Universitat Oberta de Catalunya

uoc.edu

Modelització de dades longitudinals amb efectes aleatoris de intercepció/pendent i presència de dades perdudes en el context dels estudis d'estabilitat

Estudiant: Francesc Bernad Martin
M0.178 TFM-Estadística i Bioinformàtica 2
Màster universitari en Bioinformàtica i bioestadística UOC-UB
Àrea: Bioestadística / Bioinformàtica

Nom Consultor/a: Nuria Perez Alvarez
Nom Professor/a responsable de l'assignatura: Carles Ventura Royo

Barcelona, 04 de Juny de 2019

FITXA DEL TREBALL FINAL

Títol del treball:	*Modelització de dades longitudinals amb efectes aleatoris de intercepció/pendent i presència de dades perdudes en el context dels estudis d'estabilitat*
Nom de l'autor:	*Francesc Bernad Martin*
Nom del consultor/a:	*Nuria Perez Alvarez*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*06/2019*
Titulació o programa:	*Màster universitari en Bioinformàtica i bioestadística UOC-UB*
Àrea del Treball Final:	*M0.178 TFM-Estadística i Bioinformàtica 2*
Idioma del treball:	*Català*
Paraules clau	*stability studies, mixed models, missing values*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

La finalitat d'aquest treball és analitzar l'aplicació dels models mixtos i els models amb dades perdudes dins el context de l'anàlisi estadístic en els estudis d'estabilitat de productes biològics o similars. La finalitat és valorar l'aplicabilitat d'aquests dos punts per explicar i predir els estudis d'estabilitat.

La metodologia aplicada ha estat la cerca de informació bibliogràfica per poder realitzar un anàlisi teòric tenint en compte les condicions específiques d'un estudi d'estabilitat tipus i amb aquesta informació poder generar un model de simulació d'aplicació de models mixtos i un model de simulació per avaluar la presència de dades perdudes i alguns mètodes d'imputació de dades.

Els resultats de les simulacions dels models mixtos mostren quins són els models més adequats a aplicar segons com es comporten les dades dels conjunts veient els avantatges d'aplicar els efectes aleatoris i la utilitat de l'aplicació de les funcions de modulació de la variància.

Els resultats de la simulació de models amb dades perdudes mostren els biaixos que es poden produir amb la presència de dades perdudes combinada amb l'aplicació de models mixtos i eficàcia o desviació de dos mètodes d'imputació simple.

El resultat pràctic demostra com es reprodueix l'anàlisi de models mixtos en un cas real i com, salvant les distàncies, es reprodueixen alguns dels resultats vistos en la simulació.

Abstract (in English, 250 words or less):

The purpose of this work is to analyze the application of mixed models and models with presence of missing data in the context of statistical analysis in studies of stability of biological products or similar. The purpose is to evaluate the applicability of these two points to explain and predict stability studies.

The applied methodology has been the search for bibliographic information in order to perform a theoretical analysis taking into account the specific conditions of a usual stability study and using this information, to be able to generate a simulation model for the application of mixed models and a model of simulation to evaluate the presence of missing data and some methods of data imputation.

The results of the mixed models simulations show the most suitable models to apply according to the data behavior, presenting the advantages of applying random effects and the usefulness of the application of variance modulation functions.

The results of the models with missing data simulations show the biases that can occur with the presence of missing data combined with the application of mixed models, and the efficiency or deviation of two simple imputation methods.

The practical exercice results shows the mixed models analysis reproducibility in a real case and how, differences aside, some of the results seen in the previous simulation are reproduced.

Índex

Llista de figures

1 Introducció

1.1 Context i justificació del Treball

El punt de partida d'aquest treball és l'anàlisi estadístic en el context d'estudis d'estabilitat de productes biològics o similars. Actualment aquest tipus d'estudis, tal com recomanen les guies d'organismes oficiales, es resolen amb models lineals simples i contrastos d'hipòtesi essent d'especial interès l'anàlisi ANCOVA.

En aquest TFM es vol tractar l'anàlisi estadístic en aquest tipus d'estudis, veient com s'engloben dins l'anàlisi estadístic de dades longitudinals i dins d'aquesta categoria tan genèrica, en els aspectes concrets dels models de efectes mixtos i dels mètodes per solucionar els problemes de les dades perdudes.

Mitjançant la realització d'un anàlisi teòric i pràctic s'espera obtenir una aproximació a la l'aplicabilitat d'aquests models més complexos als estudis d'estabilitat i explorar si és adequat la seva utilització i si comporta millores respecte als models més simples.

1.2 Objectius del Treball

Es descriuen els objectius general i específics del treball en la seva versió final després de l'evolució que ha anat prenent el treball:

1. Conceptes teòrics: Cerca d'informació per tenir coneixement dels principals aspectes dins el context de l'anàlisi a realitzar.
    1.1 Cerca d'informació sobre les recomanacions i guies per abordar els estudis d'estabilitat.
    1.2 Resum i referenciació dels conceptes relacionats amb la utilització dels models d'efectes mixtos.
    1.3 Resum i referenciació dels conceptes relacionats amb la utilització dels efectes de les dades perdudes en els models i les tècniques per donar possibles solucions.
    1.4 Generar un esquema de treball preliminar pel posterior anàlisi.
2. Anàlisis teòric: Utilitzant la informació recopilada entrar en el detall de com es pot aplicar cada punt en els models dels estudis d'estabilitat i fer aplicacions teòriques mitjançant simulacions.

2.1 Anàlisi teòric de les condicions a tenir en compte en el context dels estudis d'estabilitat per l'anàlisi estadístic.

2.2 Analitzar l'aplicabilitat dels models d'efectes mixtos en els estudis d'estabilitat mitjançant anàlisi teòric i simulacions de models.

2.3 Analitzar el comportament de les dades perdudes i possibles solucions en els models anteriors aplicats als estudis d'estabilitat mitjançant anàlisi teòric i simulacions de models.

2.4 Analitzar les diferents assumpcions del model d'efectes mixtos i dels mètodes de dades perdudes i l'efecte que tenen.

3 Anàlisis pràctic: Anàlisi d'un cas pràctic per la possible aplicació de l'anàlisi teòric realitzat.

1.3 Enfocament i mètode seguit

El mètode escollit pel primer objectiu ha estat el de fer una cerca bibliogràfica inicial amb les referències bibliogràfiques bàsiques de les assignatures cursades al màster. Partint d'aquests coneixements s'ha continuat per les cerques més específiques en articles mitjançant la biblioteca de la UOC i un buscador més general que en aquest cas s'ha escollit la plataforma *Mendeley* [1] per la seva versatilitat i compatibilitat amb altres eines informàtiques a l'hora de traspassar la bibliografia o d'organitzar aquesta.

Pel segon objectiu d'anàlisi teòric, l'estratègia seguida ha estat:

- En primer lloc escollir les condicions dels estudis d'estabilitat per utilitzar en la resta d'anàlisi teòric.
- En segon lloc abordar l'anàlisi de models mixtos on inicialment s'han pres les condicions que es volien utilitzar com l'estructura inicial i les possibles variants. Una vegada fet això s'ha treballat en la simulació on s'ha partit sempre de models més senzills avançant cap a models més complexos amb anàlisi intermitjos i definint els indicadors de qualitat a utilitzar per anar avaluant el model. En aquesta part s'ha donat certa bidireccionalitat ja que en l'avanç de la simulació s'ha anat millorant també la part de l'anàlisi teòric relacionat.
- En tercer lloc prenent com a referència els coneixements dels estudis d'estabilitats s'ha definit les condicions per l'anàlisi teòric dels models amb dades perdudes i s'ha aplicat a una simulació on s'ha partit de la simulació anterior per continuar avançant en l'aspecte de les dades perdudes. Com el cas anterior, en aquesta part també hi ha hagut certa bidireccionalitat ja que en l'avanç de la simulació s'ha anat millorant també la part de l'anàlisi teòric relacionada, encara que en aquest cas no ha estat tan rellevant com el cas dels models mixtos.

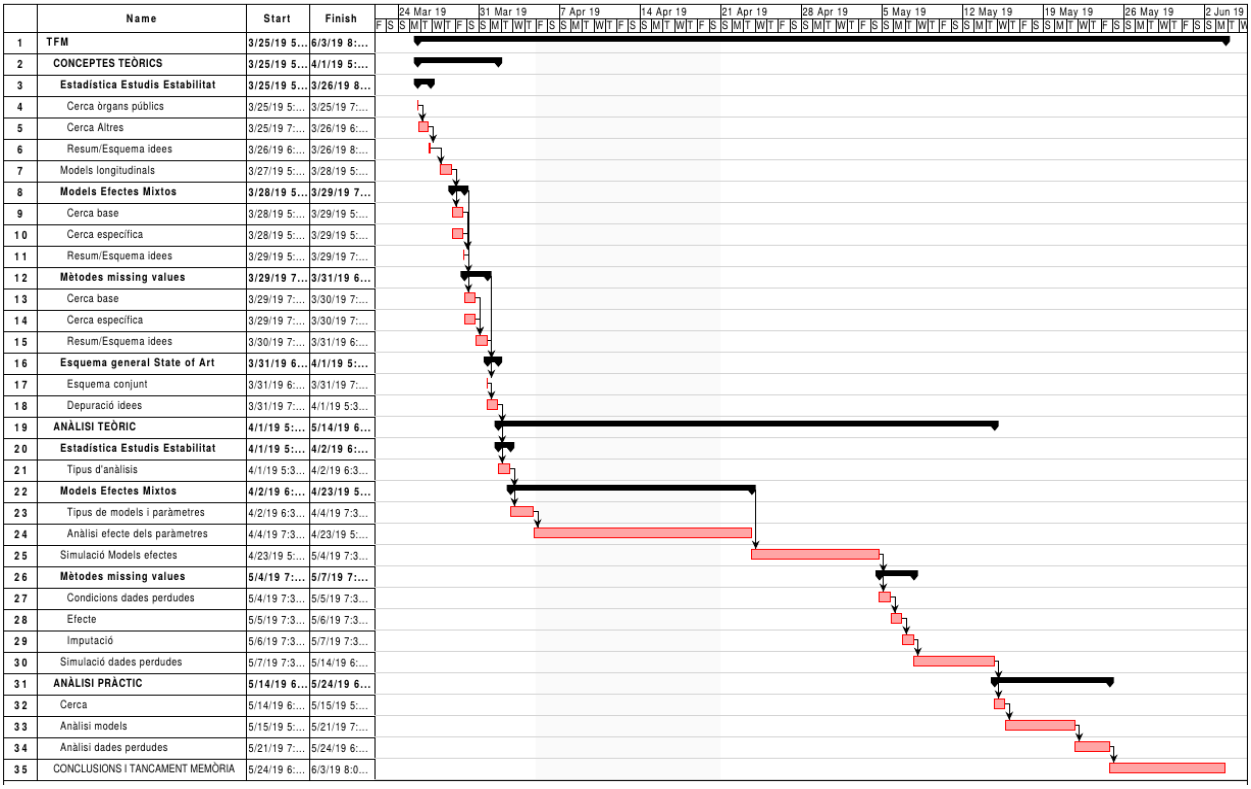
- Per la part de l'anàlisi pràctic s'ha partit des de zero per l'anàlisi inicial de les dades problema avançant sempre de models simples a models més complexos i al arribar als models practicats en les simulacions de models mixtos s'ha implementat el coneixement adquirit sempre que ha estat possible.

L'alternativa a aquesta metodologia de treball proposada podria ser fer-ho al revés de com s'ha exposat, és a dir, buscar directament les bases de dades d'estudis d'estabilitats de producte o equivalents i estudiar-los a nivell estadístic directament. A partir de l'anàlisi d'aquestes dades reals o reals adaptades (cas d'estudi en un context diferent, però adaptades per l'equivalència a un estudi d'estabilitat) es podria passar a l'anàlisi teòric tenint en compte els casos trobats i d'aquesta manera es tindria abans un coneixement previ de casos reals per veure quins casos teòrics tenen sentit i quins és poc probables que es donin.

En el meu cas no s'ha optat per aquest últim mètode ja que per l'àmbit professional de l'autor del treball ja es disposa de certa experiència en aquest context pel que s'ha decidit començar la part teòrica i després aplicar-ho a la part pràctica.

1.4 Planificació del Treball

A continuació es mostra el calendari de planificació creat mitjançant el programari obert *ProjectLibre* [2] en el qual es fa un esquema per tenir una noció dels temps que s'ha seguit:



	Name	Start	Finish
1	**TFM**	**3/25/19 5…**	**6/3/19 8:…**
2	**CONCEPTES TEÒRICS**	**3/25/19 5…**	**4/1/19 5:…**
3	**Estadística Estudis Estabilitat**	**3/25/19 5…**	**3/26/19 8…**
4	Cerca òrgans públics	3/25/19 5:…	3/25/19 7:…
5	Cerca Altres	3/25/19 7:…	3/26/19 6:…
6	Resum/Esquema idees	3/26/19 6:…	3/26/19 8:…
7	Models longitudinals	3/27/19 5:…	3/28/19 5:…
8	**Models Efectes Mixtos**	**3/28/19 5…**	**3/29/19 7…**
9	Cerca base	3/28/19 5:…	3/29/19 5:…
10	Cerca específica	3/28/19 5:…	3/29/19 5:…
11	Resum/Esquema idees	3/29/19 5:…	3/29/19 7:…
12	**Mètodes missing values**	**3/29/19 7…**	**3/31/19 6…**
13	Cerca base	3/29/19 7:…	3/30/19 7:…
14	Cerca específica	3/29/19 7:…	3/30/19 7:…
15	Resum/Esquema idees	3/30/19 7:…	3/31/19 6:…
16	**Esquema general State of Art**	**3/31/19 6…**	**4/1/19 5:…**
17	Esquema conjunt	3/31/19 6:…	3/31/19 7:…
18	Depuració idees	3/31/19 7:…	4/1/19 5:3…
19	**ANÀLISI TEÒRIC**	**4/1/19 5:…**	**5/14/19 6…**
20	**Estadística Estudis Estabilitat**	**4/1/19 5:…**	**4/2/19 6:…**
21	Tipus d'anàlisis	4/1/19 5:3…	4/2/19 6:3…
22	**Models Efectes Mixtos**	**4/2/19 6:…**	**4/23/19 5…**
23	Tipus de models i paràmetres	4/2/19 6:3…	4/4/19 7:3…
24	Anàlisi efecte dels paràmetres	4/4/19 7:3…	4/23/19 5:…
25	Simulació Models efectes	4/23/19 5:…	5/4/19 7:3…
26	**Mètodes missing values**	**5/4/19 7:…**	**5/7/19 7:…**
27	Condicions dades perdudes	5/4/19 7:3…	5/5/19 7:3…
28	Efecte	5/5/19 7:3…	5/6/19 7:3…
29	Imputació	5/6/19 7:3…	5/7/19 7:3…
30	Simulació dades perdudes	5/7/19 7:3…	5/14/19 6:…
31	**ANÀLISI PRÀCTIC**	**5/14/19 6…**	**5/24/19 6…**
32	Cerca	5/14/19 6:…	5/15/19 5:…
33	Anàlisi models	5/15/19 5:…	5/21/19 7:…
34	Anàlisi dades perdudes	5/21/19 7:…	5/24/19 6:…
35	CONCLUSIONS I TANCAMENT MEMÒRIA	5/24/19 6:…	6/3/19 8:0…

1.5 Breu sumari de productes obtinguts

Els productes principals obtinguts han estat la memòria principal del treball, la presentació resum sobre els continguts d'aquest i l'autoavaluació del treball, documents que serveixen per poder descobrir com s'ha desenvolupat aquest tema, quins resultats s'han obtingut i el grau de compliment general.

Addicionalment, les tres grans parts realitzades amb R Studio [3] de codi de programació s'han escrit amb l'objectiu d'obtenir informes que es puguin categoritzar en gran part com a informes dinàmics per la flexibilitat de futures modificacions o recàlculs amb nous conjunts de dades. Es considera la obtenció de 3 productes en aquest projecte en forma d'informe dinàmic i que es mostren en els annexos i en el material addicional:

- Simulació de models mixtos en el context dels estudis d'estabilitat: Generació de conjunts de dades simulades amb paràmetres definits i modelatge d'aquests conjunts de dades amb resultats de l'avaluació dels models testejats.
- Simulació de models amb dades perdudes: Generació de conjunts de dades simulades amb paràmetres definits, generació de dades perdudes de manera simulada amb paràmetres definits, modelatge d'aquests conjunts de dades, resultats de les comparacions per avaluar els efectes de la presència de les dades perdudes i avaluació comparativa dels models en aplicar dos models d'imputació de dades perdudes.
- Anàlisi del cas pràctic dins el context de l'aplicació dels models mixtos: Anàlisi estadístic d'un conjunt de dades longitudinals amb equivalència de condicions amb els estudis d'estabilitat, analitzant els models avançant de simples a complexos amb l'avaluació en cada pas de la bondat d'ajust i les mesures comparatives de qualitat entre models.

1.6 Breu descripció dels altres capítols de la memòria

Es resumeix de manera breu el contingut de cada capítol:

- Conceptes teòrics: Es prenen les idees generals per cada aspecte que es consideri rellevant pel posterior anàlisi teòric exposant les idees bàsiques sobre els apartats corresponents:
    - Estudis d'estabilitat
    - Dades longitudinals
    - Models mixtos
    - Dades perdudes
    - Esquema general

- Anàlisi teòric: Es divideix en una part on es prenen els conceptes teòrics del capítol anterior i es condicionen a un anàlisi més específic i una part on amb programació es simulen dades i es mostren els resultats obtinguts.
    - Estudis d'estabilitat: Especificació de les condicions que es decideixen utilitzar per l'anàlisi teòric a partir dels coneixements exposats en el capítol anterior i a la bibliografia consultada.
    - Models mixtos: Selecció i exposició de les condicions preses per l'anàlisi teòric prenent lo exposat en el capítol anterior i desenvolupament teòric d'aquestes. Exposició dels resultats de les simulacions realitzades.
    - Dades perdudes: Selecció i exposició de les condicions preses per l'anàlisi teòric també a partir del capítol anterior i desenvolupament teòric d'aquestes. Exposició dels resultats de les simulacions realitzades.
- Anàlisi pràctic: Exposició dels resultats obtinguts en l'anàlisi estadístic amb dades reals on s'ha intentat aplicar alguns dels models vistos en l'anàlisi teòric de models mixtos.
- Conclusions: Breu exposició de les conclusions extretes del treball i possibles ampliacions de futur considerant planificacions anteriors descartades o aprofundiment de coneixements en algun dels aspectes treballats.

2 Conceptes teòrics

En el present capítol es pretén descriure els conceptes teòrics dels punts principals que conformen el disseny del treball. En aquesta fase preliminar es descriuen d'una manera conceptual, sense entrar en el detall del disseny concret a utilitzar i descrits de manera relativament independent, és a dir, sense establir encara una relació directe entre ells.

En el capítol posterior 3 Anàlisi teòric de es pretén entrar més en el detall d'analitzar cada un d'ells a nivell teòric en el context del treball i en fer aplicacions simulades per analitzar-ne l'aspecte més pràctic.

2.1 Estudis d'estabilitat

Les idees del present capítol i que seran la base per la posterior anàlisi, s'han extret de guies de qualitat d'organitzacions públiques i s'exposen en el capítol 7 Bibliografia. Donat que la major part d'agències governamentals públiques tenen regulacions específiques per la realització dels estudis d'estabilitat dels productes biològics enfocats als medicaments, per enfocar el treball en els mètodes de treball més habituals s'ha considerat la utilització de les guies d'harmonització que són vàlides a una quantitat més gran de països que altres regulacions més específiques corresponents a les *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (ICH) i la *World Health Organization* (WHO).

L'objectiu principal dels estudis d'estabilitat de productes biològics comercials és el de «proporcionar evidències de com la qualitat d'un producte final o un producte semielaborat varia en el temps sota la influència d'una varietat de factors ambientals com la temperatura, la humitat i la llum, i establir un període de re-anàlisi del producte semielaborat o una vida útil pel producte final i unes condicions d'emmagatzematge recomanades» (traducció de cites de [4] / [5]). Per aconseguir aquestes evidències es dissenyen els estudis mitjançant els quals s'emmagatzemen mostres representatives dels productes d'interès a diferents condicions (habitualment de temperatura i/o humitat) durant un temps determinat. Durant el temps que dura l'estudi es testegen les mostres per fer el seguiment del seu comportament en els paràmetres d'interès de l'estudi i veure'n la seva evolució.

Per realitzar el disseny d'un estudi d'estabilitat els punts més importants a tenir en compte habitualment són:

- Les condicions de temperatura o humitat. S'estableixen tenint en compte la climatologia de diferents zones del món i també utilitzant les condicions anomenades d'estrès del producte pels estudis anomenats d'estabilitat accelerada on s'intenta identificar d'una manera més clara degradacions més específiques que pugui tenir el producte en un espai més curt de temps, i que en altres condiciones més favorables podrien

quedar ocultes a les analítiques de testeig durant l'estudi i no detectar-se fins al final de l'estudi.

- Freqüència de testeig: A partir del tipus d'estudi i duració es sol proposar la freqüència de testeig.
- Selecció dels subjectes o en el cas de indústria habitualment lots/presentacions representatives del producte.
- Paràmetres i especificacions: Paràmetres que es mesuren en cada testeig i criteris d'acceptació proposats per aquests (en cas de voler investigar únicament com es comporta el producte no és necessari tenir uns criteris establerts inicialment).

Aquestes condicions s'estableixen habitualment a partir de les recomanacions descrites en les guies ICH o WHO. En el cas de les especificacions dels tests de paràmetres químics i/o biològics habitualment queden establertes a partir de les farmacopees o documents equivalents que siguin d'aplicació en el país on es presenti l'estudi.

Per l'avaluació d'aquests estudis d'estabilitat les mateixes guies estableixen els casos més habituals d'anàlisi i donen indicis de com realitzar els anàlisis estadístics corresponents sense donar un grau molt elevat de especificitat i deixant l'elecció de la major part de condicions o tècniques a l'investigador.

Amb la combinació dels paràmetres, tipus d'estudi i recomanacions en general de les guies ICH/WHO és possible augmentar la complexitat dels estudis enormement en cas necessari. Tot i que els estudis d'estabilitat són la base d'aquest projecte, en aquest capítol no es considera necessari detallar totes les situacions possibles ja que no és l'objectiu final, pel que en l'apartat 3.1 Estudis d'estabilitat dins el capítol 3 Anàlisi teòric, es detallen algunes de les situacions més habituals i en concret es té en compte un cas concret d'aplicació.

2.2 Models amb dades longitudinals

El tipus de dades que es pretén modelitzar són les anomenades longitudinals. Abans de continuar avançant en els conceptes teòrics és important tenir clar què significa que els conjunts tinguin dades longitudinals a nivell d'anàlisi estadístic. A continuació es resumeix les característiques considerades més rellevants i extretes en la seva major part de Weiss sobre modelització d'aquest tipus de dades [6].

«Les dades longitudinals són un tipus de dades amb utilitat en una gran varietat de camps. (...) També s'anomenen mesures repetides i series en el temps. (...). Una mesura és presa repetidament al llarg del temps a cada subjecte de l'estudi» [6].

Bàsicament aquest tipus de dades que està present inclús fora dels camps científics (per tant els mètodes per analitzar-se poden extrapolar-se també en

molts camps), són una forma particular de les dades anomenades de mesures repetides on cada tipus de mesura es recull repetidament en cada subjecte o unitat experimental de l'estudi, i han de poder ser ordenades al llarg del temps, és a dir, en una dimensió per complir la característica clau de ser longitudinals.

A diferència dels conjunts de dades univariants, en les dades longitudinals es poden atribuir les següents inferències principalment:

- Inferència sobre la mitjana de la població a cada temps on es prenen dades, essent aquesta constant o amb algun sentit de tendència.
- Variabilitat individual sobre la mitjana de la població al obtenir inferència per la variància i desviació estàndard de la població per cada temps on es prenen dades.
- Els efectes de les covariants en la mitjana de la població i en la variació individual. Estan relacionats amb la variància que es calcula entre dos temps i amb la correlació de les diferents observacions en cada temps amb la mitjana.
- Prediccions de noves observacions. El fet que es tingui en compte els temps com a punts individuals d'observacions i es calculi la relació entre ells fa que es puguin fer prediccions amb els models de nous subjectes i valors futurs.

En el capítol 3 Anàlisi teòric s'analitza amb més detall aquest tipus de dades i les inferències a realitzar a nivell pràctic.

2.3 Models d'efectes mixtos

2.3.1 Definició i estructura dels models

Dins dels models generals de regressió lineal s'anomenen els models d'efectes mixtos aquells que contenen efectes de tipus fix i de tipus variable dins el mateix model. Habitualment això significa que el model conté almenys un factor de tipus fix i un factor de tipus variable, i per tant, pot contenir la interacció d'aquests dos tipus de factor donant lloc a nous efectes aleatoris.

Per l'anàlisi estadístic d'aquests efectes mixtos Oehlert [7] defineix dos conjunts d'assumpcions com a possibles estàndards de l'anàlisi d'aquest tipus de models que van relacionats també amb dos mecanismes diferents per generar les dades:

- Model restringit: En aquest model les dades del factor aleatori es prenen efectivament de manera aleatòria dins un conjunt més gran de possibilitats, però es té la certesa que de totes les possibilitats, cada una dona sempre els mateixos efectes de interacció. És a dir, que si es repetís l'experiment amb els mateixos nivells del factor escollit dels aleatoris donaria els mateixos resultats d'interacció.

- Model no restringit: En aquest cas, contràriament al model anterior, cada vegada que es fa l'experiment amb el factor aleatori hi ha un mostreig independent dels efectes de les poblacions. És a dir que si es repeteix l'experiment triant els mateixos factors, l'efecte d'interacció no resulta el mateix.

En molts casos no és clar quin model s'ha de prendre ja que es pot estar a mig camí entre un i l'altra si per exemple entre dos repeticions en un mateix factor s'espera obtenir resultats lleugerament diferents, però no significativament diferents a nivell d'efecte d'interacció. És convenient fer una reflexió a cada experiment per triar el model que més s'adequa a l'anàlisi del efectes mixtos.

En general l'estructura dels models ve donada per les següents decisions:

- Tipologia dels factors que intervenen entre fix i aleatori.
- Interacció o niatge (*nesting*) entre factors.
- Model restrictiu o no restrictiu.

Tal com descriu J.Jiang [8] hi ha algunes distincions que poden ser necessàries tenir en compte addicionalment:

- El tipus de model lineal d'efectes mixtos segons si és tracta d'un model Gaussià o no Gaussià.
- El model lineal d'efectes mixtos típic o una variació anomenada model generalitzat lineal d'efectes mixtos, tot i que aquest últim habitualment s'utilitza en models amb dades longitudinals discretes o binàries, en les condicions específiques següents:
    - La mitjana de la observació que en el model simple és funció lineal de una o més covariants en aquest cas també està associada a una funció lineal de covariants però a través d'una funció que enllaça.
    - La variància de les observacions que en el model simple es considera constant, en el generalitzat és funció de la mitjana.
    - S'inclou també moltes més distribucions possibles a part de la normal.

Les condicions de cada tipologia estan relacionades amb el compliment de las assumpcions del model pel que s'han de tenir en compte a l'hora d'avançar en l'ajust dels models.

Els diagrames de Hasse habitualment són el punt de partida per fer les estimacions dels models amb efectes mixtos, però normalment a dia d'avui aquests ajustos es fan a través dels algoritmes d'ajust preparats que ja inclouen internament aquests diagrames, pel que es focalitza més l'estimació del model (veure 2.3.2).

2.3.2 Estimació dels models

A continuació es resumeixen les propietats que tenen els mètodes més habituals d'estimació dels paràmetres sense entrar en la complexitat matemàtica, a efectes de poder escollir en l'anàlisi teòric. Per això s'utilitza com a base vàries referències de la bibliografia on es compara a nivell teòric els diferents mètodes com Everitt/Hothorn, Galecki, Jiang, Oehlert, Quinn GP, Verbeke/Molenberghs i Weiss [6–12] :

- ANOVA (mínims quadrats): En els models simples s'utilitza habitualment el mètode de minimitzar la suma de quadrats residuals, però quan apareixen efectes aleatoris s'ha de tenir en compte un nou concepte anomenat la mitjana quadràtica esperada o *expected mean squares* que té en compte que la variabilitat total del model ve de la variabilitat residual i de les variabilitats aportades pels efectes aleatoris que hi puguin haver.

