Visual inference for graphical diagnostics of linear mixed models

A thesis submitted for the degree of

MASTER OF APPLIED ECONOMICS AND ECONOMETRICS

by

Kaiwen Jin

Department of Econometrics and Business Statistics
Monash University
Australia

November 2020

Contents

Abstract

Graphical diagnostics focusing on the residual analysis are alternative methods to test the hypotheses based on the simulation approaches such as parametric bootstrap applied in the lineup protocol. We use three different data types, namely categorical, numerical, and mixture of categorical and numerical to accommodate the three different versions of Gaussian linear mixed models. Four diagnostic maps are used to detect the existence of abnormal observations from marginal residuals versus conditional residuals, and to evaluate the normality of conditional errors based on conditional residuals versus confounding conditional residuals. We create a survey to validate the methods. From the result, we find that the conditional residuals are better at checking the presence of outlying observations and exploring whether the conditional errors follow a normal distribution. Furthermore, by comparing those three diverse types of data, the mixture performs better at identifying the data plot among the null plots. R code is provided on GitHub (`https://github.com/kaiwenjanet/master`).

Chapter 1

Introduction

Statistical graphics play an important role in exploratory data analysis, model checking and diagnostic. For residual diagnostics, a plot may be used to infer the appropriateness of the fitted model such as by searching for unusual patterns in the plot. Graphical diagnostics enable the analyst to not only find unexpected results but may also uncover the reason for the result. However, inferences from plots are made often without calibration and in an informal manner, unlike the treatment of hypothesis tests where a numerical test statistic is formulated and compared against certain thresholds (e.g. p-value, confidence intervals or Bayes factors).

In a conventional (frequentist) hypothesis test, we formulate a null and alternate hypothesis, a test statistic and test this test statistic under the null distribution. The visual inference is another tool based on the application of lineup protocol which constitutes of usually, 19 null plots, and one data plot that randomly inserts among the null plots based on the experimental design setup that we created. These lineups are then presented to a number of (possibly independent) observers for evaluation. Figure 1.1 illustrates the conventional and visual inference. The lineup protocol follows closely to that of a frequentist hypothesis test except the data plot is treated as the test statistic.

Suppose there are K independent observers that participated, the p-value for the lineup of size m is calculated as follows (Majumder, Hofmann, and Cook, 2013). Consider the null hypothesis (H_0) that the data plot is not distinguishable from the null plots and the

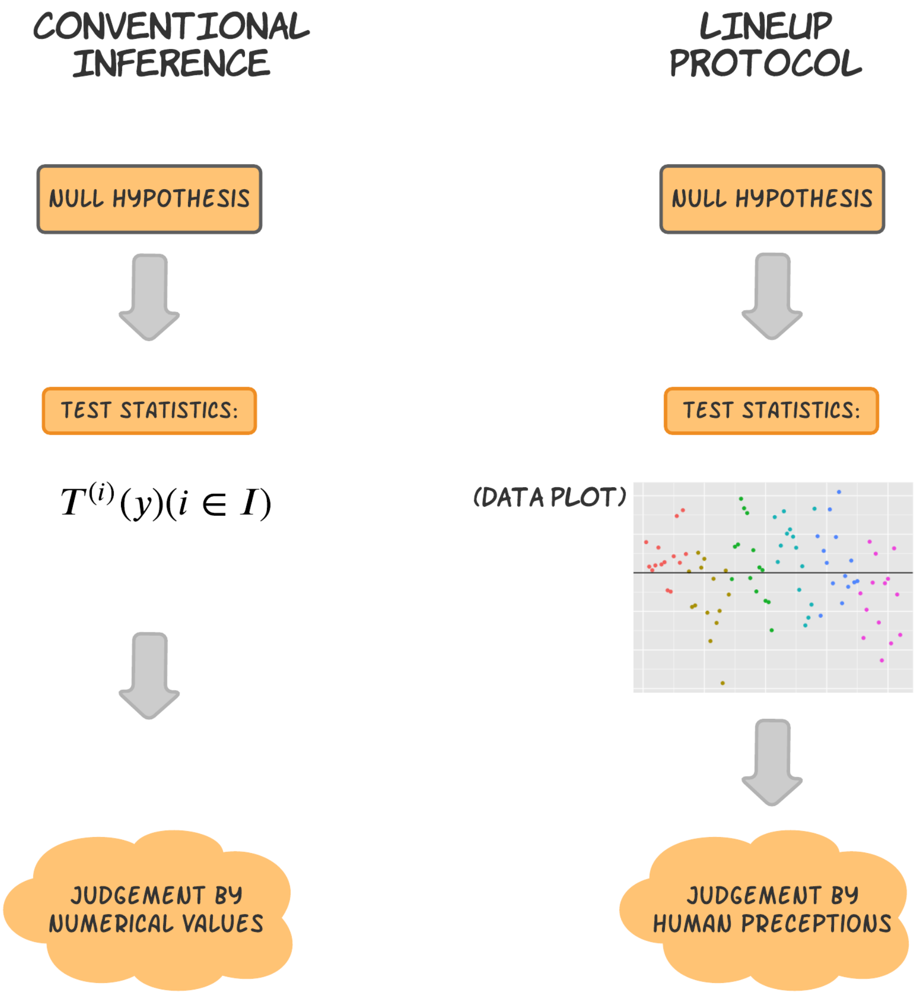


Figure 1.1: *Comparison of visual inference with conventional inference*

alternate hypothesis (H_1) that the data plot have some difference to the null plot. Assuming that observers have equal perceptive ability and that the positional placement of the plots have no effect then the probability that an observer selects the data plot in a lineup of size m is $1/m$ under H_0. If we let X be the number of observers out of K total observers who identified the data plot, then $X \sim B(K, 1/m)$ under H_0. Suppose x is the observed number of people who identified the data plot, then the visual inference p-value is calculated as $P(X \geq x)$. Majumder, Hofmann, and Cook (2013) also introduces the power of the visual test to compared the robustness of visual inference with the conventional inference.

Linear mixed model is an extension of the linear model, where the model contains additional random effects aside from the error term. The mathematical form of the model is shown in Chapter 2. The diagnostics of linear mixed models are further complicated than that of a linear model as there are considerations of multiple types of residuals. This thesis considers multiple forms of residuals that underlies the constructions of diagnostic plots. Namely, we consider marginal residual, conditional residual and least confounded residual (a linear combination of conditional residuals suggested to perform well) as numerical constructs for the residual plot, boxplot or quantile-quantile (Q-Q) plot.

The article is organized as follow. Chapter 2 introduce the linear mixed model with algebra. The brief review of the existing literature is illustrated in Chapter 3. Chapter 4 focus on the method we used to generate residual analysis based on three different data types as well as the survey that has been carried out. The results illustrated by the survey is presented in Chapter 5 and discussion and conclusion are in Chapter 6 and Chapter 7 respectively.

Chapter 2

Linear Mixed Model

Linear mixed model (LMM) are more versatile in fitting complex, correlated data structures than linear models by taking into account the dependency structures between units. Linear mixed models are special case of what is generally known as mixed-effects model; named as such because the model consists a mix of fixed and random effects. Confusingly, mixed-effects models are known with a variety of other names such as panel data model (often in econometrics), hirarchical model or multi-level model (often in social science). To elucidate the model we are referring, the mathematical form of linear mixed models is presented next.

For $i = 1, 2, \dots, n$ non-overlapping groups, a linear mixed model may be expressed as

$$\underset{(m_i\times 1)}{\mathbf{y}_i} = \underset{(m_i\times p)}{\mathbf{X}_i}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(m_i\times q)}{\mathbf{Z}_i}\underset{(q\times 1)}{\mathbf{b}_i} + \underset{(m_i\times 1)}{\mathbf{e}_i}$$

where $\mathbf{y}_i$ is a $m_i \times 1$ vector of response, $\mathbf{X}_i$ is a $m_i \times p$ fixed-effects design or regressor matrix, $\mathbf{Z}$ is a $m_i \times q$ known specification matrix corresponding to the random effects, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects describing the between-group covariance structure, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects governing the global mean structure, and $\mathbf{e}$ is an $m_i \times 1$ vector of random errors. The distributional assumptions are as follows: $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}_i)$; $\mathbf{e}_i \sim N(\mathbf{0}, \mathbf{R}_i)$; the random components $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent. The matrices $\mathbf{G}_i$ and $\mathbf{R}_i$ are symmetrical, positive definite matrices. We assume $\mathbf{R}_i = \sigma^2\mathbf{I}_{m_i}$ throughout this thesis.

We can rewrite the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

where $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top, \mathbf{X} = \oplus_{i=1}^n \mathbf{X}_i, Z = \oplus_{i=1}^n \mathbf{Z}_i, \mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$ and $\mathbf{e} = (\mathbf{e}_1^\top, \ldots, \mathbf{e}_n^\top)^\top$.

The key assumption of LMMs is that the residual errors and random effects are normally distributed, namely

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{e} \end{bmatrix} \sim \mathcal{N}_{M+N} \left(\begin{bmatrix} \mathbf{0_M} \\ \mathbf{0_N} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{0_{M\times N}} \\ \mathbf{0_{M\times N}} & \mathbf{R} \end{bmatrix} \right)$$

where $\mathbf{N} = \sum_{i=1}^n m_i$ and $\mathbf{M} = nq$.

