Commutability of control materials

Statistical methods of evaluation

Pernille Kjeilen Fauskanger

Department of mathematics
University of Bergen

This thesis is submitted for the degree of
Master's of Science in Statistics

November 2020

I would like to dedicate this thesis to my loving family and friends. Thank you for everything you have done for me in writing process …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices,bibliography, footnotes, tables and equations and has fewer than 150 figures.

Pernille Kjeilen Fauskanger
November 2020

Acknowledgements

Finally, I would like to give my sincere gratitude towards my supervisors Bård Støve and Thomas Røraas for their support, patience and inputs during the writing process. Thank you for being so available and motivating in the writing process. Also, a special thanks to Noklus for their interest and assistance in the writing process.

Abstract

Methods of commutability assessment is an ongoing discussion in the field of laboratory science. Commutability of control materials is roughly defined as control materials having the same numerical behaviour as patient samples. Opinions regarding which methods being suitable are polarized. Lack of caution concerning the linear model assumptions is often the case. Consequently, many of the currently used methods show deceitful from a statistical aspect. Transformations are frequently used as a go-to solution when the linear model assumptions are unsatisfied. Unfortunately, 'healing' model assumptions using transformations typically result in the opposite outcome for most instances, reducing the acceptance rates for the linear model assumptions rather than increasing them. Therefore, we would like to implement new methods that are the most independent of strict model assumptions. That way, we would have a more general procedure for commutability assessment. However, too few assumptions are not favorable because of the complexity of the resulting models. The art of balancing model assumptions and model complexity concerning their application in commutability assessment is crucial. A perfect balance for every situation does not exist, so a statistical approved decision algorithm or protocol for *external quality assessment* and *internal quality assessment* is vital. In some cases it is also of interest to apply commutability evaluation for certified reference materials. However, this will not be elaborated upon in this text. We find that commutability assessment procedures used today (e.g., Bland-Altman transformation combined with ordinary least squares regression, and Deming Regression) are sufficient in most cases. However, if non-linear patterns arise, parametric methods are typically inadequate. Henceforth we examine non-parametric evaluation methods such as Smoothing splines and Kernel Regression. Assuming equally distributed and independent error terms and sufficiently many clinical samples and replicated measurement, these non-parametric methods proved to be robust against non-linearity. We discovered that smoothing splines estimator were the most appropriate of the two because of fewer unavoidable subjective decisions and relatively stable uncertainty bands. Besides, smoothing splines with Bland-Altman transformed data proved much more suitable to interpret visually and reduce variability in x-direction. Thus, the combination of Bland-Altman transformation and smoothing splines is considered the most robust choice dealing with non-linearity.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

σ	Standard deviation
AQA	Analytical Quality Assessment
AR	All Replicates
AV	Analytical variation
BAT	Bland-Altman transformation
CA	Commutability Assessment
CV	Estimated Coefficient of Variation
DR	Deming Regression
EQA	External Quality Assessment
IQA	Internal Quality Assessment
KR	Kernel Regression
LLT	Log-log transformation
M	Total Number of Predictions
m	Number of Control materials
MOR	Mean of Replicates
MP	Measurement Procedure
N	Total Number of Measurements

n	Number of Clinical Samples
OLSR	Ordinary Least Squares Regression
PI	Point-wise Prediction Intervals
PRS	Parametric Regression Splines
R	Number of Measurements on a Sample
SD	Estimated Standard Deviation
SS	Smoothing Splines
TSR	Thiel-Sen Regression

Chapter 1

Introduction

This thesis's primary purpose is to utilize statistical models applied in the classification of control materials. The primary focus is on the practical use of the statistical methods, and consequently, in-depth detail regarding the underlying theory is not presented. A modest overview of the estimators is presented briefly as a necessary supplement to their application. In Chapter 1, an implementation of the discussion of the commutability concept and corresponding definitions are presented. In Chapters 2 and 3, we will review parametric regression models' theory and apply them as part of the commutability assessment. In Chapters 4 and 5, an examination of alternative regression models (non-parametric regression models) and their performance is presented.

1.1 Commutability - The concept

The Commutability of control materials concerning comparisons of multiple (≥ 2) *measurement procedures* is the central theme in this text. From the VIM3: International Vocabulary of Metrology, a measurement procedure is defined as follows:

Definition 1.1.1. - Measurement procedure

A detailed description of a measurement according to one or more measurement principles and a a given measurement method, based on a measurement model, including any calculation to obtain a measurement result.

We will restrict ourselves to consider two measurement procedures at the time. [Vore, 2014] defines commutability as follows:

Definition 1.1.2. - Commutability (of a control material)

Property of given control material, demonstrated by the closeness of agreement between the relation among measurement results for a stated quantity in this material, obtained

according to two measurement procedures, and the relation obtained among the measurement results from clinical samples.

In an everyday speech, one might say that commutability is a property of a control material where the control material acts 'similar' to clinical materials concerning two or more measurement procedures. It is essential to stress that commutability is a control material feature. Commutability is consequently not comparability. This is a misconception we want to avoid. [Braga and Panteghini, 2019] argues that the commutability of control materials appears overlooked. Especially for some decades ago, commutability was barely mentioned. The reason for little acknowledgment may be due to not realizing its importance.

However, in recent years people have started to see the potential of the commutability property. The commutability property is the most crucial property a control material possesses [G. W. Miller, Jones, Horowitz, and Weykamp, 2011]. Holding commutable control materials provides useful applications. First of all, [Braga and Panteghini, 2019] states that commutable control materials are essential for analytical quality assessment. Roughly speaking, evaluation of measurement procedures in laboratories concerning given standards, such as satisfactory *precision* and *accuracy* of the measurement procedures. *Matrix effects* will result in enlarged variability within clinical samples. They may also enlarge the difference between clinical samples and control material samples, which frequently provide non-commutable control materials or misclassified control materials. Non-commutable control materials reflect that the producer of the control materials has inappropriate measurement procedures concerning relevant performances. The definition used in this text for *matrix effects* is the same as given in Nomenclature for automated and mechanized analysis (Recommendations 1989);

Definition 1.1.3. - Matrix effect

The combined effect of all components of the sample other than the analyte on measuring the quantity.

Matrix effects commonly affect the *precision* and *accuracy* of the measurement procedures. See Section 1.3 for definitions. Consequently, the presence of matrix effects may result in erroneous medical decisions. Knowing the *reference range* of the measurand in question and performing several replications of each sample, we can disclose the clinical samples' *precision*. [Solberg, 1993] defines *reference interval* by:

Definition 1.1.4. - Reference range

The reference interval is an interval constructed by the central 95% of a reference population.

Obtaining the *accuracy* of the clinical samples is done explicitly by comparing clinical samples with control material samples. We can then tell if measurement values propose disease and other irregularities in a patient with increased trustworthiness. In opposition, non-commutable control material will leave us with potential erroneous medical decisions because of poor precision and accuracy. Diagnostic mistakes are likely to be highly dangerous. Avoiding the before-mentioned hazards is an essential argument for attaining commutable control materials.

How does one go forth assessing the commutability property of control materials? According to [W. G. Miller et al., 2018] there are strict guidelines regarding the assessment of commutability;

1. Obtain appropriate control materials.
2. Collect representative and fresh clinical samples.
3. Measure the clinical samples and control material samples by using several pairs of measurement procedures.
4. Apply a commutability evaluation method on all pairs of measurement procedures and report with all outcomes concerning a chosen acceptance criterion.

The first three points are more practical guidelines than theoretical. A complete discussion of these three points is beyond this text's scope and is consequently assumed to be performed correctly. Thus, the main focus will be on the fourth point, which is about statistical commutability assessment methods. Data from the first three points are either provided by the EQA-organisation Noklus or simulated in this text. However, even though the main focus is not on the three first steps, it will be interesting to discuss them in light detail. Both EQA- and IQA-organisations follow the list above. Obtaining control materials is usually done in laboratories, and the goal is to make the control materials similar to patient materials. In the case of external quality assessment, when the control materials are finished and measured, they are shipped to an EQA-organisation, which measures the clinical samples and control material samples themselves. Then results are reported back to the control materials' makers, and their equipment's performance is evaluated. Control materials are required to fulfill the following conditions to remain representative in the whole analytical process:

1. Homogeneity
2. Durability
3. Stability

4. Appropriate Volume

5. Affordable

6. Adequate analyte concentration

Control materials typically do not satisfy all these points and are typically processed to satisfy them, which comes with a price. Processed control materials may no longer act like clinical samples, which they were initially meant for, frequently due to matrix effects. Ultimately, processed control materials may be misclassified in the commutability assessment analysis.

When collecting clinical samples, it is crucial to ensure that clinical samples provide measurements scattered along and preferably beyond the reference range. The clinical samples' uniform dispersion is essential because some sample space regions indicate that a medical decision is necessary. Control materials in regions where clinical samples are few might lead to falsely accepting commutability because the applied acceptance criterion used in this text relies on sensible analytical precision.

Moreover, fresh and representative clinical samples are vital. Storage, transport, and other factors may process the clinical samples. Processed clinical samples are likely to be different from the fresh samples, often due to unwanted matrix effects. The clinical samples' matrix effects could again lead to a wrongly classified control material regarding the commutability property.

After obtaining control materials and collecting clinical samples, it is time to measure them. It is crucial to minimize unexplained analytical errors in the measurement phase and the transportation phase. Unexplained analytical variability, in addition to the explained analytical variation, enlarge the acceptance region for commutability, which is why matrix effects are important to control.

Classification of analytical variability is an essential part of the assessment of the commutability concerning control materials. AV is the primary source of the total variability. However, this is not the only source of variability one might consider in a commutability assessment approach. For instance, [W. G. Miller et al., 2018] uses biological variability when determining the limits of acceptance for the commutability property.

The total analytical variation contains variance connected to the measurement procedures and potential variance produced by processed clinical samples' matrix effects. Naturally, AV is minimized by minimizing the clinical samples' matrix effects, and one might, with minimized AV, correctly conclude whether commutability is accepted or else is not. By estimating the AV, which is assumed to exhibit negligible variance connected to matrix effects, regression models are applied together with the uncertainty of measurement procedures as limits of acceptance.

In the discussion of variability, it is essential to declare that we have two main types of variances, denoted by deterministic and random variance. Note that random variation is not necessarily the same as an unexplained variation. The majority of the analytical variability is considered deterministic and will solely be applied to construct the limits of acceptance regarding a control material's commutability property.

If random variability is desired to be implemented, methods to estimate both random and deterministic variation exists. Henceforward one could estimate the total analytical variation defined as the sum of deterministic and random variation. Nevertheless, the random variation is small compared to the deterministic variation by experience. Hence, no further interest is taken concerning appending the analytical variance's random component into our acceptance criterion.

1.2 Weaknesses of current evaluation methods

Currently, there exist several methods concerning the assessment of the commutability property of control materials. The problem, however, is that every one of them possesses weaknesses. Depending upon the model, weaknesses may vary. To this day, we apply models such as linear regression models potentially combined with well-behaving transformations. The problems typically arise when considering the linear model assumptions and any transformation of non-linear data patterns. These linear model assumptions sometimes appear too stringent for the clinical samples and are rejected on several occasions for our purposes.

We could, however, claim that a suitable transformation will help satisfy the model assumptions. See Section 2.3 and Section 2.4) for details concerning the transformation approaches. Nonetheless, if these transformations meet the *linear model assumptions*, they carry several additional requirements on their own, which may not be satisfied. The main point is that independently of how many smart and elegant transformations we do, the result will not produce a general and perfect assessment procedure for commutability. Besides, if the underlying clinical samples pattern regarding the measurement procedures is non-linear, it is sensible to avoid linear models altogether. In any measurement comparison procedure, we expect the relationship between the measurement procedures to be linear in theory. In reality, weak non-linear trends typically arise. To this day, we rarely use methods for dealing with non-linearity without transformations or any assumptions regarding the normal distribution. Assuming that our error terms are normally distributed and adequately linear is something we would like to relax. Henceforth, it will be of interest to look into non-parametric methods as part of the commutability evaluation. Such methods are discussed in Chapters 4 and 5.

Currently, most used methods of evaluation of control materials are parametric and some examples are presented in the list below:

1. Ordinary least squares regression models
2. Errors-in-variables regression models (e.g., Deming regression)
3. Log-log transformations
4. Bland-Altman transformations (See Section 2.4 for elaboration)
5. Piece-wise regression models

The estimators obtained from 1, 3, 4, and 5 exclusively rely on variability in one of the two measurement procedures. All five models require a linear relationship with regards to the raw or transformed clinical samples. In other words, the clinical samples must be capable of constructing a model in the form below, either directly by using them raw or by employing a transformation.

$$\mathbb{Y}|\mathbb{X} = \mathbb{X}\beta + \epsilon. \tag{1.1}$$

Above, $\mathbb{X}$ is the design matrix, and β is the vector of the theoretical regression coefficients. Ignorance of the two-dimensional variability appears unwise because it is unrealistic. Consequently, it appears naive using either 1, 3, or 5. Especially when we are not using a measurement procedure with high analytical *precision*, notice that we do not mention the Bland-Altman transformation as part of the naive approaches. The reason is that the Bland-Altman transformation will significantly reduce the variability in x-direction and is consequently appropriate when the variability in x is excluded from the statistical analysis.

A natural expansion is to include variation in both measurement procedures. Errors-in-variable models formally denote linear regression with two-dimensional variability. Depending on the information regarding the direction and magnitude of errors, we use a unique errors-in-variables model. There exist several versions of these kinds of models. Nevertheless, we will only use *Deming Regression* in this text.

1.3 Precision, accuracy, and other relevant definitions

When studying measurement errors and analytical variation, it is prudent to introduce several definitions, mainly precision and accuracy. In everyday speech, these two statistical terms appear to describe the same phenomena; nevertheless, there is a meaningful difference between the two in the fields of statistics and laboratory science. A definition of precision is displayed as presented in the VIM3: International Vocabulary of Metrology:

Definition 1.3.1. Measurement precision is the closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.

According to Definition 1.3.1, measurement precision is measurable and estimated for specific measurement procedures and measurands. There exist several estimators for precision. Two of them that we will stick to are SD and CV, the estimated standard deviation and estimated coefficient of variability. Considering replicated measurements done on a specific sample $i = I$; $\{\text{MP}_{Ir}\}$ with $r \in \{1, \ldots, R\}$, we define

$$\text{SD}_I = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\text{MP}_{Ir} - \overline{\text{MP}_I}\right)^2} \tag{1.2}$$

$$\text{CV}_I = \frac{\text{SD}_I}{\overline{\text{MP}_I}}. \tag{1.3}$$

The analytical precision may be estimated by summing the estimated standard deviations for all n clinical samples. An essential advantage of using CV is that it is scale-independent, which implies that it is more comprehensive than SD in most cases. Accordingly, whenever it is possible, the CV will be applied. Moreover, note that precision says nothing about the closeness of agreement to the true quantity value, and this is where accuracy comes in:

Definition 1.3.2. Measurement accuracy is the closeness of agreement between a measured quantity value and a measurand's true quantity value.

Thus, measurement values near the true quantity value of a measurand will enhance the accuracy and, in opposition, will decrease when measurement values are distant from the true quantity value of the measurand. In opposition to precision, accuracy is not measurable because the true quantity value is latent. However, estimated *measurement errors* reflect accuracy to some extent. A comprehensive definition of measurement errors is provided below.

Definition 1.3.3. Measurement errors (ME) is the difference between a measured quantity value and a reference quantity value. We decompose measurement errors into both random and systematic measurement errors. That is

$$\text{ME} = \text{RME} + \text{SME}. \tag{1.4}$$

Random measurement error (RME) is a type of measurement error that varies unpredictably. Systematic measurement error (SME) is a type of measurement error that is either fixed or

predictable. The estimate of the systematic measurement error is called the measurement bias.

1.4 Formal tests

In this thesis, we will use formal tests when making decisions regarding model assumptions and commutability acceptance. However, in Chapter 3, we will discuss the use of plots as assessment. The reasoning behind this is to get a solid idea of how the commutability assessment procedure works in principle. In Section 3.6 and Section 5.5 we will singularly use objective methods as evaluation. The same goes for model premises, especially linearity—Shapiro-Wilk tests, Breusch-Pagan tests, and Durbin-Watson tests *linear model assumptions*. We present a slightly detailed discussion of these three diagnostic tests in this section. For automatic commutability assessment, we will use our own R-functions. Their description and scripts are found in Appendix A.

1.4.1 Shapiro-Wilk test

Commonly we use the residuals extracted from model summaries to check for normality. This is equivalent to testing the actual data for normality because of linearity. The residuals are denoted by $\{e_{ir}\}$ with $i \in \{1, 2, \ldots, n\}$, and $r \in \{1, \ldots, R\}$. For one specific sample $i = I$ and within one specific replication $r = K$, the corresponding residual is defined by

$$e_{IK} = \text{MP}_{IK} - \hat{g}(\text{MP}_{IK}). \tag{1.5}$$

$\hat{g}$ is the fitted model. The corresponding order statistics is denoted by $\{e_{(i)}\}$. We formulate the hypotheses of interest:

$$h_0 : \ \{e_{ir}\} \text{ are from a normal distribution} \tag{1.6}$$

$$h_1 : \ \{e_{ir}\} \text{ are not from a normal distribution} \tag{1.7}$$

We formulate the following test statistic.

$$W = \frac{\left[\sum_{i=1}^{n}\sum_{r=1}^{R} a_{ir} e_{(ir)}\right]^2}{(N-1) \cdot \text{SD}} \tag{1.8}$$

Typically, we reject normality for extreme values of W. Values far from 0 are considered extreme.

1.4.2 Breusch-Pagan test

Investigating whether the regression residuals is dependent on the explanatory makes the foundation of this test. In other words, the Breusch-Pagan test checks whether the residuals are heteroscedastic. Let $\{\text{MP}_{ir}\}$ with $i \in \{1, 2, \ldots, n\}$ and $r \in \{1, \ldots, R\}$ be the measurement values of the measurement procedure acting as the explanatory variable. With the residuals denoted as in Equation (1.5) we formulate the hypotheses of interest:

$$\begin{aligned} h_0 &: \ \{e_{ir}\} \text{ is homoscedastistic} \\ h_1 &: \ \{e_{ir}\} \text{ is heteroscedastistic} \end{aligned}$$

The following theoretical regression model is estimated:

$$\frac{e_{ir}^2}{\text{SD}^2} = \beta_0 + \beta_1 \cdot \text{MP}_{ir} + \epsilon_{ir}. \tag{1.9}$$

From ordinary least squares we may estimate Equation (1.9). We denote the estimated estimator of Equation (1.9) by $\widehat{\frac{e_{ir}^2}{\text{SD}^2}}$. With the assumption that $\text{E}\left[\frac{e_{ir}^2}{\text{SD}^2}\right] = 1$ and the estimated model of Equation (1.9) we find that:

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{n}\sum_{r=1}^{R}\left(\frac{e_{ir}^2}{\text{SD}^2} - 1\right)^2 \\ \text{SSR} &= \sum_{i=1}^{n}\sum_{r=1}^{R}\left(\frac{e_{ir}^2}{\text{SD}^2} - \widehat{\frac{e_{ir}^2}{\text{SD}^2}}\right)^2. \end{aligned} \tag{1.10}$$

We may define the test statistic for the formulated hypotheses as

$$\text{LM} = \frac{1}{2}\Big(\text{SST} - \text{SSR}\Big). \tag{1.11}$$

As proven by Breusch and Pagan $\text{LM} \sim \chi_1^2$ asymptotically. Consequently, extreme estimates for LM propose rejecting the null-hypothesis given that we have a satisfactory large N.

1.4.3 Durbin-Watson test

The Durbin-Watson test is a formal assessment for the presence of auto-correlation in our data. Roughly speaking, the test uses pair-wise successive residuals and check for a correlation between them. Consequently, the text does not say much about the error terms' global dependence, ' one of the *linear model assumptions*. However, the presence of auto-correlation means that we have at least some dependent error terms, which implies that global

independence will be somewhat doubtful. We formulate the following hypotheses for the Durbin-Watson test:

$$h_0: \text{ Auto-correlation is not present } (\rho = 0)$$
$$h_1: \text{ Auto-correlation is present } (\rho \neq 0)$$

When using the notation as in Equation (1.5), we define the test statistic d by

$$d = \frac{\sum_{j=2}^{N}(e_j - e_{j-1})^2}{\sum_{j=1}^{N} e_j^2}, \tag{1.12}$$

where $e_j = \rho e_{j-1} + \epsilon_j$. For small values of d, the error terms are auto-correlated positively, and for large values of d, the error terms are auto-correlated negatively. Moderate values of d (around $d = 2$) propose that no auto-correlation is present.

1.4.4 The interpretation of the formal tests

The formal tests above will have several outputs in R, that is, both estimated test statistics and corresponding p-values. However, exact p-values will not be of importance. Degrees of significance is rather used when reporting the test results. The classical notation for statistical significance is used:

Significance degree	p-value	Symbols
Significant	>0.05	*
Very significant	>0.01	**
Extremely Significant	>0.001	***

For the acceptance of the formal tests, we accept when p-value≥ 0.05. The reason for greater or equal is because of the hypotheses formulations given above.

Chapter 2

Parametric methods of assessment

2.1 Current methods of evaluation

After considering the definition of commutability regarding control materials and cautions regarding errors, it is time to discuss parametric commutability evaluation methods. There exist several methods for assessing commutability. A simple and straightforward commutability evaluation method for control material samples is to plot clinical samples and control material samples concerning two different measurement procedures upon each other. Control materials are commutable if they fit the regression line - constructed by clinical samples - well, whereas the patient-dependent regression line should ideally follow $y = x$. Otherwise, there is some difference concerning the measurement procedures in question or some unwanted matrix effects. However, this comparability issue will not be critical because the utilized acceptance criterion for commutability is independent of comparability.

2.2 Ordinary least squares

Ordinary least squares regression may be used as part of the commutability acceptance criterion. Particularly, the rule for accepting a control material's commutability property is to check whether a control material sample lies within the $(1 - \alpha/m) \cdot 100\%$ prediction bands where m is the number of control materials found in the particular data set. Hence, when analyzing five control materials simultaneously, we must check whether the control material samples lie within the 99% prediction bands constructed by the clinical samples. The data sets considered in Chapter 3 holds five control materials; accordingly, 99% will be the applied level for the estimated prediction bands. Incorrect conclusions concerning commutability acceptance are classified into two main components: type I error and type II error. These

types of errors are defined as:

$$\begin{aligned} \text{Type I error} &= \mathrm{P}(\text{Rejecting commutability for a CM} \mid \text{The CM is commutable}) \\ \text{Type II error} &= \mathrm{P}(\text{Accept commutability for a CM} \mid \text{The CM is not commutable}). \end{aligned} \tag{2.1}$$

CM is shorthand for control material. The first error may be controlled by choosing a suitable value for α. A typical choice is Type I error $\approx \alpha$ with $\alpha = 0.05$. The second error is harder to control but is frequently reduced with suitably large sample sizes. With Equation (2.2), an analytical expression for the prediction bands is straightforward to obtain. Let $i \in \{1, 2, \ldots, n\}$ be the sample index, and let $r \in \{1, \ldots, R\}$ be the replication index. Then, the ordinary least squares model is defined as

$$\mathrm{MP}_{Air} | \mathrm{MP}_{Bir} = a + b \cdot \mathrm{MP}_{Bir} + \epsilon_{ir}, \ \epsilon_{ir} \sim \mathcal{N}(0, \sigma_A^2). \tag{2.2}$$

We assume that the random variables $\{\epsilon_{ir}\}$ are independent and identically distributed. This assumption implies that the error terms must be independent and possess equal variances (homoscedasticity). Moreover, the data patterns are required to be approximately linear since Equation (2.2) is a linear model. The proper usage of ordinary least squares regression is accordingly dependent on the following conditions:

1. Independent error terms
2. Equal variances of all error terms
3. Linearity of observations
4. Normality of error terms.

The error terms must satisfy all these four assumptions to obtain trustworthy ordinary least squares estimators. Holding unreliable estimators may produce unreliable prediction bands, which is problematic concerning our acceptance criterion for commutability. Consequently, we may not use ordinary least squares regression as part of the commutability assessment criterion. The linear model assumptions are denoted by the four requirements listed above. Assuming that the linear model assumptions are satisfied, it is straight forward to show that the 99% prediction bands can be estimated explicitly by

$$\mathrm{PI} = \mathrm{MP}_B^* \pm z_{0.995} \cdot \mathrm{SD}_A \sqrt{1 + \frac{1}{n} + \frac{(\mathrm{MP}_B^* - \overline{\mathrm{MP}_B})^2}{S_{BB}}}, \tag{2.3}$$

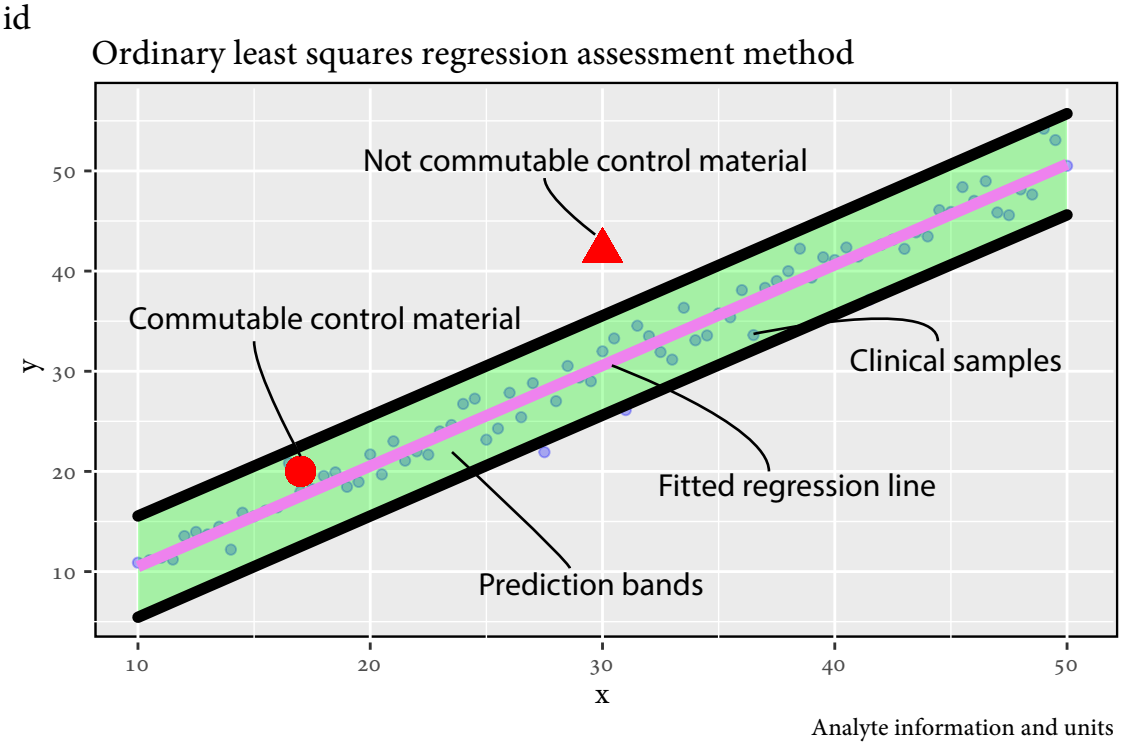

Figure 2.1 *– This figure reflects the visual commutability assessment scheme used in this text.*

where

$$S_{BB} = \sum_{i=1}^{n} (\text{MP}_{Bi} - \overline{\text{MP}_B})^2.$$

A large sample size signifies that the prediction bands' width will be approximately the same as the corresponding confidence bands. In other words, the prediction bands will be narrower when possessing more numerous clinical samples. On the contrary, large SD_A increases the width of the prediction bands, which increases the probability of making a type II error. A visualization of how the acceptance criterion for commutability functions is displayed in Figure 2.1. Figure 2.1 proposes that the visual commutability assessment procedure is simple in practice. However, control materials positioned at the boundaries may be hard to classify. As part of the linear model assumptions assessment, visual tests like residual versus fitted plots and quantile-quantile plots are used. Alternatively, one might use formal tests such as those formulated in Section 1.4 or *gvlma()* in R. For subjective interpretation purposes, visual tests regarding linear model assumptions are best to avoid. However, the visual tests will be used and discussed in some detail when testing the parametric methods in Chapter 3. How to interpret visual plots are assumed to be known to the reader and will not be elaborated further.

The real and simulated data are in this text nested. Mainly, each sample is measured several times. Replicated measurements on the samples naturally induce auto-correlation, and a discussion on the handling of replicates will be an element of the discussion. The analysis of ordinary least squares regression as part of the commutability evaluation require a discussion on disclosure of typical faults. As already touched upon, ordinary least squares regression has stringent model assumptions that must be satisfied.

Moreover, ordinary least squares regression ignores variability in x. One might, therefore, conclude that this commutability assessment method is somewhat unrealistic. In some instances, clinical samples may not follow a linear pattern, which results in poor model fits. The suggested solution to this may be some transformation of the observations.

2.3 Log-log transformation

If the clinical samples seem to follow a non-linear curve, an alternative approach is preferred. A typical approach to deal with a non-linearity is to transform the response and the predictor by applying the natural logs. The goal of doing this is to achieve a linear relationship. The appropriate transformation requires the application of the logarithm, of any base, on both measurement procedures. In this text, the natural base will be used. The log-log transformation will only produce linear relationships if the raw pattern of the clinical samples follows a curve of the form

$$\text{MP}_A = K[\text{MP}_B]^{\beta_1}, \tag{2.4}$$

where K and β_1 are constants. By this constraint, not all non-linear patterns will convert into a straight line using the log-log transformation. For instance, the only satisfactory intercept is 0, and only one-term polynomials are accepted. An illustration of this is displayed below. Taking natural logs on both sides of Equation (2.4) produces

$$\log_e(\text{MP}_A) = \beta_0 + \beta_1 \log_e(\text{MP}_B), \tag{2.5}$$

where $\beta_0 = \log_e(K)$. It is evident that Equation (2.5) is a linear relationship as desired, because it is in the appropriate form. After log-log transforming the response and prediction, ordinary least squares or any other linear on estimators may be used to estimate β_0 and β_1. To reveal the potential problems concerning the log-log transformed data, an illustration of the use of log-log transformation for several data patterns is included:

- $m_1: \ y = 1.2 \cdot x^3$
- $m_2: \ y = 50 + 1.2 \cdot x^3$

- m_3 : $y = x^2 + \exp(x)$

Log-log transformations of the simulated data constructed from the equations above produced the curves presented in Figure 2.2.

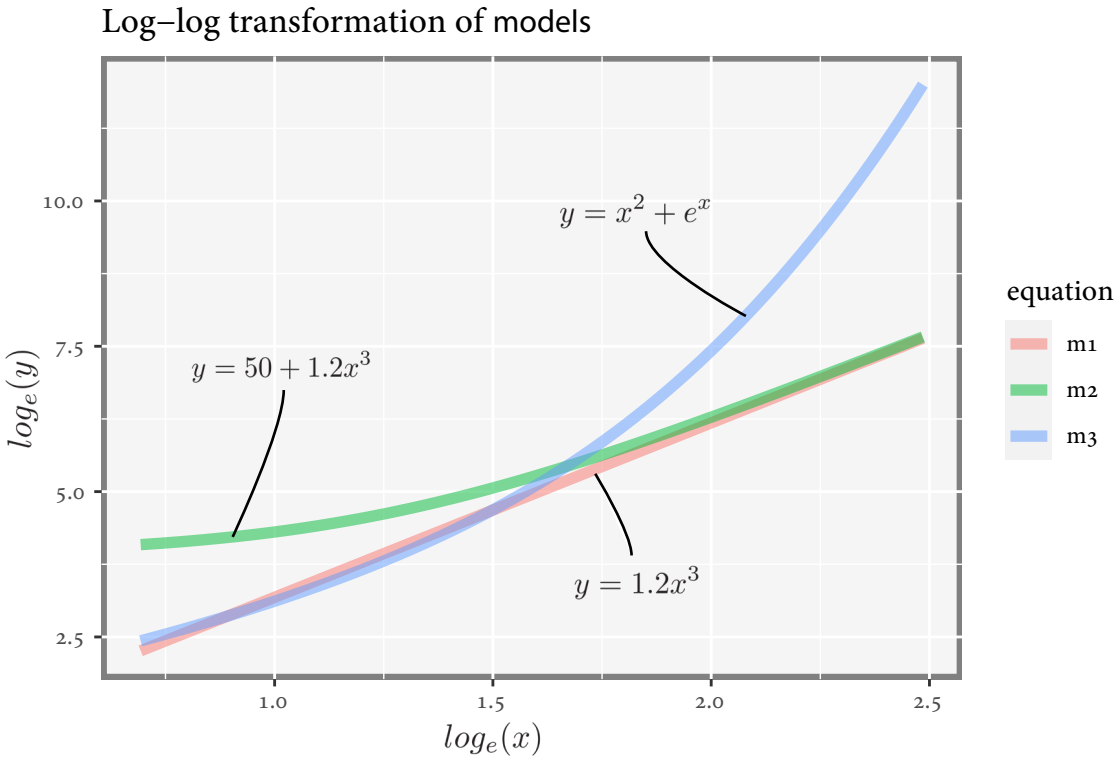

Figure 2.2 *– This figure illustrates the log-log transformations of the three models defined above. As expected, only m_1 resulted in a linear curve. Regard that we need significant intercept coefficients for it to cause problems.*

From statistical theory, log-log transforming our data may result in more symmetrical data, and heteroscedasticity may no longer be a problem. However, the log-log transformation is no guaranteed fix for the linear model assumptions. It even happens that more model assumptions are left unsatisfied for the transformed clinical samples. Accordingly, we may argue that the log-log transformation should only be used when the underlying distribution is approximately log-normal. Moreover, the clinical samples should follow Equation (2.4) well. Unfortunately, the log-log transformation, combined with ordinary least squares, will not register variability in x-direction. However, it is possible to combine the log-log transformation with *Deming regression*, which accounts for variability in both measurement procedures. More on *Deming regression* is presented in Section 2.5. Another approach dealing with variability in x is to *Bland-Altman*-transform the data. The *Bland-Altman transformation* does not directly

account for two-dimension variability but reduces it to some extent. This transformation will be discussed in more detail in the next section.

2.4 Bland-Altman transformation

Another scheme for evaluating the commutability property of a control material is the so-called Bland-Altman approach. The assessment scheme of commutability relies on the Bland-Altman transformation, which is defined by the logarithmic difference between the measurement procedures as the response (i.e., MP_A and MP_B) and the average measurement values between them as the predictor. According to [W. G. Miller et al., 2018], a plot visualizing the Bland-Altman transformed clinical samples will be favorable. The reason is that our measurement procedures' modeling bias reveal hard-to-detect trends concerning measurement magnitudes between two measurement procedures. Moreover, the Bland-Altman transformation reduces the variability magnitude in x, which is favorable when ignoring variability in x-direction. In the simulation section, we will look into this in more significant detail.

In similarity to log-log-transformations, we transform both response and predictor, as formulated above. Let $i \in \{1, 2, \ldots, n\}$. The Bland-Altman transformation with a logarithmic difference is then given by

$$\text{Response}_i = \log_e(\text{MP}_{Ai}) - \log_e(\text{MP}_{Bi}) \tag{2.6}$$

$$\text{Predictor}_i = \frac{\text{MP}_{Ai} + \text{MP}_{Bi}}{2}. \tag{2.7}$$

The goal of these transformations is to transform all samples so that, ideally, they follow a horizontal line. Clinical samples following a horizontal line mean that the log-difference of measurements between the procedures is constant. Samples following the line $y = 0$, therefore, indicates agreement between the relevant measurement procedures. However, from a commutability aspect, data following a straight line is not of notable importance since the acceptance criterion used is independent of deviations from linearity. To see how the Bland-Altman transformation works, it will be of interest to define the relationships proposed below for data patterns according to four underlying models. The essential point of this is to visualize how the Bland-Altman transformation deals with the different types of relationships. Figure 2.3 visualizes this.

- $m_1: \; y = x + \mathcal{N}(0, 0.25)$
- $m_2: \; y = 2x + 0.01x^2 + \mathcal{N}(0, 0.25)$
- $m_3: \; y = 2x + 0.001x^3 + \mathcal{N}(0, 0.25)$

- $m_4 : \ y = 20 + 2x + \mathcal{N}(0, 0.25)$.

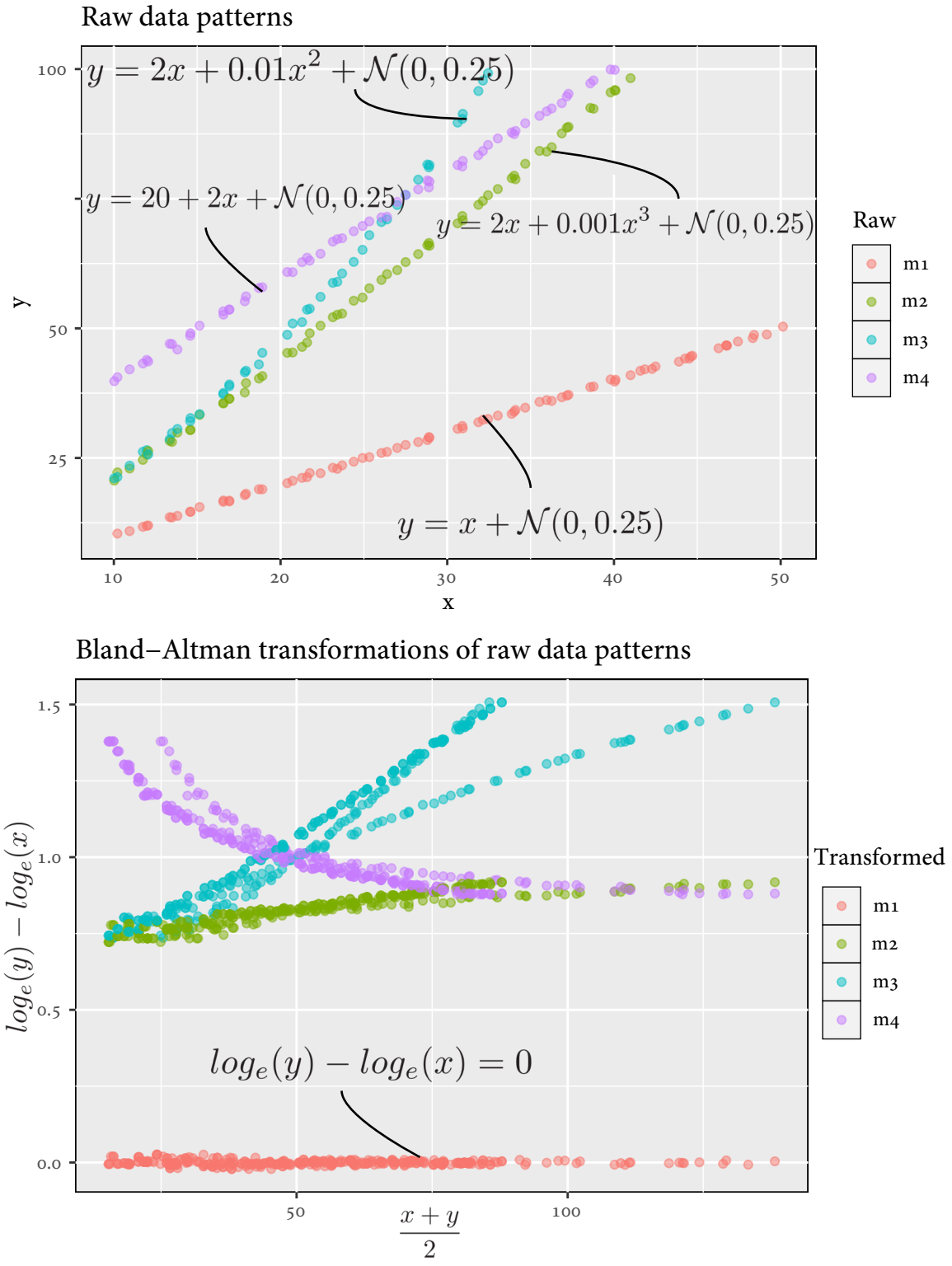

Figure 2.3 *– Raw data and Bland-Altman transformed data for the four models given is visualized. Note how the Bland-Altman transformed data differs from the different raw data.*

From Figure 2.3 we see that the first pattern, that is $y = x + \mathcal{N}(0, 0.25)$, was transformed into a horizontal line. This is because of

$$\log_e(y) - \log_e(x) = \log_e(x + \mathcal{N}(0, 0.25)) - \log_e(x). \tag{2.8}$$

The probability of this difference to be close to zero is large because the expected value of the random noise is equal to zero and possesses relatively small variance. Besides, non-zero intercepts, as implemented in m_4, results in large deviations from linearity in the start, but the pattern is approximately linear as the predictor values increase. By experience, there is merely a handful of measurement procedure comparison studies resulting in a horizontal line. Interestingly, non-linear patterns is frequently expected in the Bland-Altman plots because of the 'revealing' property. A natural approach is to model potential non-linear patterns with polynomial regression. After transforming the data, the theoretical model below is estimated by estimating the corresponding regression coefficients:

$$\text{Response}_i|\text{Predictor}_i = \sum_{j=0}^{p} \beta_j \cdot \text{Predictor}_i^j + \epsilon_i. \tag{2.9}$$

In Equation (2.9), $\epsilon_i \sim \mathcal{N}(0, \sigma^2_{Response})$. We will estimate the polynomial regression curve (with degree $p-1$) using ordinary least squares. p is currently chosen manually concerning what seems like a good fit for the data. However, $p = 4$ and $p = 5$ are often satisfactory choices. Choosing p is generally dependent upon our data, which implies that a general method for choosing p may be of interest. There exist functions in R, e.g., *stepAIC()*, which obtain such p by using a step-wise algorithm dependent on the models' estimated Akaike information criterion. Applying ordinary least squares with linearity in the regression coefficients, estimating the corresponding prediction bands is straight forward. Hence, it follows that the same acceptance criterion principle is utilized as in Figure 2.1).

The IFCC group is well known to use the Bland-Altman transformations. However, how they apply it as a commutability assessment is quite different from what this text suggests. This difference is due to the that the IFCC group typically does commutability and comparability tests simultaneously. Therefore, [W. G. Miller et al., 2018] operates with fixed boundaries that are independent of measurement procedures compared. They suggest using the clinical samples' mean bias as the golden standard, whereas we use the estimated regression line and the estimated prediction bands constructed by the clinical samples. Note that the clinical samples mean bias is equivalent to our approach if $p = 0$ is selected. Applying the prediction bands as part of the acceptance criteria will induce slightly different requirements for every

pair of measurement procedure comparisons. In opposition, they recommend the same, fixed requirements in every comparison procedure.

