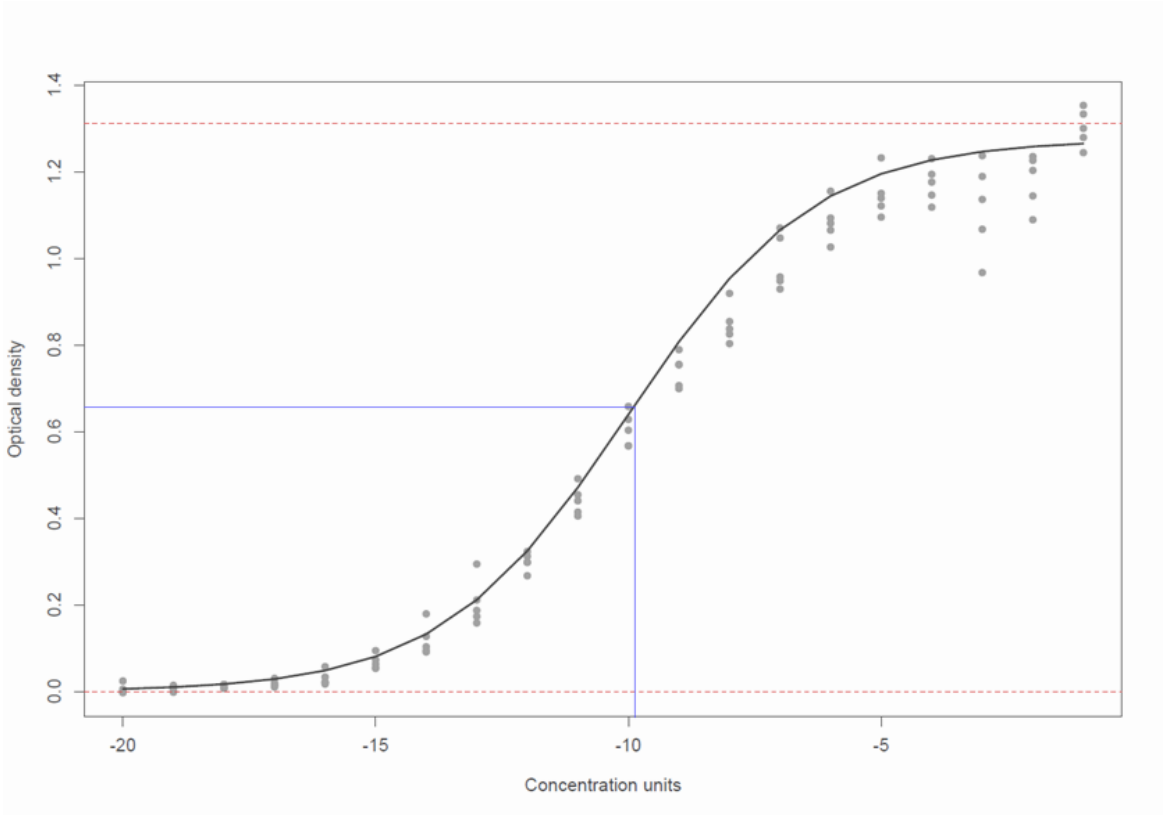

ELISA validation in the pharmaceutical industry: a mixed-effects models approach with R.

Albert Mayola Coromina

January 2019
Second Edition

HIPRA

UOC
Universitat Oberta de Catalunya

ELISA validation in the pharmaceutical industry: a mixed-effects models approach with R.

Author: **Albert Mayola Coromina**

Study plan: Màster en Bioinformàtica i Bioestadística UOC-UB

Area: Àrea 2 – Subàrea 2 – Anàlisi de dades

Director: **Núria Pérez Álvarez**

Subject coordinator**: Carles Ventura Royo**

Date: January 2019

Edition: Second proofread edition.

FITXA DEL TREBALL FINAL

Títol del treball:	*ELISA validation in the pharmaceutical industry: mixed-effects model approach with R.*
Nom de l'autor:	*Albert Mayola Coromina*
Nom del consultor/a:	*Núria Pérez Álvarez*
Nom del PRA:	*Carles Royo Ventura*
Data de lliurament:	*12/2018*
Titulació o programa:	*Màster en Bioinformàtica i Bioestadística UOC-UB*
Àrea del Treball Final:	*Àrea 2 – Subàrea 2 – Anàlisi de dades*
Idioma del treball:	*Anglès*
Paraules clau	*Mixed-effects models, ELISA, Validation*

Resum:

Els Enzyme-linked Immunosorbent Assays (ELISA) són la principal tècnica utilitzada per quantificació de molècules terapèutiques en la indústria farmacèutica. Basats en principis biològics, els ELISA estan subjectes a un ampli ventall de factors que introdueixen variabilitat en els resultats. Per tal d'obtenir suficient coneixement sobre el rendiment d'aquestes tècniques, els anomenats estudis de validació són part fonamental del procés de desenvolupament d'un nou producte. Habitualment aquests estudis s'analitzen mitjançant tècniques d'estadística clàssica com els ANOVA. Malgrat això, habitualment aquest tipus de dades són correlacionades i els dissenys no són balancejats i per tant les conclusions poden ser greument esbiaixades. Els models d'efectes mixtes són una alternativa potent i flexible que soluciona aquest tipus de problemes. La primera part del treball demostra l'aplicació de models lineals d'efectes mixtes per obtenir informació sobre l'exactitud i la precisió dels assaigs. En la segona part, s'apliquen models no lineals d'efectes mixtes per tal de comprovar l'existència de paral · lelisme entre un producte de prova i un de referència en el context d'assaigs dosi-resposta. En general, els models d'efectes mixtes han demostrat ser una eina molt complexa però potent i versàtil. S'ha obtingut informació important sobre

l'impacte de diversos factors que causen variabilitat utilitzant els protocols descrits en aquest treball. També s'ha mostrat l'obtenció d'estimacions més precises pels paràmetres rellevants que defineixen la resposta dels assaigs. El treball s'ha executat en llenguatge R que ha contribuït decisivament al potencial d'aquestes tècniques.

Abstract:

Enzyme-linked Immunosorbent Assays (ELISAs) are the workhorse of therapeutic molecule quantification in the pharmaceutical field. As a biologically based assay, ELISA results are influenced by a myriad of potential nuisance factors thus, as a new protocol is being developed, validation of its performance is a key step. To gather enough knowledge about assay performance, validation studies are conducted and the resulting data is usually analyzed by classical statistical techniques like ANOVA. However, this data usually is generated through complex designs that result in correlated and unbalanced data and thus appropriate methodology should be applied to achieve correct and robust conclusions. Mixed effects models arise as a powerful framework that overcome the limitations of classical approaches. In the first part of this work, a methodology to gain insight into ELISA performance and finally obtain accuracy and precision estimates using linear mixed effects models is presented. Also, in the second part, non-linear mixed effects models are applied as a tool to establish parallelism between test and reference product preparations when conducting full dose-response assays. Overall, both linear and non-linear mixed effects have demonstrated to be complex yet extremely powerful and versatile tools. Great knowledge on the impact of potential nuisance factors has been extracted using the workflows presented and more precise estimates of important assay parameters have been established. The work has been developed in R statistical programming language which contributes to the potential of the framework itself.

Contents

List of Figures

Chapter 1

Project information

1.1 Context and rationale

This project has been developed within the bioassay validation frame, specifically concerning the Enzyme-Linked Immunosorbent Assays (ELISA) [2] in the context of veterinary pharmaceutical industry.

Consequently, methodology applied has been subject to the requirements of two of the most influential regulatory agencies: the United States Department of Agriculture (USDA) and the European Medicines Agency (EMA)[53, 16]. To adapt the complexity of this endeavour to the available time, the degree of fulfilment of these guidelines has been lowered and only the most critical aspects have been considered.

The project and subsequent report have been organized in two parts both concerning to different aspects of the validation process of an ELISA assay but nevertheless they are closely related. The first part is centered around the analysis of validation designs intended to quantify the accuracy and precision of the assays. In this kind of assays, one to several distinct product preparations are tested at different locations, at different time points and by different laboratory technicians. Accordingly, resulting data has a longitudinal component but also other grouping structures given by the different factors considered. Due to this complex design structure the application of linear mixed effects models [44] has been tested as an alternative to classical statistical procedures with the final objective to comply with the recommended workflows and information required by the authorities [61].

The second part is devoted to the use of non-linear mixed effects models [44] to analyse ELISA data obtained from full dose-response experiments. In these experiments, the objective is to establish what is known as *relative potency* (RP) of the test preparation with respect to a pre-specified reference product of known properties. To calculate the RP, an appropriate model for the observed dose-response relationship needs to be specified. In this work, the 3-parameter logistic model has been used to describe this data generating process with the objective to obtain estimates of the defining curve parameters from whom to establish an RP. Previously though, parallelism between the test and reference serials dose-response curves should be demonstrated. The core of this part of the work has been to establish an appropriate workflow to attain this objective following the available regulatory documentation [61].

1.2 Objectives

Project global objectives express the main statistical objectives of this project. The type of statistical model used in each case has served as a basis to further divide the work in two different and independent parts. Each part has then its specific objectives which relate to particular tasks that have been executed.

1.2.1 Global objectives

- Part I: Analyse and interpret two ELISA validation studies using linear mixed effects models.
- Part II: Analyse and interpret ELISA dose-response data obtained from a single experiment using 3-parameter logistic non-linear mixed effects models.

1.2.2 Specific objectives

1. Part I:
   1.1 Correctly determine the data structure (nesting, crossing…).
   1.2 Application of linear mixed effects models to analyse the chosen data.
   1.3 Offer a correct interpretation of the results in the validation context.
   1.4 Generate an R markdown script to automatically analyse and report from the chosen study designs.
   1.5 Identify weak points in the analysed designs and propose how to improve them.
2. Part II:
   2.1 Analysis of ELISA dose-response curves using non-liner mixed effect models.
   2.2 Interpret and make inference on the model parameters.
   2.3 Calculate the asymptote and scale parameter ratios between a reference and a test serial based on a NLME model parameter estimates. Write an R script to automate this task.
   2.4 Calculate confidence intervals for the previously mentioned ratios using the delta method. Write an R script to automate this task.

1.3 Applied methodology

Three possible methodologies were identified at the beginning of the project, namely:

1. Start with a learning phase used to search and understand the theory of mixed effects models followed by a phase for experimenting with the available software packages and finally a phase where this previous knowledge is used on real data.
2. Start by applying the procedures described in software documentation to analyse sample data directly to real data and gain knowledge in a parallel way.
3. Start with a learning phase for the theory of mixed models based on a practical approach using sample data. Finally, apply the gained knowledge to the analysis of real data.

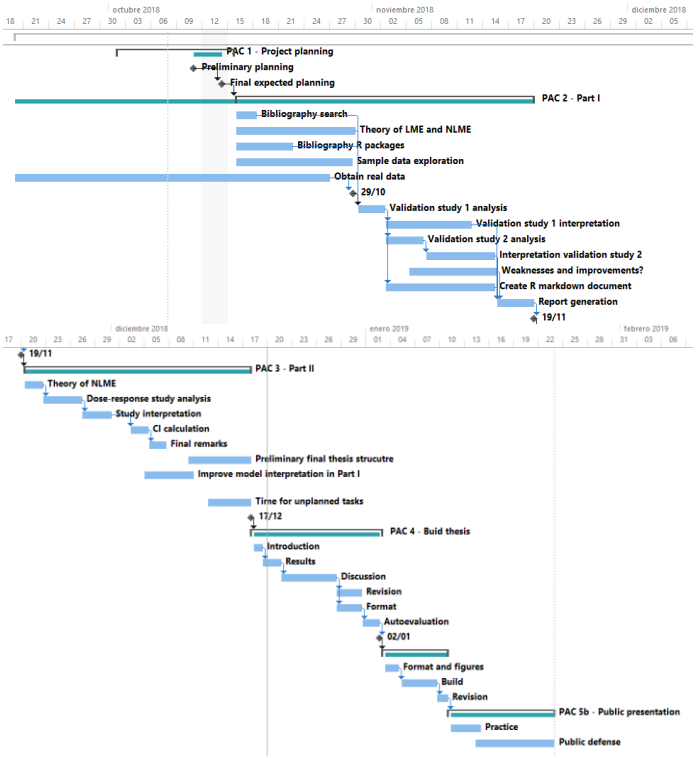

Figure 1.1: Gantt diagram for temporal planning including task names.

Initially strategies 1 and 3 were considered appropriate as both considered a learning period. Strategy 1 is a more classical approximation with practical learning following the theoretical one whereas strategy 3 is a *learning-by-doing* kind of strategy where theory and practice are considered in parallel. Strategy 2 was considered too risky from the beginning considering the theoretical challenge that mixed effects models represent. Following careful analysis and considering the director of the work recommendation, strategy 1 has been applied. It was considered that enough time was available to make this classical yet more robust approach.

1.4 Project planning

No special resources have been needed or employed to develop this project. Availability of real data to analyse during the course of the project has been granted thanks to the collaboration of Laboratorios Hipra S.A., affiliation of the author at the time of execution of this work.

Initial task planning has been subject to minor modifications in light of several events occurred during the execution, none of them critical. Justification and detailed explanation was given in the intermediate reports presented during the project development. The actual executed temporal planification and the several tasks that have conformed the project are presented in form of a Gantt diagram (Figure 1.1).

1.5 Summary of results

Results of this project include:

- This thesis, which includes an extensive bibliography review about mixed-effects models and demonstrates their implementation in distinct situations relevant to pharmaceutical validation.
- The R markdown document obtained as result of Part I containing common R syntax and custom functions to properly analyse validation studies using linear mixed effects models.
- The R script obtained as a result of Part II containing a custom function to calculate ratios and confidence regions for model parameters of a 3-parameter logistic non-linear mixed effects model. This is not included as a separate file but instead it can be found in Section 7.2.1.
- Intermediate progress reports obtained as a result of PAC 2 and PAC 3 containing relevant project follow-ups and updates.
- The final presentation that will be delivered according to the planned schedule.

1.6 Summary of chapters

As mentioned above, the work in this project has been organized in two parts each of them centered around an important topic in bioassay validation. Chapter 1 contains relevant information about the project itself as the rationale behind it and the list of global and concrete objectives that have been accomplished. Also, the real executed task planification is reported accompanied with the final list of products obtained.

Chapter 2 is an introductory text built through an in-depth bibliography review. Starting from the basic concepts related with ELISA bioassays the text guides the reader through the most important regulatory recommendations and obligations and ends with an extensive text covering the basic concepts related with mixed models. As an statistical centered project, this last part is extensive and covers the rationale behind the need for mixed models and its mathematical formulae. Also several important concepts are explained such as the meaning of random effects or nesting structures in designs. Most importantly, the end of the chapter covers the issue of testing the statistical significance of the model parameters which is an area of active research far from being closed or well established.

Chapter 3 covers the analysis of two ELISA validation studies each of them with some particularities, such as different number of random effects or model structure. Appropriate model fitting of linear mixed effects models with the R *lme4* package and model diagnostics is explained followed by a section on how to practically test statistical significance of both fixed and random effects, knowing that this is not a closed issue. Finally, a validation section covers the main objective of the analysis: check the performance of the assay based on accuracy and precision criteria.

Chapter 4 covers the fitting of a 3-parameter logistic mixed effects model for ELISA raw dose-response curves accounting for the grouping structure. Model fitting with the *nlme* R package is covered alongside the basic model diagnostics and inference on the parameters. Once the model is

obtained, parameter estimates for both a test and a reference serial preparations are used to demonstrate how to calculate the asymptote and scale factor ratios and their accompanying confidence intervals to check if curve parallelism can be assumed.

Chapter 5 contain the final thoughts regarding the lessons learned from the project, the degree of fulfilment of the initial objectives and a critical comment with respect the task planning and project execution. Also, some comments on future issues to explore are detailed there.

Chapter 7 contain all custom R code used during this work. All functions are properly annotated explaining possible dependencies, requirements and intended use.

Chapter 2

Introduction

2.1 Principles of ELISA assays

Biologically-based analytical methods to quantify specimens (bioassays) are the group of techniques that rely on the use of biological reagents, such as antibodies, live cells, etc., to quantify a substance of interest. This substance of interest is usually called analyte and, in the pharmaceutical vaccine industry, the analyte is in turn usually an antigen. The antigen is the substance that, when appropriately formulated, triggers an immune response from the immune system of a given animal [59]. The term antigen is a wrapper as the chemical structure of antigens is diverse ranging from purified proteins to lipids or, for example, whole cells of a given bacterium [59]. As diverse as the nature of the antigen is the type of immune response. Two basic types are antibody-based or cell-mediated and which type is generated greatly depends on the type of antigen and its accompanying adjuvants, which substances that are used to stimulate or boost the immune response. Also, immune responses are complex in that they are usually not exclusively of one type or another [23].

Enzyme-Linked Immunosorbent Assays or ELISA(s) are a type of bioassay in that they are based on the antigen-antibody recognition. Antibodies are generated such as they are specific against a given antigen or family of antigens. Thus, as the antigen-antibody interaction is specific, this property can be used to selectively find a molecule in a mixture of constituents, such as a vaccine. ELISAs make use of this property to specifically quantify analytes of interest in complex solutions, even in very low quantities [2].

Several types of ELISA exist that are used preferentially for different purposes as their properties, also varying, makes each type best suited for a particular task [2]. In the pharmaceutical industry, all types are used to quantify antigens in vaccines as the type performance is difficult to predict and selection is usually product dependent made by trial and error. Despite this, here the *sandwich* type will be explained as it is favoured by some authorities [52] and it is also widely implemented due to its outstanding sensivity [2]. This type is based in the use of two antibodies which are added in a sequential order and trap the antigen in between them to achieve a selective recognition. The second antibody is pre-labelled with a marker which is able to produce some sort of detectable signal (colour, light or fluorescence) by itself or when a substrate is added [2].

To gain a better understanding of the explained protocol, Figure 2.1 depicts the fundamental method to execute a *sandwich* ELISA. First, a solid support is coated with the capture antibody (1). After that, the product which contains the analyte to be detected is added (2). This causes the

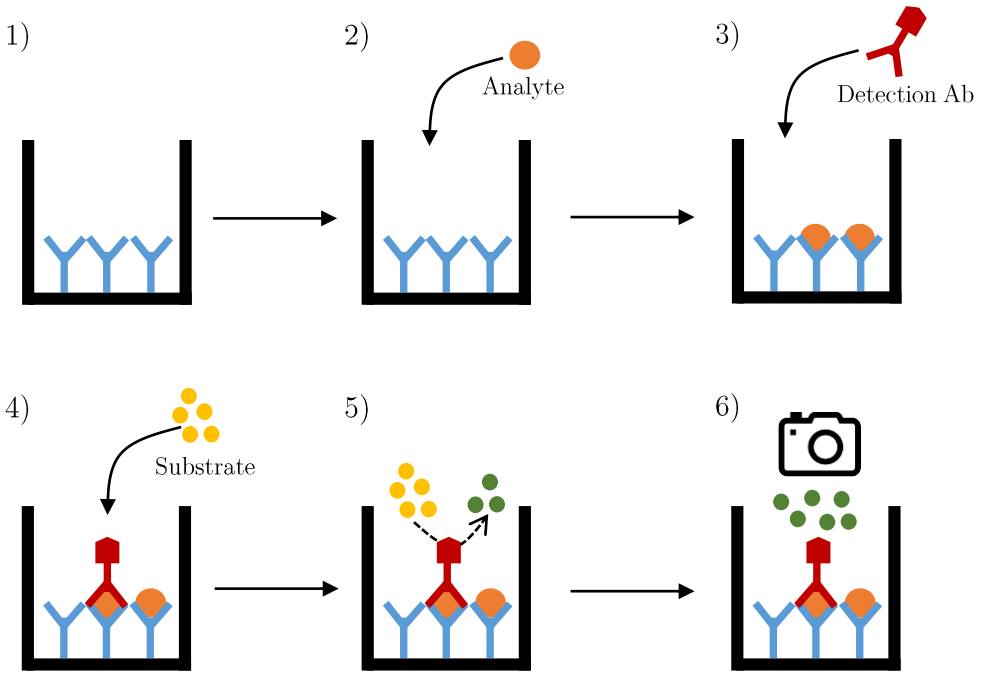

Figure 2.1: Execution phases for a *sandwhich* type ELISA.

formation of antigen-antibody complexes throughout the coated surface but, as capture antibody is used in excess, it is expected that some binding sites remain empty. Then, the detection antibody is added (3). Before that, this antibody has been chemically bounded to an enzyme which is able to produce a detectable signal by some type of chemical method. Usually, these enzymes chemically decompose a colourless substrate into a coloured product. Once the detection antibody-antigen complexes have formed, the *sandwhich* is obtained. Finally, the substrate is added (4), allowed to react during a specified amount of time (5) and the resulting signal is then measured (6) using an imager or some other detector. It should be noted that the arrows between phases indicate in-between procedures that are omitted for clarity. These procedures are incubations, usually conducted at a given temperature for a given period of time, and washes, that are used to remove unused products at the end of a step before proceeding to the next. As the solid support used is usually a microtiter plate of at least 96 wells, the black "U" shaped solid lines represent each one of this wells.

The final result of a usual assay are signal lectures for each well in a plate. The intensity of the signal of a well is proportional to the amount of analyte present during the assay, thus, by combining this procedure with dilution techniques, analysts are able to produce saturation to extinction response curves depending on the concentration of the analyte which finally can be used to quantify it. A close example of a quantification procedure albeit for other purposes, will be explained in Section 4.

2.2 Bioassay validation

ELISAs are a crucial tool in the manufacturing and quality control of pharmaceutical products. The production of pharmaceutical products is strictly regulated under the auspices of the Good Manufacturing Practices, or GMP, which demand all quantification methods used to characterize a product to meet certain standardized accuracy and precision criteria [61]. Validation is the process of demonstrating and documenting that a certain type of assay, which has been designed for a particular purpose (e.g. determine an analyte concentration), it is in fact suitable for the task and results in accurate and reliable results which can be appropriately reproduced [61, 16].

Validation process is extensive [20] and it is intended to characterize every aspect of a certain quantification method. Regulatory agencies, sometimes through child agencies or through pharmacopeial compendiums, issue their own guidelines on how to conduct the tests and report the results ensuring, at least, that common standards are meet inside their influence zones. These guidelines include a mixture of recommendation and unavoidable requirements and constitute the regulatory framework to which the pharmaceutical industry is bounded [61, 16]. At least 10 parameters need to be evaluated to complete a full validation protocol [1] but in this work only accuracy, precision and parallelism will be considered.

2.2.1 Accuracy

Accuracy is defined as the closeness of the value obtained by a certain method to the real value or a theoretically expected value so, it is a performance measure that quantifies any systematic bias the assay might have. In this work, formulation provided by the United States Pharmacopoeia (USP) [62], which is usually more restrictive than the European Pharmacopoeia (Ph.Eur.), is used. Thus, accuracy is to be reported in the form of a relative bias using percent units, % RB, using the following expression:

$$RB(\%) = 100 \left(\frac{Actual\ value}{Expected\ value} - 1 \right) \tag{2.1}$$

Usually the RB value is reported alongside a confidence interval. As the actual value used in the calculations is a parameter estimate of a mixed-effects model and the expected value is the theoretical value used to prepare the product and is a known constant, upper and lower confidence interval limits for the actual value, which are easily extracted from the model summary, can be used to calculate approximate confidence intervals by using the same formula.

2.2.2 Precision

Precision is defined as the closeness of repeated individual measures of the analyte and thus it is a performance value that relates to random variation introduced by potential nuisance factors. Is a more complex term than accuracy and it is usually evaluated at three distinct levels: repeatability, reproducibility and intermediate factor precision [45].

Repeatability is the variability observed when the assay is independently conducted several times under the same conditions within a short period of time. By the “same” conditions it must be understood to held as many factors as possible at a constant value. On the contrary, reproducibility is the variability observed when the assay is independently replicated several times but under

changing conditions and over a longer period of time. In this case, the assay protocol is maintained but the rest of factors potentially affecting its performance, e.g. laboratory and analyst, are allowed to change. Finally, intermediate factor precision is the variability observed when all factors except one are allowed to change. For example, intermediate laboratory precision would be measured by repeating the assay several times in the same laboratory under normal operating conditions but by different analysts and using different lot reagents [1, 62, 45].

Estimation of precision values is difficult [1] but the use of mixed-effects models makes it easier as variance components for a given model are a natural outcome of the analysis. In this way, estimates of the three types of precision measures considered can be obtained by combining the different variance component estimates [1, 65]:

- Repeatability (r) is estimated by the residual model variance: σ_r^2.
- Intermediate factor precision (*ifp*) is estimated from the sum of the residual variance and the variance component estimate for the factor of interest: $\sigma_r^2 + \sigma_{factor}^2$. Obviously the factor of interest should have been considered in the model specification.
- Reproducibility (R) is estimated as the overall variance taking into account all variance component estimates derived from the model: $\sigma_r^2 + \sigma_{factor_1}^2 + \sigma_{factor_2}^2 + ... + \sigma_{factor_n}^2$.

