1568.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics*, 19(1):58–65.

Murrell, P. and Potter, S. (2013). *gridSVG: Export grid graphics as SVG*. R package version 1.0-1.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, Thousand Oaks, 2nd edition.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., Lee, E.-K., and Toth, A. (2014). Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Computational Statistics*, pages 1–24.

Skene, S. S. and Kenward, M. G. (2010). The analysis of very small samples of repeated measurements I: An adjusted sandwich estimator. *Statistics in Medicine*, 29(27):2825–2837.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1177.

Unwin, A., Theus, M., and Hofmann, H. (2006). *Graphics of Large Datasets: Visualizing a Million*. Springer.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, pages 1–17.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2012). *nullabor: Tools for graphical inference*. R package version 0.2.1.

A Data sets

All of the data sets used in this paper are publicly available: the General Certificate of Secondary Education Exam data set is available in the R package mlmRev (Bates et al., 2011); the Dialyzer data set is available in the R package MEMSS (Bates et al., 2012); all other data sets can be found in the R package HLMdiag (Loy, 2013).

A.1 General certificate of secondary education exam data

We make use of a subset of examination results of 4,065 students nested within 65 inner-London schools discussed by Goldstein et al. (1993). The original analysis explored school effectiveness as defined by students' performance on the General Certificate of Secondary Education Exam (GCSEE) in both mathematics and English. This exam is taken at the end of compulsory education, typically when students are 16 years old. To adjust for a student's ability when they began secondary education, the students' scores on the standardized London Reading Test (LRT) and verbal reasoning group (bottom 25%, middle 50%, or top 25%) at age 11 were recorded. Additional information contained in the data set includes student gender, school gender, and the average LRT intake score for each school.

A.2 Autism study

In an effort to better understand changes in verbal and social abilities from childhood to adolescence, Anderson et al. (2007, 2009) carried out a prospective longitudinal study following 214 children between the ages of 2 and 13 who had been diagnosed with either autism spectrum disorder or non-spectrum developmental delays at age 2. The Vineland Adaptive Behavior Interview survey was used to assess each child's interpersonal relationships, play time activities, and coping skills, from which the Vineland Socialization Age Equivalent (VSAE) was computed as an overall measure of a child's social skills. Additionally, expressive language development at age 2 was assessed using the Sequenced Inventory of Communication Development (SICD) and the children were classified into three groups (high, medium, or low). Assessments were made on the children at ages 2, 3, 5, 9, and 13, however, not all children were assessed at each age. Additional information collected on each child includes: gender, race (white or non-white), and initial diagnosis at age 2 (autism, pervasive development disorder (pdd), or non-spectrum). We restricted attention to models concerned with the changes in social skills for subjects diagnosed with autism spectrum disorder having complete data. This results in a reduced data set of 155 children. For more detailed analyses we refer the reader to Anderson et al. (2007, 2009).

A.3 Methylprednisolone study

Carithers et al. (1989) conducted a four week longitudinal study to investigate the effectiveness of methylprednisolone to treat patients with severe alcoholic hepatitis. The researchers randomly assigned 66 patients to receive either methylprednisolone (35 patients) or a placebo (31 patients). Over the study duration, each subject's serum bilirubin levels (in μmol/L) were measured each week, with the first measurement taken at the start of the study (week 0).

A.4 Dialyzer study

Vonesh and Carter (1992) describe a study characterizing the water transportation characteristics of 20 high flux membrane dialyzers, which were introduced to reduce the time a patient spends on hemodialysis. The 20 dialyzers were studied *in vitro* using bovine blood at flow rates of either 200 or 300 ml/min, and the ultrafiltration rate (ml/hr) for each dialyzer was measured at seven transmembrane pressures (in mmHg). Vonesh and Carter (1992) use nonlinear mixed-effects models to analyze these data; however, they can be modeled using polynomials in the linear mixed-effects framework (see Littell et al., 2006, Section 9.5).

