Are you Normal?
The problem of confounded residual structures in hierarchical models.

Adam Loy and Heike Hofmann
Department of Statistics, Iowa State University

April 1, 2013

Abstract

We encounter hierarchical data structures in a wide range of applications. Regular linear models are extended by random effects to address correlation between observations in the same group. Inference for random effects is sensitive to distributional mis-specifications of the model, making checks for (distributional) assumptions particularly important. The investigation of residual structures is complicated by the presence of different levels and corresponding dependencies. Ignoring these dependencies leads to erroneous conclusions using our familiar tools, such as Q-Q plots or normal tests. We first show the extent of the problem, then we introduce the *fraction of confounding* as a measure of the level of confounding in a model and finally introduce minimally confounded residuals as a solution to assessing distributional model assumptions.

1 Introduction

There is a wide range of application areas—from the biological and physical sciences to the social sciences—in which we encounter nested data. Whether it is quality control in a manufacturing process that involves the monitoring of a set of components over time or students' performances in different schools across the country, analysts have to account for the correlation between observations in the same group. Hierarchical linear models allow us to do exactly that—but they also require us to make distributional assumptions on both the error terms and the random effects. These assumptions must hold to ensure the validity of the model. Inference for the fixed effects in linear mixed models is fairly robust against model mis-specification (Butler and Louis, 1992; Verbeke and Lesaffre, 1997). This is different for random effects: they are sensitive to distributional mis-specifications and therefore have to be checked carefully.

Quantile-quantile (Q-Q) plots (Wilk and Gnanadesikan, 1968) are our main graphical tool for evaluating a specific distributional assumption. For that, we plot the empirical distribution against expected quantiles. In hierarchical models the investigation of residual structures is complicated by the presence of different levels. The nested structure of the data is reflected in the residual structure. And just as there is dependency between different levels in the data, we can expect dependencies between different levels in the residual structure. Q-Q plots, weighted (Dempster and Ryan, 1985; Lange and Ryan, 1989) or unweighted, do not account for this, which might lead us to erroneous conclusions in evaluating normality based on the plots (or any other test).

In this paper, we address the problem of distributional assessment due to confounding in residual structures. First, we illustrate the inadequacy of existing methods based on the predicted residuals. We then introduce the concept of least confounded residuals for the random effects and present a general method to obtain least confounded residuals for residuals at all levels of the model. We demonstrate how this enables an appropriate (graphical) assessment of distributional assumptions.

Radon Data To illustrate the effect of confounding between different levels of residuals, we consider the data set presented by Gelman and Pardoe (2006). This data set consists of a stratified random sample of 919 owner-occupied homes nested within 85 counties in Minnesota, for which the authors suggested a hierarchical model of the form

$$\log(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + b_{0i} + b_{1i} + \varepsilon_{ij} \tag{1}$$

where $\log(y_{ij})$ denotes the (log pCi/L, i.e log picoCurie per litre) radon measurement for house j $(1 \leq j \leq n_i, 1 \leq i \leq 85)$ within county i, x_{1ij} is a binary variable describing the level at which the measurement was taken (0 for the basement and 1 for a higher level) and x_{2i} denotes the average soil uranium content for county i. Further, we assume i.i.d. normal errors $\varepsilon_{ij} \sim \mathcal{N}(0,\ \sigma_\varepsilon^2)$ and $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{D})$, where $\boldsymbol{D}$ allows for correlation between b_{0i} and b_{1i}. We also assume independence between random effects and errors.

A map of counties in Minnesota is given in figure 1. The color shading represents average radon activity in a county. For two counties no data is available. Generally, more southern locations exhibit higher levels of Radon activity. Figure 2 focuses on the two counties of Hennepin (home to Minneapolis) and Winona (home to the city of the same name). Radon levels are plotted by floor level. Radon levels are usually the highest at the basement level of a house. The within-county sample sizes, n_i, are extremely unbalanced, ranging from one house to 116 houses, with 50% of the counties having between three and ten houses. Such unbalanced designs are common in applications, and result in a high degree of pooling for the predicted county-level random effects. This leads to dependence between predicted random effects and error terms (cf. eqns. 3 and 4), which in turn can lead us to draw erroneous conclusions for corresponding residual quantities.

