UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Statistical approaches to correct for baseline values in clinical trials

Matilde Martins da Palma Francisco

Mestrado em Bioestatística

Trabalho de projeto orientado por:
Professor Doutor Klaus Langohr
Professora Doutora Maria Helena Mouriño Silva Nunes

2023

Acknowledgements

At the end of another really important chapter of my academic chapter, there are some people that I would like to thank without whose this would have been far more complicated.

To my advisor Professor Klaus Langohr, not only for the knowledge transmitted to me during the elaboration of this master thesis but also for believing in me, and encouraging me to do more and better and to get out of my comfort zone. Your guidance through all this time meant a lot, not only at an academic level but also for indirectly teaching me to look into things from a different perspective. To Professor Jordi Cortés, Professor Daniel Fernandez, and Professor Xavier Piulachs for the support and contribution to the good environment. Finally, to Professor Guadalupe Gomez for welcoming in such a good way to the research group and actively participating in the activities, always being available, and being an example. Also for encouraging the long morning coffees. You all contributed a lot, not only to my professional but also to my personal growth.

To my Co-advisor Professor Maria Helena Nunes for always supporting my decision to do the master thesis abroad and facilitating the process. Also, for the constant support during this process. To all the professors of the Master in Biostatistics for all the knowledge transmitted and for helping to enter this area that in the beginning was unknown.
To Professor Pedro Oliveira, for being the first contact I had with Biostatistics and always encouraging me to follow this path. It was definitely a key in my academic and future professional journey.

To Andrea and really missed Ida, for the long talks about everything. I was so lucky to have you around during this time. Also, I am really happy for have taught you really valuable lessons such as the importance of hugs and phone calls.

Ao meu Porto. Sou tão agradecida por a vida vos ter posto no meu caminho. À Babá, de melhor girl boss a conselheira. À Bárbara,amiga de todas as horas, por conseguir trazer a cima o lado bom de tudo. À Catarina, que no terceiro dia de aulas do primeiro ano me deu casa e desde aí atura as minhas crises existenciais. Ao David, por todas a frases (des)motivadoras. À Maria (forever partner), pelas horas e horas a conversar sobre tudo. À Olívia (IDOM), por me ajudar a ver um lado meu que sempre desconheci. Obrigada por me fazerem acreditar que ia correr bem, mesmo quando tinha tudo para correr mal. Sei que

estarão sempre comigo seja onde quer que for.

À Margarida, a minha ride or die. Obrigada por estares sempre lá para festejar comigo os meus sucessos e apoiar nos momentos menos bons. Para toda a vida a uma chamada de distância.

À Daniela, sem a qual fazer este mestrado tinha sido sem dúvida muito mais doloroso. A todas as chamadas para resolver exercícios e fazer trabalhos de grupo com muito gossip pelo meio. Ainda bem que te encontrei.

Ao Flávio, pelo apoio incondicional e acreditar sempre em mim, e me encorajar a dar o meu melhor. Obrigada por teres esse poder de melhorar dias maus, e tornares ainda melhores os bons. A vida é mais bonita contigo ao lado.

Aos meus pais, sem os quais nada disto teria sido possível. Obrigada por me encorajarem e ajudarem sempre a seguir os meus sonhos. Pela confiança que sempre depositaram em mim me permitiu chegar onde quieria. À Madalena, ser o teu role model é a minha motivação. Obrigada por saberes exatamente quando tudo o que preciso é ser desconcentrada e quando preciso de uma motivaçãozinha extra. Mal posso esperar por continuar a viver a vida ao vosso lado.

Matilde Francisco, Barcelona, Setembro de 2023

Abstract

Longitudinal studies allow the repeated monitoring of health outcomes or risk factors, and the identification of differences in outcomes. Baseline measurements are demographic characteristics or measurements taken at the beginning of the study of the response variable or variables correlated with it. Whether to consider baseline as a covariate or a dependent variable is a frequently asked question. Not accounting for baseline, can not only affect the magnitude of differences detected in a study, but also the direction of these differences, which can result in different clinical conclusions. The lack of consistency in the literature around this topic contributes to the difficulty to establish a standard statistical approach, so studies' specific characteristics influence the decision on what statistical approach should be used.

When adjusting a model, it is possible to adopt different strategies regarding the use of baseline. It can be included in the model as a covariate, and the post-baseline values are the response variable or, assuming that the randomization of the subjects involved was efficient, baseline and post-randomization values can be treated as dependent variables.

In this work, two different methods, constrained longitudinal data analysis (cLDA) and analysis of covariance (ANCOVA) were applied to a real data set from a clinical trial and simulated data, in order to study the behaviour of the methods under different conditions and try to figure out what would be the best approach. The obtained results indicate that cLDA can be appropriate in the cases where data follows a normal distribution, and its application can bring advantages especially in the presence of missing data. However, when there is a deviation from normality, ANCOVA showed to be a better approach regardless of the other conditions.

Keywords: Baseline-value adjustment, linear mixed models, longitudinal studies, ANCOVA, cLDA.

Resumo alargado

Na era em que se vive hoje em dia, em que os avanços na medicina são realizados com base em evidência, os clínicos conduzem estudos de modo a avaliar o efeito de intervenções, tratamentos ou exposição de indivíduos de modo a apoiar decisões relativamente aos doentes. Estudos longitudinais são muitas vezes utilizados, pois permitem a monitorização repetida de resultados relativos à saúde de um paciente ou de fatores de risco, e a identificação de diferenças nestes. Neste contexto, características de linha-base de um paciente, características demográficas ou as variáveis medidas no início do estudo da variável de resposta, ou de variáveis correlacionadas com esta têm vários propósitos, incluindo a avaliação do efeito do tratamento com base na alteração a partir do valor de linha-base. A questão de considerar o valor inicial das variáveis resposta como covariável ou variável dependente é frequentemente levantada.

Não considerar características de linha-base como covariável pode não só afetar a magnitude das diferenças detetadas, como também a direção dessas diferenças, o que pode resultar em conclusões clínicas diferentes e incorretas. No entanto, corrigir os efeitos da linha-base também pode introduzir viés, o que é problemático, especialmente em casos onde o tamanho amostral é pequeno. A falta de consistência na literatura sobre este tópico contribui para a dificuldade em estabelecer uma abordagem estatística padrão. As características específicas dos estudos influenciam então a decisão sobre qual abordagem estatística deve ser usada.

Em ensaios clínicos, estudos realizados de modo a comparar em humanos a eficácia e segurança associadas à utilização de diferentes intervenções médicas, a aleatorização é um processo fundamental. Esta ajuda a prevenir o viés associado à seleção dos candidatos que recebem cada tratamento o garante que seja possível comparar o efeito dos diferentes tratamentos entre os grupos, uma vez que eles são semelhantes em quase todos os outros aspetos críticos.

Ao ajustar um modelo, é então necessário adotar estratégias diferentes relativamente ao uso dos valores de linha-base. O uso da Análise da Covariância (ANCOVA) tem sido recomendado quando há necessidade de corrigir os efeitos desta. Neste modelo, sob o pressuposto da existência de uma correlação entre as medidas de linha-base e pós-linha-base, os valores de linha-base são incluídos no modelo como covariável, e os valores pós-linha- base são a variável resposta. A Análise de Dados Longitudinais Condicional (cLDA) é construída sob o pressuposto de que a aleatorização dos sujeitos envolvidos foi eficiente, portanto, assume-se que as médias na linha-base são idênticas para os grupos em comparação. Devido a isso, tanto os valores de linha-base quanto os valores pós-aleatorização são a variável dependente. Ambos os modelos podem ser aplicados tanto para análises pré-pós, como para análises longitudinais. Enquanto na

primeira, a cada indivíduo foram medidos dois valores da variável de interesse, na segunda características de cada indivíduo são medidas em vários momentos, pelo menos 3.

Quando na variável resposta, há mais que uma medida de cada indivíduo, é necessário adotar uma abordagem onde se tem em consideração que estas estão correlacionadas. Deste modo, a utilização dos modelos de regressão linear, que podem ser utilizados para modelar a variável de interesse quando há uma resposta de cada indivíduo, não é adequada. É necessário aplicar modelos mistos lineares já que este tipo de modelos tem em consideração tanto a variabilidade intra como inter indivíduos. Isto é possível através da introdução no modelo de efeitos fixos e efeito aleatórios. Os modelos aplicados ao longo do trabalho são adaptações de modelos de regressão linear e modelos lineares mistos.

Neste trabalho, os modelos previamente referidos foram aplicados a um conjunto de variáveis selecionadas dentro dos dados recolhidos ensaio clínico *Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome* (TESDAD), com o fim de avaliar se abordagens diferentes relativamente à linha-base têm impacto nos resultados. Neste ensaio clínico, os pacientes foram divididos em dois grupos, onde, simultaneamente com treino cognitivo, um dos grupos recebeu o tratamento em estudo e o outro um placebo durante um período de doze meses. Testes para avaliação neuro fisiológica foram aplicados no início, 3, 6 e 12 meses após o início da administração do tratamento e 6 meses após a descontinuação deste. Nesses mesmos instantes algumas características fisiológicas foram também medidas. Foram realizadas duas análises com os dados, uma análise pré-pós com os resultados registados inicialmente e aos 18 meses e uma análise longitudinal, com os dados recolhidos no início e após 3, 6 e 12 meses. Foi realizada uma análise exploratória de modo a averiguar se seria necessário a introdução no modelo de alguma variável medida ao longo do estudo, no entanto, todas se encontravam equilibradas entre os dois grupos de tratamento. A única covariável inserida foi o sexo por recomendação da equipa clínica. Também nesta análise se calculou a percentagem de dados omissos. O alvo do estudo, parâmetro avaliado, foi a estimativa do efeito de tratamento.

Em seguida, para estudar a validade das conclusões retiradas previamente sob diferentes condições, realizou-se um estudo de simulação. Neste estudo também foram testados ambos os cenários de análise pré-pós e longitudinal. Foram simulados dados, com base em parâmetros calculados para uma das variáveis do estudo TESDAD. De modo a obter cenários com diferentes características, os parâmetros tamanho da amostra, percentagem de dados omissos e distribuição seguida pela linha-base foram alterados. Posteriormente os modelos cLDA e ANCOVA foram aplicados a estes, de modo a verificar qual o comportamento nos diferentes cenários. Para avaliar o comportamento do modelo, os parâmetros enviesamento, desvio padrão, quadrado médio do erro e percentagem de cobertura fornecida pelos intervalos de confiança a 95% foram calculados. Todas as análises foram realizadas com recurso ao software R.

Por fim, analisou-se tanto dos resultados obtidos pela análise dos dados do ensaio clínico como do estudo de simulação. No estudo realizado com dados reais, verificou-se que apesar de se verificar pequenas diferenças nas estimativas do efeito de tratamento, a conclusão retirada sobre a existência de diferenças entre os dois tratamentos é igual independentemente do método utilizado. Verifica-se que na presença de

percentagens mais elevadas de dados omissos, as estimativas obtidas utilizando o método cLDA são mais eficientes que quando utilizando o método ANCOVA. Pela análise dos gráficos de resíduos, verificamos que ambos os modelos são adequados, uma vez que o perfil dos resíduos é adequado. Quanto aos resultados obtidos no estudo de simulação, as estimativas obtidas por ambos os modelos nos vários cenários não são enviesadas. Quando a linha-base segue uma distribuição normal, a cobertura dos intervalos de confiança a 95% para a estimativa da diferença de tratamentos é adequada, no entanto, face a desvios à normalidade, o comportamento do modelo cLDA passa a não ser adequado. Nestas condições, mesmo na presença de dados omissos, as estimativas dadas pelo modelo ANCOVA continuam a ser melhores. O referido indica então que a utilização do cLDA pode ser apropriado, especialmente em casos em que os dados seguem uma distribuição normal, e pode mesmo trazer vantagens, especialmente na presença de dados omissos. No entanto, no caso dos dados terem um desvio à normalidade, o método ANCOVA mostrou ser uma melhor alternativa independentemente das outras condições.
Verifica-se então que nem aplicar o modelo cLDA, nem ANCOVA é uma resposta direta à questão de qual metodologia deverá ser utilizado. É necessário ter em consideração características específicas dos dados.

Palavras-chave: Estudos longitudinais, modelos lineares mistos, ajustamentos ao valor de linha-base, ANCOVA, cLDA

Contents

List of Figures

List of Tables

List of Aberviations

AIC - Akaike Criterion
ANCOVA - Analysis of Covariance
BIC - Bayesian Information Criterion
BMI - Body Mass Index
cLDA - Constrained Longitudinal Data Analysis
DYRK1A - Dual Specificity Tyrosine Phosphorylation Regulated Kinase 1A
DS - Down Syndrome
EGCG - Epigallocatechin Gallate
LMM - Linear Mixed Model
LRM - Linear Regression Model
LSC - Least Squares Criteria
MSE - Mean Squared Error
SMD- Standardized Mean Difference
SSE - Sum of Squares Error

Chapter 1

Introduction

1.1 Motivation of the thesis

In the evidence-based medicine era where we live in, clinicians conduct studies to evaluate the effect of an intervention, treatment, or exposure in individuals to support patient management and prescribing decisions (Ligthelm et al. 2007). In these studies, if several measurements across time are taken from the same individual, a measurement is often taken at baseline before a certain treatment starts to be administrated. When it comes to analyse the results obtained at the clinical trial, different approaches to this measurement can be used. The main ones studied so far are the use of baseline measurement as a covariate or as part of the response vector, by applying Analysis of Covariance (ANCOVA) (Fisher 1970) or Constrained Longitudinal Data Analysis (cLDA) (Liang and Zeger 2000) respectively. The approach used can have an impact on the efficiency and accuracy of the estimates obtained (Liu, Lu, et al. 2009), that is why it is important to access the main advantages and disadvantages of the application of both methods. Several authors studied the different ways to treat baseline in an analysis, and some advantages and drawbacks of each one were identified. Coffman et al. (2016) and Liu, Lu, et al. (2009) reached the conclusion that cLDA model is more appropriate for the analysis of data in clinical trials, however, the studies conducted by Lu (2010) found no significant differences between the methods. The lack of consistency on the published literature about the topic, contributes to the difficulty to establish a standard statistical approach.