Based on the model, the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})$$

with $\mathbb{E}(\mathbf{y}) = \boldsymbol{X\beta}$ and $\mathbb{V}(\mathbf{y}) = \boldsymbol{\Omega} = \boldsymbol{Z\Gamma Z}^\top + \mathbf{R}$, and the conditional distribution of $\mathbf{y}$ given $\mathbf{b}$ is given by

$$\mathbf{y}|\mathbf{b} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R})$$

Suppose we know the covariance matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{R}$, the estimated best linear unbiased estimators (BLUE) of the fixed effects $\boldsymbol{\beta}$ and best linear predictors (BLUP) of the random effects $\mathbf{b}$ can be obtained. Given the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{y}|\mathbf{b}$, the joint density function of $\mathbf{y}$ and $\mathbf{b}$ is defined as

$$\begin{aligned} f(\mathbf{y}, \mathbf{b}) &= g(\mathbf{y}|\mathbf{b})h(\mathbf{b}) \\ &= const \times \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^\top \mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) - \frac{1}{2}\mathbf{b}^\top \boldsymbol{\Gamma}^{-1}\mathbf{b} \right). \end{aligned}$$

The parameters $\mathbf{b}$ and $\boldsymbol{\beta}$ can be estimated by maximising this function. Taking the first derivative and equating the derivatives to zero, the result is shown below.

$$\begin{aligned}
\frac{\partial f(\mathbf{y},\boldsymbol{b})}{\partial \boldsymbol{\beta}} &\propto \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})^\top\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})+\mathbf{b}^\top\boldsymbol{\Gamma}^{-1}\mathbf{b}}{\partial\boldsymbol{\beta}} \\
&\propto \frac{-\mathbf{y}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}+\mathbf{b}^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}}{\partial\boldsymbol{\beta}} \\
&= -2\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y}+2\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+2\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}=0 \\
&\Rightarrow \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}=\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f(\mathbf{y},\boldsymbol{b})}{\partial \mathbf{b}} &\propto \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})^\top\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})+\mathbf{b}^\top\boldsymbol{\Gamma}^{-1}\mathbf{b}}{\partial\mathbf{b}} \\
&\propto \frac{-\mathbf{y}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}-\mathbf{b}^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}+\mathbf{b}^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+\mathbf{b}^\top\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}+\mathbf{b}^\top\boldsymbol{\Gamma}^{-1}\mathbf{b}}{\partial\mathbf{b}} \\
&= -2\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}+2\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+2\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}+2\boldsymbol{\Gamma}^{-1}\mathbf{b}=0 \\
&\Rightarrow \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}\mathbf{b}+\boldsymbol{\Gamma}^{-1}\mathbf{b}=\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}
\end{aligned}$$

Based on the two equations above, the mixed model equations (MME), which first proposed by Henderson (1973), is listed as:

$$\begin{pmatrix}\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}+\boldsymbol{\Gamma}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}\end{pmatrix}$$

And $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ refer to as “mixed model solutions”. However, the question is that how we can get the results. Next we will explain the procedure of solving the question.

Based on the method provided by Smith (1999), we write the coefficient matrix as:

$$\mathbf{C}=\begin{bmatrix}\mathbf{C}_{XX} & \mathbf{C}_{XZ} \\ \mathbf{C}_{ZX} & \mathbf{C}_{ZZ}\end{bmatrix}$$

Then, the model can be shown as,

$$\mathbf{C}_{XX}\boldsymbol{\beta} + \mathbf{C}_{XZ}\boldsymbol{b} = \mathbf{c}_{Xy}$$
$$\mathbf{C}_{ZX}\boldsymbol{\beta} + \mathbf{C}_{ZZ}\boldsymbol{b} = \mathbf{c}_{Zy}$$

where the $\mathbf{c}_{Xy} = \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y}$ and $\mathbf{c}_{Zy} = \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}$.

From the second equation, we can get

$$\mathbf{b} = \mathbf{C}_{ZZ}^{-1}\mathbf{c}_{Zy} - \mathbf{C}_{ZZ}^{-1}\mathbf{C}_{ZX}\boldsymbol{\beta}$$

By substituting **b** into the first equation,

$$(\mathbf{C}_{XX} - \mathbf{C}_{XZ}\mathbf{C}_{ZZ}^{-1}\mathbf{C}_{ZX})\boldsymbol{\beta} = \mathbf{c}_{Xy} - \mathbf{C}_{XZ}\mathbf{C}_{ZZ}^{-1}\mathbf{c}_{Zy}$$

where

$$\begin{aligned}
&\mathbf{C}_{XX} - \mathbf{C}_{XZ}\mathbf{C}_{ZZ}^{-1}\mathbf{C}_{ZX} \\
&= \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{X} - \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X} \\
&= \mathbf{X}^\top(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{Z}^\top\mathbf{R}^{-1})\mathbf{X} \\
&= \mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}
\end{aligned}$$

Then the best linear unbiased estimators (BLUE) of the fixed effects $\boldsymbol{\beta}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}^\top\mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y}) \\
&= (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Gamma}^{-1})^{-1}\mathbf{Z}^\top\mathbf{R}^{-1})\mathbf{y} \\
&= (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y}
\end{aligned}$$

Substituting the result for $\hat{\boldsymbol{\beta}}$ into $\mathbf{b} = \mathbf{C}_{ZZ}^{-1}\mathbf{c}_{Zy} - \mathbf{C}_{ZZ}^{-1}\mathbf{C}_{ZX}\boldsymbol{\beta}$, the best linear unbiased predictors (BLUP) of the random effects **b** is shown as,

$$\hat{\mathbf{b}} = \mathbf{C}_{ZZ}^{-1}\mathbf{c}_{Zy} - \mathbf{C}_{ZZ}^{-1}\mathbf{C}_{ZX}(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y}$$

$$
\begin{aligned}
&= \mathbf{C}_{ZZ}^{-1}(\mathbf{c}_{Zy} - \mathbf{C}_{ZX}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{y})\\
&= (\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}+\mathbf{\Gamma}^{-1})^{-1}(\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{y})\\
&= (\mathbf{Z}^\top\mathbf{R}^{-1}\mathbf{Z}+\mathbf{\Gamma}^{-1})^{-1}\mathbf{Z}^\top\mathbf{R}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1})\mathbf{y}\\
&= \mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1})\mathbf{y}\\
&= \mathbf{\Gamma}\mathbf{Z}^\top\mathbf{Q}\mathbf{y}
\end{aligned}
$$

where

$$\mathbf{Q} = \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}$$

This implies that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \mathbb{V}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\Omega^{-1}\mathbf{X})^{-1}$$

$$\mathbb{E}(\hat{\mathbf{b}}) = \mathbf{0}, \mathbb{V}(\hat{\mathbf{b}}) = \mathbf{\Gamma}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\mathbf{\Gamma}$$

Besides, the estimates of residuals are given by $\hat{\mathbf{e}} = \mathbf{R}\mathbf{Q}\mathbf{y}$, the proof is shown as,

$$
\begin{aligned}
\hat{\mathbf{e}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}\\
&= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\\
&= (\mathbf{I} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1})\mathbf{y} - (\mathbf{I} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1})\mathbf{X}\hat{\boldsymbol{\beta}}\\
&= (\mathbf{I} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1})\mathbf{y} - (\mathbf{I} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1})\mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{y}\\
&= (\mathbf{I} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top\mathbf{\Omega}^{-1})(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1})\mathbf{y}\\
&= (\mathbf{\Omega} - \mathbf{Z}\mathbf{\Gamma}\mathbf{Z}^\top)(\mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{\Omega}^{-1})\mathbf{y}\\
&= \mathbf{R}\mathbf{Q}\mathbf{y}
\end{aligned}
$$

Robinson et al. (1991) gives an excellent introduction to the estimation of random effects and discusses its various fields of application.

After introducing the linear mixed model, there are several literatures explain the residuals applied in the model. Similar to simple linear model, residuals are significant to diagnostics of linear mixed model. There are three various types of residuals based on Singer, Rocha, and Nobre (2017), that is, marginal residuals $\hat{\boldsymbol{\xi}}$, conditional residuals $\hat{\mathbf{e}}$ and random effect residuals $\mathbf{Z}\hat{\mathbf{b}}$. Besides, given that $\hat{\mathbf{e}} = \mathbf{RQe} + \mathbf{RQZb}$, the conditional and random effects residuals may be confounded. It suggests that $\hat{\mathbf{e}}$ may not be adequate to check the normality of $\mathbf{e}$ because when $\mathbf{b}$ is grossly non-Gaussian, $\hat{\mathbf{e}}$ may not presented a Gaussian behaviour even when $\mathbf{e}$ is Gaussian. The further application of these residuals will be listed in Chapter 3 and Chapter 4.

Chapter 3

Visual inference

3.1 Residual diagnostics

Residual diagnostics are common approach to infer the appropriateness of statistical models, however this is complicated in linear mixed models (LMMs) due to different types of residuals. There are three basic types of residuals for a linear mixed models, that is, marginal residuals $\hat{\xi}$, conditional residuals $\hat{\mathbf{e}}$ and random effect residuals $\mathbf{Z}\hat{\mathbf{b}}$ (Haslett and Haslett, 2007). More specifically, a marginal residual is the difference between the observed data and the estimated marginal mean,

$$\hat{\xi} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

A conditional residual is the difference between observed data and predicted value of the observation,

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}$$

A random effects residual is the difference between the predicted responses and the population average, $\mathbf{Z}\hat{\mathbf{b}}$. Singer, Rocha, and Nobre (2017) summarised the areas of application for each type of residual defined by LMMs such as detecting the linearity of effects, assessing the normality and homoscedasticity of errors, checking outliers and accessing the covariance structure for individual subjects. There are various ways of judging deviations from normality for residual errors and for random effects, although it should be noted that none confirm normality and in many cases, normality is assumed to be granted (due to either the limitation of sample size or neglect for diagnosis). A common approach to infer normality is to use quantile-quantile (QQ) plot (Pinheiro and Bates (2006)). Marginal residuals are useful in checking the linearity of fixed effects, outlying observations and the covariance structure of $\mathbf{\Omega}_i$, while the conditional residuals can be used to detect the outlying observations, homoscedasticity and normality of the conditional errors. Moreover, the random effects can be applied to examine the outlying subjects and the normality of the random effects (Singer, Rocha, and Nobre (2017)).