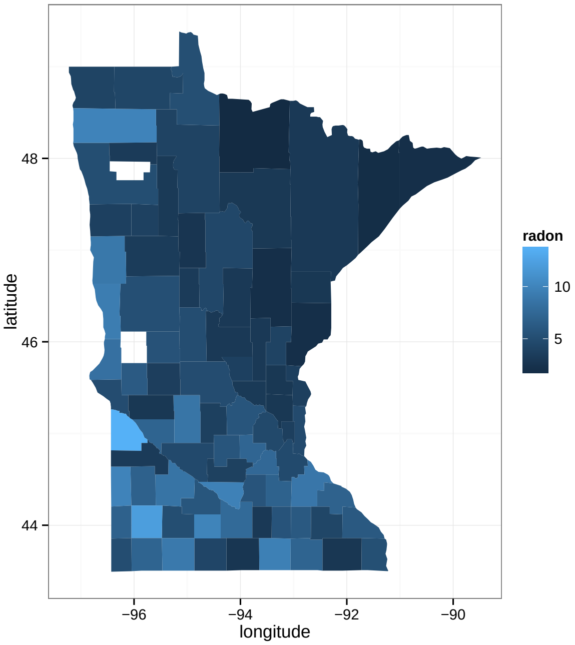


Figure 1: Map of counties in Minnesota, color shading is average Radon activity.

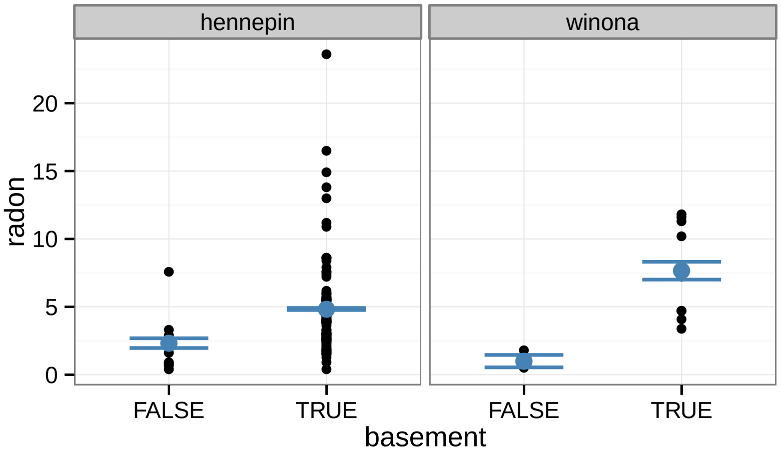


Figure 2: Activity of radon levels for the two counties of Hennepin and Winona at basement levels and higher. Radon levels at the basement level are usually higher.

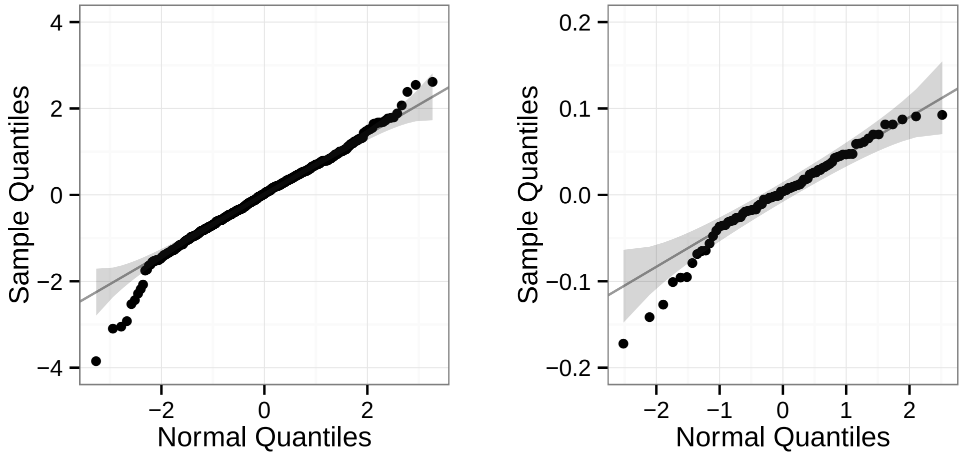
In this example, Q-Q plots (Figure 3) for the residuals show that normality seems to be violated for the error terms and both random effects. But is this cause for concern? There is little pooling at the observation level, and we would therefore expect the distributional assessment of the error terms to be reliable, but for the random effects the high degree of pooling casts doubt on the reliability of their Q-Q plots. We find our doubts increasing in a simulation-based assessment of distributions for residuals and random effects.