A.5 Radon study

The data consist of a stratified random sample of 919 owner-occupied homes in 85 counties in Minnesota. For each home, a radon measurement was recorded (in log pCi/L, i.e., log picoCuries per liter) as well as a binary variable indicating whether the measurement was taken in the basement (0) or a higher level (1). Additionally, the average soil uranium content for each county was available. The number of homes within each county varies greatly between counties ranging from one home to 116 homes, with 50% of counties having measurements from between 3 and 10 homes. Gelman and Pardoe (2006) suggest a simple hierarchical model allowing for a random intercept for each county and a random slope for floor level. This is the model from which we simulate predicted random effects.

B Experimental Setup and Results

B.1 Experimental Setup and Calculation of p-values

For each of the lineup designs described in the paper we constructed five replicates consisting of the same data plot and different sets of nineteen null plots for a total of 75 different lineups. These were evaluated by 487 participants in altogether 4927 evaluations. For each lineup, observers were instructed to identify the plot most different from the set and asked what feature led them to their choice. These choices came in the form of four suggestions (in checkboxes) and one text box for a free-form answer. For each lineup the time taken to answer was recorded and observers were asked for their confidence level (on a scale from 1=weak to 5=high). Observers were also asked to provide their demographics: age category, gender, education range, and geographic location (from parts of the ip address).

The results of the evaluations for all lineups are displayed in Table 2 and the observers' reasons for identifying plots are summarized in Table 3. The significances in Table 2 are based on the number of evaluations and the number of times that the data plot was identified. For that, we the introduce the random variable Y as the number of evaluations of a lineup in which the observer identifies the data plot. Assume that the lineup has size $m = 20$, and it is shown to a total of K independent observers. Then Y has a Visual distribution $V_{K,m,s=3}$ as defined in Hofmann et al. (2015), where s delineates scenario III—i.e., the same lineup is shown to all K observers. The p-values in Table 2 for each of the five replicates are calculated this way. For the overall p-value, we use a simulation based approach to combine the five results. Treating the number of evaluations $(K_1, ..., K_5)$ as fixed, we simulate assessments of lineups without signal as follows: we assume that the signal in a plot is complementary to its p-value, which is i.i.d. $U[0, 1]$ under the null hypothesis. We further assume that the probability an observer picks a plot is allocated proportionally to its signal. For each "data plot" we create five sets of null plots to be evaluated simultaneously $(K_1, ..., K_5)$ times. p-values are then based on a comparison of the sum of data picks from five no-signal lineups and the observed number of data picks from the actual lineups. The column on the right of Table 2 shows p-values based on 10^5 simulation runs.

B.2 Additional lineups included in the study

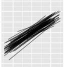
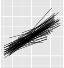
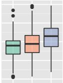
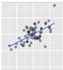
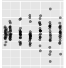
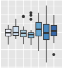
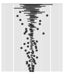
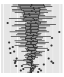
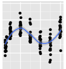
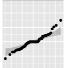
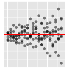
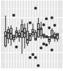
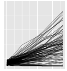
This section includes two lineups that were included in the MTurk study, but were not discussed in the paper.

Figure 11 contains a box plot representation of the same data as Figure 5, but categorizes pressure into seven categories, and shows residuals in the form of box plots. In order to preserve the appearance of continuity on the x-axis we used a color scheme to fill the boxes with deepening shades of blue from left to right. In this form 23 out of 70 observers identify the plot of the data. This is consistent with the other design.

Figure 12 displays another lineup testing the adequacy of the random effects specification (see Section 3.2) using data from the autism study. The null plots were generated from a model containing only a linear random slope, so if the true plot in panel $\#(\sqrt{16} + 12)$ is identified it provides support for the inadequacy of this specification, and the need for additional random effects.

Figures 13 and 14 show another example of testing for homogeneity in the variance following the approach taken in Section 4.1. Both of these lineups are based on the dialyzer data. Level-1 residuals are plotted by subject. Subjects are ordered by variance—i.e., we get some structure that might be taken for differences in variability, that are really just differences due to the imbalance in group size. If any panel of this lineup is considered separately, an analyst may come to the conclusion that the within-group variance increases across the x axis. However, inserting the true plot into the lineup forces the analyst to consider this particular feature as inherent to the data structure rather than evidence against a hypothesis of homogenous variance. The dot plot version is not significant, but the box plot version is. When participants in the box plot design identify the data plot,

