Catarina João Mesquita da Silva

Efeito da hipertensão arterial crónica sobre o índice de massa corporal e pressão arterial média durante a gravidez

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2014

Catarina João Mesquita da Silva

Efeito da hipertensão arterial crónica sobre o índice de massa corporal e pressão arterial média durante a gravidez

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática

Orientador: Prof.ª Doutora Ana Rita Gaio

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2014

Agradecimentos

A realização deste trabalho só foi possível graças ao apoio, colaboração e encorajamento de algumas pessoas, as quais merecem todo o meu agradecimento e sincero obrigado.

Em particular, agradeço:

À Prof.ª Doutora Ana Rita Gaio por todo o apoio, paciência, todas as oportunidades dadas para aprender e evoluir, por toda a exigência e disponibilidade demonstrada ao longo deste ano para a realização deste estudo. Agradeço ainda tudo o que me ensinou e que em muito contribuiu para o meu conhecimento.

Ao Dr. Luís Guedes, médico ginecologista-obstetra no Centro Hospitalar Materno Infantil do Porto, pela proposta e cedência dos dados para análise deste estudo e por toda a paciência e compreensão revelada.

À Mariana Carvalho pela disponibilidade sempre mantida e ajuda na compreensão dos dados.

Aos cinco magníficos: Márcio, Ana, Rita, Catarina e Vânia por estarem sempre presentes e por todo o carinho e incentivo extra nesta fase.

Às grandes amizades reveladas neste mestrado, por toda a preocupação e por todas as tardes de ajuda permanente.

Aos meus pais, pelo carinho, apoio e todo o sacrifício suportado nesta etapa. Em especial, agradeço à pessoa que melhor me conhece e compreende, que sempre esteve comigo, que sofreu tanto ou mais que eu, e a quem devo tudo o que sou hoje. Obrigada Mãe!

A todos aqueles não mencionados que contribuíram de alguma forma para a finalização de mais uma etapa importante na minha vida, obrigada.

Resumo

Os estudos em dados longitudinais surgem quando um indivíduo é observado várias vezes ao longo do tempo. Desempenham um papel fundamental nas mais variadas áreas sendo possível estudar as alterações dentro do indivíduo e relacioná-las com fatores externos. Estes estudos constituem uma importante estratégia na investigação clínica, fornecendo conhecimentos sobre o desenvolvimento e persistência de doenças.
Este trabalho cumpre dois objetivos distintos. O primeiro consiste no estudo de modelos de regressão para dados longitudinais. O segundo é a aplicação destes modelos a dados reais na área da saúde: estudar a evolução do índice de massa corporal (IMC) e da pressão arterial média (PAM) ao longo da gestação em mulheres normotensas e com hipertensão arterial crónica. Os dados foram recolhidos pelo Dr. Luís Guedes-Martins, médico ginecologista-obstetra, do Centro Hospitalar Materno Infantil do Porto, que aliás também sugeriu o estudo.
Metodologicamente este trabalho estuda e aplica dois modelos distintos: o modelo linear de efeitos mistos (MLEM) e o modelo de regressão com estimação feita pelo método dos mínimos quadrados generalizados (MMQG). O primeiro modelo permite a inclusão de efeitos aleatórios, para além da consideração usual de efeitos fixos. Por outro lado, o modelo estidao pelo MMQG apenas inclui efeitos fixos. Ambos os modelos permitem uma modelação da matriz de variância-covariância dos erros.
Relativamente ao estudo da PAM, o modelo que melhor se ajustou aos dados foi um modelo com uma progressão temporal cúbica estimado pelo MMQG, que revelou ter todas as variáveis explicativas estatisticamente significativas. Em particular, a variável hipertensão tem um efeito aditivo sobre a PAM, estatísticamente significativo. No estudo do IMC, foi necessário remover algumas observações para garantir a normalidade dos erros. O modelo considerado mais adequado foi o modelo linear misto com progressão temporal quadrática e com efeito aleatório na constante. Este modelo apresentou significância estatística nas variáveis explicativas: $tempo$, $tempo^2$ e hipertensão. Para a população hipertensa, o seu IMC é mais alto do que o da população normotensa ao longo de toda a gravidez. Na relação entre a PAM e o IMC, o modelo que melhor se ajustou aos dados foi o modelo linear MMQG. Este modelo prevê que, para o mesmo instante de tempo, por cada aumento de uma unidade no IMC espera-se um aumento de 0.25 na PAM. Nas gestantes hipertensas verificou-se, como esperado, valores de PAM significativamente mais altos do que nas gestantes normotensas.
Palavras-chave: DADOS LONGITUDINAIS, MODELO LEM, EFEITOS ALEATÓRIOS, EFEITOS FIXOS, MODELO MMQG, GESTAÇÃO, HIPERTENSÃO ARTERIAL CRÓNICA, PAM, IMC.