Figure 4 shows a lineup (Buja et al., 2009) of 20 Q-Q plots for the predicted random slope. The Q-Q plot of the observed random slopes is placed among nineteen decoy plots of parametric bootstrap samples based on model (1) satisfying the normal distribution assumptions. The simulation parameters were set to the maximum likelihood estimates of model (1). The observed Q-Q plot (panel $12 + 2^2$) is virtually indistinguishable from the field of null plots. This suggests that the predicted random slopes from the data do not deviate significantly from model (1). Note that in practice we must blind ourselves from the true plot for proper use of lineups. In order to not violate this, we did not show the Q-Q plot of random slopes earlier.

What becomes apparent from the lineup, is that, astonishingly, *none* of the null plots conforms to normality. Further investigation revealed that a proportion far above the nominal 0.05 of distributions of random intercepts fail the normality tests, too.

Table 1 shows results from 1000 parametric bootstrap samples of model 1 under normal errors. For each simulated data set, we evaluated the assumption of normality for both the level-1 and -2 residuals using the Anderson-Darling (AD), Cramér von Mises (CVM), Kolmogorov-Smirnov (KS), and Shapiro-Wilks (SW) tests for normality. Type I error rates are hugely inflated for both random effects, making an assessment of normality based on the empirical distribution impossible. In the example we are able to use the empirical distribution for assessing normality of level-1 residuals as pooling is minimal at this level. In situations with higher levels of pooling, this will not be the case.

In the remainder of this paper we investigate the root of concern that leads to the distributional deviations, and derive residuals that address the issues introduced by pooling, allowing for a familiar



(a) Predicted error terms

(b) Random intercepts

Figure 3: Q-Q plots of predictions at different levels for model (1). Note that random slopes (not shown) exhibited the largest deviation from normality – see figure 4.

Table 1: Proportions of tests rejecting the null hypothesis of normality of the predicted error terms and random effects at the nominal .05 significance level. Type I error rates are hugely inflated.

	Test			
Residual	SW	AD	CVM	KS
Error term	0.06	0.06	0.06	0.05
Random intercept	**0.48**	**0.48**	**0.46**	**0.35**
Random slope	**0.75**	**0.75**	**0.75**	**0.68**

graphical assessment of these distributions again.

2 Rotated Residuals

The general stacked representation of the hierarchical linear model is given by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}, \tag{2}$$
$$\mathrm{E}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \boldsymbol{0},\ \mathrm{Cov}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\boldsymbol{D} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of observed responses, $\boldsymbol{X}$ $(n \times p)$ and $\boldsymbol{Z}$ $(n \times q)$ are design matrices, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{b}$ is a $q \times 1$ vector of unobserved random effects, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of unobserved errors, and $\boldsymbol{R}$ and $\boldsymbol{D}$ are positive definite covariance matrices.

Using this specification, the predicted error terms and random effects are given by

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{RPy} = \boldsymbol{RPZb} + \boldsymbol{RP\varepsilon} \tag{3}$$
$$\widehat{\boldsymbol{b}} = \boldsymbol{DZ'Py} = \boldsymbol{DZ'PZb} + \boldsymbol{DZ'P\varepsilon} \tag{4}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})$. This set of equations reveals the inherent dependence between the the residuals defined above. Additionally, it is easily seen that both (3) and (4) lead