Usually, variance components are reported in the form of a percent coefficient of variation (CV) which is basically a way to standardize precision estimates to make them comparable between assays [1, 16]. A CV is a formally a ratio between a variance estimate and a related expected value, e.g. the CV for repeatability when a single sample is tested in the assay would be the residual variance estimated by the model divided by the mean quantification value of the sample in question. If percent transformed, then the % CV is obtained. Formulation to obtain the CV is:

$$CV(\%) = 100 \left(\frac{\sqrt{\sigma_i^2}}{E(analyte)} \right) \tag{2.2}$$

where i indexes the type of precision CV to be obtained so, as stated above, $i = r$, *ifp* or R and thus the computation of σ_i^2 changes accordingly. Note that the expression $\sqrt{\sigma_i^2}$ is used to calculate the variance in standard deviation units thus, if variance components are readily available in standard deviation units, this expression can be directly substituted by σ_i. *E(analyte)* represents an estimate of the expected value for the analyte used in the assay and it will usually be, for example, the mean concentration value for a given product preparation.

However, it is not always the case that reporting a CV is appropriate. When data is log-transformed and response is assumed to follow log-normal distribution the percent geometric standard deviation (% GSD) is recommended by the USP [62]. In this case formulation to obtain the % GSD is:

$$GSD(\%) = 100(exp^{\sqrt{\sigma_i^2}} - 1) \tag{2.3}$$

where σ_i^2 has the same interpretation as before. Other formulations are proposed that are demonstrated to work better in certain circumstances [65] but, as the above formulation is explicitly stated in the official guidelines, its use is recommended.

Finally, the above presented point estimates for either CV or GSD are usually reported with a confidence interval. Confidence intervals for ratios are difficult to calculate and, although several approximations exist [25], here they are calculated using bootstrap as it will be demonstrated in Section 3.

2.2.3 Parallelism

Quantification of an analyte using an ELISA is usually based on establishing a relative potency (RP) of a test product with respect to previously characterized reference product. Thus, RP is simply a ratio of the estimated potency of the test and reference products and it takes the following form [18, 60]:

$$RP = \frac{Potency_{(test)}}{Potency_{(reference)}} \tag{2.4}$$

The terms labelled *Potency* can be estimated by several means but, in this work, they will be estimated as the location parameters of a logistic model equation fitted to full dose-response data. This is extensively explained in Sections 2.3.4.2.1 and 4.

A key aspect for the potency of an unknown product to be established in this way is to ensure that dose-response curves between products are "equivalent". This requirement, present in all regulatory documents [61, 62, 63, 16, 53], is known as *parallelism* and is an unavoidable requisite to compute valid RP estimates. Conformance to this requirement is demonstrated by reporting the ratios between the same model parameters individually estimated for both the reference and the unknown products. If this ratios are not found to be significantly different from 1, then parallelism is granted and RP estimates are considered valid. On the contrary, if these ratios fall outside a predefined acceptability zone, parallelism can not be assumed and no RP estimate should be calculated [53, 63]. For a 3-parameter logistic model (3PL) these parameters are the upper asymptote and scale factor (see Section 2.3.4.2.1).

As usual, point estimates of validation parameters are to be reported alongside a confidence interval. The confidence interval can be then used to establish if the parameter is likely to fall within the acceptability region. However, to calculate confidence intervals for ratios is not easy. For example, as ratios are the quotient of two quantities, a problem arise if the denominator value is close to zero as it leads to an undefined situation. Moreover, this in turn causes distributional complexities and, for example, neither an expected value nor a variance are defined for a ratio [25, 42, 54]. Due to these issues, among others, confidence intervals for ratios are calculated through several approximations. Two methods will be employed here: the Fieller and delta methods. Whereas the Fieller method is considered to be the standard solution by the USP [63], the delta method, based on a Taylor expansion series, it is far simpler if appropriate assumptions are met. Those assumptions are: 1) denominator significantly different from zero and 2) denominator low standard error.

Formulas shown below are based on the article by Franz, H.V. (2007) [25]. The USP provides a different formulation for the Fieller confidence intervals that otherwise is equivalent to the one described here [63]. For the delta method, mathematical expression of the ratio and its confidence region is:

$$CI_{upper/lower} = \hat{\rho} \pm t_q|\hat{\rho}|\sqrt{\frac{\hat{\sigma}_x^2}{\hat{x}^2} + \frac{\hat{\sigma}_y^2}{\hat{y}^2} - 2\frac{\hat{\sigma}_{xy}}{\hat{x}\hat{y}}} \tag{2.5}$$

where $\hat{\rho} = \hat{y}/\hat{x}$ is the ratio and $\hat{y}$ and $\hat{x}$ are (unbiassed) estimators for the quantities of interest, in this case, the location parameters for the test and reference serials respectively. $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are the estimators for the variance and $\hat{\sigma}_{xy}$ is the estimated covariance between $\hat{x}$ and $\hat{y}$. Term t_q represents the $(1 - \alpha/2)$ quantile of a t-distribution with degrees of freedom equal to the number of plates minus 3 according to CVB-USDA [51].

For the Fieller method formulation is expressed as:

$$CI_{upper/lower} = \frac{(\hat{x}\hat{y} - t_q^2\hat{\sigma}_{xy}) \pm \sqrt{(\hat{x}\hat{y} - t_q^2\hat{\sigma}_{xy})^2 - (\hat{x}^2 - t_q^2\hat{\sigma}_x^2)(\hat{y}^2 - t_q^2\hat{\sigma}_y^2)}}{\hat{x}^2 - t_q^2\hat{\sigma}_x^2} \tag{2.6}$$

where the all the terms have the same interpretation as in Equation (2.5).

It should be noticed that both formulations are only valid if the denominator term of the ratio is significantly different from zero at the same significance level used to calculate the term t_q. Other limitations exist but they are out of the scope of this work so the reader is referred to the original reference for a complete discussion [25].

2.3 Mixed effects models

2.3.1 Rationale behind mixed models

The standard general linear model, also called ordinary least squares (OLS) regression, allows to describe the relationship between one or more predictor variables (categorical or continuous) and a response in a linear way. The modelled relationship however depends on the fulfilment of several assumptions made during the statistical derivation of the regression method. One of the most critical assumptions is that model residuals (error estimates) are **independent and identically distributed** (henceforth *iid*) with an expected value of zero and constant variance. This can be mathematically expressed as $\epsilon_i \sim iid(0, \sigma^2)$ [21]. This condition can be separated in two parts. The *identically distributed* errors implies that samples used to build the model were drawn from the same underlying distribution. Stronger than the former statement is the *independent* errors assumption. This states that the correlation among residuals is zero or, alternatively, the probability of a residual taking some value is independent of the values the other residuals have [56, 21].

However, in biological fields it is common to have highly structured experimental designs with clustered data or to take several measures on the same experimental unit over time rendering a *repeated measures* design [30]. In those situations it is expected that residuals (and hence observations) will have some degree of correlation and thus the assumption of independence is not held. For example, in the context of an ELISA assay, if a particular sample is tested at two different laboratories, replicates within each one location are expected to be more similar than replicates between different locations. Violations of the *iid* assumption leads to an increased type I error rates and standard statistical methods such as ANOVA are not robust against it [56].

A classical method for dealing with correlated observations is the well known *repeated measures* ANOVA (*rm*ANOVA). This method is technically simple and is based, like the traditional ANOVA,

in partitioning the variance of the response through sums of squares obtaining the expected mean squares (EMS). Then, by algebraic manipulation variance components can be analytically obtained, at least for the simpler designs [33, 22]. However, this method albeit simple has some major limitations. It requires a perfectly balanced design for the sum of squares decomposition to be unique, otherwise the decomposition method affect the variance estimates. It does not tolerate missing values in the data; usually this situation translates into a complete elimination of the observation. Finally, it is not expandable to complex designs with complex clustering structures often found in biological sciences [22, 33, 32].

In this context, mixed effects models arise as a powerful and flexible methodology that can overcome the limitations of classical methods by accommodating a variety of data structures, balance situations and is also capable of dealing with missing values. These advantages come at a cost of a complex methodology which requires more computational power but the latter should not be an issue nowadays, at least for the vast majority of situations [56, 33].

2.3.2 Fixed and random effects

In statistical modelling, the response is related to a set of measured predictor variables through several model *parameters*; one for each variable if they are treated as quantitative or one for each level if they are qualitative (also known as factors). These parameters are values estimated from the data and constitute our most plausible guess on how each variable or variable level relates with the response. These parameters are commonly called *effects*, but this terminology is more common when they are associated to levels of a factor rather than for continuous predictors [6].

These effects can then be subdivided into *fixed* or *random*. In mixed models framework, it is necessary to specify in the model structure if the effect of a particular variable is to be considered fixed, random, or in some cases both. This specification is crucial to correctly interpret model results and make inferences. A model parameter should be considered to be a fixed effect if it is associated to an entire population or to specific and reproducible levels of a factor. Put in another way, if for example the final interest is to make inferences about the specific levels of a factor without any aim to generalize, then this factor should be considered to have a fixed effect. On the other way, if considered as a random effect, the parameter is then a random variable which captures the random variability from known or expected sources of variation (e.g. subject) in the data that would otherwise be considered residual model variance. Following the same logic as before, if the levels of a factor can be considered as a random subset of a larger population and we wish to make inferences about that population then this factor should be considered to have a random effect. Note that, unlike fixed effects, random effects are always related to qualitative variables and in the context of mixed models these variables are usually referred to as *grouping variables* or *grouping factors* [56, 6, 22, 44].

2.3.3 Nested and crossed effects

As explained above, one of the key features of mixed models is the capacity to model a response as a dependency of several variables which can be considered to have fixed or random effects. In particular, categorical variables are usually considered to have random effects as they serve as grouping factors. When several grouping factors are present in a design, their relationship determines the properties of the possible models that can be fitted.

The effects of two variables can be nested, crossed or partially crossed and this relationship is a property of the experimental design [30]. Effects are considered to be crossed if one level of one of the variables is associated with more than one level of the other variable. When there is lack of balance, probably not every possible level combination among factors will actually be present in the dataset; this situation is known as partial crossing. Finally, a nested relationship exists if one level of a variable is uniquely associated to a particular level of the other. The nested variable has therefore a lower hierarchical position compared to the variable which contains it so these designs are also called hierarchical [50]. A good way to determine the correct structure of any design is to cross-tabulate the observations or use level plots but sometimes it is not obvious at all, specially with complex designs.

One of the main consequences of nesting is that it is no longer possible to calculate an interaction between factors and this will in turn affect how the variance is partitioned and the interpretation of variance estimates. This issue is extensive and beyond the scope of this introductory text but an excellent dissertation can be found in Schielzeth and Nakagawa (2013) [50].

2.3.4 Statistical model formulation

2.3.4.1 Linear mixed effects models

In statistical modelling, a model is called linear when its parameters enter linearly into the model equation. This means that the parameters appear in the model with a power of 1 and are not multiplied or divided by any other parameter. The predictors themselves do not need to be entered in the formula with these constraints for it to be a linear model [21].

Linear mixed effects models are simply linear models statistically formulated to allow both fixed and random effects to coexist. The most common matrix formulation for a linear mixed effects model with a single grouping level is, as described by Laird and Ware (1982) [64] and adapted from Pinheiro and Bates (2002) [44]:

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \epsilon_i \\ b_i &\sim N(0, \Psi) \\ \epsilon_i &\sim N(0, \sigma^2 I) \end{aligned} \tag{2.7}$$

where y_i is the N x 1 response vector for group i being N the total number of observations, X_i is an N x p design matrix of p parameters, β is a p x 1 column vector of fixed effects, Z_i is an N x q design matrix of q random effects (the random equivalent of X_i), b_i is a q x 1 column vector of random effects and ϵ_i is an N x 1 column vector describing the residual error term.

For the model to be complete, the assumed distribution of both random effects and residual error must be specified. In both cases, a Gaussian distribution with mean zero is assumed. To fully define a Gaussian distribution the variance-covariance matrix structure in each case must be also specified. For the residual error, ϵ_i, its variance-covariance matrix it is defined as $\sigma^2 I$ which implies constant variance and no within-group correlation of the residuals. This assumption is quite restrictive and can be effectively relaxed using appropriate tools to model heteroscedasticity and correlation structures such as in the *nlme* R package [43].

For the random effects, Ψ is the variance-covariance matrix. In its simpler form, e.g. when only a random intercept is considered, it is just a 1 x 1 matrix containing only the variance of the random

intercept. On the contrary, if a second random effect is added, e.g. another random intercept term representing another factor, this variance-covariance matrix is then a 2 x 2 matrix defined by both individual variances in the diagonal and their respective covariances as off-diagonal elements.

Random effects, b_i, and residual errors, ϵ_i, are assumed to be independent for different groups and for the same group.

Finally, it should be noticed that, despite the formulation, not all terms in the model should be estimated from the data. In fact, only the β vector parameters and the variance-covariance matrix for the random effects, Ψ, are to be estimated. However, the number of variances and covariances to be estimated grows quickly with the number of random effects and this is the main reason it could become computationally burdensome to fit these models [7, 56].

This formulation can be easily extended to accommodate more grouping levels as described elsewhere (see [44]; [56]; [3] for extensive discussions). It must be noted that, when several grouping levels exist, there also exist a mathematical expression to model each grouping level. Formulation presented here and in most texts refers to the lowest (observational) level which is usually the level of interest.

For example, suppose a simple unreplicated random intercept linear mixed model containing one fixed effect variable (e.g. sample) and two crossed random effects variables (e.g. analyst and day). Each variable is a factor with two levels. The matrix formulae describing such a model would be:

$$\begin{aligned} y_{ijk} &= X_i\beta + A_j a_{0j} + D_k d_{0k} + \epsilon_{ijk} \\ a_{0j} &\sim N(0, \Psi_1) \\ d_{0k} &\sim N(0, \Psi_2) \\ \epsilon_{ijk} &\sim N(0, \sigma^2 I) \end{aligned} \tag{2.8}$$

where $i = 1$ or 2 indexes the fixed effects variable, j and k index the grouping factors each of them taking values 1 or 2, y_{ijk} is the N x 1 response vector for the i-th element (sample) of j-th group of grouping variable analyst and k-th group of grouping variable day being N the total number of observations. The fixed effects design matrix, X_i, is an N x p matrix of p parameters and β is a p x 1 column vector of fixed effects. Being q_1 the number of random effects associated to grouping factor analyst and q_2 the number of random effects associated to grouping factor day, A_j is an N x q_1 random effects design matrix for analyst grouping factor and a_{0j} is q_1 x 1 column vector of random effects linked to them. Likewise, D_k is an N x q_2 design matrix for day variable and d_{0k} is a q_2 x 1 column vector of random effects linked to them. ϵ_{ijk} is an N x 1 column vector describing the residual error term. Finally, note also the definition of the assumed distribution for every random variable included in the fromula.

The common equation formulae for a given observation can be easily obtained by doing algebraic calculations with the general matrix formulation. For example, if matrix notation for the current example is fully expanded the model takes the form:

$$\underbrace{\begin{pmatrix} y_{111} \\ y_{211} \\ y_{121} \\ y_{221} \\ y_{112} \\ y_{212} \\ y_{122} \\ y_{222} \end{pmatrix}}_{y_{ijk}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}}_{X_i} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}_{A_j} \underbrace{\begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix}}_{a_{0j}} + \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}_{D_k} \underbrace{\begin{pmatrix} d_{01} \\ d_{02} \end{pmatrix}}_{d_{0k}} + \underbrace{\begin{pmatrix} \epsilon_{111} \\ \epsilon_{211} \\ \epsilon_{121} \\ \epsilon_{221} \\ \epsilon_{112} \\ \epsilon_{212} \\ \epsilon_{122} \\ \epsilon_{222} \end{pmatrix}}_{\epsilon_{ijk}} \tag{2.9}$$

Then, doing appropriate matrix operations the predicted model response for observation corresponding to sample 2 on day 1 by analyst 1 can be expressed as:

$$\begin{aligned} y_{211} &= \beta_0 + \beta_1 sample + a_{01} + d_{01} + \epsilon_{211} \\ &= (\beta_0 + a_{01} + d_{01}) + \beta_1 sample + \epsilon_{211} \end{aligned} \tag{2.10}$$

where it can be seen the global intercept definition for this observation as affected by a fixed intercept term (β_0) modified by a random quantity depending on the analyst doing the assay (a_{01}) plus another random quantity depending on the day the assay is conducted (d_{01}). The final response is then determined by a fixed slope (β_1) related to which sample is used and finally a random quantity (ϵ_{ijk}) is added representing the random error.

2.3.4.2 Non-linear mixed effects models

An excellent dissertation on non-linear mixed effects models can be found in Pinheiro and Bates (2002) [44] which will be summarized here. Linear mixed effects models presented in Section 2.3.4.1 are useful to describe how a response variable varies with a set of given covariates influenced for some grouping factors **within the observed range** of data points available. Thus, LMM (and all linear models) are empirically derived functions used to approximate the real behaviour of a complex and usually non-linear response inside a given, usually small, interval. As such, LMM models parameters do not have a direct physical meaning and they can only be interpreted as *the way variables relate to the response.*

By contrast, non-linear (mixed effects) models usually incorporate some kind of knowledge of the underlying mechanism generating the response. Most notably, non-linear models arise as a derivation of some physical law describing a phenomenon. In those cases, the model is said to be mechanistic and model parameters usually have a direct physical meaning relevant to characterize the phenomenon in question. Other non-linear models are empirically derived; this is, the exact mechanism producing the response is not theoretically defined but the model takes into account some theoretical properties that define the response and are of interest, e.g. asymptotes or inflection points.

Statistical theory behind non-linear mixed models is more complex than for their linear counterparts and algorithms used to fit the models are slightly more difficult to implement although they are based on the same tools, such as maximum likelihood procedures. Two clear advantages that justify the use of NLME models despite the increased complexity are: 1) non-linear models use

less parameters than linear approximations and 2) they can be valid outside the observed response range.

Statistical formulation for a simple non-linear mixed effects models with a single level grouping structure is:

$$\begin{aligned} y_{ij} &= f(\phi_{ij}, v_{ij}) + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned} \tag{2.11}$$

where j indexes an observation pertaining to the i-th group whose value is described by the non-linear and differentiable function $f()$ of several group specific model parameters ϕ_{ij} and possibly the value of some covariate or covariates expressed by the term v_{ij}. The residual error term is specified as being normally distributed with a mean zero, σ^2 variance-covariance matrix and are independent within them. It is not necessary that all model parameters are entered in the model non-linearly but at least one should be. Each group specific parameter is then modelled as:

$$\begin{aligned} \phi_{ij} &= A_{ij}\beta + B_{ij}b_i \\ b_i &\sim N(0, \Psi) \end{aligned} \tag{2.12}$$

where β is a vector of fixed effects, b_i is a group dependent vector of random effects which has Gaussian distribution with zero valued mean and Ψ variance-covariance matrix. Terms A_{ij} and B_{ij} are design matrices containing the grouping structure and covariate dependence of each observation respectively. Observations in different groups are assumed independent and residual error is assumed independent of the b_i component. As seen in Equation (2.11), errors are assumed independent and homocedastic but both assumptions can be relaxed if needed.

2.3.4.2.1 3 and 4-parameter logistic model

A special case of non-linear model and particularly important in bioassay modelling is the 4 parameter logistic model (4PL). This model has several formulations, the two most common being [60, 63]:

- 4PL formulation used by the US Pharmacopoeia and commercial software

$$y_x = \phi_1 + \frac{\phi_2 - \phi_1}{1 + (x/\phi_3)^{\phi_4}} \tag{2.13}$$

- 4PL formulation proposed by Pinheiro and Bates [44]

$$y_x = \phi_1 + \frac{\phi_2 - \phi_1}{1 + exp[(\phi_3 - logx)/\phi_4]} \tag{2.14}$$

where ϕ_1 is the upper asymptote, ϕ_2 is the lower asymptote, ϕ_3 is the x value at the inflection point and ϕ_4 is the scale parameter determining the slope of the curve. This interpretation is true

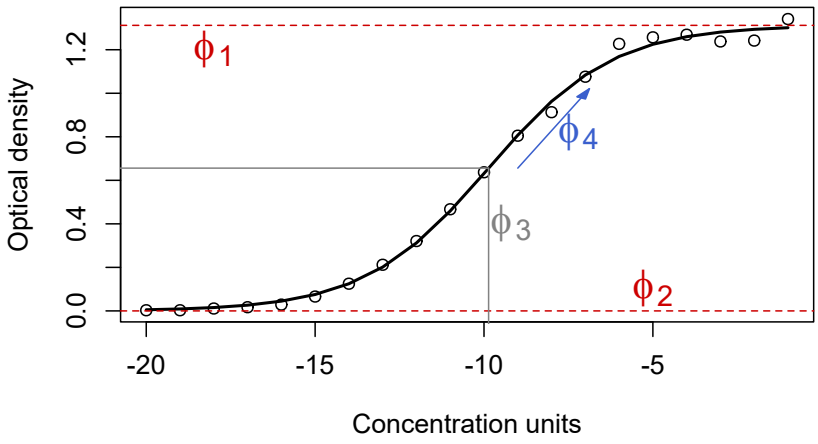

Figure 2.2: Example of a 4 parameter logistic function fitted to real data. The function defining parameters are shown identified by the greek letter ϕ plus a subindex.

provided the value of ϕ_4 is positive which means the response increases when moving to the right in the x axis. If the sign of this parameter is reversed, then the asymptotes interpretation is also reversed. The value of ϕ_3 is invariant to these changes. This parameter is also known in some texts and software by the name EC_{50} and represents the value of x where the response is halfway between its minimum and maximum asymptotical values. In quantitative immunoassay context, the quotient of estimated ϕ_3 parameters for an unknown and a reference preparation respectively is known as relative potency or RP. For a better understanding of these equations, a graphical interpretation of the four parameters using a real fit is shown in Figure 2.2.

The main difference between Equations (2.13) and (2.14) resides in the type of independent variable needed to fit them. In the context of this work, Equation (2.13) requires the x to be the raw dilution or concentration whereas Equation (2.14) requires the logarithm of the dilution or concentration to be used as x variable.

Usually in regulatory documentation the use of a 3 parameter logistic model (3PL) instead of a 4PL formulation is recommended [51, 60]. A 3PL model is simply a 4PL model with the lower asymptote constrained to be zero. This assumption is easy to meet in practice as usually the response is the optical density which is blank corrected by default so it has a naturally occurring zero asymptote. By using this easy method one less parameter needs to be estimated which usually results in less algorithmic convergence problems and stronger parameter estimates.

The 3 parameter versions of the previous model formulations can be expressed as:

- 3PL formulation used by the US Pharmacopoeia and commercial software

$$y_x = \frac{\phi_1}{1 + (x/\phi_2)^{\phi_3}} \tag{2.15}$$

- 3PL formulation proposed by Pinheiro and Bates [44]

$$y_x = \phi_1 + \frac{\phi_2 - \phi_1}{1 + exp[(\phi_3 - logx)/\phi_4]} \tag{2.16}$$

2.3.5 Parameter estimation

2.3.5.1 Contrast scheme for categorical variables

To properly estimate model parameters, regression methods require the independent variables to be numerical [21]. As a consequence, an *n*-level factor should be properly translated into a set of numerical-type *dummy* variables that, when considered together, express the same information. To this end, each one of the *n* levels should be converted to an independent column of the model matrix (also called design matrix); *n-1* new numerical variables (columns) are required to fully traduce a one column factor information [21].

Consider the following example.A 3 level factor with categories A, B and C will require two columns containing numerical values 0 and 1 to define the same categories. For example, an observation pertaining to category A can be expressed as having a value of 0 and 0 in both columns. A category B observation can be represented by having a 0 value on the firs column and a 1 on the second. Consequently, category C must be 1 and 1. Note that a third column is unnecessary.

The coding scheme used is technically called a *contrast* and its values determine the interpretation estimated regression coefficients [21, 41, 44]. In R, several pre-defined contrast schemes exist (treatment, sum-to-zero, Helmert,…) to fulfil a range of different needs but in specific situations where special needs may arise, one can specify custom contrasts [46].

The contrast scheme is usually irrelevant to interpret a raw regression output if the user knows exactly which codification has been used. Nevertheless, some contrasts are more well suited for particular situations than others like in ANOVA, the contrast scheme choice really matters [56].

As explained in most introductory texts [41] several types of sums of squares exist, at least: types I or sequential, II or hierarchical and III or marginal, following SAS nomenclature. A detailed discussion of each type and its particularities is beyond the scope of this text but, in summary, for balanced designs they all give the same results but for unbalanced designs, particularly when active factor interactions are present, the results differ as each decomposition is different in nature and assumptions [31, 35]. Except for type I, which is generally regarded as not appropriate for analysing experimental designs, the use of type II or III is not generally established but in general type III seems to be the recommended standard even though type II has shown to be more powerful under some circumstances [35].

In R, the default contrast scheme is called *treatment* contrasts (each factor level is coded with a combination of 0 and 1) and the default sums of squares decomposition is type I. Neither of these options is well suited to generalize to common analytical situations so it is advisable to change to more appropriate settings when working with regression models [56]. These changes are:

- Set the default contrast scheme to *effects* contrasts (contr.sum, each factor level is coded with a combination of -1 and 1). The code *options(contrasts = c("contr.sum", "contr.poly"))* is used to change the default behaviour of R to use *effects* contrasts for unordered factors and polynomial contrasts for ordered ones.