  A diferència dels mètodes de màxima versemblança (exposats a continuació) i com a part positiva, matemàticament és el més simple i no té la condició de la distribució normal (excepte si s'han de calcular intervals de confiança o contrastos d'hipòtesis). També és vàlid amb la condició de no correlació d'errors residuals (assumpció menys estricta que independència d'errors). Com a part negativa, a diferència dels mètodes de màxima versemblança, hi ha el perill de que les estimacions calculades pels mínims quadrats puguin resultar negatives al estimar la variància residual el qual pot ser un indicador d'estar aplicant un model inadequat o de tenir *outliers* importants.

- Estimacions de màxima versemblança o *maximum likelihood estimates* (MLE): Es tracta de derivar les equacions del model lineal i les solucions iteratives en un procés de càlcul més complex. L'estimació ML es considera esbiaixada al dependre de la mitjana de les observacions el càlcul de l'estimador de variància i tendeix a subestimar els components de variància.

  A diferència de ANOVA com s'ha comentat necessiten assumir distribució normal pel càlcul de les estimacions.

- Estimacions de màxima versemblança restringida o *restricted maximum likelihood estimates* (REML): Modificació del modelatge MLE. Com a millora respecte al MLE s'exclou el terme de la mitjana de la funció de versemblança per corregir el biaix en els estimadors de la variància i corregeix la subestimació de variància dels components al proporcionar una estimació més robusta (les correccions entre REML i MLE prenen especial importància en mostres de mida reduïda o en mostres grans si les diferències entre mida i nombre de covariants d'efecte fix són petites). En contrapartida, si es vol comparar dos models utilitzant un

ràtio de versemblança, només permet la comparació entre models estimats per REML amb la mateixa estructura d'efectes fixos.

En el cas de dissenys balancejats produeix les mateixes estimacions de variància que el mètode ANOVA, pel que aquestes estimacions tampoc depenen de la normalitat. Tot i així hi ha una part del model REML calculada igual que l'estructura MLE pel que en part sí que dependran d'aquesta assumpció.

2.3.3 Modelació de la matriu de covariància

Com s'ha comentat en apartats anteriors del present capítol, els models d'efectes mixtos contenen com a mínim un factor d'efecte aleatori, és a dir que aquest factor (per exemple subjecte) té un efecte que es dona per fet que pot variar segons el nivell del factor que s'utilitzi a l'estudi. En el cas dels models amb dades longitudinals on es produeix la repetició de mesures en diferents temps, habitualment es dona també un factor de correlació entre les mesures preses en un mateix subjecte, ja que per exemple la dada d'un temps molt probablement pot tenir relació amb la dada del temps anterior o del temps posterior.

Per intentar tenir en compte aquestes condicions tan específiques és important modelar la matriu de covariància que defineix de quina manera es comporten els factors aleatoris. Especialment en la situació comentada de mesures repetides és important tenir en compte la matriu de correlació que sovint va directament relacionada amb la de variància covariància.

Aquest és un dels punts obscurs a dia d'avui en el modelatge de models d'efectes mixtos i que segons els experts encara és molt desconegut. En el present apartat s'ha pres com a referència l'article de Kincaid [13] i el llibre de Weiss [6], ambdós tractant el modelatge d'aquest tipus de matrius i models. Aquestes referències plantegen alguns models bàsics d'estructura de matriu de covariància a tenir en compte per escollir. Es descriuen breument els que es solen associar més a dades longitudinals:

- Components de variància o *Variance Components* (VC): L'estructura per defecte dels components de variància estàndards.
- Autoregressiva o *Autorregressive* (AR): Estructura amb variància homogènia i tenint en compte la matriu de correlacions. En aquest cas està dissenyada perquè la correlació disminueixi exponencialment segons la distància de les observacions.
- Simetria composta o *Compound Simmetry* (CS): Estructura tradicionalment utilitzada en els dissenys amb variància homogènia. Té en compte la correlació entre mesures separades, però s'assumeix com a constant independentment de la distància entre les mesures.

- Intercepció aleatòria o *Random Intercept* (RI): Quan s'assumeix una pendent zero i les observacions depenen de la variació aleatòria que es produeix en l'intercepció de la regressió. El resultat és un model similar al CS on hi ha variància constant.
- Intercepció i pendent aleatòries o *Random Intercept and Slope* (RIS): És un cas generalitzat de l'anterior on segons el subjecte es produeix un efecte aleatori tant a l'intercepció com a la pendent de la regressió. És un model més complex que CS, RI i AR ja que afegeix dos paràmetres corresponents a la variància de pendents i la covariància entre intercepcions i pendents.

Kincaid [13] proposa possibles estratègies per seleccionar l'estructura de les quals es descriuen les bàsiques:

- Segons parsimònia: Si el conjunt de dades ho permet es pot escollir el model UN i anar ajustant individualment. En aquest cas també depèn de si s'està ajustant un model més simple o més complex amb un nombre més elevat de paràmetres a ajustar.
- Segons significat: El nombre d'estructures estudiades és molt major de les exposades aquí, pel que no és una bona tàctica intentar aplicar-les totes i veure quina és la que dona més bon resultats. Intentar entendre el tipus de dades a manejar del disseny, les estructures dels tractaments i els tipus de covariàncies habitualment restringeixen les opcions a escollir.
- Segons criteri d'informació o *Information Criteria* (IC): Les IC estadísticament són indicadors de la eficàcia del model. El criteri de Akaike o *Akaike's Information Criteria* (AIC) o el criteri Bayesià o *Bayesian Information Criteria* (BIC) en serien dos exemples.

  En aquesta estratègia es tracta de ajustar models amb estructures de covariància oposades i comparar els indicadors IC. Tot i així els experts aconsellen tenir en compte que aquest criteri no garanteix que s'estigui utilitzant l'estructura correcta.

Finalment i relacionat amb l'apartat 2.4 Models longitudinals amb dades perdudes i com comenta Weiss [6], un bon model de matriu de covariància garanteix una millor aproximació en la imputació de valors perduts.

2.4 Models longitudinals amb dades perdudes

Les dades perdudes en els models estadístics són bastant freqüents en els casos amb dades reals, i els estudis amb dades longitudinals no en són una excepció. Tenint en compte que en aquest tipus d'estudis s'està realitzant observacions en punts de temps determinats, habitualment amb més d'un subjecte i més d'un paràmetre, es poden donar moltes situacions on es produeixin dades perdudes. Per exemple que un temps no s'hagi analitzat un

subjecte, que un subjecte no hagi acabat l'estudi o que un dels paràmetres no s'hagi pogut analitzar. En resum, totes les dades que no s'han obtingut per una causa fora del control de l'investigador.

En el present capítol es fa un resum de la teoria a l'entorn de l'existència de les dades perdudes d'una manera resumida. Posteriorment en l'apartat 3.4 Models amb dades perdudes es detalla amb més profunditat quines proves es fan per comprovar-ne la possible aplicació en casos més específics.

2.4.1 Tipus de dades perdudes

En resum i tal com explica Faraway [14] i també exposa J.J.Curto [15], els tipus de dades perdudes més típiques en els models de regressió són:

- Casos perduts: Es dona quan falla un cas complet, és a dir quan hi ha dades perdudes en totes les variables d'un individu. És important conèixer el mecanisme pel qual s'ha donat el cas perdut ja que en depèn poder fer les correccions i inferències adequades.
- Valors incomplets: Aquest cas és més habitual quan l'estudi finalitza abans que passi alguna cosa específica que s'espera, el que estadísticament es defineix com un esdeveniment (*event*) generant casos censurats. Aquest tipus de pèrdua de dades es pot solucionar habitualment amb mètodes d'anàlisi de supervivència o anàlisi de fiabilitat.
- Valors perduts: Es dona quan hi ha dades perdudes en una o més variables sense l'absència completa d'un registre (individu o subjecte d'anàlisis).

Pels casos de valors perduts és important analitzar les mecanismes pels quals s'han perdut. A nivell genèric tal com explica Little [16] i també comenta Faraway [14] es distingeixen els següents tipus:

- Dades perdudes completament a l'atzar (MCAR, *missing completely at random*): Es dona quan la probabilitat que un valor es perdi és la mateixa per tots els casos. Una possible solució és eliminar els casos on hi hagi valors perduts ponderant si és assumible el risc que comporta de pèrdua d'informació.
- Dades perdudes a l'atzar (MAR, *missing at random*): Es dona quan la probabilitat de tenir valors perduts depèn de un mecanisme conegut per l'investigador. Tot i que és possible una eliminació de casos com en el cas anterior, en aquest depèn de si el model inclou el mecanisme que provoca aquesta pèrdua com a factor del model, i també s'ha de tenir en compte si es pot assumir la pèrdua d'informació.
- Dades perdudes no a causa de l'atzar (MNAR, *missing not at random*): Es dona quan la probabilitat de tenir valors perduts depèn d'una variable

no observada o d'una possible observació verdadera no feta. És el tipus de pèrdua més difícil de tractar.

2.4.2 Cas de models no balancejats

En el cas on no hi ha el mateix nombre de respostes per cada factor-nivell del model es diu que el model és no balancejat. Aquesta condició es pot donar habitualment quan es dissenya un estudi de manera que estigui previst obtenir les dades de manera balancejada i es produeix l'efecte de les dades perdudes (en el cas de casos perduts si es considera el subjecte com un factor o en el cas de valors perduts en general).

Com explica Oehlert [7], quan es dona aquest cas l'anàlisi estadístic es complica. A diferència de les dades balancejades on els contrastos son ortogonals i per tant els resultats independents, en les dades no balancejades els resultats obtinguts per cada terme depenen d'altres termes en el model i per tant perden precisió si no s'analitzen de manera simultània. Principalment s'ha de tenir en compte al generar els models, calcular les sumes de quadrats i els contrastos d'hipòtesis.

2.4.3 Mecanismes per solucionar la presència de dades perdudes

Per tractar amb la pèrdua de dades comentada en els apartats anteriors s'exposen els mecanismes més habituals a utilitzar :

- Eliminació de casos: Com s'ha anat comentant hi ha el risc de pèrdua d'informació important. Habitualment depèn de factors com la mida del conjunt de dades. Abans d'aplicar aquest mecanisme és habitual fer alguns anàlisis i gràfics per comprovar quin és el possible efecte d'aquesta eliminació.
- Mètodes d'imputació simple per substituir els valors perduts. Dins d'aquesta categoria hi ha varis sistemes tal com comenta Jorge J. Curto García [15], dels quals es descriuen alguns dels bàsics més relacionats en el context d'aquest TFM:
    - Substitució per la mitjana condicional mitjançant regressió simple
    - Substitució pel veí més proper amb el càlcul de la distància.
    - Imputació per la observació prèvia (LOCF *Last Observation Carried Forward*): Utilitzar la observació del valor precedent en el temps del mateix individu (específic per dades longitudinals).
    - Aplicació de la regressió estocàstica bayesiana: Predicció semblant a la feta per regressió, però tenint en compte la distribució de la variància de l'error i la incertesa de l'estimació dels coeficients (i la distribució d'aquests coeficients de regressió).

Els mètodes d'imputació simple encara que ajuden a reduir la variància, tenen el risc de provocar un biaix que afecti el model en generar els nous valors. En el cas concret dels que utilitzen la regressió lineal, un dels factors que determina l'efecte d'aquest esbiaix és la col·linealitat, pel que és important analitzar si és millor fer la imputació o eliminar el predictor que acumuli els valors perduts al no suposar una pèrdua gran d'informació pel mateix fet de tenir una elevada col·linealitat amb altres predictors.

Els conjunts de dades que es manegen en el context d'estabilitats solen ser de dades limitades pel que habitualment no s'utilitzaran mètodes més complexos. Tot i així val la pena tenir en compte que existeixen de manera resumida els d'imputació múltiple per intentar reduir el biaix produït pels mètodes d'imputació simple. La manera de aconseguir això habitualment és la de reintroduir la variació de l'error al fer la substitució no tingut en compte en els mètodes d'imputació simple, i fer-ho de manera múltiple.

Com descriuen Huque/Simpson/Lee [17] aquests mètodes inclouen la generació de múltiples còpies del conjunt de dades substituint els valors perduts en cada una per valors imputats a partir de la distribució predictiva de les dades observades. Les aproximacions que exposa l'article i descrits de manera molt resumida serien:

- Modelació conjunta o *Joint modelling* (JM): Procediment que assumeix una distribució multivariant conjunta entre totes les variables del model d'imputació.
- Especificació completament condicional o *Fully conditional specification* (FCS): Procediment que imputa les dades utilitzant un model condicional univariant per cada una de les variables on hi ha dades perdudes.

Dins els mètodes d'imputació múltiple es poden utilitzar varis mètodes per optimitzar els resultats. Els algoritmes més utilitzats serien:

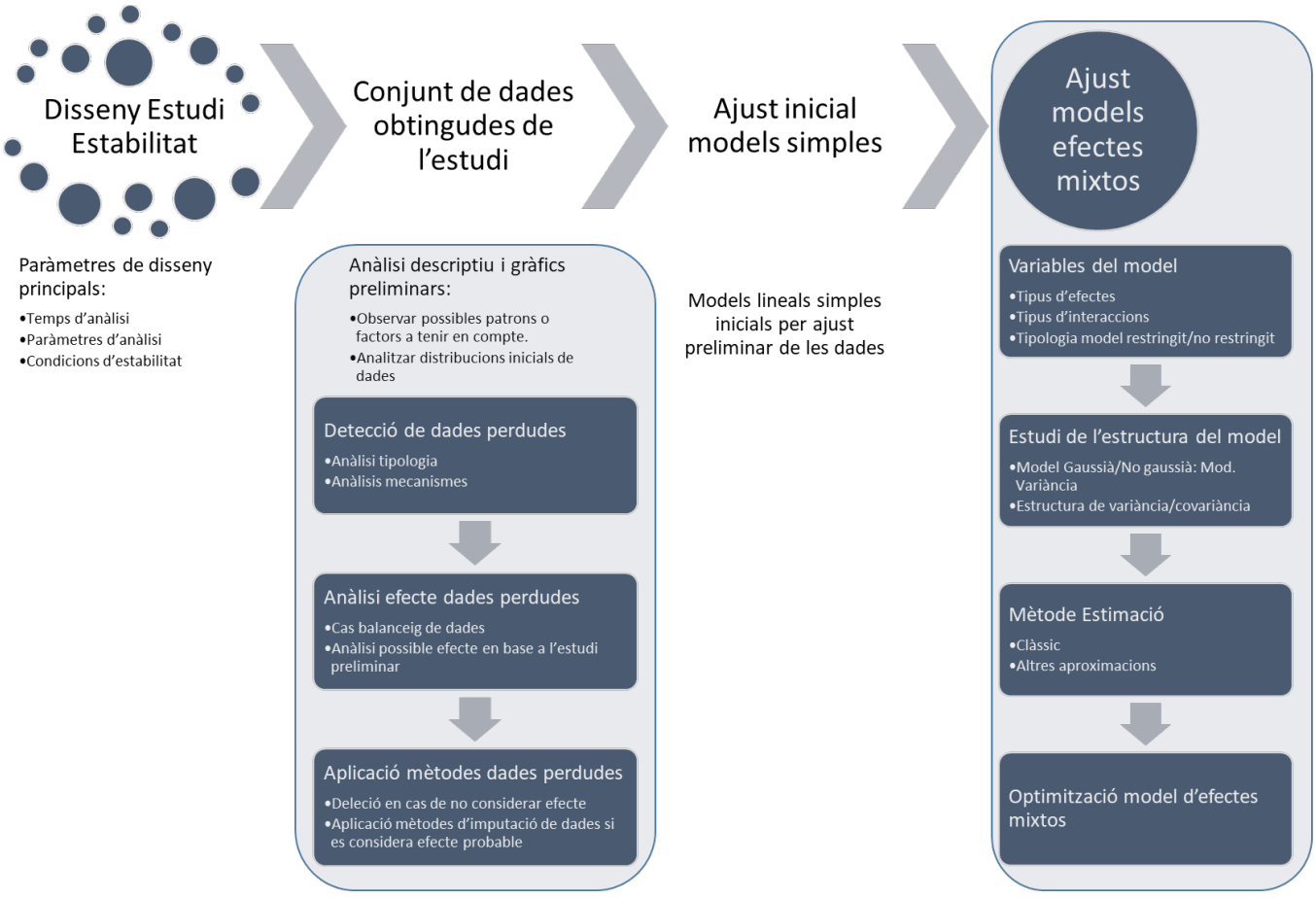
- Algoritme de expectació-maximització (EM): Com indica Peña [18] al realitzar un anàlisi multivariant l'algoritme EM és una bona opció per estimar amb fiabilitat les dades perdudes.
- Màxima versemblança amb informació completa o *Full Information Maximum Likelihood* (FIML): Com s'exposa en el TFM de Jorge J. Curto García [15] és molt semblant a l'algoritme EM, però en aquest cas l'algoritme de màxima versemblança s'aplica per separat a cada cas o individu fent un procés semblant a l'anterior de predicció, maximització i iteració.
- *Monte Carlo Markov Chain* (MCMC): Es descriu habitualment con una de les principals alternatives al mètode EM, inclús es pot veure comparacions entre els dos analitzant les diferències com fa Lin [19]. Aquest mètode es basa en la cadena Markov o *Markov Chain* que és

una seqüència de variables aleatòries en les quals la distribució de cada element depèn únicament en el valor de l'anterior.

Existeixen múltiples algoritmes i sistemes per tractar les dades perdudes, en alguns casos variacions dels existents adaptats a aplicacions molt específiques. Per l'aplicació del present TFM es considera suficient amb el resum dels més utilitzats. En l'apartat 3.4 Models amb dades perdudes s'analitza amb més detall les possibles aplicacions en els models específics proposats.

2.5 Esquema general o punt de partida de l'anàlisi

De manera molt resumida i tenint en compte els diferents apartats anteriors del present capítol de Conceptes teòrics es mostra un esquema de possible enllaç de les diferents idees per tenir un mapa conceptual de les idees principals per avançar en els anàlisis posteriors:



3 Anàlisi teòric

En el present capítol es pretén realitzar un anàlisi teòric passant pels punts principals que conformen l'anàlisi tipus que es vol estudiar. Utilitzant els conceptes descrits en el capítol anterior 2 Conceptes teòrics es pretén intentar veure amb més detall el disseny concret a utilitzar i la relació entre les condicions que es consideren d'aplicació en el context dels tipus d'estudis a analitzar.

Addicionalment es generen mitjançant codi de programació aplicacions simulades per analitzar-ne l'aspecte més pràctic i treure'n possibles conclusions.

3.1 Estudis d'estabilitat

En aquest apartat es descriu a nivell teòric les condicions o sistemes per analitzar estadísticament els estudis d'estabilitats que es proposa des de les guies o *guidelines* dels organismes oficials considerats (ICH; WHO). Al haver idees molt similars es pren en concret ICH per aquesta capítol d'aplicació en anàlisi teòric.

Per començar extret dels conceptes globals de les guies i de l'experiència personal de l'autor del TFM es descriu els termes principals que es consideren dins d'un estudi d'estabilitat:

- Observacions: A cada subjecte o unitat experimental es realitza una sèrie de mesures repetitives de un o més paràmetres. En aquest cas els resultats d'aquestes mesures són les observacions. Aquestes observacions habitualment són la variable resposta o objectiu del model, encara que en cas de múltiples paràmetres a mesurar, es pot tenir en compte alguns paràmetres com a variables predictores del model relacionades amb altres dels paràmetres. Es pren la consideració de tenir-les en format numèric de tipus continuu.
- Factor temps covariant: El factor dels temps a estudiar marcats per l'investigador com a norma general és la covariant o covariant principal del model, ja que un dels objectius habitualment és comprovar com varia la variable resposta en front d'aquesta covariant.
- Factor condició d'emmagatzematge: L'efecte d'aquest factor s'assumeix amb seguretat que sigui del tipus fix ja que es tracta d'una condició fixada per l'investigador. Es pren el model més simple amb una sola condició i per tant sense aparèixer al model en aquest cas.
- Factor subjecte/lot: En cas que al model intervingui més d'un lot de producte és un factor a tenir en compte. Com que els estudis habitualment tenen la pretensió de representar la població, habitualment

s'hauria de tenir en compte com un efecte aleatori atès que segons el lot que agafi el resultat de l'estudi pot variar.

- Factors relacionats amb el lot: Les característiques en les quals es puguin agrupar els lots poden formar part del model si són d'interès en algun contrast o predicció. Per aquest anàlisis no es tindran en compte.

A continuació es procedeix a extreure de cada guia consultada els conceptes que en facin referència i a desenvolupar-los mínimament en aquesta fase preliminar de l'anàlisi teòric per assentar les condicions de base a seguir que siguin d'interès:

3.1.1 Stability testing of new drug substances and products Q1A(R2) – ICH [4] /Evaluation for stability data Q1E – ICH (ICH, 2004) [20]

L'establiment d'una caducitat, vida útil o període de reanàlisi a un producte s'estableix amb l'anàlisi mínima de 3 lots del producte avaluant la informació que en resulti de l'estudi d'estabilitat complet. Pels models habituals es considera un rang que pot anar dels 3 fins als 10 lots.

Segons el que es mostri de les dades obtingudes de l'estudi es pren la decisió de utilitzar models més simples analitzant per exemple només la variabilitat de les repeticions o utilitzar models més complexos. Habitualment el procés que es segueix és el d'ajustar els models de més simples a més complex per veure'n la conveniència d'aplicar o no aplicar segons quines estructures.

En el cas que l'anàlisi mostri poca variabilitat entre lots, és viable combinar les dades per fer un únic anàlisi. Pel test de combinació de lots, una de les possibilitats és l'ús de l'anàlisi de covariància ANCOVA amb el temps com a covariant per testar les diferències en pendents i intercepcions a l'origen. Es demana a les guies per aquests tests un nivell de significància del 0.25 per compensar l'efecte habitual dels estudis d'estabilitat de una baixa mida mostral.

Per testejar altres factors o combinacions de factors es procedeix de la mateixa manera que amb les comparacions de lots tot i que les guies en el cas de factors no relacionats amb lots diferencien el nivell de significació de la comprovació de agrupabilitat o *poolability* dels lots (0.25) amb la dels factors no relacionats amb els lots (0.05).

La naturalesa de les possibles relacions de degradació determinen també si és necessari aplicar algun tipus de transformació a les dades per l'anàlisi de regressió lineal. Habitualment poden ser del tipus lineal, quadràtica, cúbica i en una escala aritmètica o logarítmica.

Es recomana utilitzar mètodes estadístics per testejar la bondat de l'ajust (*goodness of fit*) de les dades en tots els lots o combinació de lots (si s'escau) a la suposada línia o corba de degradació. Es prenen els mètodes de bondat d'ajust habituals com els coeficients de determinació, les proves de regressió,

els indicadors de qualitat de criteris d'informació i les proves de les assumpcions com a paràmetres habituals de bondat d'ajust.

3.1.2 Tipus d'estudis a analitzar

Tenint en compte els conceptes teòrics presentats a l'apartat 2.1 i les situacions d'anàlisi estadístic descrites segons la guia Q1E [20] , es poden establir els anàlisis habituals que són necessaris en els estudis d'estabilitats com els següents:

- Prova de agrupabilitat per determinar la combinació de factors utilitzant les dades de referència. Habitualment utilitzat per comprovar l'equivalència dels individus o lots que representen el mateix tipus de producte.
- Aplicació de diferents factors o combinacions de factors característics del producte a part del factor lot per comprovar-ne l'equivalència.
- Aplicació del factor condició d'emmagatzematge per tenir un model de predicció.
- Predicció o extrapolació de les dades d'estudis de referència per proposar un nou període de re-anàlisi o de vida útil del producte o per predir el comportament d'un lot o individu en estudi o per tenir un model genèric per predir el comportament del producte en general.

Per veure un anàlisi teòric més concret en aquest treball es considera l'aplicació del cas possiblement més genèric, ja que en certa mesura engloba part dels altres, i que és el modelatge del comportament específic d'un producte a partir de les dades d'estudis d'estabilitat finalitzats, per tal d'obtenir un model que serveixi tant per descriure el comportament d'aquest producte com per utilitzar en predir el comportament de futurs lots en estudi.

3.2 Models d'efectes mixtos

En aquest treball es contempla el cas específic d'aplicar el model d'efectes mixtos. Per fer-ho es pretén començar a partir dels models més simples i avançar cap als més complexos analitzant les diferències per intentar trobar el model més adequat en cada cas.

3.2.1 Model base d'efectes mixtos

Es comença pel model bàsic amb covariants que és el típic dels estudis d'estabilitat i en molts casos dels estudis amb dades longitudinals atès que un dels objectius principals és veure l'efecte d'aquesta covariant en les respostes:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij} \quad \textit{Eq 3.2.1.a}$$

On:

i *: Individus/lots del model. En el cas dels estudis d'estabilitat es pot considerar que hi pot haver un mínim de 3 i un màxim de 10 lots per generar un model.*

j *: Repeticions o en el cas dels estudis d'estabilitat els valors possibles de la covariant que són els temps d'estudis. Es considera per aquest tipus d'estudi que poden ser entre 4 i 10 repeticions segons la condició dels estudis i el paràmetre a estudiar.*

y_{ij} *: Matriu o vector de respostes corresponents a cada condició de tractament i repetició.*

μ *: Mitjana global de les respostes.*

α_i *: Efecte del tractament.*

β *: Pendent del model. Inicialment s'assumeix que és el mateix per cada tractament.*

x_{ij} *: Valor de la covariant temps per cada tractament i repetició. Tot i que hi ha estudis a petita escala al laboratori on es fan estabilitats de minuts o hores, les estabilitats de producte es solen fer comptant per mesos la covariant.*

ϵ_{ij} *: Part desconeguda o d'error del model.*

3.2.2 Restriccions model

Tal com expliquen Everitt i Hothorn [9], l'aplicació dels models d'efectes mixtos per dades amb mesures repetides reflexa d'alguna manera la idea que el patró de respostes individuals és probable que depengui de moltes característiques d'aquest individu incloent les no observades. En el cas dels estudis d'estabilitat es pot reflexionar que encara que els lots segueixin el mateix procés exacte de fabricació sempre poden tenir variables no tingudes en compte que els diferencien entre ells, igual que si es pren el factor auxiliar que realitza l'assaig per obtenir la observació pot variar encara que es segueixi una metòdica concreta. Totes aquestes variables no observades o no tingudes en compte es consideren en el model poden ser variables aleatòries que provoquen efectes aleatoris.

La reflexió exposada en el paràgraf anterior també pot estar directament relacionada amb l'assumpció del tipus de model d'efectes mixtos que es planteja sobre si és restringit o no restringit sobre la qual es poden fer algunes reflexions:

- A priori si només es té l'efecte del lot es podria encaixar els models d'estudis d'estabilitat en els models restringits ja que els lots teòricament són característics i se suposa que totes les mostres procedents del lot són iguals al haver una homogeneïtat de producte.
- A nivell pràctic i com s'ha comentat, solen haver molts factors ocults que modulen la resposta i la suposada homogeneïtat de mostres es sol mantenir dins un nivell raonable, però sempre mostrant un cert grau d'heterogeneïtat. Això se suma al fet que l'única manera de demostrar que l'efecte aleatori lot correspon a un model restringit seria fer dos estudis en paral·lel que comencessin alhora, ja que el lot en el context d'un estudi d'estabilitat depèn del temps (és a dir que un mateix lot no és el mateix al cap d'una setmana de la data de fabricació que al cap d'un mes).

A efectes pràctics doncs no es considera l'assumpció dels efectes aleatoris restringits i per tant la suma dels efectes del terme d'interacció no s'assumeix que sumi 0, és a dir que un exemple de la interacció de l'efecte lot amb l'efecte temps o qualsevol altre factor que intervingui de tipus fix no té perquès sumar 0

els efectes d'interacció al llarg de l'efecte fix i com molt bé s'il·lustra en el material de *Penn State University* [21] no s'imposa la restricció:

$$\sum_{1}^{n} (\alpha\beta)_{ij} = 0 \quad \textit{Eq 3.2.2.a}$$

on α i β serien dos efectes un aleatori i un fix.

3.2.3 Assumpcions de les distribucions

Respecte a la condició de Gaussià, es pren en aquest anàlisi sempre l'assumpció inicial que es segueix la normalitat en el model. En el cas que el diagnòstic d'aquest demostri que no es segueix l'estratègia que es segueix és intentar aplicar eines per acostar el model a la normalitat. En aquest treball no s'aborda els models no Gaussians ja que com comenta Jiming Jiang [8] hi ha menys estudis realitzats i les inferències en resulten d'una complexitat més elevada. Per l'assumpció de homoscedasticitat es segueix la mateixa estratègia de contemplar el compliment i en cas necessari aplicar mesures de correcció.