In this thesis, the methods ANCOVA, that includes baseline in the model as a covariate and cLDA, where baseline is treated as a response, are going to be compared by analysing pre-post and longitudinal data with different characteristics.

Liang and Zeger (2000) proposed cLDA in 2000 as an alternative approach to ANCOVA previously established by Fisher (Fisher 1970) to deal with baseline data. Although it was first presented to deal with pre-post data, extensions to longitudinal data have been made through the years. Both models analytical

estimations are extensively described in the literature, as well as the analytical demonstrations of the comparisons of the estimators and respective variances (Liu, Lu, et al. 2009), so this is out of the scope of this work. The primary focus of this thesis is the application of the previously mentioned methods to pre-post and longitudinal data under different scenarios in order to try to reach some conclusions about what is the most adequate method trough empirical results.

1.2 TESDAD Study

Trisomy for human chromosome 21 results in Down's Syndrome (DS) is one of the most common genetic perturbations, affecting more than 5 million people worldwide (Dierssen 2012). Individuals with DS have, among other characteristics, intellectual disability, known to be related with the over expression of some genes. Tyrosine-(Y)-phosphorylation regulated kinase 1A (DYRK1A) is described to have an impact on the intellectual and cognitive deficit of individuals with this condition (Dierssen et al. 2009). In previous studies, epigallocatechin-gallate (EGCG), a green tea flavonoid has been shown to be an inhibitor of DYRK1A (Bain et al. 2003; Guedj et al. 2009). Based on that evidence, De la Torre, Sola, et al. (2016), conducted a study in order to assess if the administration of that DYRK1A inhibitor could rescue cognitive deficit. This was done first in mice over expressing DYRK1A trissomic and dissomic, to ensure that the EGCG administration had the capability to reduce both the DYRK1A kinase activity in the hippocampus and the plasma homocysteine levels (a biomarker of the presence of DYRK1A). The same research project also included a pilot study with human individuals with Down's Syndrome. The results obtained were favourable to the hypothesis presented (De la Torre, Pons, et al. 2014).
Due to the results from that previous study, the safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD): a double-blind, randomized, placebo-controlled, phase-2 trial was conducted (De la Torre, Sola, et al. 2016). It included patients with DS from both sexes, aged 16-34 years old and with a body mass index (BMI) between 18.5 kg/m^2 and 29.9 kg/m^2. Eighty-seven patients were randomized into two arms: placebo and treatment. The intention-to-treat groups had 43 people in the treatment group and 41 in the placebo group (there were 3 dropouts before the start of the treatment). During a period of 12 months, the two groups of the study received either a green tea extract supplement containing 45% EGCG or a placebo, both combined with cognitive training sessions. The patients' allocation to the treatment groups was balanced by sex and intellectual quotient (IQ). It is important to take into account this last factor since IQ could have a confounder effect. Neuropsychological characteristics of interest studied were: psychomotor speed, attention, visual episodic memory and learning, executive functioning, verbal word fluency, working memory, planning capacity, expressive and receptive language, adaptive behaviour, and quality of life. The biomarkers plasma homocysteine and transthyretin concentrations were also measured, since their concentration is known to be correlated with the expression of DYRK1A (De la Torre, Pons, et al. 2014).

The main objective of the study was to identify differences in those outcomes between treatment and control groups. After the 12 months, it was concluded that the administration of EGCG was having a positive effect on adaptative behaviour in functional academic skills, immediate visual memory recognition tasks, and inhibition control. The improved results on the last measurement were still identifiable 6 months after treatment discontinuation.

Study Data

In order to evaluate change over time for the group as a whole and particular individuals, measurements were taken at baseline, 3, 6 and 12 months after treatment initiation, and 6 months after treatment discontinuation. A battery of neuropsychological tests was applied to the individuals at the mentioned time points in order to assess the individual's progress through time. The results of the tests among other plasma biomarkers concentrations were treated as the primary outcome variables. Information about patient's sex, intellectual quotient, weight, height, age, and scholar education was also collected since these variables are referred in the literature as a possible confounders. The measured variables used for this master thesis were the inhibitory control, obtained by applying the Cats-and-Dogs test and measuring score and time to complete, the homocysteine levels, and adaptive behaviour obtained, by applying ABAS-II functional academics score. Cats-and-Dogs test consists on presenting images of cats and dogs combined with white noise, and the participants have to indicate whether the image is a cat or a dog (Weil et al. 2017) and the ABAS-II functional academics is a questionary where the respondent has to indicate if the individual being assessed is able to perform an activity independently and if so, how frequently (Rust and Wallace 2004). For both tests, higher scores correspond to improvement of behaviour and the shorter the times, the faster the individuals were to complete these tests. Concerning homocysteine levels, it is a biomarker of the presence of activity of the dual-specificity tyrosine-(Y)-phosphorylation-regulated kinase 1A (DYRK1A) (Noll et al. 2009). Being EGCG an inhibitor of DYRK1A, the decrease of homocysteine concentrations represents the existence of an effect of EGCG. Its concentration was measured in the patient's blood.

1.3 Simulation Studies

As previously mentioned, after the application of the selected methods to real data, they were also applied to simulated data with certain characteristics of interest.
Simulation studies enable the generation of data by pseudo-random sampling from a previously known distribution. They are particularly valuable in the context of statistical research since they enable to obtain empirical results regarding the performance of certain statistical methods under different scenarios. Some examples of its main uses are the comparison of the behaviour of different methodologies, the calculation

of sample sizes for a study with certain characteristics, the evaluation of a new statistical method, among others.
Similarly to what happens in empirical studies, when planning a simulation study, some key steps must be taken. The ones described on the Aims, Data-generating mechanisms, Methods, Estimandes, Performance measures (ADEMP) (Morris et al. 2019) approach are:

- Identify the simulation's specific aims and targets
- Determine the data generation mechanisms
- Search appropriate methods to be evaluated
- List the performance measures to be estimated
- Compute the code and execute it
- Analyse the results obtained

Focusing on a study with main target is an estimator, several parameters are either previously established or calculated through simulation. Let's denote θ, as the true value of the estimand, n_{obs} the sample size, n_{sim}, the number of simulations preformed, $\hat{\theta}$, the estimator of θ, $\hat{\theta}_i$, the estimate of θ from the ith repetition and $\overline{\theta}$ the mean of $\hat{\theta}_i$ across repetitions. For this kind of simulation to achieve its aims, some properties of the estimator of interest $\hat{\theta}$ must be accomplished. It must be consistent and unbiased, its estimated variance needs to be a consistent estimate of the true variance and as small as possible. Finally, the percentage of simulated confidence intervals that contain θ should be at least $100(1-\alpha)\%$, with α beings the significance level considered.
Regarding the data generation mechanisms used, data can be simulated either by producing parametric draws from a known model or by repeated re-sampling with replacement from a specific data set. By using a parametric simulation, several scenarios (even unrealistic ones) can be explored, and it is used to ensure the coverage of this different scenarios. When using resampling methods, it usually explores only one scenario.
Most of the simulation studies are conducted in order to compare statistical methods and the estimators they provide, being the target θ the estimand of a parameter of the data generating model. However, the study can also have the purpose to analyse null hypotheses, models, predictions or experimental designs and depending on the simulation's target. It is important to denote that one simulation study can have several targets and to evaluate it there are several performance measures that can be applied. When the target is an estimator, bias (difference between the true value of the parameter of interest and the expected value of the estimator) is the measure that is usually calculated, among with precision, mean squared error (MSE) and coverage of confidence intervals. However, as previously mentioned, depending on

the study's aim it can also be of interest to calculate measures such as power or type-I error rate. The formulas of some of these statistics are presented below.

- **Bias:** $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)$
- **Standard Error:** $\sqrt{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \overline{\theta})^2}$
- **MSE:** $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \overline{\theta})^2$
- **Coverage:** $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{low,i} \leq \theta \leq \hat{\theta}_{upp,i})$

Regarding the computational aspects, statistical packages that preform Monte-Carlo simulation use a pseudo-random generator, where each random-number is the deterministic function of the current state of the random-number generator, and the production of each random number results in a change of the state and becomes ready to produce the next random number. The first state can be set using a "seed" in order to obtain, after a large number of random-number draws, repeated states. This also makes the process reproducible. A data set with n_{obs} must be created through this process, followed by the generation of a data set containing the summary information for the n_{sim}performed. Finally, the performance measures are calculated for each data generating mechanism. In Chapter 3 more information will be included on how to compute simulation studies, specially on R software.

1.4 Objectives

The aim of this thesis is to study of different types of linear mixed models, that have different approaches to baseline, apply them to real data and identify which different conclusions are reached. An additional aim is to extend the same analysis to data simulated under different conditions. In order to achieve this, the specific objectives are the following:

- Study different linear mixed models for the analysis of longitudinal data for clinical trials;
- Study of different computational approaches (packages available in R) to model longitudinal data;
- Explore the data from the training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD) study and understand the structure of the variables measured;
- Apply the selected models to that data;
- Generate simulated data followed by application of the models on it;
- Comment the results obtained on the last two items.

1.5 Outline

The work done in order to achieve the objectives is explained in detail in this master thesis. In the first chapter, the concepts and questions that motivated the study of the topic, as well as the objectives that are expected to be accomplished, are explained. In Chapter 2, the theoretical framework behind the practical work is explained, including the models to be applied. Then, in Chapter 3, the study from where the data analysed was taken will be presented as well as the implementation of the previously described methods to that data. In Chapter 4, it is explained how the simulation study was performed and finally, on Chapter 5, the results obtained were discussed taking into account what was expected based on the findings described in the literature about the topic and some conclusions were made.
A schematic representation of the steps that were taken and are going to be presented can be found in Figure 1.1.

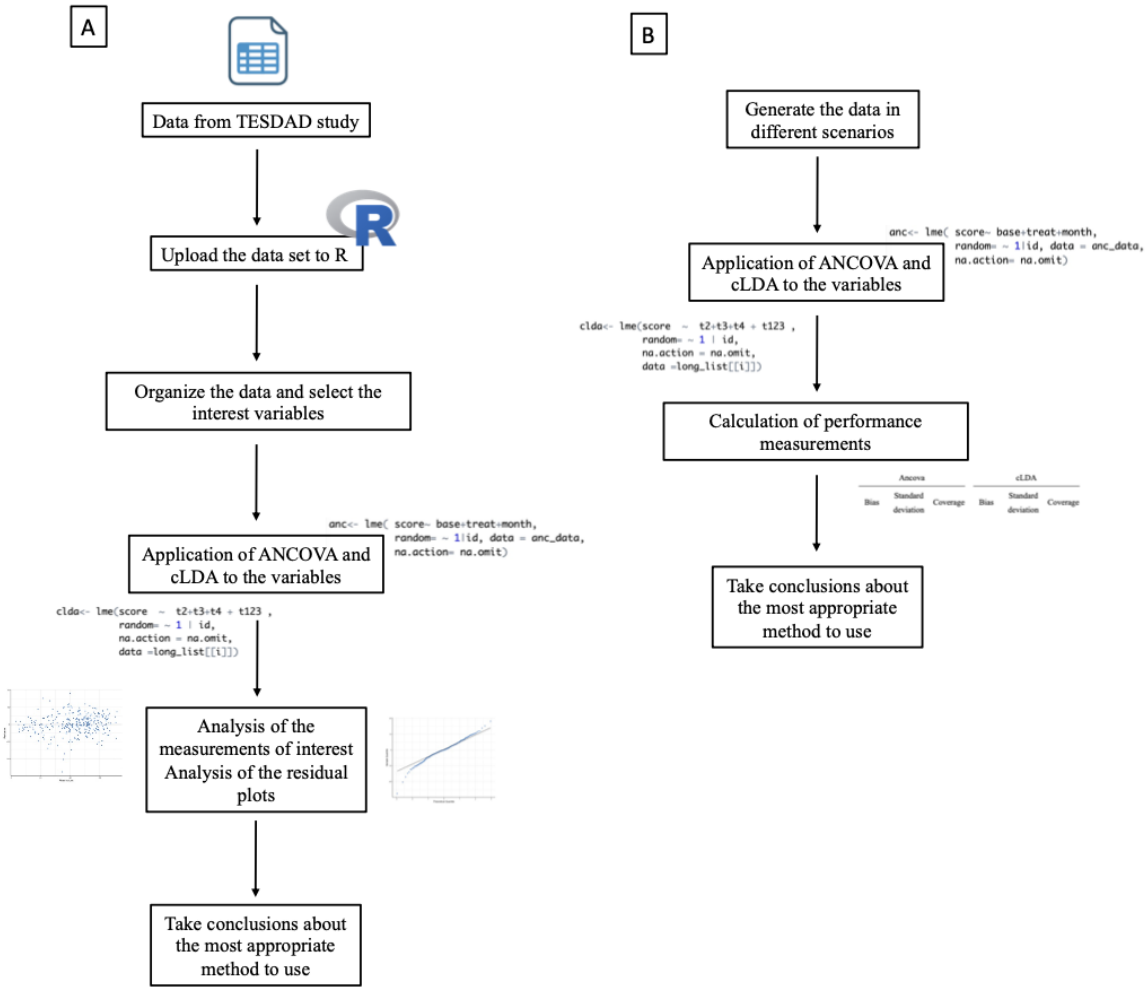


Figure 1.1: Schematic representation of the steps taken on the analysis with real data (A) and with simulation data (B)

Chapter 2

Theoretical framework

2.1 Randomized controlled trials

A clinical trial is a controlled experiment designed to compare in human individuals the efficacy and safety of different medical interventions. Its main objective is to find the most appropriate treatment to future patients with a certain medical condition and to do that, the results obtained in a group of individuals receiving the treatment being tested is compared with the results obtained in a group of identical individuals receiving a control treatment. The concept of "treatment" can be applied to a new drug, surgical procedure or therapy administered to the patients. A randomized controlled trial (RCT), is considered the "gold standard" approach for estimating the effects of treatments, interventions, and exposures on outcomes (Akobeng 2005).