- When performing statistical tests, such as ANOVA, in situations that do not specifically require to use a sequential decomposition of sums of squares, use type III tests by default. As the default *anova* function form R *base* package do not allow it, to conduct type III tests one can resort to the *Anova* function in *car* package [24], *aov_car* from *afex* package [55] or *anova* in the *lmerTest* package [34]. The two later options being specifically developed for mixed-effects models.

The combination of this coding scheme with the use of type III tests should ensure to get meaningful and the more correct results in all standard situations.

2.3.5.2 Likelihood estimation

In classical OLS regression parameters are estimated by algebraic procedures involving several matrix calculations [21]. LME and NLME models however can be mathematically so challenging that trying to obtain a closed algebraic solution would be a tedious endeavour at best; therefore, algorithmic fitting is preferred.

Several methods have been historically described to for parameter estimation in mixed effects models but maximum likelihood (ML) and restricted maximum likelihood (REML) are by far the two most widely applied by common statistical packages [44]. Statistical understanding and derivation is complex and far from the objective of this work but a summary of the principal traits of each procedure is given below.

Maximum likelihood procedures are based on the idea of trying to find parameter values for a given model that *maximize the likelihood* of the observed data. A more plain way to describe the process could be: given a data generating process, some observed data is available. A model that is thought to appropriately describe the underlying data generating process is then selected to be fit. This model of course will have some set of defining parameters whose values are initially unknown. So, a maximum likelihood algorithm will be applied such that some parameter values are found. This parameter values are called maximum likelihood estimates and if used to generate data through the selected model it is likely that the observed data could be actually obtained.

Statistically, this is translated in a procedure to maximize the value of what is known as *likelihood function*, which is an expression for the probability density or mass function of the parameters given the data. Usually it is simpler to work with the logarithmic version of the likelihood function so the term *log-likelihood* arises but conceptually, as the logarithm is a monotonically increasing function, the same results will be obtained either way [22, 44].

A disadvantage of the common ML estimation concerning to the mixed effects model framework is that variance component estimates tend to be biased, concretely they can be be underestimated [44]. The REML method modifies the form of the likelihood function such as the final variance component estimates are no longer biased [44]. However a new problem arises as the REML criterion incorporates a parameter that depends on the fixed effects structure of the model [39, 44]. This has an impact on model inference as discussed in Section 2.3.6.

2.3.6 Evaluating significance

2.3.6.1 A note on p-values for mixed effects models

Evaluate the significance of either fixed or random effects through p-values in mixed effects modelling context is a controversial topic among statisticians and an active research area nowadays [39]. The controversy is of such magnitude that even Douglas Bates, one of the authors of two of the most used R packages devoted to mixed models, the older *nlme* and the newer *lme4* [43, 8], has been pushed to give explanations as to why the *lme4::lmer* function output does not provide any kind of p-values [4]. The lack of consensus is again demonstrated by the fact that PROC MIXED routine, SAS alternative to the aforementioned R packages and one of the most important commercial statistical packages in use, do report p-values using several methodologies [48]. Nevertheless due to the commented concerns those p-values should not be understood as an absolute truth.

Despite this statistical debate, the use of p-values as a method to express the relevance of findings is nearly unavoidable in the vast majority of scientific fields and this include pharmaceutical reporting. This is generally accepted and even the *lme4* package authors included some guidance on how to externally obtain the desired p-values; type *?lme4::pvalues* in R console [8].

As the significance testing issues are distinct considering if the interest is on fixed or random effects, each particular situation will be briefly outlined in the respective sections.

2.3.6.2 Significance of fixed effects

In general linear models whose parameters are estimated by OLS, parameter significance can be tested by a simple ANOVA using F-tests. In this case, the computed F statistics are known to follow an F distribution requiring the numerator and denominator degrees of freedom to calculate the critical value [41]. However, in mixed models inference two problems arise which are interconnected. First, the distribution of parameters obtained through ML or REML usually is complex and unknown and, although they are asymptotically normal this is not the case for common sample sizes [7, 44]. Second, as a consequence of the distributional problem of the likelihood estimates and the complex model structure given by the random effects part, there is not an accepted methodology to compute the degrees of freedom for t-tests or the denominator degrees of freedom for F-tests [3, 7, 39]. Nevertheless, several well known methods to calculate p-values for fixed effects are briefly outlined below.

Markov Chain Monte Carlo sampling

The method that is perceived to be the most reliable is the *Markov Chain Monte Carlo* (MCMC) sampling because it avoids the need to calculate any degrees of freedom [3]. The downsides of this lack of dependence are a huge computational cost and the algorithmic complexity leading to gaps in its implementation. This method was "briefly" implemented as an option in the *lme4* package but has been removed in the latest releases because of concerns over its wide-spread reliability [8, 39].

Its implementation is difficult but it can be accomplished in R using the *MCMCglmm* package [28].

Likelihood ratio tests

Likelihood ratio tests (LRT) are another alternative to conduct hypothesis testing on fixed effects. These tests are based on the idea of model comparison. Fundamentally, a model including the parameter of interest is compared against a reduced, less complex model called the null model. In the case of LRT for fixed effects the null model should be a model with the same parametrization except for the parameter representing the fixed effect of interest. Thus, LRT aim to determine if, given the data, the fit of a more complex model which includes a particular parameter is better than its alternative null model [44]. As in the case of MCMC sampling in this method no calculation of degrees of freedom is required and they can be used even for complex design structures [39]. However, they also have its downsides. These tests can only be used to compare models fitted using ML but not with REML. Also they tend to be anti-conservative; this is the calculated p-value is lower than it should be, so its use is discouraged [44].

LRT are implemented in R using the *anova* methods. Both *nlme* and *lme4* fitted models have this method available [8, 43]. Note that in this case the interest is in the sequential decomposition of the sums of squares so these functions use type I decomposition.

Wald *t* and *F*-tests

A third way of obtaining p-values for each fixed effect parameter is simply to use the Wald *t*-values reported for example in the *lme4* output and contrast them against a the t or z distributions depending on sample size [39]. Also, ANOVA-like F-tests could be used to make inferences regarding the whole term [44]. It should be noted that this tests are 1) conditional on the random effects structure and 2) they require the calculation of degrees of freedom [22].

As explained before, the calculation of degrees of freedom in mixed models framework may be problematic and several options are available depending on the software used, none of them free of controversy. For example, R package *nlme* and SAS use similar "inner-outer"/"within-between" rules like those used in classical ANOVA [13, 44, 48] but R *lmerTest* package and newer SAS procedures allow the calculation to be made by either the Kenward-Roger or the Satterthwaite method. The Satterthwaite method can be applied to both ML and REML models whereas the Kenward-Roger method can only be applied to the latter [29, 39].

In R, *lmerTest* or *afex* packages implement these methods based on *lme4* outputs [34, 55]. Specifically, F-tests for fixed effects factors can be obtained using the respective ANOVA functions. Those are also the primary tests implemented in the *nlme* function outputs [43]. Note that as stated in section 2.3.5.1, it is required that models were specified using the correct contrast scheme and to use type III tests in these functions to obtain meaningful and correct statistical tests.

Parametric bootstrap

Previously explained LRT rely on test statistics that asymptotically have a χ^2 distribution. The term asymptotically here means that those tests provide only approximate p-values and rely on several assumptions and situations arise where these approximations may be poor [22, 44]. Parametric bootstrap is a re-sampling technique that allows for the estimation of p-values from LRT without making any specific assumptions about the test statistic distribution or degrees of freedom but at a high computational cost[39]. Usually bootstrap is referred to as a non-parametric method but, since LME models assume some distribution for both residuals and random effects, this means it effectively becomes a parametric approach [22].

In R, parametric bootstrap for mixed models is implemented through in the package *pbkrtest* or *afex* [57, 55]. Also, although not a formal hypothesis tests, confidence intervals for parameter estimates can be obtained through the *confint* function of the *lme4* [8].

Which one to choose?

The work published by S. Luke (2017) [39] provides a good discussion comparing all of this methods except MCMC sampling. The author found that simple t-tests against the z distribution and the LRT approach both give anti-conservative p-values being the former alternative marginally worse than the latter. Also he reports that both methods were sensitive to small sample sizes so their use is only advised for sample sizes over 40 or 50 subjects or replicates.

The use of F-tests based on the Satterthwaite or Kenward-Rogers methods to approximate the denominator degrees of freedom produced close results when used with a REML fitted model. Those methods are reported to be slightly anti-conservative but they are somewhat robust against variations in sample size and thus they are preferred when sample sizes are small. When Satterthwaite correction was applied to ML fitted models the results showed an increased type I error rate so the use of REML is again stressed. When the design is complex and sample size is small, the Satterthwaite approximation might be more robust.

Parametric bootstrap performed well, better than LRT or t-tests but was found to be sensitive to small sample sizes. Its performance, therefore, was found to be no better than the F-tests using the Satterthwaite or Kenward-Rogers methods on REML fitted models, at least in the conditions the comparison was made.

Taking all into account, either Wald t-tests or F-tests using the Satterthwaite or Kenward-Rogers approximations should be preferred but it is advised to further confirm the results by, for example, computing parametric bootstrap confidence intervals. Models should be primarily fitted by REML and the t as z approach and LRT should be avoided when possible.

2.3.6.3 Significance of random effects

The previously commented issues regarding the calculation of degrees of freedom and unknown exact distributions still apply to random effects. Thus, to test random effects one could theoretically resort to the same tools already presented for fixed effects. As the tests are briefly explained above, the description shall not be repeated here; instead it follows a brief discussion on the particular problems of random effects testing and which methods are more generally accepted.

For random effects the interest usually lies in testing hypothesis of the form $H_o : \sigma_i^2 = 0$. Taking into account that variance values are strictly positive by definition, this kind hypothesis cause what is known as *boundary problem* [6, 14]. This situation arises because many tests, including Wald t and F-tests and LRT, assume that null values used for hypothesis testing do no take extreme values in their allowable range. The consequence are erroneous p-values which tend to be too conservative [30].

Due to the specific statistical derivation of each test related to the number of assumptions they require, LRT are preferred over Wald t or F-tests [14]. Specifically, Wald tests require the calculation of standard error for the variance components and assume that the test statistics asymptotically

converge to a chi-square distribution. Whereas LRT also assume an asymptotic chi-square distribution, they do not require the calculation of standard errors which are known to be extremely biased in most cases [5, 13, 10].

As previously explained for fixed effects (section 2.3.6.2), LRT work by comparing a full model including the random effect in question against a null model without it; its estimated value is supposed to be 0. Despite being the preferred method, some corrections on the p-values might still be required to address the *boundary effect*; for example dividing the p-value by 2 to make its value more close to the "real" one but this it is only a reasonable approach when testing a simple single random effect [14, 44]. Generally, if the p-value is sufficiently above or below compared to the decision rule (e.g. $\alpha = 0.05$), LRT should provide a good idea of the significance of an effect regardless of any correction [22]. Contrary to the situation for fixed effects, LRT to compare models differing in the random structure can be used when fitted by ML or REML but the former is not recommended due to the intrinsic biased nature of its variance estimates [44].

LRT tests for random effects can be implemented with the *anova* method for *lme4* fits [8] or by using the *ranova* function in the *lmerTest* package [34] which is more convenient.

If a more precise estimate of the p-value is needed the best approach are numerical methods, namely bootstrap. In R, two types of bootstrap procedures to obtain p-values for random effects are available: *fast* and *slow* depending on the computing resources. The *slow bootstrap* is the same re-sampling method considered above for fixed effects. As stated it is accurate and requires the least amount of assumptions but this comes with a high computational cost that translates into long waiting times, hence the *slow* adjective [38]. This method is implemented in R in the same way as for fixed effects: through *pbkrtest* package for a formal test or, if confidence intervals are desired, the *confint* function in the *lme4* package [57, 8]. The *fast* method is also a numerical method partly relying on bootstrap and uses the method described in Crainicieanu and Ruppert (2004) [17] and Greven et al. (2008) [26]. Statistical derivation is complex and far from easily understandable but, in short, the authors describe methods to obtain the exact null distribution of (restricted) LRT statistics under the *boundary* conditions explained allowing for fast and precise hypothesis testing [38]. This method is implemented in the *RLRsim* package in R [49].

Chapter 3

Part I: Analysis of validation studies

3.1 Validation study 1

3.1.1 Data structure

The first study was performed to validate a sandwich ELISA developed to estimate the relative potency of vaccine formulations against a reference control. Four different vaccine formulations differing in their antigen content were analysed by three different analysts at three different time points. Five replicates of each sample were run each time.

The dataset has the following structure:

It contains a total of 80 observations with no missing data and 7 variables, where:

- **sample**: A four level factor representing the four vaccine candidates used in the study.
- **rp**: A continuous variable representing the relative potency of each sample relative to a common control formulation.
- **analyst**: A three level factor representing the three distinct analysts enrolled in the study.
- **day**: A three level factor representing the tree distinct and consecutive time points at which the experiments were actually performed.
- **replicate**: This variable codes each replicate of each sample in five levels (1 to 5).

The first step in the design analysis is to determine the correct design structure. It is clear by variable definitions that *rp* is the response variable whose behaviour shall be modelled.

The *sample* variable is a factor representing the four distinct vaccine formulas tested in the study. Each formulation, from M1 through M4 contain an increasing amount of antigen and thus a different response is expected for each level. In fact, one of the interests of the analysis is to make inferences about possible differences between this particular set of levels. For this reason this variable should be considered a fixed effects factor.

On the other hand factors *day* and *analyst* should be considered as random effects factors as neither the three specific days nor the three specific analysts themselves are of any interest. This factors

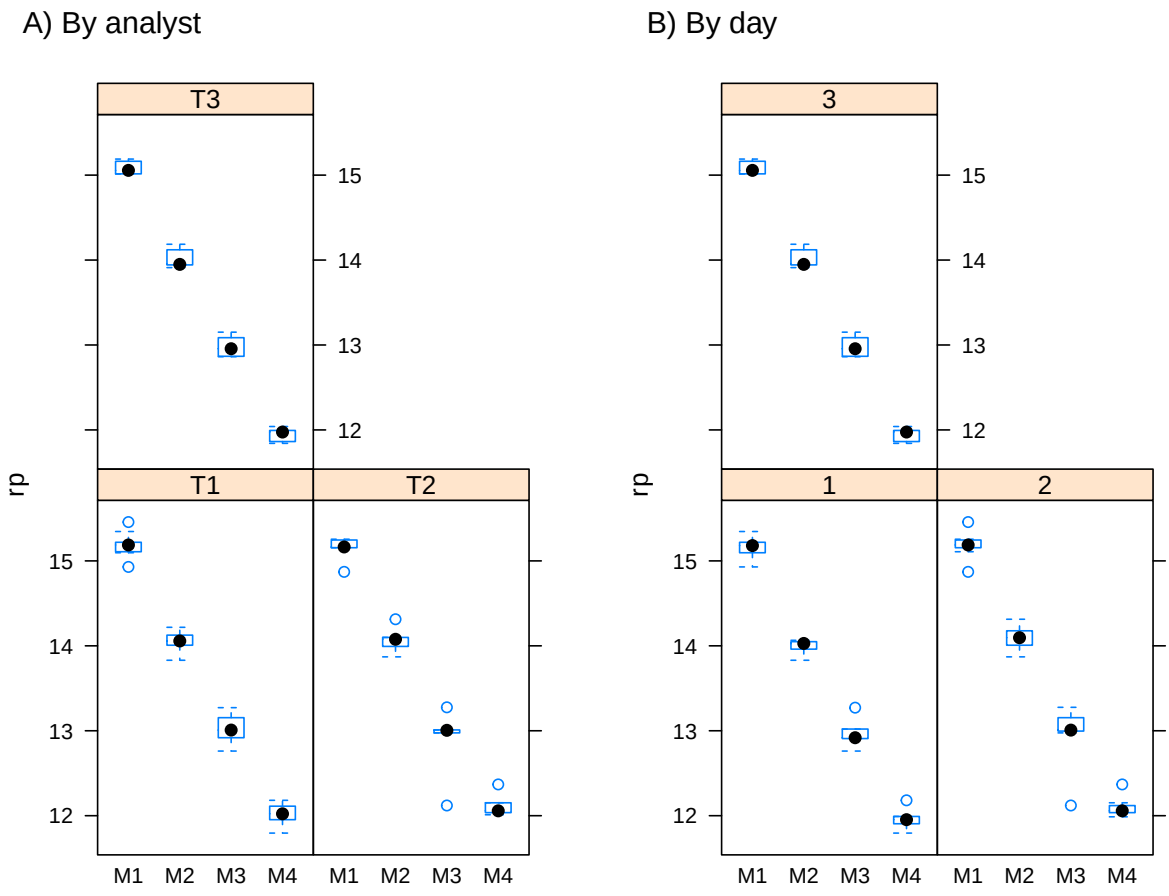

Figure 3.1: Boxplot of raw data by grouping factor analyst (A) or day (B).

represent a random selection of all the days and analysts that could have been chosen and the main research question is weather there is a substantial amount of variation in the response that could be explained by their inclusion in a model.

The goal of a validation study is to uncover the lack of robustness of an assay by controlling for known sources of variation that constitute the grouping factors, thus the ideal outcome is to find out a non substantial contribution of this factors to the overall variance. As it can be seen from the by-grouping-factor box plots in Figure 3.1, there is not a dramatic variation in the response attributable to any of the factors, at least that can be easily revealed by simple plotting.

Apart from deciding which variable effects will be specified as fixed or random, the main difficulty at the beginning of the analysis is to uncover the relationship existing between factors (crossing, partial crossing or nesting). This property is not defined by the model one would like to fit but instead is an inherent property to the design itself. The use of level plots, such as the ones displayed in Figure 3.2, is a helpful tool to uncover the underlying relationship in the data.

Panel A of Figure 3.2 reveals some interesting features. A plot showing only squares in the diagonal but not in the off-diagonal positions would be an indicative of a nesting relationship between factors, this is, levels of the nested factor (e.g. analyst) happens exclusively at one level of the parent factor

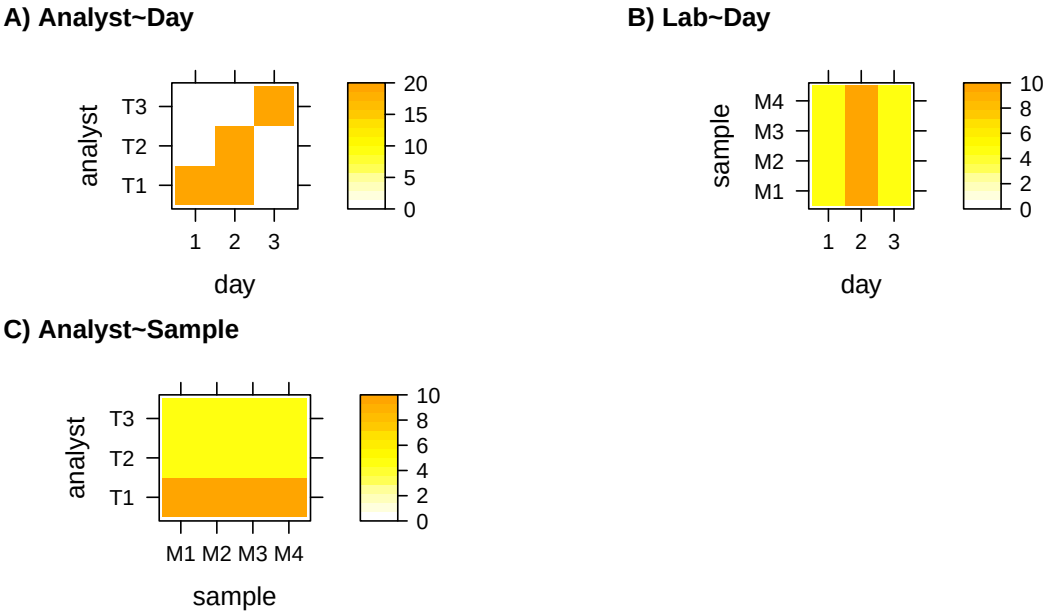

Figure 3.2: Levelplots displaying the number of observations at each variable combination.

(e.g. day) so for example at day 1 only analyst 1 was executing the experiments, at day 2 only analyst 2 and at day 3 only analyst 3. As it can be observed, this is nearly the case but an off-diagonal element exists as analyst 1 participates both at day 1 and 2. This excludes a nesting structure so factors must be considered crossed. In this case, however, they are said to be partially crossed as the design does not include samples at every possible level combination.

Panels B and C show the same plot for the relationship between sample and day and analyst and sample respectively. In both cases the plot clearly shows that samples at each factor-level combination are available so this implies a fully crossed structure among each pair.

Another interesting feature of this design uncovered by the level plots is the severe lack of balance. As can be interpreted through the colour scale there are substantial differences in the number of observations at each level combination.

3.1.2 Analysis

Once the data set has been inspected and the design structure has been defined it is possible to start the modelling phase. First a random intercept LME model is fitted to the full dataset by using the *lmer* function of the R *lme4* package [8].

Statistical formulation of the fitted model is:

$$
\begin{aligned}
y_{adsr} &= \beta_0 + A_{0a} + D_{0d} + \beta_1 sample_1 + \\
&\quad \beta_2 sample_2 + \beta_3 sample_3 + \epsilon_{adsr} \\
A_{0a} &\sim N(0, \sigma^2_{A_0}) \\
D_{0d} &\sim N(0, \sigma^2_{D_0}) \\
\epsilon_{adsr} &\sim N(0, \sigma^2)
\end{aligned}
\tag{3.1}
$$

where r = 1,...,5 define the replicate, a = 1, 2 or 3 defining the analyst, d = 1 or 2 defining the assay day and s = 1, 2, 3 or 4 defining samples M1 through M4 respectively. Vector y is the responses vector. Terms β_0 and β_1 to β_3 are fixed intercept and slopes terms respectively, defining the fixed effects part of the model. The β 1 through 3 represent the slopes with respect to the *n-1* dummy variables needed to represent the 4-level *sample* factor. Terms A_{0a} and D_{0d} are the random slopes associated with the analyst and assay day respectively defining the departure from the fixed intercept term depending on the value of each grouping factor. The within-group error term, ϵ_{adsr}, describes the random variation associated to each observation and together with the random intercept terms they define the random effects part of the model. To completely specify the model, it is assumed that random effects including the error term follow a zero centered *(multivariate)*-normal distribution defined by their respective standard deviations.