L'aplicació del model del tipus lineal generalitzat no es contempla aquest anàlisi. Aquests models s'utilitzen per respostes discretes i/o binàries i tot i que és possible realitzar un estudis d'estabilitat amb una variable resposta d'aquesta categoria, el més habitual és la modelització sobre una variable numèrica normalment contínua, que et permet estudiar la tendència de la resposta al llarg del temps.

3.2.4 Tipus d'estimacions

En quant a l'aproximació per l'estimació del model, habitualment el mètode ANOVA sol ser menys apropiat per models lineals complexos com poden ser els d'efectes mixtos pels quals solen ser d'aplicabilitat més àmplia. En l'anàlisi teòric es parteix de *datasets* relativament simples pel que no es descarta cap de les tres aproximacions més utilitzades: ANOVA, MLE i REML. En aquests tres casos es pot donar variància no homogènia i això habitualment es pot resoldre pel cas ANOVA amb ANOVA amb factors de pes o weighted least squares (WLS) i pels mètodes de màxima versemblança modificats per estimar també els paràmetres de modulació de la variància, però amb les mateixes diferències entre ells ja descrites. Per aquest anàlisi es decideix prendre directament els models de modulació d'estimacions de màxima versemblança si es dona el cas que s'hagi de modular, per no entrar en la complexitat dels models WLS.

Com a dades addicionals és important tenir en compte segons comenta Everitt [22] que, si s'utilitzen els mètodes de màxima versemblança, el mètode MLE tendeix a subestimar els components de variància comparat amb el mètode REML, però com a contrapartida els models estimats per REML poden comparar-se amb altres models amb tests de ràtio de versemblança únicament si els dos models tenen els mateixos efectes fixos.

3.2.5 Matriu de variància-covariància

La part clau de la matriu de variància-covariància en aquest cas s'opta com a mètode de selecció el mètode «segons significat» dins el context dels tipus d'estudis concrets que es treballen. Pel cas dels estudis d'estabilitat es considera que els dos models més factibles a tenir en compte són el model de intercepció aleatòria i el model de intercepció i pendent aleatòries que habitualment són els que reflecteixen millor els possibles comportaments de les tendències de diferents lots o individus en aquest tipus d'estudi. De fet Everitt/Hothorn [9] , Everitt [22] o Galecki [10] ja proposen en el cas de dades longitudinals amb mesures repetides i en exemples d'aquest tipus l'ús d'aquest tipus de matriu com una de les estructures habituals.

Tot i que no s'utilitza cap altre criteri de selecció en la part de simulació sí que es pren, encara que a posteriori, el criteri d'informació, prenent com a indicadors de qualitat del model els criteris AIC i BIC que resumint com els descriu Peña [18] :

- AIC: Pel Criteri d'informació d'Akaike es pretén adaptar la fórmula de la distància de Kullback-Leibler per comparar la funció de densitat real i una aproximació de màxima versemblança. Per fer això s'ha de minimitzar la distància que equival la equació descrita per Akaike:

  $$AIC = -2L(M_i) + 2i = D(M_i) + 2i \quad \textit{Eq 3.2.5.a}$$

  On la D(Mi) suposa la desviació del model respecte a la realitat i el paràmetre 2i que és la correcció pel nombre de paràmetres afegir al model.

- BIC: És l'enfoc Bayesià del criteri d'informació amb el model de major probabilitat «a posteriori» mitjançant la maximització del producte de la probabilitat del model i de la versemblança marginal arribant a una expressió semblant a l'anterior:

  $$BIC(M_j) = -2L_j\hat{\theta}|X + p_j \log n \quad \textit{Eq 3.2.5.b}$$

  On la desviació del model el pren el conjunt $-2L_j\hat{\theta}|X$ que es minimitza si augmenta el nombre de paràmetres o individus i que es corregeix amb el factor $p_j \log n$.

Seguint d'alguna manera un procés semblant d'anàlisi al que realitza Galecki [10] en els seus exemples de models mixtos, es segueix un procés semblant per analitzar amb més detall què significa la utilització de la matriu de variància-covariància d'intercepció aleatòria i de intercepció i pendents aleatòries. Addicionalment també s'analitza l'ús de les funcions de modulació de la variància en cas de heteroscedasticitat.

3.2.5.1 Model amb intercepció aleatòria

A continuació s'explica un dels modelatges que es podria considerar adequat pel cas dels estudis d'estabilitat basat en el model explicat per Everitt/Hothorn [9] i Weiss [6]. En aquest model s'assumeix que les observacions varien al voltant d'un valor diferent per cada lot/individu i que aquest valors són les intercepcions. S'assumeix que les intercepcions dels lots/individus del model són una mostra de la població de intercepcions en paral·lel a la assumpció principal de que els lots/individus són una mostra de la població del producte. En aquest model s'està assumint que la pendent és zero.

La diferència entre el model base (una variant de l'equació anterior) i el model amb l'intercepció aleatòria:

- base:

$$y_{ij}=\beta_0+\beta_1 x_j+\epsilon_{ij}$$ *Eq 3.2.5.1.-a*

- intercepció aleatòria:

$$y_{ij}=(\beta_0+u_i)+\beta_1 x_j+\epsilon_{ij}$$ *Eq 3.2.5.1.-b*

Els dos models contenen els paràmetres bàsics d'intercepció a l'origen (β_0), pendent o coeficient de regressió de la covariant (β_1) i valor de la covariant de predicció (x_j). En el primer model base s'assumeix que les observacions repetides són independents entre sí cosa que és poc realista en els estudis d'estabilitat i en els estudis amb dades repetides en general del tipus longitudinal. El segon model podria considerar-se més realista ja que afegeix el component u_i aleatori específic de lot/individu que es considera constant al llarg de la covariant (temps) i que en el primer model formava par de l'error general ϵ_{ij} que varia al llarg de la covariant (temps), és a dir que les respostes tenen dos fonts de variació, la variació de mitges al voltant de la mitjana de la població i la variació d'observacions al voltant de la mitjana específica per subjecte. Es pot considerar que s'ha introduït una primera estructura de correlació per les mesures repetides.

D'aquest segon model s'assumeix $u_i \sim N(0,\sigma_u^2)$ i $\epsilon_{ij} \sim N(0,\sigma^2)$ assumint que són independents un de l'altre i de la x_j .

En aquest model s'assumeix l'efecte de la covariant com a fixa entre els individus/lots, assumint cert paral·lelisme amb una variabilitat en la intercepció sobre els lots, però no en la pendent.

L'equació d'efectes mixtos Eq 3.2.5.1.-b també es pot definir de la següent manera:

$$y_{ij}=(\beta_0+u_i)+\beta_1 x_j+\epsilon_{ij}=Z_i b_i+\beta_1 x_j+\epsilon_{ij}$$ *Eq 3.2.5.1.-c*

On Z_i i b_i són la matriu de covariants i el corresponent vector dels efectes aleatoris i on es pren les assumpcions $b_i \sim N(0, D)$ on en aquest cas D correspon només a $\sigma_u^2 = d_{11}$, la variància de l'efecte aleatori de l'intercepció.

La matriu de variància-covariància i la correlació entre repeticions en aquest nou model es considera que per cada lot/individu:

- La variància total de cada mesura repetida és la suma de la variància entre subjectes/lots i la variància residual del propi subjecte:

$$Var(u_i + \epsilon_{ij}) = \sigma_u^2 + \sigma^2 = \tau^2 \quad \textit{Eq 3.2.5.1.-d}$$

- La covariància entre els residus totals a dos nivells diferents del mateix individu no és 0 en el cas que la variància de l'efecte aleatori entre subjectes no és 0:

$$Cov(u_i + \epsilon_{ij}, u_i + \epsilon_{ij}') = \sigma_u^2 \quad \textit{Eq 3.2.5.1.-e}$$

- La correlació com es mostra correspon a la proporció de variància entre subjectes sobre la variància total

$$Cor(u_i + \epsilon_{ij}, u_i + \epsilon_{ij}') = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = \rho \quad \textit{Eq 3.2.5.1.-f}$$

Tenint en compte que la matriu de variància de la resposta es defineix com a:

$$V = Z_i D Z_i' + \sigma^2 I \quad \textit{Eq 3.2.5.1.-g}$$

S'obtenen les següents matrius de variància-covariància que defineixen la variància del model entre diferents temps tenint en compte com es relaciona amb la matriu de correlacions:

- Model base:

$$V = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{pmatrix} \quad \textit{Eq 3.2.5.1.-h}$$

  En aquest cas la part aleatòria és 0 i quedaria definida la variància residual amb una matriu que donaria la mateixa variància per cada mesura repetida i una relació entre repeticions de variància 0.

- Model intercepció aleatòria:

$$V=\sigma^2\begin{pmatrix}1&\rho&\dots&\rho\\ \rho&1&\dots&\vdots\\ \vdots&\vdots&\ddots&\rho\\ \rho&\dots&\rho&1\end{pmatrix}=\begin{pmatrix}\sigma_u^2+\sigma^2&\sigma_u^2&\dots&\sigma_u^2\\ \sigma_u^2&\sigma_u^2+\sigma^2&\dots&\vdots\\ \vdots&\vdots&\ddots&\sigma_u^2\\ \sigma_u^2&\dots&\sigma_u^2&\sigma_u^2+\sigma^2\end{pmatrix}$$ *Eq 3.2.5.1.-i*

En el nou model queda definida la matriu que explica la variància del model mitjançant la suma de la variància residual i la variància entre subjectes que actua de covariància.

En aquest cas es dona el que es diu una estructura de simetria composta (CS) ja comentada en l'apartat 2.3.3 Modelació de la matriu de covariància i que considera la variància de cada repetició la mateixa i la covariància entre qualsevol parell de mesures igual.

Com comenta Everitt [9] en aquest punt el model encara es situa en la situació que tant pel cas d'estabilitats com pel cas de longitudinals amb mesures repetides pot distar de la realitat, ja que té més lògica que dos valors més propers estiguin més altament correlacionats en aquests casos. Per acostar-nos més a la realitat del que pot passar realment amb les observacions i amb una estructura de variància-covariància més realista és recomanable considerar el model amb intercepcions i pendents aleatoris que es mostra en el següent apartat.

3.2.5.2 Model amb intercepció i pendent aleatòries

Basat també en el model explicat per Everitt/Hothorn [9] i Weiss [6] s'explica aquesta segona matriu de variàncies-covariàncies:

En aquest cas el model que es dóna és de la següent forma:

- intercepció i pendent aleatòria:

$$y_{ij}=(\beta_0+u_{i1})+(\beta_0+u_{i2})x_j+\epsilon_{ij}$$ *Eq 3.2.5.2.-a*

D'aquest segon model s'assumeix les dos efectes aleatoris amb una distribució normal bivariada amb una mitjana de zero i variàncies de σ_{u1}^2 i σ_{u2}^2 , i covariància de σ_{u1u2} essent aquesta vegada el total de la part residual $u_{i1}+u_{i2}x_j+\epsilon_{ij}$ amb variància total ja no considerada constant per els diferents valors de covariant x:

$$Var(u_{i1}+u_{i2}x_j+\epsilon_{ij})=\sigma_{u1}^2+2\,\sigma_{u1u2}x_j+\sigma_{u2}^2x_j^2+\sigma^2$$ *Eq 3.2.5.2.-b*

En aquest cas la covariància entre els dos residus totals per un mateix lot/individu quedaria com:

$$Cov(u_{i1}+u_{i2}x_j+\epsilon_{ij}, u_{i1}+u_{i2}x_{j'}+\epsilon_{ij'}) = \sigma^2_{u1}+2\,\sigma_{u1u2}(x_j+x_{j'})+\sigma^2_{u2}x_j x_{j'}$$

Eq 3.2.5.2.-c

És a dir que no és el mateix per totes les parelles de repeticions (j j') ja que depèn del valor numèric de la covariant temps.

Tenint en compte l'equació ja vista de la matriu genèrica de variància de la resposta Eq 3.2.5.1.-g. En aquest cas l'expressió teòrica de la matriu de variàncies-covariàncies és més complicada ja que inclou més elements i en el cas de la matriu D que abans era la variància única d'un sol efecte aleatori aquí pren la forma real de matriu amb els valors de variància dels efectes de la intercepció i de la pendent (i la seva covariància):

$$D=\begin{pmatrix}\sigma^2_{u1} & \sigma_{u1u2}\\ \sigma_{u1u2} & \sigma^2_{u2}\end{pmatrix}=\begin{pmatrix}d_{11} & d_{12}\\ d_{12} & d_{22}\end{pmatrix}$$ *Eq 3.2.5.2.-d*

Aquesta estructura de la matriu D implica la consideració de que les intercepcions i pendents aleatòries tenen els efectes correlacionats. En els casos justificats o en la cerca del model òptim es pot considerar que no hi ha correlació entre aquests efectes i considerar la matriu D amb la següent forma:

$$D=\begin{pmatrix}\sigma^2_{u1} & \sigma_{u1u2}\\ \sigma_{u1u2} & \sigma^2_{u2}\end{pmatrix}=\begin{pmatrix}d_{11} & 0\\ 0 & d_{22}\end{pmatrix}$$ *Eq 3.2.5.2.-e*

En aquest últim supòsit les distribucions dels efectes serien independents i els càlculs de la matriu de variàncies-covariàncies entre diferents temps quedarien també modificades com es mostra en les següents variacions de les equacions Eq 3.2.5.2.-b i Eq 3.2.5.2.-c:

$$Var(u_{i1}+u_{i2}x_j+\epsilon_{ij}) = \sigma^2_{u1}+\sigma^2_{u2}x^2_j+\sigma^2$$ *Eq 3.2.5.2.-f*

$$Cov(u_{i1}+u_{i2}x_j+\epsilon_{ij}, u_{i1}+u_{i2}x_{j'}+\epsilon_{ij'}) = \sigma^2_{u1}+\sigma^2_{u2}x_j x_{j'}$$ *Eq 3.2.5.2.-g*

3.2.5.3 Model amb intercepció i pendent aleatòries i funció de modulació de la variància residual

Els dos models proposats en els dos apartats anteriors 3.2.5.1 i 3.2.5.2 assumeixen que en les diferents repeticions es manté la variància residual, és a dir que s'assumeix homoscedasticitat. En el present apartat s'actualitza el model afegint una funció que permet diferenciar variàncies residuals en diferents punts de temps de la covariant.

Galecki [10] suggereix en els casos d'heteroscedasticitat l'aplicació de les anomenades funcions de variància com alternativa més flexible al mètode dels

pesos coneguts de variància o *Known Variance Weights* que depèn de tenir un coneixement més ampli de les dades. Aquest tipus de funcions pretenen modular la variància residual afegint una funció addicional λ ampliant l'equació de la variància residual de la següent manera:

$$\lambda(\delta,\mu,\nu) \rightarrow Var(\epsilon_i) = \sigma^2 \lambda^2(\delta,\mu_i,\nu_i) \quad \textit{Eq 3.2.5.3.-a}$$

On δ sol ser un vector que conté una conjunt de paràmetres de variància comuns a totes les observacions, μ_i és el valor esperat de la resposta en aquella observació *i* i ν_i és un vector de covariants conegudes que defineixen la funció de variància. Dins les possibles funcions λ es fa una diferenciació important amb un tipus que no depèn de μ_i ja que aquestes darreres no suposen la complexitat d'estimació i tècniques d'inferència que suposen les de models que sí que la inclouen. Deixant com a única opció les que no depenen de μ_i a continuació es defineixen algunes de les més habituals i com afecten la matriu R_i essent la matriu de variància del residu $\epsilon_i \sim N(0, R_i)$ suposant que es pren habitualment el temps com a modulador de la variància:

- Funció *varIdent* (δ, s_i): La més simple i que simplement defineix diferents variàncies segons un estrat com pot ser el factor temps expressant-se de la següent manera:

$$\lambda_i(varIdent) = \delta_{s_i} \quad \textit{Eq 3.2.5.3.-b}$$
$$\lambda_1 \equiv 1, \lambda_s > 0 \; for \; s \neq 1$$

  És a dir que la variància comença en 1 i va augmentant a mesura que augmenta l'estrat escollit com per exemple el temps. En el cas de la matriu R_i aquesta funció pren els ràtios entre les variàncies segons el temps on es situa:

$$R_i = \sigma^2 \begin{pmatrix} \delta_1^2 & 0 & \dots & 0 \\ 0 & \delta_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \delta_i^2 \end{pmatrix} \quad \textit{Eq 3.2.5.3.-c}$$
$$\delta_t \equiv \sigma_t / \sigma_1 (t = 1 \dots i)$$

  Calculant la proporció de variància respecte al primer temps.

- Funció *varExp* (δ, ν_i, s_i): Funció calculada amb l'exponencial de la variància de la covariant:

$$\lambda_i(varExp) = \exp(\nu_i \delta_{s_i}) \quad \textit{Eq 3.2.5.3.-d}$$

En aquest cas en la matriu de variàncies residuals queden els elements en relació directa amb l'exponencial del factor numèric del temps:

$$R_i = \sigma^2 \begin{pmatrix} \exp(2\cdot\delta\cdot x_{i1}) & 0 & \dots & 0 \\ 0 & \exp(2\cdot\delta\cdot x_{i2}) & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \exp(2\cdot\delta\cdot x_{it}) \end{pmatrix} \quad \textit{Eq 3.2.5.3.-e}$$

- Funció *varPower* (δ, v_i, s_i): Funció calculada amb la potència de la variància de la covariant:

$$\lambda_i(varPower) = \left| v_i^{\delta_s} \right| \quad \textit{Eq 3.2.5.3.-f}$$

En aquest cas en la matriu de variàncies residuals queden els elements en relació directa amb el quadrat del factor temps per un corrector δ :

$$R_i = \sigma^2 \begin{pmatrix} (x_{i1})^{2\delta} & 0 & \dots & 0 \\ 0 & (x_{i2})^{2\delta} & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & (x_{in})^{2\delta} \end{pmatrix} \quad \textit{Eq 3.2.5.3.-g}$$

Aquesta modificació del paràmetre σ^2 de la variància residual també entraria en el càlcul de la variància i covariància entre temps que ja s'ha vist modificada degut a l'estructura de variància-covariància de intercepcions aleatòries, i intercepcions/pendents aleatòries. És a dir que les equacions Eq 3.2.5.1.-d i Eq 3.2.5.2.-b es substituiria la forma que fos per la part de la variància residual.

Tot i que Galecki [10] en els diversos exemples que mostra habitualment l'utilitza amb conjunció amb el model de intercepcions/pendents aleatòries, també podria ser possible aplicar-ho en el model d'intercepcions fent servir una variable diferent del temps per modular la variància, tot i que les funcions de variància es tornarien més simples al perdre alguns dels paràmetres de modulació.

3.3 Simulacions dels models d'efectes mixtos

Per la realització de les simulacions de models mixtos s'ha utilitzat el programari *Rstudio* [3] (treballant en el sistema operatiu *Linux* distribució *Ubuntu [23]*).

Per la simulació dels models mixtos es pren la suposició que es disposa de les dades de varis subjectes/lots que han acabat els seus estudis d'estabilitat i en base a aquests es vol construir un model que expliqui el seu comportament i serveixi com a model de predicció per futurs subjectes/lots, assumint que les dades dels lots es poden combinar.

Per facilitar l'anàlisi teòric de les simulacions s'ha pres models senzills amb una variable resposta, una covariant temps i una variable d'identificació de

lot/individu. S'ha considerat un rang en nombre de lots entre 3 i 10 i un rang en nombre de temps d'anàlisi entre 4 i 10. Per les simulacions s'ha pres majoritàriament un conjunt simulat en el valor promig d'aquests rangs.

A continuació es resumeixen les etapes del procés de la simulació i es mostren els resultats i figures més rellevants obtingudes. S'ha generat 3 grups de conjunts de dades o *datasets* amb diferents característiques i l'anàlisi complet dins de cada grup de *datasets* anomenats *Ho* (simulació del primer grup de datasets amb dades homogènies, no confondre amb el terme d'hipòtesis nul·la), *He* (simulació del segon grup de datasets amb dades heterogènies) i *So* (simulació del tercer grup de datasets simulació de pendent en les dades) respectivament.

3.3.1 Simulació Models mixtos Datasets Ho

Pels datasets *Ho* s'ha intentat simular conjunts de dades amb una resposta sense tendència significativa, amb relativa homogeneïtat i possible homoscedasticitat.

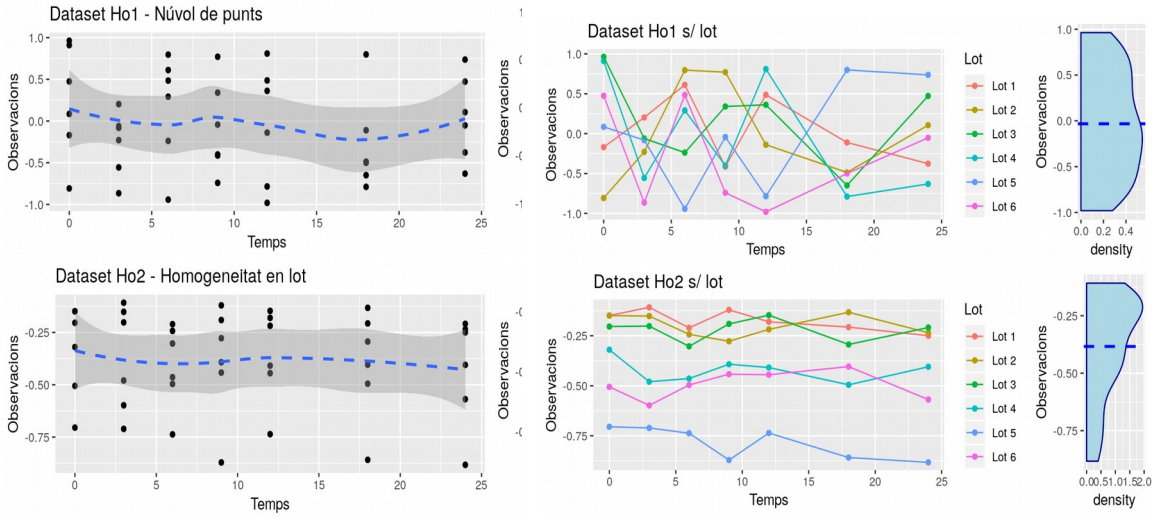
3.3.1.1 Generació i exploració inicial - Datasets Ho

Amb mètodes semialeatoris reproduïbles s'han simulat dos conjunts de dades:

- Ho1: Amb un núvol de punts entre un rang petit de possibles respostes.
- Ho2: En aquest cas s'ha generat aleatòriament les mitges de cada lot i s'han generat els punts de les diferents repeticions tenint en compte cada lot per separat amb un rang de resposta petit al voltant de cada mitjana.

A continuació es mostra visualment els *scatterplots* corresponents amb gràfics de densitats auxiliars de la generació d'aquests punts dels dos conjunts de dades amb dos variants diferents de visualització obtinguts:

Figura 1: Scatterplot/Density plot Ho amb i sense distinció de lot



Mentre que el primer gràfic de la Figura 1 no s'aprecia la diferència entre els dos conjunts, aquesta queda clara al diferenciar per lot com es veu en el segon gràfic de Figura 2, a part que en el cas de centrar per lot el gràfic de densitat es mostra més esbiaixat al dependre de la mitjana de cada lot amb un conjunt tant petit de subjectes.

3.3.1.2 Ajust i diagnòstic models - Datasets Ho

S'ha ajustat cada conjunt amb el model simple per l'ajust clàssic dels mínims quadrats utilitzant un model només amb intercepció (I) i un model amb intercepció i variable predictora temps (IS) per comparar-los amb els següents resultats:

Ho1: En el cas del model IS la prova de significació de la variable temps ha resultat no significativa amb un p-valor de 0.56.

Ho2: En el cas del model IS la prova de significació de la variable temps ha resultat no significativa amb un p-valor de 0.59.

Al comprovar que la variable predictora no tenia significació, és a dir que no tenia sentit el model amb pendent, pel següent pas s'ha ajustat el model I afegint l'efecte aleatori de l'intercepció depenent del factor lot. Per aquest ajust s'ha utilitzat l'ajust clàssic *ordinary least squares* (OLS) o també anomenat ANOVA, l'ajust MLE i l'ajust per REML, i en els models MLE i REML es realitza la prova d'hipòtesis de la significació de l'efecte aleatori mitjançant els test de ràtio de versemblança restringida o *restricted likelihood ràtio tests* (RLRT), obtenint els següents resultats:

Ho1: No s'han obtingut grans diferències entre el càlcul del model amb els 3 sistemes si es compara la desviació residual estàndard ni tampoc comparat amb les desviacions dels models més simples sense efectes aleatoris. La prova de significació RLRT ha donat no significativa en els 2 models MLE i REML amb un p-valor de 1.

Ho2: Igual que en Ho1 no hi ha diferències significatives entre els 3 sistemes d'estimació del model, però sí que s'observa una disminució possiblement significativa de la desviació residual comparat amb el model sense efecte aleatori RI. En aquest cas la prova de significació RLRT ha donat significativa en els 2 models MLE i REML amb un p-valor de 0.

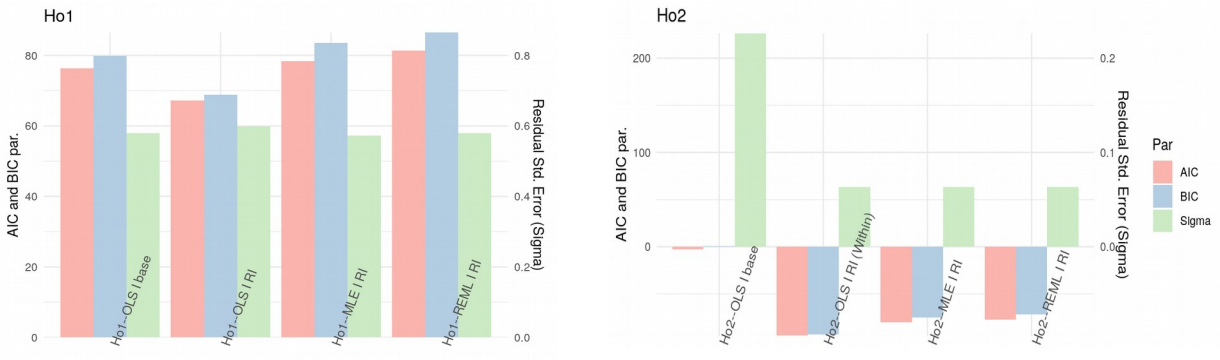
3.3.1.3 Resum - Datasets Ho

En el cas de tenir models ajustats amb diferents metodologies d'ajust i amb presència o no d'efectes aleatoris, complica la comparació clàssica de models ANOVA la qual no es pot aplicar de manera directa com a prova d'hipòtesi. Per compensar això s'ha fet la comparativa mitjançant el càlcul dels criteris d'informació AIC i BIC, i amb el càlcul de la desviació residual en forma de taula i de gràfic de barres per comparar els diferents models. Aquest mètode es té en

compte en totes les simulacions d'aquest treball pel que en endavant no es justificarà la seva utilització.

Es mostra gràficament la comparativa:

Figura 2: Barplot per comparació de models Ho



Es denoten aquí encara més les diferents característiques dels *datasets* que en els càlculs inicials dels models ajustats:

Ho1: S'observen resultats dels indicadors molt similars. A nivell de comparació visual podria semblar que el model ajustat per ANOVA clàssic seria el que minimitzaria els indicadors, mentre que els errors estàndards residuals queden molt igualats. No sembla haver una raó de pes per triar un o altre model.

Ho2: Tant l'error estàndard residual com els indicadors d'ajust del model AIC i BIC donen a entendre les diferències entre el model bàsic i els models amb efectes aleatoris que semblen ser clau per millorar l'ajust del model. En aquest cas concret no hi ha grans diferències entre els 3 models amb efecte aleatori RI, tot i que a la vista dels gràfics podria semblar que el millor model seria l'ajust pels mètodes clàssics amb efecte aleatori. Probablement al existir un patró que modula la localització dels punts per cada lot en Ho2, encara que els lots tinguin una ubicació aleatòria, l'efecte aleatori sembla que permet ajustar aquesta característica per augmentar la certesa del model.

Com a última prova rellevant s'ha obtingut la matriu D de l'efecte aleatori per cada model i el factor de correlació i s'ha pogut comprovar clarament que pel model Ho2 hi ha una proporció major de variància explicada per l'efecte aleatori i un grau molt més alt de correlació entre les mesures repetides. Els resultats complets es troben a 8.4 Annex 4: Resultats R de les matrius de variància covariància i matrius de correlació en la simulació de models mixtos.