2.5 Deming regression

In previous methods, it was assumed no error in the x-direction. In reality, this assumption is rarely satisfied. There will be errors in both x- and y-direction. In practical terms, analytical variability is expected in both measurement procedures, which are realistic and intuitive. There exist methods taking care of this supplementary variability. One such method is *Deming regression*, which is the generalization of orthogonal regression. Generally speaking, the orthogonal regression estimator is a linear regression estimator where we minimize the orthogonal projections to obtain the estimated regression coefficients. With $i \in \{1, 2, \ldots, n\}$, [Dunn, 1989] defines the system of linear equations as:

$$\begin{aligned} \text{MP}_{Bi} &= \tau_i + \delta_i \\ \text{MP}_{Ai} &= c + b\tau_i + \epsilon_i \end{aligned} \tag{2.10}$$

Moreover, the two error-term components are assumed to be normally distributed, that is

$$\begin{bmatrix} \epsilon_i \\ \delta_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{bmatrix} \right) \forall\, i \tag{2.11}$$

Furthermore, assume that $\{\delta_i\}$ and $\{\epsilon_i\}$ are independent. As imposed in Equation (2.11), the covariance between the two error terms must be equal to zero. It is crucial to note that τ_i and $c + b\tau_i$ are the true but latent values within a clinical sample i concerning the two measurement procedures. The latent values within the clinical samples are assumed to be fixed realizations of a random variable. Consequently, the magnitudes of c and b determine the linear relationship between the measurement procedures. In the real world, however, it is natural to generalize Equation (2.10) as

$$\begin{aligned} \text{MP}_{Ai} &= \tau_i + \delta_i \\ \text{MP}_{Bi} &= f(\tau_i) + \epsilon_i \end{aligned} \tag{2.12}$$

for any real function f. The reasoning of this statement relies on the likely possibility of measurement procedures having non-linear relationships. However, for the Deming models presented in this chapter, it is assumed that the underlying relationship between the measurement procedures is linear. In Section 3.6, we assume $\{\tau_i\}$ to be realized values of the random

variable

$$U \sim \mathcal{U}(\alpha, \beta). \tag{2.13}$$

What differs Deming regression from orthogonal regression is that

$$\lambda = \frac{Var[\epsilon_i]}{Var[\delta_i]} = \frac{\sigma_A^2}{\sigma_B^2} \tag{2.14}$$

not necessarily are equal to 1. Hence, the variance in one measurement procedure might be larger or smaller than the other measurement procedure variance. To successfully estimate the regression coefficients, information on λ is required. In most cases, the true value of λ is unknown and is rarely fixed. Since the Deming regression model is dependent on knowing λ, constructing an estimator of λ will be vital. From Equation (2.10) and the fact that $\{\epsilon_i\} \overset{d}{=} \{-\epsilon_i\}$ and $\{\delta_i\} \overset{d}{=} \{-\delta_i\}$ it follows that the theoretical Deming regression model takes the form

$$\text{MP}_{Ai}|\text{MP}_{Bi} = \beta_0 + \beta_1(\text{MP}_{Bi} + \delta_i) + \epsilon_i. \tag{2.15}$$

From Equation (2.15), it is clear why $\{\epsilon_i\}$ and $\{\delta_i\}$ are required to be independent. Alternatively, it would be necessary to include a covariance term into the variance of Equation (2.15). Holding n clinical samples measured by two measurement procedures and by applying the method of moments results in

$$\begin{aligned} \hat{\beta}_1 = b_1 &= \frac{S_{AA} - \lambda S_{BB} + \sqrt{(S_{AA} - \lambda S_{BB})^2 + 4\lambda S_{BA}^2}}{2S_{BA}} \\ \hat{\beta}_0 = b_0 &= \overline{\text{MP}_A} - b_1 \cdot \overline{\text{MP}_B}. \end{aligned} \tag{2.16}$$

This set of estimators assumes that either λ is known or that both σ_A^2 and σ_B^2 are known [Gillard, 2010]. At least, we need some trustworthy preliminary information on this ratio of variances. A sufficiently large sample size and a suitable number of replicates provide this information, which results in a reliable estimator for λ. The estimated covariances are defined as

$$\begin{aligned} S_{BB} &= n^{-1} \cdot \sum_{i=1}^{n} (\text{MP}_{Bi} - \overline{\text{MP}_B})^2 \\ S_{AA} &= n^{-1} \cdot \sum_{i=1}^{n} (\text{MP}_{Ai} - \overline{\text{MP}_A})^2 \\ S_{BA} &= n^{-1} \cdot \sum_{i=1}^{n} (\text{MP}_{Bi} - \overline{\text{MP}_A})(\text{MP}_{Ai} - \overline{\text{MP}_A}). \end{aligned} \tag{2.17}$$

As formulated by [Dhanoa et al., 2011], the estimation scheme relies on first estimating σ_A^2 and σ_B^2 and calculating their ratio. That is

$$\hat{\lambda} = \frac{\hat{\sigma_A^2}}{\hat{\sigma_B^2}}. \tag{2.18}$$

[Dhanoa et al., 2011] estimates the variability components σ_A^2 and σ_B^2 as

$$\begin{aligned}
\hat{\sigma_A^2} &= \frac{1}{N-n}\sum_{i=1}^{n}\sum_{r=1}^{R}(\text{MP}_{Air} - \overline{\text{MP}_{Ai}})^2 \\
\hat{\sigma_B^2} &= \frac{1}{N-n}\sum_{i=1}^{n}\sum_{r=1}^{R}(\text{MP}_{Bir} - \overline{\text{MP}_{Bi}})^2
\end{aligned} \tag{2.19}$$

Note that $\overline{\text{MP}_{Ai}}$ and $\overline{\text{MP}_{Bi}}$ are the means of replicates within clinical sample i concerning our specified measurement procedures. The variability between replicates should reflect the precision of the measurement procedures solely. Equation (2.19) requires several (≥ 2) replicates, or else there is little or nothing to say about the variation within the samples. After estimating the variance concerning the two error terms and apply them to derive the estimate of λ, we can estimate the slope and intercept using Equation (2.16) by substituting λ with $\hat{\lambda}$:

$$\begin{aligned}
b_1 &= \frac{S_{AA} - \hat{\lambda}S_{BB} + \sqrt{(S_{AA} - \hat{\lambda}S_{BB})^2 + 4\hat{\lambda}S_{BA}^2}}{2S_{BA}} \\
b_0 &= \overline{\text{MP}_A} - b_1 \cdot \overline{\text{MP}_B}.
\end{aligned} \tag{2.20}$$

Accordingly, our estimated linear Deming regression model is given by

$$\hat{g}(\text{MP}_{Bi}) = b_0 + b_1 \cdot \text{MP}_{Bi}. \tag{2.21}$$

Since variability is accounted for in both x and y-direction, it may seem like a challenging job constructing prediction bands for the fitted regression lines. Luckily, it is possible to construct these uncertainty bands analytically. Components from [Fuller, 2009] are applied to construct the 99% prediction bands for our fitted Deming regression models. For conventional purposes, we define

$$\mathbf{MP}_B = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \text{MP}_1^* & \text{MP}_2^* & \dots & \text{MP}_M^* \end{bmatrix}^T \tag{2.22}$$

with $\{\text{MP}_j^*\}$ theoretical predicted values, where

$$\mathbf{MP}_B^* = \begin{bmatrix} \text{MP}_1^* & \text{MP}_2^* \dots \text{MP}_M^* \end{bmatrix}. \tag{2.23}$$

Furthermore we define $\mathbf{I}$ as the $M \times M$ identity matrix and $\mathbf{V}$ as the estimated covariance matrix for b_0 and b_1.

$$\begin{aligned} \text{Var}(\hat{g}(\widehat{\mathbf{MP}_B})|\mathbf{MP}_B^*) &= diag\Big(\mathbf{MP}_B^T \mathbf{V} \mathbf{MP}_B\Big) + (\hat{\sigma}_A^2 + b_1^2 + \text{Var}[b_1] \cdot \hat{\sigma}_B^2)\mathbf{I} \\ &= (a_{ij}) \in \mathbb{R}^{M \times M}. \end{aligned} \tag{2.24}$$

Lastly, we define

$$\begin{aligned} \text{SD}(\hat{g}(\widehat{\mathbf{MP}_B})|\mathbf{MP}_B^*) &= (\sqrt{a_{ij}}) \in \mathbb{R}^{M \times M} \\ \mathbb{1}_{M \times 1} &= \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T, \end{aligned} \tag{2.25}$$

which provide the estimated 99% prediction bands:

$$\text{PI} = \hat{g}(\mathbf{MP}_B^*) \pm t_{0.995,n-2} \cdot \text{SD}(\hat{g}(\widehat{\mathbf{MP}_B})|\mathbf{MP}_B^*)\mathbb{1}_{M \times 1}. \tag{2.26}$$

The estimated ratio of variances, λ, will include the risk of sparse estimation if the advised assumptions are not met regarding our samples; the number of clinical samples, n, should ideally be more extensive than 50 [Gillard, 2010]. The real data sets presented in Chapter 3 typically contain between 20 and 25 clinical samples. Henceforward, there is still room for enhancement concerning the study design. The obvious solution to this problem is to adjust the study design suitably. However, it is challenging and costly to get hold of adequately many clinical samples for our studies. It is more effortless to increase the number of replicated measurements because replicated measurements come from the same clinical sample.

From the fact that $n \geq 50$ is unsatisfied for most instances, it will be essential to review if any other estimators for β_0 and β_1 is more appropriate than Equation (2.20). If β_0 is taken for granted to be known, [Gillard, 2010] debates for the use of

$$b_1 = \frac{\overline{\text{MP}_A} + \beta_0}{\overline{\text{MP}_B}} \tag{2.27}$$

as the slope estimator, directly deduced from the second equation in Equation (2.20). When comparing measurement procedures, β_0 is intuitively expected to be zero. However, this is not certainly supported statistically because significant non-zero intercepts are obtained in most Deming models. Typically, dependent on the clinical samples, few or none measurements

lie near zero. Consequently, there is no trustworthy evidence that $\beta_0 = 0$ is appropriate to assume in the light of commutability assessment. Nevertheless, a sufficiently large sample size typically diminishes the magnitude of the bias of Equation (2.27).

[Gillard, 2010] formulates that there exist two additional alternative estimators. These two estimators rely on knowledge of observed variability in x- or y-direction. Firstly, if the information on σ_A^2 is satisfactory, the regression slope coefficient may be estimated by

$$b_1 = \frac{S_{AA} - n \cdot \sigma_A^2}{S_{BA}}. \tag{2.28}$$

Secondly, holding sufficiently enough information on σ_B^2 permits estimation of the slope estimator by

$$b_1 = \frac{S_{BA}}{S_{BB} - n \cdot \sigma_B^2}. \tag{2.29}$$

These estimators are suitable if possessing prior information on one of the measurement procedures. Providentially, we can estimate one of the measurement uncertainties and use Equation (2.28) or Equation (2.29) to estimate the slope using replicated measurements. Nevertheless, [Dunn, 1989] affirms that applying Equation (2.29) implements a more reliable slope estimator opposed to Equation (2.28). However, the latter slope estimates should be treated with caution. Usually, for small sample sizes (typically $n < 40$), Equation (2.28) is underestimated, whereas Equation (2.29) is overestimated. If prior information on the analytical variability is negligible, it appears unwise to utilize two latter estimators. The conclusion is that the three alternative estimators for the regression slope are unlikely superior to the original one; thus, Equation (2.20) will be used in the rest of this text.

2.6 Parametric regression splines

A single regression line might not be enough to describe the entire pattern provided by the clinical samples. In these unfortunate cases, we have an opportunity to partition our data-set into intervals and fit regression lines for these sections separately. In other words, we have more than one regression line as part of our model. We must, nevertheless, demand that the fitted lines connect such that the overall curve is continuous. Also, we impose that the first and second derivatives of the model be continuous for a smooth fit. The locations where the regression lines connect are called *breakpoints* or *knots*. Splines are the general name of this regression approach. A piece-wise regression like this is appropriate when modeling abrupt changes in the relationship between predictors and response, that is if relationships between measurement procedures suddenly shift drastically. Even when relationships alternate slightly,

it could be advantageous to use splines in favor of a global fit. This is typical because splines are more flexible and do not require a model fitted by an enormous polynomial degree [James, Witten, Hastie, and Tibshirani, 2013]. Generally, if the two comparable measurement procedures exhibit distinct relationships concerning different clinical samples' measurement values, regression splines will be competent. [James et al., 2013] defines the theoretical model for regression splines model (with K knots and polynomial degree j) by

$$\text{MP}_{Aik}|\text{MP}_{Bi} = \sum_{l=0}^{j} \beta_{lk} \text{MP}_{Bi}^{l} + \epsilon_i, \tag{2.30}$$

with $k \in \{1, 2, \ldots, K\}$ and $i \in \{1, 2, \ldots, n\}$. Fixing a clinical sample, e.g, $i = I$ and considering $k = \kappa$, Equation (2.30) produces the j degree polynomial fitted value in the region between the knots $\kappa - 1$ and κ. We impose that

$$\epsilon_i \sim N(0, \sigma_A^2), \ i \in \{1, 2, \ldots, n\}. \tag{2.31}$$

The error terms are independent and identically distributed. Thus, we assume homoscedasticity and independence as before. We also require linearity in the regression coefficients, as usual. Later we will try to avoid assumptions regarding distributions such as the normal-assumption presented in Equation (2.31). However, choosing to constrain ourselves with the normality assumptions, estimates of prediction errors, model errors, and more are provided for free in the specified regions. For the model fitting in this text, we apply polynomial regression with degree $j = 3$ (natural cubic model fits) between successive knots, estimated with the ordinary least squares method. The linear model assumptions are obliged to be satisfied, therefore we may apply Equation (2.3). Note that we require the linear model assumptions to be true for every single fitted regression line. The partition of the sample space implies that we have fewer observations for the local fits. Consequently, the test statistics for the linear model assumptions may be wrongly estimated. The *Splines* packages will be used in R. Particularly; we will use *lm()* with *ns()* to fit natural cubic splines models.

Chapter 3

Applying Commutability Assessment with Parametric Methods

We have been through some of the commutability assessment methods used today. The theory and possible faults of every method have been discussed. Accordingly, it will be interesting to discuss each method's actual performance using real data gathered by Noklus or simulation. See Section 3.6 and Section 5.5 for the latter.

3.1 The data sets

In this section, five different data sets will be considered. Table 3.1 summarizes the collection of data sets.

The data sets are similar, and every one of them is built up as presented in Table 3.2.

Every data set includes 25 clinical samples and five control material samples with three replicates. All but one of the data sets contain NA-values, which results in fewer clinical samples. To get a conception of how the methods presented in Chapter 2 work in practice, they

Table 3.1 *– Here is a rough description of the real data sets used in this text. MP is shorthand for measurement procedure.*

Name	Analyte	Measurement units	Number of MP's
EPK	Erythrocytes	$10^12/\mathrm{L}$	4
HB	Hemoglobin	g/dl	4
LPK	Leukocytes	$10^9/\mathrm{L}$	4
MCV	Mean Erythrocyte Volume	fl	4
TPK	Thrombocytes	$10^9/\mathrm{L}$	4

Table 3.2 *– This table describes the structure of the data sets. PS is shorthand for patient samples or clinical samples, and CM is shorthand for control materials. The specified numbers following PS are the particularized clinical samples that are manifested as NA-values.*

Data set	Number of PS's	Number of CM's	NA samples
EPK	25×3	5×3	PS: 1, 7
HB	25×3	5×3	PS: 1, 15
LPK	25×3	5×3	PS: 15, 19, 20
MCV	25×3	5×3	PS: 4, 16
TPK	25×3	5×3	None

will be tested separately and as part of a step-wise algorithm where the most suitable method is selected. The commutability assessment approaches are first considered separately, where shortcomings and strengths are discussed by turn. This design is executed by investigating the *MCV* data set, which resembles a relatively large variability in the clinical samples. An offset plot with all measurement procedure comparisons of the clinical samples is presented to get a rough idea of how the data set looks. See Figure 3.1.

Note that the offset plot also includes estimated Pearson correlation coefficients, which is interpreted as the strength of the two particularized measurement procedures' linear relationship. However, this estimator states nothing regarding the comparability of the measurement procedures [Bland and Altman, 2010]. Besides, the estimated densities for every measurement procedure comparisons are included and are at least approximately gaussian.

3.2 Commutability assessment with Ordinary least squares regression

Before going forth with the commutability assessment using the ordinary least squares estimator, it is vital to test the model residuals regarding the normality assumption. The densities displayed in Figure 3.1 suggests that the clinical sample measurements are approximately normally distributed. Even though the observations seem normally distributed, it is good practice to test the normality requirement for the model residuals. It is not straightforward to state something substantial regarding the rest of the linear model assumptions from Figure 3.1. Thereupon, the residuals vs. fitted plots constructed using ordinary least regression models are presented. The residual vs. fitted plots explain what to expect from the formal evaluation tests performed subsequently. See Figure 3.2. There are no apparent indications that we are dealing with heteroscedasticity, dependent error terms, or non-linearity in Figure 3.2.

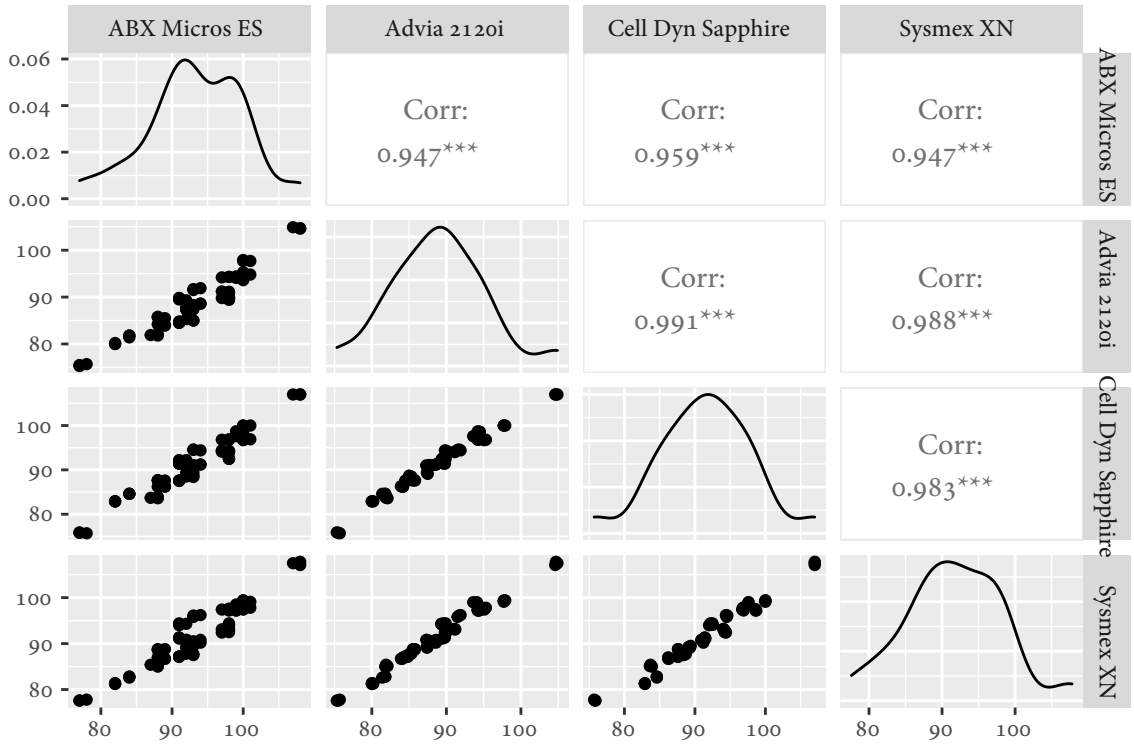

Figure 3.1 *– An offset plot that presents a full overview of the patterns of the clinical samples in the MCV data set. Note that density estimates are implemented as well as Pearson correlation coefficients with corresponding test significances.*

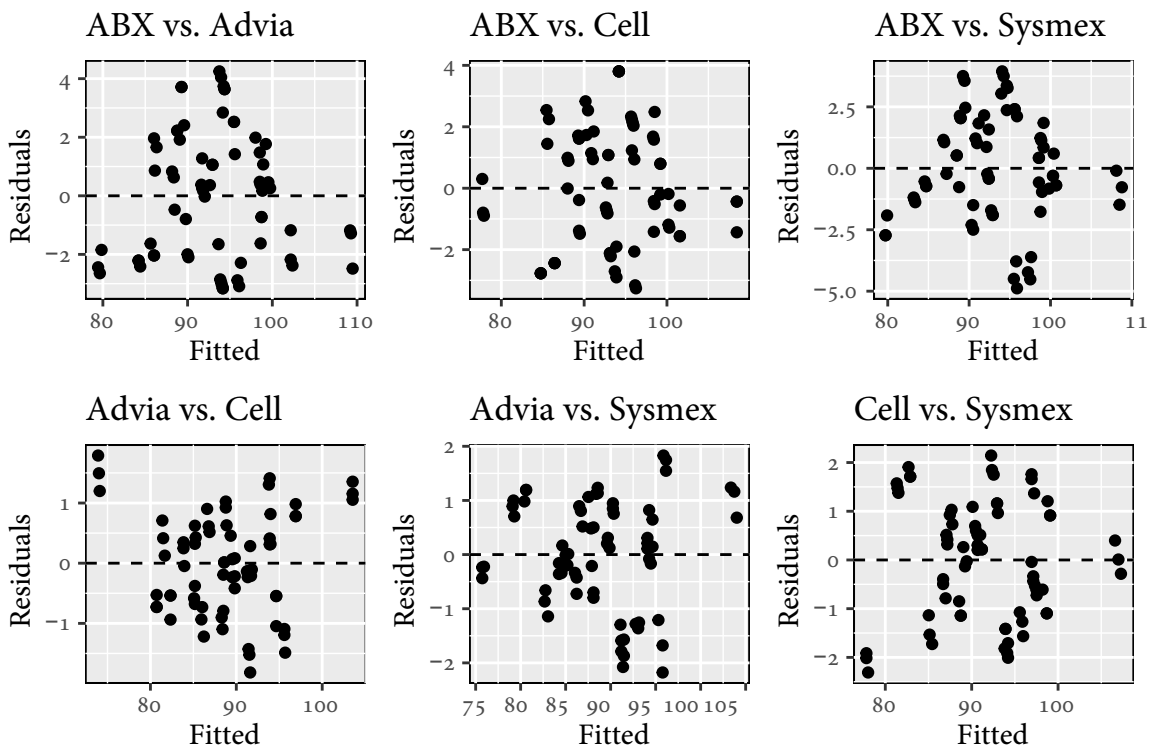

Figure 3.2 *– Residual vs. fitted plots for all six ordinary least squares regression models.*

However, to get an objective interpretation, the formal tests described in Section 1.4 will be used. Table 3.3 outlines the outcomes of the linear model assumption tests.

From Table 3.3, one observes that auto-correlation is a significant problem. Again, this is not uncommon because replicated measurements are utilized in every sample, which induces auto-correlation. Thus, using the mean of replicated measurements is the most natural way to obtain trustworthy estimators. In the rest of this chapter, only MOR will be used because of the persistent auto-correlation issue. The only place we have practical use of the replicates is when estimating the ratio of variances, λ, for Deming regression. However, the mean of replicates is also the best guess on the true value of the analyte. See Figure 3.3 for residuals vs. fitted plots for the analogous MOR-models: There is no apparent difference among the two sets of residual plots than, of course, fewer points. However, the linear model assumptions are now satisfied. See Table 3.4.

Accordingly, one might proceed with the commutability assessment since all the linear model assumptions are satisfied for all the measurement procedure comparisons. Consequently, the combination of MOR and ordinary least squares regression is adequate as part of the commutability assessment approach. The prediction bands' estimation is done by implementing Equation (2.3). The commutability assessment plots are exhibited in Figure 3.4 and

Table 3.3 *– Linear model assumptions of ordinary least squares with AR. The green fields are accepted linear model assumptions, whereas the red fields are discarded linear model assumptions. The stars signify the degree to which the assumptions are rejected.*

Linear model assumptions	*MCV*			
OLSR + AR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i	**		***	
ABX Micros ES vs. Cell Dyn Sapphire	*		***	
ABX Micros ES vs. Sysmex XN			***	
Advia 2120i vs. Cell Dyn Sapphire			***	
Advia 2120i vs. Sysmex XN			***	
Cell Dyn Sapphire vs. Sysmex XN	*	**	***	

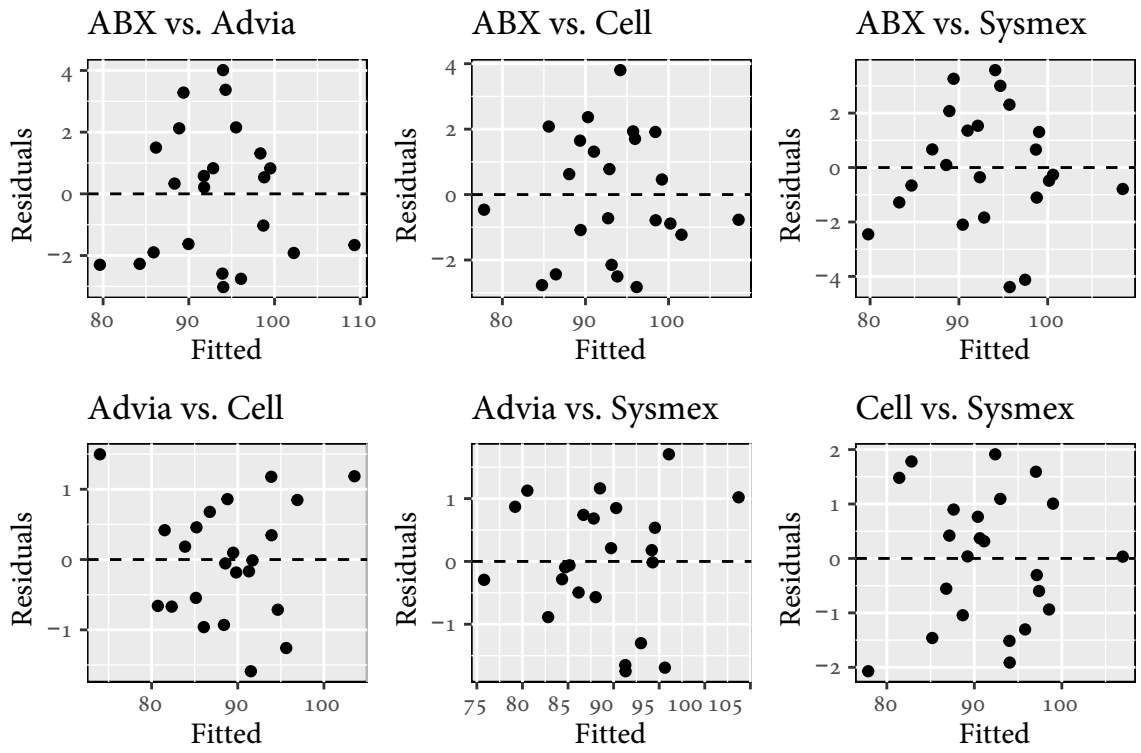

Figure 3.3 *– Residual vs. fitted plots for all six ordinary least squares regression models employing the mean of the replicates.*

Table 3.4 *– Evaluation of the linear model assumptions of ordinary least squares constructed by means of replicates.*

Linear model assumptions	*MCV*			
OLSR + AR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i				
ABX Micros ES vs. Cell Dyn Sapphire				
ABX Micros ES vs. Sysmex XN				
Advia 2120i vs. Cell Dyn Sapphire				
Advia 2120i vs. Sysmex XN				
Cell Dyn Sapphire vs. Sysmex XN				

Figure 3.5. The control material samples with yellow fill are concluded as non-commutable. By looking at Figure 3.4 and Figure 3.5, it is clear that *LQ* is concluded as non-commutable for five of the six measurement procedure comparisons. Besides, *2CPD* is concluded as not commutable in the measurement procedure comparison between *Advia 2120i* and *Cell Dyn Sapphire*. The ordinary least squares approach appears to be appropriate concerning the MCV data set. However, variability in the x-direction is neglected, which is the main fault of this evaluation method.

The log-log and Bland-Altman transformations combined with ordinary least squares regression as part of the commutability assessment design are presented in the next section.

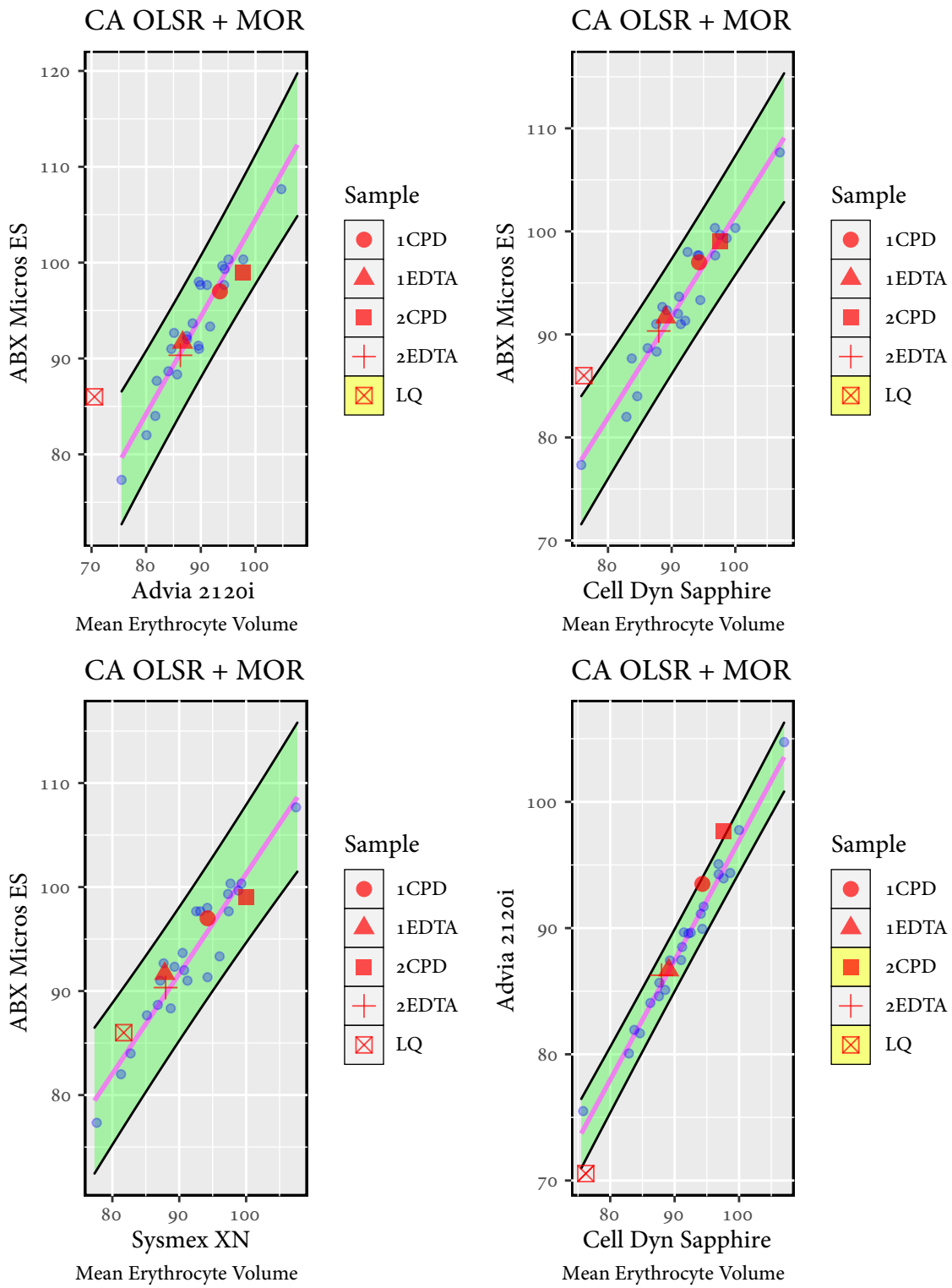

Figure 3.4 *– Visual commutability assessment evaluation using ordinary least squares regression constructed by mean of replicates.*

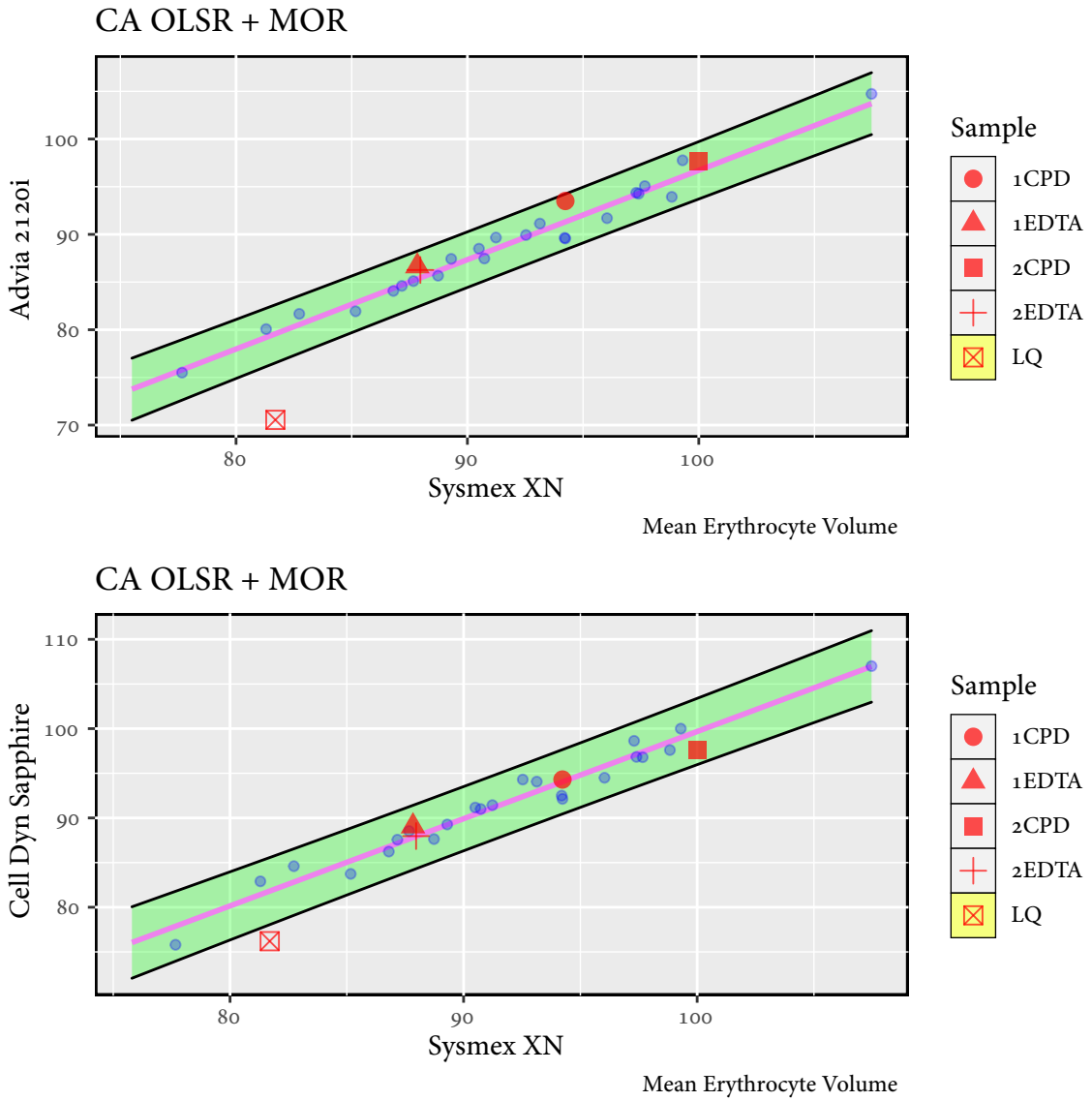

Figure 3.5 *– Visual commutability assessment evaluation using ordinary least squares regression constructed by mean of replicates.*

3.3 Commutability assessment with transformation combined with linear and polynomial regression

We observed that the ordinary least squares regression worked fine on the MCV data set. The linear model assumptions were satisfied for the replicates' mean, and it was appropriate to apply the ordinary least squares regression model as part of the commutability assessment. Nevertheless, the MCV data set is re-evaluated by utilizing the two transformation approaches as part of the commutability assessment. An offset plot consisting of measurement procedure comparisons of the clinical samples that are log-log transformed is presented:

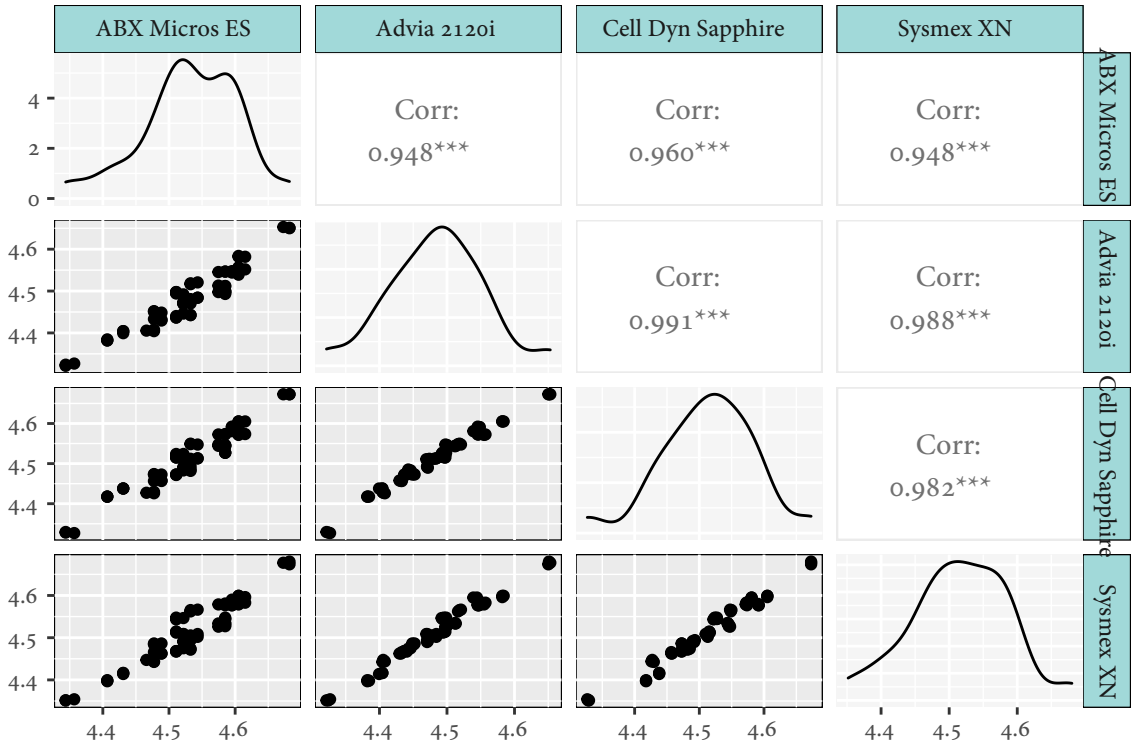

Figure 3.6 *– An offset plot, which represents an overview of the log-log transformed patterns of the clinical samples.*

There are no apparent departures from linearity observed in Figure 3.6. The resulting scatter plots appear quite similar in shape to the raw data, which implies that everything seems adequate. Likewise, an offset plot for the Bland-Altman transformations is given by;

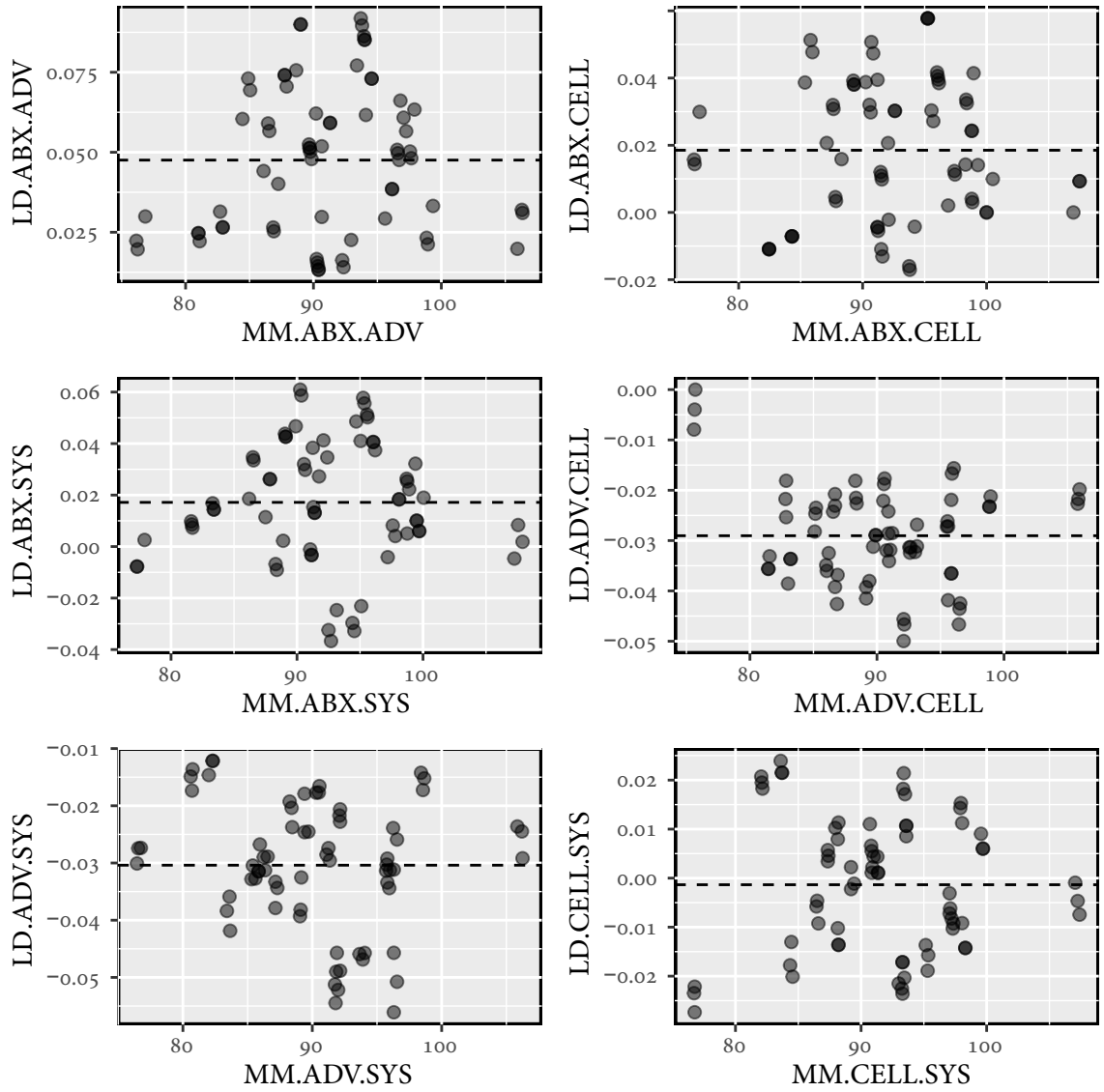

Figure 3.7 *– An offset plot, with an overview of the Bland-Altman-transformed patterns of the clinical samples. LD and MM is shorthand for logarithmic differences and means of measurement procedures, respectively.*

There is no pressing deficiencies or obvious non-linearity issues, concluded from Figure 3.7. Hence, it appears unnecessary to use polynomial regression on these Bland-Altman transformed clinical samples. The dashed horizontal lines follow the Bland-Altman transformed clinical samples adequately. Residual vs. fitted plots for both log-log and Bland-Altman models are constructed below. Recollect that we will use MOR here, as illustrated as good practice in the previous section.