The model is fitted with the following command:

```
data11.mod1 <- lmer(rp ~ sample + (1 | analyst) + (1 | day),
    data = data11)
```

The R output for this first model fit, labelled *data11.mod1*, is shown below as an example but it will not be shown by default as it is deemed to complex to be routinely useful. Instead, for routine and printer friendly use, some functions to extract the most useful parts have been written.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: rp ~ sample + (1 | analyst) + (1 | day)
##    Data: data11
##
## REML criterion at convergence: -44.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4055 -0.4065  0.0385  0.4236  1.9884
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  analyst  (Intercept) 0.000000 0.0000
##  day      (Intercept) 0.001102 0.0332
##  Residual             0.026407 0.1625
## Number of obs: 80, groups:  analyst, 3; day, 3
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 13.54450    0.02679  2.20134  505.51 1.28e-06 ***
```

```
## sample1      1.60042    0.03147 74.31244   50.86  < 2e-16 ***
## sample2      0.49947    0.03147 74.31244   15.87  < 2e-16 ***
## sample3     -0.57317    0.03147 74.31244  -18.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) sampl1 sampl2
## sample1  0.000
## sample2  0.000 -0.333
## sample3  0.000 -0.333 -0.333
```

The first line gives information about the fitting routine used (REML) and the method used to calculate the approximate p-values for the effects when required (Satterthwaite's method) as the *lmerTest* package cover for the *lmer* function is used in this case [8, 34].

The next lines return the call to the *lmer* function and the RMEL convergence criteria which is an indicator of model fit and could be regarded as a kind of deviance for those readers familiar with generalized linear models.

Next, the numerical report of residuals distribution is shown. As in general linear models, in LME models framework residuals are assumed to have a $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$ distribution so, if this assumption holds, a mean (estimated by the median) close to zero and a kind-of symmetrical distribution is expected.

After this preamble, the model output shows the random effects estimates in both variance and standard deviation units (square root of variance). As this is a random intercept model, there is only an intercept term for each grouping factor. The function also returns the grouping structure that was understood by *lmer* and is extremely important to contrast this with the expected structure to detect and correct any miss specification.

Finally one can see the section summarizing to fixed effects. Estimates for each level of the factor *sample* and the intercept alongside their estimated standard errors and hypothesis tests are reported first followed by the estimated correlation structure among this fixed effects.

Instead of displaying this output every time a model should be described and in an effort to simplify model interpretation during this work, functions to extract the fixed effects, random effects and confidence intervals for the parameters have been written. This are independent from one another and allow for a more comfortable presentation of the information as shown below for fixed effects:

Table 3.1: Fixed effects for data11.mod1 model.

Effect	Estimate	Std.error	DF	t-value	p-value
(Intercept)	13.540	0.030	2.200	505.510	<0.001
sample1	1.600	0.030	74.310	50.860	<0.001
sample2	0.500	0.030	74.310	15.870	<0.001
sample3	-0.570	0.030	74.310	-18.210	<0.001

For random effects, apart form displaying the variance estimate and the corresponding value in standard deviation units, variance estimates for each term are transformed to percent values and represent the percent variance accounted for by the term in question:

Table 3.2: Variance component estimates for data11.mod1 model.

Variance component	Variance	% Variance	Std. deviation
analyst	0.000	0.000	0.000
day	0.001	4.007	0.033
Residual	0.026	95.993	0.163

As can be seen in the tables, all parameters for fixed effects of factor *sample* are clearly significant. However, when looking at the variance component table it is easy to see that variance accounted for by factor *analyst* takes the value 0; this is an indicative of non-significance. The amount of variance accounted for by *day* factor is close to 4 %, not a big value either. Variance accounted for by the residual term in the model is obviously 96 %. This term should be regarded in this case to be representative of the variance contained in the sample replicates which are not specifically defined in the model. Doing so, considering the replicates explicitly in the model formulation, caused a fitting routine error due to have a large number of data groups that consisted of only one observation and a small amount of observations (not shown).

Before jumping into any conclusion by using this model is necessary to evaluate the model fit and check whether it meets the LME model assumptions. This is more easily accomplished by diagnostic plots of the residuals and a check of data points influence on the model through Cook's distance as shown in Figure 3.3.

Diagnostic plots clearly show some problems with the model. Panels A and B are both residuals vs. fitted plots. They are equivalent to the traditional plots obtained from the *lm* function [46] and should be interpreted the same way. Thus, panel A clearly shows no lack of linearity but one can identify an odd point at a fitted value of 13 approximately. Panel B is another way to look at the same residuals vs. fitted plot but it further transforms the residuals by applying a square root function which has the effect of magnifying any trend in their magnitude [21]. As can be observed, homoscedasticity should not be a concern except for the already mentioned observation. The normal QQplot on panel C also shows a significant deviation from a normal distribution specifically at the left tail. Finally on panel D the Cook's distance for each observation is calculated as a measure of influence on the model. Again, it is possible to distinguish an odd observation whose Cook distance value greater than the suggested cut-off value of *4/number observations* marked by the red line [19]. This odd and influential observation can be identified as observation 51.

Caution should be exercised and is recommended when assessing influential points in data and even more caution should be used before considering any point an *outlier*. Figure 3.4 depicts the response values (*rp*) for *sample* level M3. It is clearly visible that observation 51 constitutes an atypical value with a behaviour off the general trend probably due to some kind of technical problem. Thus is reasonable to label it as an outlier and, as such, this point alone has a big potential to affect the model fit and to obscure important conclusions that could be drawn. It is decided to asses the model fit using a dataset without observation 51.

This diagnostic procedure has been conducted only on the residual part of the model, namely the random error term, but nothing has been said of *day* and *analyst*, the other random terms. This is because, albeit similar assumptions must hold in these cases, they are not easily checked with the extremely low number of levels each factor have so this part is simply omitted and assumed to be correct.

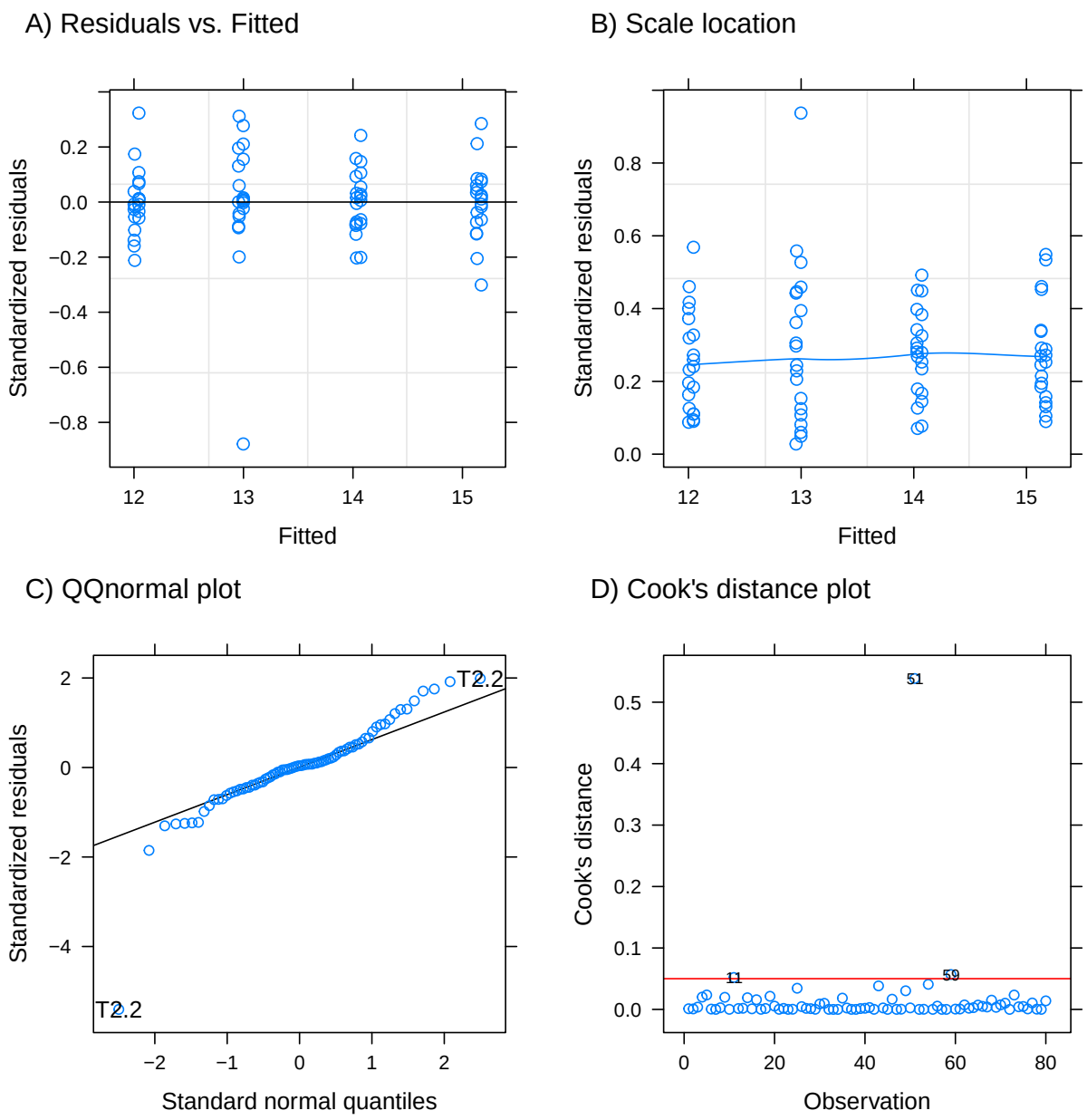

Figure 3.3: Diagnostic plots for data11.mod1 model.

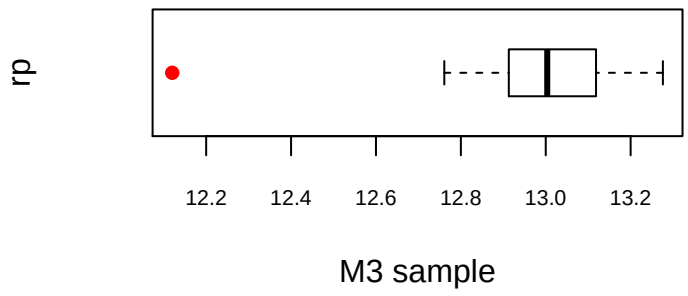

Figure 3.4: Response values (*rp*) for M3 samples with observation 51 in red.

This model is labelled as *data11.mod2* and has exactly the same model specification but observation 51 has been removed.

Tables for both fixed and random effects are shown below:

Table 3.3: Fixed effects for data11.mod2 model adjusted without observation 51.

Effect	Estimate	Std.error	DF	t-value	p-value
(Intercept)	13.550	0.030	2.120	403.160	<0.001
sample1	1.590	0.020	73.130	66.560	<0.001
sample2	0.490	0.020	73.130	20.430	<0.001
sample3	-0.540	0.020	73.140	-22.140	<0.001

The change in fixed effects estimates resulting from the removal of observation 51 from the dataset are very small only affecting the decimal part. Estimate values are nearly the same but it is worth noting that the uncertainty associated to the parameters for *sample* factor levels (std. error) has decreased marginally in all cases.

Table 3.4: Variance component estimates for data11.mod2 model adjusted without observation 51.

Variance component	Variance	% Variance	Std. deviation
analyst	0.000	0.000	0.000
day	0.003	15.453	0.053
Residual	0.015	84.547	0.123

The most noteworthy change is in variance components estimates. Despite the estimated variance for *analyst* factor remains collapsed to null, indicating the meaningless effect of this factor on the response, the value for *day* factor has increased from an accounted variance of 4 % to a value of approximately 15 %. Consequently the percent variance associated with the residual term has decreased from nearly 96 % to approximately 85 %. This is rather an important change and shows

the dramatic effect a single odd observation alone can have on the model fit and the resulting interpretation.

As usual, before going deeper into any conclusions model assumptions shall be checked. Figure 3.5 shows the diagnostic plot panel for *data11.mod2* object.

In this case, with the removal of observation 51, the residuals plots show that the assumption of a linear relationship is not violated. Moreover homoscedasticity seems reasonable across fitted values except for values around fitted value 13. This deviation however is small and should not compromise model interpretation. Normal QQplot now has a much better agreement with the diagonal line except for a small deviation on the right tail. In this case some kind of response transformations might help solving the issue. Nevertheless, as transforming variables may be a controversial topic [27, 37] and the impact of such a small deviation is deemed insignificant, no transformation will be applied on this occasion. The Cook's distance plot now shows a more reasonable range of values and distribution of points. Despite some points falling above the red line, indicating that they potentially have a high influence on the fit, the magnitude of the statistic is comparable to that of the cut off value so their influence may not significant. Nevertheless, further investigation of these points (not shown) indicates that they do not constitute atypical values and there is no documented evidence of technical errors. As such, removal is not justified and it is decided to maintain them in the analysis.

Thus, diagnostics now show a model with a good agreement with model assumptions and it is assumed to be valid for inference.

3.1.3 Inference

3.1.3.1 Tests on fixed effects

As explained in the introduction section 2.3.6, to evaluate significance of fixed and random effects in the mixed effects context is not trivial. For *data11.mod2*, it is clear from Table 3.3 that all levels of factor *sample* have a significant impact on the model fit as evaluated via Wald t -test using Satterthwaite method for degrees of freedom. Usually this marginal t-tests are difficult to interpret by itself so, on the grounds of clarity, an omnibus type III ANOVA test, using the same degrees of freedom calculation method (*lmerTest* package), is performed showing the overall significance of including the *sample* predictor:

```
anova(data11.mod2, ddf = "Satterthwaite", type = 3)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##        Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
## sample 108.06  36.019     3 73.134  2381.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.1.3.2 Tests on random effects

Table 3.4 reports the variance component estimates but no formal significance testing. In this case, testing is done by LRT provided by the *lme4 anova* method keeping in mind the limitations explained in the introductory Section 2.3.6.3. Each test requires at least two models to compute the

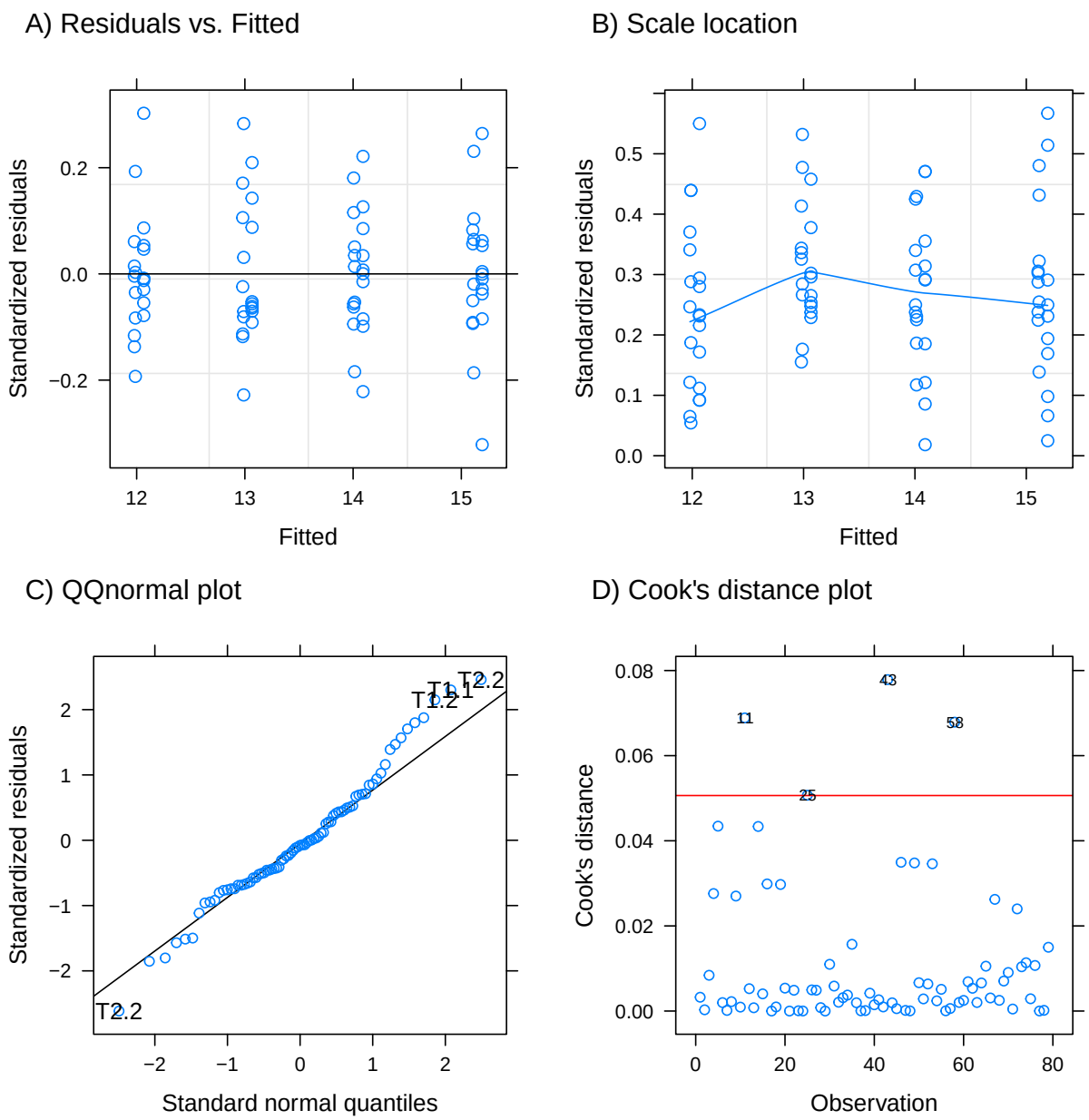

Figure 3.5: Diagnostic plots for data11.mod2 model adjusted without observation 51.

LRT so the first step is to fit what is known as the null model. In this case this is the same model but lacking the term whose significance is to be evaluated. The option *refit=FALSE* overrides the default behaviour of the *lme4 anova* method to refit the models using ML previously to compute the LRT. This is crucial when comparing models differing in fixed effects structure but it is not necessary in this case. The same results can be obtained with the *ranova* function of the *lmerTest* package [34]. The great advantage of using *ranova* is that it is not necessary to hand-build the null models as this step is a background task. Also, its output provides the significance for all random effects at once, requiring only one function call. On the grounds of clarity when reporting results, the custom function *tableanova* has been written. This function is dependent on the *ranova* function and it further formats output. Note that *anova* and *ranova* not necessarily return the same exact results and in general *ranova* is to be considered safer as it is less "manual" (type help(ranova, package="lmerTest") in R console for examples).

A second method using parametric bootstrap provided by the *pbkrtest::PBmodcomp* is used to validate results.

For the *day* factor significance testing, several approaches are presented:

LRT using the classical ANOVA

```
data11.mod2null <- update(data11.mod2, . ~ . - (1 | day))
anova(data11.mod2null, data11.mod2, refit = FALSE)
```

```
## Data: data11b
## Models:
## data11.mod2null: rp ~ sample + (1 | analyst)
## data11.mod2: rp ~ sample + (1 | analyst) + (1 | day)
##                 Df     AIC     BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## data11.mod2null  6 -66.088 -51.871 39.044  -78.088
## data11.mod2      7 -69.421 -52.834 41.710  -83.421 5.333      1    0.02093
##
## data11.mod2null
## data11.mod2     *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT using the *lmerTest::ranova* function

```
ranova(data11.mod2)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## rp ~ sample + (1 | analyst) + (1 | day)
##                npar logLik     AIC   LRT Df Pr(>Chisq)
## <none>            7 41.710 -69.421
## (1 | analyst)     6 41.710 -71.421 0.000  1    1.00000
## (1 | day)         6 39.044 -66.088 5.333  1    0.02093 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT using the custom function *tableanova*

```
tableanova(data11.mod2, capt = "LRT for random effects of data11.mod2 object.")
```

Table 3.5: LRT for random effects of data11.mod2 object.

Effect removed	Num. parameters	logLik	AIC	LRT	DF	p-value
	7	41.7103	-69.4206	NA	NA	NA
(1 | analyst)	6	41.7103	-71.4206	0.000	1	1.0000
(1 | day)	6	39.0438	-66.0876	5.333	1	0.0209

LRT using the *PBmodcomp* function (parametric bootstrap)

```
vcdaytest <- PBmodcomp(data11.mod2, data11.mod2null)
summary(vcdaytest)
```

```
## Parametric bootstrap test; time: 28.23 sec; samples: 1000 extremes: 3;
## Requested samples: 1000 Used samples: 407 Extremes: 3
## large : rp ~ sample + (1 | analyst) + (1 | day)
## small : rp ~ sample + (1 | analyst)
##             stat      df     ddf   p.value
## PBtest    5.0047                 0.0098039 **
## Gamma     5.0047                 0.0071574 **
## Bartlett 13.3155  1.0000         0.0002632 ***
## F         5.0047  1.0000 -1.2044
## LRT       5.0047  1.0000         0.0252785 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown, the inclusion of *day* effect seems justified as strictly speaking is significant at a 0.05 alpha level. Both the LRT implemented through *anova* and *ranova* functions or *PBmodcomp* function reach the same qualitative result with different p-values. The output of *PBmodcomp* function is slightly complex and the information in the *PBtest* line should be considered. It is worth noticing that the p-value obtained by the LRT is, as expected, more conservative than its parametric bootstrap counterpart. Note that in the parametric bootstrap output the function specifies that not all samples generated are used for calculations. This issue is related with the possibility of obtaining negative LRT statistic values which should not be theoretically possible. A good explanation in the package documentation is available [57] and should not be a problem if the number of used samples is sufficiently large and the LRT statistic is positive, as it is the case.

As a technical feature, note the concordance between ANOVA results of the three functions exemplified and differences between its outputs.

In this case, the use of the *RLRsim* package returned an error probably related with the small number of random effects levels [9].

NOTE:*In the previous tests, R code has been shown alongside its output. This code is applied in the same way in the next sections and will not be displayed again to enforce clarity in text reading.*

The same procedure is applied to the *analyst* factor, first using the classical ANOVA approach:

```
## Data: data11b
## Models:
## data11.mod2null.2: rp ~ sample + (1 | day)
## data11.mod2: rp ~ sample + (1 | analyst) + (1 | day)
##                   Df     AIC     BIC logLik deviance Chisq Chi Df
## data11.mod2null.2  6 -71.421 -57.204  41.71  -83.421
## data11.mod2        7 -69.421 -52.834  41.71  -83.421     0      1
##                   Pr(>Chisq)
## data11.mod2null.2
## data11.mod2                1
```

And then results are confirmed using the parametric bootstrap approach:

```
## Parametric bootstrap test; time: 31.37 sec; samples: 1000 extremes: 243;
## Requested samples: 1000 Used samples: 483 Extremes: 243
## large : rp ~ sample + (1 | analyst) + (1 | day)
## small : rp ~ sample + (1 | day)
##          stat         df     ddf p.value
## PBtest      0                     0.5041
## Gamma       0                     0.9678
## Bartlett    0 1.0000e+00          0.9998
## F           0 1.0000e+00 -1.6199
## LRT         0 1.0000e+00          0.9999
```

In this case the effect of factor *analyst* is clearly not significant. As before, both the bootstrap test and LRT agree in the qualitative outcome but differ in p-value being the LRT the most conservative as expected.

A good way to complement the above analysis is to calculate bootstrap confidence intervals for each parameter included in the model. Table 3.6 summarizes the calculations obtained using a 95 % confidence level with the *lme4::confint* function in R.

Table 3.6: 95 % parametric bootstrap confidence intervals for data11.mod2 parameter estimates adjusted without observation 51.

Parameter	Estimate	Lower limit	Upper limit
.sig01	0.000	0.000	0.061
.sig02	0.053	0.000	0.117
.sigma	0.123	0.102	0.143
(Intercept)	13.549	13.487	13.623
sample1	1.589	1.541	1.634
sample2	0.488	0.441	0.532
sample3	-0.538	-0.590	-0.491

In the table, parameter names are shown in the standard *lmer* function notation. For variance components, *.sig01* corresponds to the variance associated to the first random term entered in the model, in this case *analyst*. The *.sig02* parameter corresponds to the variance component associated with *day* factor. The residual variance of the model is always called *.sigma*. The rest of terms correspond to fixed effects parameters.

By using the standard knowledge on confidence intervals it can be seen that among variance components only the residual variance does not include the value zero so by integrating this information with the preceding tests the most appropriate results lecture might be that, given the data, the most relevant source of variation by far is the replication error. A null to moderate contribution from day-to-day error it is likely but it should not represent a concern. Error due to distinct analysts performing the assay is null to negligible under the evaluated circumstances. Nevertheless, the issue of variance intervals containing the zero value will be discussed in Section 3.3.

3.1.4 Validation

In section 3.1.3 a complete discussion about statistical significance of fixed and random effects based on common hypothesis tests was provided accompanied by a final interpretation of model results in the context of a validation study. Nevertheless, this interpretation might not suffice and, in light of regulatory requirements, a more specific discussion must be made providing estimates of bioassay accuracy and precision.

3.1.4.1 Accuracy

As stated in Section 2.2, accuracy is reported by calculating the percent RB which is expressed as the percent ratio between the estimated and expected potencies of the analyte at each concentration tested. Formulae to obtain point estimates is provided by the USP in its chapter 1033 [62].

Relative potency expected values for each sample tested in this assay were 15.15, 14.15, 13.15 and 12.15 for samples M1 through M4 respectively. Estimated group means can be obtained from the *data11.mod2* model fit using the *lmerTest::lsmeansLT* function which provides the *least squares* means (LS means, also called *marginal means*) for each group. It is important not to confound LS means with observed means; whereas the latter is a raw arithmetic mean for each group, the former is calculated based on the model fit ant thus it accounts for design characteristics such as imbalance. Expected relative potency values and model estimated relative potencies can be shown in Table 3.7:

Table 3.7: Model estimated and expected relative potency values for each sample.

Sample	Estimated	Expected
M1	15.14	15.15
M2	14.04	14.15
M3	13.01	13.15
M4	12.01	12.15

Immediately it is obvious that bias is small which indicates a high accuracy should be expected.

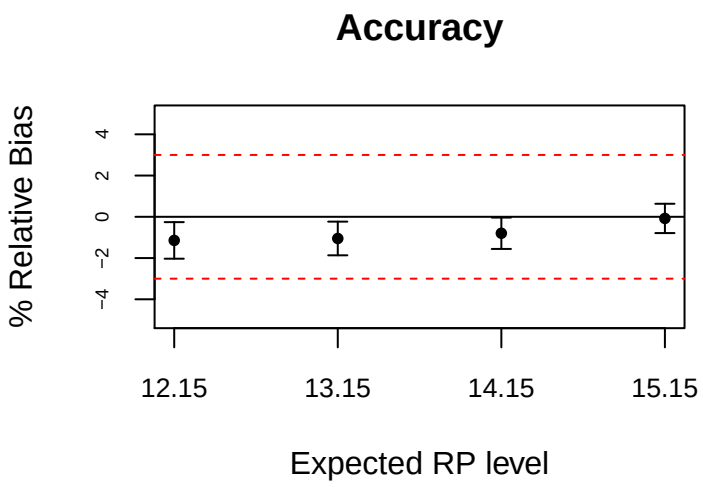

Figure 3.6: Relative bias plot for validation study 1.

Usually, accuracy point estimates are accompanied by a confidence interval as a measure of its trustworthiness. Several methods exist to calculate approximated confidence intervals but the *lsmeansLT* function output automatically returns a *t*-distribution based confidence interval using Satterthwaite degrees of freedom for each group mean. This group means can then be converted to % RB following the same formulation as to obtain the point estimate. An elegant way of reporting the % RB point estimates alongside its confidence intervals is to use a plot as the one showed in Figure 3.6.

The plot can then be used to compare the span of each confidence region with some pre-specified acceptance criteria. Target % RB is defined for each individual assay on the grounds of state of the art of the technique and prior knowledge [62] but, its derivation is not straightforward and it is out of the scope of this work. If a 15 % RB limit is assumed as an example, in this case all confidence intervals for the concentrations tested are far from this value so this assay would pass the test.

3.1.4.2 Precision

As explained in section 2.2, three distinct levels of precision can be reported: repeatability, intermediate factor precision and reproducibility. As stated, precision estimates are obtained directly by combining variance estimates from the mixed model analysis and reported as a % CV which, as a reminder, is the ratio between a given standard deviation and a related mean value [62].

The first step is to extract variance estimates from the mixed model analysis and report them directly in standard deviation units as shown in Table 3.8.

Table 3.8: Variance components for data11.mod2 object in standard deviations units.

Factor	Estimated Standard deviation
analyst	0

Factor	Estimated Standard deviation
day	0.05
Residual	0.12

Once obtained, this standard deviations can be combined to obtain the different precision estimates. Just as a reminder, repeatability is linked to the residual variance (σ_r^2) of the model. Intermediate precision for a given factor is calculated from the sum of the residual variance with the factor variance component estimate ($\sigma_r^2 + \sigma_{factor}^2$). Likewise, reproducibility is calculated from the total variance which corresponds to the sum of all variance components ($\sigma_r^2 + \sigma_{day}^2 + \sigma_{analyst}^2$).

To calculate the % CV for each precision estimate, a mean or expected value of the relative potency for the substance being analysed is necessary. In a dilutional assay like the one performed in validation study 1, several dilutions of the same analyte were tested giving place to several samples (M1 to M4).

Thus, in this case, there is not a single expected value estimate that can be used for calculations but instead there are four samples each one having its own expected relative potency value. Also, as a consequence of variance component analysis using mixed models, an overall variance component is estimated for each factor plus the residual term. Consequently, precision estimates can not be easily obtained for each *sample* level (potency level). A way to report an overall precision estimate in the form of % CV is by using the *grand mean* as a reference value which, if an appropriate contrasts setting has been defined, corresponds to the model intercept.

Table 3.9: Precision for validation study 1 assay (object data11.mod2) as % CV.

Precision	% CV
Repeatability	0.91
IP day	1.30
IP analyst	0.91
Reproducibility	1.30

Calculations are reported in Table 3.9. It is easy to see that repeatability and intermediate precision for analyst factor (IP analyst) have the same value. This is because the variance estimate associated with the analyst factor is zero, as shown in section 3.1.2. Likewise, reproducibility and intermediate precision for factor day (IP day) have the same value as reproducibility variance used in calculations is simply the sum of all variances and analyst factor associated variance is zero.

As explained above, it is customary to report validation statistics point estimates alongside its confidence intervals. Documentation consulted do not specify the method to be used to calculate this confidence region. Despite the sampling distribution of the coefficient of variation is known under some assumptions, the analytical approximation usually requires the calculation of the sample size [42]. As extensively explained, this is not easy in mixed model analysis as several grouping levels are present and sample size may have several interpretations depending of the level considered. For this reason, a confidence region for the coefficient of variation is approximated by the custom built R function shown below, which makes use of the *boot* and *boot.ci* functions of the R *boot* package [15]. In this specific example, the function estimates the 95 % confidence interval for IP

day precision but it can be easily modified to accommodate all the precision estimates.