3.3.1.4 Proves de variació de paràmetres - Datasets Ho

Addicionalment i per comprovar l'efecte del nombre de individus i nivells de la covariant temps s'ha fet una bateria de proves comprovant el rang total establert de lots i temps d'anàlisi en el model amb efectes aleatoris, i s'han representat en un tipus de gràfic anomenat *Heatmap*. Els resultats pels diferents models (OLS, MLE, REML) són semblants, per tant es mostra a

continuació per la correcció de biaix comentat en capítols anteriors el cas REML:

Figura 3: Heatmap Ho1 de sigma, AIC i BIC en el rang de lots i temps

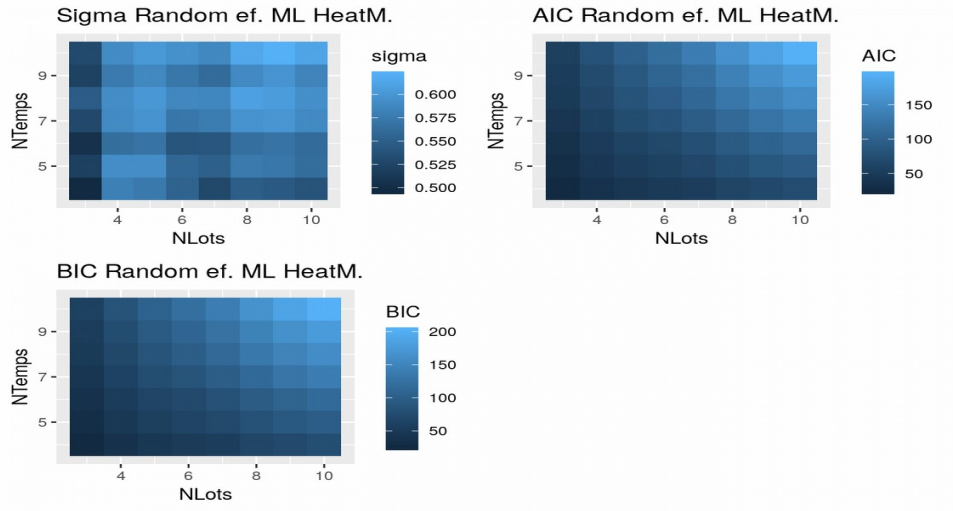
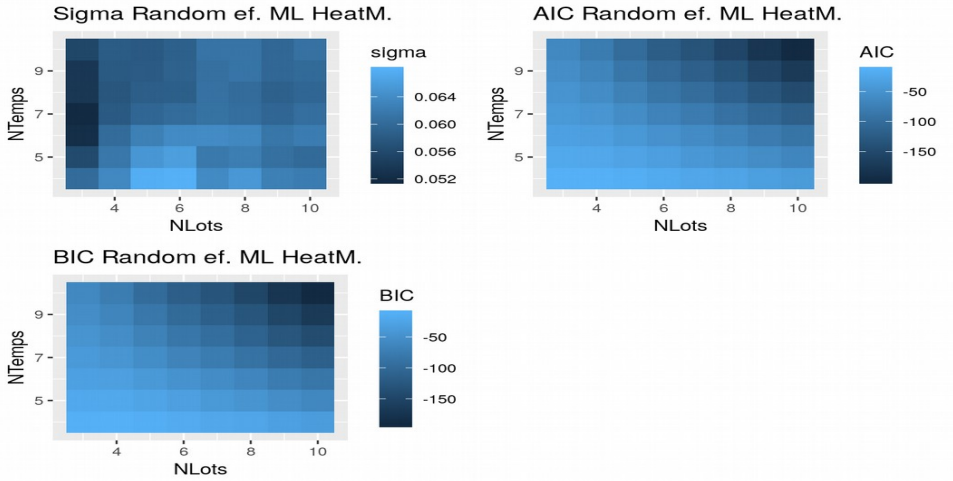


Figura 4: Heatmap Ho2 de sigma, AIC i BIC en el rang de lots i temps



S'observa que per l'error estàndard residual sembla no haver un patró concret per tenir una variància residual més baixa o més alta possiblement pel fet de treballar amb una simulació de dades tipus núvol de punts en el cas Ho1 o els lots en posicions aleatòries en Ho2. Sí que s'observa clarament com s'inverteix el comportament pel cas dels AIC i BIC segons el tipus de *dataset* que s'està tractant:

Ho1: Al contrari del que habitualment s'esperaria, quan el model disminueix la mida tant en lots com en temps es calcula uns factors més baixos i per tant suposaria un millor ajust per les dades. Probablement es deu també al tipus de simulació de *dataset* utilitzat on un major nombre de punts simplement afegeix més variància al tractar-se d'un núvol de punts.

Ho2: AIC i BIC disminueixen significativament quan s'augmenta el nombre de lots o de punts d'anàlisi possiblement degut a que un major nombre de mostres

permet ajustar amb més precisió l'efecte aleatori provocat per la variació dels lots.

3.3.2 Simulació Models mixtos Datasets He

Pels *datasets He* s'ha intentat simular conjunts de dades sense tendència significativa, amb heterogeneïtat significativa i possible heteroscedasticitat.

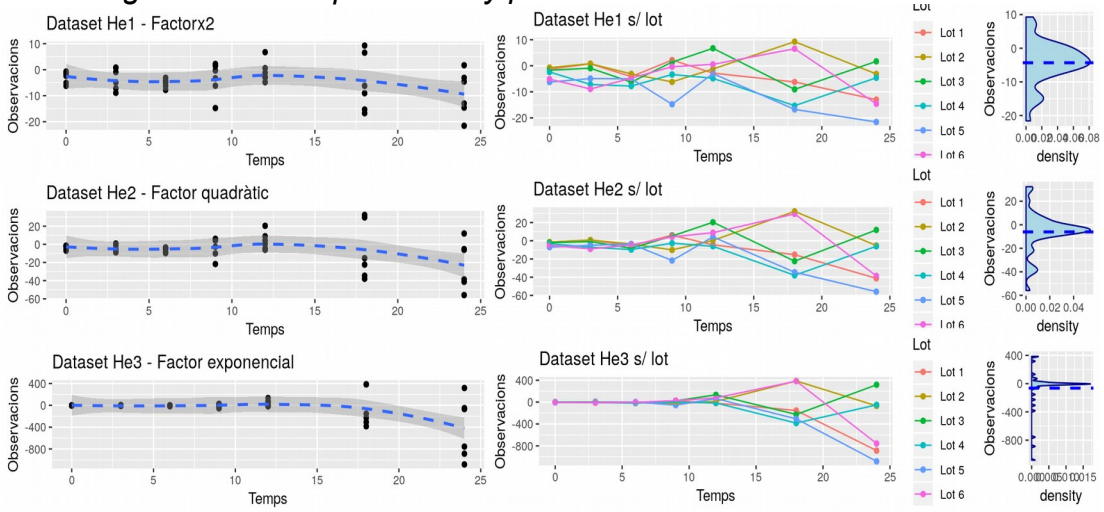
3.3.2.1 Generació i exploració inicial - Datasets He

Amb mètodes semialeatoris reproduïbles s'han simulat tres conjunts de dades intentant simular l'augment de variància en el temps per aconseguir l'efecte de l'heteroscedasticitat. Els tres conjunts s'ha modulat l'augment de variància de diferent manera:

- He1: Augment de rang de valors possibles amb un factor multiplicador de 2 en el temps.
- He2: Augment de rang de valors possibles amb un factor de la potència del temps al quadrat.
- He3: Augment de rang de valors possibles amb un factor del factor exponencial del temps.

A continuació es mostren visualment els *scatterplots* corresponents amb gràfics de densitats auxiliars de la generació d'aquests punts dels tres conjunts de dades amb dos variants diferents de visualització obtinguts:

Figura 5: Scatterplot/Density plot He amb i sense distinció de lot



En la Figura 6 queda clar com va augmentat la variància a mesura que s'avança en el temps. Addicionalment es veu com es comporta el gràfic de densitat on amb un major factor d'augment de variància es denoten més les diferències entre la part concentrada i la part dispersa de dades.

3.3.2.2 Ajust i diagnòstic models - Datasets He

S'ha ajustat cada conjunt amb el model simple per l'ajust clàssic dels mínims quadrats utilitzant un model I i un model IS amb els següents resultats:

He1: En el cas del model IS la prova de significació de la variable temps ha resultat no significativa amb un p-valor de 0.105.

He2: En el cas del model IS la prova de significació de la variable temps ha resultat no significativa amb un p-valor de 0.056.

He3: En el cas del model IS la prova de significació de la variable temps ha resultat significativa amb un p-valor de 0.008.

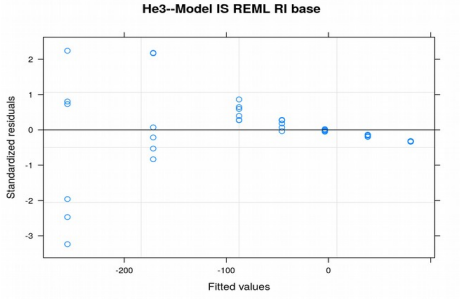
En aquest cas no queda tan clar que es pugui descartar la variable predictora pel que, pel següent pas s'ha ajustat tant el model I com el model IS afegint l'efecte aleatori RI depenent del factor lot. Per aquest ajust s'ha utilitzat com abans OLS, MLE i REML, i en els models MLE i REML es realitza la prova RLRT amb els següents resultats:

Dataset	p-valor RLRT
He1 I RI MLE	0.033
He1 I RI REML	0.0296
He1 IS RI MLE	0.0257
He1 IS RI REML	0.0257
He2 I RI MLE	0.2656
He2 I RI REML	0.2431
He2 IS RI MLE	0.2143
He2 IS RI REML	0.2002
He3 I RI MLE	1
He3 I RI REML	1
He3 IS RI MLE	1
He3 IS RI REML	1

Donaria la sensació pels resultats obtinguts que en els conjunts on augmenta més ràpid la dispersió també tendeix a tenir els p-valors de la provar RLRT més alts i per tant menys significació en l'efecte aleatori. Per altra banda en aquest contrast no es detecten diferències a aparents entre utilitzar la covariant o no.

Mitjançant la prova de visual dels gràfics de residus s'ha pogut comprovar l'heteroscedasticitat prevista dels models, la qual és major com més ràpid és l'augment d'heterogeneïtat com es pot veure d'exemple a la Figura 6 pel model REML i el conjunt He3:

Figura 6: Prova d'homoscedasticitat residual amb residus estandarditzats He3

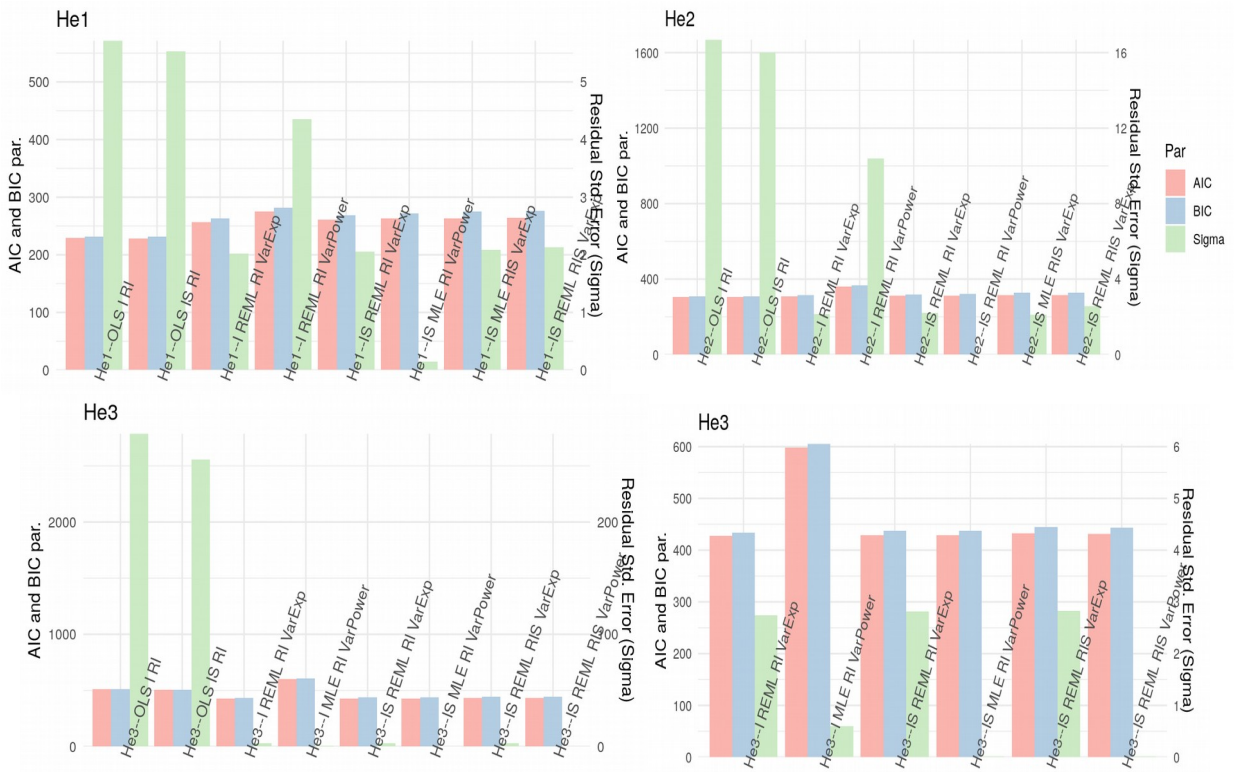


Per aquest motiu s'ha fet un pas més i s'han ajustat els conjunts a models amb les funcions per modular la variància de l'error. A nivell de programació i d'anàlisi com s'ha comentat s'ajusten els models MLE i REML. Les funcions que s'han aplicat son les varIdent, varExp i varPower depenent del factor temps sempre que ha estat possible (en casos on els algoritmes no han convergit en cap solució s'ha omès el model). Per mesurar l'efectivitat d'afegir aquestes funcions no s'ha utilitzat cap mètode directe si no que s'ha pres els indicadors que es veuen en el proper apartat 3.3.2.3.

3.3.2.3 Resum - Datasets He

Veiem els *barplots* comparatiu AIC/BIC i desv. std residual:

Figura 7: Barplot per comparació de models He



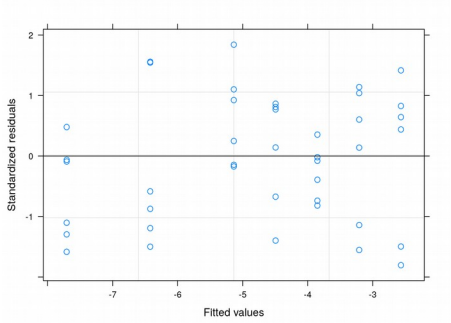
Es denoten aquí encara més les diferents característiques dels *datasets* que en els càlculs inicials dels models ajustats:

He1/He2: Tot i que s'observen resultats dels indicadors AIC/BIC molt similars, els errors estàndards residuals sembla que es redueixen significativament en els models on s'afegeix la funció de modulació de variància, i especialment en el model MLE IS RI amb funció de modulació *varPower*.

He3: Aquest cas s'ha vist una diferència més gran en els errors estàndards residuals, pel que es mostra el gràfic centrat en els models amb modulació de variància (Figura 23) on s'aprecia que el millor cas semblaria ser el modulat per *varPower*.

S'ha realitzat el diagnòstic d'homoscedasticitat veient que al aplicar les funcions RI habitualment millora tan la normalitat com l'homoscedasticitat. Es mostra un dels gràfics de millora d'exemple a la Figura 8:

Figura 8: Prova d'homoscedasticitat residual res. estandarditzats He3 VarPower al model IS MLE RI



En les extraccions de matrius de variància-covariància residual i correlacions dels models on s'han aplicat les funcions s'ha observat com es modula la variància en cada punt de manera diferent i també com es correlacionen les repeticions de manera diferent segons la relació. Els resultats d'aquestes matrius es troben a 8.4 Annex 4: Resultats R de les matrius de variància covariància i matrius de correlació en la simulació de models mixtos.

Dels resultats obtinguts (els principals i en alguns casos d'exemple exposats aquí) es poden extreure algunes reflexions:

Al haver simulat una variància que varia amb el temps al voltant de la mateixa mitjana s'esperaria que el predictor Temps no fos significatiu per explicar la mitjana de la resposta. Tot i així, en algun dels models aparentment dona significació. Possiblement al tenir un nombre de lots limitat, provoca que l'augment de variabilitat emmascari la mitjana real sobre la qual es reparteixen i dona un fals efecte de tendència. És per aquest motiu probablement que s'observa que afegir la variable temps al model pren més significació al tenir un model amb un factor d'augment de variància més gran.

En aquest cas no seria tan senzill com descartar el model amb l'efecte fix de covariància del predictor Temps ja que aquest fals efecte de tendència pot servir també per millorar el model i veient els resultats no quedaria clar ja que:

- En la majoria dels models tot i veure que augmenta la significació del predictor, aquest no té la força suficient per provocar un canvi molt significatiu si es comparen els errors estàndards residuals dels mateixos *datasets* en models amb i sense la pendent S, cosa que ens donaria pistes en cas que fos un *dataset* desconegut de que hi ha algun factor addicional amagat en el model que no és un factor de tendència.
- En el cas concret dels models amb *varPower* es nota la diferència al afegir la variable temps a la part d'efectes fixos en relació al model on només compte la intercepció com efecte fix.

En general l'estimació i errors residuals augmenten a mesura que s'estudien models més dispersos ja que van lligats a la incertesa de predicció de les dades.

Segons el conjunt no s'ha pogut aplicar la *varPower* a totes les variants del model ja que en algun cas s'arribava a resultats inconsistents, tot i així no es denota una millora significativa al modificar els efectes aleatoris de la intercepció per efectes aleatoris RIS en aquests casos, que ja quadraria amb els conjunts simulats. S'entén que la importància d'afegir la pendent en els efectes fixos en conjunts petits depèn molt de com es comporten els lots.

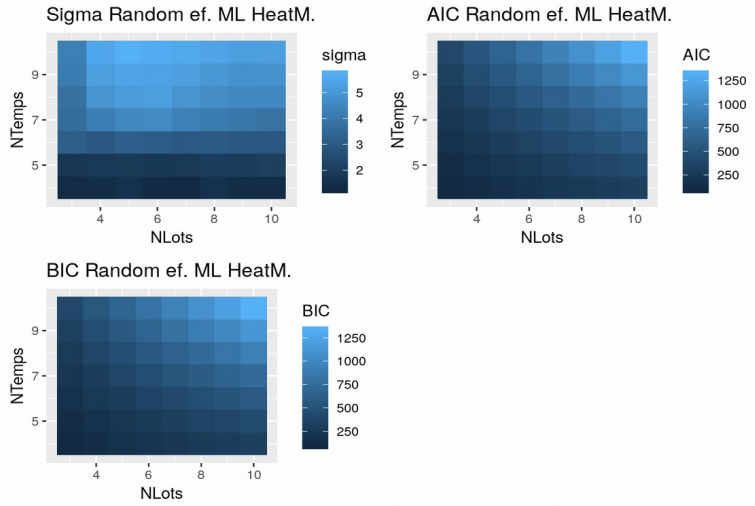
Pel que fa a les matrius variància i covariància es veu com la funció addicional recalcula de manera que a mesura que s'avança en els punts de temps augmenta l'error estàndard propi del temps (i si el conjunt és més dispers també es calculen errors més grans). També es calcula la correlació de variàncies segons quin temps es relaciona amb quin, però s'han obtingut baixes en general. Probablement perquè les funcions aplicades no han aconseguit un model amb alta eficàcia.

A la vista dels resultats i observacions, es podria deduir que en aquest tipus de conjunts de dades es considera important aplicar la funció de modular l'error estàndard i en cas necessari per compensar la mida dels lots aplicar el factor de la pendent, tot i que en aquest aspecte s'ha d'anar amb compte si es vol utilitzar com un model de predicció per futurs lots.

3.3.2.4 Proves de variació de paràmetres - Datasets He

Es mostra en aquest cas els resultats de la variació dels paràmetres de lots i temps únicament en el model He3 REML IS RI varExp (per evitar els problemes de convergència del varPower) ja que les tendències són semblants en tots els models:

Figura 9: Heatmap He3 de sigma, AIC i BIC en el rang de lots i temps



A part de les diferències ja vistes entre els 3 *datasets* que provoquen l'augment de dispersió de dades en aquest cas s'observa que:

Pels paràmetres indicatius de la qualitat del model AIC i BIC sembla que al tenir més lots/Temps augmenta aquest paràmetre disminuint l'eficàcia del model. Possiblement l'efecte de dispersió augmenta al afegir més lots i temps en aquest tipus de conjunts i per tant fa més difícil l'ajust d'un model eficient amb els recursos que s'han utilitzat fins ara.

El paràmetre de l'error estàndard residual semblaria que segueix una patró similar a AIC i BIC, però és curiós com sembla que pren més importància el paràmetre del nombre de temps. És a dir al variar el nombre de lots sembla no haver un rang molt gran de canvi mentre que en el nombre de temps sembla haver-hi dos zones molt diferenciades amb sigmes alts i baixos.

Tot i que en aquest apartat es podria treure la conclusió que és millor tenir un nombre més limitat de lots i temps, en la realitat sempre és millor disposar del màxim de dades possibles ja que habitualment les dades d'un mateix lot no es comportaran amb una dispersió tan exagerada com s'ha simulat en aquests models. Si es donés aquest cas amb aquest diagnòstic, potser s'hauria de començar a sospitar que hi ha algun factor que no s'ha tingut en compte o algun error en les dades o en la metodologia.

3.3.3 Simulació Models mixtos Datasets So

Pels *datasets So* s'ha intentat simular conjunts de dades amb tendència significativa i amb diferent grau d'homogeneïtat de pendents. No es té en compte el grau d'homogeneïtat i es centra la simulació en la fórmula que correspongui a la tendència.

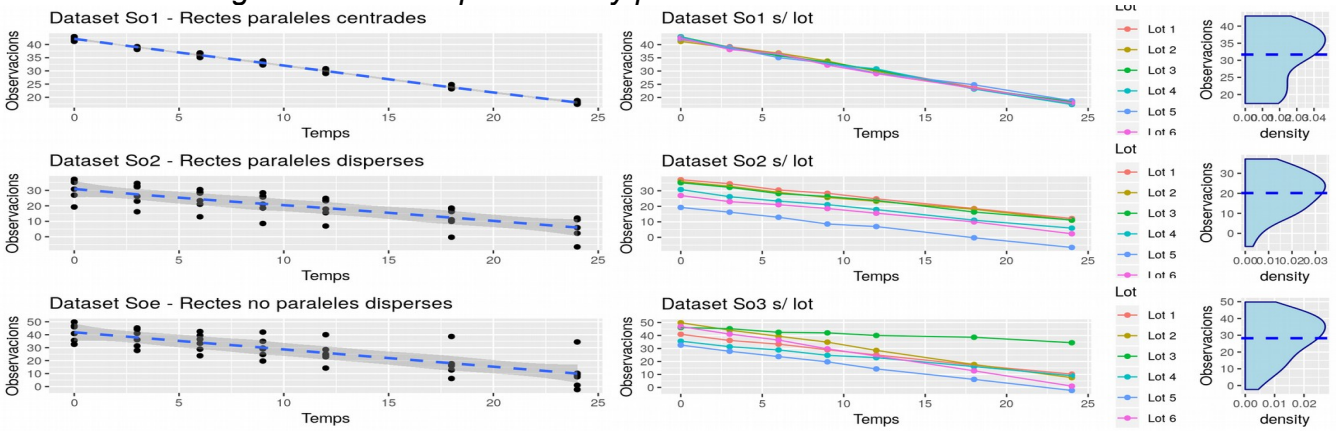
3.3.3.1 Generació i exploració inicial - Datasets So

Amb mètodes semialeatoris reproduïbles s'han simulat tres conjunts de dades intentant simular diferents casos en els models amb tendència:

- So1: Tendència concreta amb els punts centrats en una recta concreta amb poca dispersió.
- So2: El mateix cas que l'anterior, però suposant que hi ha dispersió més gran entre els lots.
- So3: Certa dispersió en la intercepció dels lots i una variabilitat en la pendent segons el lot dins un rang definit.

A continuació es mostren visualment els *scatterplots* corresponents amb gràfics de densitats auxiliars de la generació d'aquests punts dels tres conjunts de dades amb dos variants diferents de visualització obtinguts:

Figura 10: Scatterplot/Density plot So amb i sense distinció de lot



A la Figura 10 queda clar com és la tendència a mesura que s'avança en el temps. Addicionalment es veu com es comporta el gràfic de densitat sense aportar possiblement informació molt rellevant.

3.3.3.2 Ajust i diagnòstic models - Datasets So

S'ha ajustat cada conjunt amb el model simple per l'ajust OLS utilitzant un model només I i un model IS amb els següents resultats:

So1: En el cas del model IS la prova de significació de la variable temps ha resultat significativa amb un p-valor de 0.

So2: En el cas del model IS la prova de significació de la variable temps ha resultat significativa amb un p-valor de 0.

He3: En el cas del model IS la prova de significació de la variable temps ha resultat significativa amb un p-valor de 0.

Per començar s'observa que, com és d'esperar, afegir la variable temps al model és una diferència significativa i una suposada millora al model per explicar la resposta.

El següent pas que s'ha calculat és habitual en aquests anàlisis i és el càlcul de si l'efecte d'interacció lot/temps és significatiu per saber si les pendents són equivalents. S'han obtingut com era d'esperar resultats no significatius als models So1 i So2 (p-valor > 0.05) i significatiu al So3 (p-valor de 0).

En aquests models un dels contrastos que han estat d'interès i relacionat amb el paràgraf anterior ha estat la prova ANOVA diferenciant els models amb efectes aleatoris RI i RIS per veure la significació d'afegir l'efecte aleatori de les pendents obtenint els resultats de p-valors no significatius en el cas de So1 i So2 (p-valor > 0.05) i significatiu en el cas So3 (p-valor de 0).

Finalment al intentar ajustar amb funcions de modulació de variància els resultats ANOVA calculats han donat un resultat no significatiu en tots els casos, pel que en principi no seria significatiu afegir aquest terme a aquestes simulacions.

3.3.3.3 Resum - Datasets So

Veiem els *barplots* comparatius AIC/BIC i desv. std residual:

Figura 11: Barplot per comparació de models So1-So2

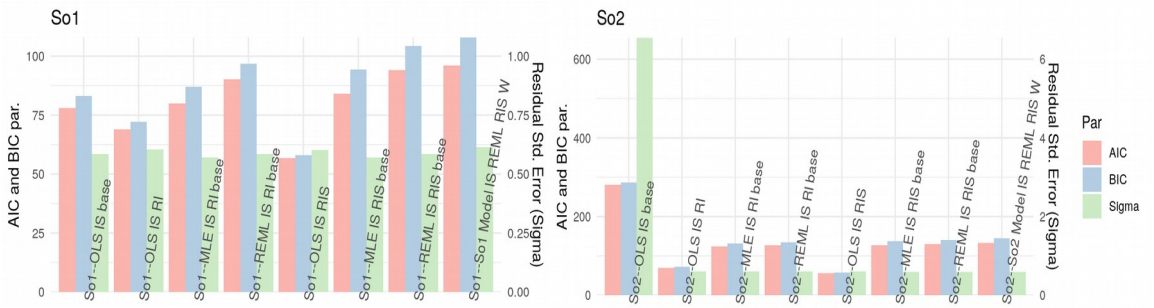
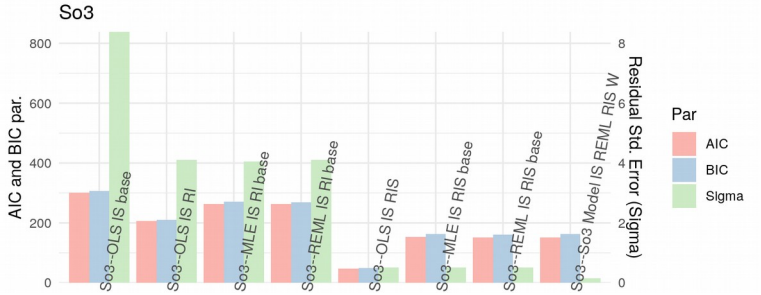


Figura 12:Barplot per comparació de models So3



Els indicadors donen prova del que s'ha anat veient en els diferents models:

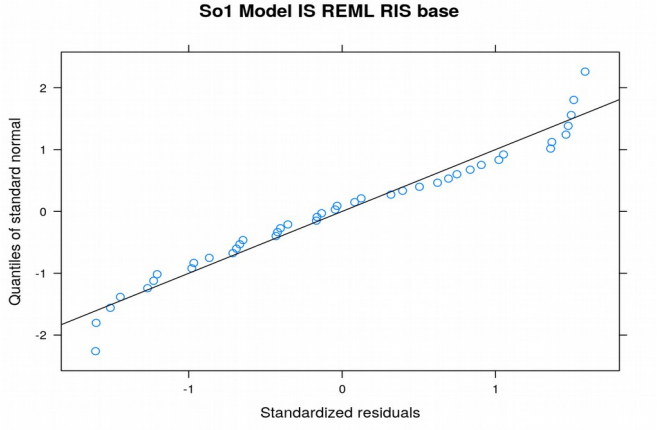
So1: Els models tenen poca diferència de qualitat en variància residual. Possiblement les diferències en AIC/BIC són per la proximitat de les distribucions a la normalitat.