There are generally, several variables that are known for having an interaction or confusion effect on the relationship between the treatment and the outcome. These variables depend on the problem being studied and to participate in the clinical trial, the patient must meet some eligibility criteria, previously defined, which are usually baseline characteristics regarding those variables that can influence the results. Considering these baseline characteristics in the analysis, bias can, generally be avoided, ensuring that it is possible to compare the effect of the different treatments between groups since they are similar in almost every other critical aspect.

Depending on the scientific/medical questions that one wants to answer with the study, the experimental design of the clinical trial can differ. The most simple and common one is the parallel group design. In these cases, there are two or more groups, each receiving a different treatment, that can include control, the treatment being tested in one or more doses, among others. The treatments are randomly assigned to each individual, and each one will receive only one treatment. This design has been indicated by the International Conference of Harmonization as a standard to study the effect of different therapeutics and should be used whenever possible (Singh 2015).Although this is the most applied, designs such as

crossover, sequential, factorial, among others are specific alternatives whose use depends on the specific study objectives. A schematic representation of the parallel design can be found in Figure 2.1.

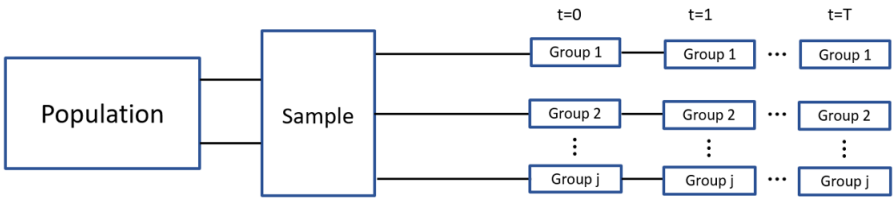


Figure 2.1: Schematic representation of a parallel clinical trial

Achieving proper randomization is fundamental when conducting a RCT, since it helps to prevent bias associated with the selection of the candidates receiving each treatment.
A RCT has five sequential phases that are classified according to their complexity level. In a pre-clinical phase, the drugs are synthesized and experimented in animals. In the phase I, starts the first tests on human individuals, in order to study the safety, toxicity and pharmacokinetic characteristics of the new treatment. Then, in phase II, the therapeutic activity of the treatment is studied, and if its efficacy and safety are demonstrated. The new treatment goes to phase III where, in a larger scale, the effects in the long and short term are studied. Finally, the treatment becomes available in the market and the phase IV starts, where motorization of the rare adverse effects is done (Sims and Miracle 2002).

2.2 Pre-post analysis

In some trials, it can be of interest to use pre-post measurements instead of repeated measures, not only because sometimes it is expensive to measure the response, but also because it can lead to a more intuitive interpretation of the effect detected (O'Connell et al. 2017). In those cases, we are dealing with a pre-post design, and most of the time, the complexity of the statistical analysis required is smaller when compared, for example, to the statistical methods required for longitudinal data.
In a study conducted to assess the effectiveness of a therapy, it is quite common to measure the response variable of the individuals before the beginning of the study and at one defined time point after treatment administration. In the case of a parallel design, with two arms, for example, the baseline and post-baseline measurements of individuals in both treatment and control groups are compared, and the difference detected between the two groups is assumed to be the treatment effect. The estimation of this treatment effect can be done either by using the change scores or the measurements taken at the two-time points.
When performing a pre-post analysis, different approaches can be used. In this kind of analysis, each individual has two measurements of the variable of interest. If they are both part of the response vector,

it will be necessary to use a linear mixed model (LMM) to model it - this method will be explained in detail in the next section. However, if one of the measure is considered a covariate and the other one a response or if the analysis is done with the change score, a linear regression model can be used. In these two cases, a linear regression model (LRM) can be applied.

Linear regression models (LRM)

Linear regression is used in the attempt to model the relationship between two or more variables. One of them is a continuous dependent variable, and the others are the explanatory ones (Draper and Smith 1998).

Notation

When we have n independent observations that we admit are dependent of a set of p explanatory variables, part of vector X_j, $(X_1, X_2, ..., X_p)$, it is common to model their relationship using linear regression. The values of the independent variables, Y are given by a linear combination of p predictor variables:

$$Y_i = \mathbf{X}_i\beta + \epsilon_i \tag{2.1}$$

where,

- $\mathbf{Y}_i$ is the response value for the subject i;
- $\mathbf{X}_i$ is a vector of predictor variables $j = 1, ..., p$ for individual i;
- $\boldsymbol{\beta}$ is vector of the (non random) $p + 1$ parameters of the model;
- $\boldsymbol{\varepsilon}_i$ is the n-dimensional vector of error components $\varepsilon \sim \mathcal{N}(0,\, \sigma^2)$.

The random variable ε does not mean mistake, it represents the random fluctuations around the mean values that are not within statistical control.

Model estimation

In order to estimate the β parameters, the method used will be the Least Squares Criteria (LSC). According to this criteria, the estimates obtained are the ones that produces a collection of errors whose sum of squares is minimal. This will be done by minimizing the distance between two vectors:

- $\widehat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + \hat{\beta}_2\mathbf{x}_2 + ... + \hat{\beta}_p\mathbf{x}_p$, a vector in space defined by $\mathbf{X}$
- $\mathbf{y}$, the vector of observations of the response variable

As mentioned before, using the LSC, the estimations are done in order to minimize the sum of squares of the residuals (SSE)

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p)^2 \end{aligned} \tag{2.2}$$

Using a matrix notation, it takes the form:

$$\begin{aligned} S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \ \varepsilon'\varepsilon &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \end{aligned} \tag{2.3}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

being $\hat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we can simplify by writing $\hat{Y} = \mathbf{H}Y$, where $\mathbf{H}$ is the hat matrix. Then, the vector of residuals can be calculated as: $\hat{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (1 - \mathbf{H})\mathbf{Y}$.

2.3 Longitudinal Analysis

In longitudinal studies, individual measures are taken at several time points to allow the understanding of the degree and direction of change of a certain characteristic over time. This allows the repeated monitoring of health outcomes or risk factors and the identification of differences in outcomes. In Figure 2.2 an example of a graphical representation of results obtained from a longitudinal study is showed.

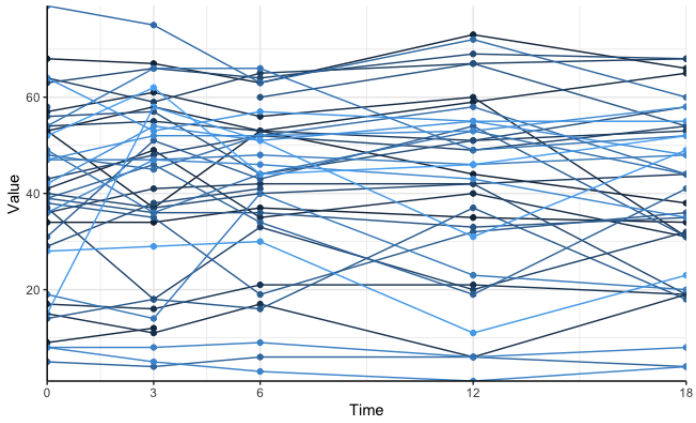


Figure 2.2: Individual profiles in a longitudinal study

Longitudinal studies allow the study of the development of individual profiles. It also increases the precision of the estimated treatment effect, and increases the power to detect such an effect (Zeger and Liang 1992). However, several observations throughout the time of the same individual are not independent, and that can pose as a challenge in data analysis (Albert 1999; Liu, Rovine, et al. 2012). If the statistical approach used, fails to account for the intra-individuals measurements' correlation, it can result in biased point estimates and standard errors that fail to reflect the uncertainty of the data (Mallinckrodt and Lipkovich 2016). The heterogeneity between the individuals' profiles may arise from subject-specific factors that sometimes are unknown, such as genotype or environmental factors. Generally, the biggest concern in longitudinal clinical trials is not the correlation between but within subjects (Mallinckrodt and Lipkovich 2016). The appropriate statistical tests must therefore be applied in order to evaluate change over time for the group as a whole or for particular individuals (Caruana et al. 2015; Liang and Zeger 1993).

Linear Mixed Models

Verbeke (1997) proposed a statistical model that would account for both within and between-subject variability. With between subject variability, one is referring to the heterogeneity between subject's profiles and within-subject variability is the fluctuation of the individual measures over time. This can be modelled in a two stages statistical analysis, where in a first phase the subject-specific longitudinal profiles are approximated by a linear regression function including the fixed effects, and in a second step, the variability between subjects is explained by using a multivariate regression, where the random effects enter. Fixed effects represent population-specific parameters that have the same effect for all the participants, and random effects are cluster-specific parameters, equal for every observation belonging o the same cluster or individual. Unlike fixed effects which are considered to have a constant effect on the

outcome, random effects have a varying effect on the outcome across the individuals. Random effects can be introduced either in the intercept or slope. In the first case, the most common one, we are assuming that each cluster/ individual starts from a different point, and in the second one, that each cluster has its own relationship with the response. Neither of these effects are estimated.

Notation

Considering the random variable Y_{it}, to be the response for the i^{th} individual, $i = 1, 2, ...N$, where N is the total number of individuals. This response was taken at time t with $t = 0, 1, ..., T$, where $t = 0$ is the first measurement taken, before the beginning of the study, also called the baseline, and $t = T$ is the last measurement taken. Each individual can be represented by a vector Y_i of length $T + 1$. A set of p covariates is also related to the subject at each time point, x_{ijt}, where $j = 1, ..., p$. Laird and Ware (Laird and Ware 1982) described the LMM as a model containing some population-specific and some subject-specific parameters, that can be written with the following formulation:

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}) \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sum_i) \end{cases} \tag{2.4}$$

where

- $\mathbf{Y}_i$ is the $(T + 1)$- dimensional response vector for the subject i;
- $\mathbf{X}_i$ is a matrix of $(T + 1) \times (p + 1)$ responses for the fixed effects;
- $\mathbf{Z}_i$ is a matrix of $(T + 1) \times q$ responses for the random effects;
- $\boldsymbol{\beta}$ is a $(p + 1)$-dimensional vector containing the parameters of fixed effects;
- $\mathbf{b}_i$ is a q-dimensional vector containing the random effects;
- $\boldsymbol{\varepsilon}_i$ is the $(T + 1)$-dimensional vector of error components;
- $\mathbf{b}_1, ..., \mathbf{b}_N$ and $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_N$ are independent.

In the context of longitudinal studies to assess treatment effects, for example, the fixed effects represented by the β coefficients can represent the effect of time, treatment, interaction between this two covariates as well as the effect of other variables considered important for the case in study. Each individual is associated with a set of values of the parameters above and can be represented as:

$$\mathbf{Y}_i = \begin{pmatrix} y_{i0} \\ y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}; \quad \mathbf{X}_i = \begin{pmatrix} 1 & x_{01} & x_{02} & ... & x_{0p} \\ 1 & x_{11} & x_{11} & ... & x_{1p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{T1} & x_{T2} & ... & x_{Tp} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}; \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}$$

Based on the equation, it is possible to write the hierarchical model for the distribution of Y_i given the subject-specific random effect.

$$\begin{cases} \mathbf{Y}_i \mid \mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b_i}, \sum_{\boldsymbol{i}}) \\ \mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}) \end{cases} \tag{2.5}$$

Based on the hierarchical model, it is possible to obtain the marginal model, that will allow the estimation average profile. From:

$$\text{Var}(\text{Y}_i) = \text{Var}(E[Y_i|b_i)] + E[\text{Var}(Y_i|b_i)] = \text{Var}(X_i\beta + Z_ib_i) + E[\textstyle\sum_i] = Z_iDZ_i' + \textstyle\sum_i \tag{2.6}$$

it comes,

$$\text{Y}_i \sim \mathcal{N}(X - i\beta, Z_iDZ_i' + \textstyle\sum_i) \tag{2.7}$$

The matrix D can have several structures, depending on the way in which the observations within a subject are correlated with each other. Some examples are:

- **Compound symmetry**: it is assumed that, for a subject, the correlation of a pair of any two observations is the same.
- **Autoregressive**: it is assumed that the correlation between two observations of subjects tends to decrease with the increasing of the time that passes between those two observations.
- **Unstructured**: it is assumed that the correlation between a pair of observations can be different from any other pair of observations.
- **Diagonal**: it is assumed that the correlation between any pair of observations is zero.

Model estimation

When estimating the parameters of a LMM, as long as a Bayesian approach is not used, the inference is done using the marginal and conditional distribution for Y_i. The marginal model is used to estimate the fixed effects, the resulting coefficients of this estimation will be similar to the ones estimated when using a linear regression model and allows to make inference about the average profiles.
Assuming that α is a vector containing all the variance and covariance present in $V_i = Z_i D Z_i' + \sum_i$, and $\theta = (\beta', \alpha')$ is a vector with all the parameters present in the marginal model that are going to be estimated. The inference about the average profile will be done by maximizing the marginal likelihood function with respect to θ

$$L(\theta) = \prod \{(2\pi)^{N/2} |V_i(\alpha)^{-1/2} \exp(-\frac{1}{2}(Y_i - X_i\beta)' V_i^{-} 1(\alpha)(Y_i - X_i\beta)\} \tag{2.8}$$

Assuming that α is known, the maximum likelihood estimator of β is given by:

$$\hat{\beta}(\alpha) = \Big(\sum_{i=1}^{N} X_i' W_i X_i\Big)^{-1} \sum_{i=1}^{N} X_i' W_i y_i \tag{2.9}$$

where $W_i = V_i^{-1}$ and N the number of individuals in the sample.
In the cases where α is not known, it is necessary to estimate its value, which can be done using either maximum likelihood estimation or restricted maximum likelihood estimation.

Maximum likelihood estimations

In order to obtain the maximum likelihood estimations of the parameter α, the expression (2.8) must be maximized with respect to α and the β be replaced by (2.9). When μ is known an unbiased estimator of σ^2 can be obtained.