According to Hilden-Minton (1995), a residual is confounded when it is dependent on the other types of errors. Since the conditional and random effects residual are not independent of one another, then the conditional residual may not be adequate to check the normality of conditional errors. They try to minimise the fraction of confounding for the k-th conditional residual, namely,

$$0 \leq \frac{\mathbf{u}_k^\top \mathbf{RQZ\Gamma Z}^\top \mathbf{QRu}_k}{\mathbf{u}_k^\top \mathbf{RQRu}_k} = 1 - \frac{\mathbf{u}_k^\top \mathbf{RQRQRu}_k}{\mathbf{u}_k^\top \mathbf{RQRu}_k} \leq 1$$

where $\mathbf{u}_k$ is the k-th column of $\mathbf{I}_N$. Furthermore, Hilden-Minton (1995) proposed a linear transformation of the conditional residuals to get the $(n-q)$ least confounded conditional residuals $(\mathbf{c}_k^\top \hat{\mathbf{e}})$. The least confounded residuals are given by,

$$\mathbf{c}_k^\top \hat{\mathbf{e}} = \lambda_k^{-1/2} \boldsymbol{\ell}_k^\top \mathbf{R}^{-1/2} \hat{\mathbf{e}} = \lambda_k^{1/2} \boldsymbol{\ell}_k^\top \mathbf{R}^{-1/2} \mathbf{y}, k = 1, 2, \ldots, N-p$$

where $1 \geq \lambda_1 \geq \ldots \geq \lambda_{N-p} > 0$ are the ordered values to $\mathbf{\Lambda}$, obtained from the spectral decomposition

$$\mathbf{R}^{1/2}\mathbf{QR}^{1/2} = \mathbf{L\Lambda L}^\top, \mathbf{L}^\top \mathbf{L} = \mathbf{I}_{N-p}$$

and $\boldsymbol{\ell}_k$ represents the k-th column of $\mathbf{L}$. However, the linearly transformed residuals do not correspond to individual observations anymore. Besides, they may amplify the supernormality effect, which tends to look more normal than the underlying effects actually are (Schützenmeister and Piepho, 2012). Schützenmeister and Piepho (2012) prefer to use the studentized conditional residuals.

In Singer, Rocha, and Nobre (2017)'s paper, they not only consider the residual analysis, they also explore the global influence analysis, such that, leverage analysis, case deletion analysis and local influence analysis based on details in Beckman, Nachtsheim, and Cook (1987), Banerjee and Frees (1997), Christensen, Pearson, and Johnson (1992), Lesaffre and Verbeke (1998), among others.

In Loy, Hofmann, and Cook (2017)'s literature, rather than three different types of residuals, they used two basic types of residuals of LMM that they termed level-1 (observation level) residuals and level-2 (group level) residuals. These are equivalent to the conditional residual and predicted random effects, respectively. Although both Singer, Rocha, and Nobre (2017) and Loy, Hofmann, and Cook (2017) use the conditional residuals to check the homogeneity of residual variance, Singer, Rocha, and Nobre (2017) used the standardised conditional residual versus the fitted value to detect the homoscedasticity of the conditional errors, whereas Loy, Hofmann, and Cook (2017) checked the relationship between conditional residual and one of the model's covariate. They also check the homogeneity of conditional residual variance between groups by comparing the conditional residual and grouping variable. Loy, Hofmann, and Cook (2017) tested the linearity of conditional residual by plotting the explanatory variable versus the conditional errors. However, the linearity diagnostic only occurred in the fixed effect as Singer, Rocha, and Nobre (2017) mentioned by plotting the standardised marginal residuals against the fitted value. Moreover, the distributional assessment of the random effect for Loy, Hofmann, and Cook (2017) is based on the Q-Q plot where the random effects follow a t_3 distribution. Whereas, plotting the χ_q^2 QQ-plot for the Mahalannobis's distance between $\hat{\mathbf{b}_i}$ and $\mathbb{E}(\mathbf{b_i}) = 0$ is the way to check the normality of the random effects ($\mathbf{b}_i$) denoted by Singer, Rocha, and Nobre (2017). Rather than residual analysis, Loy, Hofmann, and Cook (2017) was motivated by model selection. They also used the residuals to detect the significance of a fixed effect by plotting a residual quantity from the model without the variable of interest with the values of that variable. They also detected if the model needs the random effect, if the random effects include both random intercept and random slope, and whether the random effects need to be correlated.

3.2 Conventional hypothesis vs Visual inference

People always use classical statistic inference such as p-value or the t-statistic. But these methods only tell that there is a problem with the model. However, graphical diagnostic can not only tell the problem but also can tell the cause of the problem. Buja et al. (2009a) outline the parallelism between quantitative testing and visual discovery and the steps for the visual inference. Both methods are started with the same set of hypotheses. As we all know, the conventional statistical inference makes up of 1) formulating the null and alternative hypotheses, 2) calculating the test statistic from the observed data, 3) comparing the test statistic based on a null distribution, and 4) making a decision for any rejections. While, for visual inference, the test statistic is the plot of the data. And the plot of the data is compared with a set of plots of samples drawn from the null distribution. Accordingly, the null distribution of plots refers to the infinite collection of plots of null datasets sampled under the null hypothesis(Buja et al. (2009a), Majumder, Hofmann, and Cook (2013), Loy, Hofmann, and Cook (2017)). If the plot of data is distinguishable, that is selected as the most different, then the null hypothesis will be rejected. Meanwhile the visual discovery is based on the viewer's cognition when they see the plots.

Buja et al. (2009a) states the crucial difference between conventional hypothesis test and visual inference: quantitative testing needs the explicit prior specification of the features; by contrast, the range of visual discoveries under the exploration data analysis and model diagnostic is not pre-specified explicitly. However, this will result in rejection that will be known. Furthermore, the important divergence between conventional and visual testing is that lineups will almost rely on the viewers' evaluation (Majumder, Hofmann, and Cook (2013)).

Since a graphical representation of the data is chosen to display the strength of the parameter of interest θ in visual inference, a definition of *visual statistic* has been carried out by Buja et al. (2009a). In details, the plots are drawn from the data generated consistently with the null hypothesis are called null plots which denoted as $T(y_0)$ and data plot, $T(y)$, maps the actual data to the plot. Meanwhile, there is a tool called power of a visual test is developed by Majumder, Hofmann, and Cook (2013), which is useful in comparing the performance of visual inference and conventional test.

3.3 Protocols

Buja et al. (2009a) introduces two protocols for the inferential use of null plots based on null datasets drawn from null hypothesis, that is 'the Rorschach' and lineup. The Rorschach is taken as the cognitive experimentation. To measure a data analyst's tendency to over interpret plots in which there is no or only obvious structure is the goal. Although it can be biased based on the analysts' knowledge, the aim of this training is to improve the awareness of the features when they detect. For the lineup, they asked the viewer to identify the most different plot among 20 plots where the data plot is randomly allocated among the 19 null plots (Buja et al. (2009a), Majumder, Hofmann, and Cook (2013), Loy, Hofmann, and Cook (2017)). Therefore, there is a one in 20 chance that the data plot will be pointed out.

The steps are constructed by Loy, Hofmann, and Cook (2017):

1. *Create lineup data*: Assuming that the proposed models are correct, we use the parametric bootstrap to simulate new responses, refit the model to these simulated responses, and extract the residuals of interest from the proposed model. For each lineup, this process is used to obtain $m - 1 = 19$ simulated null datasets.
2. *Render lineups*: Draw small multiples of each of the null datasets and randomly insert the observed data among the nulls. Each plot is labelled by a number from 1 to m. These IDs are used for identification and later evaluation of results.
3. *Evaluate lineups*: Present the lineups to independent observers, instructing them to identify the plot most different from the set and asking them what feature led to their choice. Theses choices came in the form of four suggestions (in checkboxes) and one text box for a free-form answer.
4. *Evaluate the strength of evidence*: For a lineup of size $m = 20$ that has been evaluated by K independent observers, the number of evaluations of a lineup in which the

observer identifies the data plot, Y, has a Visual distribution $V_{K,m,s=3}$ as defined by **Hofmann (2015)**

If the observers cannot find any distinguishable features, they may select one based on a random guess. Some observers may select the plot based on identifying features different to the rest (Buja et al. (2009a)). Buja et al. (2009a) also carries out the characteristic of lineup that not only choosing single plot among the lineups but also selecting the rank of the difference level of the plots. If this protocol is repeated to multiple independent observers, the p-value will be realized by tabulating the number of independent investigators chosen the data plot among the 19 null plots. Afterwards, the definition of p-values of lineup is implemented by Buja et al. (2009a) and Majumder, Hofmann, and Cook (2013). For K independent investigators, let X be the number of observers picking the test statistic from the lineup. Under the null hypothesis, $X \sim B(K, 1/m)$ (for a lineup of size m). Hence, Majumder, Hofmann, and Cook (2013) introduce the visual p-value of a lineup of a size m evaluated by K observers is given by

$$P(X \geq x) = 1 - \text{Binom}_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^{i} \left(\frac{m-1}{m}\right)^{K-i}$$

where x is the number of the observers pointing out the actual data plot. Then with the significance level α, the author indicates decision rule that we can reject the null if out of K observers at least χ_α correctly identify the data plot, otherwise, fail to reject the null. While the lineup size will affect the probability that the viewer correctly points the data plot out proved by Majumder, Hofmann, and Cook (2013), such that with larger m, the probability of correctly distinguishing the data plot decreases. However, we often choose the lineup size as 20.