So2: S'aprecia clarament tant en els indicadors AIC/BIC com l'error estàndard residual la reducció al afegir el factor aleatori. Dins els models amb factor aleatori en canvi no s'aprecia diferència entre si és l'efecte RI o RIS. Finalment les diferències amb AIC/BIC entre el tipus de càlcul de model s'observa possiblement pel compliment o no de normalitat.

So3: Disminució significativa al afegir els models amb efecte RI i encara més significativa al afegir els RIS. Quedaria bastant clar quin és el camí per enfocar aquest tipus de models amb efectes aleatoris. Tot i l'aparent millora que afegeix el terme varPower s'estudiarà cada cas concret si realment aquesta funció dona una millora a l'ajust del model.

Es destaca en aquest cas la posterior prova de normalitat realitzada al conjunt So1 sense veure una exagerada asimetria o kurtosis, però que podria ser l'origen de les petites diferències comentades en els termes AIC/BIC de la comparació de models (es visualitza el cas REML del model IS RIS):

Figura 13: Prova de normalitat al conjunt So1 aplicat al model IS RIS REML



En les extraccions de matrius de variància-covariància residual i correlacions dels models on s'han aplicat les funcions s'ha observat diferents resultats en els tres conjunts:

En el primer conjunt So1 sembla tenir poca significació la covariant corresponent a l'efecte aleatori i de fet la correlació entre els temps és molt baixa, propera a zero. Contràriament en els conjunts So2 i So3 augmenta considerablement l'efecte aleatori i també es reflexa en la correlació entre els temps que és relativament gran. Aquest fet en principi seria causat pel fet que al primer conjunt possiblement no és necessari l'efecte aleatori per explicar el model al tenir un model molt centrat i paral·lel, cosa que no passaria en els models So2 i So3 on (de diferent manera) hi ha una major dispersió que requeriria de l'efecte aleatori per ser explicada.

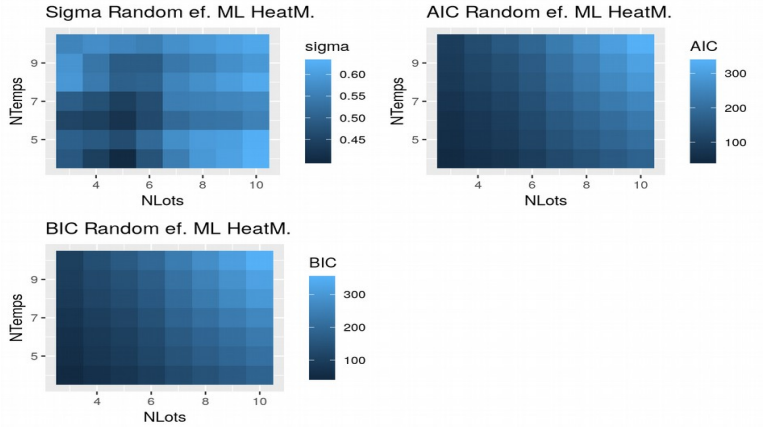
Tot i que en el conjunt So1 i So2 no varia la covariància i correlació entre temps, sí que ho fa en el conjunt So3, pel que en aquest cas sí que la funció ha aconseguit trobar certa correlació per poder reflectir en aquesta matriu. En els conjunts So1 i So2 la variació s'entén que és de intercepció i per tant no hi hauria diferències segons el temps. El conjunt So3 on s'aplica també l'efecte aleatori de la pendent dona per fet que els diferents temps estaran correlacionats per poder construir la pendent que toqui en cada moment (de fet ja es veu l'estructura típica on els temps més propers estan més correlacionats que els temps més llunyans i que coincidiria amb un comportament habitual dels models en els estudis d'estabilitat).

Els resultats complets d'aquestes matrius es troben a 8.4 Annex 4: Resultats R de les matrius de variància covariància i matrius de correlació en la simulació de models mixtos.

3.3.3.4 Proves de variació de paràmetres - Datasets So

Es mostra en aquest cas els resultats de la variació dels paràmetres de lots i temps únicament en el model So3 REML IS RI varExp (per evitar els problemes de convergència del varPower) ja que les tendències són semblants en tots els models:

Figura 14: Heatmap So3 de sigma, AIC i BIC en el rang de lots i temps



A part de les diferències ja vistes entre els 3 datasets en aquest cas s'observa que:

Pels paràmetres indicatius de la qualitat del model AIC i BIC sembla que al tenir més lots/Temps augmenta aquest paràmetre disminuint l'eficàcia del model. Possiblement com en altres conjunts l'efecte de dispersió augmenta al afegir més lots i temps i per tant fa més difícil l'ajust d'un model eficient amb els recursos que s'han utilitzat fins ara, tot i així sempre és millor tenir més lots en aquests casos de mides petites.

El paràmetre de l'error estàndard residual no sembla seguir un patró concret, probablement degut a que es dóna una mida mostral no molt gran i el càlcul d'aquests varia molt segons quines dades es posen o treuen.

3.4 Models amb dades perdudes

Tot i haver un ventall gran de possibilitats d'aplicació de tècniques per models amb dades perdudes, en el cas dels models d'interès pel projectes s'analitzen de manera menys complex i pels casos més habituals que puguin aparèixer en el context dels estudis d'estabilitats.

3.4.1 Condicions anàlisi teòric

Per ara és important concretar entre els coneixements vistos en el capítol anterior l'apartat 2.4 Models longitudinals amb dades perdudes per determinar les condicions a tenir en compte en el present anàlisis teòric:

- Tipus de dades perdudes: El cas dels valors perduts per absència de dades d'alguna repetició seria el cas més habitual que es podria donar i el que es considera pel present anàlisi.

- Mecanismes de pèrdua de dades: Els estudis d'estabilitat habitualment es realitzen amb un disseny previ i seguint un protocol o procediment determinat que permet realitzar l'estudi. Com que els casos o individus solen ser els lots en estudi o alguna altra condició com la presentació, no s'espera que hi hagi un mecanisme concret o un factor ocult que determini quan es produiran valors perduts ja que en la majoria (per no dir tots) els casos són per errors del protocol/procediment o errors en el seguiment del protocol/procediment, pel que es pren com a mecanisme característic les dades perdudes a l'atzar completament MCAR.

  En casos específics es podria tenir un mecanisme que no fos MCAR al dependre d'alguna variable real com el tipus de lot, on es situa temporalment el temps d'anàlisi o l'auxiliar que gestiona la mostra, però en el present projecte no es contemplen al no ser els casos habituals, tot i que serien interessants per futurs estudis relacionats amb dades perdudes.

  Per donar exemples utilitzant els casos més habituals:

  - Identificació de mostres: Quan el laboratori d'anàlisi rep les mostres identificades incorrectament moltes vegades lligat als problemes de creuament de mostres. Excepte en els casos on es pugui utilitzar algun mitjà que et permeti reidentificar amb total fiabilitat les mostres, es solen anul·lar els resultats obtinguts per no afegir possibles resultats falsos a l'estudi.
  - Errors en el sistema de gestió de l'estudi: Habitualment errors que provoquen la no realització de l'anàlisi en algun dels temps del disseny de l'estudi.
  - Pèrdua o ruptura de mostra: Si no es detecta o es produeix a temps és possible que hagi passat massa temps per poder repetir l'anàlisi ja que en els estudis d'estabilitat respectar el factor temps és clau per tenir un disseny robust.
- Balanceig de dades: Les dades perdudes dels estudis d'estabilitat solen coincidir amb la pèrdua del balanç de dades en el conjunt, pel que com es comentava en el capítol anterior pot ser d'interès veure les diferències en el càlcul de sumes de quadrats i contrastos d'hipòtesi.
- Mecanismes de solució per les dades perdudes: Com s'ha comentat es tracta el tema de manera més simple, pel que s'opta per tenir en compte únicament els següents mètodes:
  - Veí més proper
  - Regressió estocàstica bayesiana

3.4.2 Anàlisi de l'efecte de tenir dades perdudes

Abans d'abordar les solucions a la pèrdua de dades, és igualment important o inclús més, calcular l'efecte estadístic real de l'absència d'aquestes dades ja que, en cas que no tinguin una afectació real o significativa pot significar no haver d'aplicar cap solució o optar per les solucions més simples.

Com comenta Graham [24] hi ha varis mètodes per estimar l'efecte i biaix que pot produir aquest fet. En conjunt es recomana realitzar els anàlisis anomenats de sensibilitat tenint en compte:

- Percentatge de pèrdua de dades: El càlcul simple del percentatge de respostes perdut per veure l'efecte a nivell de proporció de dades.
- Estimació correlació ZR: Utilitzat habitualment amb dades longitudinals on s'utilitza variacions seqüencials en els predictors per veure el canvi en el coeficient de determinació $R^{2.}$
- Anàlisis gràfiques: Les anàlisis gràfiques dels models permeten veure de manera més clara les diferències entre els resultats del model segons l'efecte de l'absència de dades.

Tot i que l'anàlisi depèn molt de cada cas específic, si es focalitza en el cas del context d'estudis d'estabilitat es poden establir alguns objectius d'anàlisi lleugerament més acotats addicionals o complementaris als ja comentats:

- Centrant-nos en els models de predicció a desenvolupar, pot ser un indicador també com es comporten els indicadors qualitatius tinguts en compte de AIC, BIC i variància residual en relació al model complet.
- Habitualment dels models de predicció sol interessar la mitjana i els intervals de confiança a un temps concret (temps de caducitat del producte) pel que es pot fer una comparativa gràfica i calculada de mitges i intervals de confiança (CI).
- En el context dels models generats amb els efectes aleatoris podria ser d'interès veure la variació de paràmetres propis d'aquests models com és la matriu de variància-covariància i la matriu de correlació entre mesures repetides, així com la matriu de relació entre els efectes aleatoris quan apliqui.

La quantitat d'anàlisis a fer per calcular l'efecte de la pèrdua de dades és molt gran, pel que només s'han escollit en aquest apartat els casos que poden interessar més en el context del tipus d'estudi que es pretén modelitzar.

En el context de la simulació amb R, el mateix programa proporciona varis paquets especialitzats per l'anàlisi de dades perdudes que es poden utilitzar en l'anàlisi teòric presents a partir de la simulació de dades a part dels punts ja proposats.

3.4.3 Imputació de dades

Els mecanismes que es proposen tenir en compte per l'anàlisi teòric són alguns mètodes d'imputació simple i altres d'imputació múltiple:

- Veí més proper: Hi ha dos variants, utilitzar el veí més proper o utilitzar un algoritme de classificació habitualment utilitzat per conjunts multivariants i anomenat també k-nearest neighbours o kNN que el que fa és, per qualsevol punt x (en el nostre cas de temps) que es vol estimar, observa les k respostes dels punts veïns més propers i pren una mitjana per obtenir una estimació funció d'aquests punts $f(x_{k1}; x_{k2};..)$.
- Regressió estocàstica bayesiana: Sol donar millor resultat que fiar-se de la regressió sense el valor en sí, ja que té en compte un factor d'error per estimar més correctament el valor perdut. De fet el terme estocàstic és el que fa referència a la incertesa en la predicció mentre que el terme bayesià es refereix a una metodologia concreta que es basa en les probabilitats subjectives.

Per la major part de la simulació es segueix Kowarik [25] per la imputació de variables amb VIM.

3.5 Simulacions de models amb dades perdudes

Per la realització de les simulacions de models mixtos s'ha utilitzat el programari *Rstudio [3]* (treballant en el sistema operatiu Linux distribució *Ubuntu [23]*.

Per la simulació dels models amb dades perdudes s'ha utilitzat els models generats en la simulació de models mixtos (apartat 3.3) i sobre aquests s'ha fet el forçat manual de generar les dades perdudes. Per fer això es tenen en compte tres nivells segons el % de dades perdudes (20%, 40% i 60%) i en base a cada nivell es generen dos vectors, un a nivell de lots i l'altre de temps d'anàlisi (TA) amb els punts a eliminar escollits de manera semialeatòria i reproduïble.

A continuació es resumeixen les etapes del procés de la simulació i es mostren els resultats i figures més rellevants obtingudes. El codi de programació s'ha dividit en 3 etapes un cop generats els *datasets*, la visualització i anàlisi preliminar, la modelització amb el càlcul dels efectes de les dades perdudes i una petita prova d'imputació de dades.

3.5.1 Anàlisi i visualització dels conjunts de dades perdudes

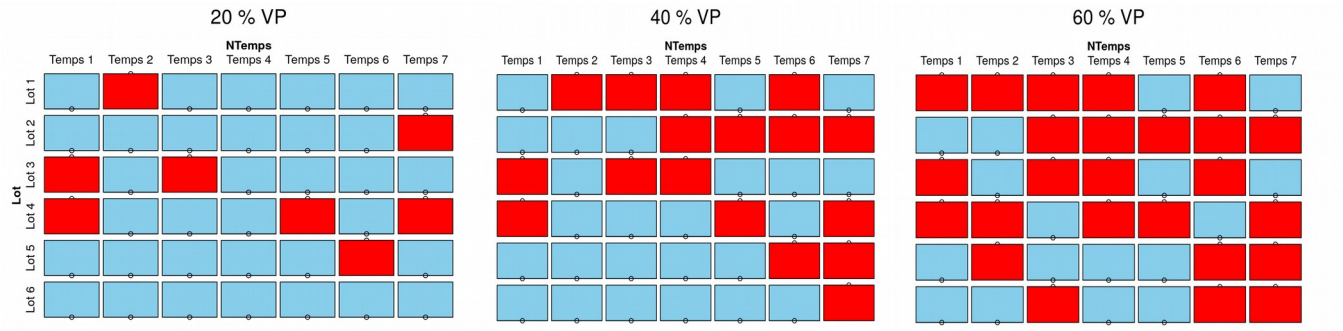
Per l'aplicació de la generació de dades perdudes simulades s'ha utilitzat el mateix vector de coordenades a generar les dades perdudes per tots els conjunts. Per aquest motiu hi ha una part de la visualització i anàlisi inicial que és genèric per tots els conjunts descriptiu de les dades perdudes que contenen i un altre específic que mostra dins el conjunt concret les dades perdudes.

3.5.1.1 Anàlisi general de dades perdudes

Els següents gràfics mostren les visualitzacions més rellevants en quant a proporcions en general que s'apliquen de dades perdudes als punts sense tenir en compte a quina variable s'aplica. Encara que es vegi aplicat a Ho1 com a conjunt per defecte, són els mateixos per la resta de variables.

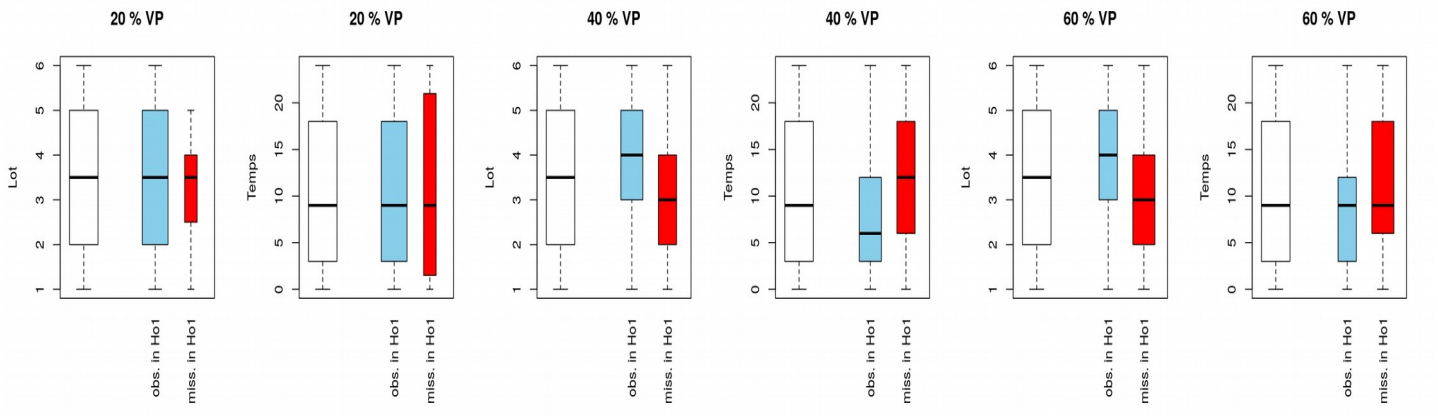
En el gràfic de mosaic s'aprecia per les dos variables temps i lot a quins punts s'han generat les dades perdudes pels diferents nivells:

Figura 15: Gràfic tipus mosaic de dades perdudes pels diferents nivells de pèrdua de dades



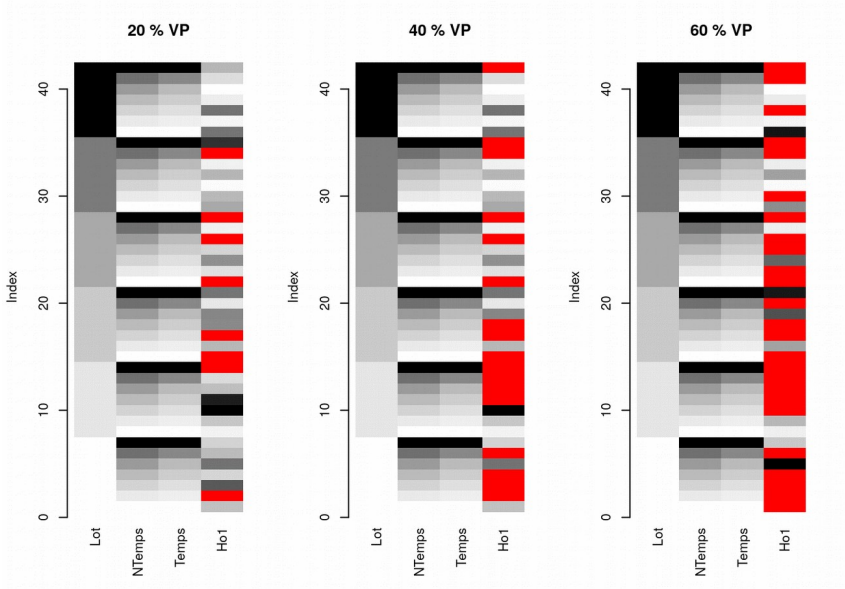
Els gràfics de caixa també són interessants al mostrar com es reparteixen tant pel factor lot com pel factor temps les dades perdudes:

Figura 16: Gràfic tipus caixa de dades perdudes pels diferents nivells de pèrdua de dades en front de cada variable Lot / Temps



Finalment en els gràfics genèrics també és interessant el gràfic anomenat *Matrixplot* on es veu també com està repartida la localització de les dades perdudes:

Figura 17: Gràfic tipus matriu de dades perdudes pels diferents nivells de pèrdua de dades



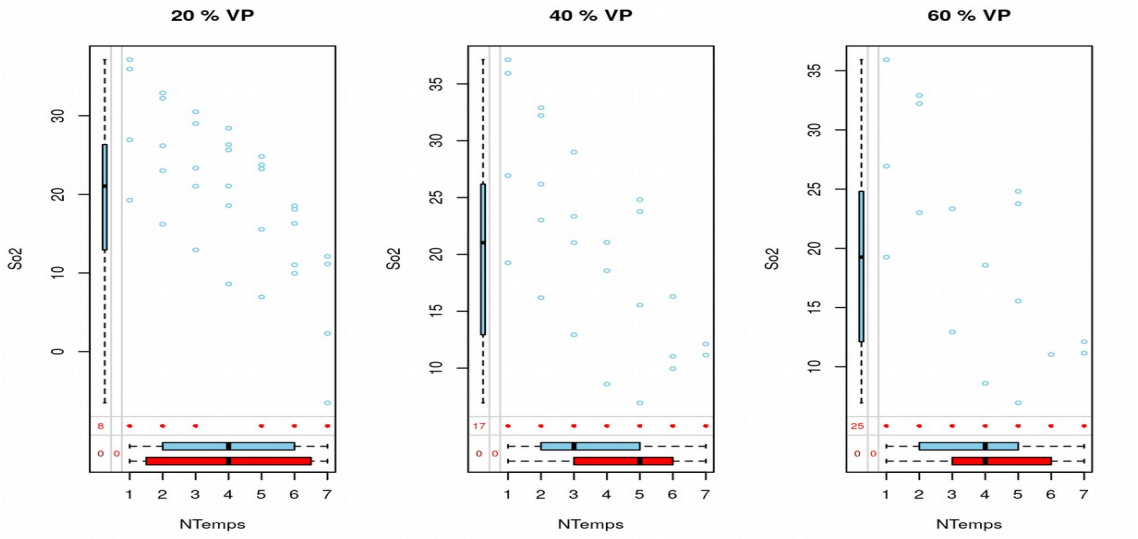
En els diferents gràfics obtinguts i també als mostrats anteriors es reflecteix com s'ha aplicat d'una manera aproximadament igualada les absències de dades. Evidentment com més petita és la mida mostral és més fàcil que es donin diferències i veure possibles patrons que realment amb una mida mostral més gran no hi serien.

3.5.1.2 Anàlisi específic de dades perdudes per conjunt

Relacionat amb els gràfics anteriors, en aquest apartat es pot veure com es reparteixen les dades perdudes en els diferents conjunts per veure en els casos específics com poden afectar al model. Com l'apartat anterior, els següents gràfics mostren les visualitzacions més rellevants en quant a proporcions que s'apliquen de dades perdudes als punts dels diferents conjunts.

Un dels gràfics que pot donar informació interessant és una variació de *scatterplot* amb diagrames de caixes laterals per mostrar la pèrdua de dades. Com a exemple del tipus de informació visualitzada es mostra So2:

Figura 18: Scatterplot/Boxplot de dades perdudes en funció del temps en So2

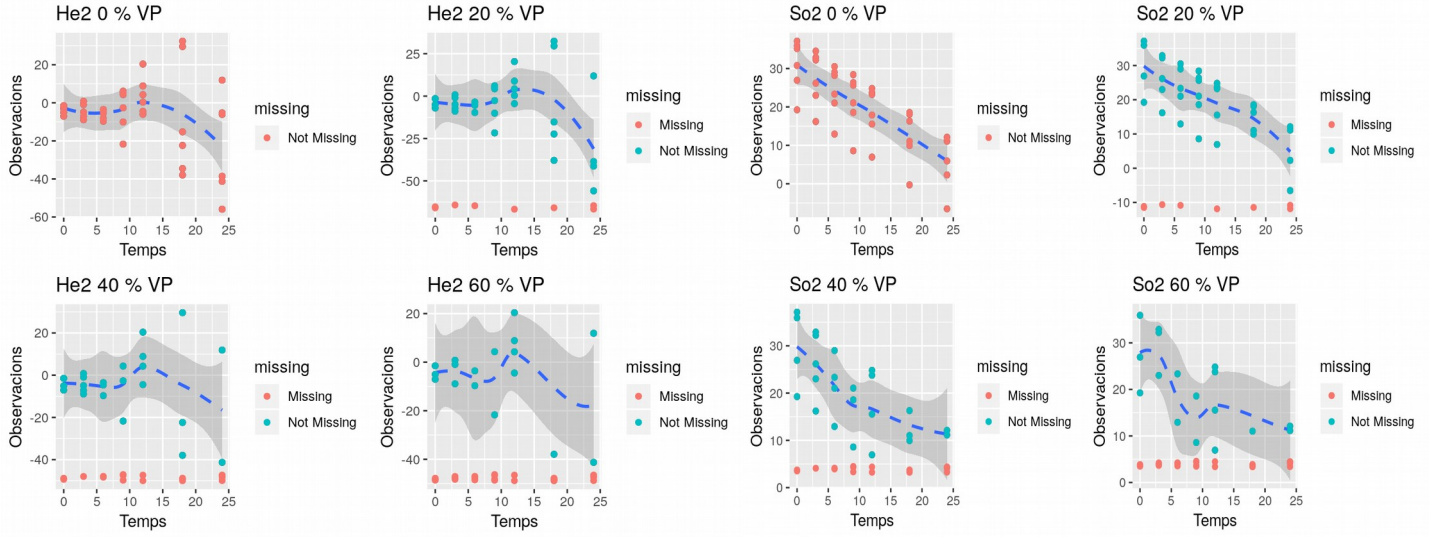


Els gràfics específics per variable ens mostren com canvia la distribució de dades perdudes en front del factor Lot i del factor Temps (en aquest cas es mostren en el factor temps). Al estar aplicant una simulació de MCAR pur es pot veure que:

- En els gràfics sense patró lògic al tenir absència de dades segueix aproximadament una distribució similar, excepte potser algun cas del % més alt de dades perdudes.
- En els gràfics que segueixen un patró també aparentment es manté la tendència tot i perdre observacions.

A continuació els gràfics on es pot veure una línia de tendència suavitzada amb intervals de variància potser és on es pot veure més clar els canvis que provoca visualment la pèrdua de dades. Es mostra pels conjunts He2 i So2:

Figura 19: Gràfics scatterplot amb línia de tendència dels punts i variància associada per He2 i So2



És interessant observar els gràfics al complet en el material addicional l'informe complet i també els gràfics addicionals on es veu la diferenciació per lot.

Dins la quantitat d'informació generada en el codi original la qual s'ha plasmat de manera parcial en els gràfics anteriors, com a idees generals podria extreure's que:

- Les dades perdudes afectarien menys quan les dades estan realment molt centrades.
- En els casos de mesures ja disperses potser tampoc tindrien un efecte rellevant al ja seguir essent disperses.
- En els casos amb patrons molt marcats, però amb dades més disperses possiblement es podrien generar patrons diferents que falsejarien les prediccions (per exemple a So2 o Ho2).

3.5.2 Efectes en la modelització amb dades perdudes

Per aquest apartat es prenen els models ajustats en l'apartat de simulació de models mixtos (3.3) amb dades mixtes per fer els comparatius dels conjunts generats en aquesta simulació i a partir d'aquests es fa servir les eines ja utilitzades de diagnòstic per veure els efectes de tenir dades perdudes (en el cas de funcions de modulació de variància s'ha optat per *varExp* encara que *varPower* sol donar resultats més bons al ser aquesta última menys inestable en l'algoritme de càlcul).

Aquest apartat pot mostrar resultats interessants al haver mesclat les condicions dels models amb efectes aleatoris i mixtos amb l'efecte de pèrdua de dades.