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \mu)^2}{N} \tag{2.10}$$

However, when the parameter μ is unknown, the MLE estimator obtained is biased.

$$E(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 \tag{2.11}$$

In the context of LMM this situation can be problematic in the cases where β is unknown and also has to be estimated from the data. Although the previous expression approaches the true variance values when sample sizes are large, a distance from the real value of the variance is noted when the size is smaller.

Restricted maximum likelihood estimations

To obtain an unbiased estimate, it is necessary to use an estimator that does not depend on the mean value. This can be done with the restricted maximum likelihood estimator. Harville (Harville 1977) proposed the following expression:

$$\begin{aligned} L(\alpha) = (2\pi)^{-\frac{(n-p)}{2}} & \left| \sum_{i=1}^{N} X_i' X_i \right|^{1/2} \\ & \times \left| \sum_{i=1}^{N} X_i' V_i^{-1} X_i \right|^{-1/2} \prod_{i=1}^{N} |V_i|^{-1/2} \\ & \times \exp\Big(-\frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i\hat{\beta})' V_i^{-1} (Y_i - X_i\hat{\beta}) \Big) \end{aligned} \tag{2.12}$$

where, $\hat{\beta}$ is given by (2.9).

In the cases of LMM, however, it is also necessary o estimate the random effects, in order to identify how much the subject specific profiles deviate from the average model.

The random effects are assumed to be random variables, so a Bayesian approach comes naturally when deciding what technique should be use to estimate them. Based on this method, we have that the prior distribution of b_i (the one that does not depend on Y_i), is $\mathcal{N}(0, D)$. Once the values of Y_i are observed, the posterior distribution of b_i can be obtained.

$$\hat{b}_i(\theta) = E[b_i|Y_i = y_i] = \int_i f(b_i|y_i) db_i = DZ_i' W_i(\alpha)(y_i - X_i\beta) \tag{2.13}$$

$$\text{Var}(\hat{b_i}(\theta) = DZ_i'\{W_i - W_i X_i(\sum i = 1^N X_i' W_i X_i)^{-}1 X_i W_1\} Z_i D \tag{2.14}$$

2.4 Models quality assessment

After fitting the models, we can evaluate them by assessing the quality of each one using metrics based on maximum likelihood such as Akaike information criterion (AIC), Bayesian information criterion (BIC) or Likelihood ratio tests. If the models being compared are nested, likelihood ratio tests are usually an option. If this is not the case, AIC and BIC are often used. These methods evaluate how well the model fits the data penalizing by the number of independent variables used to build the model, they favour more parsimonious models. The lower these scores are, the better.

The formula for calculating AIC is: $AIC = 2K - 2\ln(L)$ being K the number of independent variables used and L the value of log-likelihood function.

The formula for calculating BIC is: $BIC = \ln(n)K - 2\ln(L)$, where n is the sample size, k the number of parameters which the model estimates. Either the restricted maximum likelihood or the full likelihood can be used.
As previously mentioned, there are some assumptions that must hold when applying linear mixed models: normality, linearity, homocedasticity and independence. Residual plots are a useful tool to examine these assumptions.

- Linearity: plot of the residuals against x_i.
- Independence: a plot of the residuals in the order the observations were obtained.
- Normality:plot of the empirical quantiles of the residuals against the theoretical quantiles of the standard normal distribution.
- Homocedasticity: graphic representation of the residuals against the fitted values.

2.5 Data modelling approaches

2.5.1 Analysis of change scores

The change score is the result obtained by subtracting from the follow-up score, the baseline score. It is not possible to calculate the change score for an individual who is missing one of those two values, so in those cases, subjects are excluded from the analysis (Vickers and Altman 2001). This method is only used to analyse data from pre-post studies since when applied to longitudinal data, a lot of information is lost. Analysis of change scores does not correct for baseline differences, since if the correlation between baseline and follow-up scores is low, the regression toward the mean effect expresses. This effect first described by Galton (Galton 1886) refers that, there is a tendency that even though the baseline score may be extreme, the follow-up score is closer to the mean of the group. In the case of a low correlation, change score became less reliable than scores themselves since, for example, if by chance the baseline values are lower in the treatment group, the scores would be bigger and the treatment effect might be overestimated (Vickers and Altman 2001).

2.5.2 Analysis of Covariance (ANCOVA)

Analysis of Covariance (Fisher 1970)is used to analyse data with continuous numeric response. Over the last years, published literature advocates the use of ANCOVA when there is the need to adjust for baseline. In this method, response value can be the post baseline measurements or the change scores, and the result obtained for the expected difference between groups is the same regardless of the values used. In studies where more than one post-randomization measurement is taken, the longitudinal ANCOVA

model can be applied.
Subjects who only have baseline values, or who do not have them, cannot be introduced in the analysis. Together with that one, the use of ANCOVA can introduce problems. In some cases, it is possible that the probability of dropout depends on the baseline value, and the exclusion of that individuals from the analysis can introduce bias on the individual profile estimations. When the baseline is the only value missing, since it must be missing completely at random, it does not introduce bias in the estimations. However, since the individuals without baseline measurement must be excluded from the analysis, the information contained in those post-baseline values is not considered and the power to detect a difference between treatments may be reduced (Lu 2010). Generally, the power to detect differences between treatments is higher when using ANCOVA, comparing for example with the change of scores ANOVA. However, these differences in power tend to disappear when the correlation between the baseline and post-baseline value is high ($r > 0.8$) (Vickers and Altman 2001).
Under the assumption of bivariate normality for baseline and post-baseline measurements, estimates and statistical tests from the ANCOVA model conditional on baseline values are unbiased and valid even when the baseline is a random variable. Depending on the kind of study one is dealing with, ANCOVA can be an adaptation of a linear regression model or of a linear mixed model, to fit data from pre-post or longitudinal studies respectively.

2.5.3 Constrained Longitudinal Data Analysis (cLDA)

Proposed by Liang and Zeger (2000), cLDA is a conditional data analysis method built under the assumption that the randomization of the subjects involved was efficient. It is assumed that means at baseline are identical for the groups being compared, which, in the context of RTCs is reasonable to assume. Due to this, both baseline and post-randomization values are part of the dependent variable (Liang and Zeger 2000).
The question of whether baseline can or not be included in the outcome vector is controversial. Since the value of baseline of an individual is measured before randomization and treatment initiation, some authors argue that it should not be used to model the treatment effect (Senn 2014). Furthermore, in some studies, baseline values are an inclusion criteria, so they are likely to be truncated (Dinh and Yang 2011). However, others argue that treating baseline as a fix effect, as happens in ANCOVA can result in loss of efficiency. Despite this discussion, both cLDA and ANCOVA are referred as being the most efficient ways to model longitudinal data (Dinh and Yang 2011).
In cLDA randomized subjects with at least one observation can be included in the analysis, which is an advantage of this method compared with ANCOVA, since even without imputation is possible to include all individuals with a measurement in the analysis (Lu 2010). Considering a case where there is no missing data, the estimates for the group differences obtained by ANCOVA and cLDA are expected

to be similar since estimation of fixed effects is made based on the maximum likelihood for both methods (Liang and Zeger 2000). However, the variance of the point estimate obtained by cLDA is smaller than the one obtained by ANCOVA, being the first method more efficient. In cLDA, the baseline and post-baseline values follow a jointly multivariate normal distribution (Liu, Lu, et al. 2009).

2.6 Source bias

When conducting a study, there are a set of systematic errors that can occur in the design, conduction, or analysis and can lead to incorrect conclusions. These are called bias, and they can have several natures, including selection and confounding bias (Schlesselman 1982). In the context of clinical trials, the first ones are related to bad or non-randomization of the subjects included in the study, and the second ones are related to a non-adjustment between groups of the variables that are related to both the exposure and outcome being studied. The occurrence of bias is related to background covariates that may be strong predictors of the outcome. In order to deal with these covariates and prevent bias, there are a set of procedures that can be applied when designing a study. Randomization of the individuals, restrict the enrolment of individuals with certain characteristics in the study, and balance the groups regarding certain individuals' characteristics.
In clinical trials these procedures are followed, however in some situations, the inclusion of background covariates in the analysis may be important since it can increase the precision of the estimated effect. As an example, in cases where patients come from several centres, for example, they tend to be more similar to each other, so in order to have more accurate results it can be of interest to adjust for this variable.

2.7 Missing data

In longitudinal data, the occurrence of missing data is very common. This happens when an individual misses one scheduled visit but attends a subsequent one. It is important to denote the difference between patients with missing data and dropouts. The last one happens when, after a visit, the individual does not return for the subsequent. In order to decide what approach should be used to deal with the missing data we are faced with, it is important to identify the reason why the data is missing. Data is considered missing completely at random (MCAR) when missing data is independent of the observed and unobserved outcome, it means that there should not have been differences between individuals with missing data and the completers. When data is missing at random (MAR), it means that missing data is related to the observed but not the unobserved outcomes, it means that a characteristic of the individual could justify the fact that she/he did not attend to the visit but is not related to the information that was going to be collected. This is the most usual, and consequently important, case of missing data. In the presence of these two types of missing data, although some power may be lost, the estimates of the parameters

should remain unbiased. Finally, the data missing not at random (MNAR) results from the cases when the missingness is directly related to the unobserved data (Kang 2013; Verbeke 1997). In the case of MCAR and MAR, the analysis of the observed data does not introduce bias.

Chapter 3

TESDAD study

In this chapter, the statistical methods previously described are applied to the data from the TESDAD study. The main goal is to assess whether different models applied lead to different results. In order to do that, the estimations obtained for the coefficient representing treatment differences and respective variances will be compared. Model assumptions will also be analysed. The variables that were selected for this analysis are the ones which showed to be significantly different for individuals taking EGCG, compared to the ones taking placebo. The variables in study as primary outcomes will be the Cats-and-Dogs test score and time, ABAS-II functional academics score and homocysteine levels. The analysis on the published study (De la Torre, Sola, et al. 2016) was done using ANCOVA.

3.1 Exploratory Analysis

An exploratory analysis was conducted in order to understand and obtain information from the data being studied. Some baseline measurements of demographic and clinic characteristics of the subjects were taken at the beginning of the study in order to assure that the variables that could have an impact on the treatment effect detection were balanced in both groups. In Table 3.1 is presented a summary of those measurements.

Table 3.1: Summary characteristics of TESDAD study

	Placebo plus cognitive training (n=44)	**EGCG plus cognitive training (n=43)**
Sex		
Male	21 (48%)	24 (46%)
Female	23 (52%)	19 (44%)
IQ category		
<40	18 (41%)	18 (42%)
40-69	26 (59%)	25 (58%)
Age (years): mean	23.4	23.1
Body mass index (kg/m^2): mean	25.9	25.6

In an RTC, baseline variables in the treatment groups should be similar. However, in some cases, it can be sensible to adjust for some potential confounders. These confounders are factors associated with both the treatment and the outcome, for example, a disease or death that are not part of the causal pathway from treatment to outcome (Kahlert et al. 2017). Analysing Table 3.1 it is possible to see that the variables are balanced, so this would be a reason not to include them in the analysis. In randomized controlled trials, when the covariates are balanced in the treatment groups, its inclusion in the model to detect the treatment effect, might inflate the variance of the estimates without adding important information.
In this case, although the variables are balanced, it was denoted that there were relevant differences of the values of some of the studied variables for the two different categories of the variable sex. For that reason, sex was included in all the analysis performed. The presence of missing values in the analysis was also studied, and the results can be found on Table **??**. Although from the analysis of table it is possible to denote that for some variable the percentage of missing data is slightly bigger than 5% it was assumed that the values were MAR and methods based on maximum likelihood have good behaviour under the assumption of data being MAR. Due to the fact that the methodologies to be applied are based on maximum likelihood, there was not done any kind of imputation. Despite that, one has to be careful with the missing values when analysing the data with ANCOVA. In this kind of modelling, when baseline or all post-baseline measurements are missing, the individual is not included in the analysis. While the first situation does not pose as a problem, the fact that some post-baseline measures are missing may depend on the baseline values. This implies that the baseline mean of the individuals after excluding those individuals is different from the baseline mean of all individuals enrolled, which may lead to biased

Table 3.2: Missing values in the data

		Complete cases	Missing baseline	Missing post-randomization	Total
Cats-and-Dogs test score					
	EGCG	41 (95.3%)	1	1	43
	Placebo	33 (80.5%)	5	7	41
Cats-and-Dogs test time					
	EGCG	41 (95.3%)	1	1	43
	Placebo	33 (80.5%)	5	7	41
ABAS-II functional academic score					
	EGCG	39 (90.7%)	0	4	43
	Placebo	35 (85.4%)	0	6	41
Homocysteine levels					
	EGCG	40 (93.0%)	1	3	43
	Placebo	38 (92.7%)	0	3	41

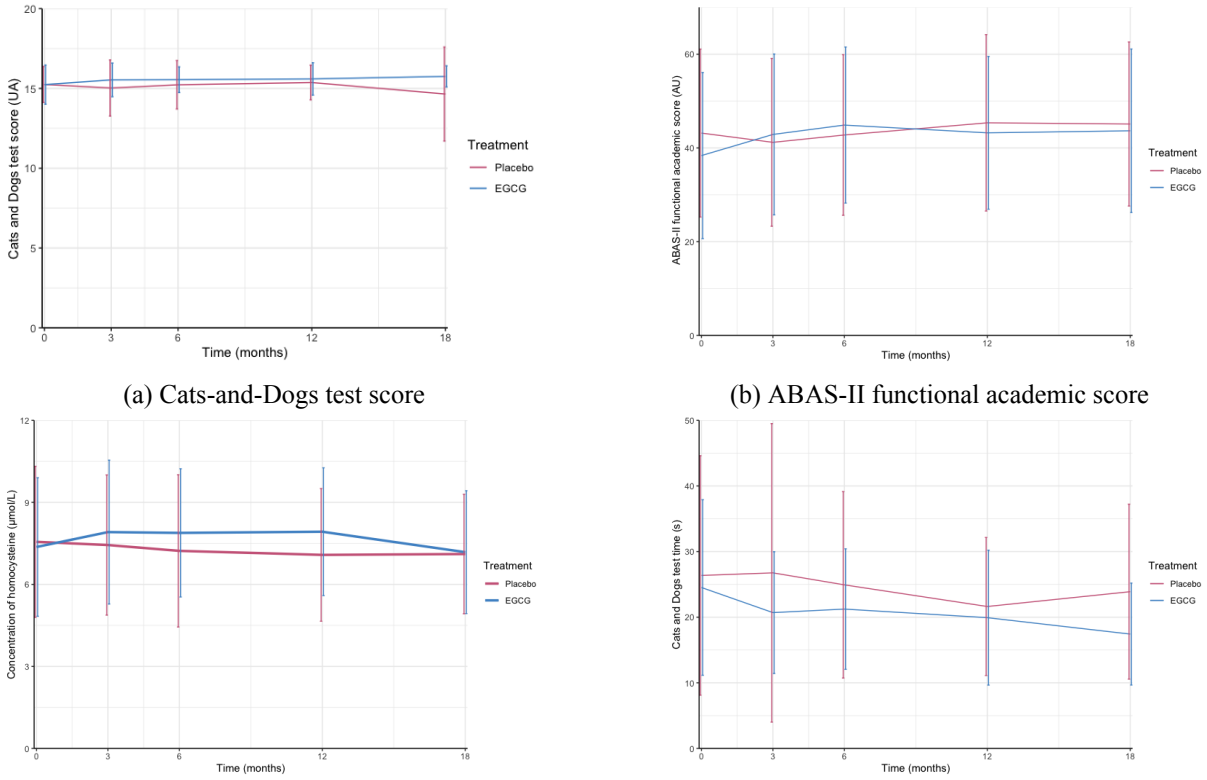
results of treatment effect detection. It is important to denote that the evaluation of the two methods in the presence of missing values was also one of the main goals with the analysis, so that is another reason why imputation was not considered.