```
# Function to calculate the % CV of a sample (see boot
# package examples)
calccv <- function(formula, data, indices) {
    dat <- data[indices, ]
    m <- lmer(formula, data = dat)
    mu <- fixef(m)[1]
    # Changing row subsetting indices for VarCorr allows for
    # other % CV to be calculated
    sigma <- sum(as.data.frame(VarCorr(m))[2:3, 5])
    cv <- 100 * sigma/mu
    return(cv)
}

# Actual bootstrap function, R specifies the number of
# bootstrap samples.  The more samples the best the estimated
# interval. The more samples the higher the computational
# cost. Formula specifies the model to be fitted.
bo <- boot(data = data11b, statistic = calccv, R = 500, formula = rp ~
    sample + (1 | analyst) + (1 | day))

# Conficence intervals
boci <- boot.ci(bo, type = "bca")
boci
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bo, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.993,  1.599 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

This function adjusts a mixed model using the same specification than the original one for a sub-sample of the experimental sample of the validation study. Then it extracts the variance components estimates and the grand mean and calculates the % CV of the sub-sample. With the bootstrapped statistics, then *boot.ci* function retrieves the bias corrected percentile confidence interval. The downside of this function is the high computational cost as it needs to adjust a model for each R bootstrap samples. The result shows that IP day % CV 95 % confidence region falls between 0.99 and 1.6 approximately. This can then be compared to a pre-specified value to accept or reject the performance of the assay.

The computational cost of this function is shown in Figure 3.7 as the system computation time dependent on the number of bootstrap samples, R. From the plot, it can be clearly seen that computational time linearly increases with increasing values of R and it takes around 40 seconds to compute the intervals with a reasonable R value of 500. Due to this, no further examples are provided.

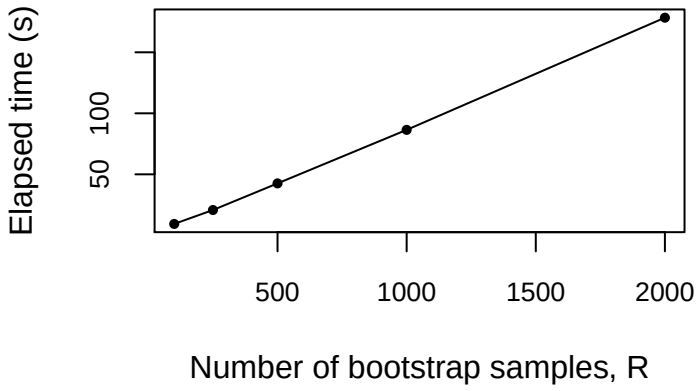

Figure 3.7: Computational cost for the % CV bootstrap CI calculation.

3.2 Validation study 2

3.2.1 Data structure

The second study was conducted to validate a sandwich ELISA developed to estimate the relative potency of vaccine formulations against a reference control. In this case, different vaccine formulations (*samples*) differing in their antigen content were analysed by three different analysts at three different time points and at two different locations. Five replicates of each sample were run each time.

The dataset has the following structure:

It contains a total of 125 observations with no missing values and 5 variables, where:

- **sample**: A five level factor representing the four vaccine candidates used in the study. Samples V050 through V200 contain an increasing amount of antigen. Thus, an increase in response should be expected in a dose-response fashion.
- **rp**: A continuous variable representing the relative potency of each sample relative to a common control formulation.
- **analyst**: A three level factor representing the three distinct analysts enrolled in the study.
- **day**: A three level factor representing the tree distinct and consecutive time points at which the experiments were actually performed.
- **lab**: A two level factor representing the two laboratories where the assay validation protocol was executed.

As in the previous example, first step is to determine the correct data structure to correctly specify the model. Concerning the variables, this design is exactly the same as before with the exception

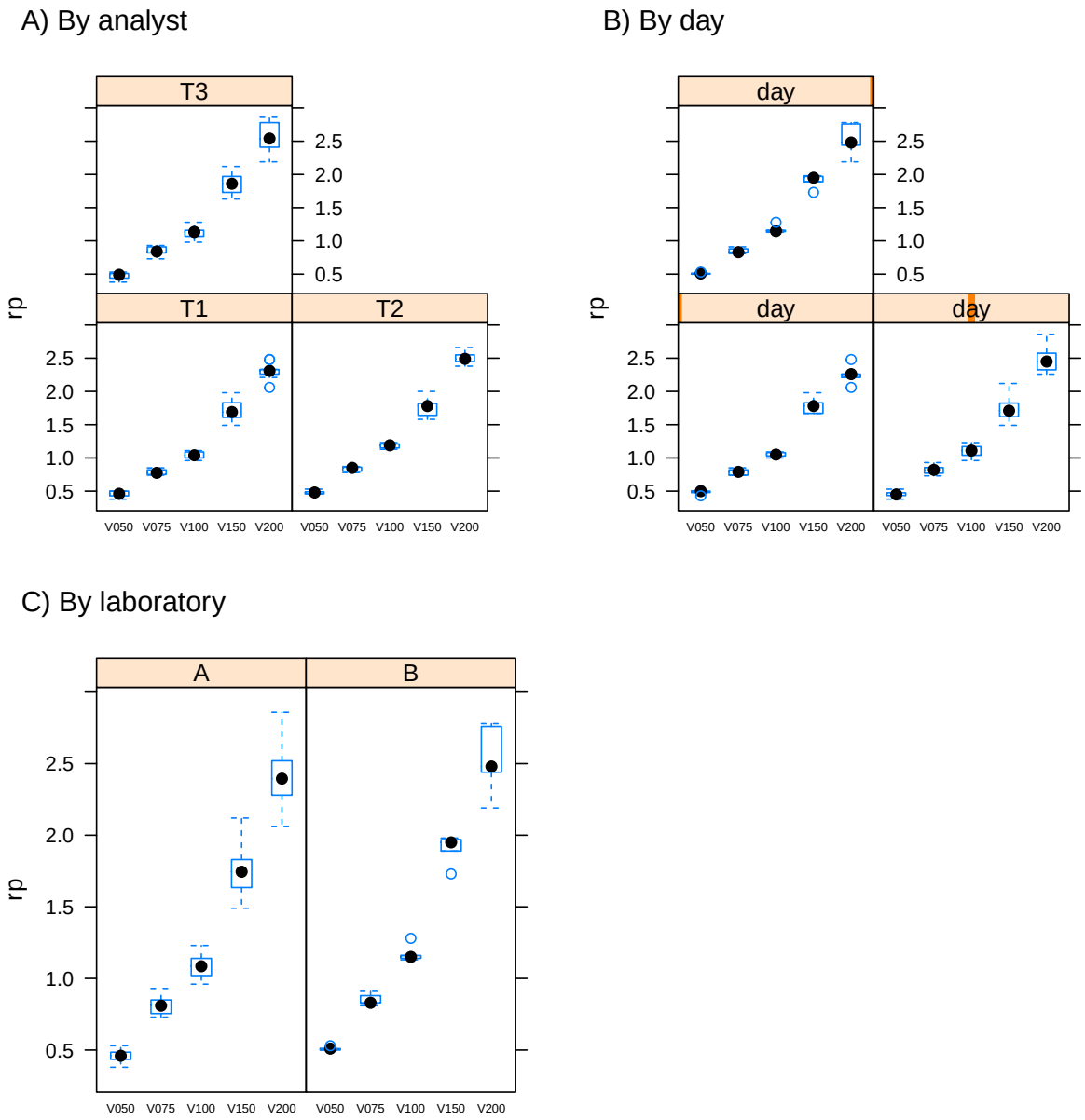

Figure 3.8: Boxplot of raw data by grouping factor analyst (A), day (B) or laboratory (C).

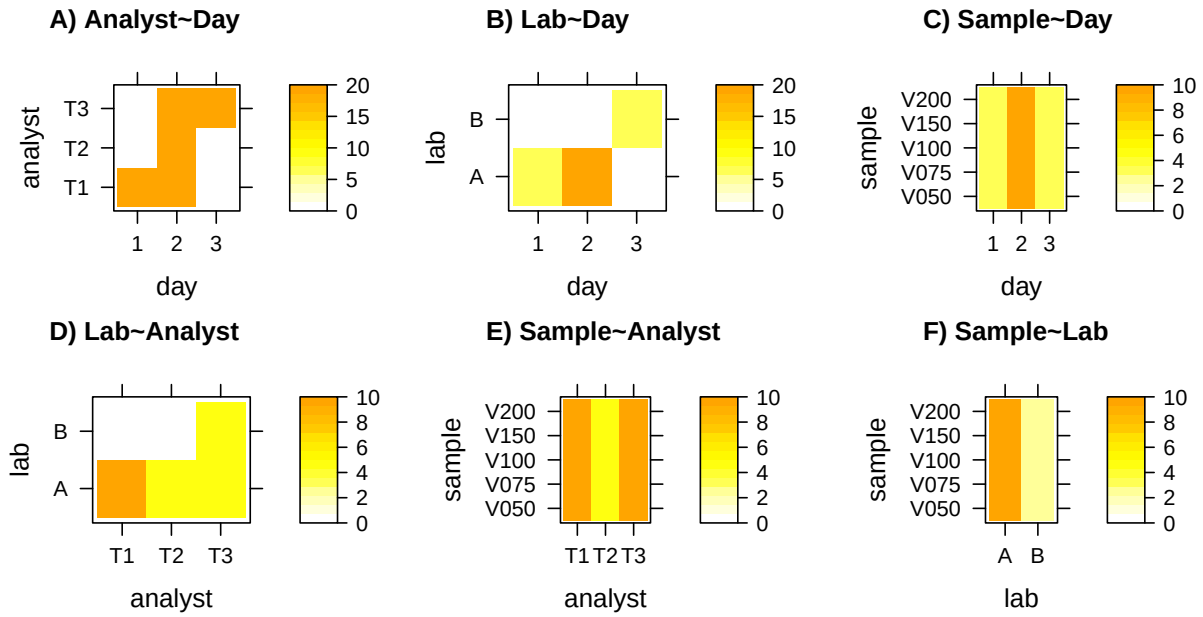

Figure 3.9: Levelplots displaying the number of observations at each variable combination.

of the *lab* variable. This variable is trickier than *day* or *analyst* factors as, contrary to them, it makes sense to model its effect as both fixed or random. Again, following the definition, if we were interested in drawing conclusions for the specific levels tested (e.g laboratory 1 and 2) a fixed effects approach should be used. On the other hand, if both laboratories are only a sample of all possible locations where the assay could be performed then a random effects approach is needed. Also, when a variable has a reduced number of levels it can make more sense to model it as a fixed effects factors (see Section 3.3). In this case however the *lab* variable will be initially modelled as a random effects factor.

The analysis starts by looking at the raw box plots of the replicates by grouping factor to get a general impression on data trends. As it can be observed from Figure 3.8, some variation should be expected in *rp* estimates by analyst, day or lab specially for the V150 and V200 presentations. Also, the box plots clearly show that replicates tend to have more scattered values around the median as the dose of antigen increases, regardless of the grouping factor considered. This is an early warning for hetereoscedasticity.

Again, level plots depicting available factor combination observations (Figure 3.9) will be used to determine the underlying relationship between data factors.

Rules to interpret the plots were clearly stated during study 1 analysis (Section 3.1.1) and shall not be repeated here but it can be seen that a similar situation arises. This situation is further complicated by the inclusion of the *lab* factor. In short, the design is irregular and highly unbalanced. Panel A shows the typical partially crossed relationship between *day* and *analyst* variables. The same relationship is observed for the *lab* and *analyst* relationship depicted in panel D. Panels C, E and F show the usual fully crossed structure pattern between its factors.

Interestingly in this case a nesting situation might be found. Panel B shows the underlying relationship between *lab* and *day* factors. As it can be seen, laboratory A only participates during the

firs two days of the study whereas laboratory B is restricted to day three. Following the nesting definition factor day is nested within factor laboratory as levels 1 and 2 of *day* factor happens exclusively at level A of *lab* and level 3 of *day* happens only at level B of *lab*.

3.2.2 Analysis

The modelling phase starts with the specification of the statistical model which will be used in the *lmer* function of the *lme4* package [8].

Statistical formulation of the fitted model is:

$$\begin{aligned} y_{adlsr} &= \beta_0 + A_{0a} + D_{0d} + L_{0l} + \beta_1 sample_1 + \beta_2 sample_2 + \\ &\quad \beta_3 sample_3 + \beta_4 sample_4 + \epsilon_{adlsr} \\ A_{0a} &\sim N(0, \sigma^2_{A_0}) \\ D_{0d} &\sim N(0, \sigma^2_{D_0}) \\ L_{0l} &\sim N(0, \sigma^2_{L_0}) \\ \epsilon_{adsr} &\sim N(0, \sigma^2) \end{aligned} \tag{3.2}$$

where r = 1,...,5 define the replicate, a = 1, 2 or 3 define the analyst, d = 1 or 2 define the assay day, l = 1 or 2 define the laboratory and s = 1, 2, 3, 4 or 5 define samples V050 through V200 respectively. Vector y is the responses vector. Terms β_0 and β_1 to β_3 are fixed intercept and slopes terms respectively, defining the fixed effects part of the model. The β terms, 1 through 4, represent the slopes with respect to the *n-1* dummy variables needed to represent the 5-level *sample* factor. Terms A_{0a}, D_{0d} and L_{0l} are the random slopes associated with the analyst, assay day and laboratory respectively defining the departure from the fixed intercept term depending on the value of each grouping factor. The within-group error term, ϵ_{adlsr}, describes the random variation associated to each observation and together with the random intercept terms they define the random effects part of the model. To completely specify the model, it is assumed that random effects including the error term follow a zero centered *(multivariate)*-normal distribution defined by their respective standard deviations.

The model is fitted with the following command, where it can be seen that a term defining a random intercept for the nested factor *day* within *lab* has been included as *1/lab/day* using the standard *lme4* notation. This notation expands to *1/lab + 1/lab:day* which denotes varying intercepts for *lab* factor and for *day* within *lab* factor.

```
data12.mod1 <- lmer(rp ~ sample + (1 | analyst) + (1 | lab/day),
    data = data12)
```

Model diagnostics should be conducted to confirm whether the residual part meets the assumptions made. As explained in the previous example, diagnostics for other random terms apart from the error are possible but not informative due to the few levels available. For this first attempt to model the *data12* dataset diagnostics are shown in Figure 3.10.

Although the relationship between the response and fixed effects might be linear, it can be observed that the homoscedasticity assumption does not hold as for larger fitted values the residual dispersion increases. Also, the residuals distribution has longer tails when compared to a normal distribution. Usually, a log transform of the response is a good option to re-conduct this situation. The base

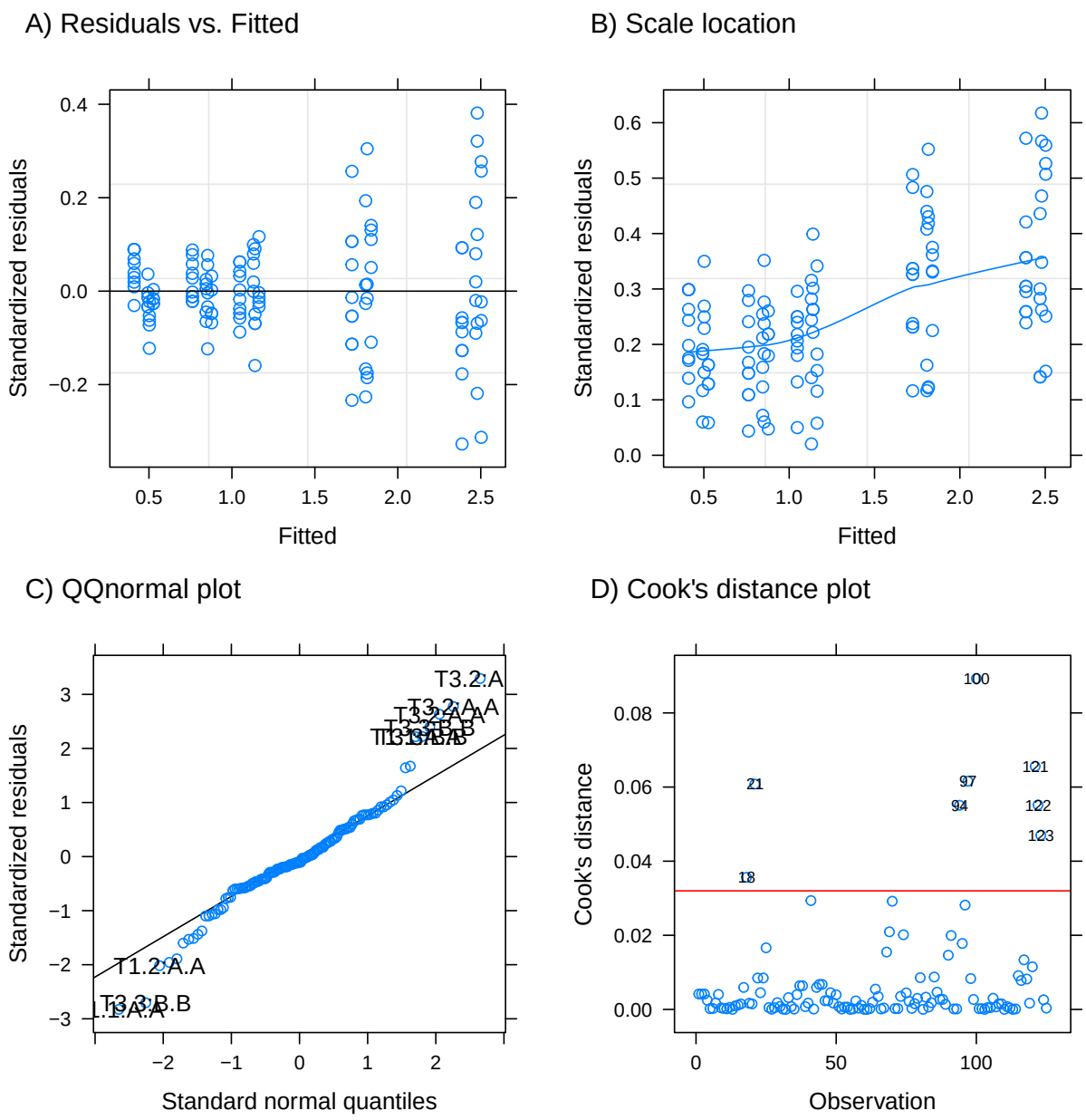

Figure 3.10: Diagnostic plots for data12.mod1 model.

of the logarithm is irrelevant if consistency is maintained during the analysis so in this case the same model is adjusted by applying a natural logarithm. The new model is coded as *data12.mod2*. Alternatively, other, more complex transformations such as the Box-Cox could be explored [27].

```
data12.mod2 <- lmer(log(rp) ~ sample + (1 | analyst) + (1 | lab/day),
    data = data12)
```

The new diagnostic plots for the logarithmic model are shown in Figure 3.11. As it can be seen, now the heteroscedasticity problem has been almost solved by the log transformation. Residuals QQplot now shows a better agreement with a normal distribution. In the Cook's distance plot several observations are detected as possibly influential. Several warnings related to having few observations for the relatively complex random effects structure arose during influence calculations; this situation will be latter discussed.

Regardless of the warnings, point 80 appears as the most influential based on the arbitrary cut-off value used. This particular observation has been studied as described in Section 3.1.2. As recorded data do not support a technical error as a main cause for this odd response value, the observation is considered an outlier. Model *data12.mod3* is fitted without this particular observation.

```
data12b <- data12[-c(80), ]
data12.mod3 <- lmer(log(rp) ~ sample + (1 | analyst) + (1 | lab/day),
    data = data12b)
```

Model diagnostics are not shown but again, a handful of warnings appear related to influence calculations. Moreover, heteroscedasticity and residuals normality do not improve dramatically so it is not justified to eliminate an observation.

Even though no warnings appear when fitting the model with the *lmer* routine with the full data, the recurrent warnings appearing when calculating influence must not be taken lightly. In this case, calculation algorithm sequentially adjusts models lacking one observation at a time to calculate its effect on the model fit; the error returned means that the influence displayed for some observations has been calculated based on low quality model fits and thus it is potentially misleading. Beyond this lecture, one might interpret that eliminating only one data point causes the fitting routine to struggle to estimate all model parameters and this situation is observed several times.

This exemplifies that an analytical dead end might have been reached. The usual cause is related with a too complex random effects structure given the amount of levels available for random effects factors and/or too few observation to support parameter estimation [11, 13]. If this situation happens only by dropping one observation, this should be considered a warning and alternative model specifications should be considered.