3.5.2.1 Comparació paràmetres qualitatius AIC/BIC i error estàndard residual

Es mostren els gràfics més rellevants de comparació dels diferents models ajustats i els mateixos models amb les pèrdues de dades:

Figura 20: Comparativa models amb AIC/BIC/error std a conjunts Ho1;2 segons pèrdua de dades

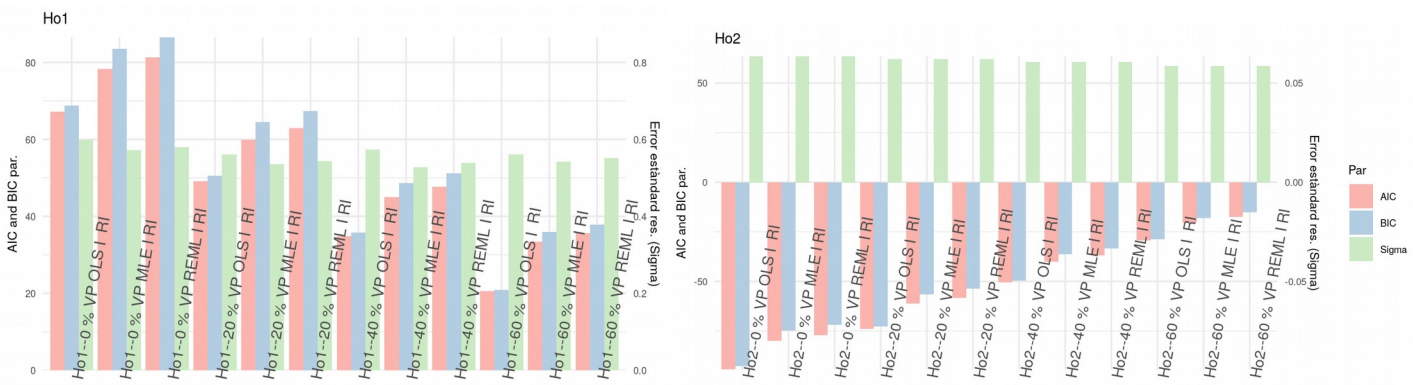


Figura 21:Comparativa models amb AIC/BIC/error std a conjunt He3 segons pèrdua de dades

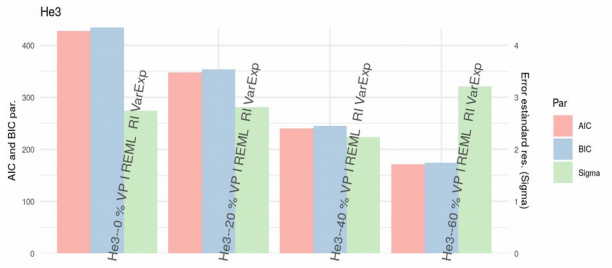
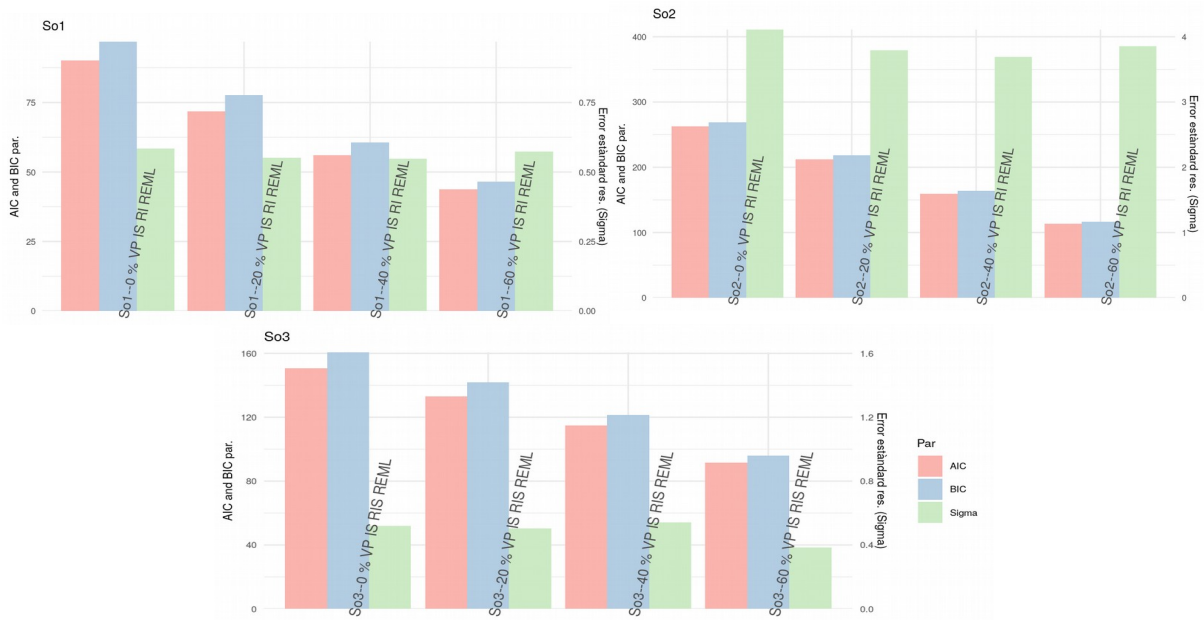


Figura 22: Comparativa models amb AIC/BIC/error std a conjunts So segons pèrdua de dades



Excepte en Ho2, la resta de conjunts sembla haver un patró clar de baixada dels indicadors AIC/BIC a mesura que hi ha més absència de dades. Possiblement en aquest cas el model s'ajusta més correctament a les dades que queden tot i que és probable que el resultat sigui més esbiaixat.

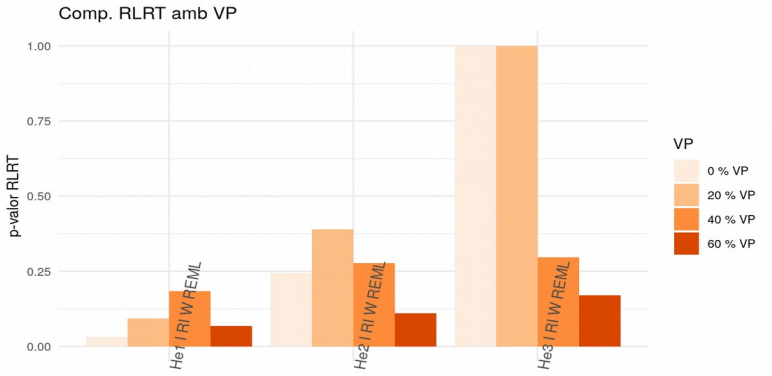
3.5.2.2 Comparació contrastos RLRT

Realitzat amb els models ajustats per ML REML. En aquest cas possiblement el gràfic que ha aportat més informació ha estat el dels conjunts He atès que la resta ha conservat aproximadament els mateixos nivells de significació:

He1: L'efecte aleatori perd significació al perdre dades el conjunt. El patró que detecta inicialment el model de la lleugera dependència del lot no es manté al augmentar la dispersió que d'inici ja hi havia.

He2/He3: L'efecte aleatori ja no era significatiu degut probablement a la alta dispersió dels punts i el canvi de punts fa variar aquesta significació. No es considera que sigui un patró genèric de significació en relació a la pèrdua de valors ja que pot ser més fruit de l'aleatorietat concreta dels punts retirats que ajusten millor l'efecte aleatori dels lots o pitjor.

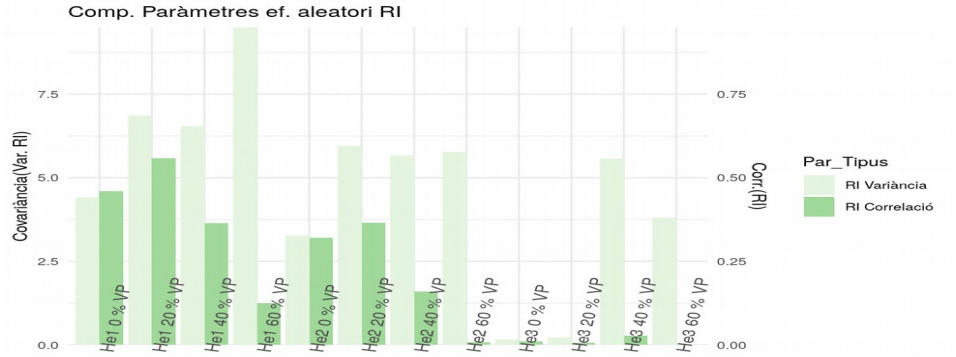
Figura 23: Comparació entre nivells de significació de la prova RLRT als conjunts He



3.5.2.3 Comparació paràmetres efectes aleatoris

Prenent igualment el model REML com a vàlid, s'han calculat els canvis en els components de la variància de l'efecte aleatori (covariància de la matriu de variància-covariància entre repeticions) i la correlació que en resulta entre repeticions. Es representa únicament aquí els conjunts He3 ja que els altres dos grups de conjunts no semblen donar una informació molt rellevant:

Figura 24: Comparació dels paràmetres de variància i covariància i correlació models He dades perdudes



He1/He2: Curiosament semblaria que la covariància augmenta amb la pèrdua de valors i la correlació disminueix. És a dir que possiblement el model dona una càrrega de l'explicació de la variabilitat més alta a l'efecte aleatori, però és incapaç d'explicar-lo a nivell de correlació entre repeticions. S'ha de tenir en compte que en aquest cas també s'afegeix la funció de modulació de la variància.

He3: La dispersió d'aquest conjunt fa que la correlació entre repeticions es mantingui sempre baixa tot i que en els models amb més pèrdua de dades el model calcula una covariància de l'efecte aleatori alta. Possiblement pot ser que estigui detectant un fals patró com s'ha anat comentat en alguns casos.

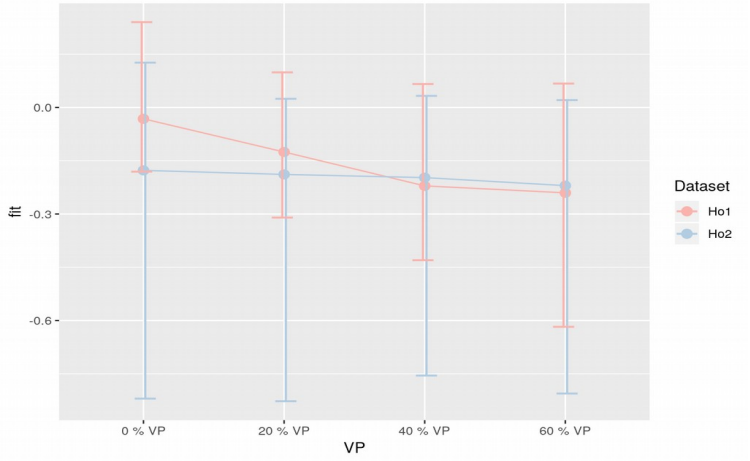
3.5.2.4 Comparació CI (intervals de confiança) en predicció de mitges

No és senzill calcular els CI de predicció en els models mixtes ja que la variabilitat depèn de més factors a part del factor residual com ara el factor o

factors aleatoris. Per duu a terme aquesta tasca es segueix les recomanacions de Duursma [26] que proposa el càlcul de CI mitjançant la generació dels valors de predicció del model amb *bootstrap* prenent com a punt de referència de temps pel càlcul el temps de 24 mesos com a referència d'exemple d'un producte amb una caducitat de 2 anys (per fer-ho s'ha hagut de variar la tècnica de modelatge a nivell de programació utilitzant la funció *lmer*).

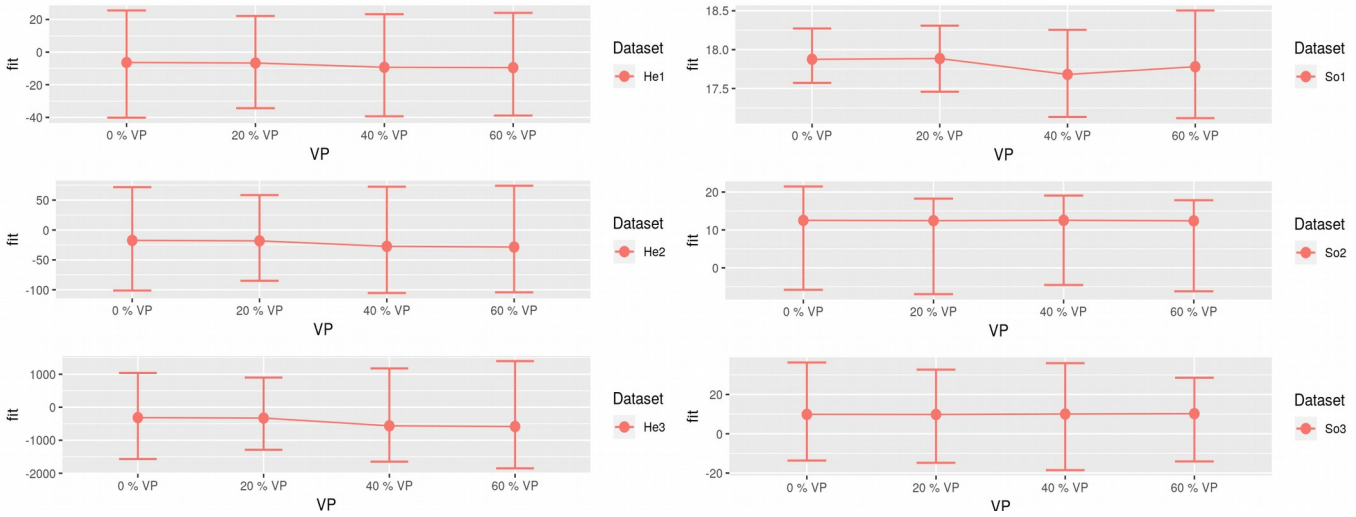
Els resultats es mostren a continuació pels diferents conjunts:

Figura 25: Comparació dels CI de predicció obtinguts a 24 mesos pels conjunts Ho



Tot i la variació de conjunt per les dades perdudes semblaria que el CI per predir la mitjana d'un lot qualsevol no variaria enormement. **Ho** sempre sembla variar més probablement pel factor aleatori aplicat que s'ha anat comentant i que modula la resposta amb els patrons o falsos patrons que es produeixen al perdre dades.

Figura 26: Comparació dels CI de predicció obtinguts a 24 mesos pels conjunts He i So



He: No s'aprecia una variació molt gran entre els intervals de confiança de les prediccions tot i que seria recomanable aprofundir en aquest anàlisi per veure efecte en variació de paràmetres.

So: S'aprecien lleugeres diferències en la predicció, però sembla que l'efectivitat del model compensa la pèrdua de dades en els algoritmes utilitzats en el codi almenys.

3.5.3 Imputació de dades perdudes

Per simplificar es tindrà en compte únicament alguns dels mètodes que s'han fet servir per diagnosticar els efectes dels resultats:

- *Scatterplots* i Gràfics de densitat inicials on es comparen les observacions.
- Anàlisi de indicadors de qualitat AIC, BIC i errors estàndards residuals.
- Predicció de valors en el temps final.

Es simplifica també les bases tractades prenent els casos on s'han vist possibles desviacions interessants de tractar essent Ho2, He2, He3, So2 i So3 i per cada una d'elles només el cas de pèrdua de dades del 40%.

Per la programació del codi R s'han utilitzat els 2 models d'imputació comentats:

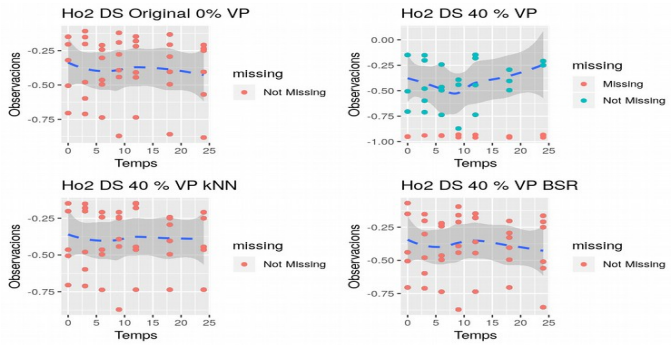
- kNN: L'algoritme utilitzat en R utilitza la formula de la distància i i es pren la opció de tenir en compte la distància per calcular el pes del veí.
- S'utilitza l'algoritme amb opció de regressió estocàstica dins el paquet MICE de R i anomenat BSR o Bayesian stochastic regression.

Posteriorment s'han tornat a ajustar els mateixos models extraient els resultats de diagnòstic utilitzats en les anteriors simulacions.

3.5.3.1 Exploració preliminar conjunts

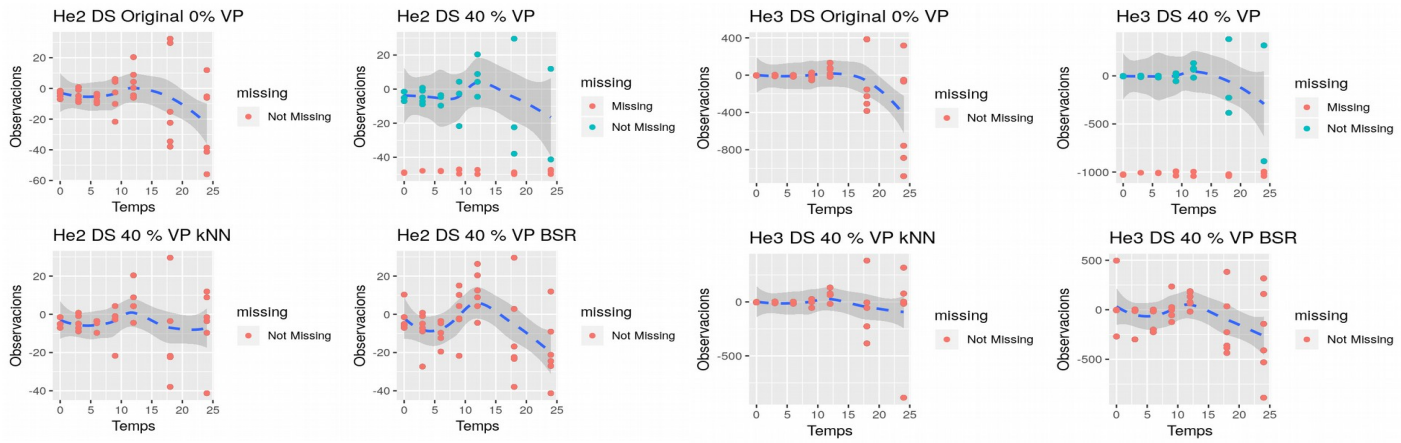
En els següents gràfics es mostra la comparativa en la tendència dels punts en els conjunts originals, amb dades perdudes i les dos opcions d'imputació de dades:

Figura 27: Comparació de tendències entre el model original, amb dades perdudes i amb imputació de dades per Ho2



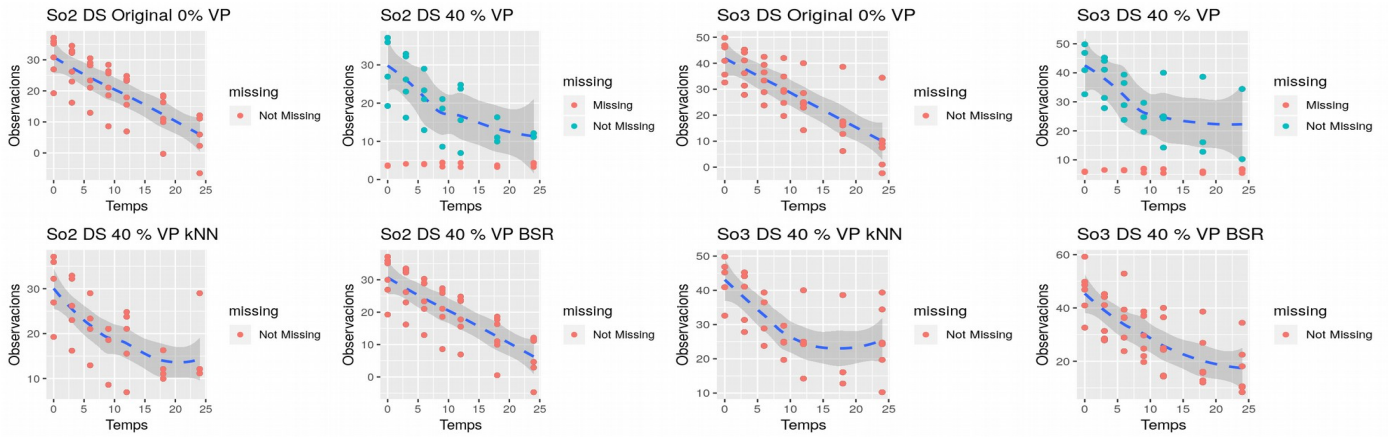
Sembla que els dos tipus d'imputacions resolen de manera correcte les pèrdua de dades en un conjunt amb dispersió de lots, però homogeneïtat.

Figura 28: Comparació de tendències entre el model original, amb dades perdudes i amb imputació de dades per He



Semblaria que en el cas dels conjunts amb simulació d'heteroscedasticitat en el temps, el model de imputació per regressió és el més encertat, tot i que en aquest cas no sembla que la simple deleció de dades modifiqui molt el perfil original.

Figura 29: Comparació de tendències entre el model original, amb dades perdudes i amb imputació de dades per So



En aquest cas es veu amb més claredat la superior eficàcia de la imputació per regressió, essent evidentment simulacions de models amb regressió significativa.

En general és interessant veure com es resolen els punts segons el mètode. Donaria la impressió que en els conjunts amb patrons de regressió més definits o més centrats en cada lot, l'aproximació per BSR aconsegueix més bon efecte, mentre que en altres conjunts sense patrons de regressió dona més bon resultats kNN.

Addicionalment en el codi original també s'ha extret els gràfics diferenciant per lots i les densitats, que no aporten informació tan rellevant, però també són interessants per veure els efectes de cada cas.

3.5.3.2 Diferències models ajustats

Es visualitza la comparació dels models amb AIC/BIC i error estàndard residual en els gràfics de barres per veure l'efectivitat de la imputació:

Figura 30: Comparació AIC/BIC/Error std. residual entre el model original, amb dades perdudes i amb imputació de dades (Ho2;He2)

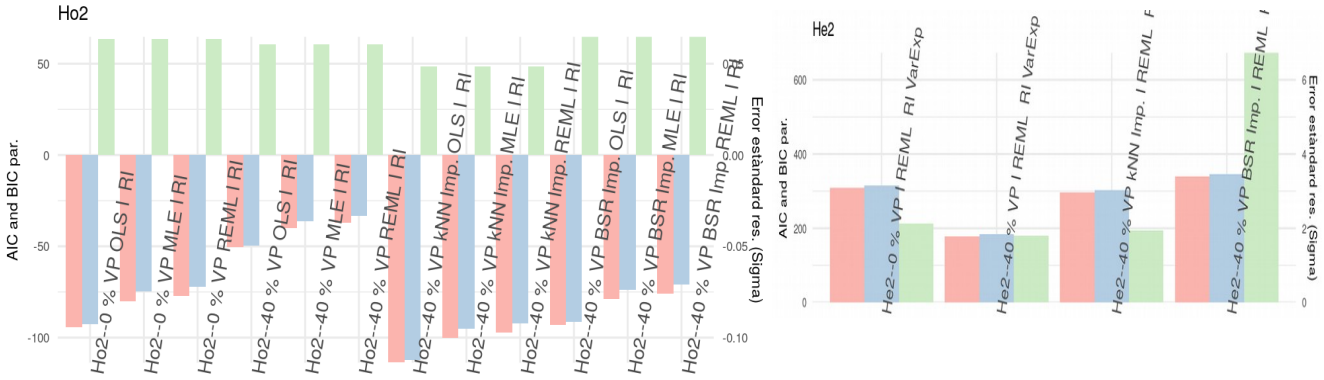
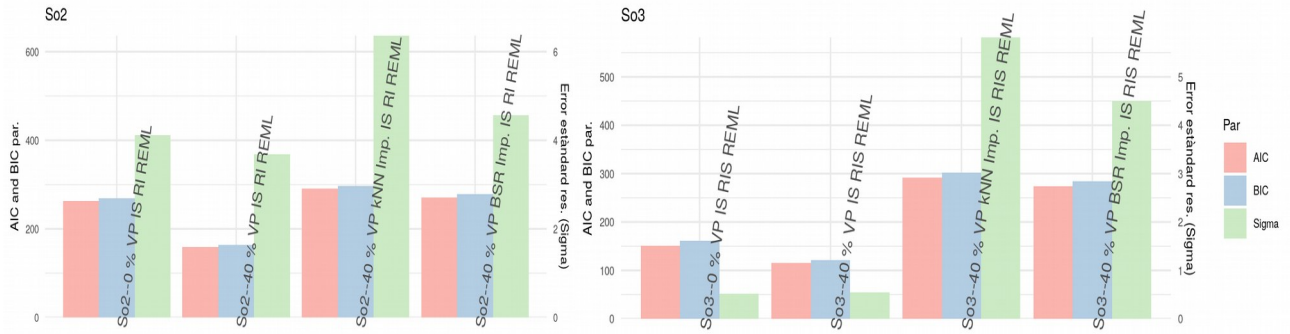


Figura 31: Comparació AIC/BIC/Error std. residual entre el model original, amb dades perdudes i amb imputació de dades (So2;So3)



Ho2: Tant per la part de AIC i BIC com la error estàndard residual, sembla que en el cas de BSR compensa millor les dades perdudes en aquest cas, és a dir que en aquests paràmetres hi ha més proximitat al model original amb aquest tipus d'imputació.

He2/He3: Semblaria ajustar amb més proximitat el model amb kNN, sobretot a nivell de error estàndard residual, tot i que per part dels indicadors AIC i BIC estarien situats a nivells similars els dos sistemes d'imputació.

So2: Semblaria clar en aquesta part que el model que s'aproxima més al model original és el corregit per BSR basat en l'error estàndard, tot i que a nivell AIC/BIC no s'aprecien diferències. De fet és una correcció basada en una regressió, cosa que tindria sentit per aquest model.

So3: La aleatorietat de pendents fa que els mètodes utilitzats per suplir les dades perdudes no siguin efectius en aquest cas donant resultats dels 3

indicadors significativament més diferents que l'ajustat amb eliminació dels valors perduts en comparació al model original.

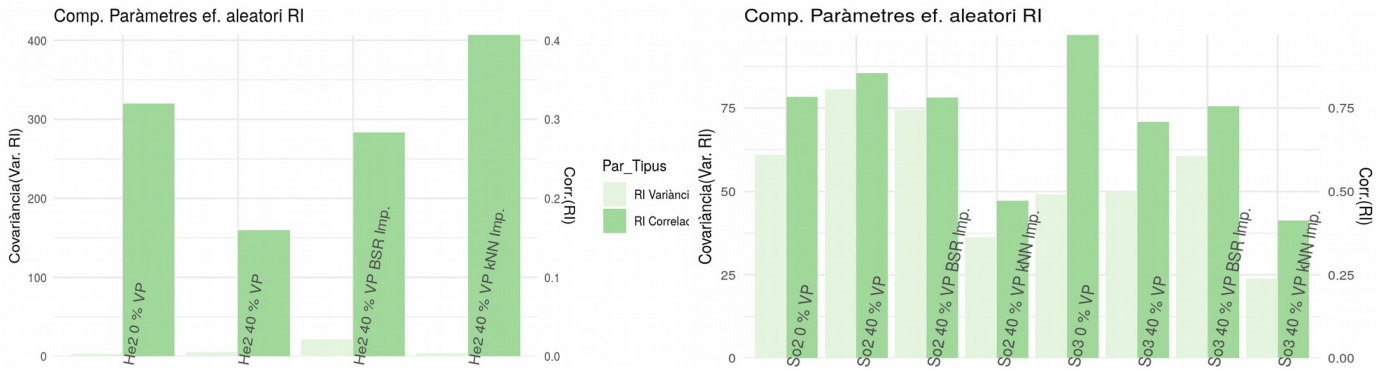
3.5.3.3 Comparació contrastos RLRT (proves de ratio de versemblança restringida)

No s'ha trobat conclusions clares en les visualitzacions i resultats (veure informe complet a materials addicionals).

3.5.3.4 Comparació paràmetres efectes aleatoris

En aquest cas els únics resultats obtinguts que semblen aportar algun tipus d'informació serien els conjunts He1 i So:

Figura 32: Comparació de paràmetres variància i covariància i correlació de matriu residual per comparació de models d'imputació de dades perdudes (He2;So2)

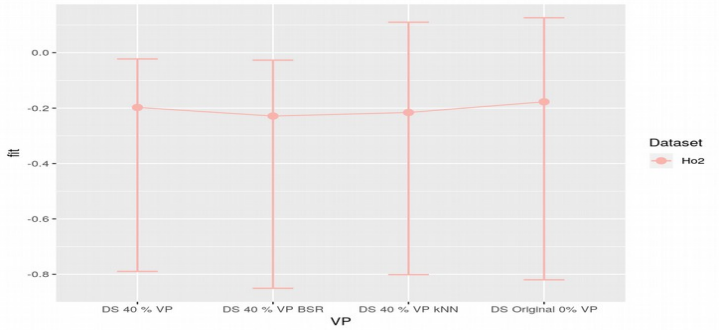


Sembla que el model amb imputació BSR dona millors resultats en el cas He2 i So2 (el cas He possiblement per la falsa tendència). En el model So3 no queda clar si és millor BSR o deleció atès que depèn de si es té en compte la correlació entre mesures repetides o la variància corresponent a l'efecte aleatori s'optaria per un model o l'altre. Possiblement la complexitat del model fa que sigui necessari en aquest cas un model d'imputació més complex per fer una bona correcció de dades perdudes.