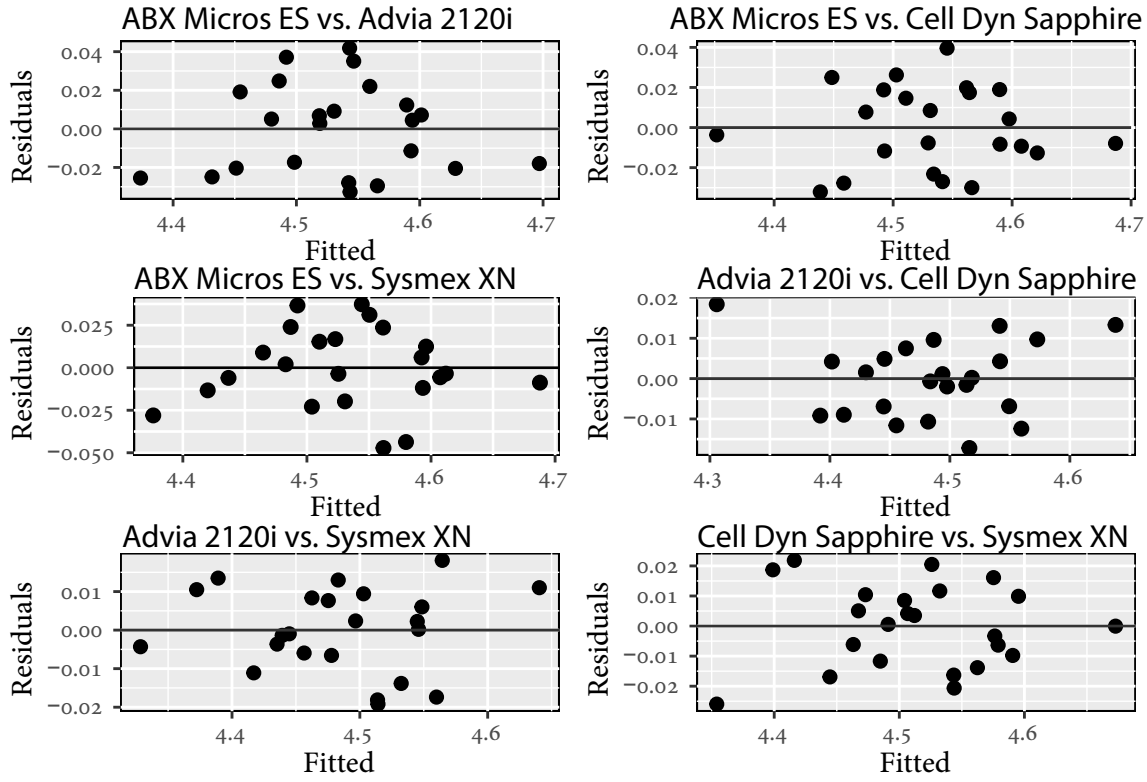

Figure 3.8 *– Residual versus fitted plots for all six proposed ordinary least squares models constructed by the log-log transformed clinical samples—no clear evidence of the residuals' heteroscedasticity or dependence are present.*

Table 3.5 *– Linear model assumptions of log-log transformed clinical samples fitted by ordinary least squares. The mean of replicates is applied.*

Linear model assumptions	*MCV*			
LLT + OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i				
ABX Micros ES vs. Cell Dyn Sapphire				
ABX Micros ES vs. Sysmex XN				
Advia 2120i vs. Cell Dyn Sapphire				
Advia 2120i vs. Sysmex XN				
Cell Dyn Sapphire vs. Sysmex XN		*		

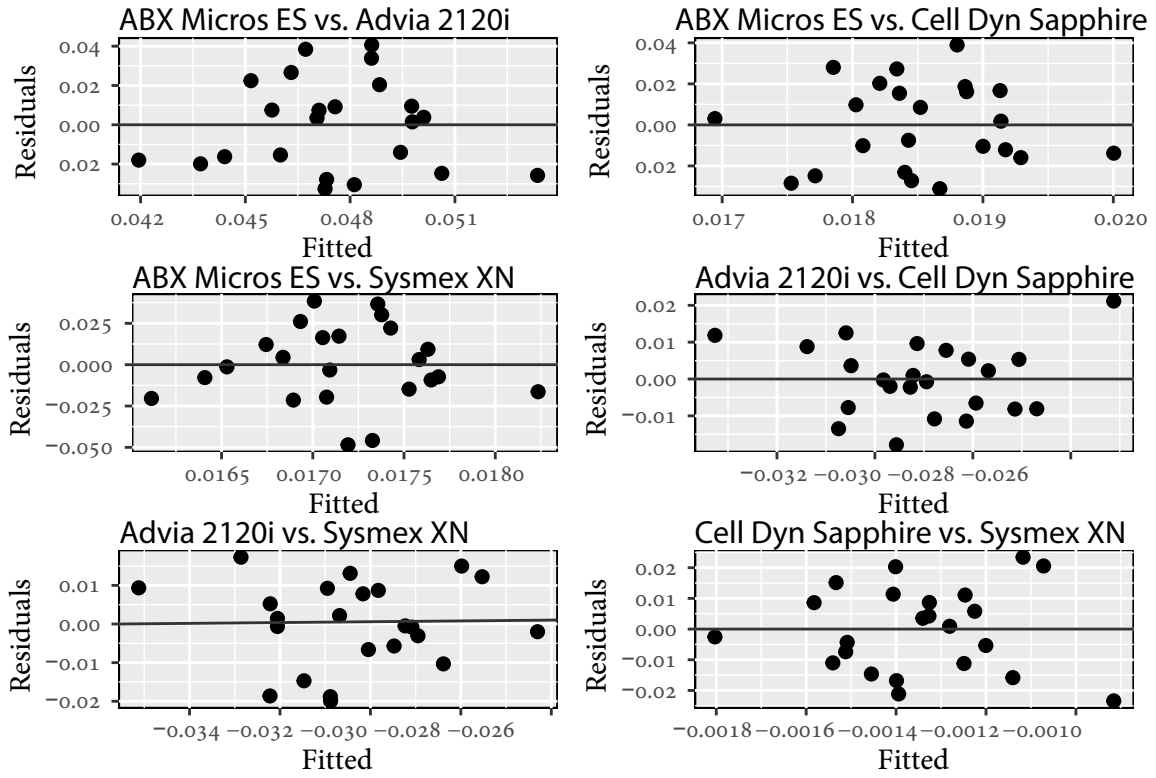

Figure 3.9 *– Residual versus fitted plots for all six proposed ordinary least squares models constructed by the Bland-Altman-transformed clinical samples. There is no clear evidence of heteroscedasticity or dependence of the residuals.*

The linear model assumption tests outlined in Section 1.4 are applied, and their results are displayed in Table 3.5.

Homoscedasticity is not satisfied with the two last measurement procedures compared. Therefore, parametric methods, such as ordinary least squares or Deming regression, may

not be used to evaluate the control materials. At least not with log-log- and Bland-Altman transformed clinical samples. Nevertheless, as formulated in the preceding section, ordinary least squares of the raw data could be applied, considering that this led to accepted linear model assumptions. Firstly, commutability assessment will be executed by using log-log transformation with ordinary least squares. Secondly, Bland-Altman transformation combined with ordinary least squares is utilized.

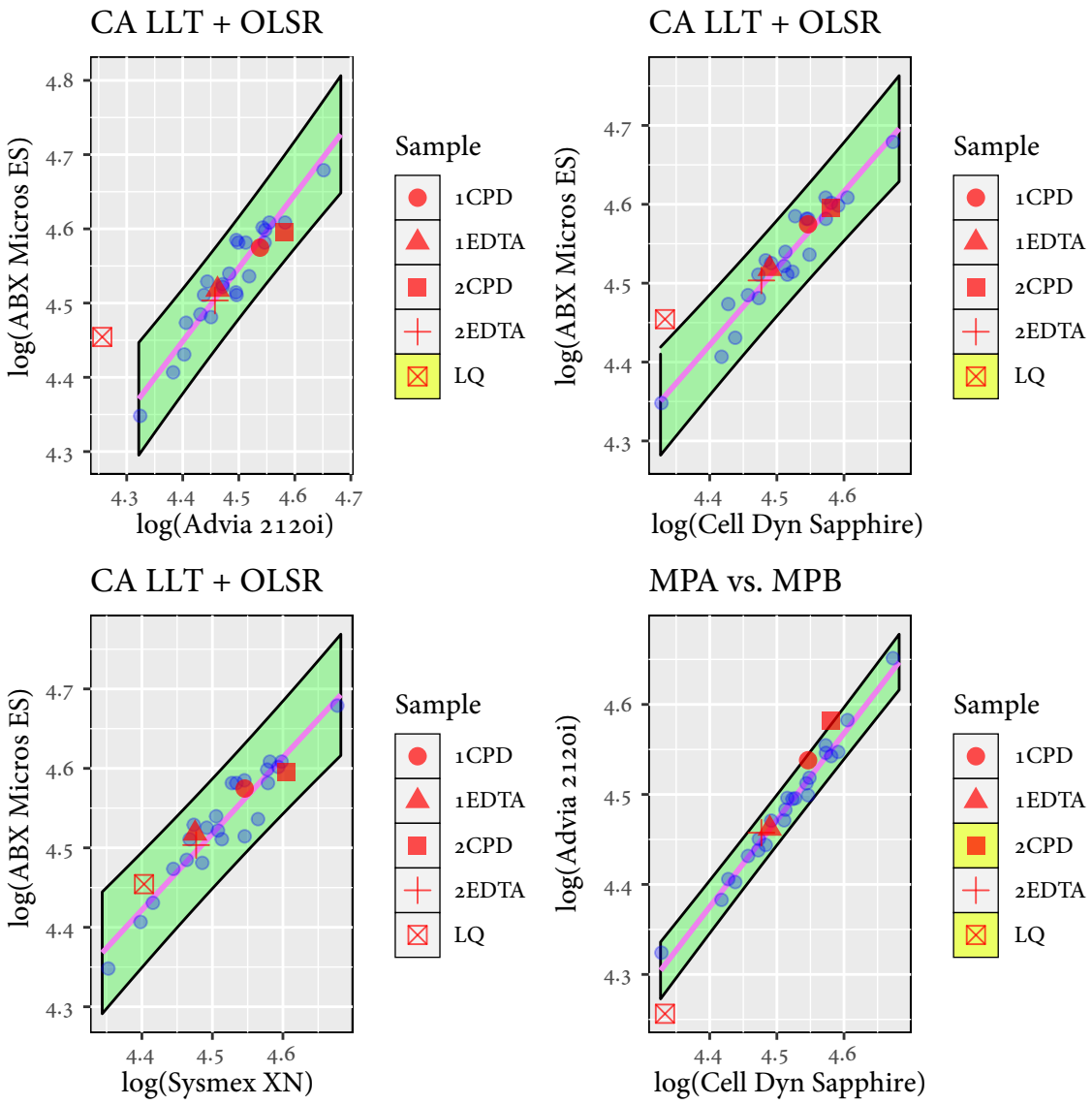

Figure 3.10 *– Visual commutability assessment using log-log transformed clinical samples combined with ordinary least squares regression. The means of replicates are used.*

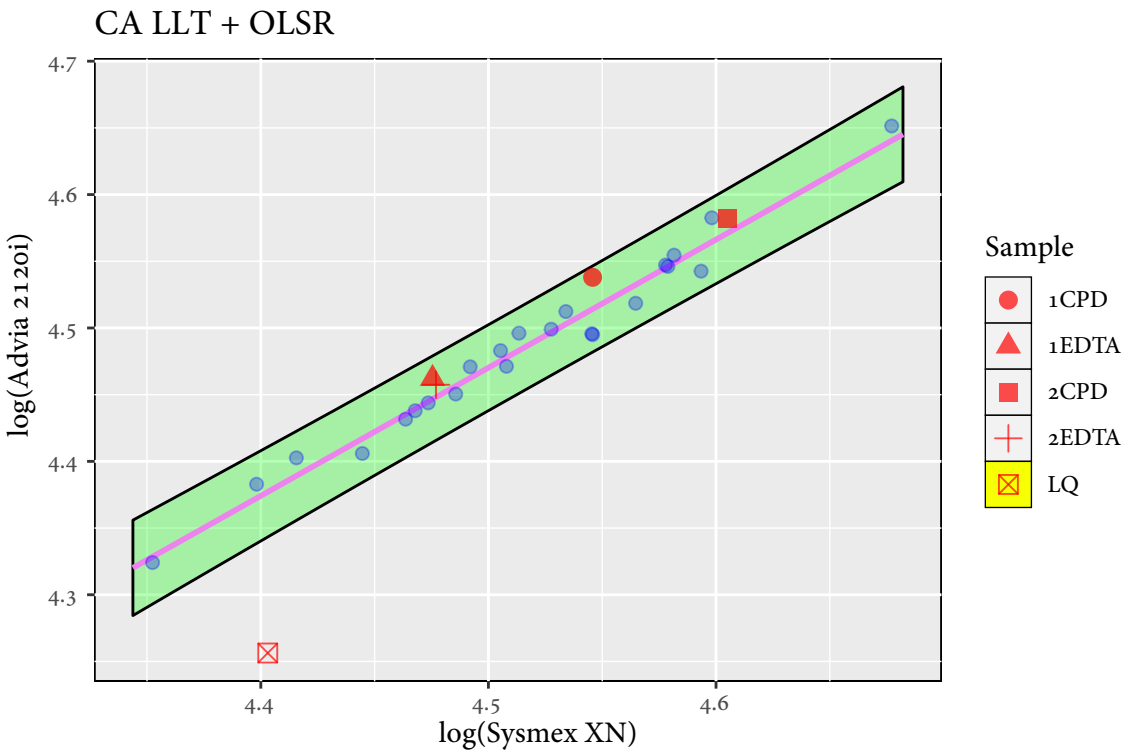

Figure 3.11 *– Visual commutability assessment using log-log transformed clinical samples combined with ordinary least squares regression. The means of replicates are used.*

As observed in Figure 3.10 and Figure 3.11, the same conclusion is drawn regarding commutability acceptance of the control materials, as with raw clinical samples combined with ordinary least squares regression. The Bland-Altman transformation with ordinary least squares regression could preferably be applied. Considering that the Bland-Altman transformation reduces the x-direction variability, this transformation might be slightly more realistic and favorable. In this case, $p = 1$ is selected.

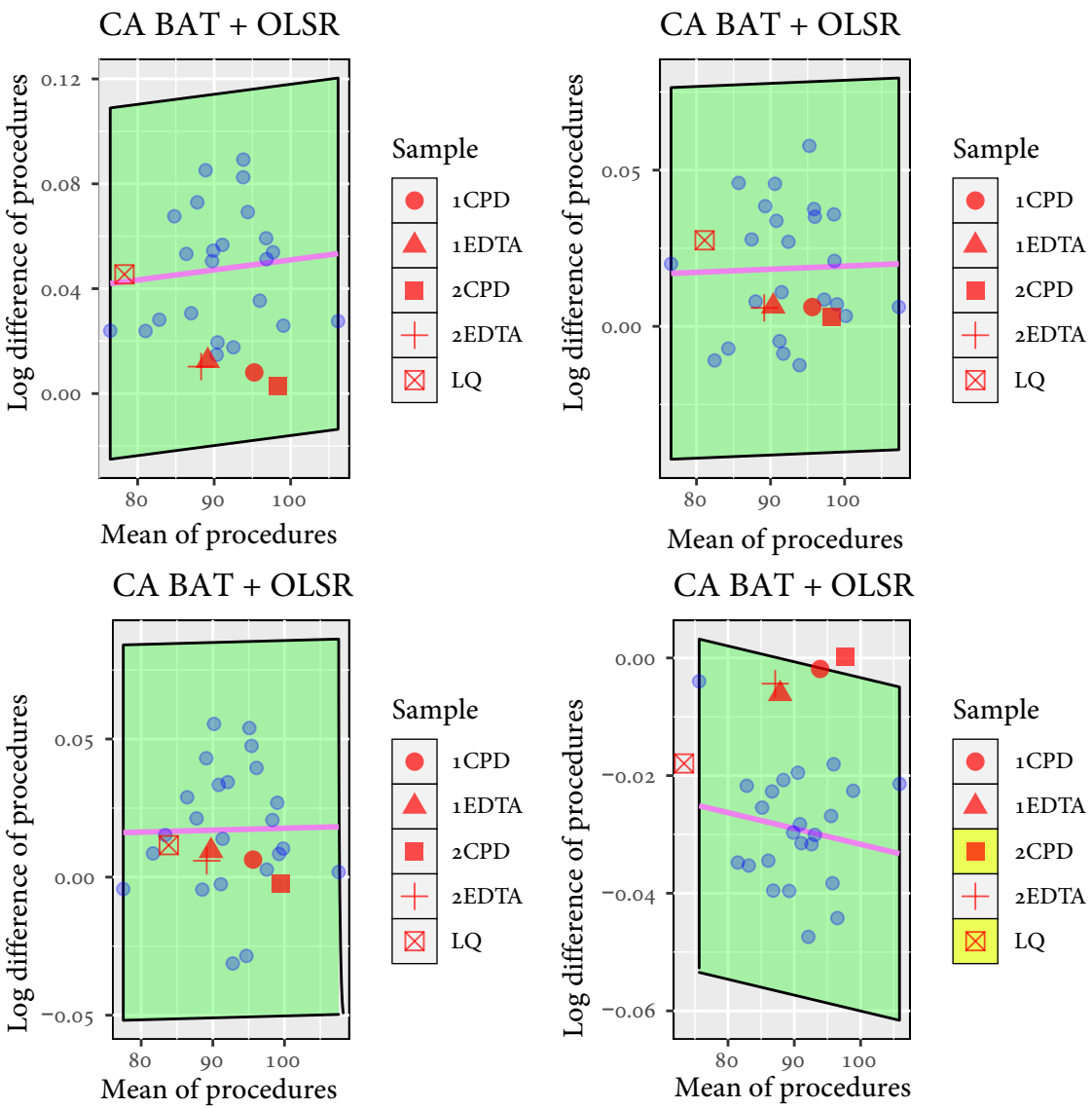

Figure 3.12 *– Visual commutability assessment using Bland-Altman-transformed clinical samples combined with ordinary least squares regression. The means of replicates are used.*

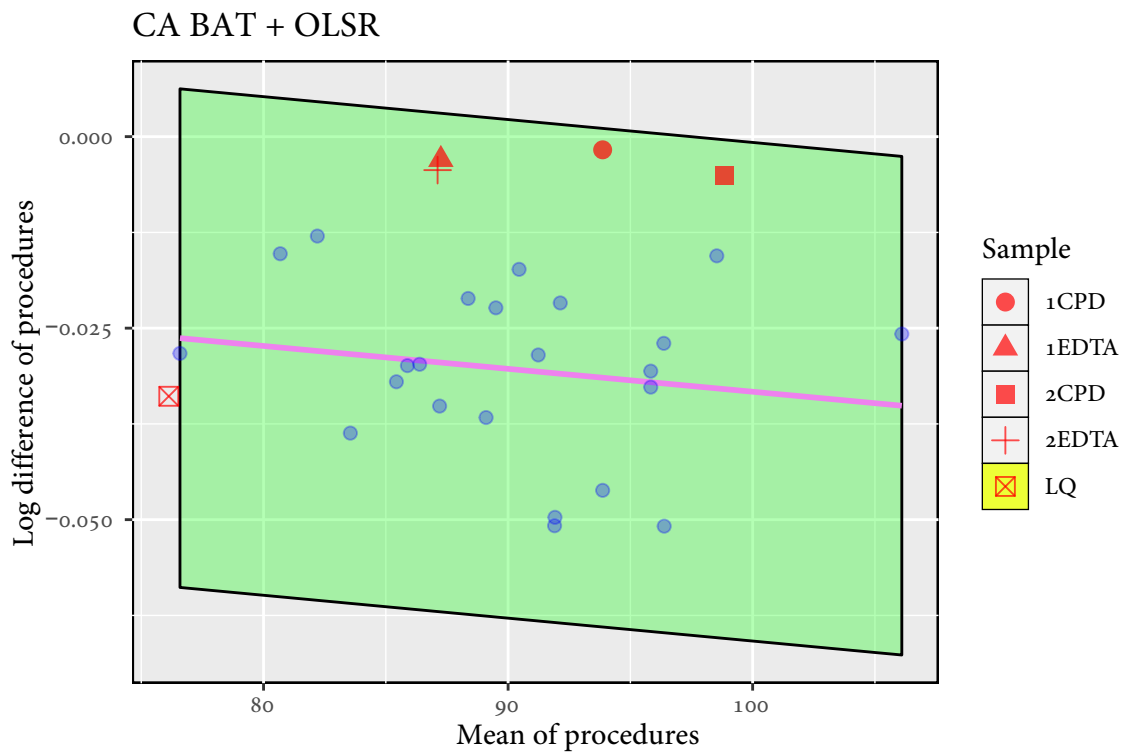

Figure 3.13 *– Commutability assessment using Bland-Altman transformed clinical samples and ordinary least squares regression*

Regard that the Bland-Altman transformation linked with ordinary least squares regression concluded differently to what log-log transformation did. The basis for this may be due to the Bland-Altman transformation inhibits relatively larger prediction errors because of the lack of statistical significance for the estimated slope coefficients. Only the intercepts showed statistically significant. Therefore the bands must be too large here. Simultaneously, as [Bland and Altman, 2010] describes, the prediction bands are underestimated because the replicates' mean is used, implying that the prediction bands should be even more expansive than proposed.

3.4 Commutability assessment with Deming regression or Parametric Regression splines

When the linear model assumptions are satisfied for the raw data, the ordinary least squares regression estimator was concluded to be appropriate as part of the commutability assessment with the *MCV* data set. However, as discussed in Section 3.2, the ordinary least squares regression model is not estimated accounting for the variability in x-direction. Therefore

Deming regression could have been applied because Deming regression relies on the same linear model assumptions as ordinary least squares. As proposed, these assumptions were satisfied for the ordinary least squares estimator. The calculation of the Deming regression coefficients is estimated by Equation (2.20). The resulting assessment plots are displayed in the two succeeding figures below.

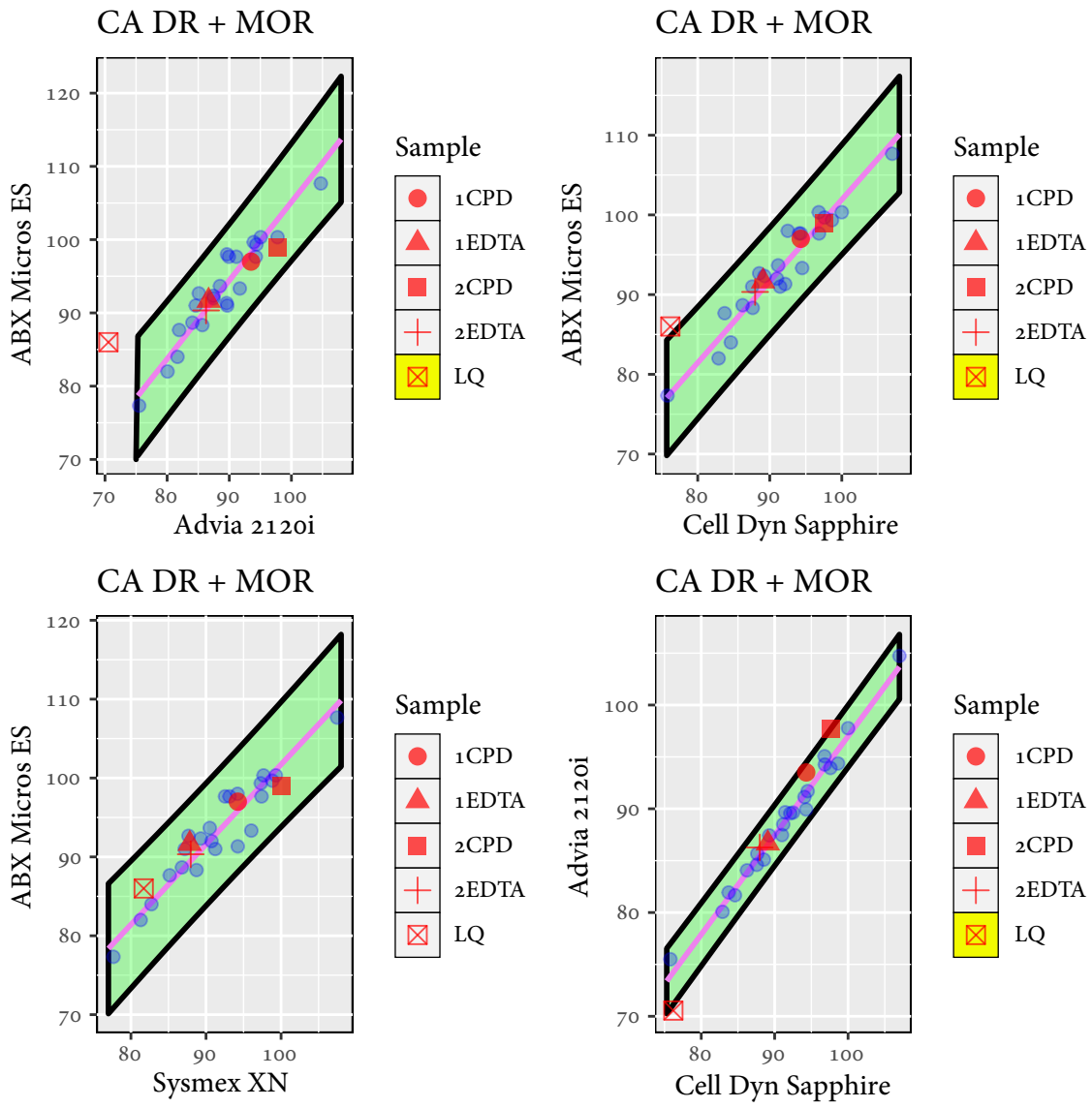

Figure 3.14 *– Commutability assessment using Deming regression fitted by the mean of the replicates.*

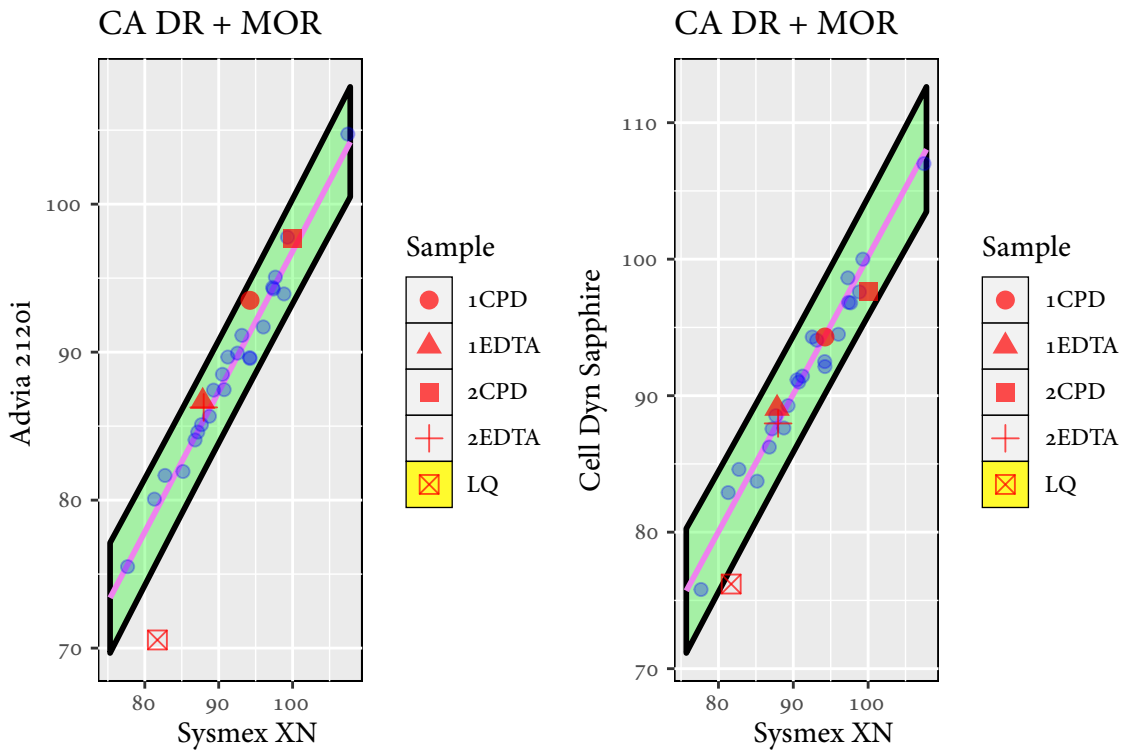

Figure 3.15 *– Commutability assessment using Deming regression fitted by the mean of the replicates.*

Remark that the control material *2CPD* now is commutable regarding our acceptance criterion. This fact must imply that the *Advia 2120i* vs. *Cell Dyn Sapphire* comparison's prediction bands are more extensive than for the ordinary least squares estimator. The enlargement source is most likely linked with the inclusion of two-dimensional variability, which is directly deduced from Equation (2.26). The regression coefficients estimators Equation (2.20) are probably overestimated because of too few clinical samples and replicates in our data sets. If the latter is the primary source of enlargement, it would arguably be better to use ordinary least squares regression. In Section 3.6, it is proposed that the ratio of variances is approximately overestimated by $6 - 7\%$ concerning our data sets. False-positive classification of the commutability property may have drastic consequences, as formulated in Chapter 1. As part of the commutability assessment, parametric regression splines may meet the corresponding linear model assumptions. In that instance, the following commutability assessment plots are provided:

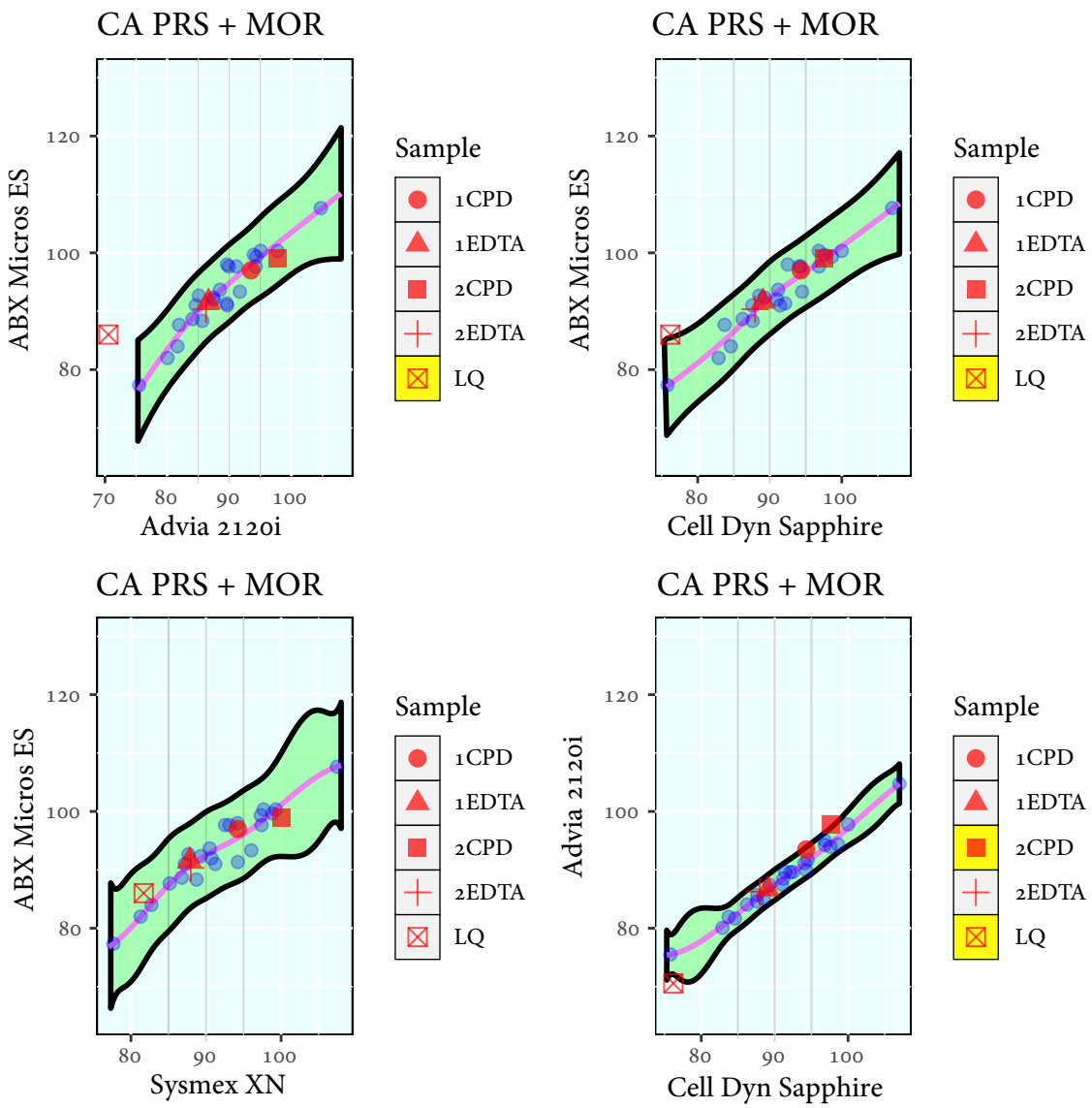

Figure 3.16 *– Visual commutability assessment using parametric regression splines. Natural cubic splines are used as the basis, and the mean of the replicated measurements are employed to construct to model. The gray lines are the locations of the three knots. Three knots implies that we get four models across the global range.*

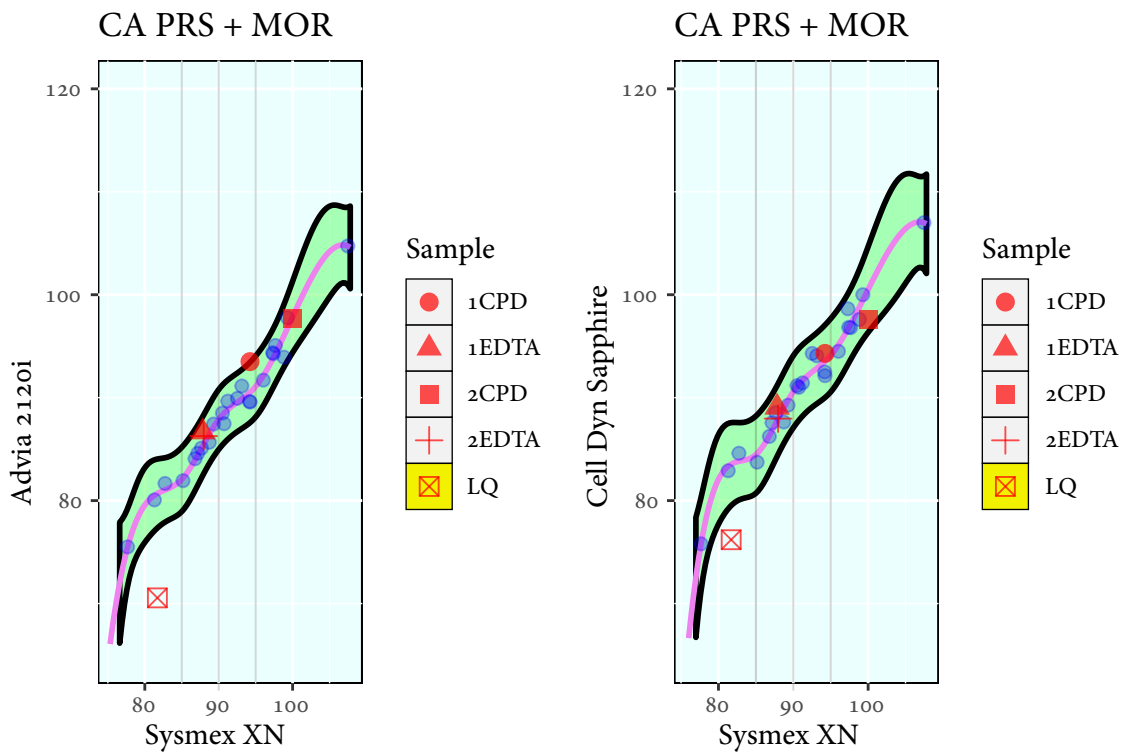

Figure 3.17 *– Commutability assessment using parametric natural cubic splines with mean of replications.*

The same conclusions are drawn regarding commutability, as for ordinary least squares when parametric regression splines are used. Compared to the Deming regression approach's outcomes, the only difference is that *2CPD* was accepted as commutable. In contrast, it is not accepted for parametric regression splines and the ordinary least squares approach.

3.5 Step-wise commutability assessment with parametric methods

With all methods discussed in magnificent detail, it is time to evaluate control materials as part of a step-wise algorithm. A data set containing strong linear relationships between the measurement procedures are considered. The *EPK* data set is such a data set. The strength of the linear relationships regarding the different measurement procedure comparisons is powerful. After investigating the data set visually, and looking into the linear model assumptions, the most reliable assessment method for commutability is applied. As given above, the analyte units in the *EPK* data set are the number of erythrocytes per liter. Measurements of the true quantity within the clinical samples are expected to lie between $4 \cdot 10^{12}/L$ and $6 \cdot 10^{12}/L$.

The *EPK* data set contains 25 clinical samples and five control material samples. Because of NA-values, two of the clinical samples (1 and 7) are dropped, which results in $n = 23$ remaining clinical samples. As with the other data sets, *EPK* contains the four measurement procedures:

- ABX Micros ES
- Advia 2120i
- Cell Dyn Sapphire
- Sysmex XN

An offset plot investigates the relationships of the different measurement procedures and is consequently included. The estimated correlation coefficients with corresponding significance degrees are included as well as the density estimates.

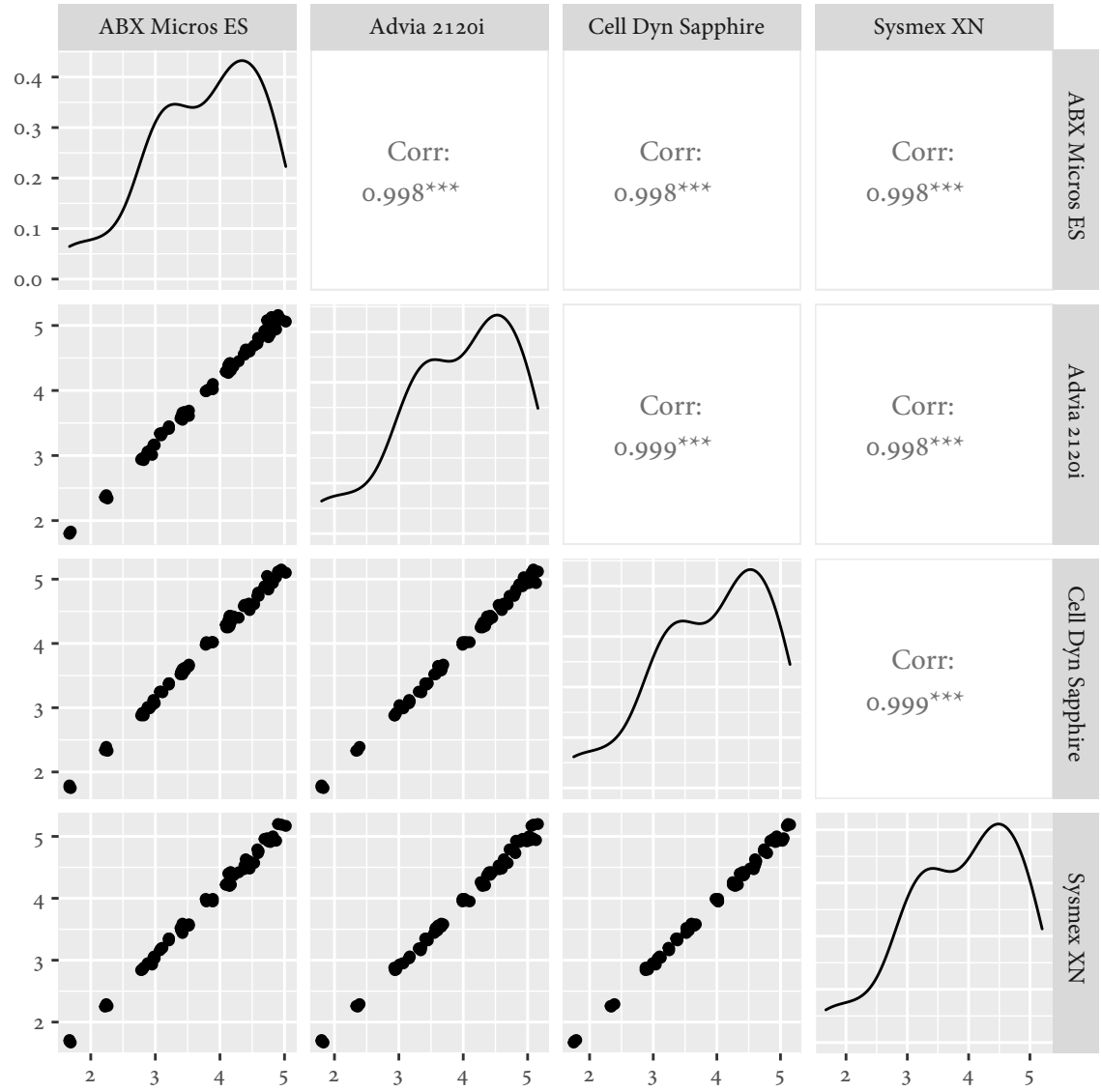

Figure 3.18 *– An offset plot that is visually describing the six relationships between the four specified measurement procedures in the EPK data set. Density estimates and Pearson correlations are also implemented.*

As observed in Figure 3.18, there seems to be a close to perfect linear pattern among every measurement procedure. The Pearson correlation coefficients confirm this by estimates near one and three stars of significance degree. Thus, linearity is definitively provided here. Furthermore, if the rest of the linear model assumptions are satisfied, parametric estimators may be used as part of the commutability assessment. The formal tests from Section 1.4 are applied to appraise the linear model assumptions. Linearity is evaluated by the glvm() function in R. The results are given in

Linear model assumptions	*EPK*			
OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX vs. Advia		**		
ABX vs. Cell		**		
ABX vs. Sysmex		**		
Advia vs. Cell				
Advia vs. Sysmex		**		
Cell vs. Sysmex		***		

The table above results suggests that Deming Regression and Ordinary least squares regression on raw data is applicable for only one measurement procedure comparison (*Advia 2120i vs. Cell Dyn Sapphire*). Heteroscedasticity is a critical issue for this data set, and accordingly, the subsequent step is to apply the Bland-Altman transformation scheme to transform the clinical samples. If the linear model assumptions still are unsatisfied concerning the Bland-Altman transformed clinical samples combined with ordinary least squares regression, the log-log transformation is an alternative. Parametric regression splines could also be applied because this estimator is likely to dampen heteroscedasticity present because of flexibility. Before anything else, the Bland-Altman transformation is used to fix the persistent heteroscedasticity issues potentially.