Abstract

The studies in longitudinal data arise when an individual is observed repeatedly over time. They play a crucial role in the most varied fields and it is possible to study the changes inside the individual and to relate them with external factors. These studies are an important strategy in the clinical investigation, providing knowledge about the development and persistence of diseases.
This work fulfills two distinct goals. The first one consists of the study of regression models for longitudinal data. The second one is the application of these models to real data in the health field: study the evolution of the body mass index (BMI) and the mean arterial pressure (MAP) during pregnancy in normotensive and hypertensive women. The data was collected by Dr. Luis Guedes-Martins, specialist in Obstetrics and Gynecology, from Centro Hospitalar Materno Intantil do Porto, who who actually has also suggested the study.
Methodologically speaking, this thesis considers two classes of models: Linear Mixed-Effects Models (LMEM) and regression models with estimation based on the method of generalized least squares (GLS). The first type of models allow for the inclusion of random effects taking also into consideration fixed effects. On the other hand, models estimated by GLS only include fixed effects. Both models allow for a certain degree of modelling of the errors variance-covariance matrix.
Regarding the MAP study, the model with the best goodness-of-fit was a model with a cubic time progression estimated by GLS, with effects that were all statistically significant. In particular, hypertension was shown to have a statistically significant additive effect on MAP with no significant time interactions. In the study of BMI, some observations had to be removed in order to ensure errors normality. The model considered as the most appropriate was the linear mixed effects model with a quadratic time progression and with a random effect in the constant. Variables $time$, $time^2$ and hypertension were shown to be statistically significant. The mean BMI is higher than that in the normotensive population, across the whole pregnancy. For the evaluation of the longitudinal relationship between BMI and MAP, the model that better fits the data was the linear model GLS. This model predicts that for the same instant of time, each one-unit increase in BMI is expected to increase from 0.25 in MAP. In the hypertensive pregnants it was verified higher values of MAP than the normotensive pregnants, as expected.
Keywords:LONGITUDINAL DATA, LME MODEL, RANDOM EFFECTS, FIXED EFFECTS, GLS MODEL, PREGNANCY, CHRONIC HYPERTENSION, MAP, BMI.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Este trabalho engloba dois objetivos principais: estudar modelos de regressão para dados longitudinais e a sua aplicação em dados reais. Mais precisamente, através destes modelos perceber o comportamento do índice de massa corporal (IMC) e da pressão arterial média (PAM) ao longo da gestação de acordo com o estado hipertensivo da gestante.

Os estudos em dados longitudinais surgem quando existem observações repetidas para um mesmo indivíduo ao longo do tempo. Desempenham um papel fundamental nas mais variadas áreas de estudo na medida em que é possível estudar as alterações dentro do indivíduo e relacioná-las com diversos fatores externos. Assim, os estudos longitudinais constituem uma importante estratégia na investigação clínica, fornecendo conhecimentos sobre o desenvolvimento e persistência de doenças, os fatores que influenciam a sua alteração e permitem estudar o seu comportamento ao longo do tempo. Os modelos aqui estudados são os seguintes: o modelo linear misto e o modelo de regressão com estimação dos parâmetros feita pelo método dos mínimos quadrados generalizados (MMQG).

Os dados para este estudo foram recolhidos, ao longo da gestação, de grávidas saudáveis com consulta de rotina no Centro Hospitalar Materno Infantil do Porto. Esta recolha foi realizada entre Janeiro de 2010 e Dezembro de 2012 pelo Dr. Luís Guedes-Martins, médico ginecologista-obstetra na unidade de saúde referida anteriormente. Para o problema em causa consideram-se as seguintes variáveis: o IMC; a pressão arterial sistólica e diastólica, e portanto a PAM, e o estado hipertensivo da gestante. Esta dissertação incide no comportamento destas variáveis ao longo do tempo em gestantes hipertensas *versus* normotensas.