Loy, Hofmann, and Cook (2017), Majumder, Hofmann, and Cook (2013), Buja et al. (2009a) and etc. have used Turk (2012) to run the visual inference experiments. In these experiments, the demographic information, such as age, gender, education level, will be asked. Furthermore, the time taken to complete questions for each observer is recorded and observers are asked to give the confidence score from 1 to 5 with weak to high

(Majumder, Hofmann, and Cook (2013), Loy, Hofmann, and Cook (2017)). In order to test the ability of individuals, Majumder, Hofmann, and Cook (2013) gives the easy lineup as the reference lineup. If they are able to identify the actual data plot, then the response for the reference lineup will be removed and all the following responses will be stored. If the answer for the reference lineup is wrong, then the all responses made by this observer will be removed. Based on the examples that they give and also refer to the survey design (Dawes (2000)), the wording of the instruction might help the viewers to make wiser decisions.

Chapter 4

Methodology

In this section, we describe (i) the three data sets used as a basis for the simulation settings; (ii) the construction of the lineups and (iii) the experimental set up for the survey.

4.1 Data sets

The three data sets that we used in this thesis are: reaction times in a sleep deprivation study data set in the `R` package `lme4` (Bates et al., 2015); autism study data set in `R` package `HLMdiag` (Loy and Hofmann, 2014); and linguistic data set (Winter, 2013). Each of these data sets have different characteristics. More specifically, the linguistic data set contains only categorical dependent variables; the variables in the sleep study data are all numerical variables except the subject level; and finally, the autism data contains a combination of numerical and categorical variables. We describe briefly about each data in the following sections with more detailed variable summaries in the Appendix chapter.

4.1.1 Reaction times in a sleep deprivation study

Sleep deprivation and chronic sleep restriction endanger health, safety, productivity and quality of the life. To better understand the importance of sleep time, Belenky et al. (2003) conducted a test based on 66 observers with 3 hours, 5 hours, 7 hours and 9 hours daily time in bed (TIB). The sleep deprivation data set concentrates on 18 volunteers who spend 3 hours in bed per night. On day 0, the subjects have the normal amount of sleep. But for the rest of nine nights, 3 hours of sleep time is restricted to them. The output of 180 recorded responses in Reaction variable shows the average reaction time per day for each subject. As a result, those volunteers with 3-hour TIB, speed (mean and fastest 10% of responses) on the psychomotor vigilance task (PVT) declined (Belenky et al., 2003).

4.1.2 Autism study

A prospective longitudinal study following 155 children between the ages of 2 and 13 who were diagnosed with either autism spectrum disorder or non-spectrum developmental delays at age 2 has been carried out by Anderson et al. (2009) to explore the changes in verbal and social abilities from childhood to adolescence. Assessments were made on the children at ages 2, 3, 5, 9, and 13, however, not all children were assessed at each age. Their age, gender (female or male) and race (white or non-white) were captured. Vineland Socialization Age Equivalent (VSAE) recorded the overall measurement of the child's social skill. Sequenced Inventory of Communication Development (SICD) assessed the expressive language development at aged 2, in which three groups were divided into low, median and high. In addition, the initial diagnostics at age 2 of each child has been recorded as autism or pervasive development disorder (pdd).

4.1.3 Linguistic study

Winter (2013) introduced a linguistic study case on the voice pitch by identifying the gender, subjects, scenarios and attitudes. One of the scenarios was to ask different people for help. For example, one subject was asking the professor to help with polite attitude, or asking a peer for a favour on an informal condition. The data contains 84 observations on the voice pitch (or frequency) from 6 subjects (3 females and 3 males) under 7 scenarios with 2 attitudes (informal or polite).

4.2 Exploratory data analysis

Exploratory data analysis (EDA) is a critical tool before fitting the model as this method can help us to identity the features of the raw data. In the sleep deprivation study data, we use the plot comparing the response and explanatory variable of interest using linear smoothers for each group. Figure 4.1 illustrates the random intercept and random slope in reaction time versus days from day 0 to day 9 corresponding to different subjects. Hence, we treat the explanatory variable as the designed matrix of fixed effect and the subjects as the designed matrix of random effect. To test the dependency of random effects, based on the method provided by Loy, Hofmann, and Cook (2017), a scatter plot of predicted random effects with overlaid regression lines. The null plots are simulated from the model with the fact that random effects are not correlated and the data plot is made using the predicted random effects from original model fit to the observed data. While the regression lines show the degree of correlation between random effects. According to the result, although the correlation between random effects is small, it is hard to distinguish the data plot from null plots. Therefore, there is no need for the correlated random effect, that is, the random slope and random intercept are independent.

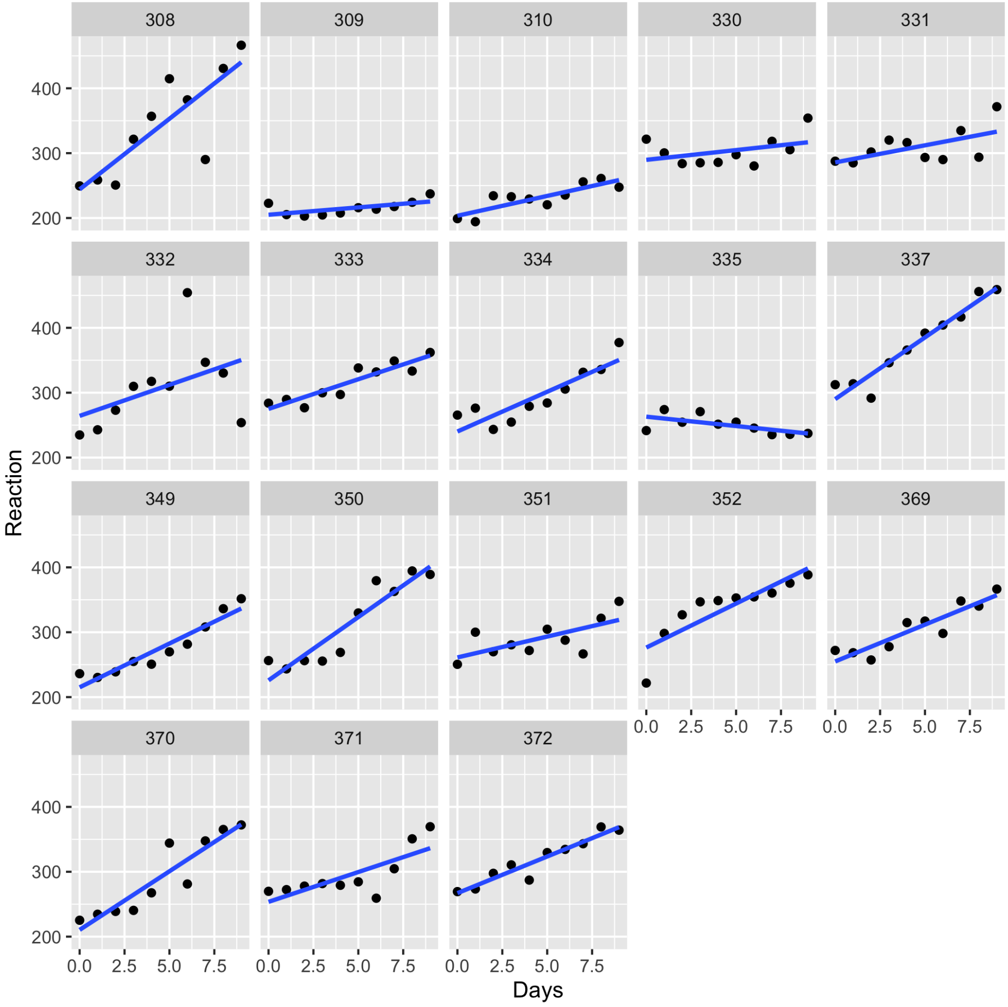


Figure 4.1: *Exploratory data analysis example in the sleep study case. It shows the random intercept and random slope with different subjects.*

4.3 Residuals Diagnostics

Residuals are used to examine model assumptions and to detect outliers and potentially influential data points. As Singer, Rocha, and Nobre (2017) listed eight uses of residuals for diagnostic purposes, we focus on three residuals with two diagnostic purposes in this thesis: (1) checking presence of outlying observations based on marginal residuals as well as conditional residuals, and (2) detecting normality of conditional error according to conditional residuals and confounded residuals. After fitting the data sets into the "best" model, we can estimate the fixed effect $\hat{\beta}$ and predict the random effect $\hat{\mathbf{b}}$, with latter generated as a realisation from a normal distribution with zero mean and variance-covariance matrix estimated from the data using either maximum likelihood or residual maximum likelihood (Patterson and Thompson, 1971). Then we apply the method that Singer, Rocha, and Nobre (2017) stated to obtain these residuals and then acquire the plots.

Given the true variance of marginal residuals that $\mathbb{V}(\hat{\boldsymbol{\xi}}_i) = \boldsymbol{\Omega}_i - \mathbf{X}_i(\mathbf{X}_i^\top \boldsymbol{\Omega}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i^\top$, the standardised marginal residuals can be generated as $\hat{\xi}^*_{ij} = \hat{\xi}_{ij}/[diag_j(\hat{\mathbb{V}}(\hat{\boldsymbol{\xi}}_i))]^{1/2}$, where $diag_j(\hat{\mathbb{V}}(\hat{\boldsymbol{\xi}}_i))$ is the j-th element of the main diagonal of $\mathbb{V}(\hat{\boldsymbol{\xi}}_i)$.

With true variance of conditional residuals, $\mathbb{V}(\hat{\mathbf{e}}) = \mathbf{R}[\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}]\mathbf{R} = \mathbf{RQR}$, Singer, Rocha, and Nobre (2017) suggested using the standardised conditional residuals, $\hat{\mathbf{e}}^*_{ij} = \hat{\mathbf{e}}_{ij}/diag_{ij}(\hat{\mathbf{R}}\hat{\mathbf{Q}}\hat{\mathbf{R}})$, where $diag_{ij}(\hat{\mathbf{R}}\hat{\mathbf{Q}}\hat{\mathbf{R}})$ representing the main diagonal element of $\mathbf{RQR}$ corresponding to the j-th observation of the i-th unit.