In Table 3.3 summary statistics (mean and standard deviation) for each response variable of interest at each time point are presented.

Table 3.3: Descriptive analysis of several variables of interest

	Placebo plus cognitive training		EGCG plus cognitive training	
	Mean	Std. Deviation	Mean	Std. Deviation
Cats and Dogs test score				
t=0	15.250	1.131	15.238	1.226
t=3	15.027	1.756	15.535	1.054
t=6	15.229	1.516	15.548	0.803
t=12	15.371	1.087	15.595	1.014
t=18	14.649	2.946	15.756	0.663
Cats and Dogs test time				
t=0	26.361	18.241	24.524	13.375
t=3	26.757	22.744	20.698	9.275
t=6	24.429	14.201	21.236	9.180
t=12	21.629	10.516	19.929	10.278
t=18	23.892	13.329	17.427	7.767
ABAS-II functional academic score				
t=0	43.171	17.916	38.372	17.721
t=3	41.195	17.899	42.878	17.157
t=6	42.763	17.127	44.881	16.641
t=12	45.361	18.820	43.220	16.296
t=18	45.105	17.497	43.667	17.438
Homocysteine levels				
t=0	7.556	2.762	7.367	2.533
t=3	7.439	2.563	7.912	2.629
t=6	7.225	2.784	7.883	2.345
t=12	7.079	2.426	7.926	2.337
t=18	7.110	2.184	7.179	2.246

A graphical representation of the mean profiles for these variables are represented in Figure 3.1.



(a) Cats-and-Dogs test score

(b) ABAS-II functional academic score

(c) Homocysteine levels

(d) Cats-and-Dogs test time

Figure 3.1: Mean profiles and 95% confidence intervals of the interest variables

When studying what is the most appropriate model to analyse the data, as mentioned before, something that should be taken into account, is how similar measures taken on the baseline of the response variable are between the two groups. With that goal, the standardized mean difference (SMD) of baseline measurements between groups of each variable was calculated. This measure, also called Cohen's D, is a useful tool to measure the difference between two group means by calculating how many standard deviations are there between the two means. To interpret it, it is said that a value is small from 0.2 to 0.5, medium from 0.5 to 0.8 and large from 0.8. Values under 0.2 are considered negligible (Cohen 1988). The results obtained are presented in Table 3.4.

Table 3.4: Standardized mean differences (SMD) for the variables analysed

Cats-and-Dogs test score	**Cats-and-Dogs test time**	**ABAS-II functional academic score**	**Homocysteine levels**
0.010	0.116	0.269	0.072

The only quantifiable value of SMD obtained was one of the differences between the baseline value between groups of the variable ABAS-II functional academic score. Although is classified as small, the values given for interpretation are only a reference. With that said, it is possible that even though the difference is small, it might have an impact when analysing the data. Analysing Figures 3.1, it is also possible to denote that the variable with the highest difference between groups at $t = 0$ is ABAS-II functional academic score.

3.2 Pre-post analysis

The models described in Section 2.2 of Chapter 2 are going to be used to perform a pre-post analysis. There will be used data from the variables Cats-and-Dogs test score and time, ABAS II functional academic score, and homocysteine levels for $t = 0$ and $t = 18$ of TESDAD study.
These variables were analysed using the methods ANCOVA, cLDA, and change scores analysis.

3.2.1 Model building

Considering all the notation presented before, in the context of this study, in the case of the pre-post analysis, the expression for Y_i for a treatment difference W_i is given by:
ANCOVA:

$$Y_i = \beta_1 T_{1i} + \beta_2 W_i + \varepsilon_i \tag{3.1}$$

Y_i is the post-baseline response for individual i.
Change scores:

$$Y_i = \beta_0 + \beta_1 W_i + \varepsilon_i \tag{3.2}$$

Y_i is the change-score for individual i.
cLDA

$$Y_{it} = \beta_0 + \beta_1 T^*_{it} W_i + \varepsilon_{i,j}, \quad t = 0, 1 \tag{3.3}$$

$$T^*_{it} = \begin{cases} 0, \text{ if t=0 (baseline)} \\ 1, \text{ if t=1} \end{cases} \tag{3.4}$$

where Y_{it} represents the response of individual i at baseline (t=0) or at post baseline (t=1)

In all cases, if there is evidence of the existence of interaction between variables in the analysis, this interaction must be also included.

3.2.2 Model fitting

The analysis in this master's thesis were done using version 4.2.0 of the software R. To perform the pre-post analysis with change scores analysis and ANCOVA, where each individual only had one response, the function `lm` from package **stats** was used. It allows the fitting of linear regression models.
Formulation: `lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`
The purpose of each argument can be found in Table 3.5.

Table 3.5: Explanation of the arguments of the `lm` function

Argument	**Function**
`formula`	Two-sided formula divided by $\sim$ where on the right we can find a vector with the response for the individuals and on the left the covariates that will be used to model the response.
`data`	Data frame containing the necessary information to build the model
`subset`	Specify the subset of observations to be included in the analysis
`na.action`	Indicates what kind of approach should be applied to the missing values detected in the data frame

To perform the analysis where the response had more than one time point, computational methods for longitudinal analysis were used. There are two available packages that can be used to model linear mixed models: **nlme** (Pinheiro et al. 2023) and **lme4** (Bates et al. 2015). Although the functions of the packages are very similar, the use of a specific one of them can be more advantageous, for example, when the distribution of the data we are dealing with is not normal, or when the data has a complex variance-covariance structure. For the analysis of the TESDAD data, as it does not have these specifications, any of the two packages could be used, so the **nlme** was chosen. The function `lme` allows to fit a linear mixed effects model as described by Laird and Ware (1982). The function in **lme4** that can be used for the same purpose is `lmer`.
Formulation: `lme(fixed, data, random, correlation, weights, subset, method, na.action, control, contrasts = NULL, keep.data = TRUE)`
The purpose of each argument can be found in Table 3.6.

Table 3.6: Explanation of the arguments of the `lme` function

Argument	**Function**
`fixed`	Two sided formula divided by $\sim$ where on the right we can find a vector with the response measure for all the individuals at all time points. On the left, can be found the variables considered to be fixed effects.
`data`	Data frame containing the necessary information to build the model
`random`	Contains the variables that should be considered as random effects and information about where do the fixed effects have impact
`correlation`	Describes the within group correlation structure
`weights`	Describes the within group heteroscedasticity structure
`method`	Indicates if should be used maximum likelihood or restricted maximum likelihood for model estimation
`na.action`	Indicates what kind of approach should be applied to the missing values detected in the data frame

When fitting the models to the TESDAD data, the guidelines given by Verbeke (1997) were followed. Among with other aspects, it is denoted that the random effects are meant to collect all variability in the data that cannot be explained by the fixed effects. In this work, random effects were added only to the intercept due to clinician's belief that the effect of treatment would be the same for all individuals.
Regarding the covariance matrix of the random effects, two options are available: we can either estimate all components from D, or we can assume independence among the random effects. In this analysis, the second approach was used, so the term correlation remained unspecified. Different correlation structures were studied, however regardless of the specified correlation the results were the same. Due to this, and because the main aim of this work was not to optimize the models in order to estimate a parameter, the models were fitted for independence of random effects.
As an example of the use of the previously explained functions for the methods being studied, the R code for the pre-post analysis of the variable ABAS-II functional academic score is presented bellow.

ANCOVA

```
anc <- lm(academfuncion_18 ~academfuncion_0 + treat + sex ,
          data = data_wide, na.action = na.omit)
```

cLDA

```
long<- subset(data_long, month %in% c(0,18))
long$month <- as.factor(long$month)
long$t1 <- as.integer(long$month == 18)
long$t <- ifelse(long$month == "18" & long$treat == "B", 1, 0)
clda<- lme(academfuncion ~ t1 + t + sex, random = ~ 1 | id,
           na.action = na.omit, data = long)
```

Change scores

```
data_wide$change <- data_wide$academfuncion_18 - data_wide$academfuncion_0
chsc_ac <- lm(change ~ treat + sex , data = data_wide, na.action = na.omit)
```

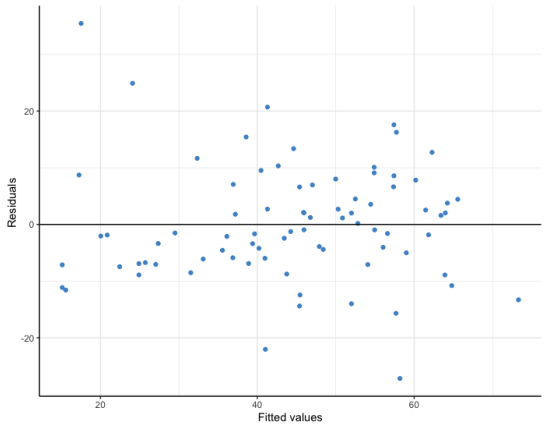
3.2.3 Results

The results obtained are represented in Table 3.7.

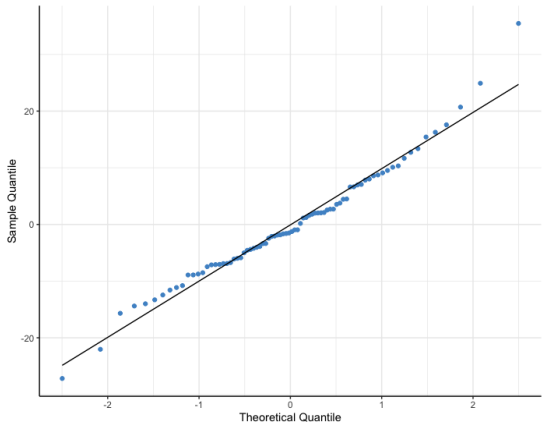
Table 3.7: Pre-post analysis' results obtained of TESDAD study variables

	Estimated difference between treatments	**Standard Deviation**	**p-value**
Cats-and-Dogs Test score			
ANCOVA	0.514	0.210	0.017
cLDA	0.771	0.267	0.005
Change score	0.605	0.268	0.027
Cats-and-Dogs test time			
ANCOVA	-5.212	1.649	0.002
cLDA	-6.338	2.135	0.004
Change score	-6.158	2.225	0.007
ABAS-II functional academic score			
ANCOVA	3.223	2.339	0.172
cLDA	3.323	2.277	0.148
Change score	4.143	2.398	0.08
Homocysteine levels			
ANCOVA	0.149	0.394	0.706
cLDA	0.147	0.427	0.731
Change score	0.280	0.478	0.560

From the analysis of the Table 3.7, it is possible to denote that, for a significance level of 0.05 conclusions about the differences between the two treatments on the different variables are the same for all the models used. For the variables ABAS-II functional academic score and homocysteine levels, the results provided by cLDA and ANCOVA are similar, and it is not possible to find any pattern in the standard deviation, making it impossible to say which method is more efficient. For Cats and Dogs test score and time, the estimations provided by cLDA are slightly more efficient, however this do not mean that they are necessarily better. In order to evaluate if some assumptions of the models hold when they were applied to the data, it was performed an analysis of the residuals. In Figures 3.2, 3.3 and 3.4 there are represented the graphs to evaluate the assumptions normality and homoscedasticity hold. The ones represented in correspond to the variable ABAS-II functional academic score.