Linear model assumptions	*EPK*			
BAT + OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX vs. Advia		*		
ABX vs. Cell	***			
ABX vs. Sysmex				
Advia vs. Cell				
Advia vs. Sysmex		**		
Cell vs. Sysmex				

As implied by the table above, the Bland-Altman transformation fixed most assumptions but broke the normality assumption for one, which implies that trustworthy least squares

estimators may not be obtained for three of the six measurement procedure comparisons. Utilizing log-log transformation results in:

Linear model assumptions	*EPK*			
LLT + OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX vs. Advia	*			
ABX vs. Cell				
ABX vs. Sysmex				
Advia vs. Cell				
Advia vs. Sysmex				
Cell vs. Sysmex				

The log-log transformation, combined with ordinary least squares regression, resulted in the most fulfilled linear model assumptions. Thus one might conclude that the combination of log-log transformation and Deming regression is suitable. However, caution must be practiced because [Gillard, 2010] recommends $n > 50$ to use the method of moments estimators defined in Equation (2.20). After removing the NA-values, only $n = 23$ clinical samples remain, which induces the overestimation of the variances' ratio, λ. The normality assumption is unfulfilled in the first measurement procedure, ABX Micros ES vs. Advia 2120i, hence we are forced to drop it from the commutability assessment:

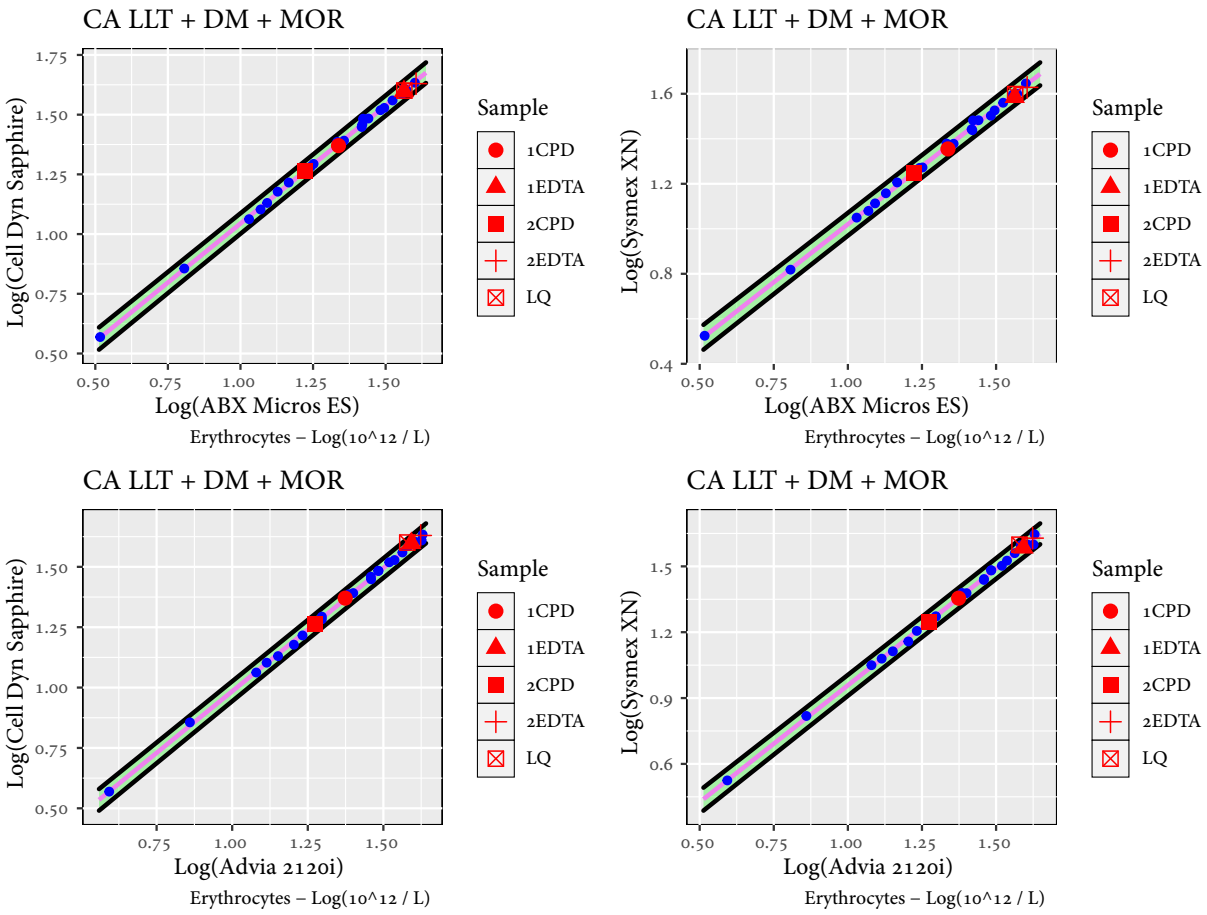

Figure 3.19 *– Visual commutability assessment for EPK data set employing the log-log transformation combined with Deming regression. The means of replicates are used to construct the models.*

From Figure 3.19, the estimated prediction bands enclose all control material samples for all measurement procedure comparisons presented. Thus all control materials are commutable regarding our acceptance rule. The figure below, Figure 3.20, shows that the last measurement procedure comparison results in all control materials' commutability.

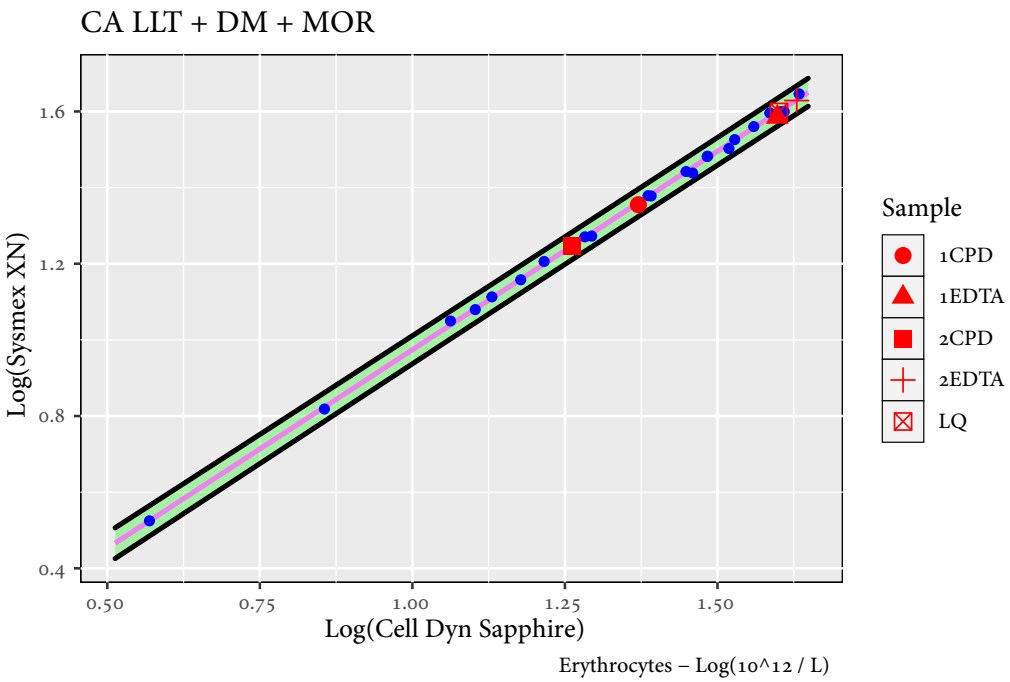

Figure 3.20 *– Visual commutability assessment for the EPK data set using log-log transformed data combined with Deming regression constructed employing the replicates' mean.*

The plots conclude commutable control materials regarding the determination of the commutability assessment procedure. The discarding of the normality assumption for *ABX Micros ES* vs. *Advia 2120i* is a strong motivation for the implementation of non-parametric assessment methods as part of the commutability assessment. Another data set, which is the *LPK* data set, is considered to strengthen the motivation for alternative approaches regarding the commutability assessment. *LPK* contains measurement procedure comparisons that are weakly non-linearly related. See the offset plot below for an illustration of the non-linear patterns of the clinical samples:

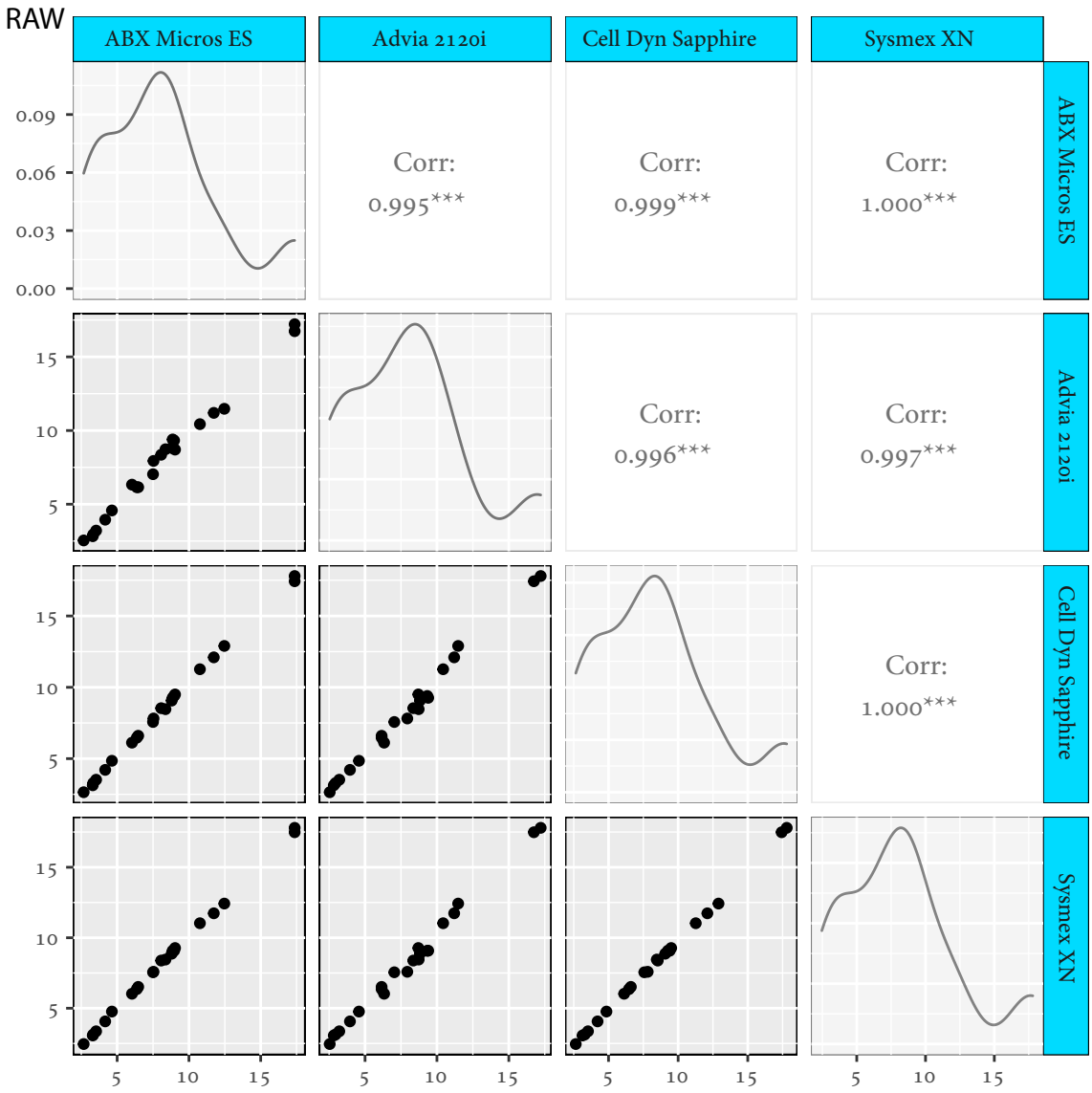

Figure 3.21 *– An offset plot that is visually describing the six relationships between the four specified measurement procedures in the LPK data set.*

Note that the degree of non-linearity is not drastic. However, some non-linear tendencies are observed. The estimated densities propose that the measurements are approximately log-normal in shape; hence the log-log transformation combined with either Deming regression or ordinary least squares is expected to be suitable. Because of data gaps in the upper regions of the measurement range, it is difficult to state anything trustworthy regarding the measurement patterns. In theory, the pattern may take any shape in such gaps. The log-log transformation reduces the absolute width of proposed data gaps; thus, the basis of selecting the log-log scheme is solid. Below an offset plot of the log-log transformed clinical samples are included.

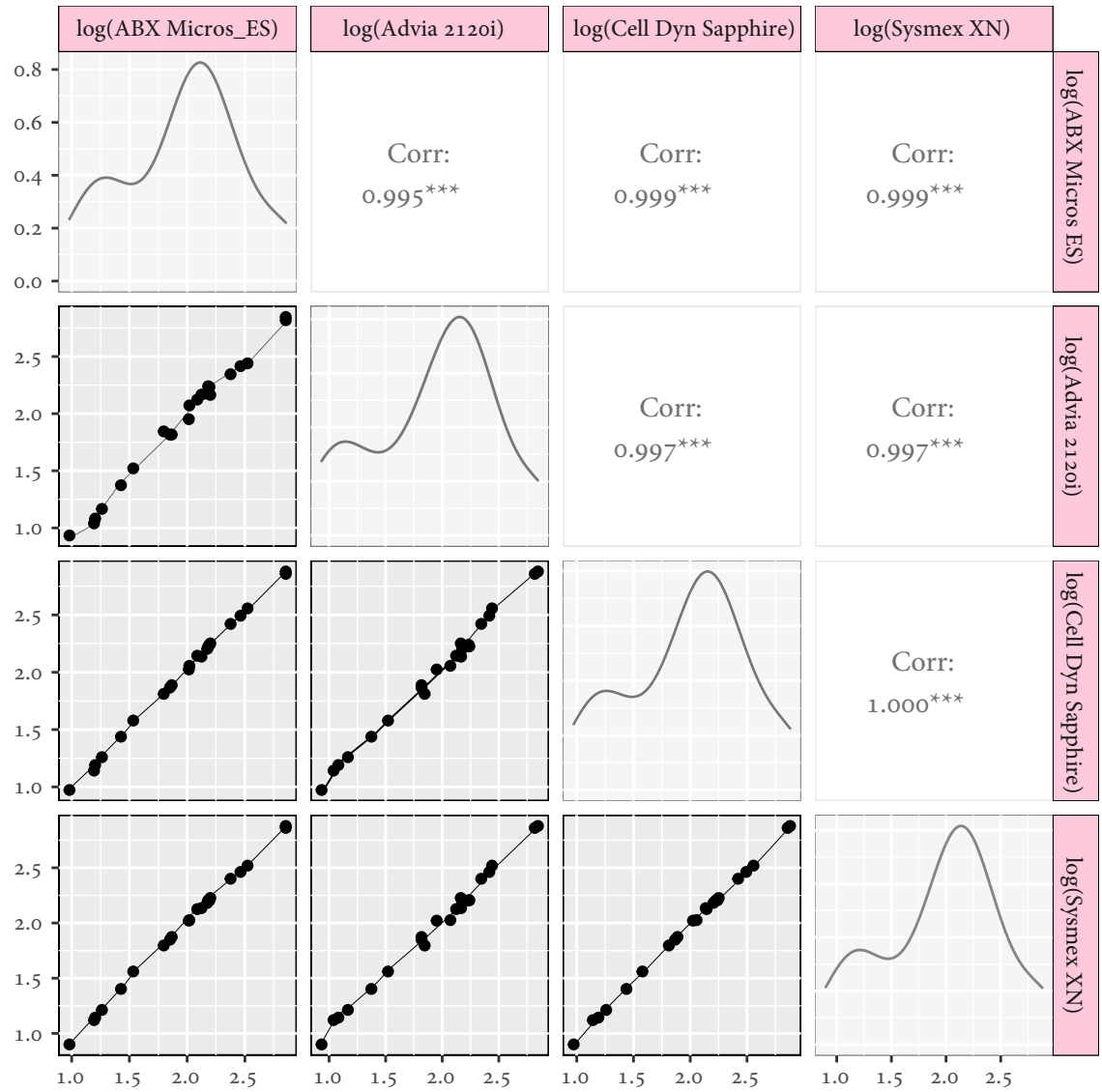

Figure 3.22 *– An offset plot that is visually describing the six relationships, where the clinical samples are log-log transformed.*

Note that the degree of non-linearity is not drastic. However, some non-linear tendencies are observed. The estimated densities propose that the measurements are approximately Gaussian in shape; hence the log-log transformation combined with either Deming regression or ordinary least squares appears suitable. Because of data gaps in the upper regions of the measurement range, it is difficult to state anything trustworthy regarding the measurement patterns. In theory, the clinical sample patterns may take any shape in these gaps. The log-log transformation reduces the absolute width of proposed data gaps, which indicates that the basis of selecting the log-log scheme is factual. The linear model assumptions below are

Table 3.6 *– Evaluation of linear model assumptions of raw data and log-log transformed data fitted by ordinary least squares. The mean of the replicated measurements is applied.*

Linear model assumptions	*LPK*			
OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i				
ABX Micros ES vs. Cell Dyn Sapphire				**
ABX Micros ES vs. Sysmex XN		*		
Advia 2120i vs. Cell Dyn Sapphire				
Advia 2120i vs. Sysmex XN				
Cell Dyn Sapphire vs. Sysmex XN				*
Linear model assumptions	*LPK*			
LLT + OLSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i				**
ABX Micros ES vs. Cell Dyn Sapphire				**
ABX Micros ES vs. Sysmex XN				***
Advia 2120i vs. Cell Dyn Sapphire				
Advia 2120i vs. Sysmex XN				
Cell Dyn Sapphire vs. Sysmex XN				

tested using the *gvlma()* function provided by the *gvlma* package in R. See Table 3.6 for displayed results.

The log-log transformation did not really help us regarding linearity. Note that the rest of the linear model assumptions are satisfied. Therefore it would be nice to have a method that did not depend on linearity. This is another motivation for implementing non-linear non-parametric methods for commutability evaluation. Since we may only use three of the six possible comparisons, we choose not to take it further for the moment.

3.6 Simulation studies

It is time to consider some simulated data to be evaluated by parametric assessment procedures for commutability. Before anything else, it is vital to discuss the premises upon which our simulations were dependent. CV will be desirable in opposition to SD. This decision comes from the fact that CV is scale-independent, whereas SD is not. Consequently, the CV will be a more comparable measure of dispersion, essential for these sections' purposes. Log-log transformation will naturally already have a scale-independent standard deviation (that is, if the log-log transformed data are approximately normally distributed), so we should,

accordingly, transform the standard deviation of the log-transformed data to the CV of the raw data by:

$$CV_A \approx \sqrt{\exp\left[\text{SD}^2_{\log_e}\right] - 1}. \tag{3.1}$$

Here $\text{SD}_{\log_e}$ is the estimated standard deviation of the log-transformed measurements. The standard deviation of the logarithmic difference of the Bland-Altman transformed clinical samples is also scale-independent, and the same relationship between the CV of raw data, and this standard deviation holds:

$$CV_A \approx \sqrt{\exp\left[\text{SD}^2_{\text{LD}}\right] - 1}. \tag{3.2}$$

In these simulation studies and the simulation studies for the non-parametric methods, we will simulate $K = 100$ to $K = 1000$ data sets for every simulation step. There will be several hundreds of these simulation steps; consequently, many of the simulations studies are computationally demanding. Parallel programming will, therefore, be applied. We generate the true values of the quantity within analytes from a uniform probability distribution with boundaries α and β:

$$\tau_{Ai} \sim \mathcal{U}(\alpha, \beta) \tag{3.3}$$

In all our simulations we will apply that $\alpha = 3.5$ and $\beta = 11$. Regard that this is the reference range of the number of leukocytes per Liter in human adults. Adopting the same sample space for all simulations to come, we may isolate the differences (defects and strengths) between assessment methods. Besides, for non-linearity purposes, we will restrict the relationship between τ_{Ai} and τ_{Bi} to be parabolic:

$$\tau_{Bi} = f(\tau_{Ai}) = a \cdot \tau^2_{Ai} + b \cdot \tau_{Ai} + c. \tag{3.4}$$

a, b, and c will be unconstrained parameters for us to choose whether we want to discuss linear relationships ($a = 0$) or else non-linear relationships ($a \neq 0$). In the real world, data following a perfect parabola is rare. However, we are interested in any non-linear pattern that we may control by specified parameters. Parabola curvatures will accordingly suffice in this text. For the uncertainty in measurement procedures, we will add some normal noise, which earlier established. Thus, if we fixate sample $i = I$ with replications $r \in \{1, \ldots, R\}$ yields

$$\begin{aligned} \text{MP}_{AIr} &= \tau_{AI} + \mathcal{N}(0, \sigma^2_A) \\ \text{MP}_{BIr} &= \tau_{BI} + \mathcal{N}(0, \sigma^2_B). \end{aligned}$$

In the simulations, $\sigma_A = 0.1$ and $\sigma_B = 0.2$. Thus the corresponding theoretical coefficients of variation is given by

$$\text{CV}_A \approx 0.014$$
$$\text{CV}_B \approx 0.028$$

Subsequently, simulated measurements within sample $i = I$ will be a little different, as in reality. We must stress that differences in replicated measurements will be more complicated than the proposed normally distributed. Nevertheless, the gaussian distribution assumption on the error terms will suffice. The table below describes the details concerning the simulations done in this section.

***Table 3.7** – Descriptive details of the simulation studies.*

Simulation details	*Parametric methods*				
CA method	Simulated data sets	Simulation steps	LMA simulations	CA simulations	Other
OLSR	1000	500	1	3	0
LLT + OLSR	1000	50	3	3	0
BAT + OLSR	1000	500	6	3	1
DR	100	50	1	3	1

3.6.1 Ordinary least squares regression - Simulation

We will first look at the most simple commutability assessment procedure, the ordinary least squares regression method. Here we are to test how robust the method is concerning different underlying adjustments. Firstly, the linear model assumptions are examined for different non-linearity coefficients. The acceptance rates of linear model assumptions are expected to be $\approx 86\%$ because multiple tests are done (assumed independent) with $\alpha = 0.05$. Secondly, the control materials are set as commutable, and we will observe how often the ordinary least squares estimator will conclude with commutability when the truth is commutable. 99% is expected because we use one control material sample and a significance level of $\alpha = 0.01$. Acceptance rates of commutability much below 99% are considered to be concerning.

In this section, the generation of clinical samples and control material samples is done using *simulate.data()*. See Appendix A for details on this function. An advantage of simulating data is that we actually know the true values, and thus, the quality of the methods is see-through. The acceptance rates of linear model assumptions are studied by implementing different magnitudes of non-linearity upon the clinical sample measurements. The linear

model assumptions are expected to reduce as the non-linearity coefficient, a, increases. The set of considered values of a is given by

$$\{a_i\} = \{0, 0.0003, \ldots, 0.14997, 0.15\}. \tag{3.5}$$

The *breakpoint* is defined as the simulation variable's approximate value when the acceptance rates of commutability or linear model assumptions are 50%.

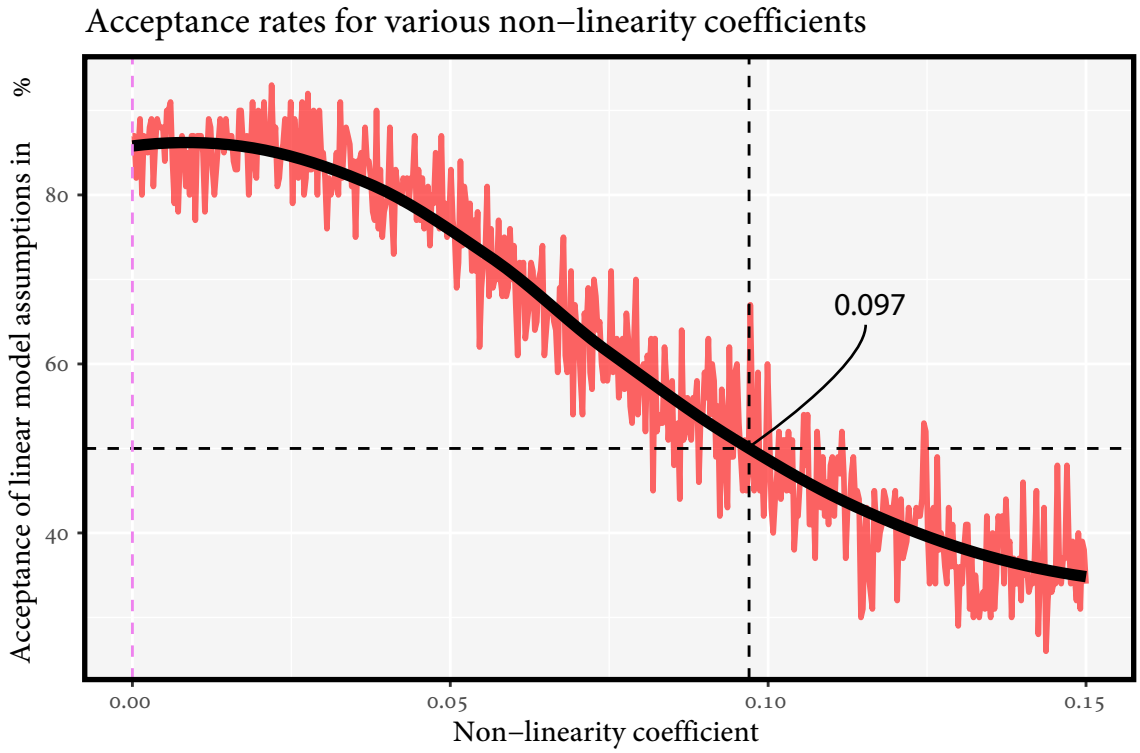

Figure 3.23 *– The accepting rates with various degrees of non-linearity.*

We can also see how much we must alter the different coefficients of a, b, and c for this evaluation method to reject all control materials' commutability. Particularly, the defined relationships between clinical samples and control material samples are presented as

$$\begin{aligned} a_{\text{patients}} &= a_{\text{controls}} + 0.0001 \cdot q \\ b_{\text{patients}} &= b_{\text{controls}} - 0.5 + 0.002 \cdot q \\ c_{\text{patients}} &= c_{\text{controls}} - 0 + 0.04 \cdot q, \end{aligned}$$

where each $q \in \{1, \ldots, 500\}$ is considered as a simulation step, where 1000 data sets are simulated.

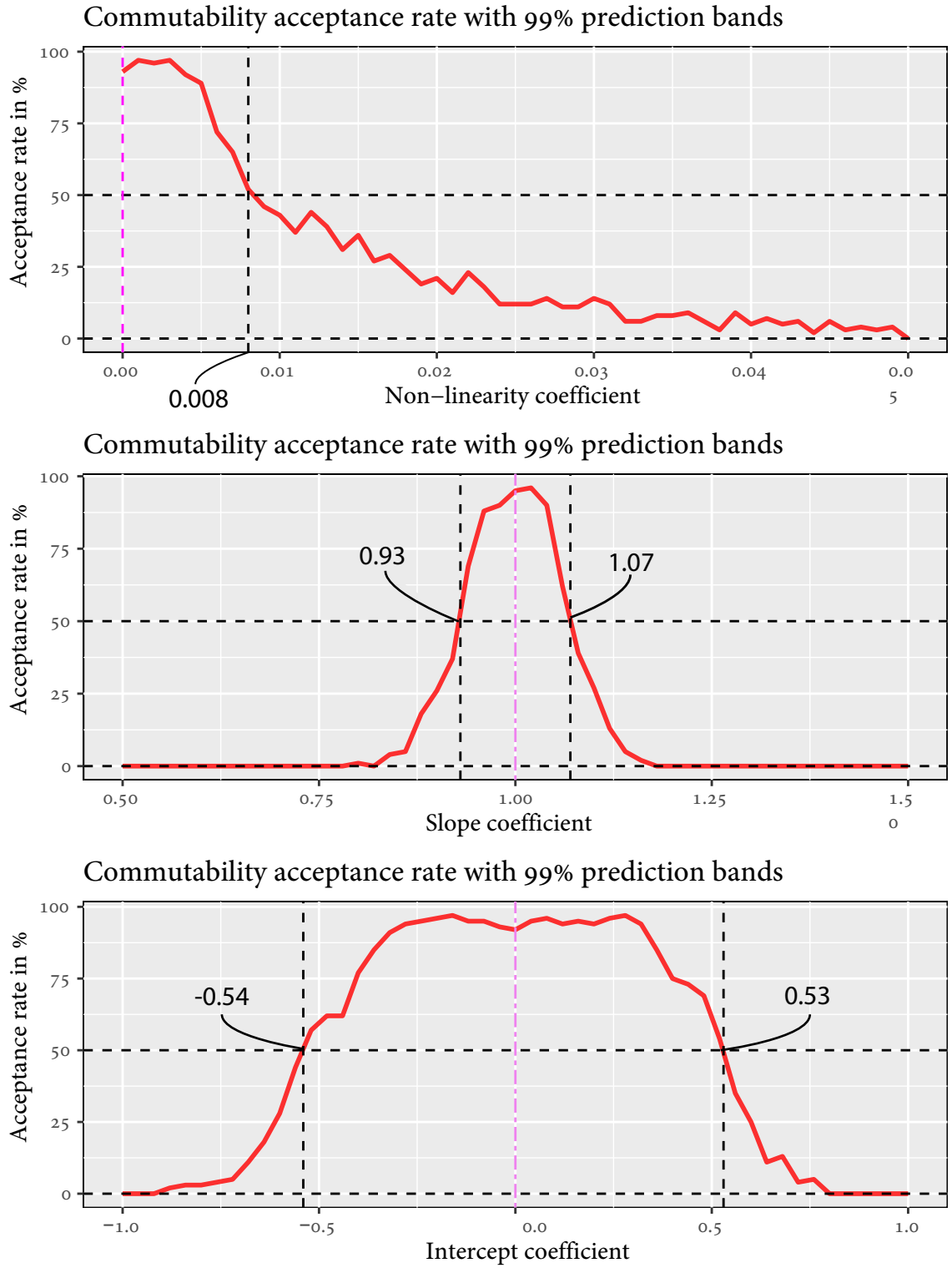

Figure 3.24 *– Different choices of a,b, and c for the clinical samples and their corresponding commutability acceptance rate while a,b, and c for control material samples are unchanged. The violet dashed lines outline the values of a, b, and c for the control material samples.*

3.6.2 Log-log transformation - Simulation

The log-log transformation approach as part of the commutability assessment, proved to have several requirements on its own. Consequently, it will be interesting to analyze the dynamics of the acceptance rates of both commutability and model adequacy tests as the non-linearity, slope intercept coefficients increase. In contrast to Section 3.6.1, two additional simulation studies is performed on the linear model assumptions in this section. The different coefficients impact on the linear model assumptions are considered. Particularly, the following simulation steps are applied:

$$\begin{aligned}\{a_i\} &= \{0, 0.001, \ldots, 0.049, 0.05\}\\ \{b_i\} &= \{0.5, 0.52 \ldots, 1, \ldots, 1.48, 1.5\}\\ \{c_i\} &= \{-5, -4.8, \ldots, 4.8, 5.0\}\end{aligned}$$

These simulation steps yielded the results in Figure 3.25

A notable difference between the log-log transformation and the ordinary least squares is that the log-log method is somewhat unstable regarding accepting the linear model assumptions. Large deviations surfaces, as we see in Figure 3.25 (red line). Note also that the acceptance rates are significantly lower than what we saw for ordinary least squares regression, even for acceptable choices of a, b, and c. Thus, we have found a more severe weakness of the log-log transformation evaluation method. Let us examine the log-log method's strictness regarding various alterations of a, b, and c concerning the patients. Strictness regarding commutability assessment is now of interest. In order to examine this, we simulate 1000 data-sets for every $q \in \{0, 1, \ldots, 50\}$ such that

$$\begin{aligned}a_{\text{patients}} &= a_{\text{controls}} + 0.001 \cdot q\\ b_{\text{patients}} &= b_{\text{controls}} - 0.5 + 0.02 \cdot q\\ c_{\text{patients}} &= c_{\text{controls}} - 1 + 0.4 \cdot q.\end{aligned} \tag{3.6}$$

Note that we chose the specifications above to get convenient and visible results. In that sense, it is not relevant how the relationships between controls and patients are specified. Ideally, we would like to reject the commutability property if the relationship between patients and controls are sufficiently extreme. Particularly, we prefer that acceptance rates of commutability decrease as deviations between patients and controls increase. Note also that this is entirely dependent on the sample space specified in the simulation studies. We use that $\alpha = 3.5$ and $\beta = 11$ here. We would need different alterations if α and β were defined differently.

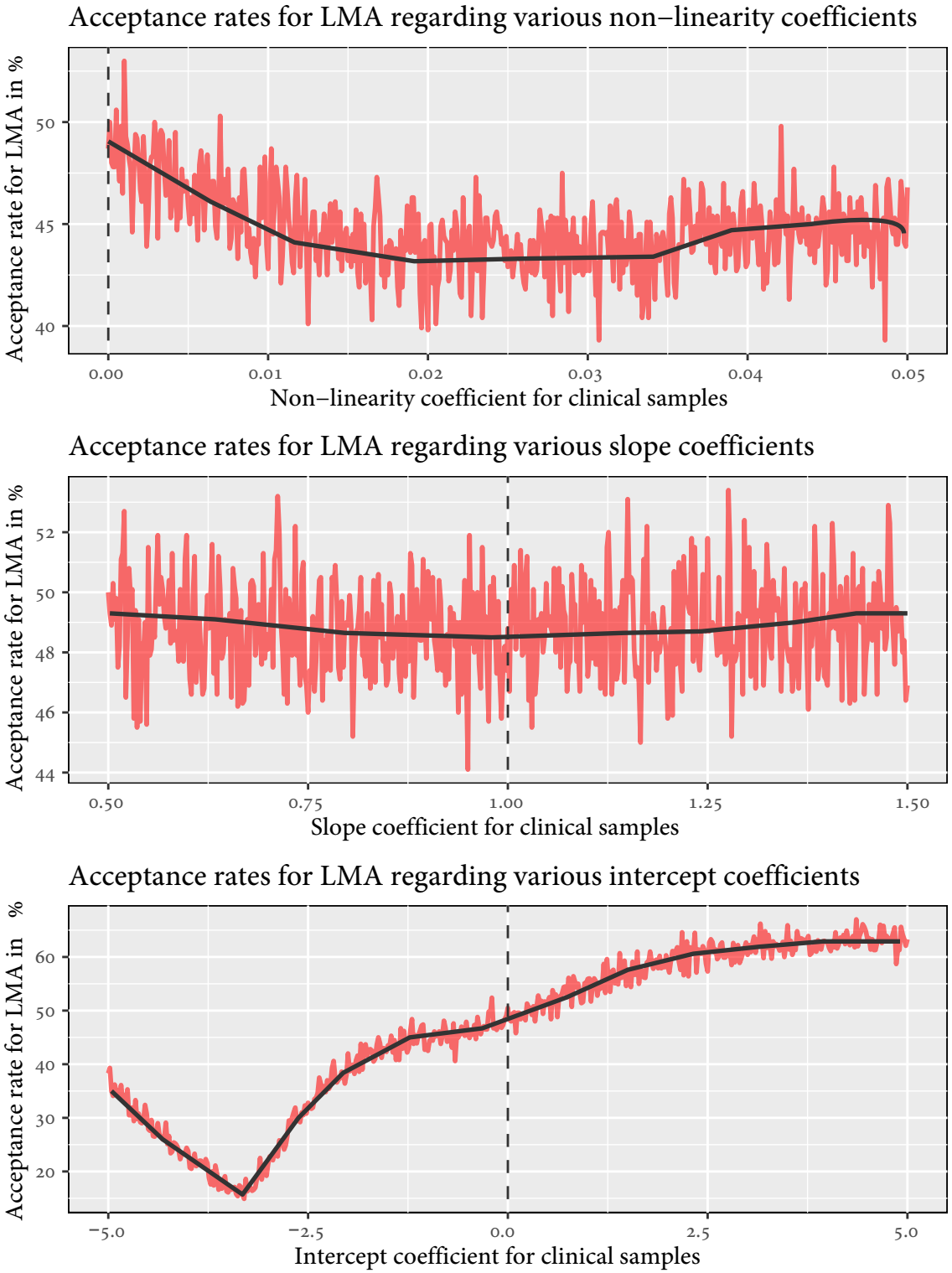

Figure 3.25 *– Acceptance rates of the linear model assumptions when changing a, b, and c for log-log transformed observations fitted by ordinary least squares.*

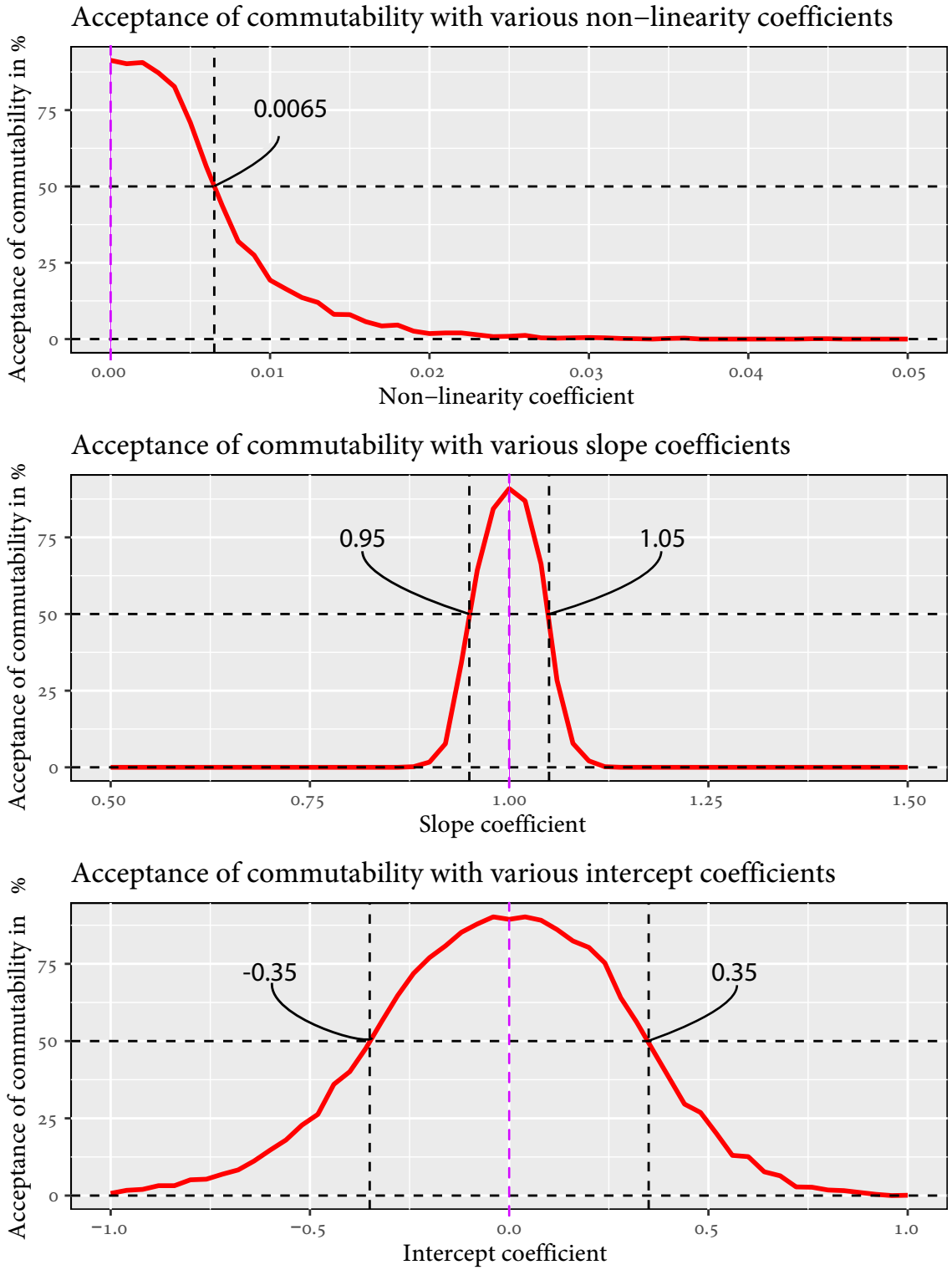

Figure 3.26 *– How much change in a, b, and c are necessary to reject the commutability property for the log-log transformed observations fitted by ordinary least squares.*

We observe that the log-log evaluation method will reject commutability for less extreme deviations between control materials and patients than with ordinary least squares regression. However, this observation is positive. This implies that the log-log transformation is more stringent than just using ordinary least squares regression. However, it was a big fault that the linear model assumptions acceptance rates went down when we did the log-log transformation. An Acceptance rate of around 50% is far from good enough. Besides, the variability of the acceptance rates concerning the linear model assumptions seems relatively vast compared to the other methods. We may therefore conclude that LLT with OLSR is somewhat incapable as a commutability assessment method.

3.6.3 Bland-Altman Transformation - Simulation

It is time to consider a simulation study regarding the Bland-Altman transformation evaluation method. As illustrated in Section 2.4, the Bland-Altman transformation combined with ordinary least squares regression, is suitable for exposing weak non-linearity and moderate non-zero intercepts in the clinical samples. Furthermore, we recognized that the prediction bands of these types of transformations yielded were relatively extensive. Far-reaching prediction bands are questionable because they may accept control materials' commutability when they are not in reality. See Figure 3.13. Also, using Bland-Altman regression with ordinary least squares regression is only appropriate if the estimated coefficients are statistically significant and were not the status in Section 3.5. A vicious obstacle concerning this assessment procedure is manually choosing the polynomial degree used in the approach. Therefore different polynomial degrees are applied, particularly $p = 1$ and $p = 4$. As discussed briefly, the Bland-Altman transformation is associated with handling the variability in x-direction by reducing its magnitude and ergo its influence. Estimating the reduction in x-variability in percent is a legitimate interest. Therefore $K = 100$ data sets are simulated, which are applied to estimate the coefficients of variation with and without Bland-Altman transformations. This simulation step is replicated 100 times, where the calculated ratio of coefficients of variances is returned. The standard formula calculates the estimated coefficient of variation of the raw clinical samples. In contrast, the equivalent converted CV is calculated by Equation (3.2). The results are presented in Figure 3.27.