A hipertensão arterial crónica é altamente prevalente em Portugal. De acordo com um estudo anterior (Macedo et al., 2007), a sua prevalência é de 42% e no Norte este número é o mais baixo (33%). Sabe-se que nas mulheres, a hipertensão complica 6-8% das gestações (Barra et al., 2012).

O uso de medicamentos anti-hipertensivos antes da gravidez e da persistência de hipertensão superior a 12 semanas após o parto é bastante comum. Esta condição está

presente em até 5% das mulheres grávidas, e pode causar morbilidade e mortalidade maternal, fetal e neonatal [1], apesar da maioria das mulheres com hipertensão arterial crónica terem uma gravidez saudável e normal (Seely and Maxwell., 2007).
Sabe-se que durante a gestação, em mulheres saudáveis, a pressão arterial diminui até às 18-20 semanas de gestação e sobe até à altura do parto com valores idênticos aos encontrados no início da gravidez (Seely and Maxwell., 2007).

Relativamente ao excesso de peso, vem sendo observado um aumento da prevalência de obesidade em mulheres em idade reprodutiva e um aumento do seu IMC na gestação. Cerca de dois terços da população adulta portuguesa é considerada com valores de IMC altos. Atualmente, estima-se que em Portugal 38% das mulheres estão acima do peso e 20% são obesas. A prevalência da obesidade está a aumentar gradualmente em mulheres com idade fértil, o que também é uma preocupação por causa da sua associação com complicações na gravidez, como por exemplo, diabetes gestacional, entre outras (S. Paiva et al., 1998). De acordo com o IMC apresentado pela gestante, o seu aumento ao longo da gravidez deve ser corretamente avaliado.

Alguns estudos mostram a relação entre o índice de massa corporal (IMC) inicial e o ganho de peso durante a gravidez, mas em mulheres hipertensas crónicas, os dados são escassos ou inexistentes na literatura científica.
Assim, um dos objetivos desta dissertação é encontrar a evolução da PAM e do IMC durante a gestação na população e descrever as diferenças entre gestantes normotensas e hipertensas, como referido inicialmente.

A implementação dos modelos supra citados foi efetuada recorrendo a bibliotecas adequadas no software R versão 3.0.3 (R Development Core Team, 2012). Ao longo desta dissertação, sempre que uma biblioteca tenha sido usada, a sua designação será explicitamente mencionada.

1.1 Organização da Dissertação

Nesta secção descreve-se a estrutura desta dissertação. O **capítulo 1** de natureza introdutória; são descritos os principais objetivos deste trabalho e o contexto dos tópicos abordados na mesma.

O **capítulo 2** refere-se à contextualização clínica dos assuntos abordados neste estudo, permitindo ao leitor ter uma perceção real dos temas, do que é conhecido a nível científico sobre a PAM e sobre o IMC em gestantes normotensas e hipertensas. É um capítulo resumido que permitirá uma fácil inserção do leitor com conhecimentos matemáticos nestes tópicos.

No **capítulo 3** é apresentada a metodologia teórica relativa aos modelos abordados no desenvolvimento do estudo. É introduzida a teoria sobre modelos de efeitos mistos

[1] Morbilidade neonatal é o número de casos de doença até aos 28 dias de vida

seguido do método dos mínimos quadrados generalizados.

O **capítulo 4** apresenta uma breve descrição dos dados. Seguidamente, é realizada uma análise exploratória de forma a perceber como se comportam os dados e o que esperar na modelação posterior.

Os resultados obtidos são apresentados no **capítulo 5**. Este está dividido em três secções, uma abordando o estudo da pressão arterial média, a seguinte o estudo do índice de massa corporal e por fim é apresentado o estudo da relação entre ambas as variáveis. Em cada secção são apresentados e interpretados os resultados da aplicação dos modelos estudados no capítulo 3.

No **capítulo 6** são mencionadas as principais conclusões deste trabalho e as limitações inerentes ao estudo.

Capítulo 2

Contextualização

Neste capítulo é apresentada alguma informação clínica relativa à hipertensão arterial crónica e IMC durante a gravidez. Os conceitos aqui referidos são relevantes para a compreensão do presente estudo na medida em que facilitam a interpretação dos resultados obtidos.