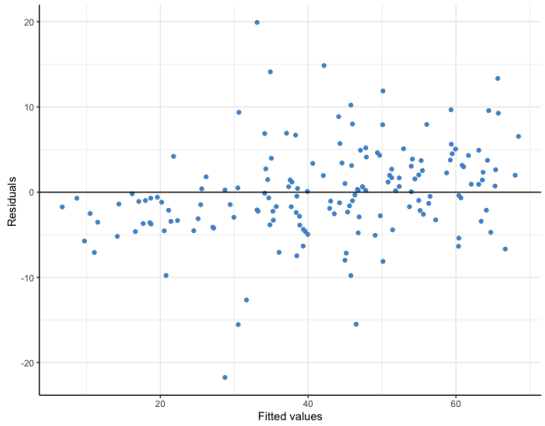


(a)

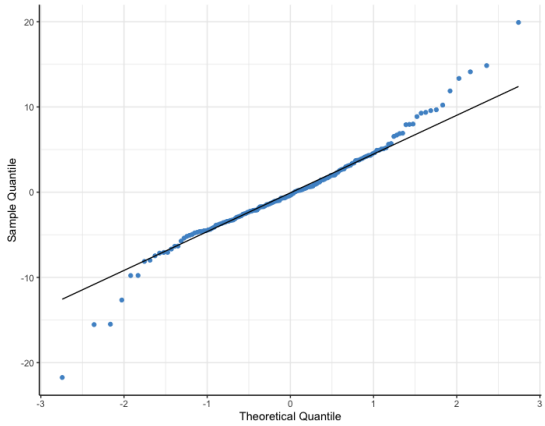


(b)

Figure 3.2: Residuals' plots for the analysis with ANCOVA for the variable ABAS-II functional academic score

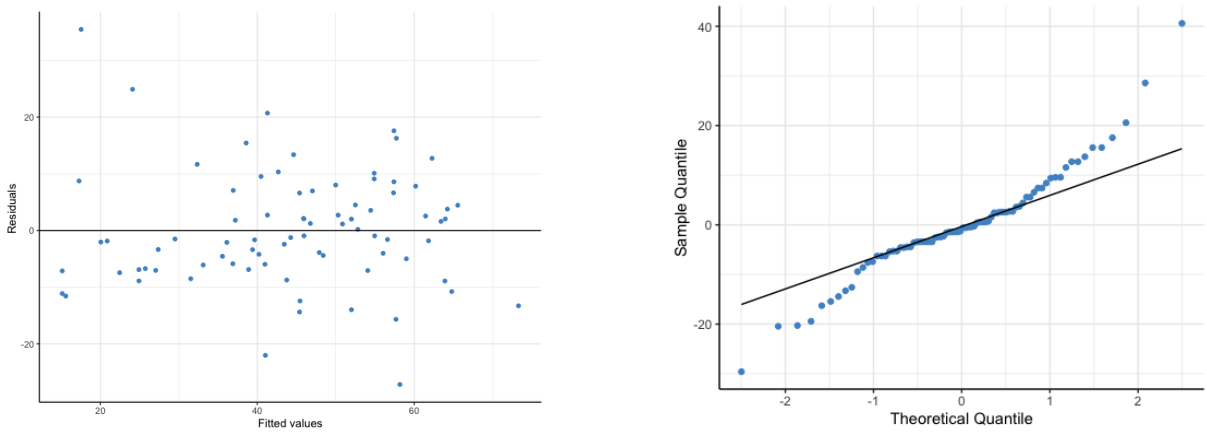


(a)



(b)

Figure 3.3: Residuals' plots for the analysis with cLDA for the variable ABAS-II functional academic score



(a) (b)

Figure 3.4: Residuals' plots for the analysis with change scores for the variable ABAS-II functional academic score

In the plots of fitted values against residuals, although it is possible to observe observations with higher values, the majority belongs to a dispersion of points around zero, without forming a specific pattern, regardless of the model used to fit the data. From the Q-Q plot, specially when using ANCOVA, the points follow a straight line indicating that they follow a normal distribution, and even using cLDA, the normality assumption of the error term seems to be reasonable.

3.3 Longitudinal Analysis

For the longitudinal analysis, the strategy used was the same as used by the authors of the published work about the study. Time was treated as a categorical variable since to use it as continuous it would be necessary to assume that the relationship between time and outcome variable is linear. Four time points were included in the analysis: $t = 0$, $t = 3$, $t = 6$ and $t = 12$.

3.3.1 Model building

Considering all the notation previously presented, in the context of this study, where four time points are included, the expression for the vector of responses for the individual i, Y_i is given by:

ANCOVA:

$$Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \beta_4 W_i + \varepsilon_i \tag{3.5}$$

cLDA

$$Y_i = \beta_0 + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{3i} + \beta_4 T_i^* W_i + \varepsilon_i \tag{3.6}$$

Under this model, β_4 is considered to be the treatment effect for all time points except from baseline since, when there is no interaction between the effect of the treatment and the time,

$$E(Y|X = 1, \text{Time } 1) - E(Y|X = 0, \text{Time } 1)$$

is assumed to be the same as

$$E(Y|X = 1, \text{Time } 2) - E(Y|X = 1, \text{Time } 2)$$

and

$$E(Y|X = 1, \text{Time } 3) - E(Y|X = 1, \text{Time } 3)$$

In both cases, if there is evidence of the existence of interaction between variables in the analysis, this interaction must be also included.

3.3.2 Model fitting

This analysis were also done with the software R, using the function lme for analysis of LMM
Different adaptations of linear mixed models had to be implemented in order to compute ANCOVA and cLDA. As an example, the code to generate the models for the variable ABAS-II functional academic score is presented bellow when using both ANCOVA and cLDA.

ANCOVA

```
anc <- lme(academic ~ base + month + treat + sex, random = ~ 1 |id,
           data = long, na.action= na.omit)
```

cLDA

```
long$t1 <- as.integer(long$month == 3)
long$t2 <- as.integer(long$month == 6)
long$t3 <- as.integer(long$month == 12)
#defining binary variables that are 1 if the measuremnet
#corresponds to that time point and 0 if not
long$t123 <- ifelse(long$month %in% c(3, 6, 12) & long$treat == "B", 1, 0)
#defining the treatment effect that is the same for all timepoints
clda<- lme(academic ~  t1 + t2 + t3 + t123 + sex, random= ~ 1 | id,
           na.action = na.omit, data = long)
```

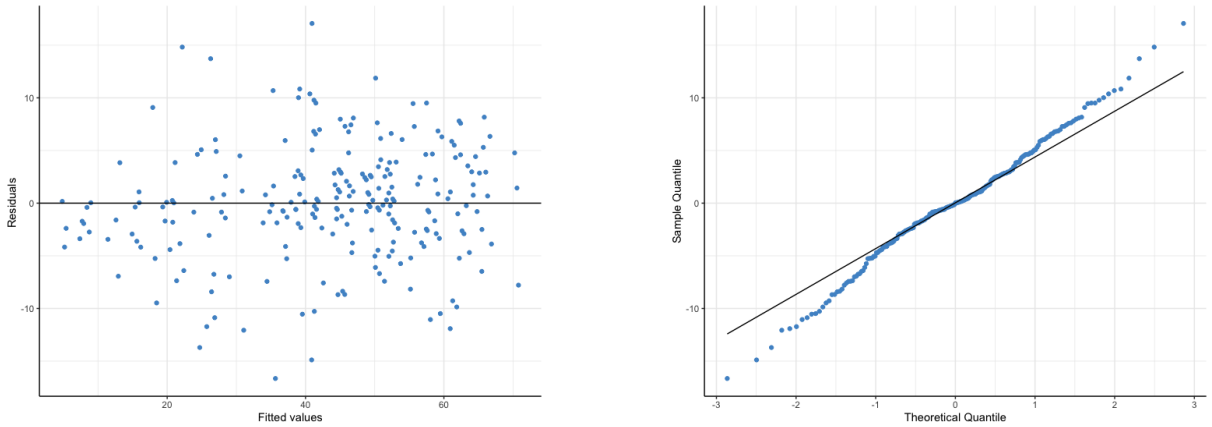
3.3.3 Results

The results obtained for the previously mentioned variables for the longitudinal analysis are represented in Table 3.8.

Table 3.8: Longitudinal analysis' results obtained of TESDAD study variables

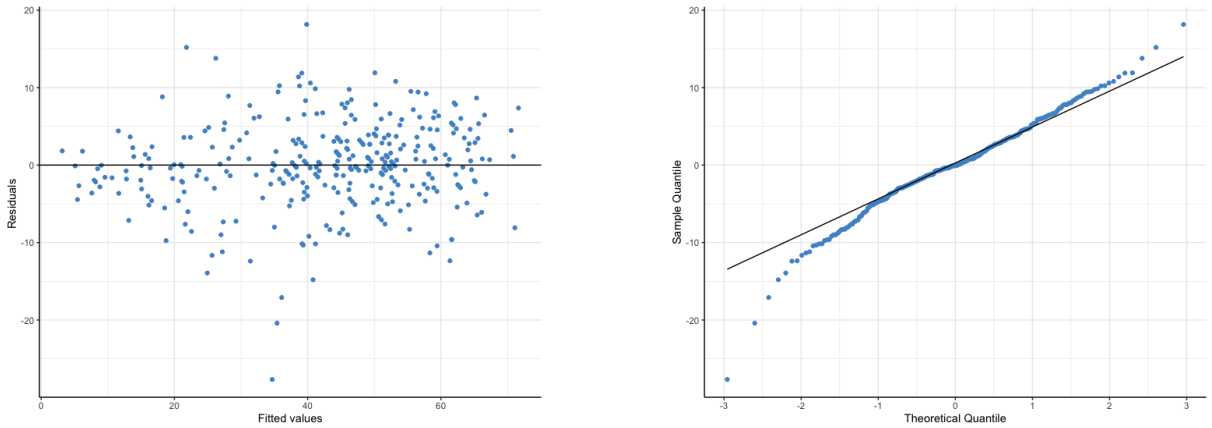
	Estimated difference between treatments	**Standard Deviation**	**p-value**
Cats and Dogs Test score			
ANCOVA	0.475	0.229	0.041
cLDA	0.427	0.184	0.018
Cats and Dogs test time			
ANCOVA	-4.580	1.987	0.024
cLDA	-4.022	1.533	0.009
ABAS-II functional academic score			
ANCOVA	5.492	1.692	0.002
cLDA	5.622	1.578	0.0004
Homocysteine levels			
ANCOVA	0.697	0.281	0.015
cLDA	0.706	0.290	0.016

For a significance level of 0.05, the conclusions taken are the same when applying either cLDA or ANCOVA. The estimations obtained by applying cLDA are slightly smaller, but the values for the estimation of treatment effect are similar. For the same reasons mentioned for the previous analysis, residuals plots were also done and in figures 3.5 and 3.6 are represented the ones for the variable ABAS-II functional academic score.



(a) (b)

Figure 3.5: Residuals' plots for the analysis with ANCOVA for the variable ABAS-II functional academic score



(a) (b)

Figure 3.6: Residuals' plots for the analysis with cLDA for the variable ABAS-II functional academic score

From the analysis of graphs 3.5a and 3.6a, representing the residuals against the fitted values, it is possible to denote that the residuals from both models do not follow a specific pattern and the variation is roughly the same all the way across the points which leads to their constant dispersion along the horizontal line centred at zero. However, in both cases several residuals present very high values. These can be outliers or influential observations that could have affected the fit of the model.

In Figures 3.5b and 3.6b the plots of the empirical quantiles of the residuals against the theoretic quantiles of the standard normal distribution are represented. At both cases, normality of the error terms appears

to be a fairly safe assumption, since the points seem to follow a straight line.

Chapter 4

Simulation study

The simulation study presented in this chapter was conducted in order to compare two different methods to deal with baseline measurements in longitudinal studies, ANCOVA and cLDA. This study aims to evaluate the behaviour of these two statistical approaches under different settings. Although settings with pre-post data were included, change-scores analysis was not applied since the simulated data did not meet the requirements to use it.

The data was generated by producing parametric draws from a multivariate normal distribution, initializing the parameters with values from real data. In this case, it was important to create realistic scenarios, since the main point of studying this models is to assess how they can be applied in real life cases. To do that, it is important that the parameters are realistic, so values from the previously presented study were used to initialize the parameters in the simulation. These values were the ones obtained when applying cLDA to the variable ABAS-II functional academic scores. The choice of the variable used was random. From there, a data set with two time points in the case of the pre-post analysis and four time points in the case of the longitudinal analysis was generates.

In order to test specific conditions, some parameters were varied related to sample size, normality assumptions and missing values.

As mentioned in Section 2.5 of Chapter 2, ANCOVA assumes that conditional on baseline measurements, the post-baseline measurements follow a jointly multivariate normal distribution. In contrast, cLDA assumes that all the measurements taken, baseline included, are jointly multivariate normally distributed. Due to this condition, we can verify that the assumptions about the normality of baseline are stronger for cLDA than for ANCOVA. In some trials, baseline measurements can be used as exclusion criteria, which makes the distribution of these variables skewed, which can lead to a violation of the normality assumption. To test the impact of this situation in the data modelling, besides from the scenario where baseline follows a normal distribution, one where baseline follows a truncated normal distribution will also be introduced.

When the data is being analysed, if ANCOVA is applied, it is necessary to remove the individuals that do not have baseline values or that only have baseline values from the analysis. In cLDA, the same does not happen, and those individuals can be included. To investigate if this can turn into an advantage to cLDA, scenarios with and without missing data are also going to be tested. In Table **??** the percentages of missing data in each time point for the two created missing data conditions are presented.

	Pre-post analysis		**Longitudinal analysis**			
	t=0	t=18	t=0	t=3	t=6	t=12
m0	0	0	0	0	0	0
m1	∼2	∼15	∼2	∼3	∼8	∼15
m2	∼4	∼25	∼4	∼6	∼15	∼25

Finally, the described conditions are going to be tested to different sample sizes ($n_{obs} = 120$ and $n_{obs} = 70$) and different study designs (longitudinal and pre-post). A summary of the generated scenarios can be found in Figure 4.1.

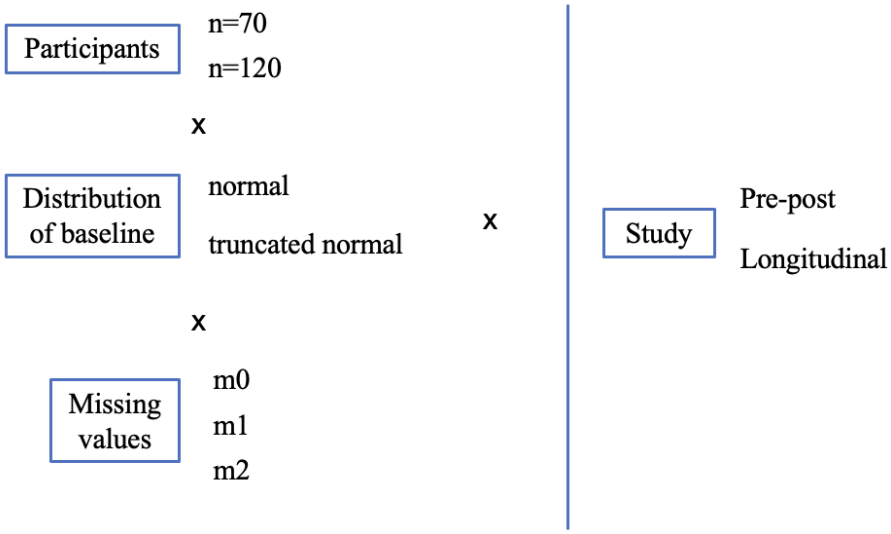


Figure 4.1: Different simulation settings generated

Under each scenario, a total of 5000 data sets were generated, so the one used for analysis is a result of the mean of all those sets. The number of simulations to be perform have to take into account that errors tend to be reduced when increasing the number of simulation replicates, however one also have to keep computing time within a reasonable limit (O'Connell et al. 2017). The resulting data sets with either four or two time points were analysed using both ANCOVA and cLDA, and the models' formulation was

the same as presented in Chapter 3.