A possible alternative specification is to treat the *lab* factor as a fixed effects factor. This is possible because although one might not be interested in inferring on the particular laboratories used, the assay will only be performed at either lab A or B but at no other. Statistical formulation for this modified model can be expressed as:

$$\begin{aligned} y_{adlsr} &= \beta_0 + A_{0a} + D_{0d} + L_{0l} + \beta_1 sample_1 + \beta_2 sample_2 + \\ &\quad \beta_3 sample_3 + \beta_4 sample_4 + \beta_5 lab_1 + \epsilon_{adlsr} \\ A_{0a} &\sim N(0, \sigma^2_{A_0}) \\ D_{0d} &\sim N(0, \sigma^2_{D_0}) \\ \epsilon_{adlsr} &\sim N(0, \sigma^2) \end{aligned} \tag{3.3}$$

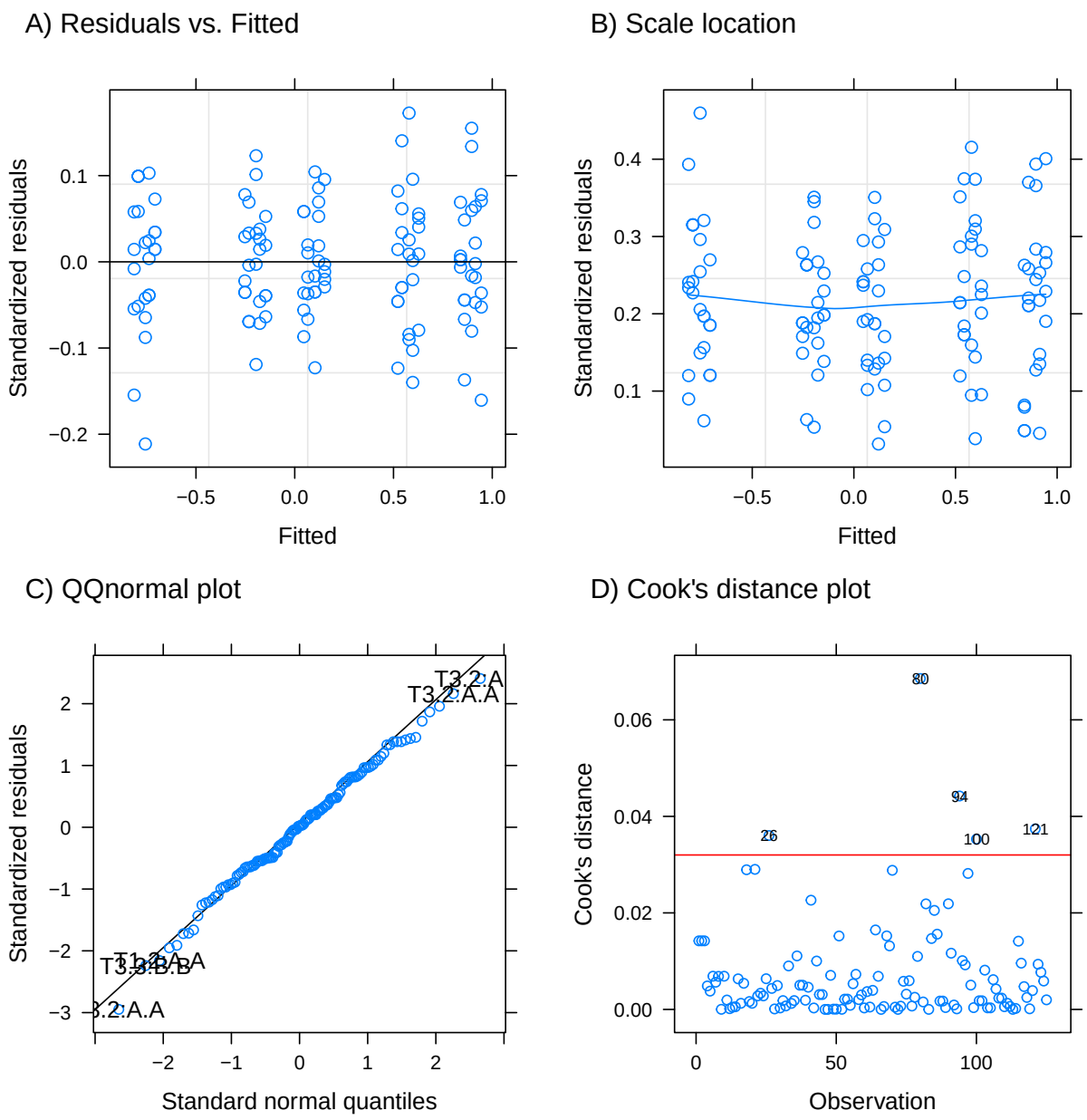

Figure 3.11: Diagnostic plots for data12.mod2 model.

where sub-indexes and parameters have the same meaning as before and the new β_5 represents the fixed slope with respect to the *n-1* dummy variables needed to code for the 2-level factor *lab*.

The model is fitted with the following command:

```
data12.mod4 <- lmer(log(rp) ~ sample + lab + (1 | analyst) +
    (1 | day), data = data12)
```

Diagnostic plots for this model are provided in Figure 3.12. It can be observed that homoscedasticity assumption is approximately satisfied together with normality of residuals. As discussed before, no observation appears to be sufficiently influential to proceed to its elimination. Finally, no warnings were issued neither in the *lmer* fit nor in the influence data calculations. Overall, simplification of the random effects structure has resulted in a more robust model.

Final model estimates for fixed effects are shown in Table 3.10 where it can be seen that some factor levels are not significant. This situation will be discussed in the next section.

Table 3.10: Fixed effects for data12.mod4 model after specifying lab as a fixed effects factor.

Effect	Estimate	Std.error	DF	t-value	p-value
(Intercept)	0.140	0.030	1.880	5.580	0.035
sample1	-0.880	0.010	114.890	-68.600	<0.001
sample2	-0.320	0.010	114.890	-24.960	<0.001
sample3	-0.020	0.010	114.890	-1.740	0.084
sample4	0.450	0.010	114.890	35.320	<0.001
lab1	-0.020	0.010	1.200	-1.840	0.284

Table 3.11 shows the variance component estimates. It should be noticed that *analyst* factor has a relatively high contribution to the overall variance.

Table 3.11: Variance component estimates for data12.mod4 model after specifying lab as a fixed effects factor.

Variance component	Variance	% Variance	Std. deviation
analyst	0.002	21.897	0.039
day	0.000	2.887	0.014
Residual	0.005	75.217	0.072

3.2.3 Inference

3.2.3.1 Tests on fixed effects

After obtaining a working model, inference can be made over parameter estimates. For fixed effects it can be seen in Table 3.10 that marginal Wald *t*-tests for *sample3* and *lab1* parameters are not significant at an $\alpha = 0.05$ value. Like for OLS regression, this *t*-tests are usually difficult to interpret by itself so for multi-level factors is better to apply an omnibus ANOVA test. As discussed in Section 2.3.5.1, type III sum-of-squares decomposition is used.

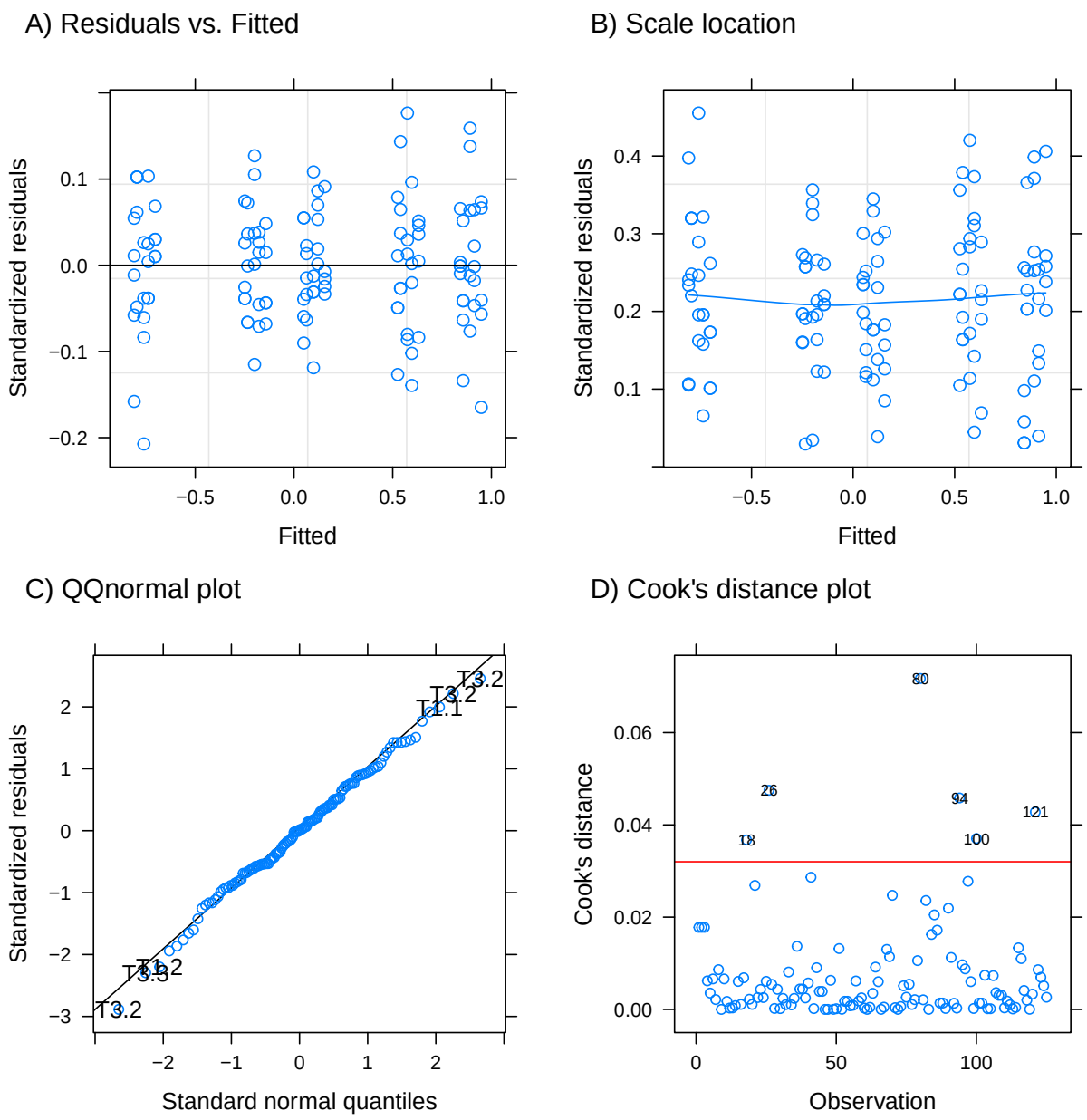

Figure 3.12: Diagnostic plots for data12.mod4 model.

```
## Type III Analysis of Variance Table with Satterthwaite's method
##        Sum Sq Mean Sq NumDF   DenDF   F value Pr(>F)
## sample 42.002 10.5006     4 114.889 2035.7433 <2e-16 ***
## lab     0.017  0.0174     1   1.204    3.3719 0.2838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result show that, as expected, *sample* factor has a significant impact on the response whereas *lab* factor is not significant. Thus, available data do not provide enough evidence to assume a dependence of the response on the specific laboratory where the assay is executed.

3.2.3.2 Tests on random effects

Significance for the random effects structure can be evaluated by conducting LRT on each individual random effect factor. In this case parametric bootstrap approach is not used to avoid computational overloading.

Tests for *day* and *analyst* factors:

Table 3.12: LRT for random effects of data12.mod4 object.

Effect removed	Num. parameters	logLik	AIC	LRT	DF	p-value
	9.000	130.054	-242.108	NA	NA	NA
(1 | analyst)	8.000	124.449	-232.898	11.210	1.000	<0.001
(1 | day)	8.000	129.925	-243.851	0.258	1.000	0.612

As in the example, these results are further confirmed by calculating 95 % parametric bootstrap confidence intervals on each parameter. Results are displayed in Table 3.13 below.

Table 3.13: 95 % parametric bootstrap confidence intervals for data12.mod4 parameter estimates.

Parameter	Estimate	Lower limit	Upper limit
.sig01	0.039	0.000	0.080
.sig02	0.014	0.000	0.046
.sigma	0.072	0.062	0.080
(Intercept)	0.144	0.094	0.199
sample1	-0.881	-0.906	-0.857
sample2	-0.321	-0.345	-0.297
sample3	-0.022	-0.049	0.004
sample4	0.454	0.428	0.479
lab1	-0.025	-0.056	0.004

For a better understanding of the results, variance estimates for *analyst* and *day* factors displayed in Table 3.11 indicate a moderate contribution of the first and a nearly negligible one for the latter. As explained in the introduction Section 2.3.6, LRT tend to be too conservative in the reported

p-values so this must be taken into account when interpreting the results. As expected by the small amount of variation it accounts for, the *day* factor is not significant and its p-value is large so, despite LRT issues, it is not likely an artefact. On the contrary, the *analyst* factor is a significant source of variation with a rather small p-value. Considering LRT are conservative, the returned value might be even smaller so the possibility of an artefact can be also ruled out. However it must be noted that the confidence intervals calculated in Table 3.13 include the zero for *analyst* (*.sig01*) and *day* (*.sig02*) variance estimates.

3.2.4 Validation

Section 3.1.4 contains a complete discussion on how to calculate and report common validation statistics such as the percent RB, which relates to assay accuracy, or several CV relating to different precision levels.

Here, the same exercise will be made for validation study 2. It should be noted that in this study a crucial analytical difference is found: response has been log-transformed during the analysis phase. As stated in Section 2.2 these changes the way precision and accuracy are calculated.

3.2.4.1 Accuracy

As before, accuracy is reported by calculating the percent RB following formulation provided by the USP in its chapter 1033 [62].

Relative potency expected values for samples V050 through V200 were 0.47, 0.82, 1.11, 1.78 and 2.45 respectively. LS group means are again obtained using the *lmerTest::lsmeansLT* function. Due to the natural logarithm transformation, the returned group means are in *log* units so they must be converted back to the original units using the the exponential transformation (*exp()* function in R) before calculating the relative biases. Final results are shown in Table 3.14.

Table 3.14: Model estimated and expected relative potency values for each sample.

Sample	Estimated	Expected
V050	0.48	0.47
V075	0.84	0.82
V100	1.13	1.11
V150	1.82	1.78
V200	2.5	2.45

Like in validation study 1, this bioassay has a relatively small bias and thus a high accuracy. Figure 3.13 shows a plot of the relative bias and its 95 % confidence region by expected potency level.

From the plot it can bee seen that this assay has a larger confidence region than the ones found in validation study 1 and this is most likely related to the logarithmic transformation and subsequent exponential back-transformation; in log scale, differences appear to be smaller. Nevertheless, a common practice is to set a % RB $\leqslant$ 15 % as the limit of acceptance and this bioassay conforms with this criteria. Also, it should be noticed that a positive bias is detected at each potency level

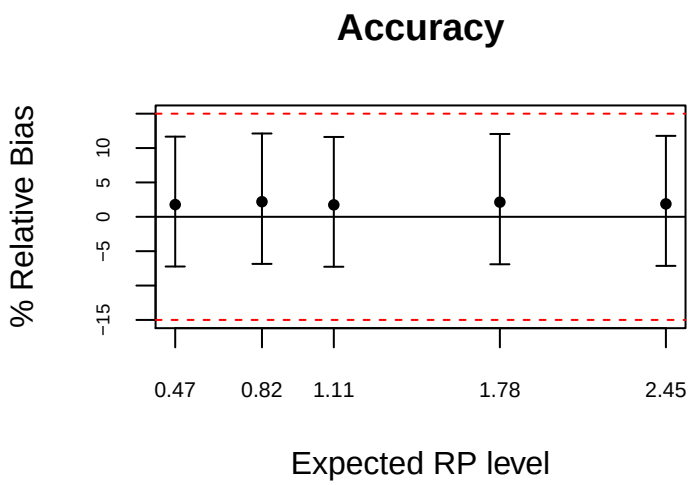

Figure 3.13: Relative bias plot for validation study 2.

so it can be stated that this biossay tends, on average, to overestimate the relative potency.

3.2.4.2 Precision

Again, the three levels of precision (repeatability, intermediate factor precision and reproducibility) will be estimated from validation study 2 model *data12.mod4*. Variance estimates reported in standard deviation units can be found in Table 3.15.

Table 3.15: Variance components for data12.mod4 object in standard deviations units.

Factor	Estimated Standard deviation
analyst	0.04
day	0.01
Residual	0.07

As a reminder of 2.2, variance estimates for each precision level can be obtained from model variance component estimates using the following formulation: residual variance (σ_r^2) stands for repeatability, residual variance plus factor associated variance ($\sigma_r^2 + \sigma_{factor}^2$) stands for factor intermediate precision and the overall variance ($\sigma_r^2 + \sigma_{day}^2 + \sigma_{analyst}^2$) stands for reproducibility.

As before, point estimates of each type of precision are calculated but in this case it is done according to the formulation provided by the USP chapter 1033 [62] for log-normal data (see Section 2.2). Thus, a percent GSD value for each estimate is finally reported.

Table 3.16: Precision for validation study 1 assay (object data12.mod4) as % CV.

Precision	% GSD
Repeatability	7.45
IP day	8.97
IP analyst	11.69
Reproducibility	13.27

From calculations reported in Table 3.16, it can be seen that both IP analyst and reproducibility values are somewhat high but in general the assay performance is good enough.

Confidence intervals for precision estimates can be calculated by bootstrap using a custom R function similar to the one used for validation study 1. The example below shows how to calculate the 95 % confidence interval for IP analyst. Again, due to computational requirements only this example is shown.

```
# Function to calculate the % GSD of a sample
calcgsd <- function(formula, data, indices) {
    dat <- data[indices, ]
    m <- lmer(formula, data = dat)
    # Changing row subsetting indices for VarCorr allows for
    # other % GSD to be calculated
    sigma <- sum(as.data.frame(VarCorr(m))[c(1, 3), 5])
    gsd <- round(100 * (exp(sigma) - 1), 2)
    return(gsd)
}

# Actual bootstrap function, R specifies the number of
# bootstrap samples.  The more samples the best the estimated
# interval. The more samples the higher the computational
# cost. Formula specifies the model to be fitted.
bod12m4 <- boot(data = data12, statistic = calcgsd, R = 200,
    formula = log(rp) ~ sample + lab + (1 | analyst) + (1 | day))

# Conficence intervals
bocid12m4 <- boot.ci(bod12m4, type = "bca")
bocid12m4
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 200 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bod12m4, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 9.72, 14.83 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

The result shows that IP analyst % GSD 95 % confidence region falls between 9.72 and 14.83 approximately. This can then be compared to a pre-specified value to accept or reject the performance of the assay, for example a 15 % GSD limit. In that case maybe it is worth to take some corrective action to improve analyst performance and repeat the validation protocol.

3.3 Comments on study designs

Both designs analysed in this work showed practically the same characteristics: 1) they are unbalanced and 2) they have few levels of grouping factors. Each of these characteristics causes different problems that will be briefly discussed below.

Lack of balance

Although balance may have several definitions, a common and simple one is to have the same number of observations for each combination of factors [58]. Obviously this is not the case for the designs analysed in the preceding sections and this causes a variety of situations during the analysis. Most notably, for unbalanced designs the standard mean-squares estimation techniques used in ANOVA to find the variance components can no longer be used and one must resort to ML or REML procedures as explained in the introductory Section 2.3.5. This leads to the inference problem that has already been discussed in Section 2.3.6 [36, 58].

Despite this, using ML or REML estimation procedures is in fact not a problem and instead, the ability of mixed models to cope with both balanced and unbalanced data is one of the reasons for its popularity in several scientific disciplines. However, as explained in Section 2.3.6, inference can be a problem when using ML or REML. As in several disciplines it is nearly a "mantra" to report statistical significance in the form of a p-value, this problem can not be easily by-passed. As stated before, this is an active research field and nowadays there is not yet a method where all statisticians agree so for a near future it is foreseeable that current approximations will still be used despite of its limitations.

Limited number of levels for groupping factors

Whereas lack of balance is "easy" to overcome by the estimation methods used to fit mixed models, these same methods struggle to estimate variance components when not enough levels of the grouping factors are present. When models are too complex given the available amount of data or there are too few number of levels of a grouping factor to get a credible estimate of its variance, fitting routines will return convergence problems or variance estimates will collapse to zero [14, 30]. Increase the amount of data or simplify the model, as when analysing validation study 2 in Section 3.2.2, can help solve the problem. Indeed, there is not a written rule on how many levels will provide good variance estimates but 5-6 levels by grouping factor seem to be a good rule of thumb [13]. Also, by increasing the number of levels it becomes possible to check for a correct specification of the random effects part by means of graphical techniques such as a QQplot of the BLUP [44].

Chapter 4

Part II: Analysis of parallelism studies

4.1 Parallelism study 1

4.1.1 Data structure

This parallelism study was performed in accordance with USDA guidelines describing the recommended experimental design to asses the conformance of an assay with respect to parallelism [53].

Two-fold serial dilutions of two different vaccine formulations, henceforth called *serials*, were assayed in five different plates in two separate days. Within a plate, each *serial* was tested in triplicates. The optical density (OD) of each well in a plate was measured as the response variable. Before proceeding with the assay, OD readings of the same *serial*/*dilution* combination were averaged and the blank substracted to account for the background OD.

The dataset contains a total of 400 observations, each corresponding to a particular blank-substracted OD reading, with no missing values and a total of 7 variables, where:

- *day*: A two level factor representing the two days when the analysis was performed.
- *plate*: A ten level factor representing the 10 different assay plates used during the study. Five plates were used each day.
- *plate2*: A five level factor representing the 10 different assay plates used during the study. Five plates were used each day, thus, plates within a day are labelled 1 to 5.
- *serial*: A two level factor identifying the two vaccines tested, the reference (*ref*) and the test (*test*) products.
- *dilution*: A continuous variable containing the dilution rate applied to the product.
- *logdilution*: A continuous variable containing the log_2 logarithm of the *dilution* variable.
- *od*: A continuous variable containing the raw OD readings of each well in a plate.
- *blank*: A continuous variable containing the background OD reading for a given plate.
- *odcorr*: A continuous variable containing the background substracted OD readings for each well in a plate.

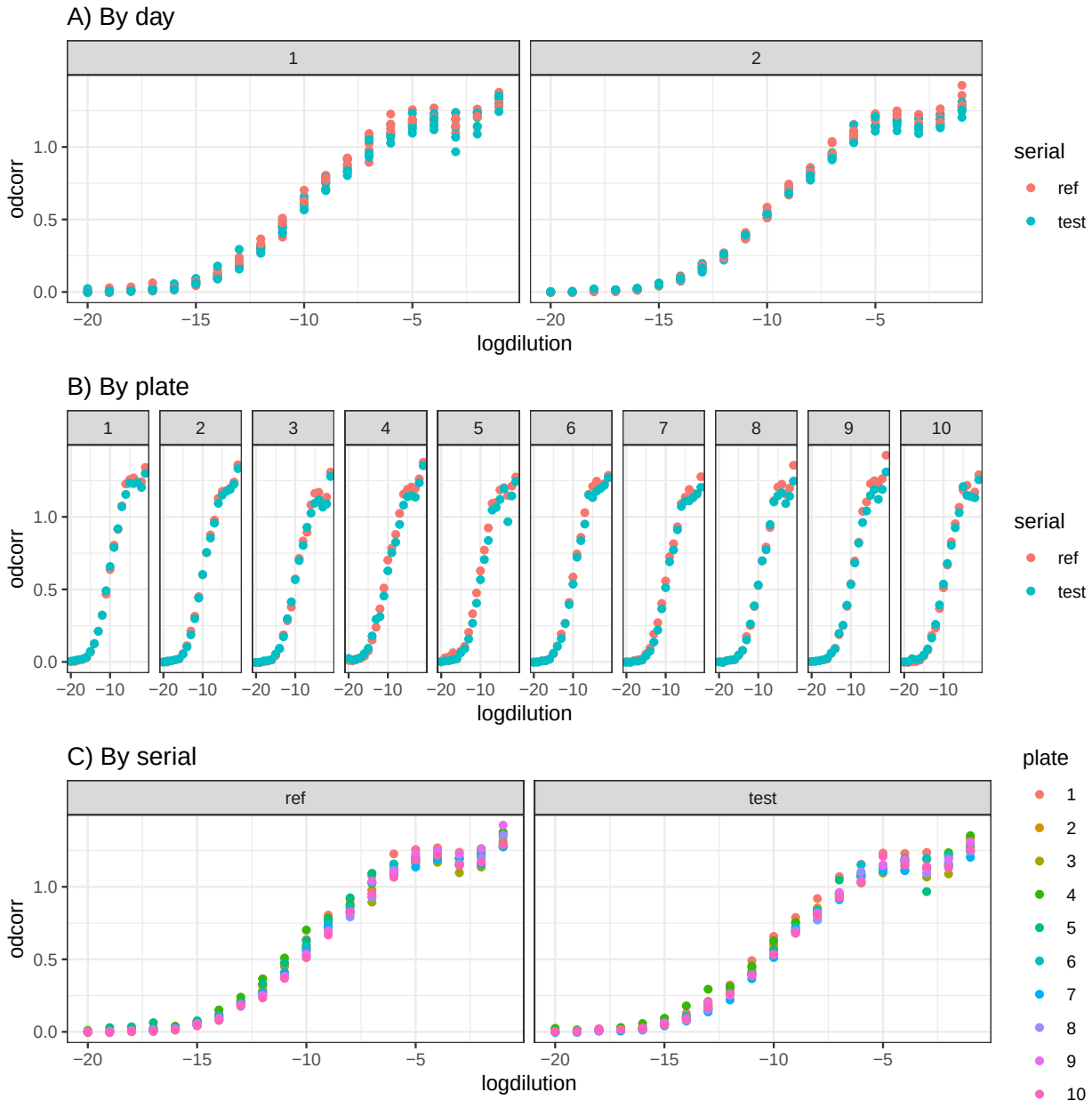

Figure 4.1: Scatter plots of raw data by grouping factor day (A) or plate (B) colored by serial type and by factor serial type colored by plate (C).

The first step before any formal analysis is to conduct a visual inspection of the raw data. It is customary to represent the response variable (*odcorr*) against the principal covariate (*logdilution*). As several grouping factors are available, they can be used to produce several informative plots depicted in Figure 4.1.

In this figure it can be observed that, regardless of the grouping factor used to plot the OD vs. dilution relationship, an evident sigmoidal shaped curve appears. Moreover, a possible heteroscedastic situation can be seen in any of the panels as it is clear that dispersion of response is greater as the value of *logdilution* increases.

Apart from the visual inspection, it is necessary to correctly determine the existing relationships between study factors to avoid incurring in erroneous model specifications. In this case factor *plate* is nested within factor *day* as 5 plates are run each day resulting in a total of 10 plates. This nesting structure can be made explicit if variable *plate* (10 levels) is selected or implicit if variable *plate2* (5 levels) is used instead. If *plate2* is selected, the nesting structure should be specified in the *groupedData* object as will be demonstrated below. Variable *serial* is crossed with *plate* and *day*.