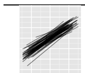
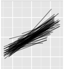
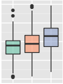
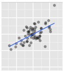
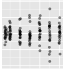
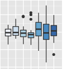
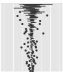
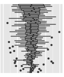
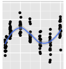
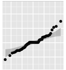
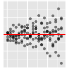
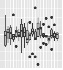
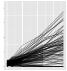
Table 2: Overview of all lineup evaluations. Ratios comparing the number correct to the total number of evaluations are shown. p-values and significances are based on the calculations as described in Section 2.1

	Replicate					Overall
Lineup	1	2	3	4	5	p-values
 fig. 1	11/73 *	7/70	9/65 *	13/66 **	7/72	0.0001
 fig. 3	0/64	7/75	11/76 *	0/69	0/60	0.4837
 fig. 2	65/74 ***	57/66 ***	62/68 ***	54/63 ***	66/76 ***	$< 10^{-5}$
 fig. 4	41/69 ***	62/73 ***	31/64 ***	57/75 ***	55/67 ***	$< 10^{-5}$
 fig. 5	29/85 ***	10/65 *	24/64 ***	7/60 .	13/72 **	$< 10^{-5}$
 fig. 11	23/70 ***	9/74 .	11/62 **	31/78 ***	25/61 ***	$< 10^{-5}$
 fig. 6	50/75 ***	44/64 ***	45/68 ***	46/76 ***	50/67 ***	$< 10^{-5}$
 fig. 7	1/59	2/79	2/68	4/62	1/73	0.7180
 fig. 8	60/63 ***	62/64 ***	57/60 ***	84/88 ***	66/69 ***	$< 10^{-5}$
 fig. 10	29/70 ***	52/79 ***	0/64	13/59 ***	6/74	$< 10^{-5}$
random intercept	0/72	1/75	0/68	0/75	2/61	0.9458
random slope	0/65	0/64	0/68	0/46	0/64	1.0000
 fig. 13	0/57	1/71	8/75	0/59	2/68	0.7078
 fig. 14	23/61 ***	38/72 ***	29/59 ***	22/70 ***	35/62 ***	$< 10^{-5}$
 fig. 12	50/79 ***	26/59 ***	31/67 ***	32/72 ***	35/71 ***	$< 10^{-5}$

Signif. codes: $0 \leq$ *** $\leq 0.001 \leq$ ** $\leq 0.01 \leq$ * $\leq 0.05 \leq$. $\leq 0.1 \leq$ ' ' ≤ 1

about 45.3% give outliers as the reason for their choice. In contrast to that, outliers, a large spread or a trend are in a three-way tie for the reason for identifying the data plot in

Table 3: Percent of data picks, given the reason for the choice of plot from the lineup.

Lineup	Outlier	Spread	Trend	Asymmetry	Other
 fig. 1	13.5	24.0	10.9	5.8	15.9
 fig. 3	2.4	10.6	3.5	5.4	0.0
 fig. 2	73.3	95.5	92.5	91.0	80.7
 fig. 4	64.5	34.3	83.4	70.7	64.9
 fig. 5	18.8	27.0	29.1	25.8	19.0
 fig. 11	26.7	24.6	30.1	43.7	0.0
 fig. 6	49.6	50.0	77.8	79.4	91.8
 fig. 7	4.2	0.6	2.4	6.7	0.0
 fig. 8	84.0	87.4	98.3	98.1	100.0
 fig. 10	37.8	49.2	19.6	4.4	10.9
random intercept	1.5	0.0	1.1	0.0	0.0
random slope	0.0	0.0	0.0	0.0	0.0
 fig. 13	2.4	3.0	7.6	4.8	0.0
 fig. 14	56.8	55.1	20.1	33.9	22.7
 fig. 12	38.9	33.8	56.8	78.5	70.0

the lineup with the dot plot design.