The x-variability is lessened by approximately 30%. The 95% percentile confidence interval is calculated to be

$$\text{CI} = [29.72\%,\ 30.35\%]. \tag{3.7}$$

As a rule of thumb, one might say that $(1 - \frac{1}{\sqrt{2}}) \cdot 100\%$ variability in x-direction is diminished when Bland-Altman transformation is applied. The next point of interest is investigating how

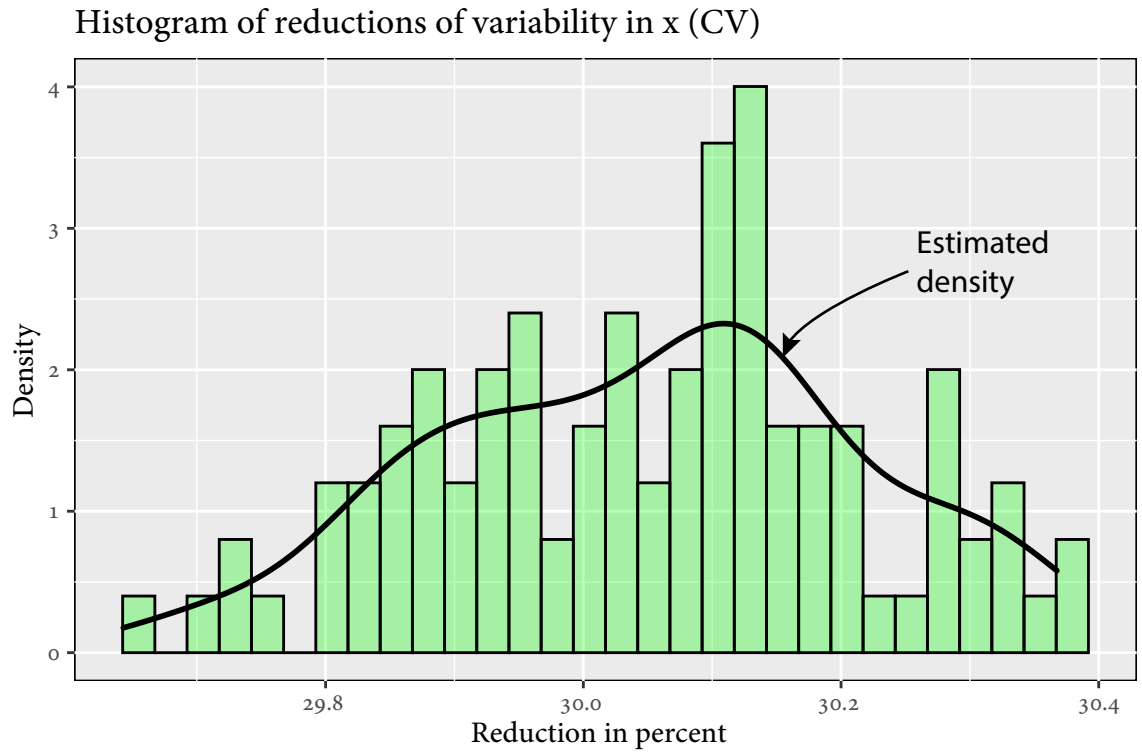

Figure 3.27 *– How much does the CV decrease using the Bland-Altman transformed data instead of raw data in x?*

stable the acceptance rates of linear model assumptions are when the clinical samples have different patterns. As before, the expected percent of accepted linear model assumptions is approximately 86%. Acceptance rates significantly lower than this is considered alarming. The alterations of the patient samples executed by using various combinations of a, b and c, are defined below:

$$\begin{aligned} a_{\text{patients}} &= 0.0001 \cdot q \\ b_{\text{patients}} &= -0.5 + 0.002 \cdot q \\ c_{\text{patients}} &= -1 + 0.004 \cdot q, \end{aligned} \tag{3.8}$$

for every $q \in \{0, 1, \ldots, 500\}$. The results of the alterations in Equation (3.8) are presented in Figure 3.28. Generally, Figure 3.28 suggest that the three separate plots' acceptance rates are inappropriately near the breakpoints for most of the adjustments presented in Equation (3.8). In the plots, one might observe peculiar relationships. For instance, the acceptance rates are declining as the non-linearity coefficient reaches approximately 0.02, but then the acceptance rates rise again. Using the same simulations, using a polynomial degree 4 produces the plots in Figure 3.29.

The results in Figure 3.29 are much better than the results in Figure 3.28. The acceptance rates for $a = 0$, $b = 1$ and $c = 0$ increased by from near breaking point to above 70%. However, it is implausible that all five coefficient estimators are significant. Thus, we cannot be sure whether the results of Figure 3.29 are reliable from a statistical perspective. For models where the five coefficient estimators are significant, we could securely move on with the commutability assessment. Since we cannot be certain here, it appears prudent to avoid the Bland-Altman transformation connected with parametric models, such as ordinary least squares or Deming. Lastly, the commutability acceptance rates are considered for various choices of a, b, and c. The acceptance rates for commutability are expected to be approximately 99% as usual. Figure 3.30 displays that results for the alterations presented in Equation (3.9)

$$\begin{aligned} a_{\text{patients}} &= a_{\text{controls}} + 0.0001 \cdot q \\ b_{\text{patients}} &= b_{\text{controls}} - 0.5 + 0.002 \cdot q \\ c_{\text{patients}} &= c_{\text{controls}} - 1 + 0.004 \cdot q, \end{aligned} \tag{3.9}$$

for every $q \in \{0, 1, \ldots, 500\}$. Figure 3.30 produced similar results to what the log-log transformation scheme did. In conclusion, the Bland-Altman transformation is superior to the log-log transformation from the following basis; Bland-Altman transformation reduces

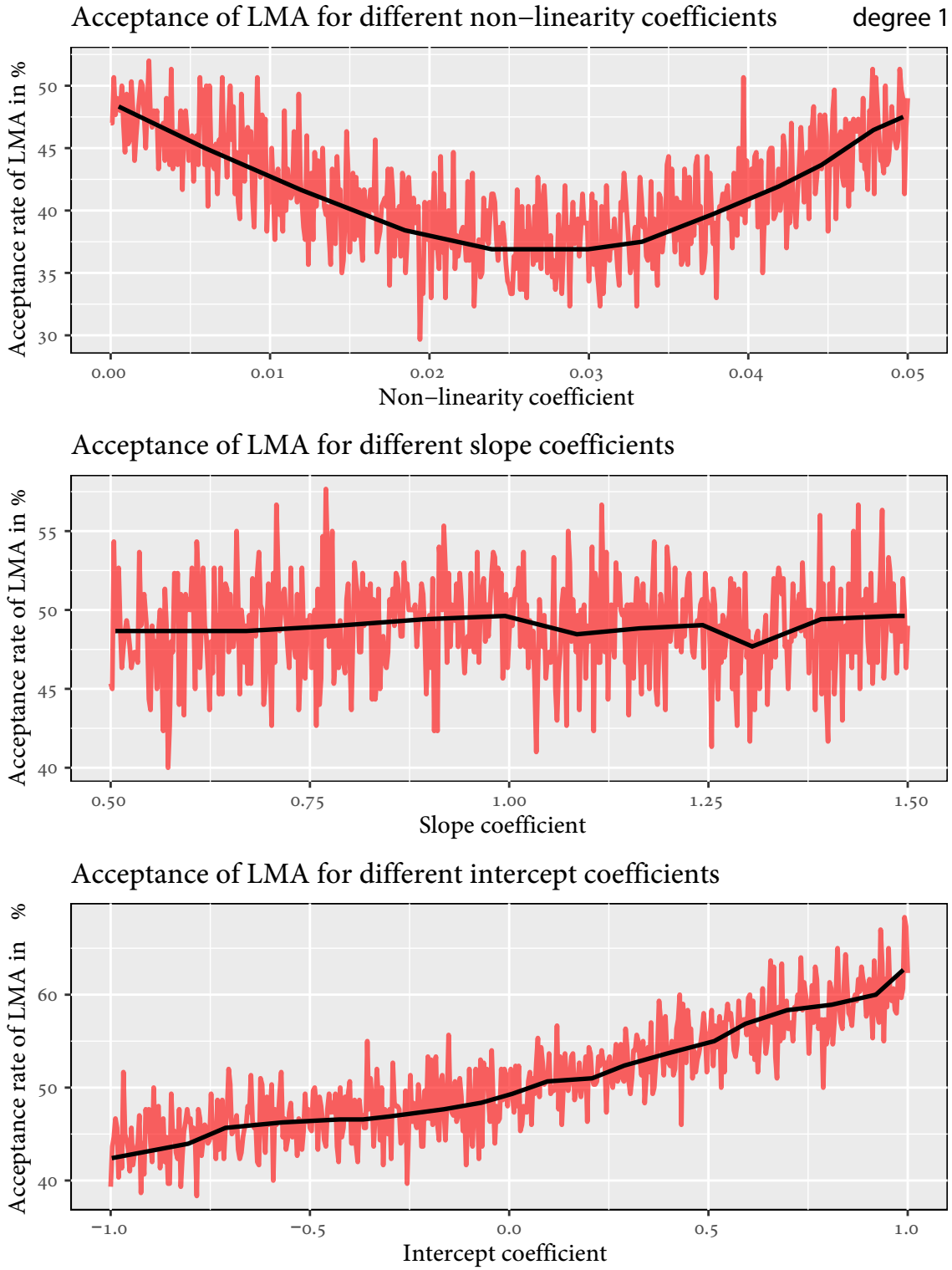

Figure 3.28 *– Acceptance rates for the linear model assumptions when increasing a, b, and c. The simulated Bland-Altman-transformed clinical samples are fitted by ordinary least squares regression.*

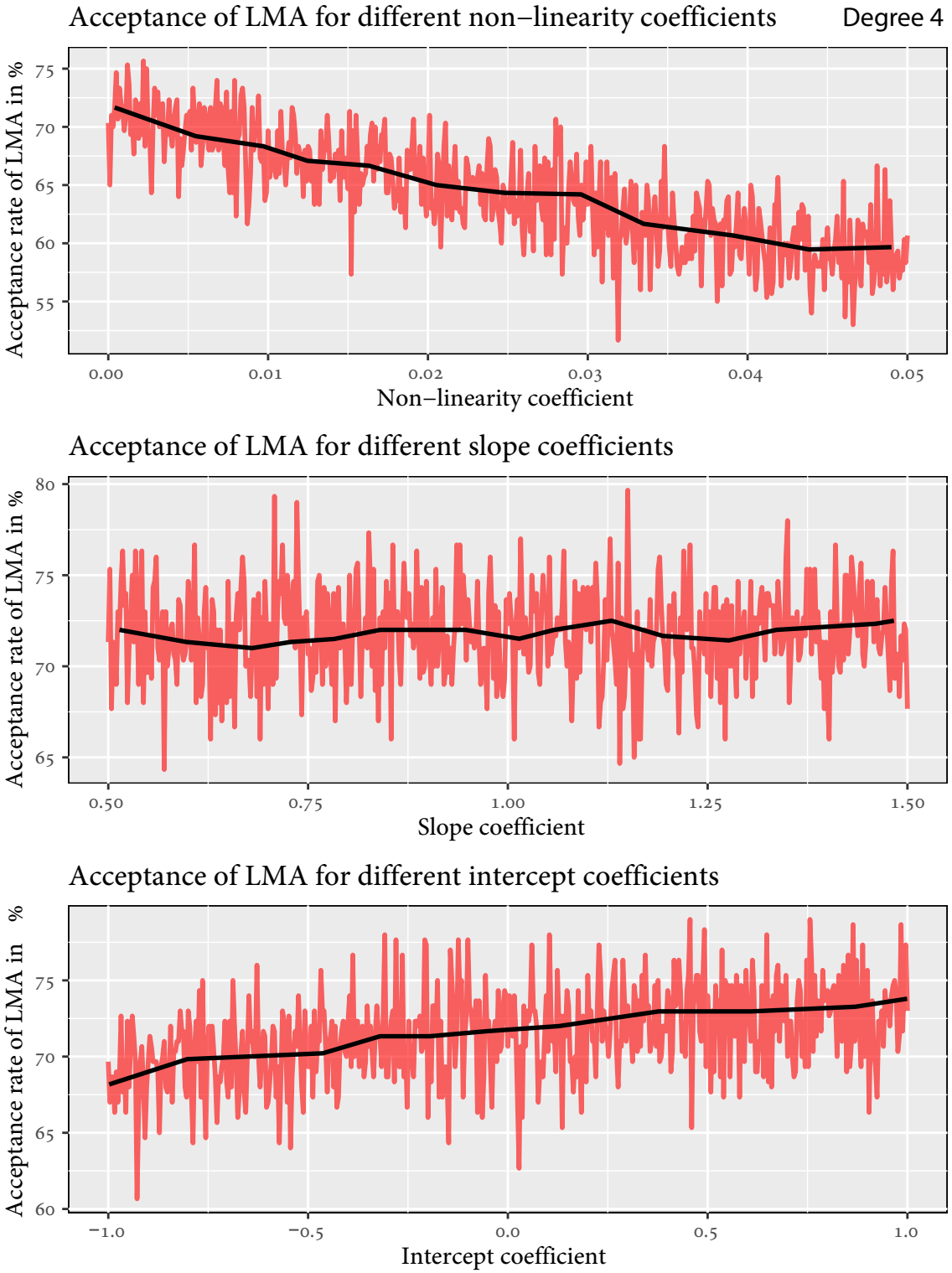

Figure 3.29 *– Acceptance rates for the linear model assumptions when increasing a, b, and c. The simulated Bland-Altman-transformed clinical samples are fitted by polynomial regression with $p = 4$.*

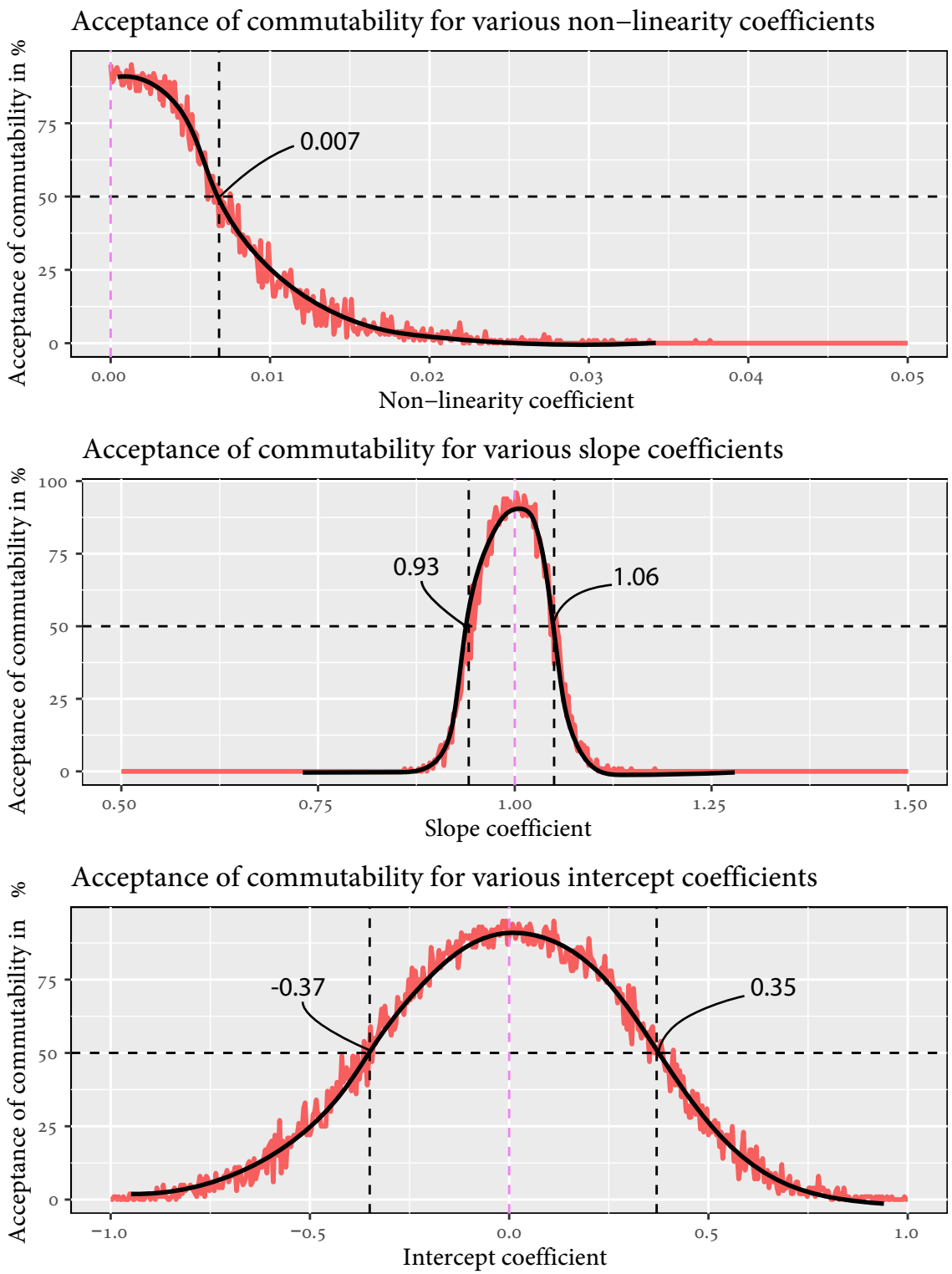

Figure 3.30 *– The assent rates of commutability for various choices of a, b, and c relative to the control materials for the simulated Bland-Altman transformed clinical samples fitted by polynomial regression with $p = 4$. The dashed violet lines are the corresponding values of a, b, and c for the control materials*

variability in x by approximately $(1 - \frac{1}{\sqrt{2}}) \cdot 100\%$ and the rates of acceptance regarding linear model assumptions were more significant for the Bland-Altman transformed clinical samples. The drawback is potentially non-significant regression estimators, which may enlarge the prediction bands too much.

3.6.4 Deming Regression - Simulation

Deming regression is unique among the parametric assessment methods. As earlier established, we saw that Deming regression considered the variability in both x- and y-direction. Therefore it was argued to be more realistic. We will do a somewhat more significant simulation study regarding DR. The main reason for this is that the DR model is more detailed than the four other evaluation methods. Firstly let us consider the acceptance rate of the linear model assumptions for different specifications of non-linearity coefficients. We expect this simulation results to be similar to what we got for ordinary least squares regression. The relationship between the linear model assumptions' acceptance rates for the range of selected non-linearity-coefficient magnitudes (same as in Section 3.6.1) is provided in Figure 3.31.

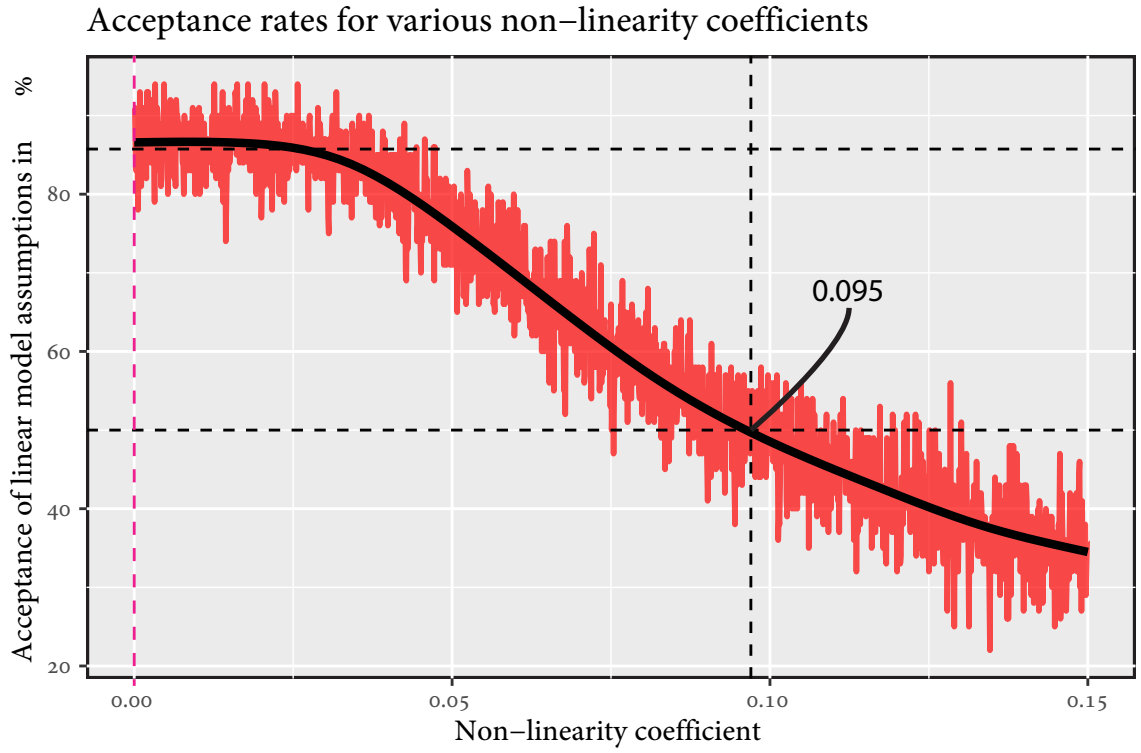

Figure 3.31 *– The linear model assumptions' acceptance rates when increasing a for simulated clinical samples fitted by Deming regression.*

A notable alteration is required to discard more than 50% of the linear model assumptions. Linear model assumptions are frequently satisfied for non-linearity coefficients up to 0.025. Regarding the commutability assessment acceptance rates, the Deming regression estimator is expected to be somewhat less stringent because of the implementation of variability in x. See Figure 3.14 and Figure 3.15 for an illustration. Accordingly, more considerable differences between control materials and clinical samples are required to reach the breakpoint. More significant alterations of a, b, and c are required to reject control materials' commutability than the transformation approaches. As described in Section 2.5, there may be problems regarding the use of Equation (2.20) when possessing few clinical samples. Consequently, the impact of having $n = 25$ samples instead of the recommendation $n > 50$ will be investigated. Analyzing a true $\lambda = 2$, we consider the estimated λ for different choices of sample sizes and the number of replications. See Figure 3.33 for the portrayed results.

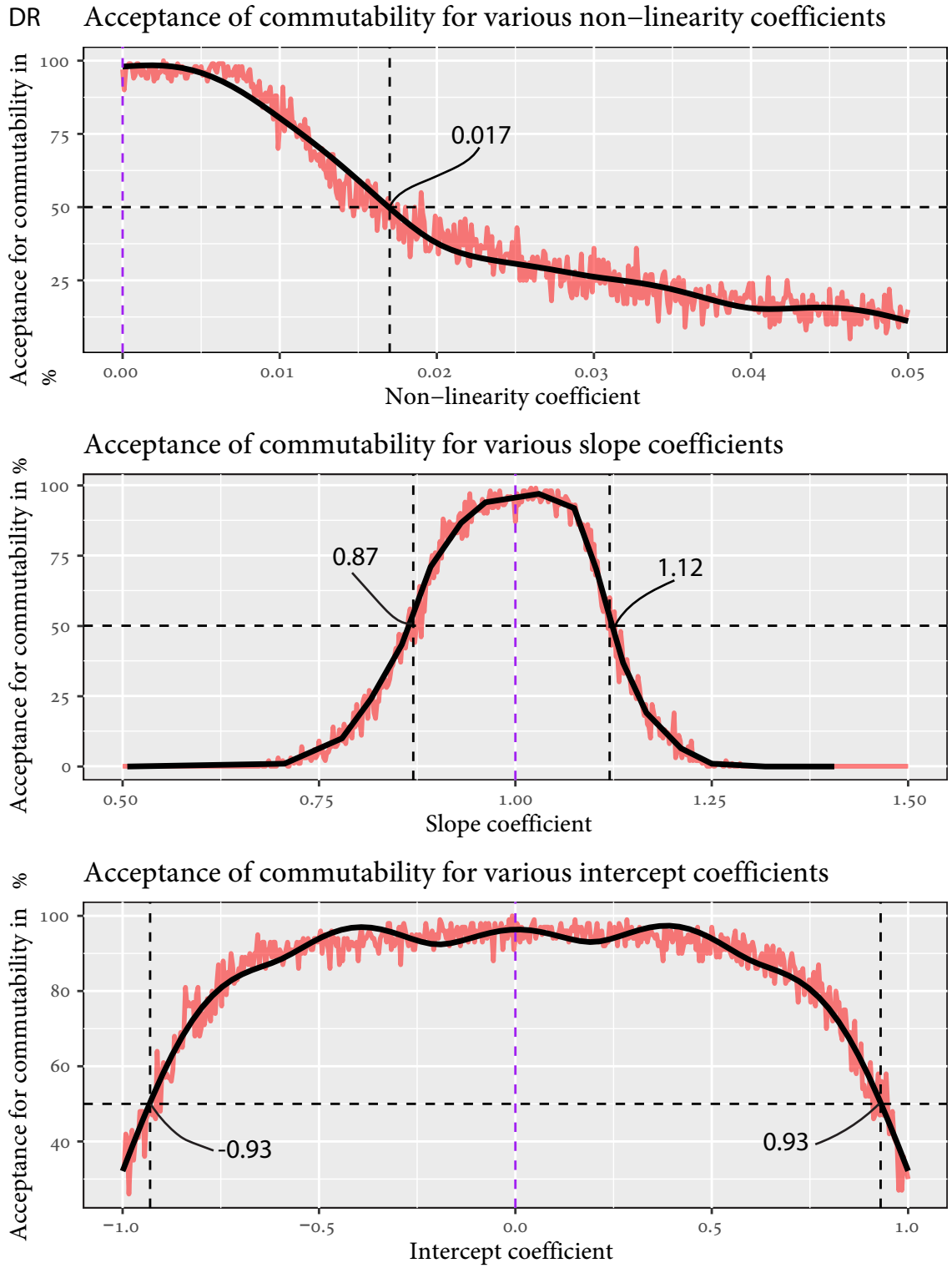

Figure 3.32 *– The commutability accepting rates for various a, b, and c, where the simulated clinical samples are fitted by Deming regression.*

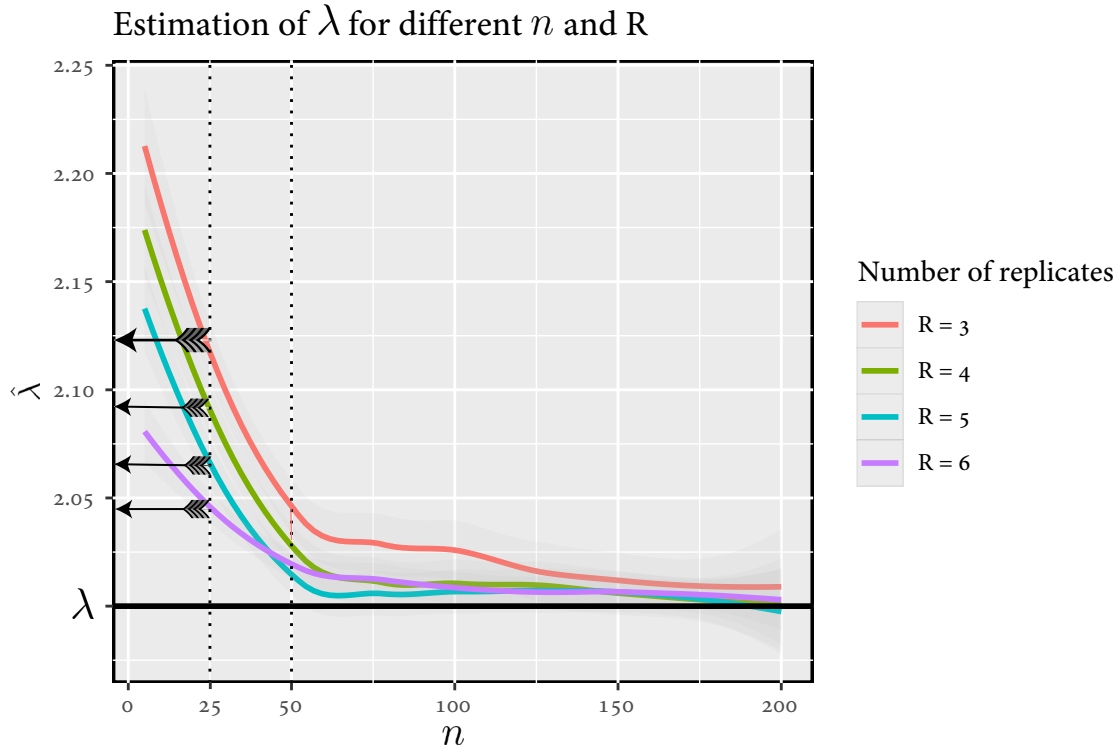

Figure 3.33 *– Estimation of λ for an increasing number of clinical samples and different numbers of replicates.*

The data sets considered in the previous sections contained $n = 25$ clinical samples and three replicates for every clinical sample. From Figure 3.33 λ is typically overestimated by approximately seven percent. Note that using five or six replicates yielded approximately the same precision as having $n = 50$ clinical samples with three replicates. The relative bias is approximately two percent for the latter. Since increasing the number of replicated measurements is much less costly than increasing the number of clinical samples, it seems prudent to use five or six replicated measurements if Deming Regression is utilized.

3.6.5 Results

These simulation study results are provided in Table 3.8.

The log-log transformation, combined with the ordinary least squares estimator, is proposed as the most strict approach to accepting commutability. In opposition, the Deming regression estimator is the least stringent. The transformation strategies did not handle the linear model assumptions well. Arguably, this is a motivation for avoiding transformations as a general commutability assessment procedure. Therefore, a possible conclusion is that only the raw clinical samples should be used for the parametric regression estimators to secure satisfactory acceptance rates for the linear model assumptions.

Table 3.8 *– The table summarizes the results from the simulation studies done in this section concerning the acceptance rates of the linear model assumptions and control materials' commutability property.*

OLSR				*BAT + OLSR*			
	LMA break point	CA break point	p		LMA break point	CA break point	p
Non-linearity coefficient	0.097	0.008	1	**Non-linearity coefficient**	0	NA	1
Slope coefficient	NA	0.93 or 1.03	1	**Slope coefficient**	NA	NA	1
Intercept coefficient	NA	-0.54 or 0.53	1	**Intercept coefficient**	NA	NA	1
LLT + OLSR							
	LMA break point	CA break point	p		LMA breakpoint	CA break point	p
Non-linearity coefficient	NA	0.0065	1	**Non-linearity coefficient**	>0.05	0.007	4
Slope coefficient	NA	0.95 or 1.05	1	**Slope coefficient**	NA	0.93 or 1.06	4
Intercept coefficient	NA	-0.35 or 0.35	1	**Intercept coefficient**	NA	-0.37 or 0.35	4
DR							
	LMA break point	CA break point	p				
Non-linearity coefficient	0.095	0.017	1				
Slope coefficient	NA	0.87 or 1.12	1				
Intercept coefficient	NA	-0.93 or 0.93	1				

3.7 Alternative acceptance criteria for commutability

In Section 3.5, we used the $(1 - \frac{\alpha}{m}) \cdot 100\%$ prediction bands (alone) as the acceptance rule for commutability. We considered the mean of the replications in control material samples and checked whether this quantity (the mean) was inside the clinical samples' prediction bands. Evaluations like this ignore the variability of the control materials. As with the patients' samples, we can estimate the variability within control materials when possessing several replicates. Firstly, in this section, we will implement the variation of control material to evaluate commutability and alter our present criterion from this addition.

3.7.1 Confidence regions for control materials

Instead of evaluating commutability concerning a specific point's location, we may assess whether a, say, 90% confidence region lies inside the prediction bands. We may use the variability within the control material samples to gather confidence intervals for both x and y-direction. This will result in the so-called confidence regions. Then we have three possible outcomes:

1. A confidence region concerning a control material is entirely inside the prediction bands

2. A confidence region concerning a control material is partly inside the prediction bands

3. A confidence region concerning a control material is fully outside the prediction bands.

From these three outcomes, we may formulate different assessment criteria for commutability. For example, we may conclude that control material samples satisfying 1 and 2 are commutable, whereas control materials satisfying 3 are not commutable. Or else we say that point 1 is good enough for assessing commutability, whereas 2 makes us unable to conclude anything. The latter is seen to be too stringent so that we would use the first one mentioned.

3.7.2 Range regions for control materials

Another way to describe the control materials' variability is to set the variability's limits to the maximum and minimum of the replications. Too see what this means, let us consider three replicated measurements $m_{A1} < m_{A2} < m_{A3}$ from measurement procedure A and $m_{B1} < m_{B2} < m_{B3}$ from measurement procedure B. Then the range region for a specific control material samples with replicated measurements specified above is given by

$$\text{RR} = \{(x, y) : (x, y) \in [m_{B1}, m_{B3}] \times [m_{A1}, m_{A3}]..\} \tag{3.10}$$

1. A range region concerning a control material is entirely inside the prediction bands

2. A range region concerning a control material is partly inside the prediction bands

3. A range region concerning a control material is fully outside the prediction bands.

As formulated for confidence regions, one might blueprint the acceptance criterion for commutability by employing 1 or 2 as acceptance limits. Alternatively, one can use 1 as the only acceptable and by defining 2 as a gray-zone where one cannot classify the control material, and lastly, 3 as ensured non-commutable.

Chapter 4

Non-parametric assessment procedures

In Chapter 3, we saw that all currently used evaluation methods of control materials' commutability depended on the linear model assumptions. Besides, some of the methods had more requirements than this. In our discussion, faults regarding many of these procedures concerning the linear assumptions' acceptance rates were established. The acceptance rates for the transformation approaches were particularly alarming. See Section 3.6 for the details. Fortunately, alternative methods, so-called non-parametric methods, do not depend on the underlying distribution's shape. In most cases, we will still require independence of error terms and homoscedasticity if weights are not applied. In this chapter, we will consider the theory behind suitable non-parametric methods. After that, the discussed models and their prediction bands will be used as part of the commutability assessment criterion.

4.1 Thiel - Sen Regression

Thiel-Sen regression is a non-parametric estimation procedure of the theoretical model

$$\text{MP}_{Ai}|\text{MP}_{Bi} = \beta_0 + \beta_1 \text{MP}_{Bi} + \epsilon_i = g(\text{MP}_{Bi}) + \epsilon_i \tag{4.1}$$

where $\{\epsilon_i\}$ are independent random variables with mean zero. The function g evaluated at MP_{Bi} consists of the two regression coefficients we need to estimate: the regression slope and intercept. The benefit of using the Thiel-Sen estimator relies on the liberation of the homoscedasticity requirement. Moreover, these estimators are less affected by outliers and skewed data, which often is problematic in parametric regression. Robustness against skewed data implies robustness against the absence of normality. Thiel-Sen is consequently considered to be a robust regression model. However, we require a linear relationship of the clinical samples; thus, the Thiel-Sen estimator will not suffice at non-linear modeling.

Besides, the Thiel-Sen estimator will not account for variability in x-direction. Let

$$\Delta = \{\Delta_{(ij)}\} \tag{4.2}$$

where $\{\Delta_{(ij)}\}$ is the ordered sample of

$$\{\Delta_{ij}\} = \frac{|\text{MP}_{Ai} - \text{MP}_{Aj}|}{|\text{MP}_{Bi} - \text{MP}_{Bj}|} \; \forall \; \{i, j\} \in \{1, \ldots, n\} \times \{1, \ldots, n\}. \tag{4.3}$$

With n as the number of clinical samples then, [Wilcox, 1998] defines the slope estimator as

$$b_1 = \frac{\Delta\left\lfloor \frac{1}{4}(n(n-1)+2) \right\rfloor + \Delta\left\lceil \frac{1}{4}(n(n-1)+2) \right\rceil}{2}. \tag{4.4}$$

Note that $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling operators, respectively. As one might notice is that Equation (4.4) is just the median of all the possible slopes between every pair of observations. Furthermore, when b_1 is obtained, the intercept is estimated by

$$b_0 = \frac{\Xi\left\lfloor \frac{1}{2}(n+1) \right\rfloor + \Xi\left\lceil \frac{1}{2}(n+1) \right\rceil}{2}, \tag{4.5}$$

where Ξ is the ordered sample of $\{\text{MP}_{Ai} - b_1\text{MP}_{Bi}\}$ with respect to i. In other quarters, this is just the median of $\{\text{MP}_{Ai} - b_1\text{MP}_{Bi}\}$. Combining Equation (4.4) and Equation (4.5) for $i \in \{1, 2, \ldots, n\}$ produces the Thiel-Sen estimator of the expectation of Equation (4.1);

$$\hat{g}(\text{MP}_{Bi}) = b_0 + b_1\text{MP}_{Bi}. \tag{4.6}$$

4.2 Smoothing Splines

In circumstances where we drop assumptions regarding the shape of underlying distribution and linearity, smoothing splines will show to be very competent. The smoothing splines estimator is a non-parametric piece-wise regression estimator. That is, in opposition to Section 2.6, smoothing splines does not require a normal distribution between successive knots. [Craven and Wahba, 1978] defines the theoretical model as

$$\text{MP}_{Ai}|\text{MP}_{Bi} = g(\text{MP}_{Bi}) + \epsilon_i \tag{4.7}$$

where the $\{\epsilon_i\}$ are independent and identically distributed error terms with mean equal to zero. [James et al., 2013] declares that fitting a model for Equation (4.7) relies on minimizing

the penalty dependent residual sum of squares

$$\text{RSS} + \text{Penalty} = \sum_{i=1}^{n} \big(\text{MP}_{Ai} - g(\text{MP}_{Bi})\big)^2 + \lambda \int \Big[\frac{d^2 g(t)}{dt^2}\Big]^2 dt \tag{4.8}$$

with respect to g. The resulting minimum, $\hat{g}(\text{MP}_{Bi})$, is then the smoothing splines estimator. Note that Equation (4.8) implies that g is a cubic fit. [Opsomer and Breidt, 2011] extends to the p-degree polynomial fit by

$$\text{RSS} + \text{Penalty} = \sum_{i=1}^{n} \big(\text{MP}_{Ai} - g(\text{MP}_{Bi})\big)^2 + \lambda \int \Big[\frac{d^{p-1} g(t)}{dt^{p-1}}\Big]^2 dt. \tag{4.9}$$

In Equation (4.8) and Equation (4.9), λ is a so-called tuning parameter or smoothing parameter, and g is the smoothing function because g guarantees that the fitted models are smooth. Now, how do we obtain λ? In other words, how do we estimate λ? [James et al., 2013] utilizes the leave-one-out procedure for estimation of λ. This cross-validation method (LOOCV) relies on leaving out one of the MP_{Bi} so that we get the corresponding leave-one-out fit $\hat{g}(\text{MP}_{Bi})^{(-i)}$ for every $i \in \{1, 2, \ldots, n\}$. In this case, the leave-one-out cross validation method seeks the λ that minimizes

$$\text{RSS}_{\text{CV}}(\lambda) = \sum_{i=1}^{n} \big(\text{MP}_{Ai} - \hat{g}(\text{MP}_{Bi})^{(-i)}\big)^2 \tag{4.10}$$

for all n fits. This type of calculation is a numerical mathematical problem and is somewhat computationally demanding. Therefore, we will apply **??** in R using appropriate packages. We use the *npreg* package for this, where it is possible to specify which cross-validation procedure seems adequate. The sm() function's method parameter specification option is set to LOOCV (CV in *R*). Another method is the general cross-validation (GCV), but these two alternatives yield very similar results; hence we will stick to the one defined in Equation (4.10).

In Section 2.6, the knots were placed manually. This fact demanded visual knowledge regarding the data before we could settle where the knots should be. We argued that a general rule of thumb is to place the knots between successive control material samples. Now, we are not to bother with this cumbersome manual placement of the knots. The reason is that the smoothing splines procedure uses every value of the mean of replicated measurements along the x-axis as knots. This automatic knot placement will not result in an overfit because the penalty term in Equation (4.9) reduces the effective degrees of freedom [James et al., 2013]. There exist built-in functions in *R* to estimate the corresponding uncertainty bands for the model estimates. However, the packages do not include the estimation of prediction bands of the smoothing splines estimator. However, tolerance bands are implemented, which are very

similar to the corresponding prediction bands. Even though smoothing splines seem robust as a commutability assessment procedure, smoothing splines does not account for variability in the x-direction. As explained in Section 2.4, using the Bland-Altman transformation reduced the variability in x-direction. Thus, combining smoothing splines and Bland-Altman transformation seems preferable instead of estimating the raw data's relationship. It is important to note that an evaluation of auto-correlation and homoscedasticity is required before implementing this procedure.

4.3 Kernel Regression

Kernel regression is another non-parametric regression scheme, somewhat related to the smoothing splines method. As previously, the goal is to determine the best fit for the theoretical relationship

$$\text{MP}_{Ai}|\text{MP}_{Bi} = g(\text{MP}_{Bi}) + \epsilon_i,\ i \in \{1, 2, \ldots, n\}, \tag{4.11}$$

where $\{\epsilon_i\}$ are independent and identically distributed error terms with mean zero and variance $\sigma_A^2(\text{MP}_B)$. Taking expectations produces

$$\text{E}[\text{MP}_{Ai}|\text{MP}_{Bi}] = g(\text{MP}_{Bi}),\ i \in \{1, 2, \ldots, n\}, \tag{4.12}$$

which is the regression that is of interest to estimate [Racine, 2007]. In the estimation procedure, we use kernels allegedly to obtain $\hat{g}(\text{MP}_{Bi})$. Kernels are non-negative real-valued integrable functions. For most purposes, kernel functions, K_λ, are obliged to meet the following conditions [Racine, 2007]:

- $0 \leq K_\lambda(u) < \infty$,
- $\int_{-\infty}^{\infty} u^s K_\lambda(u) du < \infty\ \forall\ \{s : s \in \mathbb{N}\}$,
- $\int_{-\infty}^{\infty} K_\lambda(u) du = 1$ (normalization),
- $K_\lambda(a) = K_\lambda(-a)\ \forall\ a \in \mathbb{R}$ (symmetry).