To detect outlying observations, we draw the element of the standardised marginal or conditional residuals versus the observation indices recommended by Singer, Rocha, and Nobre (2017). However, boxplot is applied for $\hat{\xi}^*_{ij}$ or $\hat{\mathbf{e}}^*_{ij}$ versus the explanatory variables for the autism data set since the data size is big.

Hilden-Minton (1995) introduced a linear transformation of the conditional residuals which is called least confounded conditional residuals $c_k^\top\hat{\mathbf{e}}$ in order to minimise the fraction of confounding. They thought the ability to check for normality of the conditional errors increases. And we used the standardised least confounded conditional residuals $c_k^\top\hat{\mathbf{e}}^*$ by dividing $c_k^\top\hat{\mathbf{e}}$ by the squared root of the corresponding element in $\mathbf{C}\hat{\mathbf{R}}\hat{\mathbf{Q}}\hat{\mathbf{R}}\mathbf{C}^\top$. Hence, we employ QQ plot of the standardised conditional residuals and standardised least confounded conditional residuals to check for normality. Whereas Pinheiro and Bates (2006) considers QQ plot of $\hat{\mathbf{e}}/\hat{\sigma}$ for checking the normality of the conditional error, we stick with the method of generating the standardised conditional residuals with Singer, Rocha, and Nobre (2017).

4.4 Experiment setup

On the basis of the data types that we have, three different versions are generated for each data type with four replicates:

1. *Generated from the "best" model*:

The estimated random effect and error term are following a normal distribution with mean $\mathbf{0}$ and variance $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{R}}$ respectively from the "best" model as well as the estimated fixed effect. The sample can be generated through $\mathbf{y}^* \sim (\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{\Omega}})$ where $\hat{\mathbf{\Omega}} = \mathbf{Z}\hat{\mathbf{\Gamma}}\mathbf{Z}' + \hat{\mathbf{R}}$. This is a scenario where the fitted model and the data generating process match and thus, we do not expect any graphical diagnostics to indicate the model is inappropriate.

2. *Added slight noises*:

i) For sleep study case, within-unit correlated error terms have been introduced. For each subject, the error terms are correlated by value 25 which is round the standard deviation of the error term.
ii) For linguistic case, we treat the error term under the t-distribution with $\nu = 1$ degree of freedom.
iii) For autism case, we also treat the error term under the t-distribution with $\nu = 15$ degree of freedom at a random choice.

In this scenario, the fitted model has a mild mis-specification there is some small chance that some diagnostics may detect this.

3. *Introduce extreme noises*:

We randomly added some extreme values to response variable for particular subjects in each data set.

i) For sleep study case, we randomly add 20% of mean value to 2 response values of each subject.
ii) For linguistic case, we firstly create four tables with respect to the category of gender and attitude, that is female with polite attitude, female with informal attitude, male with polite attitude, and male with informal attitude. Then we alter 4 reaction time value in each table by adding 20% of mean value for the overall reaction time.

iii) For autism case, we add 20% of the total median response values to roughly 20% of observations at random.

Therefore, 12 data sets are generated with corresponding 12 data plots constituting the residual plots for detecting the presence of outlying and checking the normality of conditional error.

In this scenario, the added extreme noises should not be modelled well by the fitted model so diagnostic assessments should indicate the inappropriateness of the model fit.

4.5 Generating lineups

We use a parametric bootstrap approach to generate the null data for each lineup based on the simulated data with zero to some noise added as described in the previous section. More specifically:

1. We generate the vector of random effects from $\mathcal{N}(\mathbf{0}, \hat{\mathbf{G}})$ for each group, that is, generate $\mathbf{b}_i^* \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{G}})$ for $i = 1, 2, ..., n$ where $\hat{\mathbf{G}}$ is estimated from the fit of the "best" model to the simulated data.
2. Vector of conditional residuals is generated from $\mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$ for each group, that is, generate $\mathbf{e}_i^* \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$ for $i = 1, 2, ..., n$
3. Generate a bootstrap sample $\mathbf{y}_i^*$ from $\mathbf{y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{b}_i^* + \mathbf{e}_i^*$ for each group $i = 1, 2, ..., n$
4. Refit the model to the bootstrap sample.
5. Repeat steps from 1 to 4 by 19 times.

From the null data, the null plots are produced which is consistent with the null hypothesis. Afterwards, we insert the data plot among 19 null plots to form the lineup protocol using the `nullabor` package (Buja et al., 2009b). We design the position of each data plot. According to Figure 4.2, based on the "best" model, the data plots which point the presence of outlying observations of both marginal or conditional residuals are at the position 7 in the lineup for the first two replicates. Then, for replicates 3 and 4, the position of data plot is 13. The data plot which refers to check the normality of conditional error for either standardised conditional residual or least confounded conditional residual is in

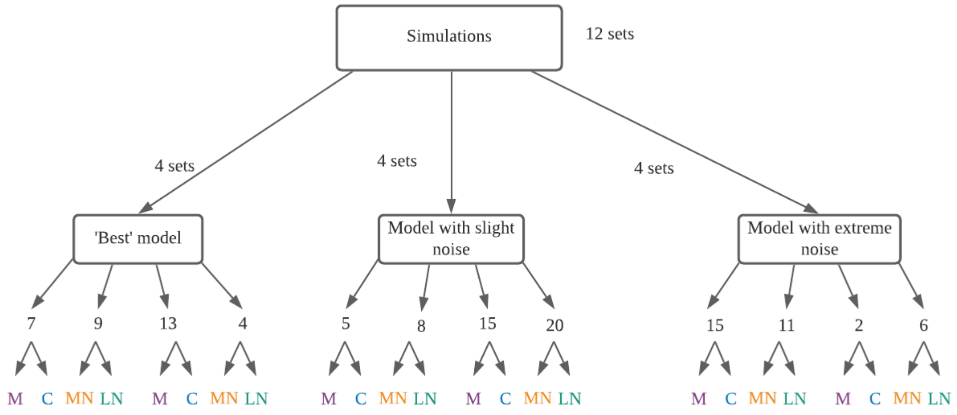
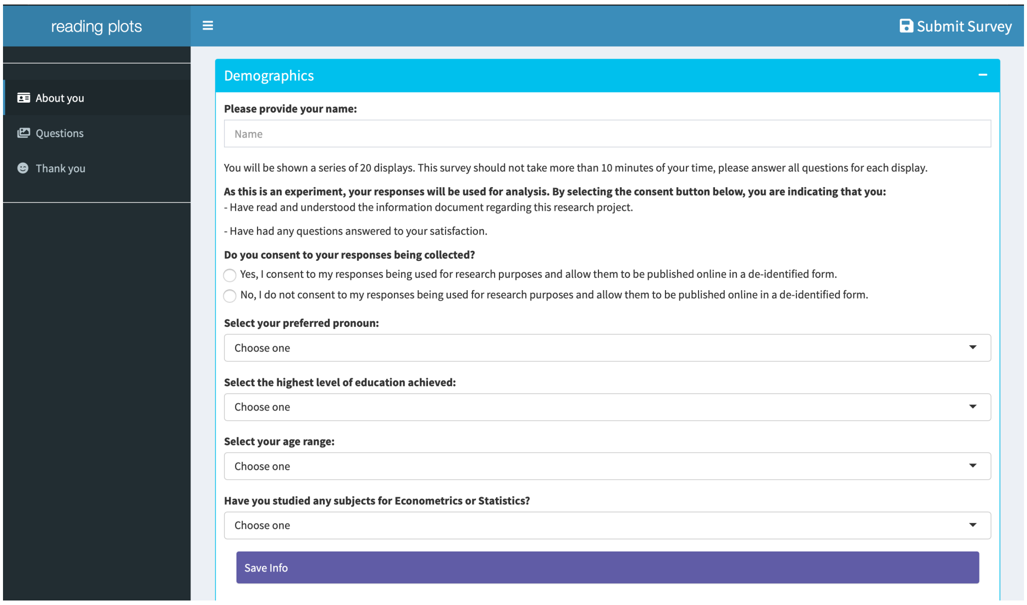


Figure 4.2: *Experimental design diagram*

panel 9 of 20 plots as the basis of the first two replicates of first version. The position for next two replicates are in 4. The location of the data plot for the remainder two versions are followed the Figure 4.2.

4.6 Survey through Shiny app

We download all the lineups to png files with first three letter for the data types (`aut`, `lin` and `slp`), followed by versions (`1`, `2`, and `3`) with replicates (`1`, `2`, `3`, and `4`). The diagnostic purposes are listed as the last letter where 1 for residual plots in marginal residuals, 2 represents the residual plots in conditional residuals, 3 stands for QQ plot under conditional residuals and 4 signals the QQ plot in confounded conditional residuals. Using the Latin Squared Design method (de Mendiburu, 2020), we list the replicates 1 to 4 as the rows and four types of diagnostic plots are the column. A, B, C and D as the variates are listed in each cell without repeating in each row and column. Once the cell in the first row has been target, the corresponding row and column will be removed. And the matching diagnostic plot with version 1 will be presented to the observers for particular data type. This process will be iterated 4 times for each data type until we achieve the 12 lineups among 144 lineups for each individual. These lineups are from 3 different data types with four diverse replicates for 2 different plot types but without restricting the versions.



Please provide your name:

Name

You will be shown a series of 20 displays. This survey should not take more than 10 minutes of your time, please answer all questions for each display.

As this is an experiment, your responses will be used for analysis. By selecting the consent button below, you are indicating that you:
- Have read and understood the information document regarding this research project.

- Have had any questions answered to your satisfaction.

Do you consent to your responses being collected?

○ Yes, I consent to my responses being used for research purposes and allow them to be published online in a de-identified form.

○ No, I do not consent to my responses being used for research purposes and allow them to be published online in a de-identified form.

Select your preferred pronoun:

Choose one

Select the highest level of education achieved:

Choose one

Select your age range:

Choose one

Have you studied any subjects for Econometrics or Statistics?