4.1.2 Analysis

From data inspection it seems clear that *serial* and *plate* within *day* grouping factors can be considered potential nuisances that one might be interested in controlling for. Thus, a mixed effects model approach it is suitable.

In this case however, a linear mixed effects model is not reasonable as it is clear from data inspection that the underlying relationship is not linear. Thus, a non-linear mixed effects model shall be used.

Usually, the sigmoid shaped relationship obtained from these kind of assays is modelled with a 3 or 4 parameter logistic equation as shown in Section 2.3.4.2. As discussed, several possible parametrizations of this equation exist (see Section 2.3.4.2.1) but the one used in this work was proposed by Pinheiro and Bates (2002) [44] and constitutes the most natural choice for its use in conjunction with the *nlme* R package [43]. The Statistics Section of the CVB-USDA and both the Ph.Eur. and the USP contemplate these parametrization [51, 60]. Also and as outlined in previous plots, the principal covariate for this assay will be the log_2 of the dilution factor which is referred to as *concentration units* by the CVB-USDA Statistics Section [60]. Finally, the Statistics Section of the CVB-USDA also recommends the use of a blank corrected OD data as response variable such that a 3 parameter logistic function could be fitted [51].

Statistical formulation of the mixed effects 3 parameter logistic model is as follows:

$$y_{(ij)k} = \frac{A_i + u_{1j}}{1 + exp[(C_i + u_{2j} - log(x)_{(ij)k})/(B_i + u_{3j})]} + \epsilon_{(ij)k} \tag{4.1}$$

where i = 1,2 as *ref* and *test* serials respectively, j = 1,...,10 is the plate number and k is the dilution (or logarithm of the dilution). Each individual observation is represented by the y term. The three parameter logistic model have both a fixed and a random effects component for each of the model parameters. The upper asymptote is determined by the fixed effect A depending on the serial and the random effect u_1 depending on the plate. Likewise, scale factor and location parameter are specified by the fixed components B and C depending on the serial and the random effects components u_3 and u_2 depending on the plate, respectively. Finally, residual error is represented by the ϵ, one for each observation.

The model specification used above however can be misleading, so it should be further explained. To conduct the study in several days is a recommendation present in the regulatory documentation [53]. Despite this, the *day* factor is not explicitly used in the analysis when performed according to the CVB-USDA procedure [51]. The term *(ij)* appears to indicate that a nesting relationship exists between *serial* and *plate* and that the higher grouping level is the i indexed; the *serial* type. However, by taking a closer look to the formula it is clear that each parameter has a fixed effects part depending only on the *serial* and a random effects part depending on the *plate*. We feel that no nesting should be specified between *serial* and *plate* as this gives place to an erroneous theoretical grouping structure that is not reflected in the data. Also, it is clear from the design that *plate* is nested within *day* and, as time is a known and potentially important nuisance factor in every assay, we feel that it must be included in the model. Thus, a *plate* within *day* nesting relationship will be specified. Note that *plate2* factor should be used for this nesting to take effect.

To fully specify the model, random effects structure will be modelled as follows:

$$\epsilon_{(ij)k} \sim N(0, \sigma^2_\epsilon) \tag{4.2}$$

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_1} & 0 & 0 \\ 0 & \sigma^2_{u_2} & 0 \\ 0 & 0 & \sigma^2_{u_3} \end{bmatrix} \right) \tag{4.3}$$

This specification states that the errors will be modelled as having a normal distribution with mean zero and constant variance σ^2_ϵ and the vector of the three random effects will be modelled as having a multivariate normal distribution with a zero valued vector of means and a diagonal variance-covariance matrix meaning that the random effects are uncorrelated to each other. This last assumption is based on the grounds that not enough data is usually available to fit an unstructured model and computational instability may arise [51].

In previous sections (Chapter 3) models were fitted using the *lmer* function of the powerful *lme4* [8]. Despite this package has the *nlmer* function to fit non-linear mixed effects models, its implementation has been found to be complex and it lacks the possibility to easily add fixed effects covariates, which is required in this case [12]. Instead, the *nlmer* function of the exceptionally well documented and proven *nlme* package [43, 44] will be used as it allows the user to specify fixed effects covariates and also the random effects variance-covariance matrix structure.

First, the *nlme* function requires the user to construct a *groupedData* object which essentially is an R data frame object containing information about design structure. This is done with the following command:

```
data21nlme <- groupedData(odcorr ~ logdilution | day/plate2,
    data21)
```

As it can be seen, the *odcorr* is defined as the response depending on *logdilution* values with a *plate* within *day* nested grouping structure defined by the syntax *day/plate2*.

Next, as maximum likelihood algorithms to fit non-linear mixed effects models work iteratively, it is required to calculate a set of initial values for the fixed effects parameters. The closer the initial values provided to the optimizer to its optimum value, the better the chances of the algorithm to converge to the appropriate solution. This initial values, that depend on the chosen parametrization, can be obtained from the data by following simple rules. These rules and an R code example are provided below.

- **Asymptote**: Simply use the maximum OD value in the dataset.
- **Location**: By definition, the location parameter is equivalent to the value the covariate takes when the response is halfway its range. A good starting point is to use the covariate value, in this case the log_2 dilution, whose response value is approximately half of the previously calculated asymptote value.
- **Scale**: Scale factor determines the rate of change of the response with respect to the primary covariate. A good starting value can be calculated by taking the difference between covariate values (*logdilution*) whose response values (OD) are 75 % and 50 % respectively of the maximum value.

```
# Calculate appropriate starting values

## For asymptote
o1 <- max(data21$odcorr)

## For location
ino2 <- which((data21$odcorr > 0.45 * o1) & (data21$odcorr <
    0.55 * o1))
o2 <- mean(data21$logdilution[ino2])

## For scale factor
ino1 <- which((data21$odcorr > 0.73 * o1) & (data21$odcorr <
    0.77 * o1))
o3 <- mean(data21$logdilution[ino1]) - o2

# Wrap starting values into a list Start values for model
# without intercept
startvals <- list(fixed = c(Asym = rep(o1, 2), xmid = rep(o2,
    2), scal = rep(o3, 2)))
```

Once data has been ordered according to the software needs and the set of starting values has been calculated, a model can be fitted. As stated before, in this case there is a need to include a fixed effects covariate in the model apart from the *logdilution* factor, which is the primary covariate. This required covariate is the *serial* factor. As shown in Equation (4.1), each parameter has a fixed and a random effects component. The fixed effects component is dependent upon the serial type and this causes the need to include the *serial* factor in the *nlme* function call. The call below creates the *data21.mod1* object containing the model fit.

```
## nlme model without intercept for fixed covariate serial
data21.mod1 <- nlme(odcorr ~ SSlogis(logdilution, Asym, xmid,
    scal), data = data21nlme, fixed = Asym + xmid + scal ~ 0 +
    serial, random = pdDiag(Asym + xmid + scal ~ 1), start = startvals,
    method = "REML")
```

Examining the call, there are several parts that need to be described:

- *odcorr* ~ *SSlogis(logdilution, Asym, xmid, scal)*: This line specifies that *odcorr* is the response that will be modelled as dependent on the *logdilution* using an *SSlogis* function. The *SSlogis* function is a pre-coded function specifying the 3PL model depending on three parameters: *Asym* (asymptote), *xmid* (location) and *scal*(scale) parameters.

- *data = data21nlme*: This line specifies that the previously generated *groupedData* object is to be used instead of the original dataset.
- *fixed = Asym+xmid+scal~0+serial*: This line specifies the fixed effects portion of the model. It tells the software that 3PL parameter estimates are dependent on the value of the *serial* factor that acts as a covariate. See below for a detailed explanation on this topic.
- *random = pdDiag(Asym+xmid+scal~1)*: This line specifies the structure of the random part of the model. It tells the fitting function that all 3PL model parameters must have a random effects component and that this random effects part must have a diagonal variance-covariance matrix (*pdDiag*). The "~1" argument indicates that a single parameter is associated with this random effects which is the common way to specify random effects in *nlme* when needed.
- *start = startvals*: These lines provide the initial values for the 3PL parameters as calculated before and wrapped inside an R list object. Two values for each parameter must be provided as the *serial* covariate has two levels. Values can be the same unless asymptote, location and scale parameters are largely different between serials or covariate specification is changed (see below).
- *method = "REML"*: This line forces the optimizer to rely on restricted maximum likelihood instead of maximum likelihood to fit the model.

Before looking at the fitted model, the fixed effects part specification must be further explained. Dependency of model parameters from a covariate is usually modelled as a simple linear relationship. Taking the *Asym* parameter as an example, this parameter would be in fact modelled as:

$$A_i = \gamma_0 + \gamma_1 serial \tag{4.4}$$

where A_i is the asymptote parameter as specified in Equation (4.1). Coefficients γ_0 and γ_1 are the intercept and slope respectively for a simple linear formula relating the asymptote parameter with the serial type (*ref* or *test*).

At least two specifications arise depending on the value of γ_0. If $\gamma_0 \neq 0$, then it is clear that A_i has a basal value, γ_0, which is shifted upwards or downwards depending on the value of γ_1 and the type of serial. If treatment contrasts were used in R (see Section 2.3.5.1), then γ_0 would represent the asymptote value for the reference level of *serial* which could be *ref*. This value would be shifted according to γ_1 to obtain the asymptote value for the *test* serial. If this was the case, the list of two starting values passed to the *nlme* function should comprise: 1) the maximum OD reading for *ref* serial as an estimate of γ_0 and 2) the difference between maximum OD values of *ref* and *test* serials as an estimate of γ_1.

On the other hand, if $\gamma_0 = 0$, as it is the case in the model call used in this work, then A_i depends only on the value of γ_1 and the type of serial. In *nlme* call, the term "+0" forces the intercept to be zero. In OLS and other types of regression in R, when the intercept is removed from the model the coding of categorical factors changes. A 2-level factor like *serial* would usually be coded by *n-1* dummy variables as stated in Section 2.3.5.1 but, when no intercept is present it is automatically coded by two zero-one variables, one for each level of the factor in a coding scheme called *level-means*. Thus, the resulting model fit do not return the value of γ_1 as it would be expected but instead it returns two parameters representing the expected values of the dependent variable for each level of the categorical variable.

In this work, the covariate *serial* is fitted in the model with the expression 0+*serial* which effectively suppresses the intercept and changes the *serial* factor coding scheme to level-means. In this way, not only separate parameters are returned for each serial type but also the variance-covariance matrix between these parameters is directly obtained. The variance-covariance matrix is crucial to obtain confidence intervals for the calculations performed in Section 4.1.4.

The resulting model can be can be explored using the *summary* command:

```
## Nonlinear mixed-effects model fit by REML
##   Model: odcorr ~ SSlogis(logdilution, Asym, xmid, scal)
##  Data: data21nlme
##         AIC      BIC   logLik
##   -1454.063 -1402.37 740.0313
##
## Random effects:
##  Formula: list(Asym ~ 1, xmid ~ 1, scal ~ 1)
##  Level: day
##  Structure: Diagonal
##         Asym.(Intercept) xmid.(Intercept) scal.(Intercept)
## StdDev:     1.059995e-05        0.2115807     2.737852e-09
##
##  Formula: list(Asym ~ 1, xmid ~ 1, scal ~ 1)
##  Level: plate2 %in% day
##  Structure: Diagonal
##         Asym.(Intercept) xmid.(Intercept) scal.(Intercept)   Residual
## StdDev:       0.03186162         0.101251     6.814071e-10 0.03507052
##
## Fixed effects: Asym + xmid + scal ~ 0 + serial
##                     Value  Std.Error  DF   t-value p-value
## Asym.serialref   1.275643 0.01222590 385 104.33943       0
## Asym.serialtest  1.226268 0.01219306 385 100.57101       0
## xmid.serialref  -9.583066 0.15886760 385 -60.32109       0
## xmid.serialtest -9.616554 0.15925208 385 -60.38573       0
## scal.serialref   1.930179 0.03417979 385  56.47135       0
## scal.serialtest  1.922531 0.03537134 385  54.35280       0
##  Correlation:
##                 Asym.srlr Asym.srlt xmd.srlr xmd.srlt scl.srlr
## Asym.serialtest 0.682
## xmid.serialref  0.110     0.000
## xmid.serialtest 0.000     0.112     0.925
## scal.serialref  0.365     0.001     0.126    0.000
## scal.serialtest 0.001     0.362     0.000    0.128    0.001
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -5.7313560 -0.5482405 -0.1729319  0.3726201  3.8588987
##
## Number of Observations: 400
## Number of Groups:
##             day plate2 %in% day
##               2              10
```

It can be seen that this model output is far from intuitive but it is relatively similar to the output from *lme4::lmer* function seen in Section 3.1.2. The first part returns the model call and

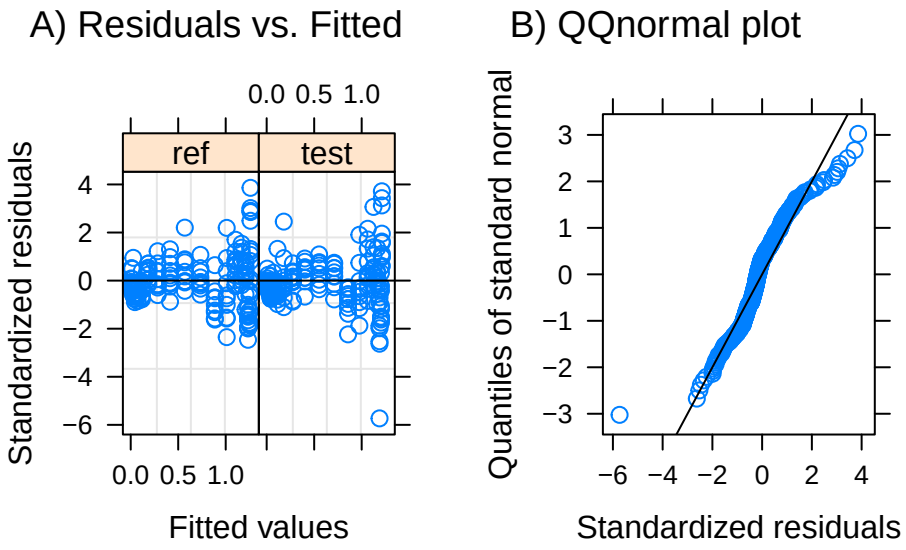

Figure 4.2: Diagnostic plots for the residuals of data21.mod1 object. A) residuals vs. fitted values, B) QQplot against a normal distribution.

characteristic statistics like AIC, BIC or log-likelihood.

Next the random effects variance estimations are presented ordered from the higher to the lower level. In this case it can be seen that the variance associated with *day* factor (in standard deviation units) is rather low, essentially zero, for the asymptote and scale parameters. In contrast to this, the *plate in day* grouping level show a bigger value for the standard deviation of the asymptote and it is comparable to that of the residual variance. This indicates that there exist some variation between plates in this parameter which has been captured in the model. Values for the location parameter are bigger but this is an expected outcome as the preparations used in this study differ in their antigenic content and thus a location shift is expected.

Following the random effects part, the summary returns the fixed effects estimates for the three parameters in the model. In this case, six lines appear and this is due to the fact that a different parameter is fitted for each serial type, as requested by the function call. Wald t-tests are used to analyse the marginal significance of each parameter. The correlation matrix between these parameters is also reported.

Finally, a brief residuals descriptive can be found followed by counts on the number of observations and groups used.

4.1.3 Model check and refit

As customary, before any in-depth analysis it must be checked if the resulting model fulfils the pre-specified assumptions for the residuals and random effects. The residuals specification is checked in Figure 4.2. It can be seen that while the QQplot is almost correct, with only some minor deviation in the right tail and an odd observation in the left one, the true problem lies in the heteroscedasticity observed in panel A. Clearly, residuals dispersion increases with increasing values of the response both for the *ref* and *test* serials.

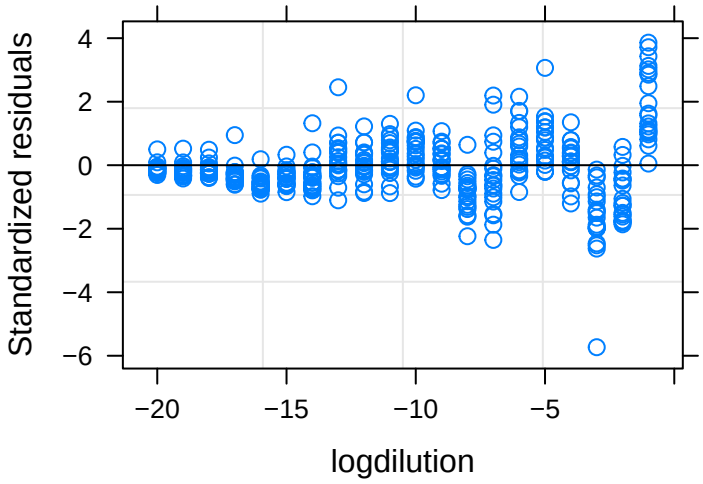

Figure 4.3: Plot of data21.mod1 model residuals against the logarithm of the dilution.

Based on the residuals analysis, it is not worth checking the random effects structure; instead, some corrective action should be considered. Failure to address problems with the distributional assumptions leads to incorrect statistic tests of model parameters and as a consequence, the conclusions drawn from them can also be wrong.

If this was an LME model, a common first option would be to transform the response variable. This is a rather complex approach as regulatory documentation clearly specifies what the authorities expect and it does not include any kind of transformation. In very specific circumstances, protocols may be modified under strong justification but generally any departure from the guidelines should be avoided.

A second option is to use the *nlme* package capabilities to model within-group error heteroscedasticity. Figure 4.3 clearly shows an increasing spread of residuals for higher (less negative) log_2 dilution values. Thus, the residual variance could be modelled as a function of this variable.

In *nlme* function calls, *wheigts* argument allows the specification of a particular model for the residual variance. Several residual variance models are readily available in the *nlme* package through *varFunc* class objects. The exact syntax and reasoning behind each variance model is out of the scope of this text but an excellent explanation can be found in Pinheiro and Bates (2002) [44]. The key point to model the residual variance is to choose the most appropriate variance model from the options provided. This can be done according to some informed knowledge about the underlying cause of the heteroscedasticity but it can also be accomplished through brute force. This strategy relies on the capacity to refit the model using several candidate residual variance model functions. The feasibility is influenced by model complexity and size of the dataset; the more complex the model, more computational power it will take.

R objects with suffix *.mod2* and *.mod3* were re-fitted using different residual variance functions based on the power family models in increasing complexity whereas in *.mod4* an exponential family model was used.

Three ANOVA (below) are then used to compare each model against the *.mod1* model using LRT.

Each comparison tests the significance of either adding a variance structure or changing between different variance structures depending on the pair of models compared.

```
##             Model df       AIC      BIC   logLik   Test L.Ratio p-value
## data21.mod1     1 13 -1454.063 -1402.37 740.0313
## data21.mod2     2 14 -1718.049 -1662.38 873.0243 1 vs 2 265.986  <.0001
```

```
##             Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## data21.mod1     1 13 -1454.063 -1402.370 740.0313
## data21.mod3     2 15 -1718.829 -1659.184 874.4147 1 vs 2 268.7669  <.0001
```

```
##             Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## data21.mod1     1 13 -1454.063 -1402.370 740.0313
## data21.mod4     2 14 -1704.983 -1649.314 866.4915 1 vs 2 252.9205  <.0001
```

It can be seen that when adding a defined variance structure improves the model, both AIC and BIC criterion tend to more negative values when compared to the first model and the log-likelihood increases, thus giving significant test results. However it is not possible to select a final model based only on p-values and a model selection criterion must be selected. The AIC value will be used here to select between those models significantly better than the first one. The one with the lower AIC should be selected but it should be noted that using this method can lead to over-complex models [30].

Based on this, *data21.mod3* is finally selected. A summary of the fit for this model is presented below where it can be seen a new section is added specifying the variance modelling structure. Also, parameter estimates have changed, most notably those of random effects.

```
## Nonlinear mixed-effects model fit by REML
##   Model: odcorr ~ SSlogis(logdilution, Asym, xmid, scal)
##  Data: data21nlme
##         AIC       BIC   logLik
##   -1718.829 -1659.184 874.4147
##
## Random effects:
##  Formula: list(Asym ~ 1, xmid ~ 1, scal ~ 1)
##  Level: day
##  Structure: Diagonal
##         Asym.(Intercept) xmid.(Intercept) scal.(Intercept)
## StdDev:     9.487025e-10        0.2049749     1.078595e-05
##
##  Formula: list(Asym ~ 1, xmid ~ 1, scal ~ 1)
##  Level: plate2 %in% day
##  Structure: Diagonal
##         Asym.(Intercept) xmid.(Intercept) scal.(Intercept)   Residual
## StdDev:         0.034592       0.07010869       0.07725416 0.04285794
##
## Variance function:
##  Structure: Constant plus power of variance covariate
##  Formula: ~fitted(.)
##  Parameter estimates:
##     const     power
```

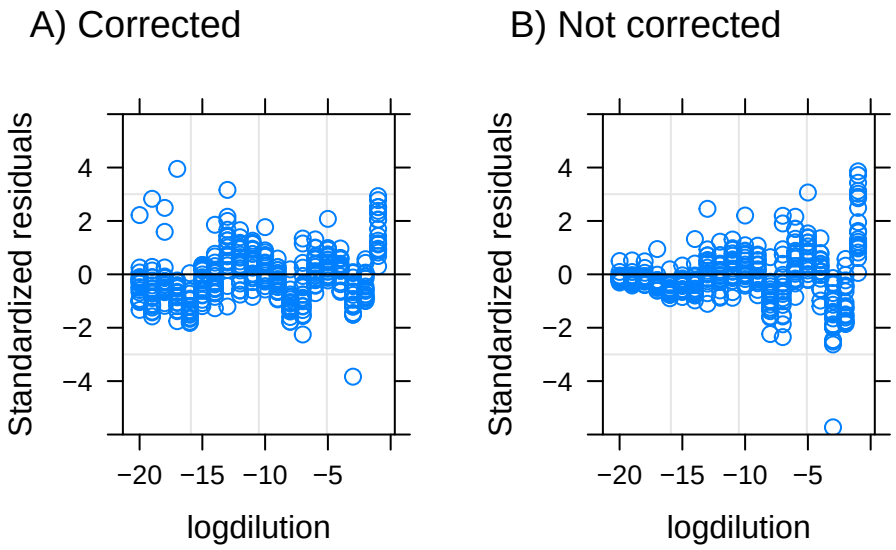

Figure 4.4: Comparison of the residuals against log-dilution plots from the data21.mod3 variance corrected model (A) and data21.mod1 non-corrected model (B).

```
## 0.1037736 0.5566015
## Fixed effects: Asym + xmid + scal ~ 0 + serial
##                     Value  Std.Error  DF   t-value p-value
## Asym.serialref   1.255150 0.01382200 385  90.80818       0
## Asym.serialtest  1.206712 0.01369179 385  88.13394       0
## xmid.serialref  -9.701999 0.15393952 385 -63.02474       0
## xmid.serialtest -9.733995 0.15413171 385 -63.15375       0
## scal.serialref   1.802525 0.03402466 385  52.97700       0
## scal.serialtest  1.793775 0.03433160 385  52.24852       0
##  Correlation:
##                 Asym.srlr Asym.srlt xmd.srlr xmd.srlt scl.srlr
## Asym.serialtest 0.633
## xmid.serialref  0.140     0.000
## xmid.serialtest 0.000     0.139     0.906
## scal.serialref  0.231     0.000     0.147    0.000
## scal.serialtest 0.000     0.228     0.000    0.150    0.511
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.8327801 -0.7597142 -0.2046779  0.4087140  3.9501681
##
## Number of Observations: 400
## Number of Groups:
##             day plate2 %in% day
##               2              10
```

A comparison between the residuals against *logdilution* plot for the initial model and the corrected model is shown in Figure 4.4. The effect of the correction is clearly visible in panel A.

Once the problems with the residual variance have been addressed it is recommended to check the structure of the random effects part of the model which was defined in Equation (4.3) to be

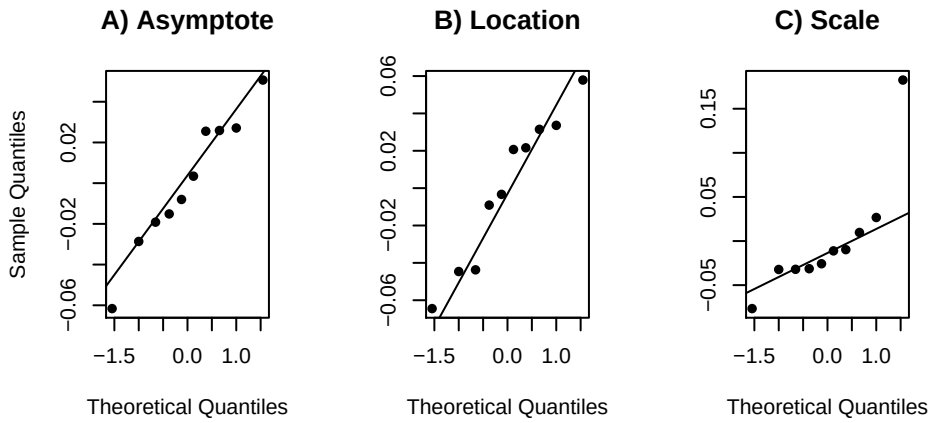

Figure 4.5: QQplots to check the random effects structure for each model parameter for the data21.mod3 model.

normally distributed. This is usually done using a QQplot of the estimated random effects for a given grouping level. The ability to confidently check the fulfilment of this assumption depends on the number of levels each grouping factor has. If the grouping factor has few levels it is virtually impossible construct a QQplot.