For example, the Epanechnikov kernel may be defined by

$$K_\lambda(u) = \frac{3}{4}[1 - u^2],\ |u| \leq 1, \tag{4.13}$$

satisfies these requirements for sure. Note that the subscript λ defines the bandwidth for the particular kernel function. Several kernel functions are currently used. We will in this

text-only use the Gaussian kernel and the epanechnikov kernel, which are defined by:

$$K_\lambda(x_0 - \text{MP}_{Bi}) = \frac{1}{\sqrt{2\pi}} \exp\Big[-\frac{1}{2}\Big(\frac{x_0 - \text{MP}_{Bi}}{\lambda}\Big)^2\Big] \tag{4.14}$$

$$K_\lambda(x_0 - \text{MP}_{Bi}) = \frac{3}{4}\Big[1 - \Big(\frac{x_0 - \text{MP}_{Bi}}{\lambda}\Big)^2\Big], \ |\frac{x_0 - \text{MP}_{Bi}}{\lambda}| \leq 1. \tag{4.15}$$

Equation (4.14) and Equation (4.15) ultimately tells us that larger weights are assigned for small $x_0 - \text{MP}_{Bi}$. This fact implies that larger deviance between x_0 and MP_{Bi} will result in less assigned weight. Moreover, larger bandwidths will flatten the graph of the kernel and vice versa. See illustration below:

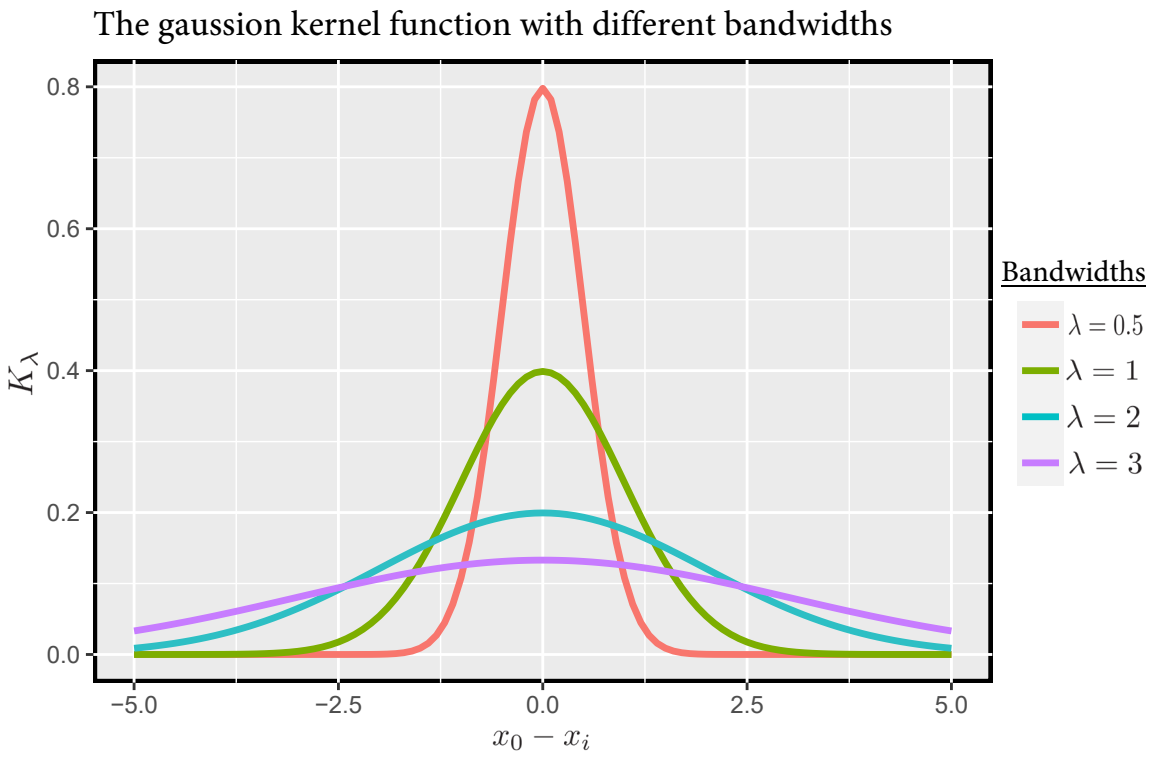

Figure 4.1 *– An overview of the Gaussian kernel for four different bandwidths.*

The bandwidth is also interpreted as the smoothing-parameter. This is because the bandwidth is directly related to the variability and the smoothness of our fitted kernel regression model. Small λ indicates wiggly fits (potential overfitting and large variance), whereas large λ proposes smooth fits (potential underfitting and large bias) [Hastie, Tibshirani, and Friedman, 2009]. It is imperative to obtain a value for λ that balances both variance and bias. These smoothing parameter interpretations are equivalent to what we pronounced for smoothing splines in Section 4.2. One might question why we need kernels. Kernels work

as a tool for calculating weights for all data points concerning a reference point x_0. That is, each data point will be assigned a weight between 0 and 1 depending on how close each point is to x_0. Since we want values from 0 to 1, it is clear that probability functions are favorable as kernels. When a kernel is chosen we use it to estimate the conditional expectation $\mathrm{E}[\mathrm{MP}_{Ai}|\mathrm{MP}_{Bi} = \mathrm{MP}_{Bi}] = g(\mathrm{MP}_{Bi})$. For instance, Nadaraya and Watson proposes the point estimator for g evaluated at x_0 as

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} K_\lambda(x_0 - \mathrm{MP}_{Bi})\mathrm{MP}_{Ai}}{\sum_{i=1}^{n} K_\lambda(x_0 - \mathrm{MP}_{Bi})}. \tag{4.16}$$

Given a specific kernel (e.g., the Gaussian kernel), several utilized estimators exist for the expectation of the theoretical relationship between the response variable and explanatory variables. Another popular estimator is the Gasser–Müller kernel estimator, which [Hastie et al., 2009] defines by

$$\frac{1}{\lambda}\sum_{i=1}^{n} y_{(i)} \int_{t_{i-1}}^{t_i} K_\lambda\Big(\frac{x_0 - y}{\lambda}\Big)\, dy. \tag{4.17}$$

Unfortunately, there is no analytical approach to estimate the prediction bands, so *variability bands* will now be used. *Variability bands* are defined by

$$\mathrm{VB} = \hat{g}(x) \pm 3 \cdot \mathrm{SD}(x). \tag{4.18}$$

With this definition in mind, we regard a commutability assessment procedure concerning measurement procedures *Sysmex XN* and *Advia 2120i*:

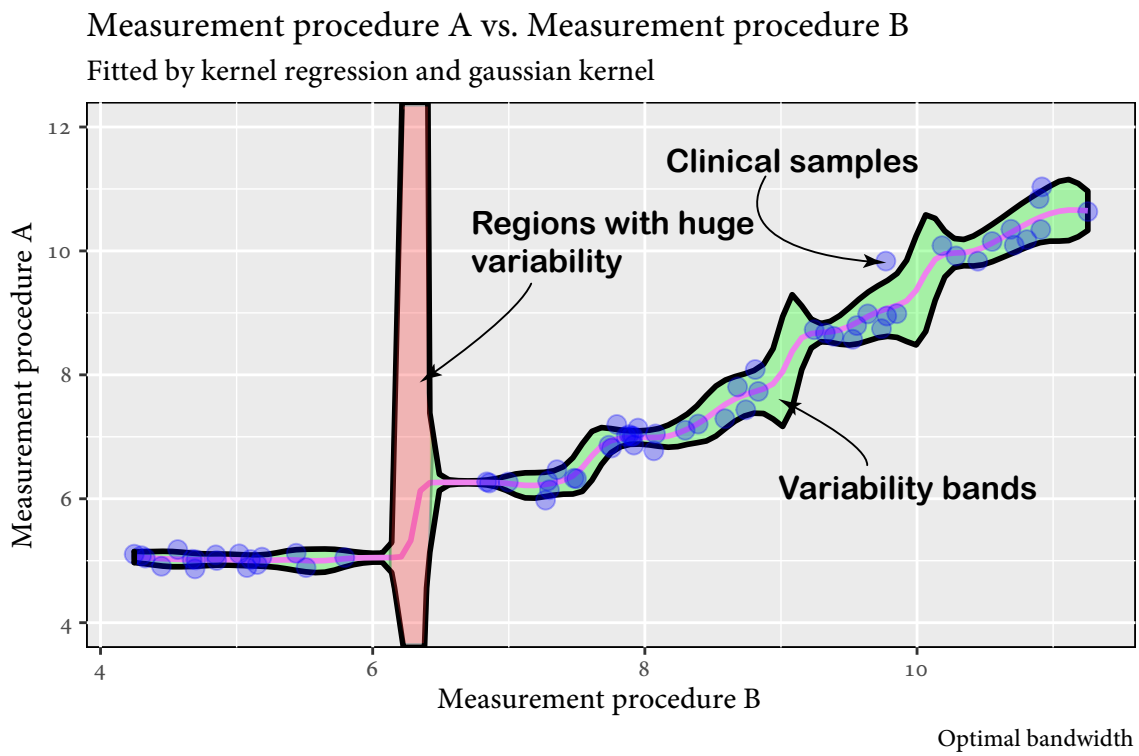

Figure 4.2 *– Nadaraya-Watson estimator constructed by clinical samples and the Gaussian kernel. The optimized bandwidth is calculated by CV and is the so-called optimal.*

As we demonstrated before, we saw that a small λ would result in wiggly fits, which significantly increases the model's variance. The fitted values' standard error was much larger than presented in Figure 4.2. However, we restricted the variability between 0 and 20 by convenience. In reality, the most massive standard errors were as significant as 10^{40}. Thus, the issue regarding choosing an insufficient λ will have drastic consequences. Let us, therefore, explore the possibilities concerning the choice of λ. In the succeeding assessment plot, we have increased the bandwidth to 1.

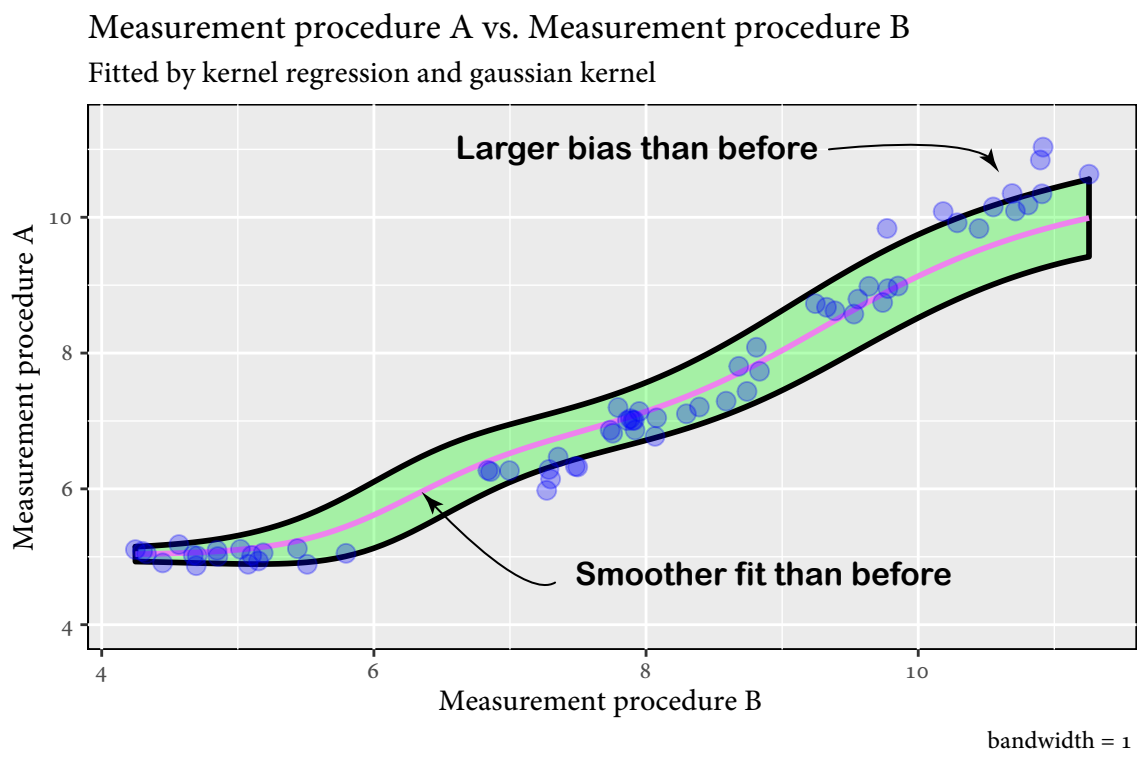

Figure 4.3 *– Nadaraya-Watson estimator constructed by clinical samples and the Gaussian kernel. Using manually chosen bandwidth - $\lambda = 1$.*

By choosing a more substantial bandwidth, we see that we get a considerably smoother fit, which diminished the prediction errors significantly at the cost of model bias. In this case, the rise in bias is considerable. The bias-variance trade-off is the reason for this because the trade-off relationship is close to quadratic. Since $\lambda = 1$ is approximately seven times larger than $\lambda_{\text{OPT}} \approx 0.14$, the bias measured by MSE will be a little less than 49 times larger than the optimal bias. See the simulation studies in Section 5.5.2 for further details on this relationship.

One might debate whether the variability bands are adequate as accepting criterion regarding the control materials' commutability property. We perceive that some clinical samples are exceeding the variability bands. Bootstrapping confidence intervals is another procedure we can implement. Alternatively, we could apply the asymptotic properties of Equation (4.16) (the Nadaraya-Watson estimator).

4.3.1 Confidence and prediction bands

Constructing prediction intervals are essential in this text because our assessment acceptance criteria rely on these. Let us firstly consider how we can construct asymptotic confidence

intervals and extend these to prediction intervals. [Racine, 2007] defines the approximate variance and bias of Equation (4.16) as

$$\begin{aligned}\operatorname{Var}\big[\hat{g}(x)\big] &\approx \frac{\hat{\sigma}^2(x)}{n\lambda \cdot \hat{f}(x)} \int K_\lambda(z)^2 dz \\ \operatorname{B}\big[\hat{g}(x)\big] &\approx \frac{\lambda^2}{2}\frac{d^2\hat{g}(x)}{dx^2} \int z^2 K_\lambda(z)\, dz + \lambda^2 \frac{\frac{d\hat{f}(x)}{dx} \cdot \frac{d\hat{g}(x)}{dx}}{\hat{f}(x)}.\end{aligned} \tag{4.19}$$

These are accurate approximations for sufficiently large n. How many observations are needed are not certain. However, we will go into this in more detail in the simulation studies of the next section. Moreover, $\hat{\sigma}^2(x)$ and $\hat{f}(x)$ are the estimated conditional prediction errors and estimated marginal density of X which [Muller, 1998] defines by:

$$\begin{aligned}\hat{\sigma}(x)^2 &= \frac{\sum_{i=1}^{n} K_\lambda\Big(\frac{\mathrm{MP}_{Bi}-x}{\lambda}\Big)\tilde{e}_i^2(x)}{\sum_{i=1}^{n} K_\lambda\Big(\frac{\mathrm{MP}_{Bi}-x}{\lambda}\Big)} \\ \hat{f}(x) &= \frac{1}{nb}\sum_{i=1}^{n} K_b\Big(\frac{\mathbf{MP}_{Bi}-x}{b}\Big).\end{aligned} \tag{4.20}$$

It then follows directly from Equation (4.20) that we can approximate the estimated confidence bands by exploring the asymptotic distribution of $\hat{g}(x)$:

$$\frac{\hat{g}(x) - \mathrm{E}[\hat{g}(x)]}{\sqrt{\mathrm{Var}[\hat{g}(x)]}} = \frac{\hat{g}(x) - g(x) - \mathrm{B}[\hat{g}(x)]}{\sqrt{\mathrm{Var}[\hat{g}(x)]}} \overset{d}{\sim} \mathcal{N}(0,1). \tag{4.21}$$

Equation (4.21) results from the central limit theorem via Lindeberg's condition. A proof of Equation (4.21) will not be presented in this text. See [Ould-Saïd and Lemdani, 2006] for details concerning the proof. By looking at the approximation

$$\mathrm{P}\Big(z_{\alpha/2} \leq \frac{\hat{g}(x) - g(x) - \mathrm{B}[\hat{g}(x)]}{\sqrt{\mathrm{Var}[\hat{g}(x)]}} \leq -z_{\alpha/2}\Big) \approx 1-\alpha, \tag{4.22}$$

one might derive that the approximate $(1-\alpha)\cdot 100\%$ point-wise confidence intervals are given by

$$\mathbf{CI}(x) \approx \hat{g}(x) - \mathrm{B}[\hat{g}(x)] \pm z_{\alpha/2}\sqrt{\mathrm{Var}[\hat{g}(x)]}. \tag{4.23}$$

As expressed in [De Brabanter, De Brabanter, Suykens, and De Moor, 2011], the approximated $(1-\alpha)\cdot 100\%$ prediction intervals are given by:

$$\text{PI}(x) = \hat{g}(x) - \text{B}[\hat{g}(x)] \pm z_{\alpha/2}\sqrt{\text{Var}[\hat{g}(x)] + \hat{\sigma}^2(x)}. \tag{4.24}$$

As a general rule, we ignore the bias term in Equation (4.23) and Equation (4.24) when possible. However, if we use larger bandwidths than the optimal bandwidth, it is clear that the prediction bands are overestimated because of ignoring the bias term. Therefore, it will be of interest to estimate this bias. Nevertheless, doing this in a precise analytical form is a challenging task. Consequently, we will use a bootstrap resampling technique to get the point-wise estimates of the bias. We may use that

$$\text{B}(\widehat{\hat{g}(\text{MP}_{Bi})}) = \frac{1}{B}\sum_{j=1}^{B}\hat{g}_{j*}(\text{MP}_{Bi}) - \hat{g}(\text{MP}_{Bi}),\ \forall\, i \in \{1,2,\ldots,n\} \tag{4.25}$$

Moreover, we will implement an R-function that calculates the prediction bands utilizing Equation (4.24). See Appendix A for a description. Furthermore, we will bias correct the prediction bands using Equation (4.25). However, the bias correction will only work properly when using MOR. So, we will primarily stick to MOR. We have to stick to MOR either way because of induced auto-correlation of replicated measurements, so this is nothing new. Note that using MOR will have a dramatic consequence; We get fewer data points, which means that n may not be large enough for achieving accurate estimates using Equation (4.24). As recently mentioned, we can also bootstrap confidence intervals. This estimation is, however, more onerous and will be impractical in the simulation studies to come. Furthermore, confidence intervals are not sufficient in a commutability assessment situation. Consequently, it will be no use in discussing this further. As with smoothing splines, Kernel regression does not account for variability in x-direction either. However, we may Bland-Altman transform the clinical samples so that the variability in x is reduced moderately. Thus a combination of Bland-Altman transformation and Kernel regression seems prudent here as well.

4.4 Kernel regression with two - dimensional variability

In the previous section, kernel regression was considered an alternative non-parametric assessment procedure for evaluating the control materials' commutability property. However, the much-discussed flaw was persistent for this method as many others, which only account for variability in y-direction. Modern research within econometric theory has developed estimators for the expectation of Equation (4.11) that include variability in x-direction. The

reader is referred to [Dabo-Niang and Thiam, 2019] for details regarding this research. In light of this being modern research, no useful methods for constructing model prediction bands yet exist. Besides, manual choosing of the bandwidths will still be required as part of commutability assessment. In future research, it will be interesting to discuss this article in further detail.

Chapter 5

Applying Commutability Assessment with Non-Parametric Methods

5.1 Data

Let us test our recently developed methods for commutability evaluation. The data sets we use are the very same as in Section 3.1. We will analyze the same data sets we considered before and compare the new results with the old results. We will also compare the performance of smoothing splines and Kernel regression as commutability assessment procedures. The two data sets *EPK*, and *LPK* are of interest. We will compare the outcomes of the commutability assessment and see how well our non-parametric assessment methods work.

5.2 Commutability assessment with Thiel-Sen Regression

We remember that the EPK data followed a linear pattern considerably. Nevertheless, recall that the linear model assumptions were troublesome to satisfy. Normality and heteroscedasticity remained a recurring problem. Therefore the next step could be to use Thiel-Sen Regression. As addressed in Section 4.1, heteroscedasticity is less of a problem for Thiel-Sen fitted models. Furthermore, we are not assuming any specific distribution. Even though auto-correlation will be less of a problem here, it is still best to avoid modeling with it. As we see in Section 3.4 we get that auto-correlation is no query for this data set when using MOR. Therefore we may proceed and draw the corresponding six commutability assessment plots. See the figures below for the assessment plots. See Figure 5.1 and Figure 5.2 for commutability assessment plots. We discern that the prediction bands are tight for these measurement procedure comparisons. We see that *LQ* lies beyond the prediction bands, and

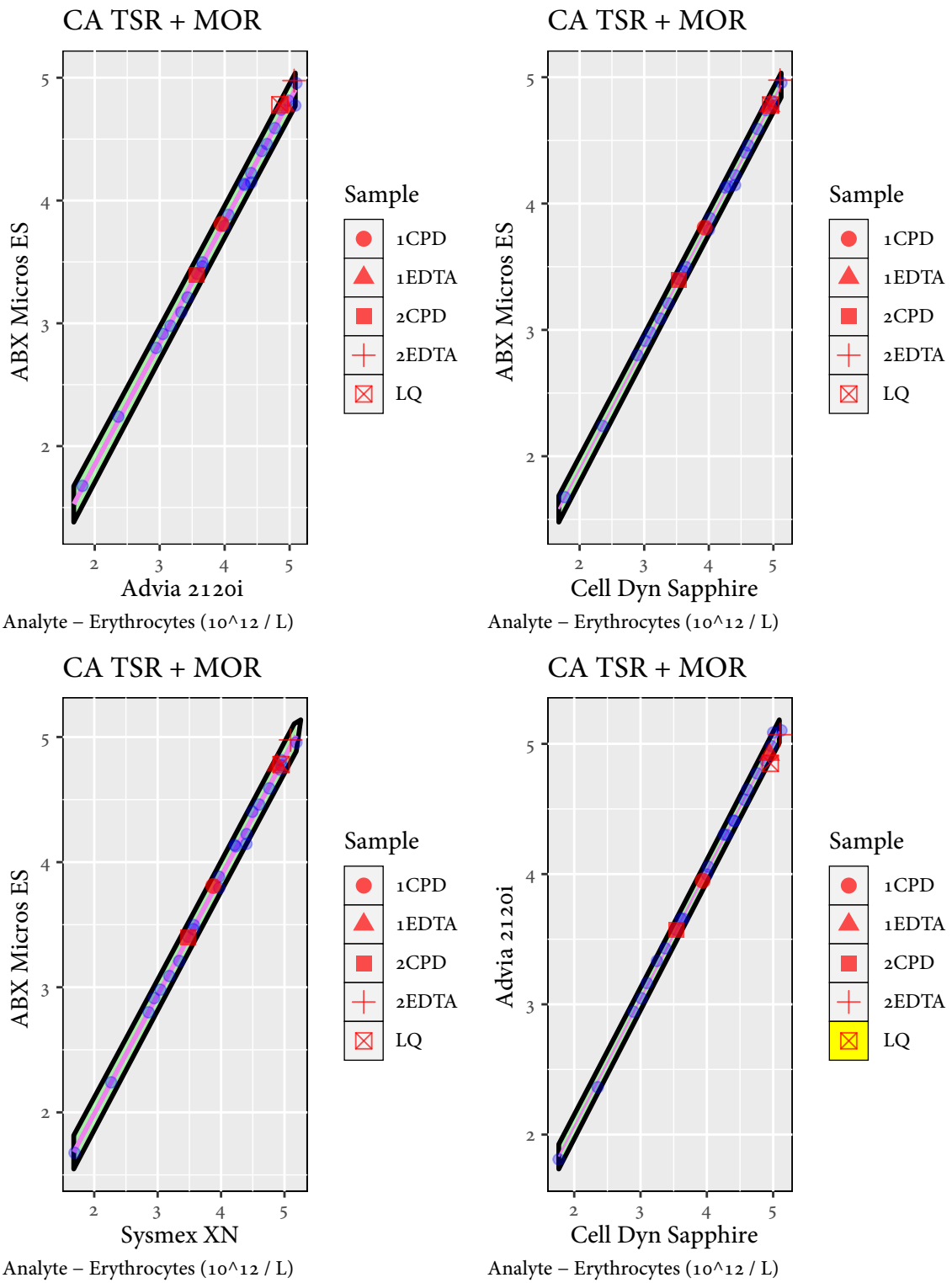

Figure 5.1 *– Visual commutability assessment with clinical samples fitted by the Thiel-Sen regression model. The means of replicates are used.*

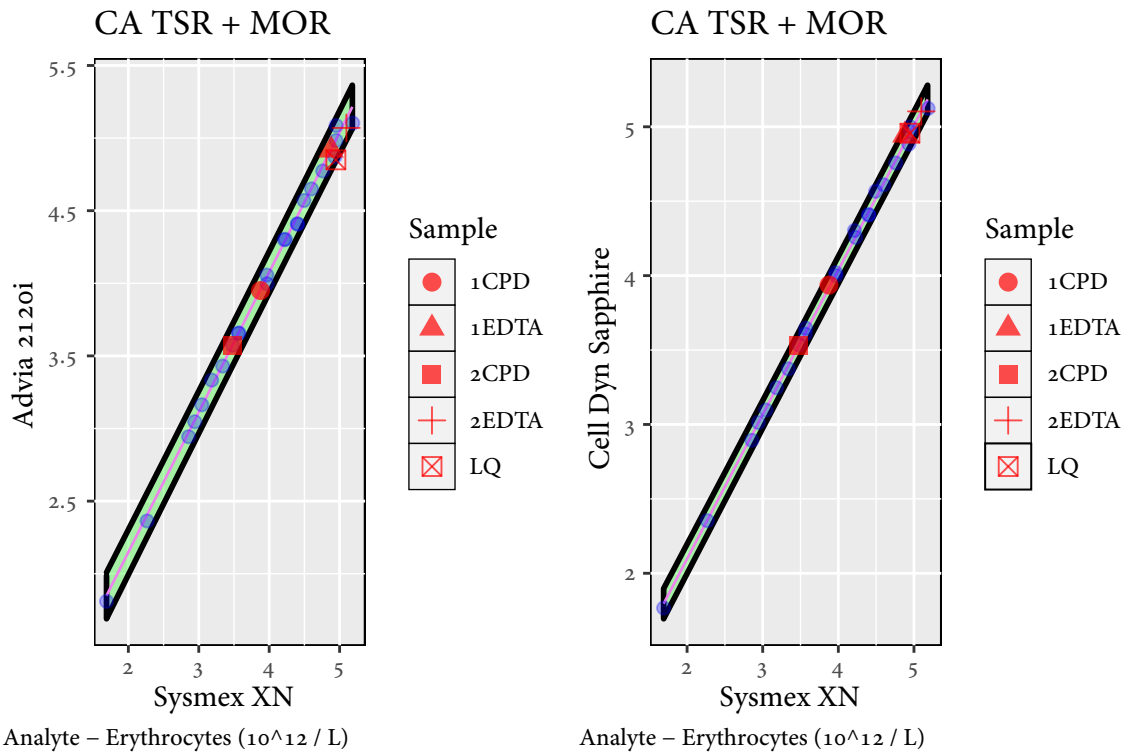

Figure 5.2 *– Visual commutability assessment with the clinical samples fitted by the Thiel-Sen regression estimator. The means of replicates are used.*

therefore concluded as non-commutable regarding *Advia 2120i* versus *Cell Dyn Sapphire*. In Section 3.4, we concluded that *LQ* was commutable. The reason might be the variability in x-direction made us conclude differently concerning the two different approaches. The rest of the plots all propose acceptance for commutability for all control materials. Relaxation of the linear model assumptions makes Thiel-Sen Regression robust as a commutability assessment procedure. The faults are not accounting for variability in x-direction and the linearity requirement. We will now consider the *LPK* data set, which is known to possess non-linear relations. It will be interesting to examine how adequately Thiel-Sen Regression copes with the non-linearity *LPK* has to offer. We

The linear model assumptions are mostly acceptable, but we face difficulties regarding the clinical samples' linearity for *ABX Micros ES* versus *Cell Dyn Sapphire* and *Cell Dyn Sapphire* versus *Sysmex XN*. The break-of for the homoscedasticity property is a minor issue since Thiel-Sen Regression is considered robust in that way. However, we can only use Thiel-Sen regression to view four of these measurement procedures as commutability assessments because of the lack of linearity. Hence, we may not believe the estimated prediction bands,

Table 5.1 *– Evaluation of the linear model assumptions of raw clinical samples fitted by the Thiel-Sen regression estimator. The means of replicates are used.*

Linear model assumptions	*LPK*			
TSR + MOR	**Normality**	**Homoscedasticity**	**Auto-correlation**	**Linearity**
ABX Micros ES vs. Advia 2120i				
ABX Micros ES vs. Cell Dyn Sapphire				**
ABX Micros ES vs. Sysmex XN		*		
Advia 2120i vs. Cell Dyn Sapphire				
Advia 2120i vs. Sysmex XN				
Cell Dyn Sapphire vs. Sysmex XN				*

resulting in erroneous conclusions regarding the control materials' commutability property. After this, we plot the remaining four comparisons with accepted linear model assumptions.

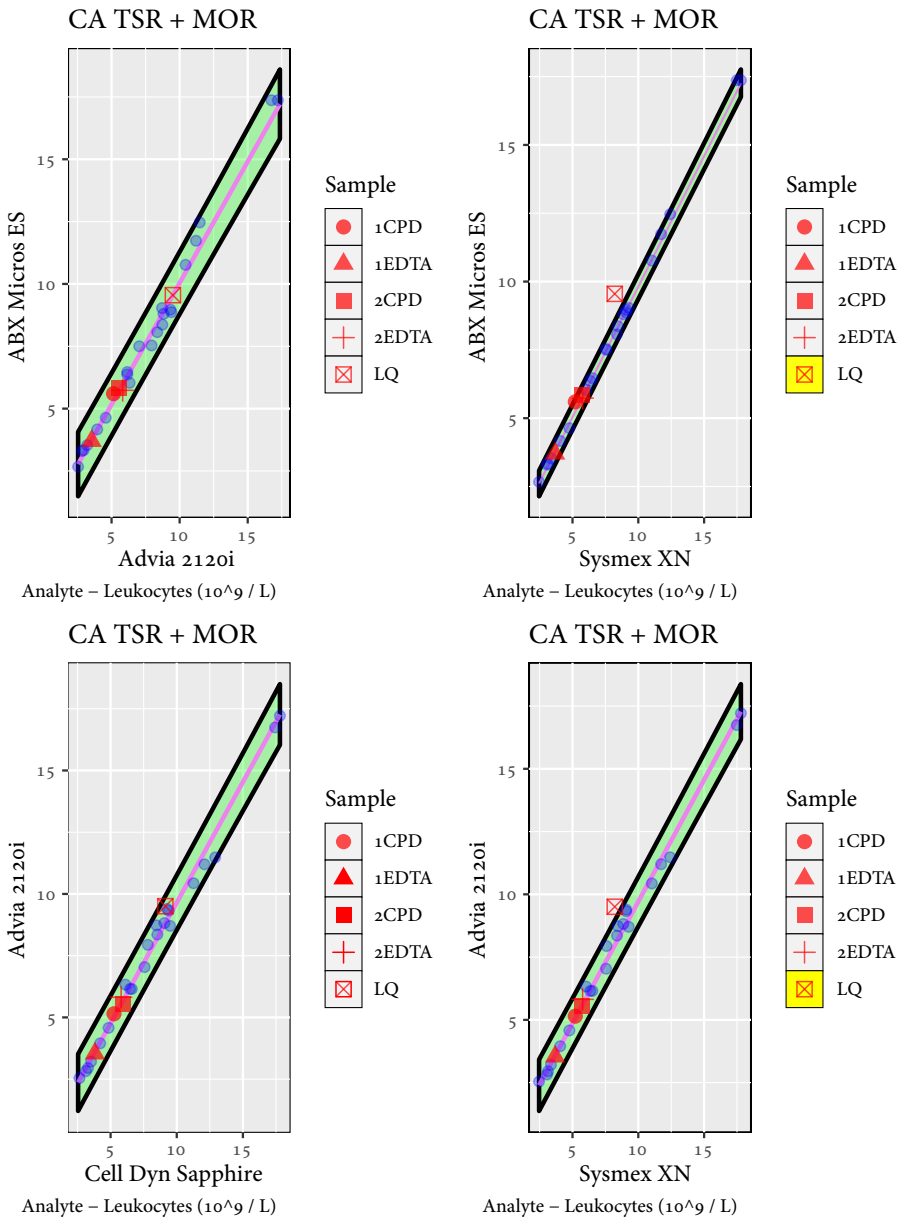

Figure 5.3 *– Visual commutability assessment for the clinical samples in the LPK data set fitted by Thiel-Sen regression estimators on those comparisons satisfying the linear model assumptions.*

We reject the commutability of *LQ* for multiple measurement procedure associations. We have learned that non-linearity is a significant issue for Thiel-Sen regression, similar to parametric regression models. The natural next step is, therefore, to consider non-parametric methods dealing with non-linearity.

5.3 Commutability assessment with Smoothing Splines

As we learned in Section 4.2, there exist comprehensive methods dealing with non-linearity in data. One of these methods is smoothing splines, which is a non-parametric piece-wise regression scheme. Moreover, we learned that we did not need to specify the locations or number of knots, as required in Section 2.6. We will use every observation along the x-axis as knots. This knot-selection procedure will make the fitted model flexible in areas where many data points are present. Consequently, not as flexible in gaps among data. Using one knot for every observation implies that we have n knots. Accordingly, we get $n - 1$ local model fits when using the minimum and maximum of x-values as boundaries. We argued that we could not use Thiel-Sen Regression to evaluate the control material samples' commutability for *ABX Micros ES* versus *Cell Dyn Sapphire* and *Cell Dyn Sapphire* versus *Sysmex XN*. The reason for this was the non-linearity of clinical samples in our measurement procedure comparisons. However, linearity is no longer an issue using smoothing splines. Thus, we may evaluate the commutability of the control materials for all measurement procedure comparisons.

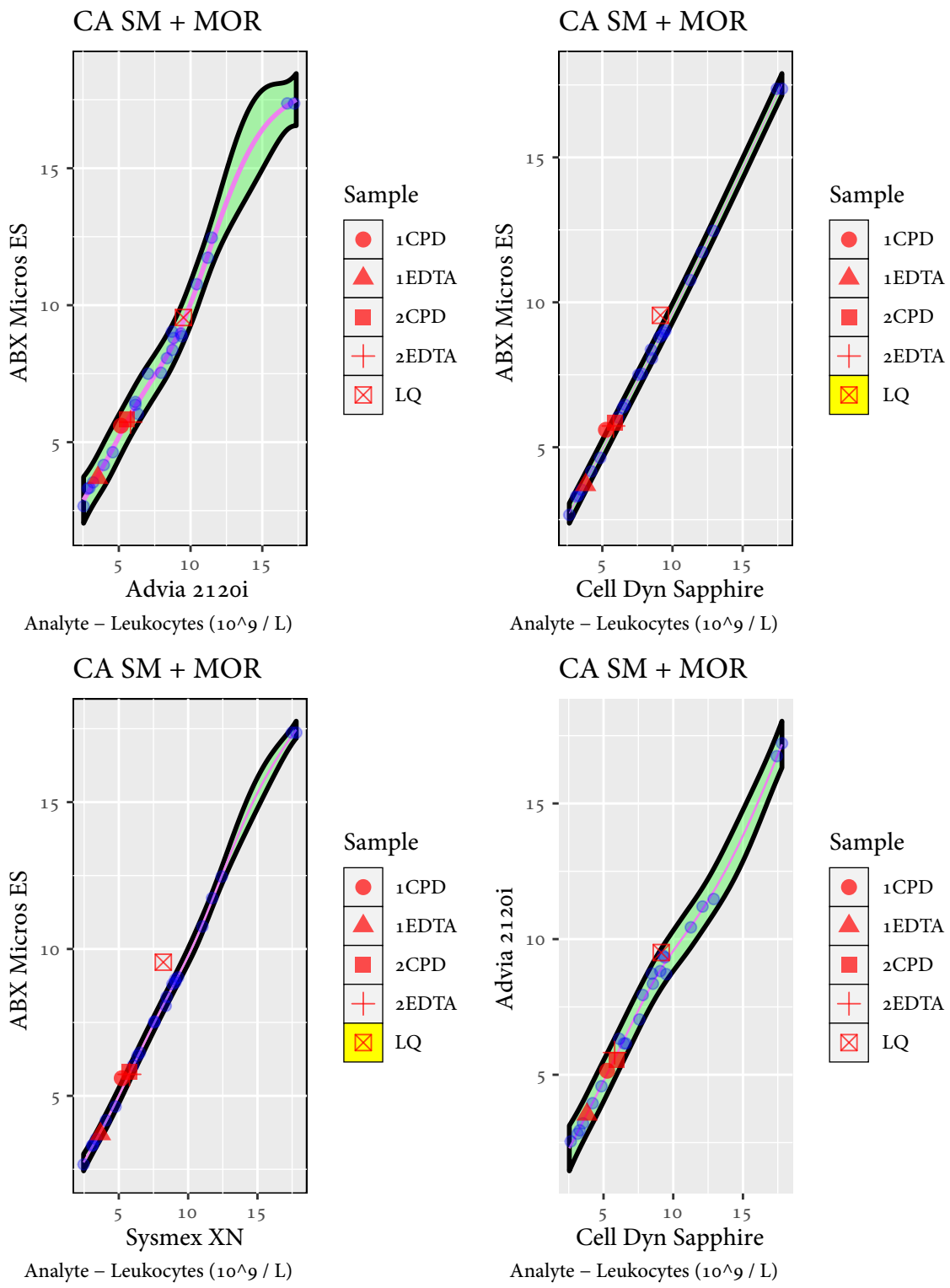

Figure 5.4 *– Visual commutability assessment with clinical samples fitted by smoothing splines regression estimators. The means of replicates are used.*

LQ is, as anticipated, still concluded as not commutable for most of the measurement procedure comparisons. LQ is only accepted for *ABX Micros ES* versus *Advia 2120i* and *Advia 2120i* versus *Cell Dyn Sapphire*. See Figure 5.5 for the two outstanding assessment plots.

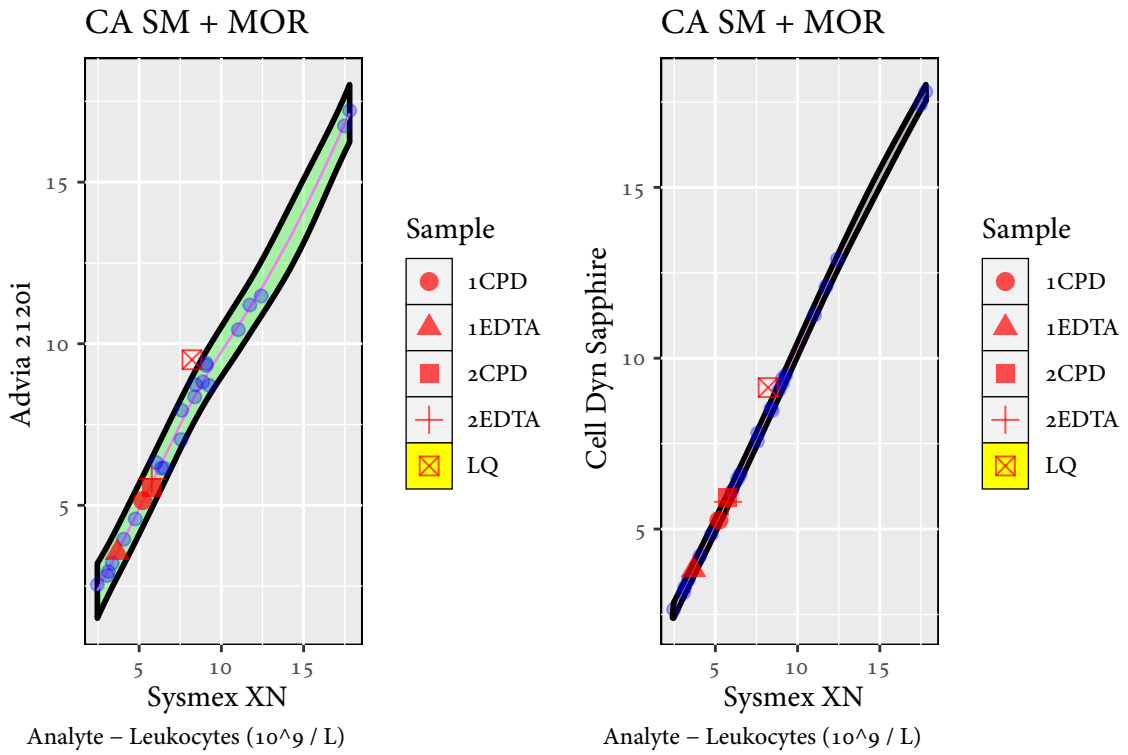

Figure 5.5 *– Visual commutability assessment with clinical samples fitted by smoothing splines regression estimators. The means of replicates are used.*

It is somewhat difficult to decide from a visual perspective whether *1CPD* is inside some of the plots above. We see that smoothing splines were capable of dealing with non-linearity with ease. However, we are facing the standard-issue again. We still have variability in x-direction, which is ignored. As mentioned in Section 4.2, we may Bland-Altman transform the raw data and use the resulting data to fit a smoothing splines model. As stated in Section 2.4, this transformation will reduce the variability in x.