Choose one

Figure 4.3: *First tab in the survey created by shiny app. It contains the demographic information about the individuals.*

We use the `taipan` package (Kobakian and O'Hara-Wild, 2018) to build our survey questions which include the demographic of audience and their responses. In the first tab, referred to Fig. 4.3, volunteer's name, gender, age, education level and whether they have studied econometrics or statistics are recorded. Next tab in Fig. 4.4, they are asked to choose the most different map from their decision and how certainty they are to choose that map. There is a process hit for them as the number of lineups they have been seen. They will iterate the process by 12 times by answering the same question but looking at different lineups. If their answers are less than 12 plots before submitting, there will be a hit popped up to reminder them to complete the survey before they leave. After they click the submit button, the final tab is shown up as the thank you page. All the responses made by each individual are securely stored into the google sheet based on the private link through `googlesheets4` package (Bryan, 2020) with authenticating. Besides, we also acquire the unique identifier for them in case they double submitted. It is also useful when participant does not leave their names. The plot name with data type, version, replicates and plot type is also recorded which will be used in the further analysis. The time period that the participant takes are displayed as the reference.

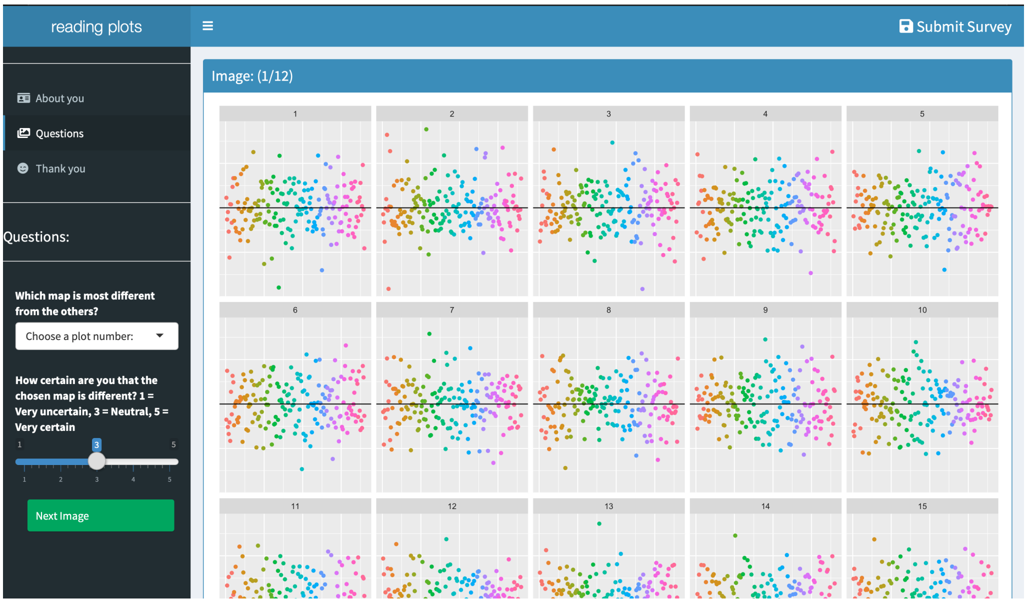


Image: (1/12)

Questions:

Which map is most different from the others?

How certain are you that the chosen map is different? 1 = Very uncertain, 3 = Neutral, 5 = Very certain

1 — 3 — 5

Figure 4.4: *Second tab in the survey created by shiny app. It includes the lineup and questions that we asked.*

Chapter 5

Results

There are 54 individuals that attempt the survey. While, we removed two tests, people who are not sincerely doing the survey such that they responded all the lineups with the same choice or they just answered several questions. Therefore, we accepted the responses from 50 valid independent individuals with 600 observations.

Among 50 independent observers, 72% of them are female and the rest are male. All most half of the participants are aged at 18 to 24 years old and second big proportion is followed by the age range from 25 to 34. There are 18 viewers who achieved Bachelor degree while 20 of them had the Master degree and 8 got the Doctorate degree. Besides, 90% of them have studied econometrics or statistics before.

Based on the responses we got, we calculated the correctness from all these 50 observers. Since we randomly select the versions of lineups based on our design, the total number of responses to each version is different. Referred to Fig. 5.1, among 600 observations, 37% of the responses correspond to version 3, and version 3 obtains the most correct responses. 194 records are presented in version 2. The rest are listed in version 1 which has the least responses. There are fewer observers who can recognize the actual data plot in the version 1, which is in line with our expectations, because version 1 is simulated from the 'best' model, which may not have any obvious features.

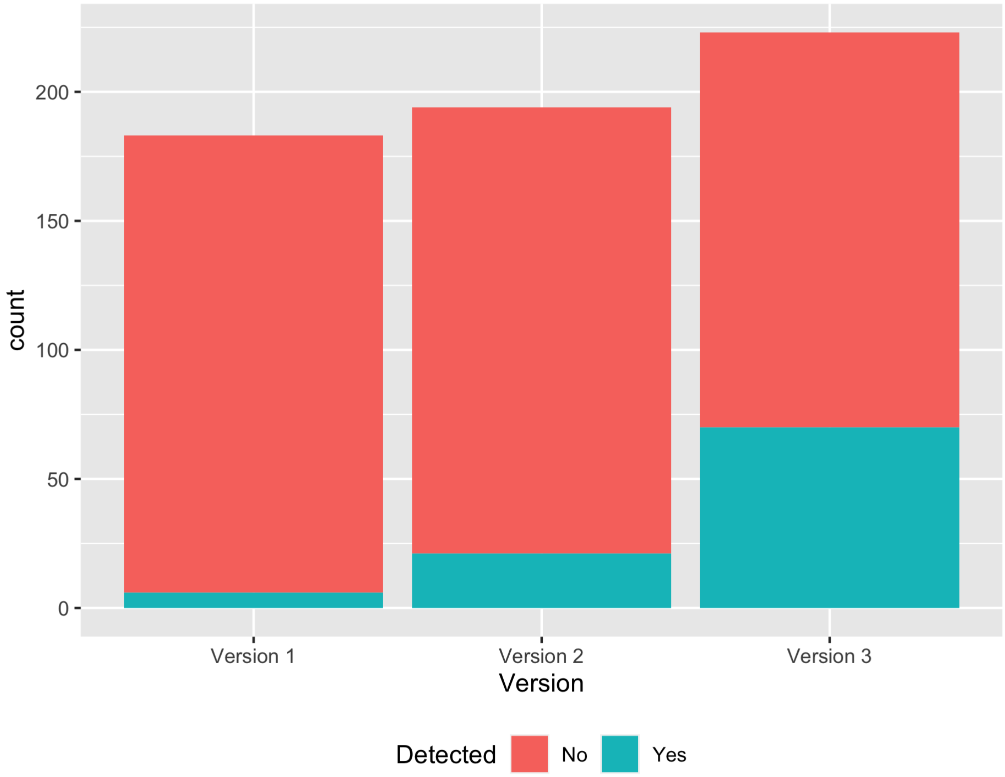


Figure 5.1: *Total amount of responses in each version*

Table 5.1 presents the summary of all lineups in evaluation. Three different data types (categorical, numerical and mixed) are combined in the result. We included four different purposes of lineups, where first two are detecting the outlying observation and the last two are checking the normality of the conditional errors and three different versions respectively. The ratio shown in the table is the division of number of observers who identify the actual data plot among the total number of participants in the same version and lineup. The visual p-value is calculated based on the method provided by Majumder, Hofmann, and Cook (2013). From the results, the increase in the ratio for the version 2 compared with version 1 may indicate that for version 2, the data plot has certain features compared with the null plots. Among the 48 viewers, 9 selected the data plot in the quantile-quantile (QQ) plot to evaluate the normality of conditional errors based on the standardised conditional residual, thereby driving the visual p-value of 0.00055. This leads us to reject the null hypothesis that represents, overall, the conditional error does not obey a normal distribution. Version 3 performs best among these three versions since it accounts for the largest proportion, and observers are able to point out the data plot

Table 5.1: *Overview of all lineup evaluations*

Linups	Version 1	Version 2	Version 3
Marginal residual plot	1/48	3/44	12/58 ***
Condtional residual plot	0/42	4/53	21/55 ***
QQ plot for conditional residual	3/49	9/48 **	26/53 ***
QQ plot for confounded residual	2/44	5/49 .	11/57 ***

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

among 19 null plots. The p-values of lineup are significant from all the results shown in the version 3, which is where we are able to reject the null hypotheses.

By comparing the ratios, we find that the standardised conditional residuals are more prominent than the standardised marginal residuals when checking the outlying observations. 21 out of 55 audiences are able to identify the actual data plot under the conditional residuals in the residual plot, while 12 out of 58 participants under the marginal residuals. The results shown are what we expected, because the conditional residuals squeeze the points which shows less information about the pattern but with more outstanding outlying observations. In addition, using quantile-quantile (QQ) plot to explore the normality of conditional errors, the performance of standardised conditional residuals is better than the performance of least confounded conditional residual, which is surprising.

Furthermore, we want to identify if the position of the data plot matters in Table 5.2. We only consider the cases from version 2 and version 3 since they have more correct responses compared with version 1. The "1" means that the first two replicates, and "2" represents the last two replicates. Since the lineups are generated as 4 rows and 5 columns by label 1 to 20 from left to right. The residual plot in version 2, the data plot is located at 5 for first and second replicates while at 15 for the rest two replicates. For the version 2 in QQ plot, the data plot is allocated at 8 and 20 respectively. From the result, we can imply that if the data plot is at the corner, then it seems to be easier to be identified. Besides, if the data plot on the first row, it will be picked out quicker for example in the version 3. The data plot of the residual plots is in 15 and 2 respectively. And 11 and 6 are the two positions that apply in the QQ plot for version 3.