This simulation study had the same target as the analysis performed with the TESDAD study,i.e., the estimation of treatment difference, denoted by θ. Dues to this, $\hat{\theta}$ was the value used for the evaluation of the models. In order to do that, the performance measures calculated were: bias, standard error, MSE and coverage. The basis on how these statistics work was previously explained in Section 1.3 of Chapter 2. These statistics were chosen since they are considered the most appropriate when evaluating the quality of an estimator.

All the process of generation and treatment of the data was done using the software R. Functions `mvnorm` and `tmvnsim` from package **tmvnsim** (Bhattacjarjee 2016), were used to generate, respectively, normal and normal with truncated baseline data. As previously stated, the arguments introduced in the functions were based on the models obtained from the TESDAD study for the variable ABAS-II functional academic score. As an example, the code for the several steps for the scenario with of the longitudinal data simulation with, $n_{obs} = 120$, the baseline following a normal distribution and no missing data is presented below. For the other scenarios, there are some differences and the code can be found in the Appendix.

```
#Data generation
gendata <- function(num_participants) {

  simulated_data <- data.frame("id" = as.character(1:num_participants))
  simulated_data$treatment <- sample(0:1, num_participants, replace = T)

  fixed_intercept <- 40.7143
  intercept_variance <- 1.9362
  cov <- c(319.4355, 263.2704, 244.0316, 251.8028,
           263.2704, 304.2826, 245.1052, 248.9508,
           244.0316, 245.1052, 282.2373, 253.3491,
           251.8028, 248.9508, 253.3491, 304.0427)
  cov <- matrix(cov, nrow = 4, ncol = 4, byrow = T)
  random_intercept <- rnorm(num_participants, mean = 0,
                            sd = sqrt(intercept_variance))
  simulated_data$intercept <- fixed_intercept + random_intercept
  mean_vector <- cbind(
    simulated_data$intercept,
    simulated_data$intercept +
```

```
    ifelse(simulated_data$treatment == 1, 5.5472, 0) + 1.10315,
    simulated_data$intercept +
    ifelse(simulated_data$treatment == 1, 5.5472, 0) + 0.54639,
    simulated_data$intercept +
    ifelse(simulated_data$treatment == 1, 5.5472, 0) + 0.76164
  )

  participant_data <- sapply(1:num_participants,
  function(p) mvrnorm(n = 1, mu = mean_vector[p,], Sigma = cov))
  participant_data <- (participant_data)
  simulated_data$Time_1 = participant_data[1,]
  simulated_data$Time_2 = participant_data[2,]
  simulated_data$Time_3 = participant_data[3,]
  simulated_data$Time_4 = participant_data[4,]
  simulated_data$intercept <- NULL
  return(simulated_data)

#Simulation retepitions
set.seed(20)
list <- replicate(5000, gendata(num_participants = 120 ), simplify = FALSE)

#Calculation of the performance measurements

fixed_coef <- vector()
confint <- list()
for (i in 1:length(long_list)){
  long_list[[i]]$t2 <- as.integer(long_list[[i]]$time == "Time_2")
  long_list[[i]]$t3 <- as.integer(long_list[[i]]$time == "Time_3")
  long_list[[i]]$t4 <- as.integer(long_list[[i]]$time == "Time_4")

  long_list[[i]]$t123 <- ifelse(long_list[[i]]$time %in%
    c("Time_2", "Time_3", "Time_4") & long_list[[i]]$treat==1, 1, 0)
```

```
  clda<- lme(score ~ t2 + t3 + t4 + t123 , random= ~ 1 | id,
  na.action = na.omit, data = long_list[[i]])
  fixed_coef1[[i]] <- coef(clda)$t123[1]
  confint[[i]] <- intervals(clda, level = 0.95)$fixed["t123", ]
}
mean(fixed_coef1)
sqrt(var(fixed_coef1))

coverage <- 0
true_value <- 5.5472 # Replace with the actual true parameter value
bias <- true_value-mean(fixed_coef1); bias

for (i in 1:length(confint1)) {
  if (confint1[[i]][1] <= true_value && true_value <= confint1[[i]][3]) {
    coverage <- coverage + 1
  }
}
coverage_rate <- coverage / length(confint1)
coverage_rate
}
```

The results obtained for the estimated performance of the methods are represented in Tables 4.1 and 4.2.

Table 4.1: Values of the performance measurements obtained from the pre-post simulation study

			Ancova				**cLDA**			
			Bias	Standard deviation	MSE	Coverage	Bias	Standard Deviation	MSE	Coverage
Normal	120	m0	0.0126	1.9943	3,9774	0.947	0.0123	1.9921	3.9686	0.951
		m1	-0.0115	2.1175	4.4839	0.9516	-0.0046	2.1022	4.4193	0.9558
		m2	-0.0108	2.2667	5.1380	0.949	0.0095	2.2402	5.0186	0.9514
	70	m0	0.0305	2.5735	6.6238	0.9494	0.0214	2.5634	6.5715	0.9504
		m1	0.0193	2.11600	4.4779	0.9506	0.0115	2.1035	4.4248	0.955
		m2	-0.0010	2.2723	5.1633	0.9506	0.0061	2.2437	5.0342	0.953
Truncated normal	120	m0	-0.0468	1.9588	3.8391	0.9514	-0.0504	1.9640	3.8598	0.9384
		m1	-0.0512	2.1081	4.4467	0.954	-0.0532	2.0975	4.4023	0.9434
		m2	-0.0386	2.2929	5.2589	0.9542	-0.0438	2.2569	5.0955	0.9434
	70	m0	-0.0155	2.5853	6.6840	0.9494	-0.0504	2.5887	6.7039	0.936
		m1	0.0114	2.1475	4.6119	0.9488	0.0103	2.1301	4.5374	0.9364
		m2	-0.0147	2.3147	5.3581	0.9518	-0.0137	2.2769	5.1845	0.9408

Analysing, Table 4.1, where the results of the simulation study with pre-post data are presented, it confirms some of the features we saw on the analysis of TESDAD data regarding the treatment difference estimations and its respective estimated standard deviation. The estimators of treatment difference of both models are unbiased in all scenarios. The standard deviations produced by both models are also similar. However, applying ANCOVA, coverage provided by 95% confidence intervals for treatment differences is adequate, and the same does not happen when applying cLDA where the coverage is slightly below 95% when baseline follows a truncated normal distribution. The models also had almost identical performances in the several missing data scenarios.

Table 4.2: Values of the performance measurements obtained from the longitudinal simulation study

			Ancova				**cLDA**			
			Bias	Standard deviation	MSE	Coverage	Bias	Standard Deviation	MSE	Coverage
Normal	120	m0	-0.0046	1.4829	2.1990	0.9538	-0.0084	1.4864	2.2095	0.9434
		m1	-0.0002	1.5098	2.2795	0.9504	-0.0027	1.5047	2.2641	0.9404
		m2	-0.0134	1.4949	2.2349	0.953	-0.0155	1.4914	2.2245	0.9404
	70	m0	-0.0424	1.9718	3.8898	0.946	-0.0382	1.9649	3.8623	0.938
		m1	0.0005	1.96197	3.8493	0.9518	-0.0004	1.9561	3.8263	0.9434
		m2	0.02493	2.01506	4.0611	0.9508	0.0323	2.0014	4.0066	0.9396
Truncated normal	120	m0	-0.0180	1.4754	2.1771	0.954	-0.0211	1.4782	2.1855	0.924
		m1	-0.0074	1.5270	2.3318	0.9414	-0.0110	1.5255	2.3273	0.9126
		m2	-0.0168	1.5214	2.3149	0.9518	-0.0209	1.5131	2.2899	0.9126
	70	m0	-0.03822	1.9315	3.7322	0.9522	-0.0401	1.9380	3.7575	0.9164
		m1	-0.0143	2.0177	4.0713	0.9472	-0.0157	2.0073	4.0295	0.9114
		m2	-0.0308	2.0154	4.0628	0.9514	-0.03172	1.9954	3.9826	0.9158

From Table 4.2, where the results of the simulation of longitudinal data are displayed, similar features are observed. The bias obtained using both models are close to zero and when baseline follows a normal distribution the results given by both ANCOVA and cLDA are close. However, when the distribution of baseline is not normal, which commonly happens in real life studies, the same does not happen. ANCOVA is robust against deviation from normality assumptions of baseline values, but when analysing the results obtained when applying cLDA, the coverage provided by cLDA is not proper in the presence of a truncated baseline.

Chapter 5

Discussion

This study was conducted with the main objective of comparing the performance of ANCOVA and cLDA when analysing longitudinal data with different characteristics. The existent literature around the topic is not consistent, which makes the decision of what approach should be used in each specific case difficult. In order to achieve this aim, the models were applied to real data from a clinical trial and then to simulated data. In both cases, first the models were applied to a simple setting, the pre-post case, and then to a setting with more time points, the longitudinal one.

Starting with the analysis of the results from the analysis of TESDAD data, in the pre-post case, apart from cLDA and ANCOVA, change-score analysis was also performed since it can also be an adequate model to apply in those conditions. However, one of its assumptions is that the correlations between change scores and baseline measurements should be high (≥ 0.8), and for the variables in analysis this was not the case. The results obtained with the change-score analysis are in conformity with the rest of the methods, so probably regression through the mean effect was not accentuated. However, since the correlations are low, the estimations obtained with this model for treatment effect can be biased and no conclusions should be taken from that analysis. The results of the differences detected between treatments were consistent between the three methods applied, when looking to the significance of detected treatment effect. For the variables ABAS-II functional academic score and homocysteine levels the estimations are very similar, probably because the percentage of missing data in these variables is very low. This goes in conformity with what is described in the literature. The lower the percentage of missing data, the closer the estimations of the parameter provided by ANCOVA and cLDA should be.

Regarding the longitudinal analysis, the results obtained are in conformity with the theoretical results obtained by Liu, Lu, et al. (2009) as well as the practical ones presented on the same work and on the one developed by Lu (2010) when cLDA and ANCOVA were applied to longitudinal data from a clinical trial. The treatment effect estimations are similar when using both methods and the estimated standard deviation for treatment effect is smaller when applying cLDA. Although there are small differences in the

estimations, assuming a significance level of 0.05, the conclusions about the existence of a treatment effect are the same regardless of the method used. The variables where the biggest differences are present are the Cats-and-Dogs test score and Cats-and-Dogs test time. These are also the variables with the highest percentage of missing data, and in the literature cLDA has been pointed as the optimal choice for providing the most precise estimations of treatment effect when missing data occur (Coffman et al. 2016), so this can be an explanation for the obtained results. There it is denoted that for the variance associated to treatment effect estimation, the one produced by ANCOVA is higher than the one produced by cLDA. Concerning the analysis of the TESDAD variables, except from the homocysteine levels, the remaining ones presented smaller standard deviations for the treatment effect estimation when using cLDA. This variable is the one where the percentage of missing values is the smallest. According to the literature (Liu, Lu, et al. 2009), variables with small percentage of missing values are likely to have similar results for the estimations when using both cLDA and ANCOVA, which may be what happens here.

Analysing the quantile plots, they do not raise any important concern with normality of residuals. The obtained Q-Q plots are similar for both models, so it does not have an impact. About homoscedasticity, the plots present a good distribution of the residuals, showing that both models are homoscedastic.

As previously mentioned, for both pre-post and longitudinal analysis, using different methods, small differences of the estimates for treatment effect were obtained. In order to know if they are relevant in the clinical context, the evaluation of these differences would have to be done by a professional of the area.

Regarding the simulation studies, on a pre-post analysis scenario, the results are in conformity with what is described in the literature. When the baseline distribution is normal, the estimations are unbiased using both models. When there is no missing data, the estimations provided are similar, and the standard deviation calculated from cLDA slightly smaller. In the presence of missing values, both bias and standard deviation provided by cLDA are smaller since, using this model, it is also possible to include individuals that are missing either baseline or post baseline measurement, while with ANCOVA this does not happen. The confidence intervals of the mean changes properly cover at a 95% level the true parameter. What was mentioned is applied for the scenarios with $n = 120$ and $n = 70$, although in the second case, standard deviations are always bigger. In the presence of a baseline following a truncated normal distribution, we observe that for all the parameters, better results are obtained when using ANCOVA. With this model, in addition to an appropriate coverage of the confidence intervals, the biases and standard deviations were smaller, while with cLDA this did not happen. This can be because, the assumptions about normality of baseline are stricter in the case of cLDA than ANCOVA, resulting in a worst behaviour of the model. Even in scenarios with missing data, ANCOVA has a better behaviour, showing that for the applied percentage of missing data and severity of truncation, this second condition has a bigger impact. In these scenarios, the coverage provided by cLDA is not adequate. In the work of Woolson et al. (Coffman et al. 2016), who studied the best methodology to apply when dealing with pre-post data, similar conclusions were reached to the ones from the normal distributed baseline scenario. In that study, cLDA had better

results at the majority of the cases, so it was selected as the best approach. However, in that study, scenarios where data did not follow a normal distribution were not considered and here it is possible to see that ANCOVA has a better result when dealing with deviations from normality.