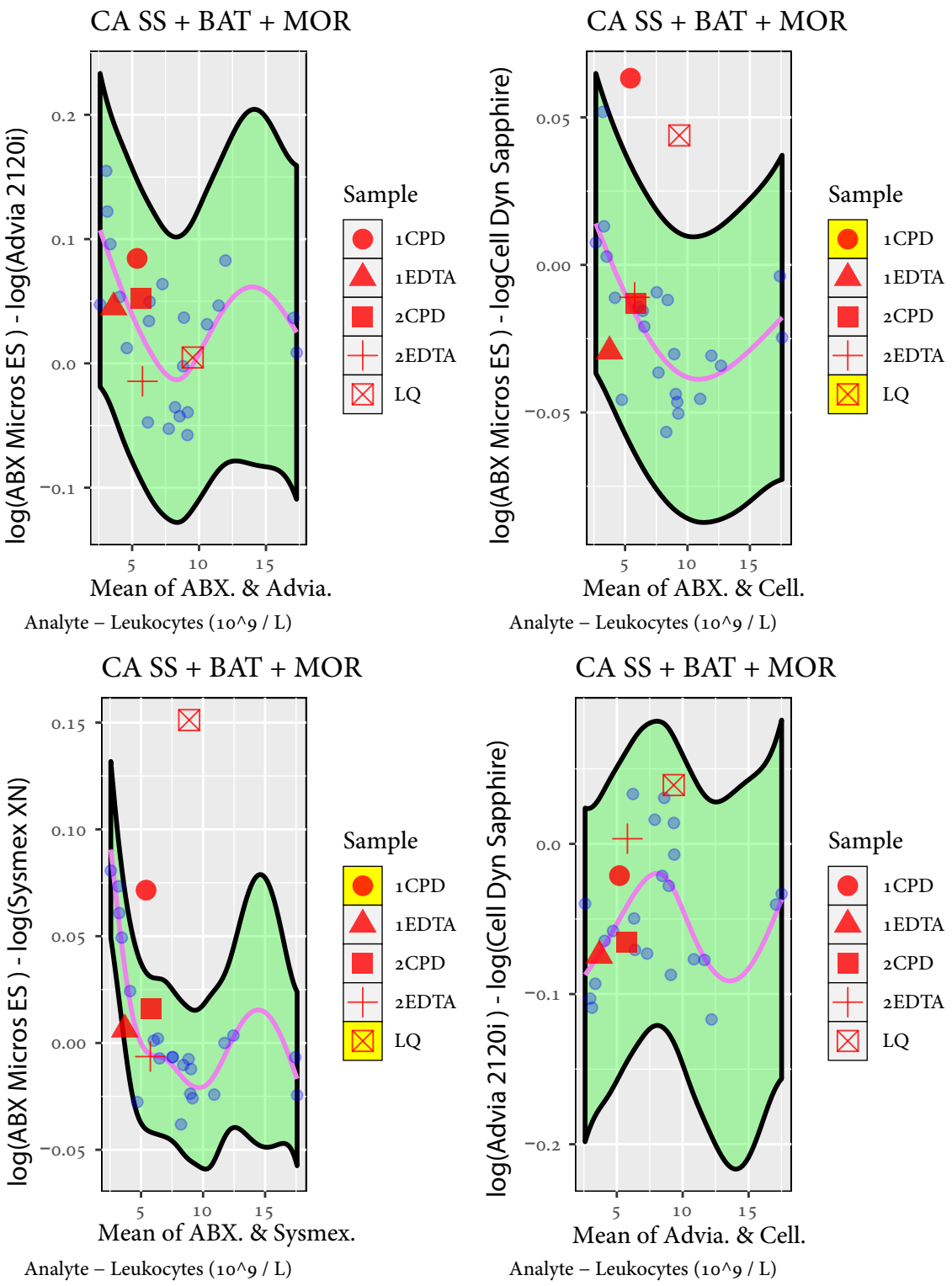

Figure 5.6 *– Visual commutability assessment with Bland-Altman transformed clinical samples fitted by smoothing splines estimators. The means of replicates are employed.*

One immediate observation is that visual interpretation is much easier because the prediction bands are relatively larger than with using raw data. It is difficult to determine whether *1CPD* is inside or outside for *ABX Micros ES* versus *Cell Dyn Sapphire* and *ABX Micros ES* versus *Sysmex XN* in Figure 5.4. Nevertheless, *1CPD* is concluded as non-commutable when using Smoothing splines with Bland-Altman transformed data. We see that *LQ* is also rejected for the remaining two assessment plots:

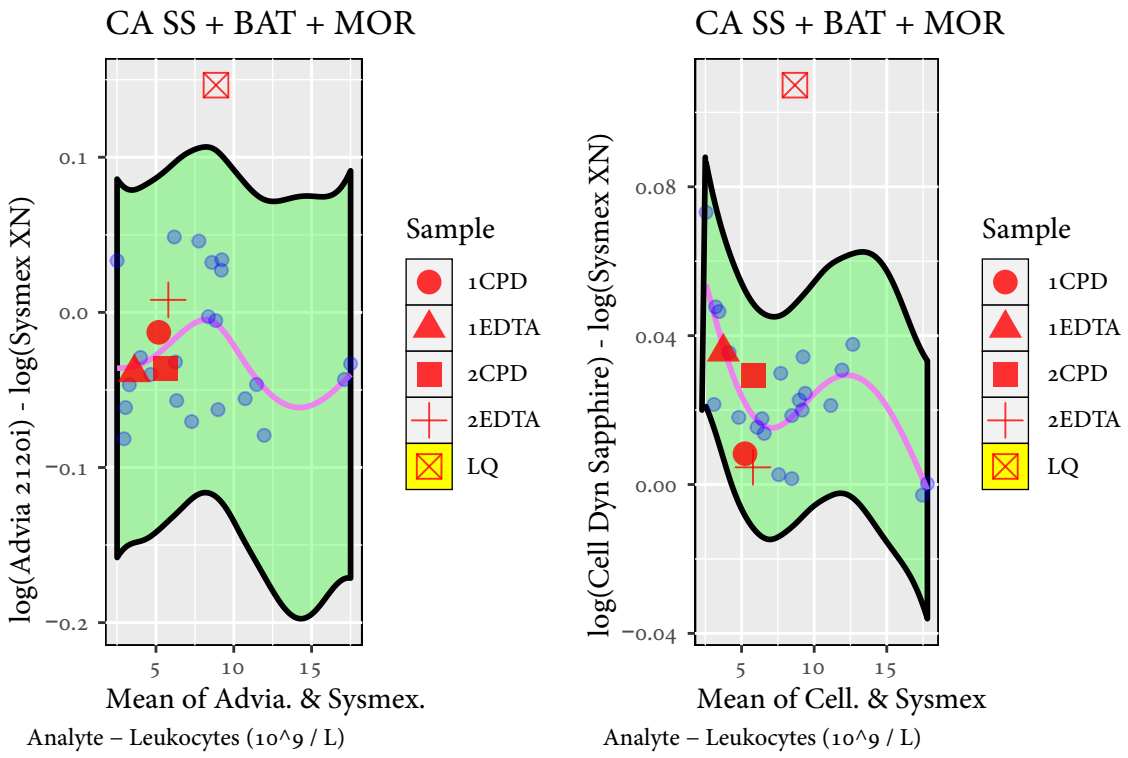

Figure 5.7 *– Visual commutability assessment with Bland-Altman transformed clinical samples fitted by smoothing splines estimators. The means of replicates are used.*

5.4 Commutability assessment with kernel regression

In the previous sections, we presented methods of fitting a regression model using kernels. We also implemented an approximation scheme for prediction intervals, which is essential in our commutability evaluation procedure regarding control materials. We will now use the theory to see how well kernel regression works as a commutability assessment method. As before, we use the 99% prediction bands as the acceptance region for the control materials. We will have a new look at the *EPK* data set we have worked on before. The *EPK* data set measurements are linear, and consequently, it will be interesting to see how well kernel

regression performs compared to the previously applied assessment methods. The visual commutability assessment plots for the data set are presented in Figure 5.8. In the plots, one might notice regions with few clinical samples, which affect the assessment performance because of vast prediction bands in these regions. The wide width might additionally be due to having too few clinical samples. As we will see in Section 5.5.2, very large n are required to classify non-commutable control materials as non-commutable concerning the applied acceptance criterion. Usually, the sample size is required to be as large as $n = 75$ for the method to reject control materials three standard deviations from the theoretical regression line. By considering the *LPK* data set, the misclassification issue arises because of non-linearity, poorly scattered data, and too few clinical samples. See Figure 5.9 for visual evidence for this claim. Recall that *LQ* was frequently rejected for this data set, which is not the case when using kernel regression on the mean of replicates with Gaussian kernel and optimized bandwidth. Note that *1CPD* and *2CPD* lie within one of the red regions of the prediction bands, which implies that these control materials' classification is unreliable. Vast prediction bands like these are unfortunately typical for kernel regression and are among the main arguments for avoiding the kernel regression approach. A proposed solution for reducing the width of the prediction bands is to use larger bandwidths. However, scaling the optimal bandwidth by scalars larger than 1 typically increases model bias. Growth of model bias is approximately a quadratic function of the scales, as we will see in Section 5.5.2.

We still end up with a significant problem. Variability in x-direction is not accounted for, which may provoke untrustworthy fits. To deal with this issue, one might Bland-Altman transform the clinical samples such that some variability in x diminishes. For amplification, the reader is referred to Section 4.2 and Section 3.6. For completeness sake, it is essential to study the combination of Bland-Altman transformation and kernel regression to observe if it fixes some of the before-mentioned faults. The plots in Figure 5.10 illustrates this approach. One might observe that the Bland-Altman transformed clinical samples of the *Cell Dyn Sapphire* vs. *Sysmex XN* comparison were much better than the non-transformation approach. Generally, as stated before, Bland-Altman plots are easier to interpret. Also, four of the plots resulted in *LQ* was rejected as commutable. From a visual perspective, Bland-Altman and kernel regression's combination appears better than the equivalent non-transformation approach. To obtain even smoother prediction bands, one could use larger bandwidths than the proposed optimal, and another kernel could be used. A more detailed discussion on these choices is described in more detail in Section 5.5.2. We consider two times the optimized bandwidth as an alternative to the optimal bandwidth.

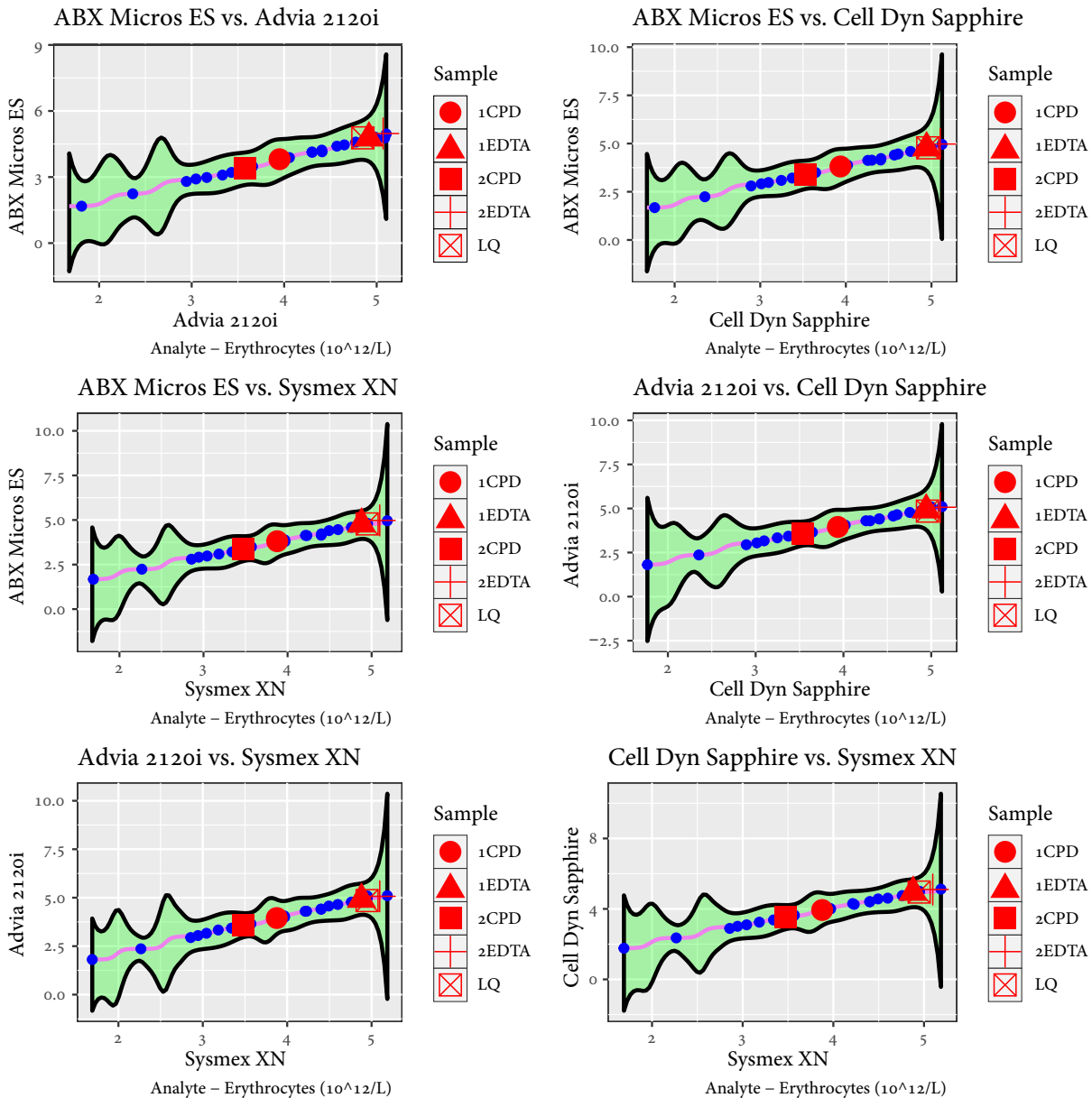

Figure 5.8 *– Visual commutability assessment with clinical samples from the EPK data set fitted by kernel regression estimators where optimal bandwidths and Gaussian kernels are used. The means of the replicates are used.*

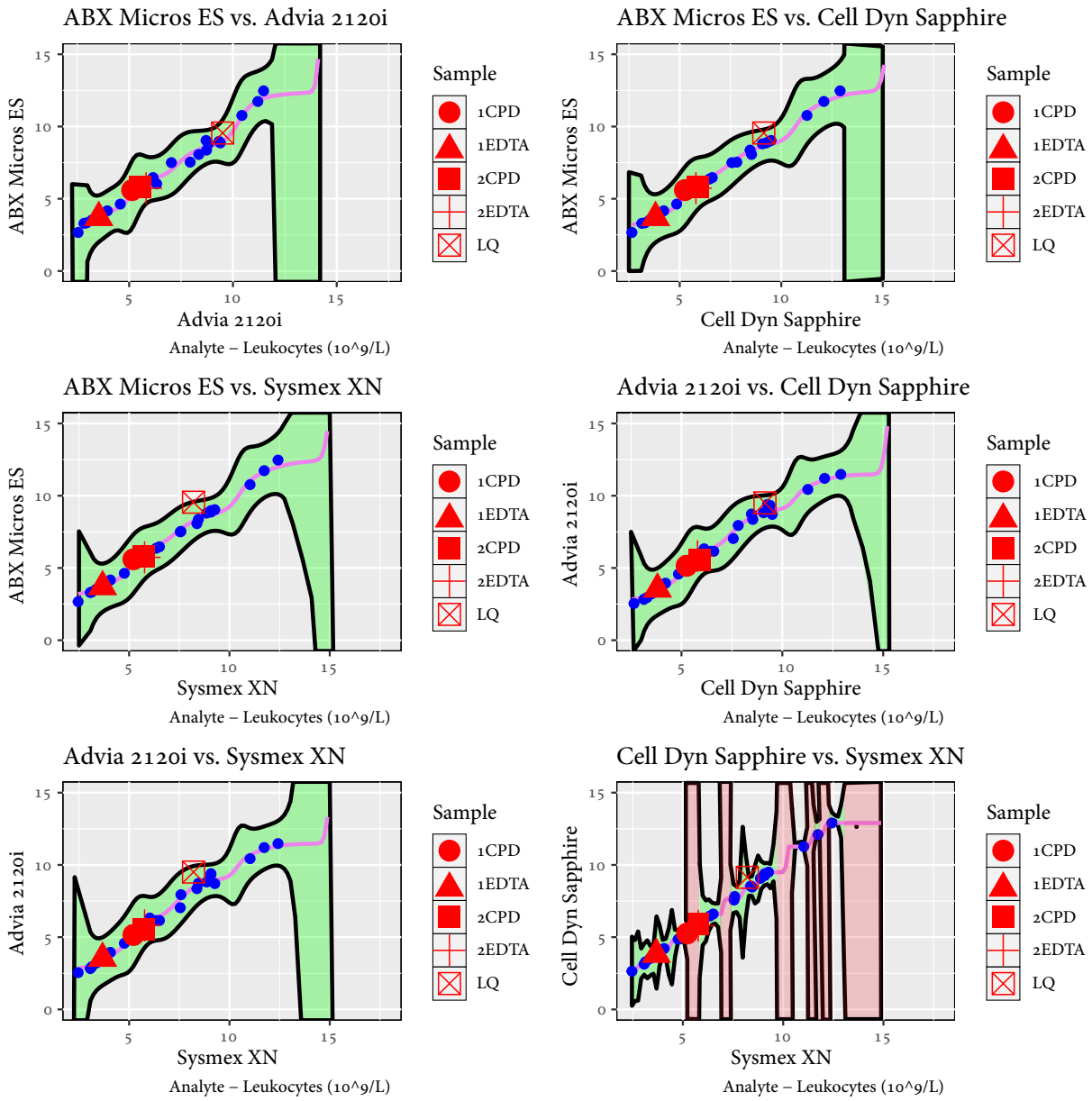

Figure 5.9 *– Visual commutability assessment with clinical samples from the LPK data set fitted by kernel regression estimators where optimal bandwidths and Gaussian kernels are used. The means of the replicates are used.*

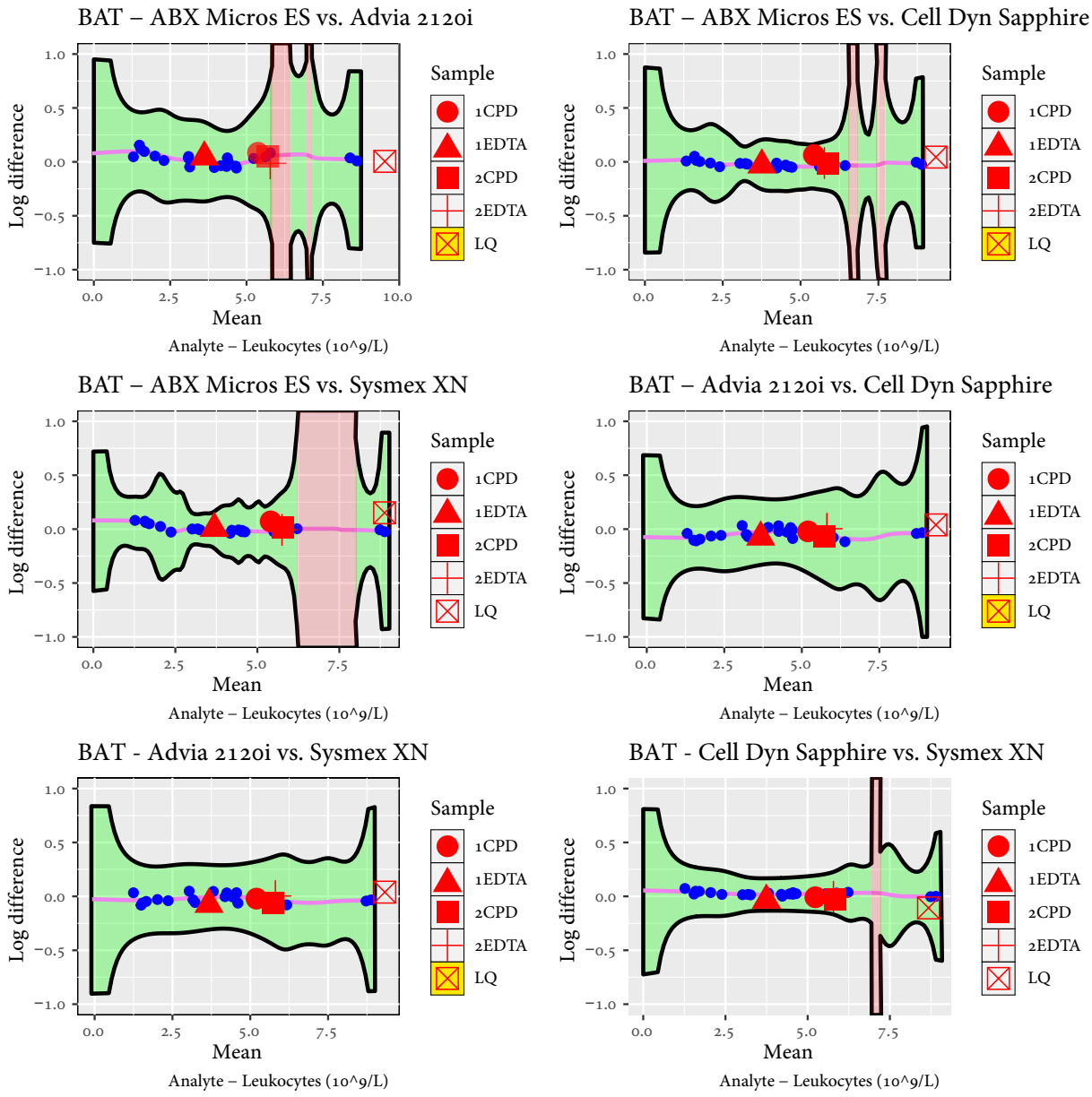

Figure 5.10 *– Commutability assessment with Bland-Altman transformed clinical samples from the LPK data set fitted by kernel regression with Gaussian kernels and optimal bandwidths.*

5.5 Non-parametric evaluation methods - Simulation

It is again time to do some simulation studies. The core will be slightly different from what we had in Section 3.6. The method of generating data is moderately the same as before. However, different levels of non-linearity will not be of importance here. Preferably, data sets are simulated, such that the clinical samples follow either a straight line or a particular non-linear curve. We will qualitatively test whether a non-linear curve will result in more inadequate assessment procedures for commutability than a linear pattern. As our non-parametric methods are meant to handle non-linearity well, it will be adequate to compare the two different pattern cases suggested in Figure 5.11. In these simulation studies, the measurements of the clinical samples following a straight line will be simulated by

$$\begin{aligned} \text{MP}_{Bir} &= \tau_{Ai} + \mathcal{N}(0, \sigma_B^2), \\ \text{MP}_{Air} &= f(\tau_{Ai}) = a + b \cdot \tau_{Ai} + \mathcal{N}(0, \sigma_A^2). \end{aligned} \tag{5.1}$$

Equation (5.1) is defined for $r \in \{1, \ldots, R\}$ and $i \in \{1, \ldots, N\}$. When data sets contain non-linear relationships of the measurement procedures are simulated, the measurements of the clinical samples are generated by

$$\begin{aligned} \text{MP}_{Bir} &= \tau_{Ai} + \mathcal{N}(0, \sigma_A^2) \\ \text{MP}_{Air} &= f(\tau_{Bi}) = f_1(\tau_{Bi}) \cdot \mathbb{1}(\tau_{Bi} < 5.4) + f_2(\tau_{Bi}) \cdot \mathbb{1}(\tau_{Bi} \geq 5.4) + \mathcal{N}(0, \sigma_A^2). \end{aligned} \tag{5.2}$$

Where f_1 and f_2 are defined by

$$\begin{aligned} f_1(\tau_{Bi}) &= -0.1(\tau_{Bi} - 5)^2 - 0.1\tau_{Bi} + 5.5 \\ f_2(\tau_{Bi}) &= 2 + \exp(0.2\tau_{Bi}). \end{aligned} \tag{5.3}$$

Let us illustrate the patterns given above:

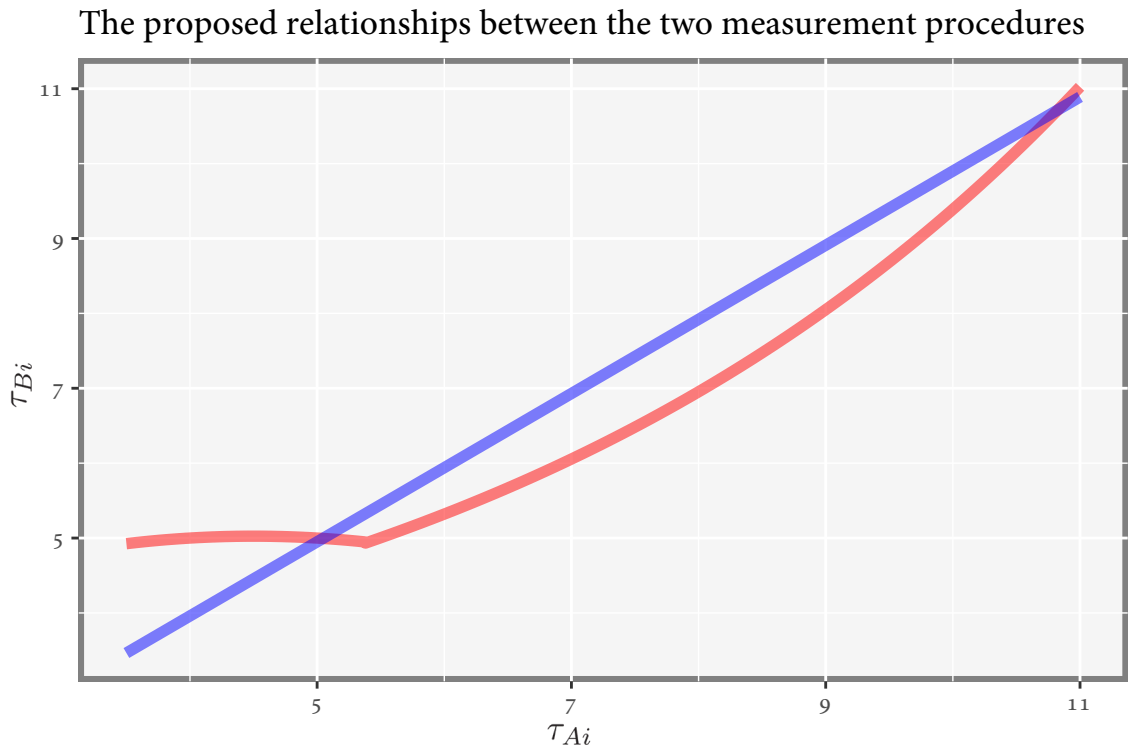

Figure 5.11 *– The relationships between measurement procedures, from which the data sets were simulated. The blue line is the linear pattern, and the red line is the non-linear pattern.*

In these simulation studies, the clinical samples are generated from the same sample range as the *LPK* data set exhibited in Section 3.1. Let $\tau_{Bi} \in [3.5, 11]$, which is the reference range of the number of billions of leukocytes per liter in human adults, for $i \in \{1, 2, \ldots n\}$.

5.5.1 Smoothing splines - simulation

To evaluate the impact of weak non-linear data patterns, we will consider how competently the smoothing splines estimator handles the before-mentioned non-linear patterns. Particularly, qualitative and visual tests will suffice, and, therefore, four separate data sets are considered. Two of them are constructed by clinical samples generated by Equation (5.2), that follow a non-linear pattern, whereas the last two follow the linear relationship defined in Equation (5.1). Bland-Altman transformation is applied in two of the cases observed in Figure 5.12.

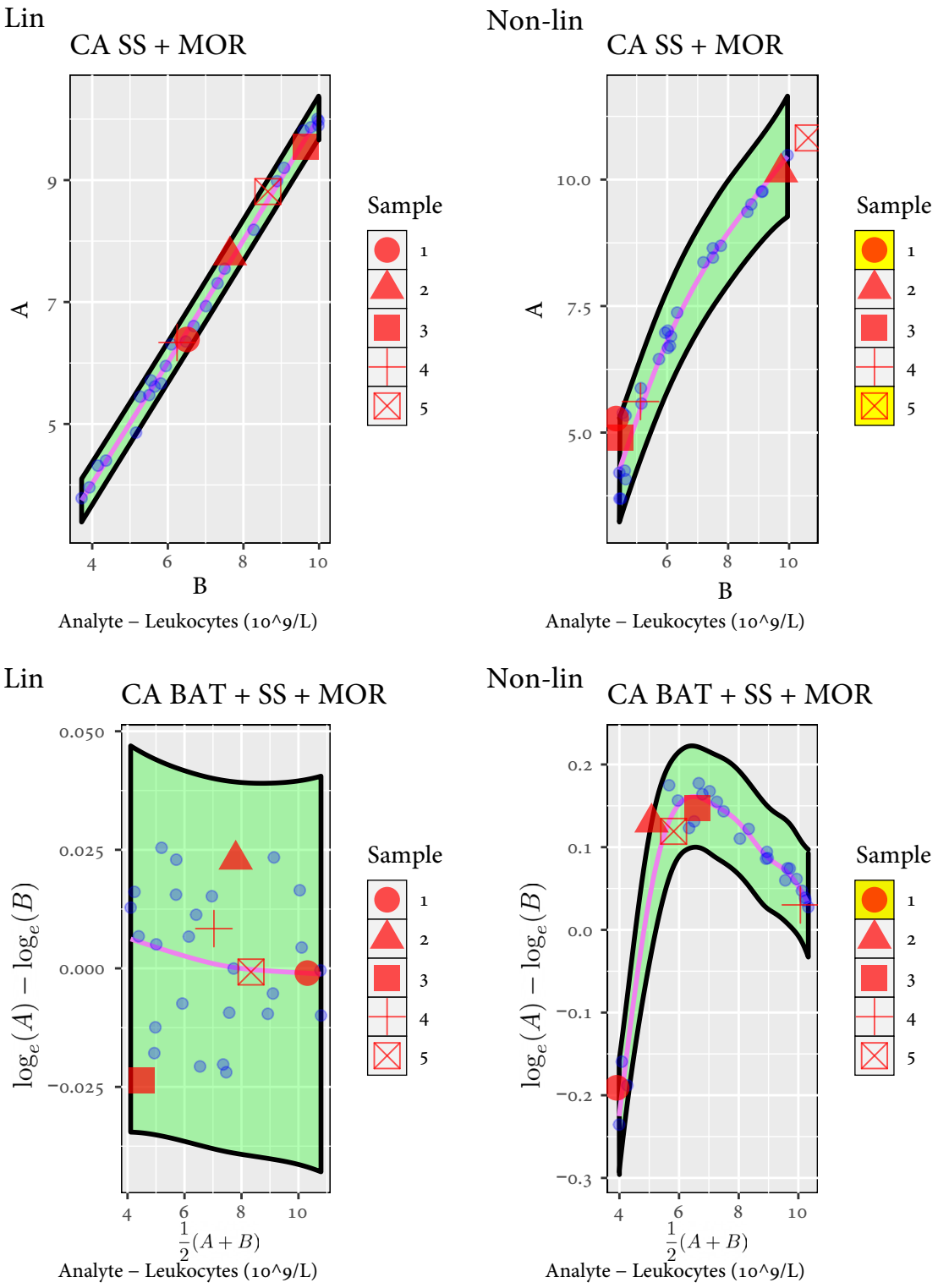

Figure 5.12 *– Simulated data sets with both linear and non-linear relationships. Two of the four models are constructed by Bland-Altman transformed clinical samples, whereas raw clinical samples construct the remaining two models.*

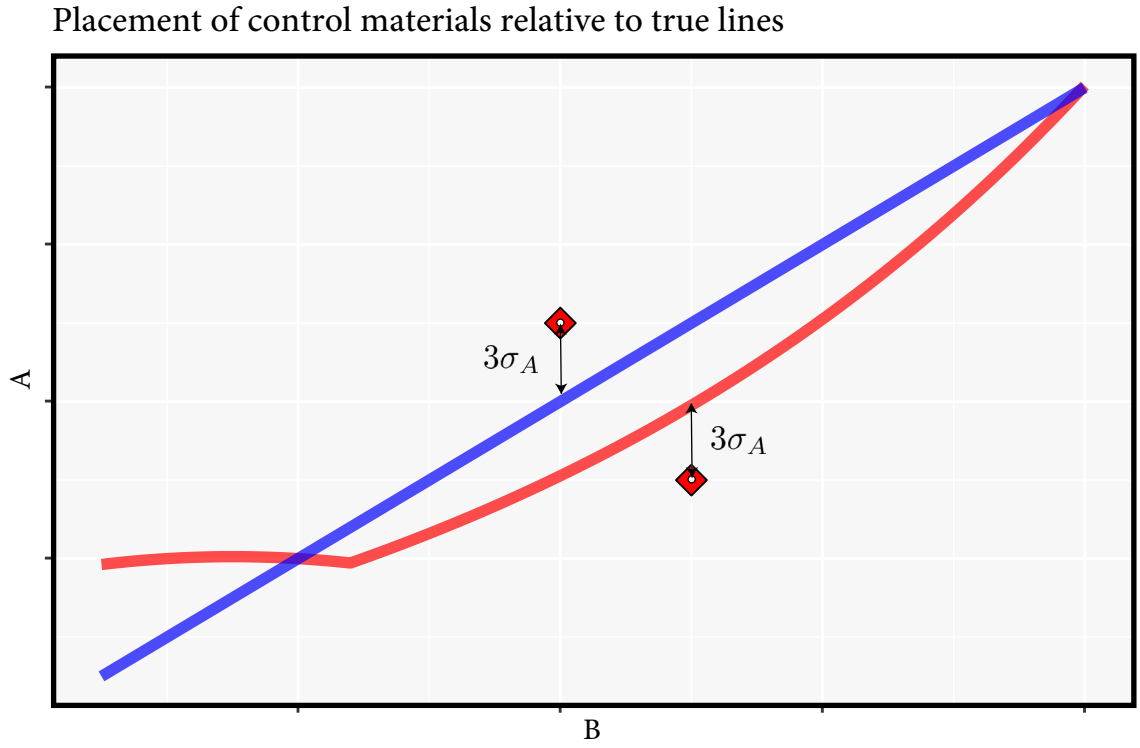

Figure 5.13 *– The placement of the defined non-commutable control materials. Concretely, the control materials are located three standard deviations from the two true relationships.*

There are no apparent barriers introducing non-linearity in the commutability assessment when we use smoothing splines estimators. The prediction bands for the non-linear case were somewhat more expansive than the linear case for the raw data. We are used to seeing that the linear model results in smaller uncertainties, so this should not be surprising. The Bland-Altman plots are, however, more comfortable to interpret visually and are consequently favorable in the cases of visual classification of control materials. This simulation's main interest is to review the commutability acceptance rate when possessing a non-commutable control material. Ideally, acceptance rates should be zero in these cases, but of course, for statistical reasons, this is not realistic. We define control materials, which are positioned $3\sigma_A$ away from the "true" lines, and denote these as non-commutable. Consequently, we expect our commutability assessment acceptance criterion to propose non-commutability for these control materials. To investigate this further, we will consider $n \in \{10, 11, \ldots, 50\}$, and it is of interest to observe how the sample size adjustments affect commutability acceptance rates for both linear and non-linear case. We repeat this simulation part for $R \in \{3, \ldots, 6\}$ to see if the number of replicates affects the commutability acceptance rates. Figure 5.13 demonstrates the control materials' location relative to the theoretical lines. With this in

mind, we simulate $K = 1000$ data sets for every pair of

$$\{n, R\} \in \{10, \ldots, 50\} \times \{3, \ldots, 6\},$$

where the estimated rate of commutability acceptance of the defined non-commutable control material is calculated. We will start with the non-linear case.

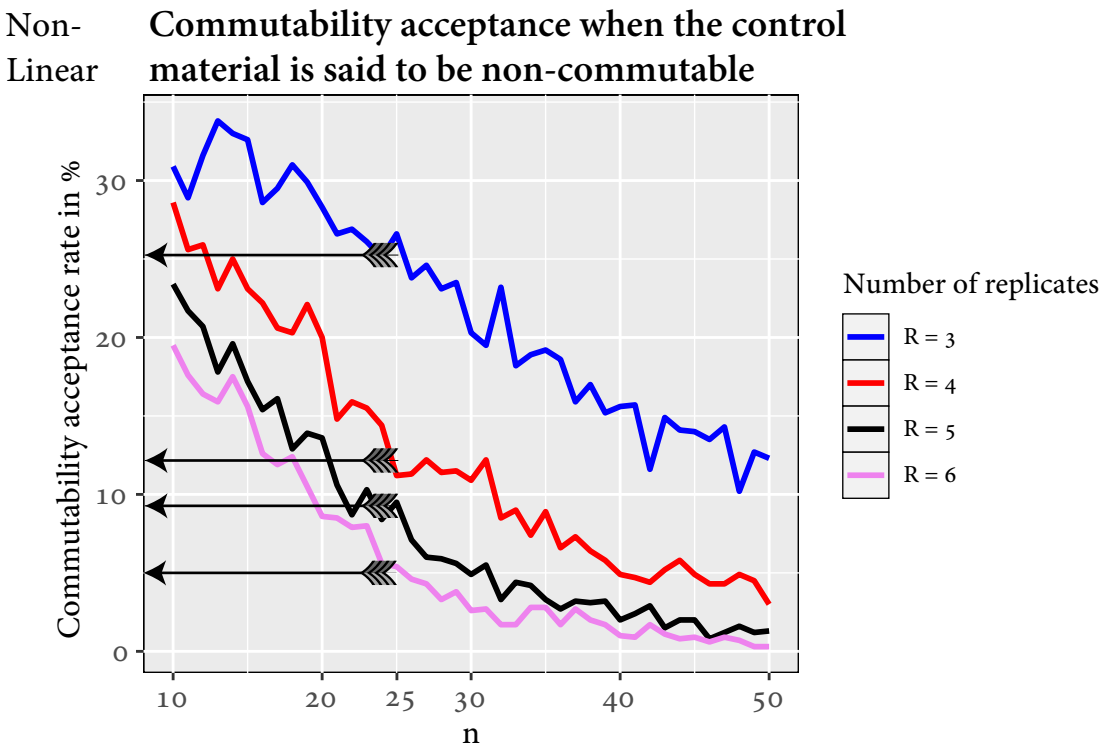

Figure 5.14 *– Acceptance rates of commutability where the clinical samples are fitted by the smoothing splines estimators and where the control material is non-commutable as a deviation of $3\sigma_A$ from the true non-linear curve.*

From Figure 5.14, we learn that the number of replicates selected is of importance regarding unsuitably accepting the control material as commutable when in reality is not. If we consider $n = 25$, as we do in Figure 5.14, we see that rate of erroneous conclusions of the control material falls from approximately 25% to 12% by just appending one replicated measurement. Generally, the acceptance rate seems to drop when the number of replicates increases. The last two acceptance-rate-dips are not as significant as the first dip, but certainly larger than zero. Besides, we observe that increasing n likewise will lead us to conclude correctly more frequently. For example, increasing the number of clinical samples from 25 to 30 reduces the absolute rate of commutability acceptance (in $\%$) of wrong conclusions by approximately 2-5%. To minimize erroneous acceptance of the non-commutable control

material, $\{n, R : n \geq 50, R = 6\}$ may be used. Alternatively, if we tolerate that approximately 5% of the instances result in misclassification, the suitable choice of study design is given by $\{n, R : n \geq 25, R = 6\}$. For the linear case, the situation is slightly changed. The prediction bands are generally narrower for linear models, yielding lower misclassification rates even for small n and R. To minimize the misclassification rates, an appropriate study design is $\{n, R : n \geq 25, R \geq 4\}$. See Figure 5.15 for portrayed simulation results for the linear case.

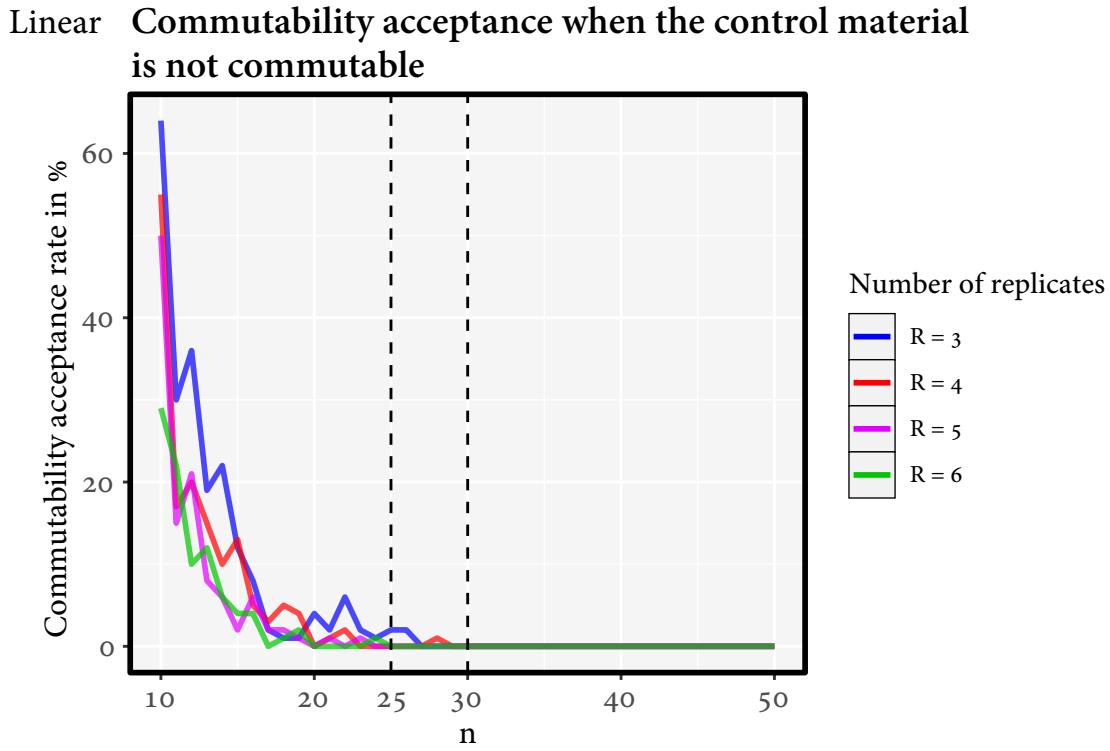

Figure 5.15 *– Acceptance rates of commutability where the clinical samples are fitted by the smoothing splines estimator and where the control material is non-commutable as a deviation of $3\sigma_A$ from the true linear curve.*

5.5.2 Kernel regression - simulation

As we did in Section 5.5.1, We will start to analyze two separate pairs of cases; One where the underlying data pattern is linear and one where it is not. Also, we consider the use of the Gaussian kernel and one where we use the epanechnikov kernel. We will include the prediction bands with(out) bias correction for differences according to model fits when increasing the bandwidth.

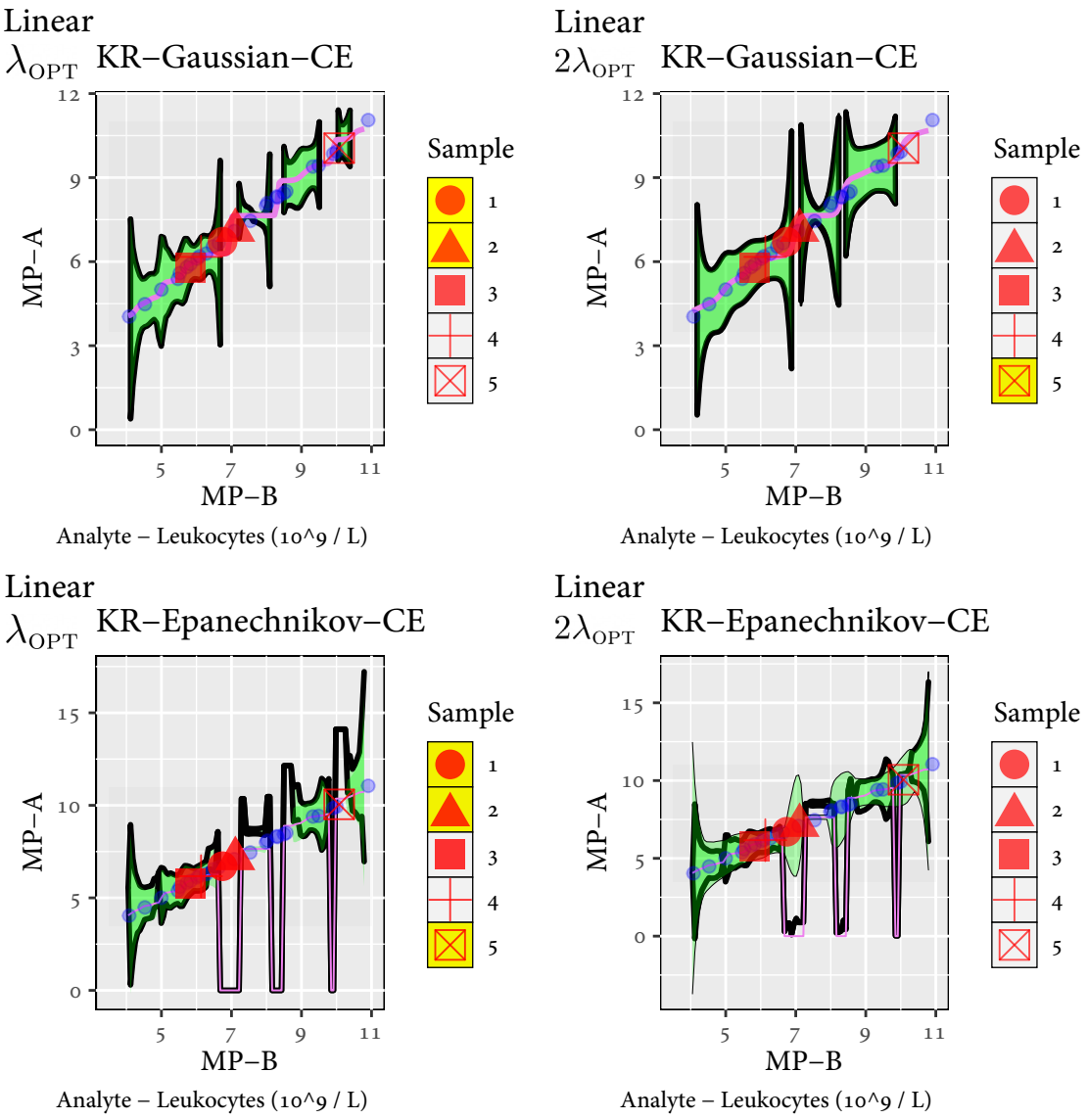

Figure 5.16 *– Gaussian and epanechnikov kernels using optimized bandwidth and two times the optimized bandwidth when the underlying data pattern is linear.*

From Figure 5.16, we recognize that larger bandwidths result in moderately smoother fits. The leftmost plots are the fitted regression lines where we applied the optimized bandwidth. The optimization procedure is done by cross-validation and is delivered by *npregbw()* in R. The rightmost plots are fitted *npreg()* models utilizing twice as large bandwidths as the corresponding optimized bandwidth. Note that breaks in the data patterns imply that we discontinuous prediction bands because of enormous variability in these regions. Note also that there are tremendous differences between the estimated prediction bands with bias correction and those not bias-corrected for the lower right plot.