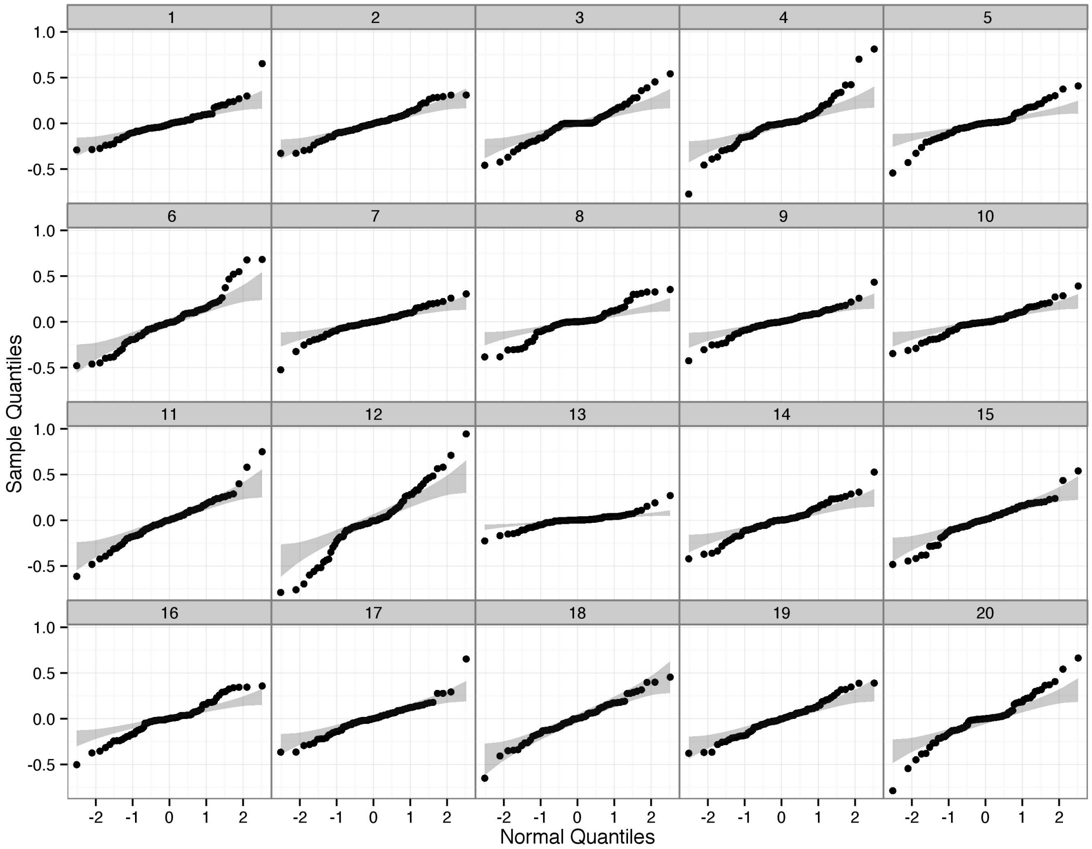


Figure 4: Lineup of normal Q-Q plots for the random slope term in model (1). The 19 null plots were obtained by simulation from the model. Can you identify the observed Q-Q plot?

to the analysis of correlated and potentially heteroscedastic disturbances as $\text{Var}(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{RPR}$ and $\text{Var}(\widehat{\boldsymbol{b}}) = \boldsymbol{DZ'PZD}$.

To combat the issues presented above, we derive a reduced set of minimally confounded residuals that are standardized, uncorrelated, and homoscedastic. Hilden-Minton (1995) presents the derivation of a similar set of minimally confounded residuals for the error terms, but did not address the random effects. The derivation is relegated to the appendix, and presents a more general method that applies to residuals at all levels of the model.

First, we define the *fraction of confounding* (Hilden-Minton, 1995) which is minimized in the result below.

Definition 1 (Fraction of confounding)**.** *For the ith element of the target residual vector,* $\widehat{\boldsymbol{e}}$, *the fraction of confounding is given by*

$$FC(\widehat{e}_i) = \frac{\boldsymbol{v_i'}\text{Var}(\widehat{\boldsymbol{e}}|\boldsymbol{e})\boldsymbol{v_i}}{\boldsymbol{v_i'}\text{Var}(\widehat{\boldsymbol{e}})\boldsymbol{v_i}} = \frac{\boldsymbol{v_i'Av_i}}{\boldsymbol{v_i'Bv_i}} \tag{5}$$

where $\boldsymbol{v_i}$ *is the ith column of the identity matrix.*

The fraction of confounding measures the contribution of the non-target residual to the variance of the target residual. FC $\in [0, 1]$, where 1 indicates there is no additional information on the target residual due to confounding, while 0 indicates no confounding. An overall measure of the amount of confounding is given below.

Definition 2. *For the target residual vectore, $\widehat{\boldsymbol{e}}$, the fraction of confounding is given by*

$$FC(\widehat{\boldsymbol{e}}) = \frac{1}{\ell}\sum_i \frac{\boldsymbol{v_i'Av_i}}{\boldsymbol{v_i'Bv_i}}. \tag{6}$$

where ℓ is the length of vector e.