3.5.3.5 Comparació CI a temps final

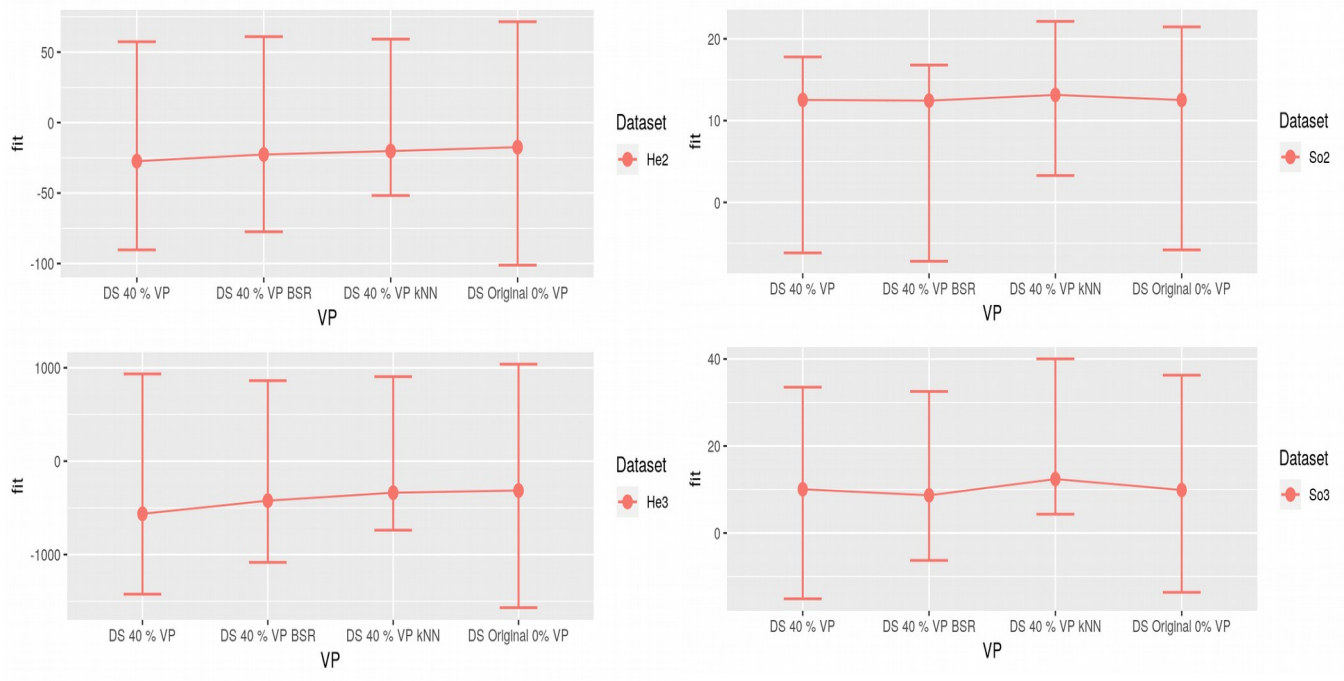
Es mostren a continuació el càlcul de CI al temps final utilitzant el mètode ja comentat en aquesta mateixa simulació de *Bootstrap* i model *lmer*:

Figura 33: Comparació CI de predicció per comparació de mètodes d'imputació Ho2



En aquest cas sembla que el sistema kNN aconsegueix generar una predicció molt semblant al model original.

Figura 34: Comparació intervals de confiança de predicció per comparació de mètodes d'imputació de dades perdudes a He ; So



He: Tot i no haver molta diferència amb el conjunt amb eliminació de valors perduts, possiblement en aquest el mètode BSR ha estat el mètode més proper al conjunt original.

So: Realment no es pot apreciar una millora significativa en aquest cas pels dos conjunts corregits en front el que ha eliminat els valors perduts, inclús potser estan més allunyats que el conjunt original.

4 Anàlisi pràctic

Per la realització de les simulacions de models mixtos s'ha utilitzat el programari *Rstudio [3]* (treballant en el sistema operatiu Linux distribució *Ubuntu* [23].

Inicialment s'ha intentat extreure un conjunt de dades que respongués al context d'estudi d'estabilitat de producte a través de les bases de dades obertes disponibles a internet. Les webs consultades per fer aquesta cerca han estat les següents:

http://archive.ics.uci.edu/ml/index.php; https://www.canada.ca; https://www.cpc.unc.edu; https://data.europa.eu; https://www.data.gov/; https://datahub.io; https://dataverse.harvard.edu/; https://data.world/; https://dreamtolearn.com/; https://explorer.opentrials.net; https://www.gse.harvard.edu/; https://healthdata.gov; https://hofmann.public.iastate.edu/data_in_r_sortable.html; http://www.issip.org/open-data-sets/; https://www.kaggle.com/; https://kilthub.cmu.edu/; http://www.lib.berkeley.edu/; http://library.missouri.edu/; https://www.math.ethz.ch/sfs; https://www.ncbi.nlm.nih.gov/; https://new.censusatschool.org.nz; https://opendatamonitor.eu; https://openmv.net/; https://www.propublica.org/datastore/; https://r-dir.com/; https://ssri.duke.edu/; https://usegalaxy.org/;

Tot i la cerca no s'ha localitzat un conjunt específic de dades corresponent a un estudi d'estabilitat de producte en règim obert. S'ha optat per utilitzar la base de dades corresponent a un estudi clínic disponible amb el *dataset* anomenat LAKE procedent d'un estudi amb malats del VIH [27].

Tot i no ser un estudi d'estabilitat de producte s'ha modificat el conjunt original per obtenir un subconjunt que tingués característiques similars a un estudi equivalent amb el qual s'ha treballat per aplicar alguns dels mètodes vistos. En aquest cas per planificació de temps s'ha modelat únicament en el context de models mixtos sense posar en pràctica la part de dades perdudes treballat en l'apartat de simulació 3.5.

En 8.1 Annex 1 es pot consultar el codi i en el material addicional descrit al capítol 9 es pot consultar l'informe complet generat mitjançant R *markdown*. A continuació es mostren les etapes i resultats més rellevants obtinguts de l'anàlisi pràctic de LAKE.

4.1 Selecció subconjunt i anàlisi descriptiu

4.1.1 Selecció subconjunt LAKE1

Per l'anàlisi inicial del conjunt original s'ha observat que està formada per una quantitat d'individus gran i suficient per analitzar de manera correcta (116 individus). Hi ha també una quantitat enorme de variables (219). És d'interès

pel model que es vol testar modelitzar amb poques variables d'inici ja que l'objectiu no és l'anàlisi multivariable en sí. Fent una identificació ràpida de les variables que componen el conjunt hi ha:

- Variables del tipus caràcter que en principi serien identificadors dels individus i de les quals es conserven *nusuario* i *npac.*
- 8 variables en el format de data: No s'utilitzen.
- 30 variables del tipus *haven_labelled*: No es tindran en compte excepte la variable *Grupo* que marca els dos tractaments de l'estudi i *sexo* que podria aportar informació interessant al tipus de model a aplicar.
- 178 variables del tipus numèric. En el present estudi, a excepció de la variable de càrrega viral que semblaria evidentment la variable a observar més important i el temps que ja es suposa una variable necessària per poder fer l'estudi longitudinal, la resta es prenen de manera semialeatòria per veure com reaccionen al model i d'aquesta manera també impedir el biaix de seguir els resultats de l'estudi original.

Amb la selecció realitzada i l'aplicació de la selecció semialeatòria de predictors (seleccionant 4 possibles predictors de tipus numèric) s'eliminen en aquest cas tots els individus que contenen en algun dels predictors seleccionats dades perdudes quedant un subconjunt que s'ha anomenat LAKE1 amb la següent estructura:

```
str(LAKE1_noNA)
## 'data.frame':    40 obs. of  11 variables:
##  $ nusuario  : Factor w/ 2 levels "aocampo","gcarosi": 1 1 1 1 1 1 1 1 1 1 ...
##  $ npac      : Factor w/ 7 levels "001","002","003",..: 2 2 2 2 2 3 3 3 3 3 ...
##  $ edad      : num  29 29 29 29 29 33 33 33 33 33 ...
##  $ Grupo     : Factor w/ 2 levels "Treatm.1","Treatm.2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sexo      : Factor w/ 2 levels "Hombre","Mujer": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Ident     : Factor w/ 8 levels "aocampo 001",..: 2 2 2 2 2 3 3 3 3 3 ...
##  $ Tiempo    : num  0 12 24 36 48 0 12 24 36 48 ...
##  $ CargaViral: num  48400 40 40 40 40 190000 87 40 47 40 ...
##  $ Albumina  : num  4.39 4.54 4.51 4.67 4.59 4.09 4.06 4.08 3.94 4.27 ...
##  $ CD4P      : num  12.6 15 26 25 26 19 22 24 29 24 ...
##  $ Cloro     : num  107 103 104 104 108 105 104 103 106 103 ...
```

S'han conservat els temps d'anàlisi originals que són 0, 12, 24, 36 i 48 setmanes, pel que les 40 observacions correspondrien a 8 individus diferents amb mesures repetides per cada un d'ells. Pel nombre de variables i d'individus es podria assimilar a un estudi d'estabilitat simple dels que s'han simulat en els apartats anteriors.

4.1.2 Anàlisi descriptiu LAKE1

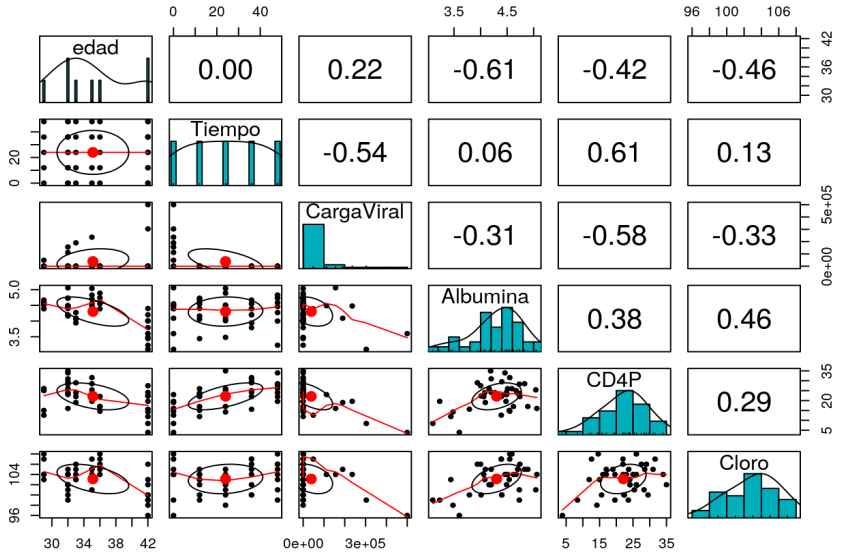
Un dels punts més importants al fer el resum d'estadístics descriptius és la diferència entre la mitjana i mitjana de la variable resposta que dóna idea de que la distribució no serà gaussiana i possiblement no s'aproximarà a la normalitat:

```
##    CargaViral
##  Min.   :    40.0
##  1st Qu.:    40.0
##  Median :    50.0
##  Mean   : 38735.7
##  3rd Qu.:   194.2
##  Max.   :500000.0
```

També s'ha comprovat que en els individus seleccionats no hi ha una proporció equivalent de homes i dones per la variable sexe si es té en compte els dos tractaments o en general pel conjunt. La proporció de sexes és 100 %, 0 % (Homes/Dones) pel tractament 1 i 75 %, 25 % (Homes/Dones) pel cas del tractament 2 (en cas que no hi hagi repartiment equitatiu pot ser un problema per analitzar el significat real del factor sexe).

De l'anomenada *scatterplot matrix* s'observa les possibles relacions entre les diferents variables escollides:

Figura 35: Scatterplot matrix de les variables del conjunt LAKE1 amb les correlacions i histogrames associats



D'aquests gràfics i càlculs interessa saber quines són les variables que tenen una correlació més alta amb la resposta de la càrrega viral que en aquest cas en resulta una correlació amb les variables *edad, Tiempo, Albumina, CD4P, Cloro*, pel que podria resultar com a predictor principal la variable de correlació més alta en valor absolut: *CD4P*. Com a prova addicional al codi original veient l'histograma de dades de la variable càrrega viral, s'ha repetit la matriu de correlació eliminant el temps 0 de tots els subjectes, però no ha resultat amb canvis aparentment significatius.

La càrrega viral semblaria que es distribuiria en una distribució diferent de la normal. Una opció seria modificar la distribució associada, però podria complicar l'anàlisi. La tècnica que s'utilitza habitualment en aquests casos i

recomanada per les guies d'organismes oficials com [20] és transformar les dades. Per aquest motiu s'han considerat dos transformacions en aquest anàlisi la transformació logarítmica i la transformació de Box-Cox que es tracta d'una funció potencial depenent d'un factor λ que es determina mitjançant una maximització de la versemblança *log-Likelihood*. A continuació es mostra els gràfics de histograma, caixa i *QQ plot* de la variable resposta *CargaViral* amb les dos transformacions:

Figura 36:Histograma, diagrama de caixa i QQ plot de la variable CargaViral amb trasnformació Cox-Box de LAKE1

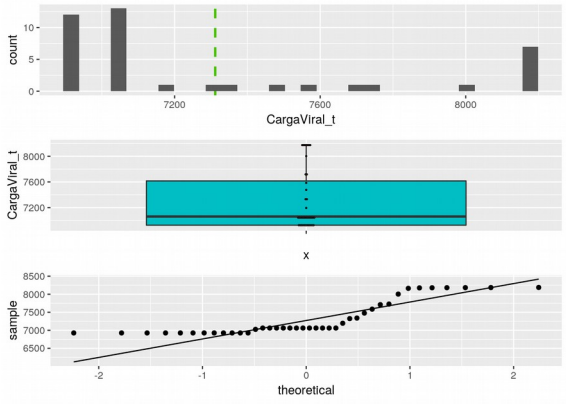
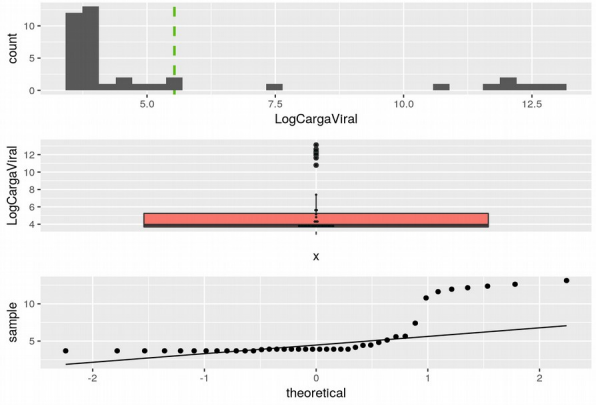


Figura 37:Histograma, diagrama de caixa i QQ plot de la variable CargaViral logtransformada de LAKE1



Sembla que amb les transformacions es produeix una millora en l'aproximació de la normalitat i en el repartiment dels valors.

Addicionalment s'ha observat les diferències entre els grups de tractament, però en el subconjunt LAKE1 no s'aprecia una diferència real entre els dos grups.

4.2 Ajust models simples

4.2.1 Ajust model

En l'ajust del model més simple per mínims quadrats s'aplica una aproximació mitjançant algoritmes que segueixen el model *stepwise* el qual és una combinació de les comparacions ANOVA entre afegir predictors al model més simple o eliminar-ne al model complet per arribar al model més òptim. Habitualment aquesta tècnica no arriba al model òptim, però serveix com a punt de partida per elaborar els models més complexos. Es mostren els models adoptats mitjançant aquesta tècnica per la variable resposta original i les dos transformacions i taules resum dels resultats dels models:

```
## lm(formula = CargaViral ~ Tiempo + CD4P + Cloro, data = LAKE1)
## lm(formula = LogCargaViral ~ Tiempo, data = LAKE1)
## lm(formula = CargaViral_t ~ Tiempo, data = LAKE1)
```

	M.Simple sense transf. 1			M.Simple log transf			M.Simple Box-Cox t		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Tiempo	-1867.5819	957.9577	0.0591	-0.1343	0.0203	0	-21.7842	2.7622	0
CD4P	-4985.2783	2505.0233	0.0542	–	–	–	–	–	–
Cloro	-6947.9474	4748.509	0.1521	–	–	–	–	–	–

Model	Mediana residual
M.Simple sense transf	-7184.6798137
M.Simple log transf	-0.0411551
M.Simple Box-Cox t	11.4023266

Els resultats complets es poden veure a 8.5 Annex 5: Resums R dels models simple i d'interaccions de l'anàlisi pràctic LAKE1.

Com es pot observar en els 3 modelatges de model simple:

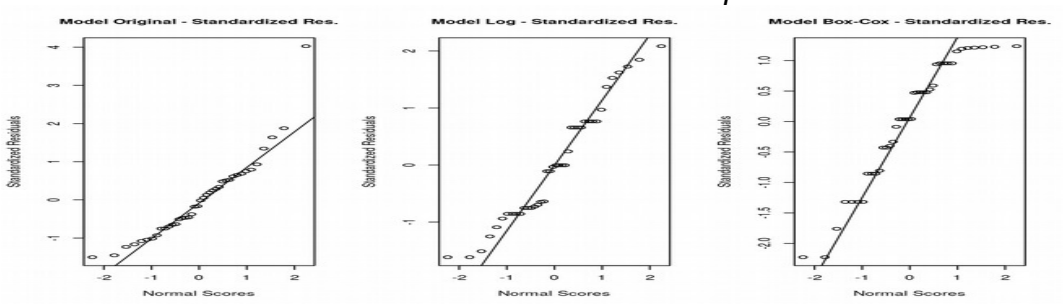
- El model amb variables originals s'ha considerat òptima la utilització de 3 predictors (*Tiempo, CD4P, Cloro*) per explicar la resposta, els 3 amb una prova de significació no significativa. Addicionalment s'ha fet el càlcul ANOVA on es calcula la significació si es treuen del model 1 a 1 i on ha resultat que l'ordre de importància dels termes és *Tiempo, CD4P, Cloro* de més a menys significatiu.
- En els 2 models transformats, el sistema decideix pels 2 un model més simple que depèn només del temps essent aquest molt significatiu per explicar la resposta (p-valors de 0).
- El que també s'observa és que tampoc s'ha considerat els factors de tractament i sexe pel que encara que en l'estudi original hauria estat un dels objectius aquí no es tenen en compte per l'ajust.

4.2.2 Diagnòstics

4.2.2.1 Normalitat

Tot i haver obtingut significació en la prova de Shapiro-Wilk pel cas original i el Cox-Box, en analitzar els gràfics QQ plot de normalitat amb residus estandarditzats es pot apreciar que la desviació respecte a la normalitat no és extrema:

Figura 38: QQ plot pels models simples LAKE1 comparant també segons la transformació de la variable resposta

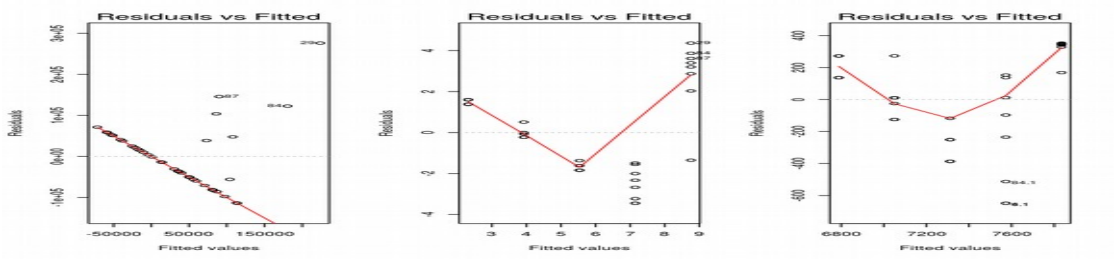


En aquests casos habitualment es pren l'assumpció de normalitat.

4.2.2.2 Homoscedasticitat

Es mostren els gràfics de residus front els valors ajustats per veure el grau d'homoscedasticitat:

Figura 39: Residus vs v.ajustats pels models simples LAKE1 comparant també segons la transformació de la variable resposta



Semblaria clarament que al model original es té un problema de heteroscedasticitat amb els residus que semblen seguir clarament una tendència negativa excepte per uns quants punts. També ho suporta el test de Breusch-Pagan amb un p-valor de 6e-04.

Al transformar logarítmicament sembla que es compensa tot i donar també aparentment un patró de comportament negatiu i amb un p-valor de BP de 0.

El model amb transformació Cox-Box seria potser l'únic model que donaria una homogeneïtat aparent residual tot i que també sembla seguir una certa forma no lineal, i amb un p-valor de BP de 0.0168.

4.2.2.3 Altres diagnòstics

Relacionat amb les proves d'homoscedasticitat i les de normalitat en el model original s'ha fet algunes proves per possibles outliers que s'observen en els gràfics. Al fer aquestes proves s'ha tornat a treballar amb el factor tractament, ja que sembla que en els dos grups hi hauria un conjunt de valors relativament gran que es mostren com aberrants i influents ja que al eliminar-los fan variar significativament el model en cada eliminació. Eliminar-los no seria un procediment correcte sense una justificació com ara una variable oculta que el convertís en una nova població. Com que no es té aquesta evidència, s'opta per no eliminar cap d'ells i amb avançar en paral·lel amb els conjunts transformats que mitigarien aquest efecte i el conjunt original que ampliacions del model es podria mitigar també.

Com a última prova també s'ha analitzat la multicolinealitat en el model original ja que s'ha trobat certa correlació entre les variables *Tiempo* i *CD4P*. Les proves realitzades del model complet no han donat un diagnòstic definitiu ja que apunten en diferents direccions S'està doncs en una situació de compromís on no hi ha una prova definitiva entre les fetes que doni una direcció molt clara. Com que l'objectiu és crear un model de predicció es decideix sacrificar tenir incertesa en els coeficients dels predictors en front de

tenir un % d'explicabilitat més gran de la resposta agafant el model amb més variables, encara que hi hagi un grau de multicolinealitat present.

4.3 Ajust models amb interaccions

És d'interès comprovar l'efecte que té el factor individu amb la covariant temps per plantejar-nos si en propers passos es considera l'efecte aleatori d'aquest. Per fer això un dels sistemes és afegir el factor individu dins el model i construir el terme creuat amb la covariant temps per veure si aquest factor és significatiu, és a dir si la variació de les observacions amb la covariant temps depenen del factor individu o no (seria l'equivalent a la prova de paral·lelisme dels models de regressió). Addicionalment es poden comprovar també si hi ha interaccions entre altres dels termes del model.

4.3.1 Ajust models

4.3.1.1 Interacció Individu Temps

Ajustant perla interacció individu temps, s'ha obtingut resultats pels 3 conjunts amb p-valors no significatius i que per tant no s'aplicarien mesures per ajustar el model ales variacions de pendents dependent del factor individu.

4.3.1.2 Altres interaccions

En el model original s'ha considerat més variables predictores a part de la variable temps. Seria interessant veure si la interacció entre aquestes variables té alguna significació que millori l'explicabilitat del model i per això es fan els ajustos corresponents per veure les proves ANOVA de significància. S'utilitza també el procés de *stepwise* amb el model complet d'interaccions i també es fa pel cas dels models transformats. Els models resultants i les taules resum es mostren a continuació:

```
## lm(formula = CargaViral ~ Tiempo + CD4P + Cloro + Tiempo:CD4P +
##     Tiempo:Cloro + CD4P:Cloro + Tiempo:CD4P:Cloro, data = LAKE1)
## lm(formula = LogCargaViral ~ Tiempo + CD4P + Tiempo:CD4P, data = LAKE1)
#Box-Cox ## lm(formula = CargaViral_t ~ Tiempo + CD4P + Tiempo:CD4P, data = LAKE1)
```

	M.Int sense transf. 1			M.Int log transf			M.Int Box-Cox t		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Tiempo	-176830.6843	58543.4804	0.0049	-0.3503	0.0694	0.0000	-47.3143	10.2070	0.0000
CD4P	-343452.6770	74069.4864	0.0001	-0.2723	0.0787	0.0014	-23.5364	11.5756	0.0494
Cloro	-60755.4316	10234.7569	0.0000	NA	NA	NA	NA	NA	NA
Tiempo:CD4P	9148.1570	2830.7465	0.0028	0.0105	0.0030	0.0011	1.1663	0.4347	0.0110
Tiempo:Cloro	1649.1938	574.9848	0.0072	NA	NA	NA	NA	NA	NA
CD4P:Cloro	3282.1583	732.5469	0.0001	NA	NA	NA	NA	NA	NA
Tiempo:CD4P:Cloro	-86.7203	27.7370	0.0038	NA	NA	NA	NA	NA	NA

Model	Mediana residual
M.Int sense transf	-1617.8894864
M.Int log transf	-0.2722391
M.Int Box-Cox t	31.5771159

Els resultats complets es poden veure a 8.5 Annex 5: Resums R dels models simple i d'interaccions de l'anàlisi pràctic LAKE1.

S'ha tingut en compte per un costat les interaccions entre els termes existents dels models simples optimitzats i s'ha comprovat l'efecte de l'aparició dels factors d'interacció per paràmetres que no estaven en el model òptim simple en els de variable de resposta transformada. Pel que en principi s'acceptarien aquests models per continuar amb la optimització.

4.3.2 Diagnòstics

Els diagnòstics que s'han calculat són els mateixos que pels models simples i els gràfics resultants són similars els que ja s'havien obtingut. No s'exposen aquí (disponibles a l'informe complet), però es resumeixen les conclusions:

Normalitat: Possiblement es podria apreciar certa millora en la normalitat com també s'aprecia en els p-valors de la prova *Shapiro-Wilk* obtinguts tot i que no d'una diferència extrema.

Homoscedasticitat: No s'aprecia una millora visualment (excepte potser el model de transf. Cox-Box), encara que les proves d'hipòtesis si que semblaria una millora en els p-valors de significació.

4.4 Ajust models amb efectes aleatoris

Pels resultats que s'han anat obtenint semblaria que es tracta en una possible variant de la simulació ja vista identificada com a He2 amb paral·lelisme i diferències entre individus, tot i que en aquest cas és possible que sigui un model més complex al afegir la heteroscedasticitat que s'ha anat comprovant i els efectes addicionals d'altres variables predictores.

4.4.1 Ajust model RI

Si s'analitza amb més detall el context en el que es troba el model s'ha de considerar que el conjunt utilitzat és un subgrup dels individus del model original, que a la vegada ja es una subpoblació de la població en general que pugui tenir el VIH. Cal tenir e compte que l'objectiu d'aquest model de predicció és fer inferència en la població general d'afectats de VIH i no només les de la mostra estudiada, i per tal de fer això s'ha de tenir en compte que si es realitzessin més experiments amb altres conjunts d'individus la resposta podria variar. Per poder representar això és necessari afegir el component aleatori que aporta l'efecte individu amb la seva pròpia variància dins el model.

Per fer aquesta actualització del model s'optarà com a les simulacions per tenir en compte 3 models amb diferents plantejaments del càlcul òptim: El model OLS amb l'error aleatori, MLE i REML. Així en els següents models es veurà si canvia molt la bondat d'ajust respecte als models amb interaccions afegint els termes d'efectes aleatoris i tenint en compte els 2 sistemes addicionals d'estimació de resultats ML.

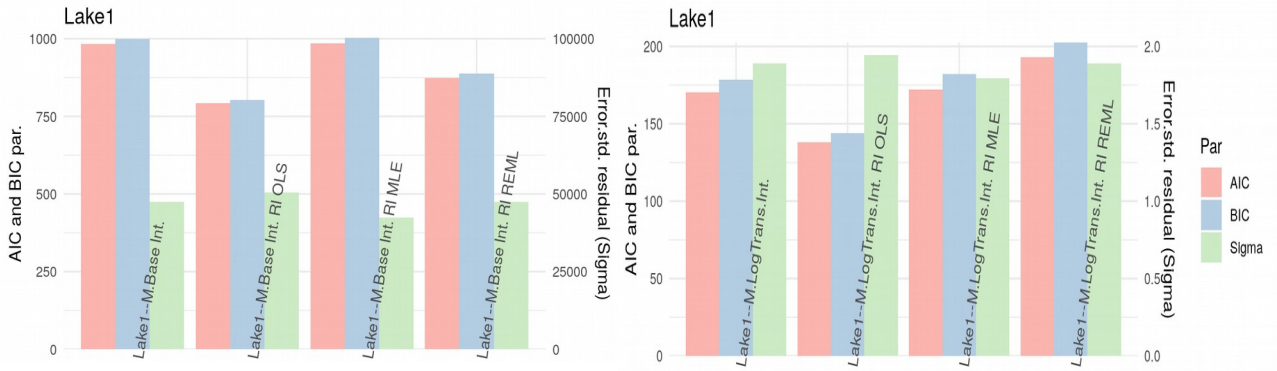
4.4.1.1 Ajust model

Al afegir el terme d'efecte aleatori RI segons individu al model i fer la prova RLRT per tots els models calculats la prova ha resultat no significativa amb un p-valor de 1. Tot i així, a nivell més concret pot ser que signifiquin algun canvi, pel que no es descarta encara afegir efectes aleatoris als models.

4.4.1.2 Diagnòstic

Tot i la prova RLRT no significativa dels models s'han fet les comparacions que ja s'havien utilitzat en les simulacions de comparació de paràmetres AIC/BIC i error estàndard residual pels diferents models. Es mostren els models obtinguts amb més bons resultats:

Figura 40:Comparació AIC/BIC i error residual models LAKE1 amb efecte aleatori RI



En general no sembla que els models amb l'efecte RI aportin una millora significativa en error residual o AIC/BIC.