The reason for the box plot design being so much more significant might not be so much of an issue of homogeneity being violated as much as a difference in the error distribution between the data and the nulls. Null data come from a parametric bootstrap where residuals are simulated under a normal error assumption. Outliers in small samples are indicative of the sample being from a distribution with heavy tails. Regardless of the reasoning, the

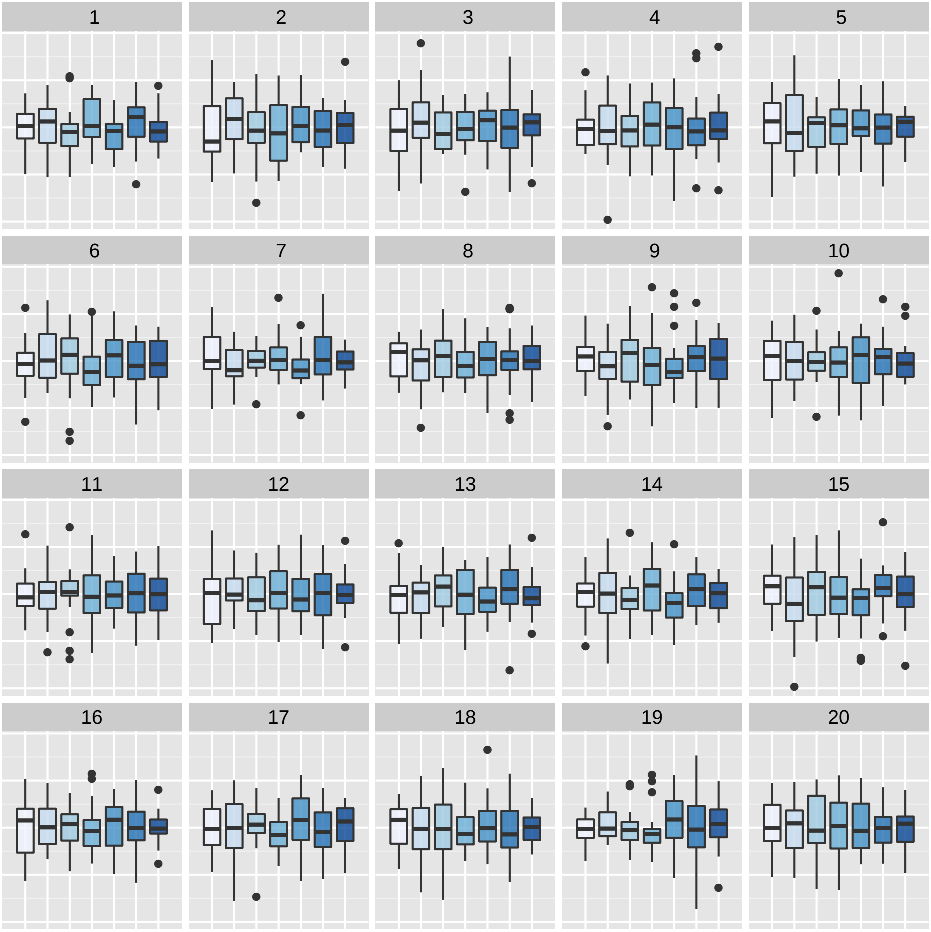


Figure 11: Alternative box plot Lineup testing homogeneity of the level-1 residuals. Which of the plots is the most different? Which feature led you to your choice?

second lineup design enables us to diagnose a problem with the model that makes the data stand out from a set of nulls. The two designs therefore represent two very similar tests with different power.