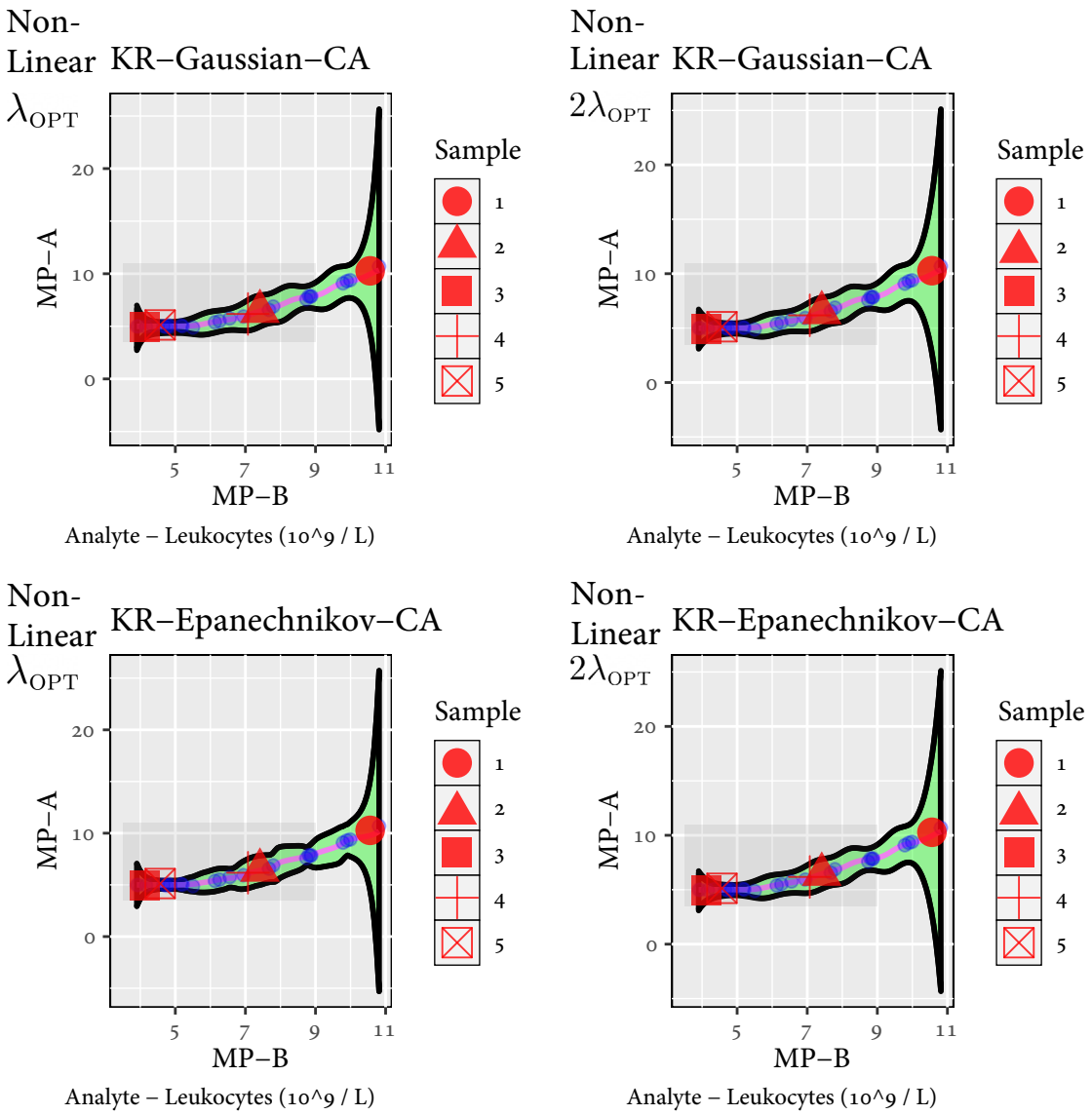

Figure 5.17 *– Gaussian and epanechnikov kernels using optimized bandwidth and two times the optimized bandwidth when the underlying data pattern is non-linear.*

As anticipated, non-linearity is not a concern for kernel regression. The non-linear patterns produce smoother fits than the linear ones. This result might, however, be a coincidence. The results in both Figure 5.16 and Figure 5.17 are equivalent from a commutability assessment viewpoint. In most cases, the discontinuity of the prediction bands mends by increasing the bandwidth. How much do we need to scale the optimized bandwidth to get something usable concerning commutability assessment is a natural question that arises. Larger bandwidths will enlarge the model bias, which will, consequently, overestimate the prediction bands when neglecting the bias term in Equation (4.24). Nonetheless, it seems like using two times the

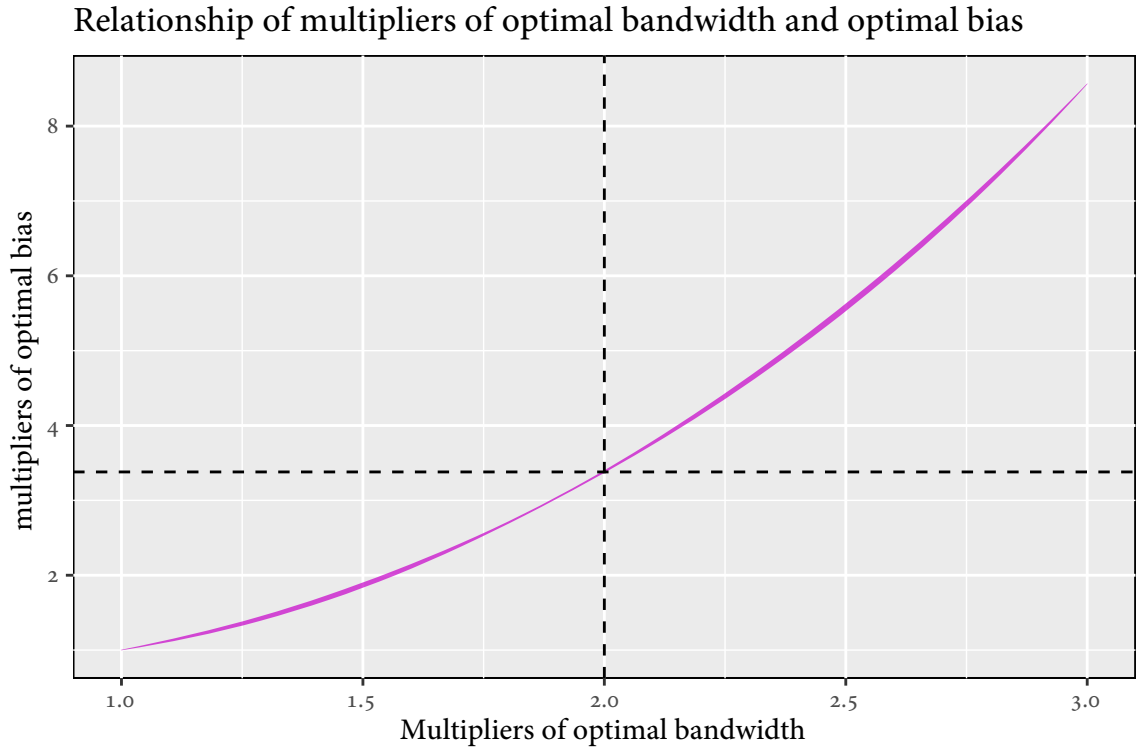

Figure 5.18 *– The relationship between multipliers of an optimal bandwidth and the corresponding optimal model bias measured by MSE. Note that the relationship is approximately quadratic.*

optimized bandwidth is sufficient in light of our simulations. A method for measuring the relative bias concerning different bandwidths will be important, and the mean square error (MSE) is a straightforward way to measure model bias. The bias relative to the optimal bias is defined by

$$\text{Relative bias} = \frac{\text{Enlarged bias}}{\text{Optimized bias}} \cdot 100\% = \frac{\text{MSE}(a \cdot \lambda_{\text{opt}})}{\text{MSE}(\lambda_{\text{opt}})} \cdot 100\%, \tag{5.4}$$

where a is the specified scalar multiplied by the optimal bandwidth, which is assumed to be defined for

$$a \in \{1, 1.01, \ldots, 3\}. \tag{5.5}$$

This simulation study will include simulations of 100 data sets for every value of a. The mean relative bias compared to the optimized bias for different multipliers of the optimized bandwidth has the relationship proposed in Figure 5.18. According to Figure 5.18, using $2 \cdot \lambda_{\text{OPT}}$ will end in model bias of approximately $3.4 \cdot \text{B}_{\text{OPT}}[\hat{g}(\text{MP}_i)]$. In other words, the bias is 3.4 times larger than the optimal bias. Consequently, caution must be practiced when using manually chosen bandwidths, which is bigger than the proposed optimal. However,

for small optimal biases, it seems appropriate to two times the optimal bandwidth. Consider a non-commutable control material as specified in Section 5.5.1. The control material is positioned $3\sigma_A$ from the true lines, as in Figure 5.13. Assuming that the underlying data pattern is non-linear and using both Gaussian and epanechnikov kernels, the acceptance rates of commutability are given as in Figure 5.19. As Figure 5.19 implies for both kernels, the estimated prediction bands are wider than $3\sigma_A$ even for large n and R, in contrast to smoothing splines. The next step is to investigate the commutability acceptance rates of the non-commutable control material when the underlying relationship is linear. The results are displayed in Figure 5.20. As witnessed in Figure 5.15, the analogous prediction bands were narrower when the true relationship was linear. Consequently, it will be more secure to use kernel regression if the true underlying relationship between the measurement procedures is linear. The drawback is that large sample sizes are required to conclude correctly, and the number of replicated measurements has no obvious effect. Besides, there are no notable differences between the Gaussian and epanechnikov kernels regarding performance, making it hard to determine if one is more preferable than the other. All in all, the performance of kernel regression is worse than smoothing splines regression. A large number of clinical samples are needed, and they are required to be evenly distributed to ensure high data density. Another approach within the Kernel regression field is local linear estimators instead of local constant estimators, which we use. This one is more robust against low data density, which is good.

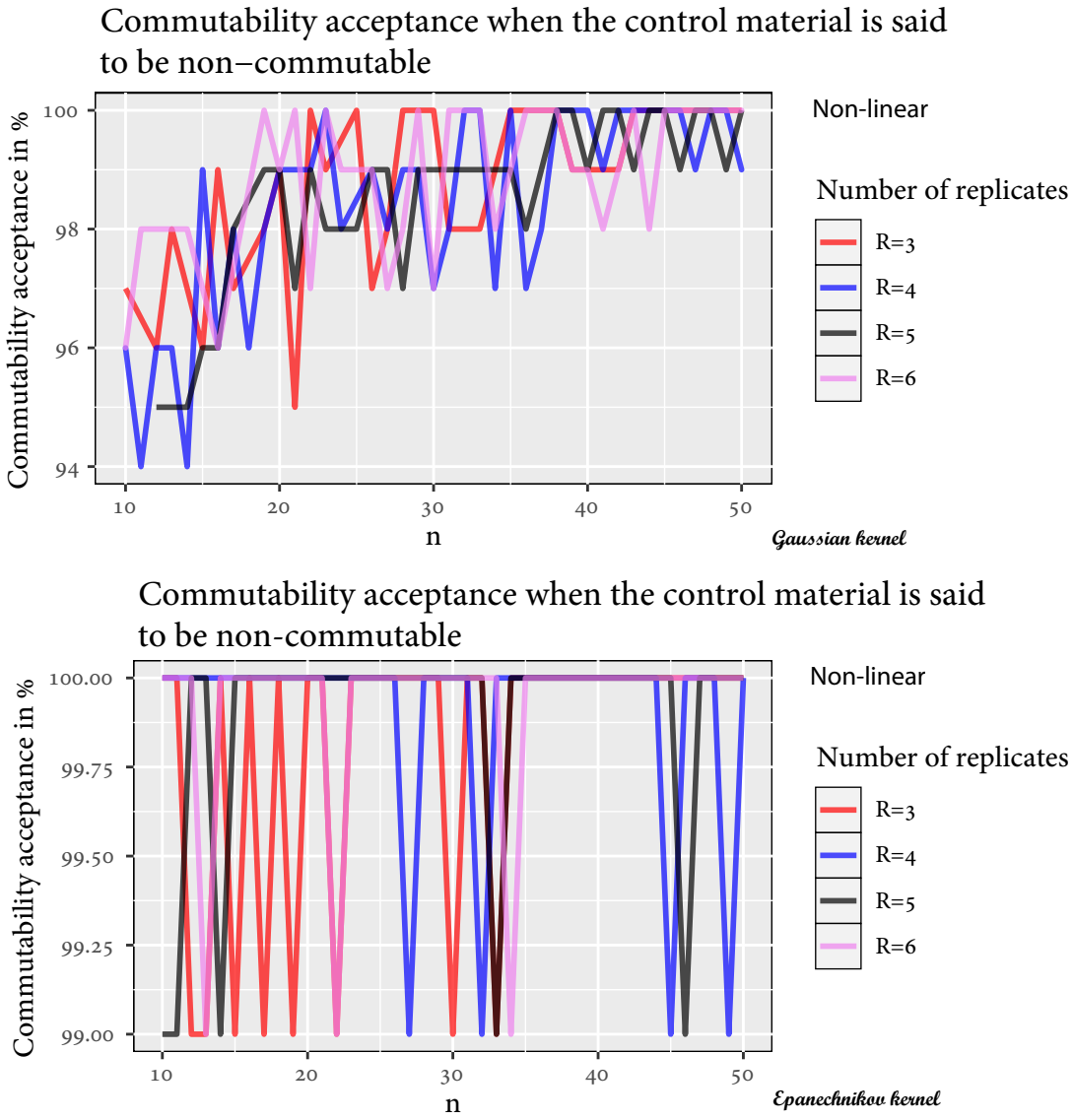

Figure 5.19 *– Acceptance of commutability where the clinical samples are fitted by kernel regression estimators and where the control material is non-commutable as a deviation of $3\sigma_A$ from the true non-linear curve.*

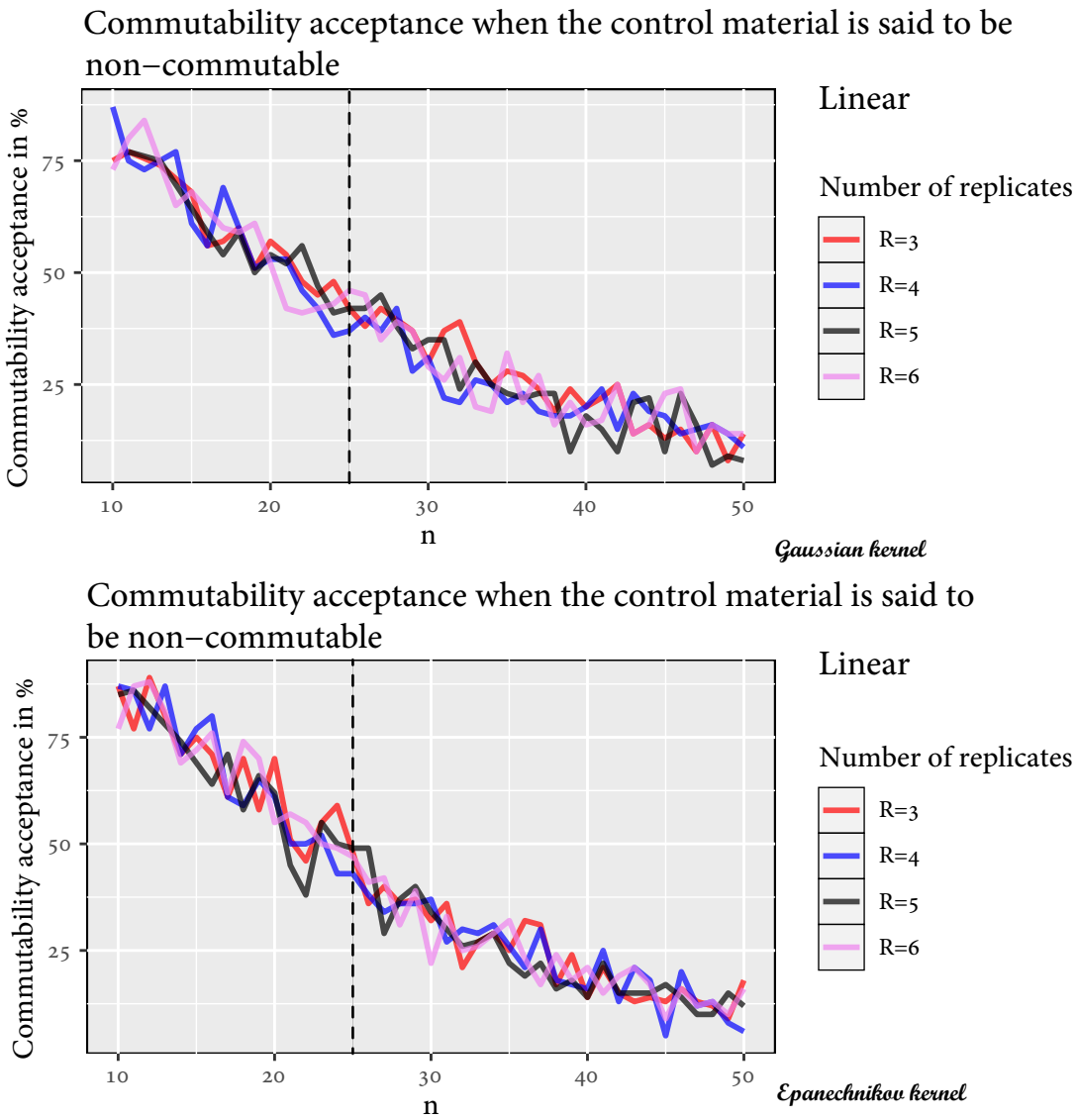

Figure 5.20 *– Acceptance of commutability where the clinical samples are fitted by kernel regression estimators and where the control material is non-commutable as a deviation of* $3\sigma_A$ *from the true linear line.*

Chapter 6

Recommendations and discussion

We have studied several statistical methods for the evaluation of commutability for control materials. We commenced with parametric assessment methods and moved on to non-parametric assessment methods. The objective is to find commutability assessment procedures with few rigorous model assumptions and robust fitted models from a statistical perspective. As only one of our methods accounts for the whole variability in x (Deming regression), it appears sensible to implement the Deming regression model as a general estimator. However, the Deming estimators' estimation when the ratio of variances is estimated requires at least 50 clinical samples when performing three replicated measurements on every one of them, which is unsatisfactory for our data sets. As we learned in Section 3.6, five or six replicates with 25 clinical samples could instead be used to significantly reduce the bias. Employing Equation (2.20) when possessing fewer than 25 clinical samples and three replicates for each clinical sample may not get us trustworthy Deming estimators. In addition to slightly unreliable Deming estimators when the study design is not suitable, we also require the linear model assumptions' fulfillment. If the linear model assumptions are not met, we may not trust the prediction bands' width. To mend flawed linear model assumptions, log-log transformation is stated as a common quick-fix. However, as observed in Section 3.6, the log-log transformation will often break the linear model assumptions instead of fixing them; if the underlying distribution is not log-normal. The next best thing is to use Bland-Altman transformation and use either smoothing splines or kernel regression to model the Bland-Altman transformed clinical samples. These combinations of transformation and regression models are robust against non-linearity and do not assume the underlying relationship's shape. Bland-Altman, combined with kernel regression, unfortunately, yielded extensive variability at the tails and dramatic unstable predictions at locations with few data points. Besides, we had to choose the bandwidths manually. The Smoothing splines estimator with cubic natural splines proved to have controlled variability at the tails and generally fewer jumping variance

occasions than kernel regression. Besides, we were not required to place the knots manually, so there is one less decision to make. Every measurement is used as a knot location. We could do this because the effective degrees of freedom were generally small compared to n, thus not considered as overfitting. With Bland-Altman, transformation combined with smoothing splines did not wholly deal with variability in x but reduced its magnitude by approximately 30%. When using these measurement procedures many times, we could go forth with a Bayesian approach. We will, however, not discuss this approach in any detail in this text. For Kernel Regression, modern research has managed to account for variability in x. Nevertheless, it is challenging to estimate the prediction errors for kernel regression with two-dimensional variability. Therefore procedures regarding estimation of prediction bands do not currently exist. Besides, we would be forced to choose appropriate bandwidths manually, which we typically do not want to do.

6.1 Recommendations

There are many alternatives regarding recommended methods to supplement the acceptance criterion for commutability. However, from a statistical perspective, we will focus on two decision algorithms for EQA-organisations. We denote $\{n, R : n \geq 50, R = 3\}$ or $\{n, R : n \geq 25, R \geq 5\}$ as *adequate study designs*. See the flowchart in Figure 6.1 for the first suggested decision algorithm.

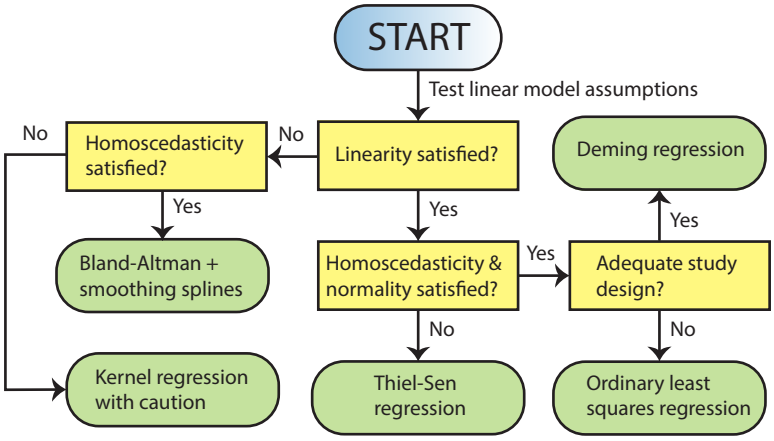

Figure 6.1 *– The first decision Algorithm, where the "optimal" statistical method is chosen from the data set properties. It is recommended to do ensure that auto-correlation is satisfied before employing the suggested evaluation method.*

In the start, an evaluation of the linear model assumptions is of order. The four linear model assumptions can for instance be evaluated simultaneously by the *gvlma()* function in R. The results from these tests are used throughout the decision algorithm implemented. All commutability assessment approaches will require independent residuals, which implies that it is important to ensure that auto-correlation is not present from the start. If *gvlma()* output proposes sufficient linearity in the data, parametric methods are the most reliable, and if the data set satisfies homoscedasticity and normality with either $\{n, R : n \geq 50, R = 3\}$ or $\{n, R : n \geq 25, R \geq 5\}$, Deming regression may be used. In opposition, if the data set does not satisfy one of the two appropriate study designs, ordinary least squares regression is used. Having linear data, but normality or/and homoscedasticity not satisfied, Section 4.1 suggests using Thiel-sen regression because of its robustness.

However, if linearity is not satisfied with the particular data set, non-parametric regression methods are recommended. Depending upon approximately equal variances for observations, either a smoothing splines estimator or a kernel regression estimator is applied. Note that extra precaution is necessary if kernel regression proves as the only valid approach. As described in Section 4.3, the kernel approach showed sensitivity when possessing too few observations or holes in the data set. More considerable non-linearity curves enlarge the prediction bands for kernel estimators significantly, as observed in Section 5.5.2. Consequently, the control materials were prone to be misclassified as commutable if the latter was true. However, if the data patterns are non-linear but not drastically different from a linear line, one could use Kernel regression if possessing at least 75 clinical samples. Doing so would minimize non-commutable control materials' misclassification rates deviating three standard deviations from the clinical samples' underlying curve.

As one might notice, the transformation approaches are little mentioned in Figure 6.1. The reason for this is elaborated in Section 3.6, where we observed that acceptance of linear model assumptions was prone to decrease if a transformation was applied to the raw data. For instance, the log-log transformed clinical samples typically proposed that only 50% of the linear model assumptions were met. The main difficulty is heteroscedasticity, which arises when the log-log transformation is used on non-log-normal distributed raw data. Generally, caution must be practiced concerning transformations if the raw data do not follow a right-skewed distribution.

An alternative and a more simple step-wise algorithm are Bland-Altman transformations combined with smoothing splines that are primarily used. See Figure 6.2 for details. This second algorithm is constructed to be more general at the cost of potential flexibility. If auto-correlation is not present and our data suggest homoscedasticity, Bland-Altman transformation combined with smoothing splines are used every time. Moreover, if possessing a satisfactory

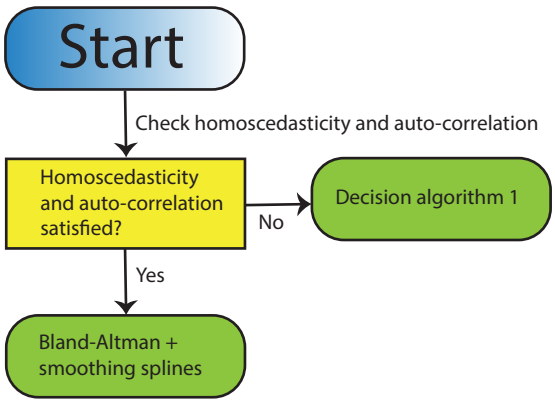

Figure 6.2 *– The second decision algorithm. This decision algorithm is less complicated than the first proposed if homoscedasticity and auto-correlation requirements are satisfied. If one of these two model requirements are unfulfilled, one uses the first decision algorithm instead.*

study design $\{n, R : 25 \leq n \leq 30, R = 6\}$ and these two linear model assumptions are accepted with $\alpha = 0.025$ significance level, smoothing splines on the Bland-Altman transformed clinical samples are used. In cases where some of the before-mentioned assumptions are not met, the other decision algorithm is rather used. For testing homoscedasticity and auto-correlation, the formal tests of *Breusch-Pagan* and *Durbin-Watson* are used at $\alpha = 0.025$. The details on these two tests are described roughly in Section 1.4.

For future studies, it will be interesting to dive deeper into non-parametric assessment methods for commutability. As touched upon in Section 4.4, methods for obtaining kernel regression estimators with errors-in-variables exist, and thereupon an algorithm for automatically selecting optimal practical bandwidths in this setting is bound to be imperative to construct. Furthermore, because tolerance intervals are more straightforward to implement than prediction bands, we could analyze their practical use in commutability assessment procedures. Also, it is not surprising that measurement procedures are used several times. It, therefore, appears wise to utilize this fact to our advantage. Methods within the Bayesian statistics field rely on prior information and may consequently provide more trustworthy models in the classification of control materials. It is well known and intuitive that models constructed with much information are more trustworthy than models with less. Regarding the commutability acceptance criterion, one could argue on how the limits of acceptance should be established. We have not addressed much about variability in the control materials, and the commutability acceptance criterion may be altered by implementing the dispersion

of the control materials as part of the assessment scheme. This was lightly discussed in Section 3.7, and may be elaborated upon in the future.

References

Bland, J. M. & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, *47*, 931–936.

Braga, F. & Panteghini, M. (2019). Commutability of reference and control materials: an essential factor for assuring the quality of measurements in laboratory medicine. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *57*(7), 967–973.

Craven, P. & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*(4), 377–403.

Dabo-Niang, S. & Thiam, B. (2019). Kernel regression estimation with errors-in-variables for random fields. *Afrika Matematika*, *31*.

De Brabanter, K., De Brabanter, J., Suykens, J. A. K., & De Moor, B. (2011). Approximate confidence and prediction intervals for least squares support vector regression. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *22*(1), 110–120.

Dhanoa, M. S., Sanderson, R., Lopez, S., Dijkstra, J., Kebreab, E., & France, J. (2011). *Modelling nutrient digestion and utilisation in farm animals*. Academic Publishers.

Dunn, G. (1989). *Statistical evaluation of measurement errors*. John Wiley & Sons.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Gillard, J. (2010). An overview of linear structural models in errors in variables regression. *REVSTAT–Statistical Journal*, *8*(1), 57–80.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Kernel smoothing methods*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York.

Miller, G. W., Jones, G. R., Horowitz, G. L., & Weykamp, C. (2011). Proficiency testing/external quality assessment: current challenges and future directions. *Clinical Chemistry*, *57*(12), 1670–1680.

Miller, W. G., Schimmel, H., Rej, R., Greenberg, N., Ceriotti, F., Burns, C., … Nilsson, G. (2018). IFCC working group recommendations for assessing commutability part 1: general experimental design. *Clinical chemistry*, *64*(3), 447–454.

Muller, H.-G. (1998). Local polynomial modeling and its applications. *Journal of the American Statistical Association*, *93*(442), 835–836.

Opsomer, J. D. & Breidt, F. J. (2011). *International encyclopedia of statistical science*. Springer Berlin Heidelberg.

Ould-Saïd, E. & Lemdani, M. (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. *Annals of the Institute of Statistical Mathematics*, *58*(2), 357–378.

Racine, J. S. (2007). Nonparametric econometrics: a primer. *Foundations and Trends® in Econometrics*, *3*(1), 1–88.

Solberg, H. (1993). A guide to IFCC recommendations on reference values. *Journal of the International Federation of Clinical Chemistry*, *5*(4), 162–165.

Vore, K. D. (2014). Evaluation of commutability of processed samples; approved guideline—third edition. *Clinical and laboratory Standards Institute*, 58.

Wilcox, R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal*, *40*(3), 261–268.

Appendix A

R-functions

Simulation of data:

```
simulate_data<-function(samp,repl=3,abc=c(0,1,0),ran, sdx=0.2,sdy
    =0.1)
{
  sams <- runif(samp,ran[1],ran[2]) ## True values from uniform RV
  tru <- rep(sams,each=repl) ## Each sample has the same true value
  samn <- rep(1:length(sams),each=repl) ## Appropriate structure
  repn <- rep(1:repl,times=samp) ## Appropriate structure
  data.frame(cbind(samn,repn,tru)) %>% rowwise() %>%
    mutate(B=tru + rnorm(1,sd = sdx)) %>%
    mutate(A= abc[1]*tru^2 + abc[2]*tru + abc[3] + rnorm(1,sd=sdy))
    %>%
    mutate_at(.vars = "samn", .funs = factor) %>% ## Change to factor
    .
    dplyr::select(-"tru",Sample = "samn",Replicate = "repn")
}
```

Automatic generation of 100 new data valaues from the range of MP_B:

```
get_newdata <- function(x)
{
  ran <- range(x)
  return(seq(from = ran[1], to = ran[2], by = abs(ran[2]-ran[1]) *
    0.01))
}
```

Bland-Altman transformation of simulated data set:

```
Bland_Altman_Transform <- function(df)
{df %>% mutate(A=log(A/B),B=(A+B)/2)}
```

Converting from AR to MOR for simulated data sets:

```
AR_to_MOR <- function(df){
df %>% group_by(Sample) %>%
summarise(A=mean(A),B=mean(B),.groups = 'drop')}
```

Automatic checks of linear model assumptions:

```
LMA_tests <- function(model)
{
  a<-shapiro.test(resid(model))$p.value>=0.05
  b<-bptest(model)$p.value>=0.05
  c<-dwtest(model)$p.value>=0.05
  unname(ifelse(a+b+c==3,TRUE,FALSE))
}
```

Automatic commutability assessment checks:

```
CA.auto <- function(clin,contr,model)
{
  ran<-range(clin$B)
  pred<-data.table(A=contr$A,predict(object=model,
  newdata=list(B=contr$B),interval="prediction",level=0.99)) %>%
    rowwise() %>% mutate(A = lwr < A && upr > A) %>%
    mutate(B = contr$B >= ran[1] && contr$B <= ran[2])%>%
    mutate(AB = A+B == 2) %>%
    dplyr::select(c("CA"="AB"))
}
```

Calculation of dispersion measures such as SD and CV

```
calculate_dispersion_measures <- function()
{
  data <- simulate_data(samp=25,repl=3,abc=c(0,1,0),
  ran=c(3.5,11))
  BA <- Bland_Altman_Transform(data)
  raw <- AR_to_MOR(data); trans <- AR_to_MOR(BA)
  SD_raw <- sd(raw$B); SD_trans <- sd(trans$B)
  CV_raw <- SD_raw/mean(raw$B)
  CV_trans <- sqrt(exp(SD_trans ^ 2) - 1)
  return(c(CV_raw,CV_trans,SD_raw,SD_trans))
}
```

Estimate lambda from replicates:

```
estimate_lambda <- function(y,x, R, n)
{
  N<-R*n
  df <- data.table(sample = rep(1:R,each=n), replicate = rep(1:n,
   times=R), A = y, B = x) %>%
```

```
  group_by(sample) %>%
  mutate(mA = mean(A), mB = mean(B)) %>% rowwise() %>%
  mutate(YmM = (A - mA)^2, XmM = (B - mB)^2)
  sigma.ee <- (sum(df$YmM)) / (N - n)
  sigma.uu <- (sum(df$XmM)) / (N - n)
  return(lambda = sigma.ee / sigma.uu)
}
```

```
#### Deming estimation -- Requires lambda to be estimated before
    using this ####
#### Inspired by https://gist.github.com/stla/5
    fcd959576413798d4cc09e7493e53e9 ##
deming_estimate <- function(y,x,lambda)
{
  n <- length(x) ## Number of clinical samples
  my <- mean(y) ## Mean of MP_A
  mx <- mean(x) ## Mean of MP_B
  SSDy <- crossprod(y-my)[,] ## S_AA
  SSDx <- crossprod(x-mx)[,] ## S_BB
  SPDxy <- crossprod(x-mx,y-my)[,] ## S_BA
  A <- sqrt((SSDy - lambda*SSDx)^2 + 4*lambda*SPDxy^2)
  B <- SSDy - lambda*SSDx
  beta <- (B + A) / (2*SPDxy) ## Slope
  alpha <- my - mx*beta ## Intercept
  sigma.uu <- ( (SSDy + lambda*SSDx) - A ) /(2*lambda) / (n-1)
  s.vv <- crossprod(y-my-beta*(x-mx))/(n-2)
  ## formula from Gilard & Iles paper ##
  sbeta2.Fuller <- (SSDx*SSDy-SPDxy^2)/n/(SPDxy^2/beta^2)
  sbeta.Fuller <- sqrt(sbeta2.Fuller)
  ## standard error alpha Fuller ##
  salpha2.Fuller <- s.vv/n + mx^2*sbeta2.Fuller
  salpha.Fuller <- sqrt(salpha2.Fuller)
  ## Let us define the covariance matrix for the estimated
  ## coefficients, V.
  V <- rbind(c(salpha2.Fuller, -mx*sbeta2.Fuller),
  c(-mx*sbeta2.Fuller, sbeta2.Fuller) )
  return(list(alpha=alpha,beta=beta,
  salpha.Fuller=salpha.Fuller, sbeta.Fuller=sbeta.Fuller,
              V=V,
              sigma=sqrt(sigma.uu*(n-1)/(n-2)),lambda=lambda)
  )
}
```

Prediction of one point of Deming model

```
#### Predict y from theoretical x ####
#### Inspired by https://gist.github.com/stla/5
    fcd959576413798d4cc09e7493e53e9 ##
deming_predict <-  function(y, x, R=3, n=25, xnew, level=0.99){
  lambda <- estimate_lambda(y=y,x=x,R=R,n=n)
  fit <- deming_estimate(y=y,x=x,lambda=lambda)
  sigma <- fit$sigma
  sigma.uu <- sigma^2*(n-2)/(n-1)
  V <- fit$V ## Covariance matrix for estimated coefficients
  a <- fit$alpha ## Intercept
  b <- fit$beta ## Slope
  ynew <- a+b*xnew ## Predicted Y given X
  Xnew <- as.matrix(c(1,xnew)) ## One row of MP^*
  sigma.ee <- lambda*sigma.uu
  t <- qt(1-(1-level)/2, n-2) ## t-score
  ## predict from an observed xnew ##
  sd.ynew.Fuller2 <- sqrt(sigma.ee + t(Xnew)%*%V%*%Xnew + (b^2+V
    [2,2])*sigma.uu)
  Lynew.Fuller2 <- ynew - t*sd.ynew.Fuller2
  Uynew.Fuller2 <- ynew + t*sd.ynew.Fuller2
  return(data.table(new=xnew, fit=ynew,lwr=Lynew.Fuller2,upr=Uynew.
    Fuller2))
}
```

Estimated prediction bands for Deming estimator

```
#### Requires registration of cluster and package doParallel ####
deming_predictInterval <- function(y, x, R, n, newdata,level=0.99){
foreach(i=1:length(newdata), .combine = rbind, .export = c("estimate_
    lambda","deming_predict","deming_estimate"), .packages = c("dplyr"
    ,"data.table")) %dopar% deming_predict(y=y, x=x, R=R, n = n, xnew
    = newdata[i])
}
```

Simulation of data concerning the non-parametric methods:

```
#### Simulation of data sets for non-parametric simulation studies
    ####
simulate_data_np <- function(samp, repl=3,lin=T,ran,sdx=0.2,sdy=0.1)
{
  sams <- runif(samp,ran[1],ran[2]) ## Generating true values
  tru <- rep(sams,each=repl) ## Appropriate structure
  samn <- rep(1:length(sams),each=repl) ## Appropriate structure
  repn <- rep(1:repl,times=samp) ## Appropriate structure
  data.frame(cbind(samn,repn,tru)) %>% rowwise() %>%
    mutate(B=tru) %>%
```

```
    mutate(A=ifelse(lin==T,B,ifelse(B<5.4,-0.1*(B-5)^2-0.1*B+5.5,2 +
    exp(0.2*B)))) %>% ## Relationship between MP's
    mutate(B=B+rnorm(1,sd=sdx)) %>% mutate(A=A+rnorm(1,sd=sdy)) %>%
    mutate_at(.vars = "samn", .funs = factor) %>%
    dplyr::select(-"tru",Sample = "samn",replicate = "repn")
}
```

Bias correction magnitudes for prediction bands estimated by bootstrap:

```
#### Will be repeated for every value of B ####
npregboot_core <- function(data,cker,bw,newd)
{
  ## re-sampling rows from the data frame
  bdf <- slice_sample(.data=data,n=nrow(data),replace=TRUE)
  ## bootstrap model for re-sampled rows in data frame
  mb <- npreg(bws=bw,data=bdf,formula=A~B,ckertype=cker)
  ## fill list with predictions of the bootstrap data frame
  return(list(predict(object=mb,newdata=data.frame(B=newd))))
}

#### Bootstrap bias ####
npregboot <- function(B=999,bw,npreg_obj,data,level=0.99,ran=c
    (3.5,11),ckertype="gaussian")
{
  ## Obtaining the model frame ##
  x <- npreg_obj$eval[,1]
  y <- as.vector(residuals(npreg_obj) + fitted(npreg_obj))
  ## Generating new data from predictor values ##
  newx <- get_newdata(x)
  ## Empty list to be filled and original model ##
  bpred <- list()
  m <- npreg(bws=bw,data=data,formula=A~B,ckertype=ckertype)
  ## Filling the list with bootstrapped models ##
  bpred <- foreach(b=1:B,.combine=append,.export="npregboot_core",.
    packages=c("np","dplyr")) %dopar% npregboot_core(data=data,cker=
    ckertype,bw=bw,newd=newx)
  ## Converting from list to data frame ##
  bpred <- data.frame(bpred)
  ## The actual bias-calculation returned ##
  return(rowMeans(bpred) - predict(object = m, newdata=data.frame(B=
    newx)))
}
```

Estimated prediction bands using the asymptotic properties of kernel estimators:

```
get_api <- function(npreg_obj,kernel="gaussian",rbw=F,lvl=0.99)
```

```
{
  ## Obtaining the model frame from model ##
  x <- npreg_obj$eval[,1]
  y <- as.vector(residuals(npreg_obj) + fitted(npreg_obj))
  ## generating new data from predictor values ##
  new <- get_newdata(x)
  ## Predicted values and asymptotical prediction errors ##
  pre <- predict(object = npreg_obj, newdata = data.frame(B = new),
   se.fit = TRUE)
  ## The asymptotic prediction errors ###
  shat <- pre$se.fit
  ## The predicted values ##
  ghat <- pre$fit
  ## Checks which kernel applied (Gaussian or Epanechnikov) ##
  cond <- kernel=="gaussian"
  ## Optimal bandwidth, b, for density estimation given kernel ##
  mdenb <- ifelse(rbw==F,npudensbw(x,y,ckertype=ifelse(cond,kernel,"
   epanechnikov"))$bw,rbw)
  ## Estimated density with optimal bandwidth calculated above ##
  fhat <- density(x=x,kernel=ifelse(cond,kernel,"epanechnikov"),bw=
   mdenb,n=length(ghat))
  ## Calculation of integral dependent on kerlen chosen ##
  k2 <- ifelse(cond,1/(2*sqrt(2)),0.6)
  ## Asymptotic variance of kernel regression model ##
  vghat <- (shat * k2) / (length(x) * fhat$y * rbw)
  ## Quantile-calculation assuming Z ~ N(0,1) ##
  z <- qnorm(p=(1-lvl)/2,lower.tail = F) ## z-score
  ## The calculation of the prediction bands ##
  predf <- data.frame(new=new,fit=ghat) %>%
    mutate(lwr = fit - z * sqrt(shat + vghat)) %>%
    mutate(upr = fit + z * sqrt(shat + vghat)) %>%
    na.omit()
  ## For bias correction, subtract bootstrap bias from this result ##
  return(predf)
}
```

Calculation of bias of chosen bandwidth compared to the bias of the optimal bandwidth:

```
relative_bias <- function(scalar,data=AR_to_MOR(simulate_data_np(samp
   =25,ran=c(3.5,11),lin=F)))
{
  ## Obtaining the optimal bandwidth of simulated model ##
  obj_bw <- npregbw(data=data,formula=A~B)$bw
  ## Model for optimized bandwidth selected above ##
  obj<-npreg(bws=obj_bw,formula=A~B,data=data)
```

```
  ## MSE of optimized model ##
  optimized_MSE <- sum(residuals(obj) ^ 2) / nrow(data)
  ## Model for scaled optimized bandwidth ##
  obj_change <- npreg(bws=obj_bw*scalar,formula=A~B,data=data)
  ## MSE of scaled model ##
  MSE <- sum(residuals(obj_change) ^ 2) / nrow(data)
  ## The ratio of scaled MSE and optimized MSE in % ##
  return((MSE/optimized_MSE - 1)*100)
}

#### Estimation of relative bias for different scales ####
repeat_relative_bias <- function(scalars=seq(from=1,to=3,by=0.01))
{
  ## Simulated data set to be used ##
  df <- AR_to_MOR(simulate_data_np(samp=25,ran=c(3.5,11),lin=F))
  ## Relative bias in percent for all scales ##
  percent_more_than_opt <- foreach(i=1:length(scalars),.packages=c("
    dplyr","np"),.export=c("relative_bias"),.combine = c) %dopar%
    relative_bias(scalars[i],df)
  ## Return a list of data frames of the relative bias ##
  return(list(data.frame(percent_more_than_opt=percent_more_than_opt)
    ))
  ## In the text we repeat repeat_relative_bias(), 1000 times ##
  ## and takes the average over each scale ##
  ## This produces the average mean relative MSE for each scale
    specified ##
}
```