In order to correct residuals for the impact of confounding, we employ weights $\boldsymbol{w_i}$ to each of the residual contributions. This leads to the following minimization problem:

$$\mathrm{FC}_{\boldsymbol{W}}(\widehat{\boldsymbol{e}}) = \min_{W \in \mathbb{R}^{\ell - r \times \ell}} \sum_i \frac{\boldsymbol{v_i'WAW'v_i}}{\boldsymbol{v_i'WBW'v_i}} \tag{7}$$

where ℓ is the length of vector $\boldsymbol{e}$ and $r = \mathrm{rank}(\boldsymbol{B})$.

Theorem 1. *The minimization problem (7) is solved by*

$$\boldsymbol{W} = \boldsymbol{T_r \Lambda_r}^{-1/2} \boldsymbol{U}$$

where $\boldsymbol{T_r\Lambda_r}^{-1/2}$ is the full rank inverse square root of matrix $\boldsymbol{B}$ found through the spectral decomposition of $\boldsymbol{B}$. $\boldsymbol{U}$ is the matrix of all nonzero eigenvectors of $\boldsymbol{A}^* = (\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{T_r'})\boldsymbol{A}(\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{T_r'})'$. The resulting residuals are standardized, uncorrelated, and homoscedastic.

A proof of theorem 1 is given in the appendix.

Radon Data Revisited: We now use the results from Theorem 1 to assess the amount of confounding in the radon example and examine distributional assumptions for the minimally confounded case. Figure 5 shows the marginal minimally confounded residuals for the random intercept (left) and slope (right). Neither Q-Q plot exhibits any indication of departures from normality. This agrees with the assessments based on the lineups.

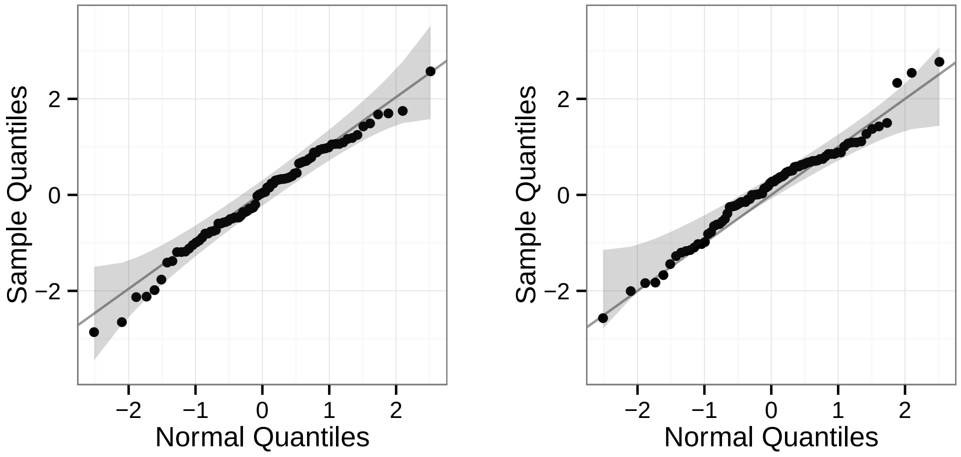
Table 2 shows the results from the previous simulations. After rotating into the space of least confounding tests for normality of the residual structures now reject at the correct nominal level of 0.05 again.

Table 2: Proportions of tests rejecting the null hypothesis of normality of the predicted error terms and random effects at the .05 significance level.

	Test			
Residual	SW	AD	CVM	KS
Error term	0.06	0.05	0.05	0.04
Random intercept	**0.04**	**0.05**	**0.05**	**0.05**
Random slope	**0.04**	**0.04**	**0.05**	**0.05**

3 Exploring the Rotation Matrix

The interpretation of minimally confounded residuals for purposes other than distributional assessment is complicated by the fact that they are linear combinations of residuals; however, the



(a) Random intercepts

(b) Random slopes

Figure 5: Normal Q-Q plots of the minimally confounded random structures. The deviation from normality are much less pronounced than before – in fact, it is questionable whether there is a significant amount of deviation; the results from the normal tests are not conclusive.

weights, $\boldsymbol{W'}$, are known, so interpretation is possible. Notice that the rows of $\boldsymbol{W'}$ give the weights specifying each linear combination and the columns give the weights assigned to each raw residual. Thus, the columns provide insight into the overall contribution of each raw residual and the rows of $\boldsymbol{W'}$ provide additional diagnostic information that, for example, can be used to investigate identified. In this section, each aspect will be investigated. To begin we explore the columns to better understand the rotation.