4.4.2 Ajust model RIS

Tot i que en el test realitzat d'interacció entre factor i covariant ha donat el resultat del paral·lelisme de pendent entre individus, seria interessant veure com es comporta si s'afegeix en el model una component aleatòria d'aquest paràmetre ja que en el supòsit que hi hagués algun factor ocult que modulés la pendent dels individus, la degradació del paràmetre observat en front al temps podria no ser igual per tots ells.

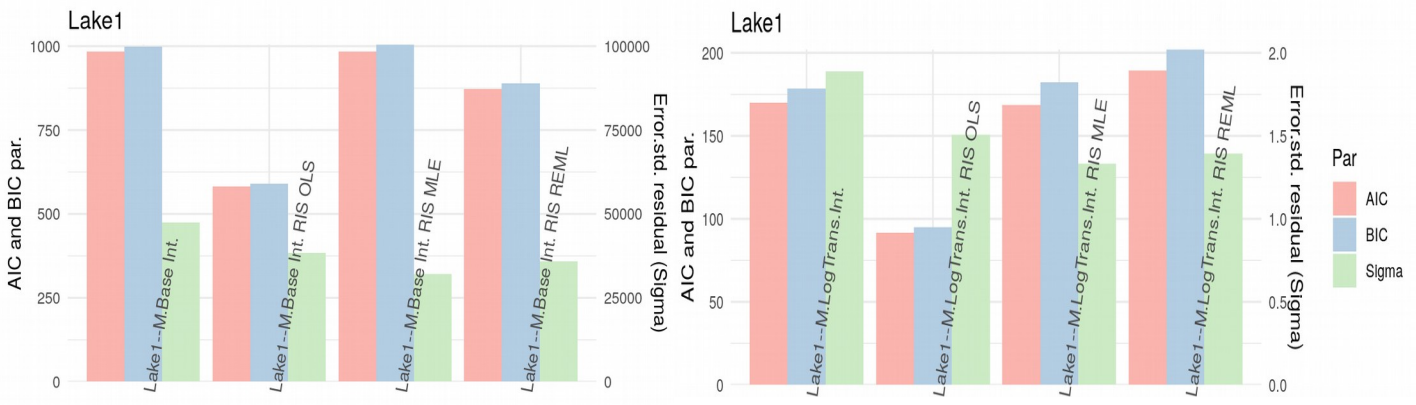
4.4.2.1 Ajust model

S'ajusten els models i, a diferència del cas anterior, al haver dos efectes aleatoris, s'ha fet la prova ANOVA de comparació dels models corresponents ML amb 1 efecte aleatori (RI) i 2 (RIS). Els resultats obtinguts són poc rellevants ja que són molt similars als obtinguts pels models afegint l'efecte RI.

4.4.2.2 Diagnòstic

Tot i la prova RLRT no significativa dels models s'han fet les comparacions de paràmetres AIC/BIC i error estàndard residual pels diferents models. Es mostren els models obtinguts amb més bons resultats:

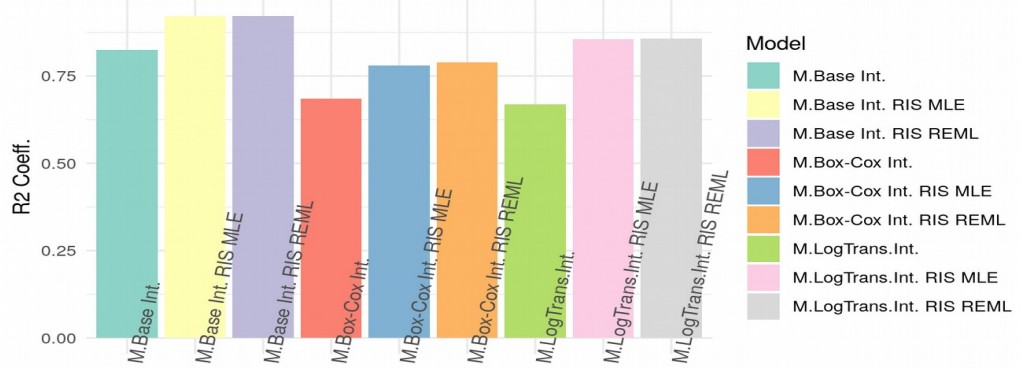
Figura 41: Comparació AIC/BIC i error residual models LAKE1 amb efecte aleatori RIS



Pels models seleccionats amb una variància residual més baixa així com els indicadors AIC/BIC més baixos curiosament es pot veure que pels 2 models que havien donat una significació alta el fet d'afegir l'efecte aleatori RIS, aparentment es nota d'una manera més efectiva la millora en aquests paràmetres al comparar-los amb la referència simple. No és una millora altament significativa, però podria ser suficient per fer prediccions més precises.

Addicionalment s'ha fet una comparativa del coeficient de determinació R^2 dels models. En el cas dels models amb efectes aleatoris es calcula a partir del coeficient de determinació condicionat que suma el que aporta els efectes fixos i els efectes aleatoris (per calcular-ho en R amb els models calculats per ML es fa una aproximació mitjançant la correlació lineal entre la resposta i els valors ajustats):

Figura 42: Comparativa R2 entre els models amb aplicació de RIS LAKE1



Sembla observar-se una millora aparent en el coeficient R^2 per tots els casos si es compara amb el model simple.

Seria interessant mencionar que en l'informe complet en aquest punt s'han tornat a aplicar els diagnòstics clàssics i tot i que no es veu diferència significativa, semblaria haver una millora aparent de normalitat en els models bàsic i log transformat.

4.5 Ajust funcions de modulació de variància

4.5.1 Ajust models

Com en la part de les simulacions s'intenten ajustar les funcions varIdent, varExp i varPower als models ML (aquesta sempre que sigui possible per convergència d'iteracions). Els models ajustats possibles han estat:

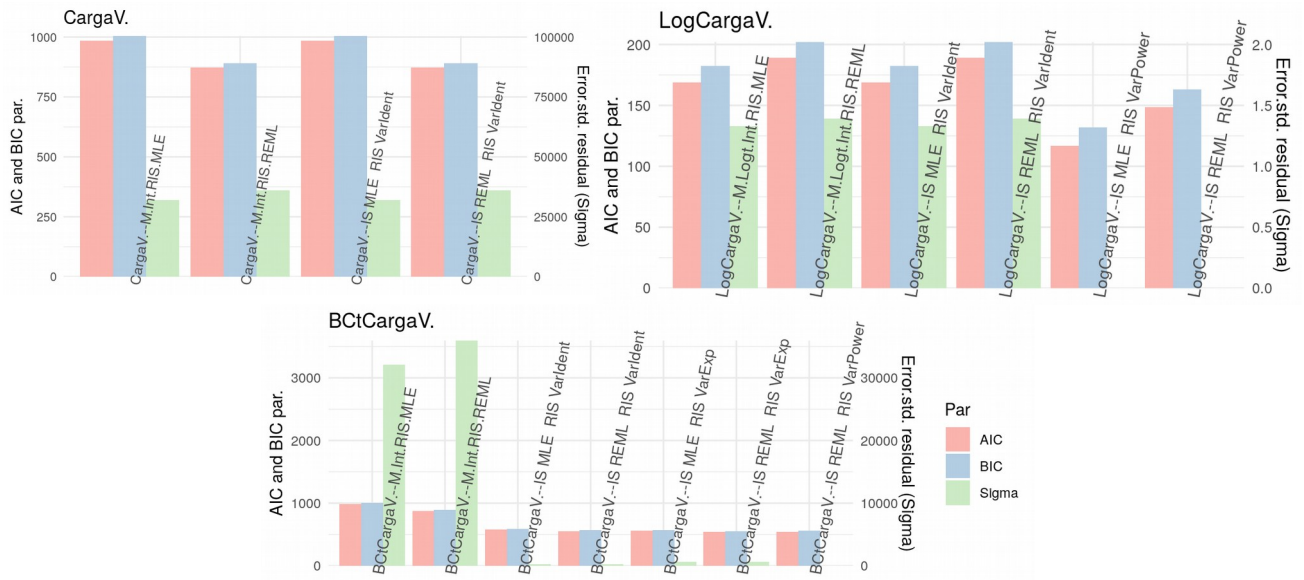
- Model sense transformar: varIdent, varExp.
- Model log transformats: varIdent, varExp, varPower
- Model transf. Box-Cox: varIdent, varExp, varPower (només a l'ajust REML)

Es mostra directament al següent apartat el resum dels resultats obtinguts.

4.5.2 Diagnòstics

Es mostren els models obtinguts amb més bons resultats pels paràmetres AIC/ BIC i error std. Residual en cada cas:

Figura 43: Comparació AIC/BIC i error residual en els diferents models amb funció de modulació de variància a LAKE1



En aquests primers diagnòstics semblaria que la funció *varPower* és la que té més efectivitat al modular la variància d'una manera molt significativa. S'ha pogut comprovar en els dos models transformats, però no en el model base on

no ha convergit en una solució real i on les altres funcions de modulació de variància no semblen aportar cap millora significativa al model.

En el càlcul que s'ha fet del coeficient de determinació R^2 com s'ha fet en el cas anterior no ha donat resultats molt concloent atès que fins i tot sembla que aquests nous models rebaixen la R^2. No és necessàriament que el model sigui pitjor ja que només està calculant la relació punt a punt, sense tenir en compte altres factors com els estrats de lot o temps.

Al fer els diagnòstics clàssics en aquest cas és on es pot veure realment el que s'ha aconseguit amb les funcions de modulació de variància ja que semblen arribar a una situació de certa homoscedasticitat, una de les propietats essencials perquè un model no tingui un biaix important en la desviació. Sobretot es produeix en els models transformats on s'ha pogut aplicar *varPower*.

4.6 Model òptim

Donat que a priori no hi ha un factor de pes per cada paràmetre de bondat d'ajust estudiada, l'estratègia a seguir ha estat tenint en compte els paràmetres Sigma, AIC/BIC, R^2 i p-valor Levene-Test(modificat). Els resultats obtinguts:

- Sigma: Models transformats amb *VarPower* amb més bons resultats.
- AIC/BIC: Models transformats amb més bons resultats, en especial els log que tenen funció de variància *varExp* i *varPower*.
- R^2: Els models transformats semblen donar més bons resultats en especial els models transformats per *varPower* de log i *Box-Cox*.
- homoscedasticitat: Semblen que els models amb més bons resultats serien els transformats amb Log i *Box-Cox* i modulació de variància.

En conjunt sembla que seria ideal utilitzar els dos models transformats Log i Cox-Box amb efectes aleatoris RIS i funció de modulació de variància *varPower*.

En relació a les propietats de variància-covariància i de correlacions. Extret de l'informe complet, com a trets interessants es podria dir que:

- En els 2 casos sembla que el sistema calcula que a mesura que avança el temps es redueix la variància residual.
- El factor de covariància de l'efecte aleatori no sembla segur un patró senzill en funció de la distància, pel que s'entén que pren un càlcul més complex en funció de les dades de què es disposa.
- Correlacions variants segons quin temps es relaciona amb quin, però baixes en general. Es podria dir que no s'ha aconseguit trobar un model que expliqui bé aquesta correlació, o simplement al haver una dispersió en funció del temps baixa la correlació d'aquesta manera.

5 Conclusions

5.1 Conclusions del treball

L'anàlisi d'aquest treball ha anat enfocat a l'entorn dels estudis d'estabilitat com es comporten al modular-los amb les estructures dels models mixtos i quines dificultats suposa tenir dades perdudes en aquests.

En l'anàlisi teòric/pràctic dels models mixtos ha estat interessant veure els conjunts generats on s'han simulat possibles comportaments que poden tenir els productes i veure com reaccionen els diferents models aplicats a aquests. S'ha pogut veure que tot i que els models lineals simples poden arribar a donar una explicació bastant aproximada a la realitat, la utilització dels efectes aleatoris en alguns casos és clau per poder aportar una millora necessària al model i li permeti explicar la variació que hi pot haver en general al descriure el comportament de la resposta d'un producte al llarg del temps. Les funcions de modulació de variància s'han mostrat com una solució bastant efectiva a l'aparèixer heteroscedasticitat al model i també s'ha mostrat flexibles per ser combinades amb transformacions del model.

La part de l'anàlisi teòric de dades perdudes tot i no haver entrat en profunditat, ha servit també per mostrar que, en nivells relativament baixos de nombre d'individus i variables com és el rang que s'ha establert com a habitual en estudis d'estabilitat en el present treball, la deleció en moltes ocasions no allunya el model amb dades perdudes del model complet, i que per les imputacions simples utilitzades és important intentar tenir coneixement de les característiques del conjunt en quant a tendència i homogeneïtat.

Com a combinació de dades perdudes i models mixtos ha estat interessant veure com segons els punts o el nivell de dades perdudes que es donen, el model d'efectes mixtos canvia el seu comportament per intentar adaptar-se i explicar el comportament sense aquelles dades perdudes, cosa que en algunes ocasions és correcte i en altres pot aportar un biaix més gran del que ja hi havia. És important anar en compte al aplicar doncs la complexitat dels models mixtos amb funcions de variància si hi ha la presència de dades perdudes.

Com a conclusió final i sense ser cap novetat ja que, s'exposa habitualment en molts dels materials d'anàlisi estadístic, s'ha pogut constatar en base al treball realitzat i als resultats obtinguts que, al ajustar models estadístics per explicar comportaments o fer prediccions, l'investigador o analista de dades sovint no ha de posar-se com a objectiu necessari el d'ajustar el model a la realitat, si no el d'ajustar un model que, en les condicions del seu estudi, li doni els resultats més pròxims possibles a la realitat. Com a exemple es pot considerar els models que s'han vist amb variables de tendència tot i no haver-se simulat

d'inici, però que realment milloren el model de cara a la explicació i predicció de comportaments.

5.2 Assoliment d'objectius

Respecte a la planificació inicial hi han hagut varis canvis, la major part degut a la replanificació en funció del temps disponible i en una altra part a la viabilitat de segons quins càlculs a nivell d'algoritmes de programació d'una manera directa i també en gran part a l'augment de dedicació en l'escriptura del codi de programació per obtenir informes dinàmics amb una base mínima d'eficiència en l'execució dels càlculs. Es mostra a continuació el resum d'alguns imprevistos o canvis realitzats durant el transcurs de la realització del treball:

- Estudis d'Estabilitat:
    - No s'ha pres en aquest apartat articles específics d'anàlisi estadístic fora de la bibliografia de fonts de organismes públics atès que ja sols la informació proporcionada per ICH s'ha considerat una font d'informació suficientment àmplia per assolir l'objectiu.
    - S'ha pres el cas genèric del model de predicció ja que es considera un dels més complets de les possibles aplicacions d'anàlisi als estudis d'estabilitat ja que n'engloba vàries dins el mateix anàlisi.
- Models amb dades longitudinals: S'ha considerat interessant afegir un apartat no previst als conceptes teòrics per aclarir com es treballa amb aquest tipus de dades.
- Models d'efectes mixtos: S'ha reduït l'anàlisi a condicions molt concretes i no s'ha entrat amb molt detall en funcions de complexitat elevada com són les de modulació de variància.
- Models amb dades perdudes: S'ha limitat el nombre de condicions a aplicar i d'imputacions possibles per adaptar el temps disponible i les situacions habituals dels estudis d'estabilitat.
- Anàlisi pràctic:
    - No s'ha aconseguit localitzar un conjunt de dades reals dins el context d'estudis d'estabilitat pel que s'ha utilitzat un conjunt equivalent.
    - No s'ha abordat dins el problema pràctic la presència de dades perdudes. Es considera tot i així que en la simulació ja es pot extreure informació interessant respecte a l'anàlisi que es volia fer.
- Compliment assumpcions: En general tant per la fase teòrica com pràctica no s'estudia amb detall l'efecte del compliment d'assumpcions, tot i que en tot moment es tenen en compte per les simulacions i anàlisi pràctic.

5.3 Anàlisi de planificació i metodologia

La planificació s'ha anat variant durant el desenvolupament del treball, sobretot en la part més pràctica corresponent a l'anàlisi amb programació R degut a la dedicació addicional del temps de programació i de la cerca de solucions a nivell tècnic de programació per poder anar assolint els diferents resultats i comparacions necessàries. Possiblement una de les dificultats principals a nivell de programació és la complexitat d'aplicar els algoritmes de models mixtos i de modulació de variància residual que comporta una necessitat d'aprofundir en les funcions i variants implicades per poder aplicar-los de manera correcte.

L'objectiu de combinar models mixtos i anàlisi de dades perdudes, els dos temes amb molta informació i aspectes interessants a investigar, ha fet que s'hagi hagut de delimitar cada un d'ells. Tot i així la combinació ha permès obtindre resultats de la combinació dels dos com a aspecte addicional en el treball.

Es considera que la metodologia utilitzada ha estat correcte i l'evolució bidireccional dels anàlisis teòrics paral·lels amb les simulacions han estat clau per poder avançar i aplicar idees que d'inici no s'havien tingut en compte.

5.4 Línies de treball futur

Les línies de treball futur que s'exposen tenen com a referència principal les diferències entre les propostes inicials de planificació del treball i l'estructura que finalment s'ha seguit juntamernt amb alguns punts addicionals dels quals el seu estudi en detall podria ser d'interès:

- Estudi de variants dels models aplicats com poden ser:
    - Models no lineals
    - Models no gaussians.
    - Models lineals generalitzats.
    - Estudi amb més detall de les aproximacions per fer l'estimació dels mètodes ja contemplats (ANOVA, MLE, REML) i afegir el sistema Minque com a mètode addicional per veure'n la comparativa respecte als altres.
- El cas concret dels models mixtos es pot aprofundir en varis aspectes no treballats:
    - Ampliar les simulacions contemplant per exemple el model de tendència i heteroscedasticitat.
    - Aplicació dels diferents models de matriu de variància-covariància i dels mètodes per seleccionar-les.

- Afegir més variants de l'estructura en la simulació com més possibilitats d'interacció de termes, combinació de efectes categòrics i de tipus numèric, niatge o aplicació de aleatorietat de blocs.
- Provar noves funcions de modulació de variància residual i en les ja aplicades aprofundir-ne el coneixement veient totes les condicions que es poden aplicar i les diferències entre elles.

- El cas concret dels conjunts amb dades perdudes es pot aprofundir en varis aspectes no treballats:
  - Dins el context dels models mixtos i les seves variants veure les diferències entre els tipus de mecanismes per generar les dades perdudes en els resultats dels models.
  - Ampliar el rang de mètodes d'imputació de dades, sobretot per veure les possibles diferències entre els de tipus simple i de tipus múltiples.
  - Ampliar l'anàlisi dels resultats combinats de models mixtos i presència de dades perdudes, com l'efecte en la matriu de variància-covariància o en la utilització de les funcions de modulació de variància.
- Ampliar l'anàlisi pràctic veient diferents conjunts de dades amb diferents propietats pel que fa a distribució de dades, nombre d'individus/variables i/o presència de dades perdudes, i veure com responen a l'aplicació dels diferents models.
- Tant per les simulacions com per l'anàlisi pràctic veure l'alternativa d'aplicar un mètode totalment diferent com podria ser la predicció a partir d'un algoritme de classificació de la tipologia *machine learning* per comparar l'eficàcia amb els models optimitzats anteriors de regressió.

6 Glossari

AIC/BIC: Criteris d'informació d'Akaike i Bayesià. Criteris calculats per equacions de ràtios de versemblança per intentar definir l'eficàcia d'un model estadístic.

ANOVA o tests de variància: Test per comparar models o per avaluar l'efecte de cada un de les variables d'un model calculant habitualment la suma de quadrats o suma de quadrats esperadas.

Balancejat / No balancejat (Model o Dades): Models amb la presència de dades en els individus equilibrada o no equilibrada en quant a repeticions.

Covariància: Variable del model que s'espera que sigui el principal regulador de la variable resposta.

Dades perdudes: Absència de dades no prevista en l'estudi.

Efecte fix: Efectes que depenen d'una variable concreta i la seva distribució sense tenir una part aleatòria i prenent únicament els valors que apareixen en el model.

Efecte aleatori: Efectes corresponents a una variable per tenir en compte la possibilitat de les respostes per altres nivells no presents en l'estudi, és a dir per fer inferència en una població més gran.

Esdeveniment (*event*): Condició que es dona per acabar un estudi o arribar a una conclusió habitualment utilitzat en models de supervivència.

Estabilitat accelerada: Condicions de l'estudi d'estabilitat portades al límit del que pot suportar el producte per provocar l'acceleració d'alguna degradació químic o canvi físic que ja es dona en el producte. Les dades d'aquests estudis solen servir per assessorar els estudis de més llarga durada en condicions més favorables pel producte (extret parcialment de [20]).

Estudi d'estabilitat: Estudi de producte habitualment destinat a determinar o demostrar la caducitat d'un producte.

***Expected mean squares* o mitges quadràtiques esperades (EMS)**: Mitges de quadrats esperades calculades pels models que contenen efectes aleatoris per tenir en compte la variància descomposada del model.

Factors creuats (*crossed*): Combinacions de nivells de dos o més factors.

Factors niats (*nested*): Combinacions de dos o més factors on un d'ells pren nivells totalment diferents pels nivells de l'altre.

Funcions de modulació de variància: Funcions per modular en un model estadístic la variància residual.

Ho: Nom donat dins d'aquest treball als conjunts de dades simulades per tenir els punts amb homogeneïtat relativament alta, no confondre amb el terma d'hipòtesis nul·la.

He: Nom donat dins d'aquest treball als conjunts de dades simulades per tenir els punts amb heterogeneïtat relativament alta.

I (definit pels models) : Abreviatura presa en aquest treball per definir els models amb la variable predictora d'intercepció sense pendent de l'efecte covariant.

IC o Intervals de confiança: Intervals de probabilitat de trobar els valors o mitges dels valors calculats a partir dels models habitualment a un nivell del 95%.

IS (definit pels models) : Abreviatura presa en aquest treball per definir els models amb la variable predictora d'intercepció i la de pendent de l'efecte covariant.

Imputació de dades: Tècniques utilitzades per substituir dades perdudes en un conjunt de dades per anàlisi estadístic.

MLE (definit pels models) : Abreviatura presa en aquest treball per definir els models estimats pels algoritmes de màxima versemblança *maximum likelihood estimates.*

Models mixtos: Models on hi ha presència d'efectes fixos i aleatoris.

OLS (definit pels models) : Abreviatura presa en aquest treball per definir els models estimats pels mètodes clàssics d'ANOVA per mínims quadrats o *ordinary least squares*.

Agrupabilitat (Poolability): Combinació de lots habitualment utilitzada en anàlisi estadístic de estudis d'estabilitat.

Producte final: Es refereix a la fase del procés de fabricació o elaboració d'un producte comercial que correspon al seu format comercial tant de composició com de condicionament.

Producte semielaborat: Es refereix a la fase del procés de fabricació o elaboració d'un producte comercial que correspon a una fase prèvia a la fase de producte final.

REML (definit pels models) : Abreviatura presa en aquest treball per definir els models estimats pels algoritmes de màxima versemblança *restricted estimated maximum likelihood.*

RI (definit pels models) : Abreviatura presa en aquest treball per definir els models amb l'efecte aleatori d'intercepció.

RIS (definit pels models) : Abreviatura presa en aquest treball per definir els models amb l'efecte aleatori d'intercepció i pendent.

RLRT o proves de ràtio de versemblança restringida (r*estricted likelihood ratio tests*) : Proves d'hipòtesis realitzades als models que contenen efectes aleatoris per comprovar la significació d'aquests.

So: Nom donat dins d'aquest treball als conjunts de dades simulades per tenir un comportament que mostri pendent en les dades.

7 Bibliografia

Per facilitar la creació i referenciació de la bibliografia s'ha utilitzat la eina *plugin* de citació del programa *Mendeley Desktop* [1] compatible amb el programa d'edició de text utilitzat per redactar el treball *Libre Office Writer [28]*.

1. Mendeley Ltd. Mendeley Desktop v.1.19.5
2. ProjectLibre ProjectLibre open source v.1.8.0
3. RStudio Inc. RStudio v.1.1.463 for Ubuntu (Affero General Public License)
4. ICH guidelines (2003) Q1A (R2): Stability Testing of New Drug Substances and Products. Int Conf Harmon. https://doi.org/10.1136/bmj.333.7574.873-a
5. World Health Organization (2009) Stability testing of active pharmaceutical ingredients and finished pharmaceutical products. WHO Tech Rep Ser. https://doi.org/10.1016/j.jmhg.2006.09.007
6. Weiss RE (2005) Modeling Longitudinal Data. Springer New York, New York, NY
7. Oehlert G (2003) A First Course in Design and Analysis of Experiments
8. Jiang J (2007) Linear and generalized linear mixed models and their applications. Springer, New York, NY :
9. Everitt B, Hothorn T (2011) An Introduction to Applied Multivariate Analysis with R. Springer New York, New York, NY
10. Galecki A, Burzykowski T (2013) Linear Mixed-Effects Models Using R: Step by Step analysis
11. Quinn GP KM (2002) (2002) Experimental design and data analysis for biologists, 1st edn. Cambridge University Press, Cambridge
12. Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York [etc.] :
13. Kincaid C (2005) Guidelines for Selecting the Covariance Structure in Mixed Model Analysis. In: SUGI 30 Proceedings
14. Faraway JJ autor (2015) Linear models with R / Julian J. Faraway, University of Bath, United Kingdom. Taylor and Francis, Boca Raton, FL :

15. Curto García JJ (2018) Missing data analysis in longitudinal data. How to analyze it?

16. Little RJA, Rubin DB (2002) Statistical Analysis with Missing Data: Second Edition

17. Huque MH, Carlin JB, Simpson JA, Lee KJ (2018) A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol 18:168. https://doi.org/10.1186/s12874-018-0615-6

18. Peña D (2002) Análisis de datos multivariantes. Editor Mc Graw Hill Interam España, SAV. https://doi.org/8448136101

19. Lin TH (2010) A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. Qual Quant 44:277–287. https://doi.org/10.1007/s11135-008-9196-5

20. ICH (2004) Guidance for Industry Q1E Evaluation of Stability Data. Evaluation

21. The Pennsylvania State University (2017) Lesson 13 : Experiments with Random Factors 13 . 1 - Random Effects Models. In: Penn State Univ.

22. Everitt B (2005) An R and S-Plus companion to multivariate analysis. J Stat Softw

23. Canonical Ltd. Ubuntu Ubuntu 18.04.2 LTS

24. Graham JW (2012) Missing data [Recurs electrònic] : analysis and design / John W. Graham

25. Kowarik A, Templ M (2016) Imputation with the {R} Package {VIM}. J Stat Softw. https://doi.org/10.18637/jss.v074.i07

26. Duursma R (2017) Confidence intervals on predictions from mixed-effects models. http://www.remkoduursma.com/post/2017-06-15-bootpredictlme4/

27. Echeverría P, Negredo E, Carosi G, et al (2010) Similar antiviral efficacy and tolerability between efavirenz and lopinavir/ritonavir, administered with abacavir/lamivudine (Kivexa ®), in antiretroviral-naïve patients: A 48-week, multicentre, randomized study (Lake Study). Antiviral Res. https://doi.org/10.1016/j.antiviral.2009.11.008

28. LibreOffice contributors Libre Office Writer v.6.2.4.2

8 Annexos

Els annexos adjunts corresponen als informes complets dels codis de programació de R utilitzats en el format d'exportació de R markdown. En el cos del treball s'ha fet un resum de la programació aplicada i els resultats obtinguts, mentre que en aquests Annexos es mostra la programació completa amb totes les proves i gràfics addicionals que s'han realitzat en el transcurs de la part pràctica de programació:

8.1 Annex 1: Exportació del codi de programació R Markdown corresponent a la simulació de models mixtos en el context dels estudis d'estabilitat

8.2 Annex 2: Exportació del codi de programació R Markdown corresponent a la simulació de models amb dades perdudes

8.3 Annex 3: Exportació del codi de programació R Markdown corresponent a resolució del problema pràctic dins el context de l'aplicació dels models mixtos analitzat

8.4 Annex 4: Resultats R de les matrius de variància covariància i matrius de correlació en la simulació de models mixtos

8.5 Annex 5: Resums R dels models simple i d'interaccions de l'anàlisi pràctic LAKE1

9 Material addicional

Com a material addicional al TFM i corresponent als productes obtinguts s'adjunten els tres informes obtinguts de la compilació completa de R *markdown* corresponents als codis adjunts en els Annexos:

- Simulació de models mixtos en el context dels estudis d'estabilitat.
- Simulació de models amb dades perdudes.
- Anàlisi del cas pràctic dins el context de l'aplicació dels models mixtos.