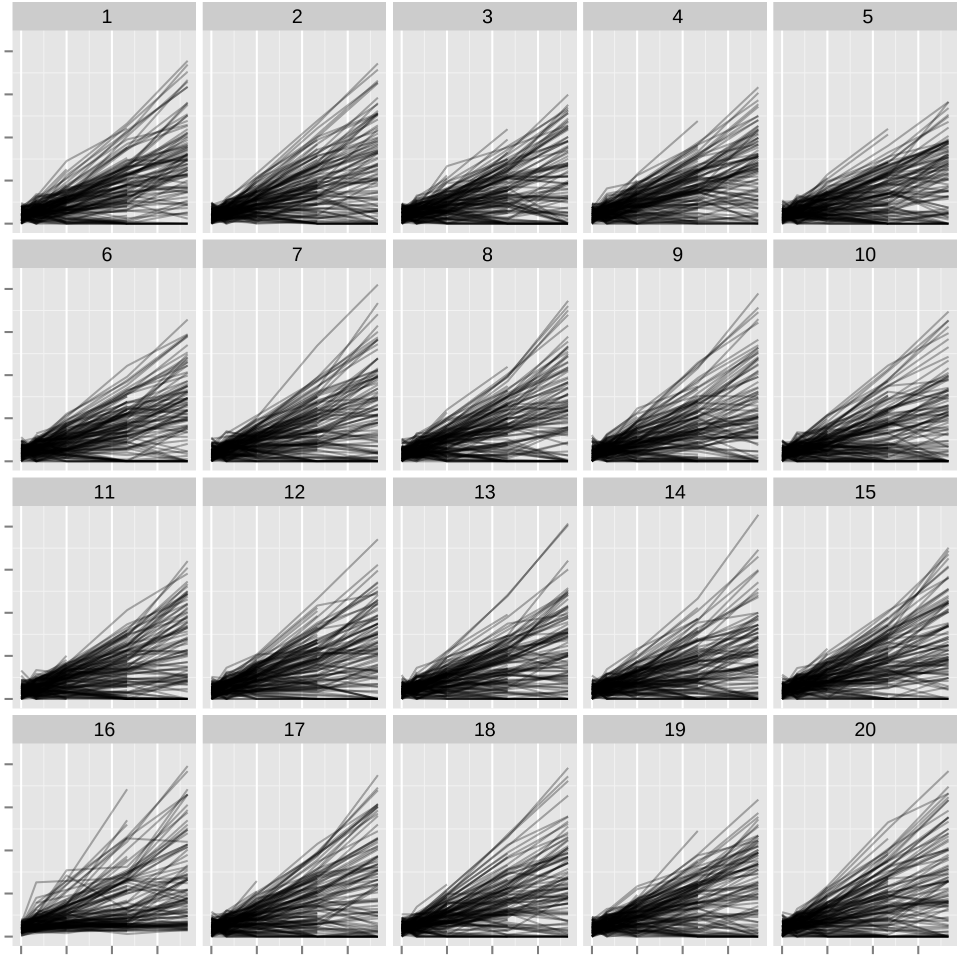


Figure 12: Which of the plots is the most different? Which feature led you to your choice?