The average weights for each raw residual (column) show that residuals receiving larger weights are located in the center of the distribution of the raw residuals, while residuals receiving less weight are in the tails of the distribution. This is illustrated in Figure 6 for the predicted random effects using a linked histogram of average weights (right) and Q-Q plot of raw residuals (left). To gain further insight, we must consider the entries of weight matrix.

Figure 7 is a heat map of the matrix of weights $\boldsymbol{W}$ for predicted random slopes ordered by the variance of the predicted random slopes (with highest variance on the right). Note that entries are scaled to their signed square root because of skewness in the weights. A block diagonal structure becomes apparent in the heat map. This indicates that terms with similar variances are treated similarly in the rotation to the least confounded direction. For example, Wilkin, Murray, and Mahnomen counties have the most variable predictions and are clustered together at the top right. Additionally, the horizontal lines that appear in the top portion of the heat map align with more confounded residuals, showing how information is combined to address this issue. Looking at the rows of the heat map will also aid in the interpretation of outlying points to determine whether a few observations are dominant or if many observations contribute equally.

4 Discussion

We have shown that minimally confounded residuals are standardized, uncorrelated, homoscedastic residuals that address the concerns arising from the use of residuals for distributional assessment

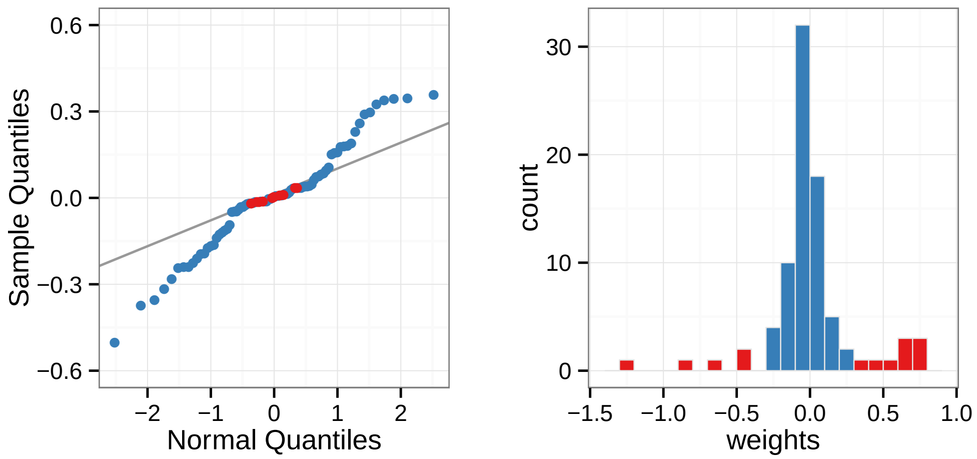


Figure 6: Q-Q plot (left) and histogram (right) of the average weight for each raw residual given by $\boldsymbol{W}'$. The largest weights (red) correspond to residuals in the central region of the distribution of raw residuals.

in the presence of pooling. The minimally confounded residuals can be used with Q-Q plots, which are familiar to analysts. While these residuals simplify the investigation of distributional assessment, they are linear combinations of residuals, and as such, are not directly applicable to the assessment of other aspects of the model, such as outlier detection; however, an investigation of the weights enables further investigation of unusual observations. It is important to note that goodness-of-fit tests are available for residuals at each level of the model (Jiang, 2001), but they do not lend themselves to graphical inspection, which is the focus of this paper, and the largely preferred method of distributional assessment.

Appendix: Proof of Theorem I

We present the proof of theorem 1. To remain general we present our derivation in terms of the residual under investigation (the target residual), $\widehat{\boldsymbol{e}}$, which can refer to either $\widehat{\boldsymbol{\varepsilon}}$ or $\widehat{\boldsymbol{b}}$.