Table 5.2: *Summary table of position for lineup*

Lineups	V2.1	V2.2	V3.1	V3.2
Marginal residual plot	3/25	0/19	4/28 *	8/30 ***
Condtional residual plot	3/29	1/24	9/26 ***	12/29 ***
QQ plot for conditional residual	2/23	7/25 ***	15/29 ***	11/24 ***
QQ plot for confounded residual	2/29	3/20 .	4/30 .	7/27 ***

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

Table 5.3: *Summary table of position for lineup*

Lineups	Categorical	Numerical	Mixed
Marginal residual plot	2/19	2/21	8/18 ***
Condtional residual plot	4/19 *	3/19 .	14/17 ***
QQ plot for conditional residual	0/11	12/21 ***	14/21 ***
QQ plot for confounded residual	1/14	2/17 .	8/26 ***

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

Based on version 3, the data type may also affect the results. As there are three data sets that have been used in this project, that is Autism case, Sleep study case, and Linguistic case. Each data set represents a particular data type, where linguistic case is the representative of categorical data type, numerical data type belongs to sleep study case and the rest, autism is the mixed data type. From the Table 5.3, it shows that the mixed data type behaves well than the others, as the observers are able to recognize the data plot at the most proportion. Nevertheless, the categorical data types perform least as there are few people point out the actual data plot even in the version 3. In addition, among the mixed data type, it is better to use the standardised conditional residual to detect the outlying observations than based on the standardised marginal residual. Otherwise, the QQ plot applied in least confounded conditional residuals for conditional error model checking is harder to identify the data plot compared with standardised conditional residuals.

We asked the independent viewers to record their confidence level when they choose the data that is the most different one. 1 means very uncertain, while 5 expresses very certain. From Fig. 5.2, it is interesting to find that, for all three versions and data types, the neutral confidence level (3) has the biggest proportion. However, there is a dramatic increase from version 2 to version 3 with more amounts on higher confidence levels. Especially, in mixed

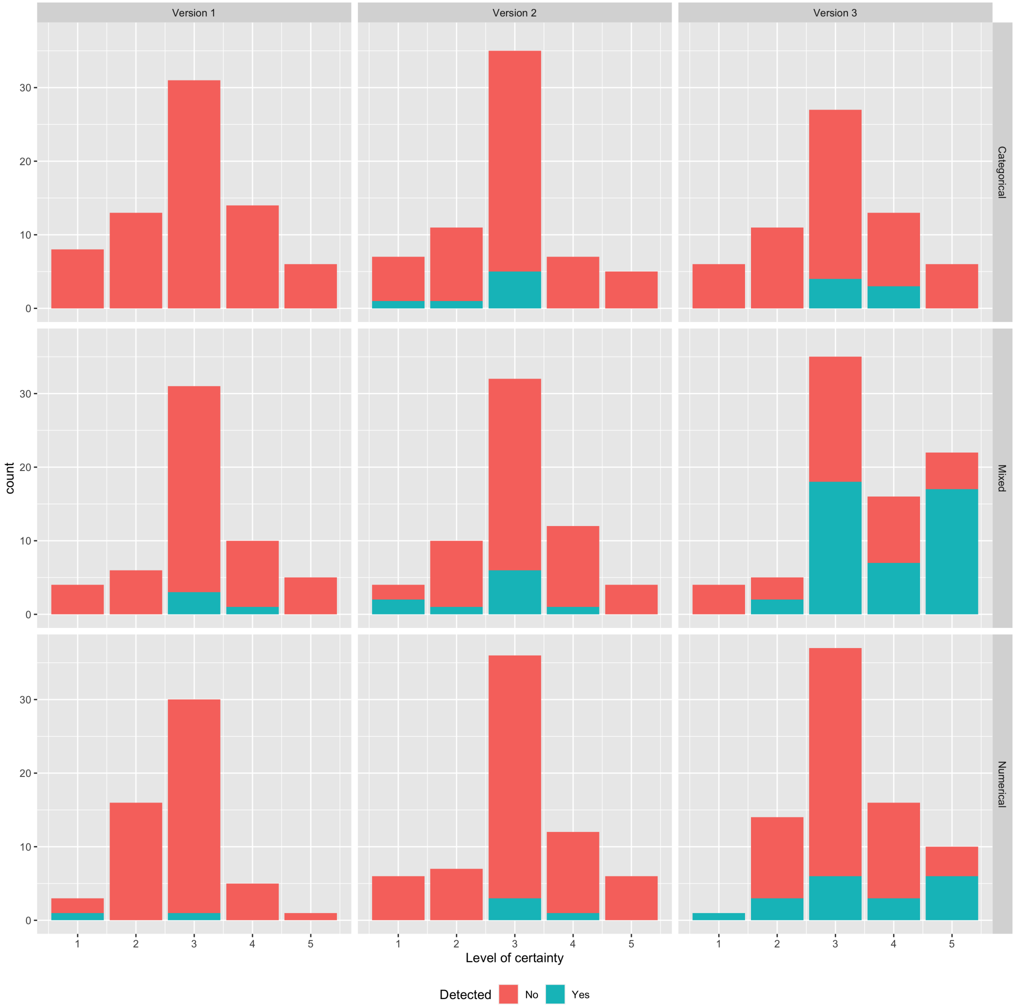


Figure 5.2: *Level of certainty that made by observers when they reading the lineups*

data type, more observers are very certain when they make the choices. Furthermore, the correctness rates are rising from version 1 to version 3. Even they are not sure when they read the lineups, the data plot can be told from the null plots. In addition, with the confidence level from neutral above, there is an excellent ratio of identifying the data plot according to the mixed data type, which is followed by the numerical data type.

Chapter 6

Discussion

At this stage, there are some limitations that occurred when we built the survey and some unexpected things happened as well.

There are three data types for three different versions with four replications each, that is 144 lineups in total. If total number of observers is small, the result will become unreliable since some images may have few or even no responses, so we tried to encourage more individuals to join us. Based on the survey that we created, we did not provide the reference lineup such that the visual ability of the individuals is not tested. It may result in the biases since we included the responses from those who are not well good at visual detecting. If the observers are able to identify the data plot from the reference lineup, their responses will be accepted otherwise, we will remove them. If we can add the tolerance interval/area for QQ-plots and residual plots, it can give us more benefits on the result since it will be better in reflecting the null distribution for a specific diagnostic feature and improving objectivity for interpreting these plots. In order to make the images clear for the viewers to read, we made the lineups as big as possible and put the questions on the sub panel side. The observers may lose patience because they need to scroll up and down the pages to see the whole picture of the lineups, make comparisons and answer the questions by repeating the process for 12 times. Also, there may some important hits are listed in the last panel of lineup. For some viewers, they will miss the information about the last 4 panels of lineups if they are not carefully reading the maps. Moreover, it is not clear for participants to see the text that "Please click the SUBMIT button at the top of the window" on the question panel for the last image. In addition, observers need to cross the whole page to click the SUBMIT button which is at the top right corner whereas the questions are listed at the bottom left. Furthermore, the position of the data plot may also matter. Since we designed the location of each data plot among the null plots for each case whereas the position for residual plots are the same as well as the QQ plots. In the survey, we haven't asked the observers the reason of choosing the distinguishable plot among the lineup. This question may bring us some hit about the model. Adding some words to explain the goal of the questions may improve the responses. Because we did not tell any information about the aim of this project, when the individuals see the residual plots which have been coloured by diverse subjects, they might treat the plots as the cluster plots which means that they were looking for the cluster pattern.

Chapter 7

Conclusion

We have presented the graphical diagnosis using the lineups generated by simulation from the model, particular for the residual plots and quantile-quantile (QQ) plots based on three different type of residuals. The graphical approach is relatively new and involves working with human viewers and let them to choose the most different plot among the lineups. From the result, we can conclude that using residual plot for identifying the outlying observations under the standardised conditional residuals is more efficient than for standardised marginal residuals. Standardised conditional residuals also perform better in diagnosing whether the conditional error follows a normal distribution than the standardised least confounded conditional residuals.

Lineup protocol is the alternative tool when we test the hypothesis. Instead of simply rejecting the null hypothesis, lineup also tells why we are rejecting the nulls. However, the graphical diagnostic does not only depend on the simulation methods, design of lineups, but also the observers. Theoretical power of the visual test may be greater than the conventional inference when the power of visual test increases with a large number of observers. We remove all the contextual information in the lineups in the survey, such as title, axis and labels in order to make sure that observers are purely picking the most different plots among the null plots.

Acknowledgments

The author would like to thank Dr. Emi Tanaka for weekly meetings with suggestions and comments, Prof. Di Cook providing useful hits and Stephanie Kobakian supports the ideas for experimental setup and shiny app codes. We are thankful for the coding of residuals in the linear mixed model and the lineup structure provided by Adam Loy and Julio M. Singer.

The analysis of the work was completed in R (Team, 2019) with the use of the following packages:

- lme4 (Bates et al., 2015) and HLMdiag (Loy and Hofmann, 2014) were used for model fitting
- nullabor (Buja et al., 2009b), ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2020), tidyverse (Wickham et al., 2019) for visualization and data analysis
- png (Urbanek, 2013), grid (R Core Team, 2019) provided the image display

Appendix A

Additional stuff

Data sets that we used in this thesis are described in the following tables (Table A.1, Table A.2, and Table A.3). All the code and figures are in the GitHub, `https://github.com/kaiwenjanet/master`.