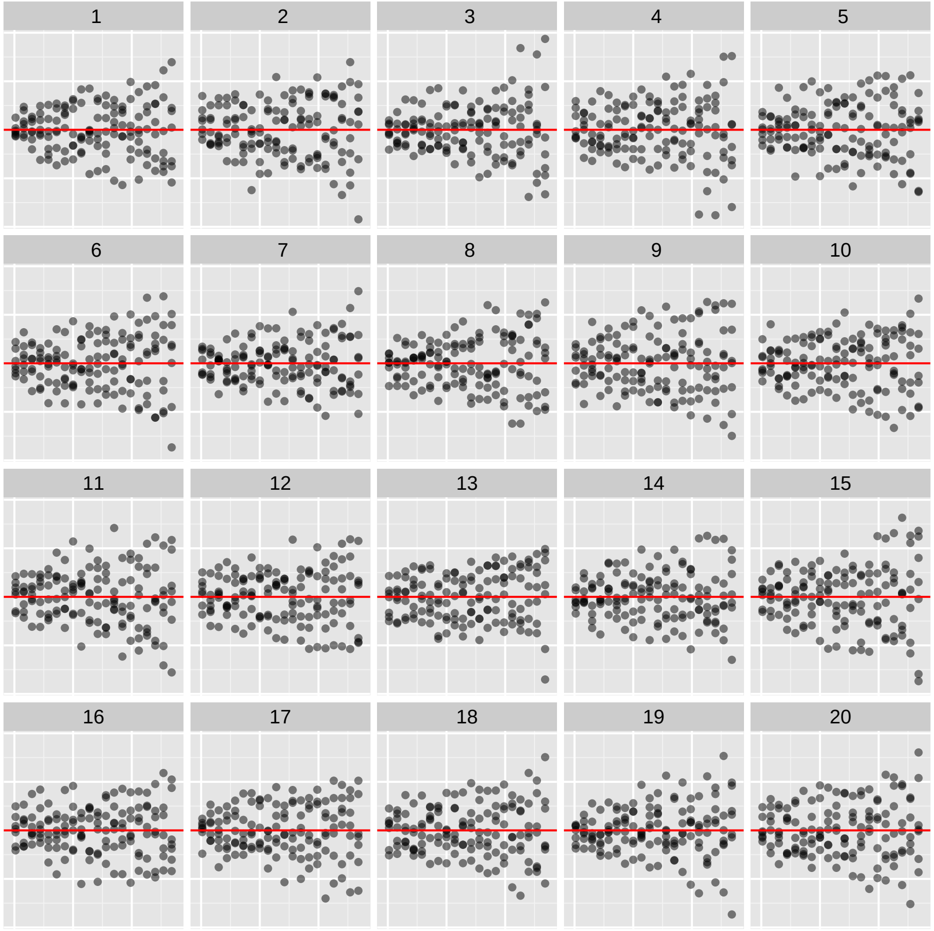


Figure 13: Lineup testing homogeneity of the level-1 residuals. Which of the plots is the most different? Which feature led you to your choice?

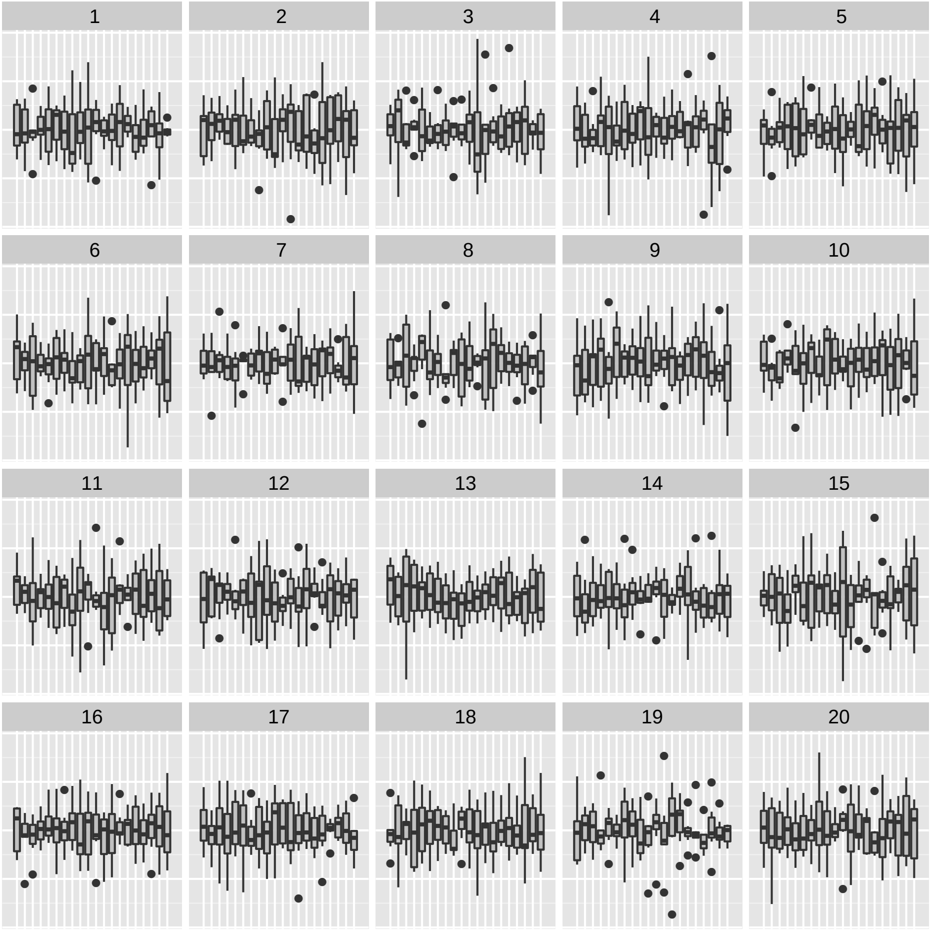


Figure 14: Lineup testing homogeneity of the level-1 residuals. Which of the plots is the most different? Which feature led you to your choice?