We want to find a full rank matrix $\boldsymbol{W}$ that minimizes

$$\min_{W \in \mathbb{R}^{\ell - r \times \ell}} \sum_i \frac{\boldsymbol{v_i'} \boldsymbol{W} \mathrm{Var}(\widehat{\boldsymbol{e}}|\boldsymbol{e}) \boldsymbol{W'} \boldsymbol{v_i}}{\boldsymbol{v_i'} \boldsymbol{W} \mathrm{Var}(\widehat{\boldsymbol{e}}) \boldsymbol{W'} \boldsymbol{v_i}} \tag{8}$$

Proof. Let $\boldsymbol{A} = \mathrm{Var}(\widehat{\boldsymbol{e}}|\boldsymbol{e})$, $\boldsymbol{B} = \mathrm{Var}(\widehat{\boldsymbol{e}})$, $r = \mathrm{rank}(\boldsymbol{B})$, and $\ell =$ the number of elements in $\widehat{\boldsymbol{e}}$. By definition $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and nonnegative definite. Write the spectral decomposition of $\boldsymbol{B}$ as $\boldsymbol{B} = \boldsymbol{T_r \Lambda_r T_r'}$, where $\boldsymbol{\Lambda_r}$ is a diagonal matrix of the nonzero eigenvalues and $\boldsymbol{T_r}$ is the matrix of associated eigenvectors.

Define $\boldsymbol{F} = \boldsymbol{T_r \Lambda_r^{1/2}}$, which is a full-rank decomposition of $\boldsymbol{B}$. The Moore-Penrose inverse of $\boldsymbol{F}$ is given by

$$\boldsymbol{F}^- = (\boldsymbol{F'F})^{-1}\boldsymbol{F'} = \boldsymbol{\Lambda_r^{-1/2} T_r'}$$

Now, notice that

$$\boldsymbol{F}^- \boldsymbol{B} (\boldsymbol{F}^-)' = \boldsymbol{\Lambda_r^{-1/2} T_r'} (\boldsymbol{T_r \Lambda_r T_r'}) \boldsymbol{T_r \Lambda_r^{-1/2}} = \boldsymbol{I}$$