In this analysis, three equation parameters are estimated, all of them having a random effects part. Thus, for each grouping level three QQplots will be available. For the higher grouping level, *day*, the plots are not meaningful as only one value for each assay day is available; this is, there are too few levels. On the contrary, the QQplots for the lower level *plate in day* has enough levels to be plotted and analysed.

The resulting plots are shown in Figure 4.5. As it can be seen, a diagonal is evident in all panels indicating close-to-normal distributions. Only a possible outlier point for the *scal* parameter is clearly deviating from the overall tendency. Nevertheless, the model is assumed to be approximately correct.

4.1.4 Parallelism validation

The primary goal of this analysis is not to establish the significance of model parameters but to estimate fixed effects while controlling the sources of variation. Regulatory documentation clearly states the form the model should take and that no model selection should be conducted in the sense of adding, eliminating or interchange random and fixed effects [63, 51].

Once the selected model is demonstrated to be correct in light of its assumptions, it can be used to calculate parameter ratios to demonstrate parallelism. As described earlier in Section 2.2, two serials can only be compared in terms of relative potency using a full dose-response assay if the assay response of products differing in their antigen content is shown to yield "parallel" curves. Dose-response for two serials are considered parallel only if the asymptote and scale factor ratios lie within a pre-specified interval, usually 0.9-1.1 [53].

The conformance of the point estimates within this interval is evaluated using 90 % confidence intervals. As disused earlier (see Section 2.2) to calculate confidence regions for ratios is not trivial

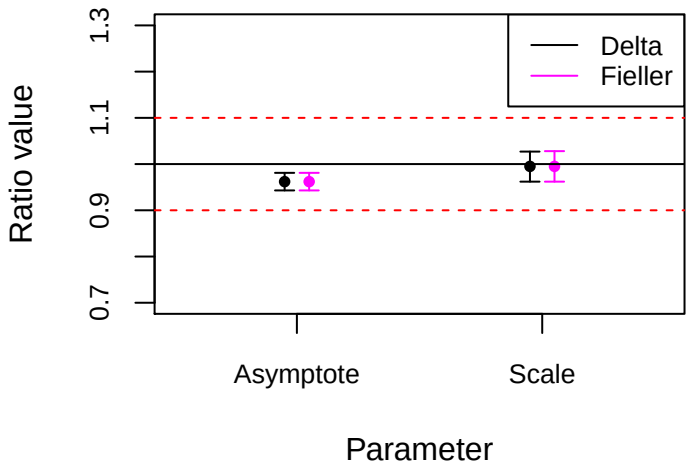

Figure 4.6: Asymptote and scale parameters ratios 90 % confidence intervals and the 0.9-1.1 acceptability region.

and two methods are used: the delta method, also called Taylor method, and the Fieller method [25, 63].

A custom function called *calcratios* has been written to calculate the ratios for models like the one analysed in this chapter. The code will be provided in the annex.

Table 4.1 provides a summary of the confidence interval regions calculated by both delta and Fieller methods and the ratio value for the *test/ref* parallelism check.

Table 4.1: 90 % confidence intervals for data21.mod4 model parameters as evaluated using the delta or Fieller methods.

Ratio	Value	Delta Lo.	Delta Up.	Fieller Lo.	Fieller Up.
Asymptote	0.962	0.943	0.981	0.943	0.981
Scale	0.995	0.962	1.027	0.962	1.028

As it is easier to visually evaluate parallelism compliance, results table is better reported as a plot (Figure 4.6). It can be observed that both the asymptote and scale ratios confidence intervals lie well within the acceptability region and thus, this assay has been validated to yield parallel dose-response curves under the tested experimental conditions.

In this specific case, as both *ref* serial asymptote and scale parameters are significantly different from zero (Section 4.1.3) both methods provide accurate enough intervals. In principle, Fieller method should be the default choice as it is more broadly accurate but the delta method provides a good alternative in situations analogous to the one presented here and is both mathematically and computationally easier to handle.

Chapter 5

Conclusions

Bioassay validation is complex due to the myriad of guidelines and requirements that apply. Each regulatory authority has its own preferences but, in the end, nearly always exists a common baseline between agencies that can be used to set a minimal common requirement. This approach has been used in this work due to the complexity it would represent to adapt to each of the requirements individually. Nevertheless, when no clear common regulation could be found or requirements were divergent, USDA requirements were prioritized for two reasons: 1) they are usually more restrictive and 2) they are also usually better well-founded from a statistical point of view.

Development of the statistical part of the project has been a though endeavour. LME and NLME models have demonstrated to be extremely powerful and versatile tools but they are not trivial to understand. Apart from the difficulty associated to their mathematical derivation, estimation methods are also complex and any of those two issues alone has enough entity to give place to several doctoral theses on their own right. Apart from the theory, software packages have a steep learning curve. Of the two packages used, *lme4*, has been found to be the easier to fit LME models. On the contrary, package *nlme* has been used to successfully fit NLME models with ease, including some special features like heteroscedasticity. In fact, flexibility is probably the best argument of the *nlme* package when compared against the more computationally efficient *lme4*. Despite the package used, both are easily implemented for most simple and common situations but, as the modelling needs become more demanding, syntax complexity exponentially increases.

Nevertheless, a comprehensive work has emerged that fulfils all the planned objectives. Moreover, several non-planned improvements are finally included in this thesis and as a result the quality is deemed superior to the initially expected. These improvements are, for example, several custom functions to generate reader friendly outputs from R functions or the implementation of the Fieller method to calculate confidence intervals for ratios during parallelism checks.

As outlined in Section 1.3, project execution has been subject to a three-phase method consisting of: 1) theory review, 2) software usage and 3) real data analysis. This strategy was captured in the task planning as shown in Section 1.4. The original task planning suffered a light modification that was reported in the Intermediate Report 1-PAC 2. This modification was due to two factors: not enough time was allocated for reviewing NLME models theory and new bibliography was available that suggested a possible improvement in the interpretation of validation studies. Both modifications demanded a replanification of the project execution planning for *PAC 3-Development phase II* period that was adequately implemented and reported. The impact on project execution, however, was deemed insignificant as enough unallocated time was left in anticipation of such issues. On

the contrary, the consequences of these modifications are tangible as both have yielded significant quality improvements over the original work. Overall, both the strategy and its implementation have allowed to complete the project and their results exceed the initially expected quality in several areas thus, planning has been adequate.

Finally, every project has some areas where further work would have been required for its completeness. Here, if time would not have been an issue, more designs could have been analysed with increasing difficulty. Also, regulatory requirements usually translate in different statistical interpretations. As shown during this text, several formulations exist for the same statistic and these formulations are obtained by making several assumptions. Apart from regulatory agencies, several academic authors have proposed possible improvements that could not be fully reviewed here. It would be, however, a very interesting work to compile and analyse the currently accepted methods and the proposed modifications to possibly achieve a common standard.

Chapter 6

Glossary

3PL	Three parameter logistic
4PL	Four parameter logistic
ANOVA	Analysis of Variance
APHIS	Animal and Plant Healch Inspection Services
BLUP	Best linear unbiased predictions
CV	Coefficient of variation
CVB	Center for Veterinary Biologics
EC50	Effective concentration 50 percent
ELISA	Enzyme-linked immunosorbent assay
EMA	European Medicines Agency
EMS	Expected Mean Squares
GMP	Good manufacturing practices
GSD	Geometric standard deviation
iid	Independent and identically distributed
LME	Linear mixed-effects
LRT	Likelihood-ratio test
LS	Least squares
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
NLME	Non-linear mixed-effects
OD	Optical density
OLS	Ordinary least squares
Ph.Eur.	European Pharmacopeia
QQplot	Quantile-quantile plot
RB	Relative bias
REML	Restricted maximum likelihood
RP	Relative potency
USDA	United States Department of Agriculture
USP	United States Pharmacopeia

Chapter 7

R code appendix

7.1 Custom functions for the *lme4* package

7.1.1 *tablefixef* function

This function extracts fixed effects estimates from a *lmerModLmerTest* object generated with the *lmer* function of the *lmerTest* package [34] and returns a pandoc style table.

```
tablefixef <- function(x, capt=""){
  fe <- cbind(row.names(coef(summary(x))), round(coef(summary(x)),2))
  df <- kable(fe,
              caption = capt,
              col.names = c("Effect","Estimate", "Std.error",
                            "DF", "t-value", "p-value*"),
              align=c("l","c","c","c","c","c"),"pandoc", row.names = FALSE)

  return(df)
}
```

7.1.2 *tablevarcomp* function

This function extracts fixed effects estimates from both *lmerMod* (*lme4* package) and *lmerModLmerTest* (*lmerTest* package) objects [8, 34] and returns a pandoc style table. It includes a new variable which represents the percent variance accounted for by each component.

```
tablevarcomp <- function(x, capt=""){
  vc <- as.data.frame(VarCorr(x),comp="Variance")
  sumvc <- sum(vc$vcov)
  vcdf <- data.frame("Group"=vc$grp, round(vc$vcov,3), round(vc$vcov/sumvc*100,2),
                     round(vc$sdcor, 3))
  df <- kable(vcdf,
              caption=capt,
              col.names=c("Variance component", "Variance",
                          "% Variance", "Std. deviation"),
              align=c("l","c","c","c"),"pandoc")
```

```
  return(df)
}
```

7.1.3 *tableanova* function

This function is a wrapper that returns the output of the *lmerTest::ranova* [34] function as a pandoc style table.

```
tableanova <- function(x, capt=""){

  ar1 <- ranova(x)
  ar <- cbind(row.names(ar1), round(ar1,4))
  df <- kable(ar,
              caption = capt,
              col.names=c("Effect removed", "Num. parameters",
                          "logLik", "AIC", "LRT", "DF", "p-value"),
              align=c("l","c","c","c","c","c","c"),"pandoc", row.names = FALSE)

  return(df)
}
```

7.1.4 *tableconfint* function

This function is a wrapper that returns the output of the *lme4::confint* [8] function as a pandoc style table.

```
tableconfint <- function(x, capt=""){
  ci1 <- confint(x, method="boot")
  ci <- cbind(row.names(ci1), round(ci1,2))
  df <- kable(ci,
              caption = capt,
              col.names=c("Parameter", "Lower bound - 2.5 %", "Upper bound - 97.5 %"),
              align=c("l","c","c"),"pandoc", row.names = FALSE)

  return(df)
}
```

7.1.5 *diagplots* function

This function makes use of different plotting methods for *lmerMod* objects to return a 2 x 2 plot layout of diagnostic plots for LME models. This function is dependent of *influence.ME* [40] and *lattice* [47] packages.

```
#Generate lmer object diagnostic plots
diagplots <- function(x){

  inf1 <- influence(x, obs=TRUE)
  cd1 <- data.frame(cd=cooks.distance(inf1), index=seq(1,dim(x@frame)[1],1))
```

```
  lim <- 4/dim(cd1)[1]
  sel <- which(cd1$cd>=lim)
  lab <- rep(" ", length(cd1$index))
  lab[sel] <- sel

  p1 <- plot(x, xlab="Fitted", ylab="Standardized residuals",
             main="A) Residuals vs. Fitted", par.settings=my.settings)

  p2 <- plot(x,sqrt(abs(resid(., type="pearson")))~fitted(.),type=c("p","smooth"),
           xlab="Fitted", ylab="Standardized residuals",
           main="B) Scale location", par.settings=my.settings )

  p3 <- qqmath(x,id=0.05, main="C) QQnormal plot", par.settings=my.settings)

  p4 <- xyplot(cd~index, data=cd1, xlab="Observation", ylab="Cook's distance",
               main="D) Cook's distance plot",
               par.settings=my.settings, panel=function(...) {
               panel.xyplot(...)
               panel.abline(h=lim, col="red")
               panel.text(cd1$index+0.1,cd1$cd,labels=lab,cex=0.7)
               })

  return(grid.arrange(p1,p2,p3,p4, ncol=2))
}
```

7.2 Custom functions for the *nlme* package

7.2.1 *calcratios* function

This function is specifically designed for models like the one fitted in Section 4 that makes use of the *SSlogis* function of the *nlme* package [43]. It extracts fixed effects estimates and their variance-covariance matrix and returns point estimates for the *Asym* and *scal* parameter ratios between a test and a reference serial. Also, it computes confidence intervals for this ratios using the delta and Fieller methods.

```
calcratios <- function(mod, plates, signif = 0.1) {
    # Extract fixed effects
    fefs <- fixef(mod)
    # Extract each value
    a.ref <- fefs[1]
    a.test <- fefs[2]
    s.ref <- fefs[5]
    s.test <- fefs[6]
    # Calculate ratios
    a.rat <- a.test/a.ref
    s.rat <- s.test/s.ref

    # Extract var-cov matrix of fixed effects Extract vars and
    # covars
    vcmat <- vcov(mod)
```

```
    a.ref.var <- vcmat[1, 1]
    a.test.var <- vcmat[2, 2]
    a.rt.cov <- vcmat[1, 2]

    s.ref.var <- vcmat[5, 5]
    s.test.var <- vcmat[6, 6]
    s.rt.cov <- vcmat[5, 6]

    # Calculate DF according to CVB-USDA Calculate t-distr.
    # critical value
    dfs <- plates - 3
    t <- qt(1 - signif/2, df = dfs)

    # Delta method
    a.d.ci <- t * abs(a.rat) * sqrt((a.ref.var/a.ref^2) + (a.test.var/a.test^2) -
        (2 * a.rt.cov/(a.test * a.ref)))

    s.d.ci <- t * abs(s.rat) * sqrt((s.ref.var/s.ref^2) + (s.test.var/s.test^2) -
        (2 * s.rt.cov/(s.test * s.ref)))


    # Fieller method
    a.f.ciu <- (1/(a.ref^2 - (t^2 * a.ref.var))) * (((a.ref *
        a.test) - (t^2 * a.rt.cov)) + sqrt((((a.ref * a.test) -
        (t^2 * a.rt.cov))^2) - (((a.ref^2) - (t^2 * a.ref.var)) *
        ((a.test^2) - (t^2 * a.test.var)))))

    a.f.cil <- (1/(a.ref^2 - (t^2 * a.ref.var))) * (((a.ref *
        a.test) - (t^2 * a.rt.cov)) - sqrt((((a.ref * a.test) -
        (t^2 * a.rt.cov))^2) - (((a.ref^2) - (t^2 * a.ref.var)) *
        ((a.test^2) - (t^2 * a.test.var)))))

    s.f.ciu <- (1/(s.ref^2 - (t^2 * s.ref.var))) * (((s.ref *
        s.test) - (t^2 * s.rt.cov)) + sqrt((((s.ref * s.test) -
        (t^2 * s.rt.cov))^2) - (((s.ref^2) - (t^2 * s.ref.var)) *
        ((s.test^2) - (t^2 * s.test.var)))))

    s.f.cil <- (1/(s.ref^2 - (t^2 * s.ref.var))) * (((s.ref *
        s.test) - (t^2 * s.rt.cov)) - sqrt((((s.ref * s.test) -
        (t^2 * s.rt.cov))^2) - (((s.ref^2) - (t^2 * s.ref.var)) *
        ((s.test^2) - (t^2 * s.test.var)))))


    ratsdf <- data.frame(ratio = c("Asymptote", "Scale"), value = round(c(a.rat,
        s.rat), 3), delta.lo = round(c(a.rat - a.d.ci, s.rat -
        s.d.ci), 3), delta.up = round(c(a.rat + a.d.ci, s.rat +
        s.d.ci), 3), fieller.lo = round(c(a.f.cil, s.f.cil),
        3), fieller.up = round(c(a.f.ciu, s.f.ciu), 3))

    return(ratsdf)
}
```

Bibliography

[1] U. Andreasson, A. Perret-Liaudet, L. J. C. van Waalwijk van Doorn, K. Blennow, D. Chiasserini, S. Engelborghs, T. Fladby, S. Genc, N. Kruse, H. B. Kuiperij, L. Kulic, P. Lewczuk, B. Mollenhauer, B. Mroczko, L. Parnetti, E. Vanmechelen, M. M. Verbeek, B. Winblad, H. Zetterberg, M. Koel-Simmelink, and C. E. Teunissen. A Practical Guide to Immunoassay Method Validation. *Frontiers in Neurology*, 6(176), Aug. 2015.

[2] S. Aydin. A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA. *Peptides*, 72:4–15, Oct. 2015.

[3] R. Baayen, D. Davidson, and D. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, Nov. 2008.

[4] D. Bates. Lmer, p-values and all that. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html, May 2006.

[5] D. Bates. Assessing the precision of estimates of variance components, July 2009.

[6] D. Bates. *Lme4: Mixed-Effects Modeling with R*. Springer, pre-print, june 25 2010 edition, June 2010.

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

[8] D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*, 2018. R package version 1.1-18-1.

[9] B. M. Bolker. Error with RLRsim package while testing for variance components. https://stats.stackexchange.com/q/63716, Sept. 13.

[10] B. M. Bolker. Standard Error of variance component from the output of lmer. https://stackoverflow.com/a/31704646, July 15.

[11] B. M. Bolker. Warning message glmer. https://stackoverflow.com/a/42729121, Oct. 17.

[12] B. M. Bolker. Algal (non)-linear mixed model example. http://rpubs.com/bbolker/3423, Jan. 2013.

[13] B. M. Bolker. GLMM FAQ. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#why-doesnt-lme4-display-denominator-degrees-of-freedomp-values-what-other-options-do-i-have, Nov. 2018.

[14] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135, Mar. 2009.

[15] A. Canty and B. Ripley. *boot: Bootstrap Functions (Originally by Angelo Canty for S)*, 2017. R package version 1.3-20.

[16] CHMP. Guideline on bioanalytical method validation., July 2011.

[17] C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, Feb. 2004.

[18] CVB. CVBSOP0102.01. Estimating Relative Potency, Mar. 14.

[19] E. Demidenko and T. A. Stukel. Influence analysis for linear mixed-effects models. *Statistics in Medicine*, 24(6):893–909, Mar. 2005.

[20] B. DeSilva, W. Smith, R. Weiner, M. Kelley, J. Smolec, B. Lee, M. Khan, R. Tacey, H. Hill, and A. Celniker. Recommendations for the Bioanalytical Method Validation of Ligand-Binding Assays to Support Pharmacokinetic Assessments of Macromolecules. *Pharmaceutical Research*, 20(11):1885–1900, Nov. 2003.

[21] J. J. Faraway. *Linear Models with R.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2004.

[22] J. J. Faraway. *Extending the Linear Model with R.* Texts in Statistical Science. CRC Press, Boca Raton, first ed. edition, 2006.

[23] C. B. Fox, R. M. Kramer, L. Barnes V, Q. M. Dowling, and T. S. Vedvick. Working together: Interactions between vaccine antigens and adjuvants. *Therapeutic Advances in Vaccines*, 1(1):7–20, May 2013.

[24] J. Fox, S. Weisberg, and B. Price. *car: Companion to Applied Regression*, 2018. R package version 3.0-2.

[25] V. H. Franz. Ratios: A short guide to confidence limits and proper use, Oct. 2007.

[26] S. Greven, C. M. Crainiceanu, H. Küchenhoff, and A. Peters. Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 17(4):870–891, 2008.

[27] M. J. Gurka, L. J. Edwards, K. E. Muller, and L. L. Kupper. Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):273–288, Mar. 2006.

[28] J. Hadfield. *MCMCglmm: MCMC Generalised Linear Mixed Models*, 2018. R package version 2.26.

[29] U. Halekoh and S. Højsgaard. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The *R* Package **pbkrtest**. *Journal of Statistical Software*, 59(9), 2014.

[30] X. A. Harrison, L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6:e4794, May 2018.

[31] A. Hector, S. von Felten, and B. Schmid. Analysis of variance with unbalanced data: An update for ecology & evolution. *J Anim Ecol*, 79(2):308–316, Mar. 2010.

[32] H. J. Keselman, J. Algina, and R. K. Kowalchuk. The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, 54(1):1–20, May 2001.

[33] C. Krueger. A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research For Nursing*, 6(2):151–157, Oct. 2004.

[34] A. Kuznetsova, P. Bruun Brockhoff, and R. Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*, 2018. R package version 3.0-1.

[35] O. Langsrud. ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and computing*, (13):163–167, 2003.

[36] R. Littel. Analysis of Unbalanced Mixed Model Data: A Case Study Comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(4):472–490, 2002.

[37] S. Lo and S. Andrews. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6(1171), Aug. 2015.

[38] J. D. Long. *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Sage Publications, Inc, Thousand Oaks, CA, first edition, Oct. 2011.

[39] S. G. Luke. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4):1494–1502, Aug. 2017.

[40] R. Nieuwenhuis, B. Pelzer, and M. te Grotenhuis. *influence.ME: Tools for Detecting Influential Data in Mixed Effects Models*, 2017. R package version 0.9-9.

[41] G. W. Oehlert. *A First Course in Design and Analysis of Experiments*. W.H. Freeman, New York, first edition, 2000.

[42] M. E. Payton. Confidence intervals for the coefficient of variation. *Conference on Applied Statistics in Agriculture*, Apr. 1996.

[43] J. Pinheiro, D. Bates, and R-core. *nlme: Linear and Nonlinear Mixed Effects Models*, 2018. R package version 3.1-137.

[44] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and computing. Springer, New York, third edition, 2002.

[45] A. Pryseley, K. Mintiens, K. Knapen, Y. Van der Stede, and G. Molenberghs. Estimating precision, repeatability, and reproducibility from Gaussian and non- Gaussian data: A mixed models approach. *Journal of Applied Statistics*, 37(10):1729–1747, Oct. 2010.

[46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[47] D. Sarkar. *lattice: Trellis Graphics for R*, 2017. R package version 0.20-35.

[48] SAS. SAS 9.2 User's Guide. https://support.sas.com/en/documentation.html, Apr. 2010.

[49] F. Scheipl and B. Bolker. *RLRsim: Exact (Restricted) Likelihood Ratio Tests for Mixed and Additive Models*, 2016. R package version 3.1-3.

[50] H. Schielzeth and S. Nakagawa. Nested by design: Model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, 4(1):14–24, Jan. 2013.

[51] S. Section. A method for analyzing ELISA parallelism data, Mar. 2015.

[52] V. Services. Veterinary Services Memorandum NO. 800.90. Guidelines for Veterinary Biological Relative Potency Assays and Reference Preparations Based on ELISA Antigen Quantification, Aug. 1998.

[53] V. Services. Veterinary Services Memorandum NO. 800.112. Guidelines for Validation of In Vitro Potency Assays, 2015.

[54] M. Sherman, A. Maity, and S. Wang. Inferences for the ratio: Fieller's interval, log ratio, and large sample based confidence intervals. *AStA Advances in Statistical Analysis*, 95(3):313–323, Sept. 2011.

[55] H. Singmann, B. Bolker, J. Westfall, and F. Aust. *afex: Analysis of Factorial Experiments*, 2018. R package version 0.22-1.

[56] H. Singmann and D. Kellen. *An Introduction to Mixed Models for Experimental Psychology in New Methods in Neuroscience and Cognitive Psychology:*. New Methods in Neuroscience and Cognitive Psychology. Psychology Press, London, pre-print edition, 2017.

[57] U. H. S. H. <sorenh@math.aau.dk>. *pbkrtest: Parametric Bootstrap and Kenward Roger Based Methods for Mixed Model Comparison*, 2017. R package version 0.4-7.

[58] J. Spilke, H. P. Piepho, and X. Hu. Analysis of Unbalanced Data by Mixed Linear Models Using the mixed Procedure of the SAS System. *Journal of Agronomy and Crop Science*, 191(1):47–54, Feb. 2005.

[59] R. Strugnell, F. Zepp, A. Cunningham, and T. Tantawichien. Vaccine antigens. *Perspectives in Vaccinology*, 1(1):61–88, Aug. 2011.

[60] C. Tong and M. Vendettuoli. Parameterizations of the four-parameter logistic curve in software for estimating relative potency, 2017.

[61] USPC. Chapter <1032> Design and development of biological assays. In *United States Pharmacopeia (USP): The National Formulary*. US Pharmacopeial convention, Inc, Rockville, MD, 34th edition, 2010.

[62] USPC. Chapter <1033> Biological assay validation. In *United States Pharmacopeia (USP): The National Formulary*. US Pharmacopeial convention, Inc, Rockville, MD, 34th edition, 2010.

[63] USPC. Chapter <1034> Analysis of biological assays. In *United States Pharmacopeia (USP): The National Formulary*, page 16. US Pharmacopeial convention, Inc, Rockville, MD, 34th edition, 2010.

[64] J. H. Ware and N. M. Laird. Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):963–974, 1982.

[65] B. Yu and H. Yang. Evaluation of Different Estimation Methods for Accuracy and Precision in Biological Assay Validation. *PDA Journal of Pharmaceutical Science and Technology*, 71(4):297–305, 2017.