Table A.1: *Description of sleep deprivation study data*

Variable	Description	Min	Mean	Max
Reaction	Average reaction time (ms)	194.3	298.5	466.4
Days	Number of days of sleep deprivation	0.0	4.5	9.0
Subject	Subject number on which the observation was made			

Table A.2: *Description of autism study data*

Variable	Description
childid	Child ID
sicdegp	Sequenced Inventory of Communication Development group (an assessment of expressive
age2	Age (in years) centered around age 2 (age at diagnosis)
vsae	Vineland Socialization Age Equivalent as the response variable
gender	Child's gender with factor male (526) and female (78)
race	Child's race with white (400) and non-white (204)
bestest2	Diagnosis at age 2 with autism (389) and pervasive developmental disorder (215)

Table A.3: *Description of linguistic case data*

Variable	Description
subject	Subject number with female or male
gender	Individual's gender, female, and male
scenario	Scenarios that each individual facing, such as aking for a favor
attitude	Their attitudes with respect to different scenario, polite or informal
frequency	Voice pitch as the response variable

Table A.4: *Summary of LMM for sleep study*

	Reaction		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	251.41	237.91 – 264.90	<0.001
Days	10.47	7.41 – 13.52	<0.001
Random Effects			
σ^2	653.58		
τ_{00} Subject	35.86		
τ_{00} Subject.1	627.57		
ICC	0.61		
N Subject	18		
Observations	180		
Marginal R^2/Conditional R^2	0.352 / 0.747		

Table A.5: *Summary of LMM for linguistic study*

	Reaction		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	256.77	225.21 – 288.34	<0.001
attitude [pol]	-19.58	-30.38 – -8.78	<0.001
gender [M]	**-108.37**	**-149.56 – -67.19**	**<0.001**
Random Effects			
σ^2	637.84		
τ_{00} scenario	216.90		
τ_{00} subject	616.77		
ICC	0.57		
N subject	6		
N scenario	7		
Observations	84		
Marginal R^2/Conditional R^2	0.676 / 0.860		

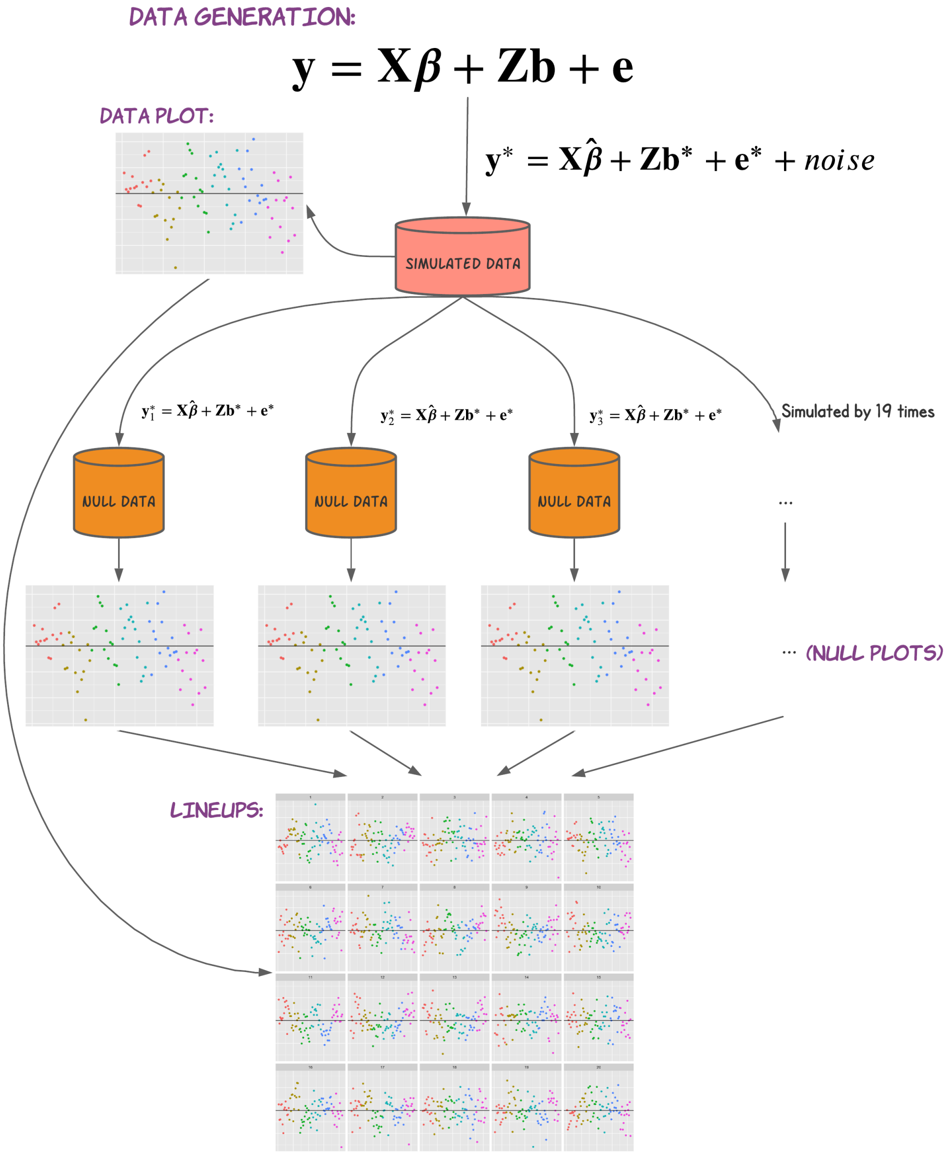

DATA GENERATION:
$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}$
DATA PLOT:
$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^* + noise$
SIMULATED DATA
$\mathbf{y}_1^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^*$
$\mathbf{y}_2^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^*$
$\mathbf{y}_3^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b}^* + \mathbf{e}^*$
Simulated by 19 times
NULL DATA
NULL DATA
NULL DATA
...
... (NULL PLOTS)
LINEUPS:

Figure A.1: *Simulation process*

Table A.6: *Summary of LMM for autism study*

	Reaction		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	7.35	4.06 – 10.64	<0.001
age2	3.80	3.09 – 4.51	<0.001
sicdegp [med]	1.01	-0.74 – 2.75	0.259
sicdegp [high]	5.75	3.75 – 7.75	<0.001
bestest2 [pdd]	1.95	0.39 – 3.52	0.015
gender [male] : racewhite	-0.18	-3.48 – 3.13	0.917
gender [female] : racewhite	2.24	-1.89 – 6.36	0.288
gender [male] : racenonwhite	-0.22	-3.65 – 3.21	0.901
Random Effects			
σ^2	39.16		
τ_{00} childid	12.77		
τ_{00} childid.1	0.12		
ICC	0.91		
N childid	155		
Observations	604		
Marginal R^2/Conditional R^2	0.359 / 0.941		

Bibliography

Anderson, DK, RS Oti, C Lord, and K Welch (2009). Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of abnormal child psychology* **37**(7), 1019–1034.

Banerjee, M and EW Frees (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association* **92**(439), 999–1005.

Bates, D, M Mächler, B Bolker, and S Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**(1), 1–48.

Beckman, RJ, CJ Nachtsheim, and RD Cook (1987). Diagnostics for mixed–model analysis of variance. *Technometrics* **29**(4), 413–426.

Belenky, G, NJ Wesensten, DR Thorne, ML Thomas, HC Sing, DP Redmond, MB Russo, and TJ Balkin (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research* **12**(1), 1–12.

Bryan, J (2020). *googlesheets4: Access Google Sheets using the Sheets API V4*. R package version 0.2.0. `https://CRAN.R-project.org/package=googlesheets4`.

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne, and H Wickham (2009a). 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.

Buja, A, D Cook, H Hofmann, M Lawrence, Ek Lee, DF Swayne, and H Wickham (2009b). Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Royal Society Philosophical Transactions A* **367**(1906), 4361–4383.

Christensen, R, LM Pearson, and W Johnson (1992). Case-deletion diagnostics for mixed models. *Technometrics* **34**(1), 38–45.

Dawes, J (2000). The impact of question wording reversal on probabilistic estimates of defection/loyalty for a subscription product. *Market. Bull* **11**, 1–7.

de Mendiburu, F (2020). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.3-3. https://CRAN.R-project.org/package=agricolae.

Haslett, J and SJ Haslett (2007). The three basic types of residuals for a linear model. *International Statistical Review* **75**(1), 1–24.

Henderson, CR (1973). Sire evaluation and genetic trends. *Journal of Animal Science* **1973**(Symposium), 10–41.

Hilden-Minton, J (1995). "Multilevel Diagnostics for Mixed and Hierarchical Linear Models, unpublished Ph. D". PhD thesis. dissertation, UCLA.

Kobakian, S and M O'Hara-Wild (2018). *taipan: Tool for Annotating Images in Preparation for Analysis*. R package version 0.1.2. https://CRAN.R-project.org/package=taipan.

Lesaffre, E and G Verbeke (1998). Local influence in linear mixed models. *Biometrics*, 570–582.

Loy, A and H Hofmann (2014). HLMdiag: A Suite of Diagnostics for Hierarchical Linear Models in R. *Journal of Statistical Software* **56**(5), 1–28.

Loy, A, H Hofmann, and D Cook (2017). Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics* **26**(3), 478–492.

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

Patterson, HD and R Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**(3), 545–554.

Pinheiro, J and D Bates (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Robinson, GK et al. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical science* **6**(1), 15–32.

Schützenmeister, A and HP Piepho (2012). Residual analysis of linear mixed models using a simulation approach. *Computational Statistics & Data Analysis* **56**(6), 1405–1416.

Singer, JM, FM Rocha, and JS Nobre (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review* **85**(2), 290–324.

Smith, AB (1999). "Multiplicative mixed models for the analysis of multi-environment trial data/Alison B. Smith". PhD thesis.

Team, RC (2019). *R: A language and environment for statistical computing (version 3.1. 2). Vienna, Austria. R Foundation for Statistical Computing; 2014.*

Turk, AM (2012). Amazon mechanical turk. *Retrieved August* **17**, 2012.

Urbanek, S (2013). *png: Read and write PNG images*. R package version 0.1-7. https://CRAN.R-project.org/package=png.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, R François, L Henry, and K Müller (2020). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr.

Winter, B (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv. *arXiv preprint arxiv:1308.5499*.