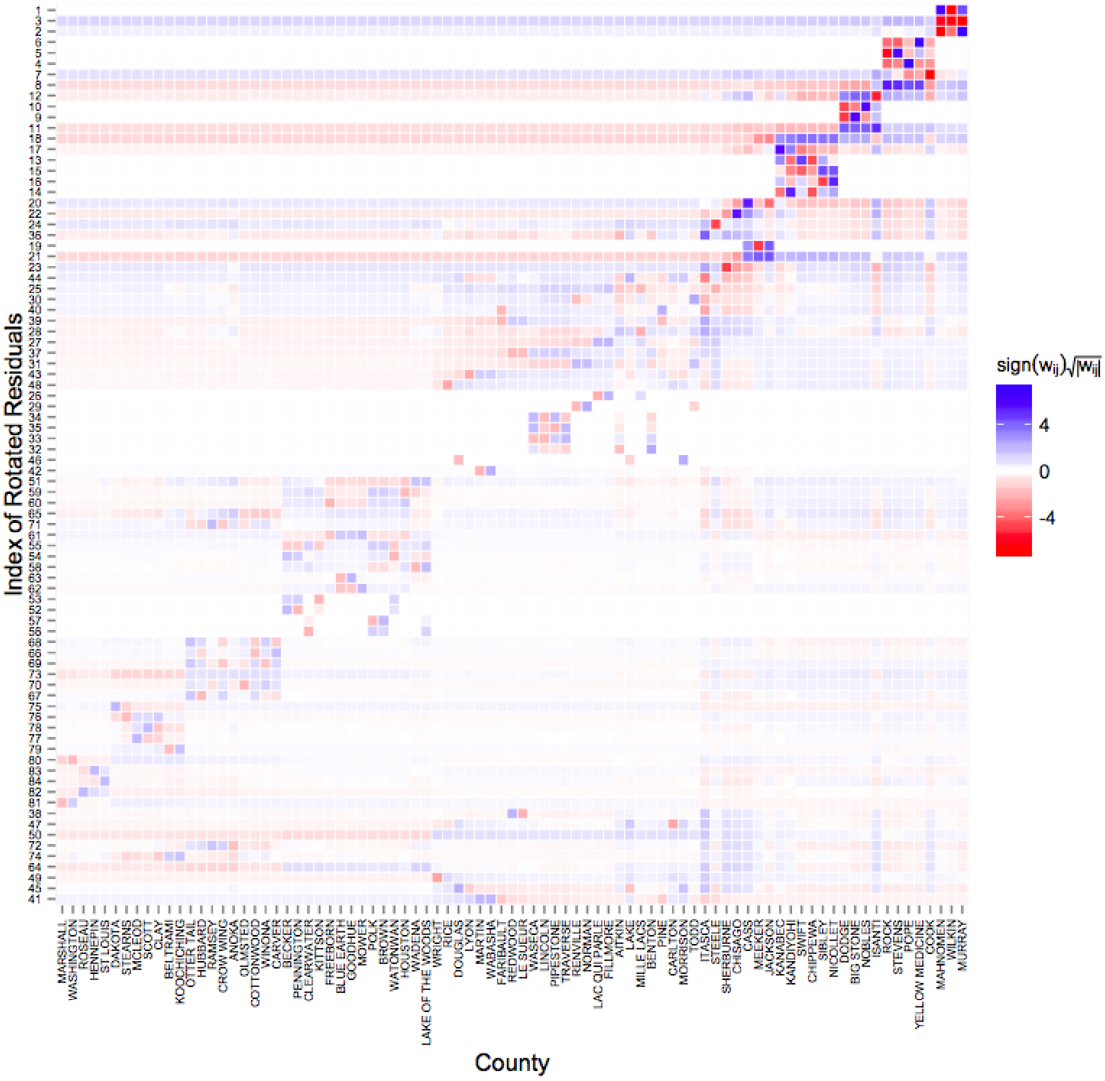


Figure 7: Heat map of $\boldsymbol{W'}$ for the predicted random slopes ordered by the variance of the predicted random effects. The entries were transformed to improve interpretability.

If we consider $\boldsymbol{w}_i = (\boldsymbol{F}^{-})'\boldsymbol{u}_i$, then we find that

$$\frac{\boldsymbol{w_i'}A\boldsymbol{w_i}}{\boldsymbol{w_i'}B\boldsymbol{w_i}} = \frac{\boldsymbol{u_i'}(\boldsymbol{F^{-}})\boldsymbol{A}(\boldsymbol{F^{-}})'\boldsymbol{u_i}}{\boldsymbol{u_i'}(\boldsymbol{F^{-}})\boldsymbol{B}(\boldsymbol{F^{-}})'\boldsymbol{w_i}} = \frac{\boldsymbol{u_i}\boldsymbol{A}^{*}\boldsymbol{u_i}}{\boldsymbol{u_i'}\boldsymbol{u_i}} \tag{9}$$

which is of the form of an eigenvalue problem, which is minimized at the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{A}^{*}$. Additionally, the value of (9) lies between the minimum and maximum eigenvalue of $\boldsymbol{A}^{*}$.

The requirement that $\boldsymbol{W}$ be of full rank results in $\boldsymbol{W} = \boldsymbol{T_r}\boldsymbol{\Lambda_r}^{-1/2}\boldsymbol{U}$, where $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{A}^{*}$. □

More general proofs can be found in McDonald et al. (1979) and de Leeuw (1982).

Next we show that the resulting residuals are standardized, uncorrelated, and homoscedastic.

Proof. Carrying through the notation from above we see that

$$\begin{aligned}\mathrm{Var}(\boldsymbol{W}'\widehat{\boldsymbol{e}}) &= \mathrm{Var}(\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}'\widehat{\boldsymbol{e}})\\ &= (\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}')\mathrm{Var}(\widehat{\boldsymbol{e}})(\boldsymbol{T}_{\boldsymbol{r}}\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{U})\\ &= (\boldsymbol{U}'\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{T}_{\boldsymbol{r}}')\boldsymbol{B}(\boldsymbol{T}_{\boldsymbol{r}}\boldsymbol{\Lambda}_{\boldsymbol{r}}^{-1/2}\boldsymbol{U})\\ &= \boldsymbol{I}\end{aligned}$$

□

References

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. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383.

Butler, S. and Louis, T. (1992). Random effects models with non-parametric priors. *Statistics in Medicine*, 11(14-15):1981–2000.

de Leeuw, J. (1982). Generalized eigenvalue problems with positive semi-definite matrices. *Psychometrika*, 47(1):87–93.

Dempster, A. P. and Ryan, L. M. (1985). Weighted normal plots. *Journal of the American Statistical Association*, 80(392):845–850.

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

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models.* PhD thesis, University of California Los Angeles.

Jiang, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics*, 29(4):1137–1164.

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

McDonald, R. P., Torii, Y., and Nishisato, S. (1979). Some results on proper eigenvalues and eigenvectors with applications to scaling. *Psychometrika*, 44(2):211–227.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556.

Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17.