Regarding the longitudinal analysis scenarios, the obtained results are similar to the ones of the pre-post scenario. ANCOVA has an adequate behaviour in the cases that baseline follows a normal and a truncated normal distribution, showing that under these conditions, the use of this method to model the data is appropriate. However, a difference between this and the previous analysis is that in a scenario where the baseline follows a normal distribution, there is no advantage in applying cLDA, even in the presence of missing data. The estimations obtained for the evaluation parameter are similar for both methods, but coverage of 95% confidence intervals is smaller for cLDA and slightly under 95%, which makes this second one less adequate. Finally, in the presence of a baseline deviated from normal distribution, the behaviour in terms of coverage is again not adequate using that method. These results are not in conformity with what is described in the literature (Liu, Lu, et al. 2009; Lu 2010). These authors describe that when analysing longitudinal data, both models have a good behaviour, even against deviations from normality and in the presence of missing data, cLDA, gives smaller standard deviations for the estimations, so it would be the selected method to use.

According to the analysis performed in this thesis with the experimental settings analysed, we can conclude that, if the data follows a normal distribution, both application of cLDA and ANCOVA are adequate, and the estimations provided by cLDA are more efficient. The application of this last method can bring even more advantages in the cases where missing data is present. However, one has to be careful when applying it since if data deviates from the normality, since although analysing the estimations, nothing unusual is noticed, when looking at the coverage, it is not adequate.

It should be noted that although several data sets were analysed, the amount of possibilities that can be explored is huge and that is definitely a limitation of this study. It was showed that the most adequate method to apply can depend on data characteristics, so it would be of interest to cover more possibilities in order to figure out the exact source of the differences detected.

Chapter 6

Conclusion

Neither application of cLDA or ANCOVA are a direct answer, since it is necessary to understand and analyse some characteristics of the data one is working with.

From what was obtained in this study, using both clinical trial and simulated data, the results indicate that in cases where the data follows a normal distribution, the use of cLDA can bring advantages, specially in the presence of missing data. Both models produce similar results, and analysing residual plots both seem to be adequate for that data. Due to this, and taking into account that the estimations provided for cLDA are more efficient, this seems to be the better approach. However, when the distribution followed by the variable deviated from the normal, the use of cLDA may not be a suitable approach. Although it continues producing smaller standard deviations when comparing with results produced by ANCOVA, the simulations indicate that the coverage of 95% confidence intervals is not appropriate.

Taking that into consideration, when deciding which model should be used, we recommend to explore first whether the model assumptions hold specially regarding normality. If a deviation from normality in the data is found, it may be adequate to use ANCOVA, since results provided by cLDA have been showed to be worst. If in the same analysis, if a high percentage of missing data is found, the use of cLDA should be considered.

It is also important to denote that in this case, the use of ANCOVA analysing longitudinal data with base-line deviated from normality was clearly better in terms of coverage, something that was not described in the literature. This could motivate a larger simulation study to prove whether the results hold under other settings.

Bibliography

Akobeng, Al K (2005). "Understanding randomised controlled trials". In: *Archives of Disease in Childhood* 90.8, pp. 840–844.

Albert, Paul S (1999). "Longitudinal data analysis (repeated measures) in clinical trials". In: *Statistics in Medicine* 18.13, pp. 1707–1732.

Bain, Jenny et al. (2003). "The specificities of protein kinase inhibitors: an update". In: *Biochemical Journal* 371.1, pp. 199–204.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bhattacjarjee, Samsiddhi (2016). *tmvnsim: Truncated Multivariate Normal Simulation*. R package version 1.0-2. URL: https://CRAN.R-project.org/package=tmvnsim.

Caruana, Edward Joseph et al. (2015). "Longitudinal studies". In: *Journal of Thoracic Disease* 7.11, E537.

Coffman, Cynthia J et al. (2016). "To condition or not condition? Analysing 'change'in longitudinal randomised controlled trials". In: *BMJ open* 6.12, e013096.

Cohen, Jacob (1988). "The effect size". In: *Statistical power analysis for the behavioral sciences*, pp. 77–83.

De la Torre, Rafael, Susana de Sola, et al. (2016). "Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD): a double-blind, randomised, placebo-controlled, phase 2 trial". In: *The Lancet Neurology* 15.8, pp. 801–810.

De la Torre Rafael De Sola, Susana, Meritxell Pons, et al. (2014). "Epigallocatechin-3-gallate, a DYRK1A inhibitor, rescues cognitive deficits in D own syndrome mouse models and in humans". In: *Molecular Nutrition & Food Research* 58.2, pp. 278–288.

Dierssen, Mara (2012). "Down syndrome: the brain in trisomic mode". In: *Nature Reviews Neuroscience* 13.12, pp. 844–858.

Dierssen, Mara et al. (2009). "Aneuploidy: from a physiological mechanism of variance to Down syndrome". In: *Physiological Reviews*.

Dinh, Phillip and Peiling Yang (2011). "Handling baselines in repeated measures analyses with missing data at random". In: *Journal of Biopharmaceutical Statistics* 21.2, pp. 326–341.

Draper, Norman R and Harry Smith (1998). *Applied Regression Analysis*. Vol. 326. John Wiley & Sons.

Fisher, Ronald Aylmer (1970). “Statistical methods for research workers”. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, pp. 66–70.

Galton, Francis (1886). “Regression towards mediocrity in hereditary stature.” In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15, pp. 246–263.

Guedj, Fayçal et al. (2009). “Green tea polyphenols rescue of brain defects induced by overexpression of DYRK1A”. In: *PloS One* 4.2, e4606.

Harville, David A. (1977). “Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems”. In: *Journal of the American Statistical Association* 72.358, pp. 320–338. DOI: 10.1080/01621459.1977.10480998.

Kahlert, Johnny et al. (2017). “Control of confounding in the analysis phase–an overview for clinicians”. In: *Clinical epidemiology*, pp. 195–204.

Kang, Hyun (2013). “The prevention and handling of the missing data”. In: *Korean Journal of Anesthesiology* 64.5, pp. 402–406.

Laird, Nan M and James H Ware (1982). “Random-effects models for longitudinal data”. In: *Biometrics*, pp. 963–974.

Liang, Kung-Yee and Scott L Zeger (1993). “Regression analysis for correlated data”. In: *Annual Review of Public Health* 14.1, pp. 43–68.

Liang, Kung-Yee and Scott L Zeger (2000). “Longitudinal data analysis of continuous and discrete responses for pre-post designs”. In: *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 134–148.

Ligthelm, Robert J et al. (2007). “Importance of observational studies in clinical practice”. In: *Clinical Therapeutics* 29.6, pp. 1284–1292.

Liu, Guanghan F, Kaifeng Lu, et al. (2009). “Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials?” In: *Statistics in Medicine* 28.20, pp. 2509–2530.

Liu, Siwei, Michael J Rovine, et al. (2012). “Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches.” In: *Psychological Methods* 17.1, p. 15.

Lu, Kaifeng (2010). “On efficiency of constrained longitudinal data analysis versus longitudinal analysis of covariance”. In: *Biometrics* 66.3, pp. 891–896.

Mallinckrodt, Craig and Ilya Lipkovich (2016). *Analyzing longitudinal clinical trial data: A practical guide*. CRC Press.

Morris, Tim P et al. (2019). “Using simulation studies to evaluate statistical methods”. In: *Statistics in Medicine* 38.11, pp. 2074–2102.

Noll, Christophe et al. (2009). “DYRK1A, a novel determinant of the methionine-homocysteine cycle in different mouse models overexpressing this Down-syndrome-associated kinase”. In: *PLoS One* 4.10, e7540.

O’Connell, Nathaniel S et al. (2017). “Methods for analysis of pre-post data in clinical research: a comparison of five common methods”. In: *Journal of Biometrics & Biostatistics* 8.1, p. 1.

Pinheiro, José et al. (2023). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-163. URL: https://CRAN.R-project.org/package=nlme.

Rust, James O and Monica A Wallace (2004). "Book review: Adaptive behavior assessment system". In: *Journal of Psychoeducational Assessment* 22.4, pp. 367–373.

Schlesselman, James J (1982). *Case-control studies: design, conduct, analysis*. Vol. 2. Oxford University Press.

Senn, Stephen (2014). "Baseline adjustment in longitudinal studies". In: *Wiley StatsRef: Statistics Reference Online*.

Sims, Jennifer and Vickie A Miracle (2002). "Phases of a clinical trial". In: *Dimensions of Critical Care Nursing* 21.4, pp. 152–153.

Singh, Jatinder (2015). "International conference on harmonization of technical requirements for registration of pharmaceuticals for human use". In: *Journal of Pharmacology and Pharmacotherapeutics* 6.3, pp. 185–187.

Verbeke, Geert (1997). "Linear mixed models for longitudinal data". In: *Linear mixed models in practice*. Springer, pp. 63–153.

Vickers, Andrew J and Douglas G Altman (2001). "Analysing controlled trials with baseline and follow up measurements". In: *BMJ* 323.7321, pp. 1123–1124.

Weil, Rimona S et al. (2017). "The cats-and-dogs test: a tool to identify visuoperceptual deficits in Parkinson's disease". In: *Movement Disorders* 32.12, pp. 1789–1790.

Zeger, Scott L and Kung-Yee Liang (1992). "An overview of methods for the analysis of longitudinal data". In: *Statistics in Medicine* 11.14-15, pp. 1825–1839.

Appendix A

Appendix

A.1 R code

A.1.1 R packages

```
library(dplyr)
library(nlme)
library(data.table)
library(MASS)
library(ggplot2)
library(mvtnorm)
library(tmvnsm)
```

A.1.2 Data preparation

The data from the TESDAD study was in a long format where the different measures from the same individual are represented in different rows. For an example of two individuals, that would be:

id	month	treat	sex	catscore	cattime	academ	hcy
1	0	A	Male	15	26	34	5.3
1	3	A	Male	15	27	34	3.7
1	6	A	Male	16	19	37	4.4
1	12	A	Male	15	21	35	3.4
1	18	A	Male	15	28	34	3.5
4	0	B	Female	14	36	37	9.3
4	3	B	Female	16	32	18	8.2
4	6	B	Female	13	47	33	6.4
4	12	B	Female	13	29	20	9.5
4	18	B	Female	13	45	32	5.3

To perform the pre-post analysis it was necessary to create a two different data sets. the firstr with the time-points $t = 0$ and $t = 18$ and the second with the time points $t = 0$, $t = 3$, $t = 6$ and $t = 12$.

```
    pre<- subset(data_long, month %in% c(0,18))
    long<- subset(data_long, month != 18)
```

For cLDA all the time points have to be in a long format, however for ANCOVA it is necessary to create a variable with the fist time point. For that, it is necessary to put that in a wide format, where all the observations of one individual are in one row. Only after that, put in long format the observations of the remaining time points.

```
data_wide <- data.table::dcast(long_selected , id ~ month ,
    value.var = c("treat", "academ", "sex", "hcy", "cat_score" ,
                  "cat_time"))
    long <- melt(wide_data, id.vars = c("id" , "treat" , "sex" , "academ0" ,
                 "hcy0", "cat_score0", "cat_time0"),
    measure.vars = c(c("academ3" , "academ6" , "academ12"),
                     c("hcy3" , "hcy6" , "hcy12"),
                     c("cat_score3" , "cat_score6" , "cat_score12"))
```

Finally, for cLDA it is necessary to create new binary variables indicating if each observation belongs or not to a category.

```
long$t1 <- as.integer(long$month == 3)
# 1 if the observation was measured at $t=3$ and 0 if not

long$t2 <- as.integer(long$month == 6)
```

```
# 1 if the observation was measured at $t=6$ and 0 if not

long$t3 <- as.integer(long$month == 12)
# 1 if the observation was measured at $t=12$ and 0 if not

long$t123 <- ifelse(long$month %in% c(3, 6, 12) & long$treat == "B", 1, 0)
# 1 if the observation was measured at $t=3$, $t=6$ or $t=12$ and from
# a patient recieving treatment B
```

A.1.3 Model fitting for TESDAD data

The process of how the data was modelled for the variable ABAS-II functional academic score is described in Chapter 3. For the remaining variables, presented above the process is the same, so it's not going to be presented here.

A.1.4 Simulation Data

In Chapter 4, it was presented the cod for the generation of longitudinal data, with baseline following a normal distribution and no missing values. It is going to be showed here the changes that needed to be done to simulate data with missing values, with baseline following a truncated normal distribution for a pre-post study.

```
#m1 and m0 correspod to the proportion of misisng data
#in $t=0$ and $t=18$ respectivelly

gendata <- function(num_participants = 120, m1 = 0.02, m2 = 0.15) {

  simulated_data <- data.frame("id" = as.character(1:num_participants))
  simulated_data$treatment <- sample(0:1, num_participants, replace = TRUE)

  fixed_intercept <- 36.6926
  intercept_variance <- 2.6023
  cov <- c(319.4355, 245.7842,
           245.7842, 301.7241)
  cov <- matrix(cov , nrow = 2 , ncol = 2 , byrow = TRUE)
  random_intercept <- rnorm(num_participants , mean = 0,
```

```
                            sd = sqrt(intercept_variance))
  simulated_data$intercept <- fixed_intercept + random_intercept
  mean_vector <- cbind(simulated_data$intercept ,
                       simulated_data$intercept +
                       ifelse(simulated_data$treatment == 1 ,3.3228 , 0) +
                       1.0199)

  participant_data <- sapply(1:num_participants, function(p)
                             tmvnsim(1, 2, lower = c(20, -Inf, -Inf, -Inf),
                             upper = c(Inf, Inf),
                             means = mean_vector[p, ], sigma = cov)$samp)
  simulated_data$Time_1 = participant_data[1,]
  simulated_data$Time_2 = participant_data[2,]
  missing_indices <- runif(nrow(simulated_data)) < m1
  simulated_data$Time_1[missing_indices] <- NA
  missing_indices <- runif(nrow(simulated_data)) < m2
  simulated_data$Time_2[missing_indices & simulated_data$Time_1 > 16] <- NA
  simulated_data$intercept <- NULL
  return(simulated_data)
}
```

Once obtained the data set, the steps to take are the same as previously mentioned in Chapter 4.