2.1 Hipertensão Arterial Crónica

A hipertensão arterial crónica é uma doença caracterizada por uma elevação da pressão sanguínea, relacionada frequentemente com uma elevação da resistência vascular periférica. Na clínica, a pressão sanguínea pode ser obtida pela avaliação de duas medidas, sistólica e diastólica, referentes ao período de contração (sistólica) ou relaxamento (diastólica) que ocorrem durante o ciclo cardíaco.

Durante a gravidez, a hipertensão arterial é considerada crónica quando é diagnosticada antes das 20 semanas de gestação ou em situações em que ela persiste para além das 12 semanas pós-parto (Seely and Maxwell., 2007). Ela é definida conforme os valores das pressões arteriais. Como tal, pode ser diagnosticada, em duas avaliações sucessivas num intervalo mínimo de 4 horas, da seguinte forma:

- Pressão arterial sistólica de 140 mmHg ou superior;
- Ou pressão arterial diastólica de 90 mmHg ou superior;
- Ou ambas as condições anteriores (Obstetricians and Gynecologists, 2012).

Na grávida normotensa, a pressão arterial diminui até às 18-20 semanas de gestação e sofre um incremento significativo até à altura do parto, em valores semelhantes aos encontrados no inicio da gravidez (Macdonald-Wallis et al., 2012). Na maioria das mulheres com hipertensão arterial crónica, a pressão arterial segue o mesmo padrão.

Existem dois tipos de hipertensão arterial crónica: a hipertensão primária e a hipertensão secundária. Quanto à hipertensão primária ela é frequentemente designada

como essencial porque a sua causa não é conhecida. Contudo, a hipertensão secundária ocorre quando uma causa específica é objectivada. De referir que esta última ocorre em apenas numa minoria dos indivíduos. Em situações concretas, algumas grávidas necessitam de tratamento anti-hipertensivo agressivo com consequências fetais importantes, nomeadamente o parto pré-termo, morte fetal no útero, malformações congénitas, entre outras (Seely and Maxwell., 2007).
Ainda assim, os síndromes hipertensivos da gravidez encontram-se entre as principais causas de morbilidade e mortalidade materno-fetal, sendo que a terâpeutica anti-hipertensiva faz parte da prevenção das suas complicações. Portanto, é consensual que a terapia anti-hipertensiva é essencial no caso de hipertensão grave (Barra et al., 2012, Hermida et al., 2001).

Por vezes a hipertensão arterial crónica é confundida com a pré-eclampsia [1]. Contudo, neste último caso, a hipertensão apenas ocorre após as 20 semanas. Sabe-se que 30% ou mais das mulheres com hipertensão crónica ou hipertensão gestacional podem desenvolver pré-eclampsia. Quando isto acontece a doença é reclassificada como hipertensão crónica com pré-eclampsia sobreposta.

Devido à existência de diversas complicações associadas à hipertensão arterial crónica na gravidez é necessário um acompanhamento diferenciado (Barra et al., 2012, Seely and Maxwell., 2007).

Quando vigiada de forma adequada, a gravidez da maioria das mulheres com hipertensão arterial crónica decorre sem complicações maiores. Este objectivo é conseguido no contexto de cuidados de saúde multidisciplinares e estruturados de acordo com políticas de saúde que ajustam as medidas às necessidades deste grupo de risco (Seely and Maxwell., 2007).

2.2 Obesidade e índice de massa corporal na Gravidez

A prevalência de obesidade tem vindo a ocorrer com uma incidência crescente em todo o mundo, constituindo um importante problema de saúde pública. A situação mundial atual é tão marcada que a obesidade é referida como uma epidemia global (Mattar et al., 2009). O agravamento deste problema deve-se a uma série de fatores, sendo que os novos hábitos alimentares e estilo de vida sedentário são as principais causas do excesso de peso na população (Latifa Mochhoury and Barkat., 2013, Shub et al., 2013). Apesar de a predisposição genética ser importante na suscetibilidade individual para o ganho de peso, o equilíbrio energético é basicamente resultante da ingestão calórica e da atividade física. Aliado aos avanços dos meios de transporte e da disponibilidade de equipamentos que facilitam o desempenho de quase todas as atividades da vida

[1] Patologia que, aparentemente, começa a ocorrer no início da gravidez e é caraterizada por um aumento da pressão arterial associada ao aparecimento de uma quantidade anómala de proteínas na urina.

diária, o acesso fácil aos alimentos têm desempenhado a principal responsabilidade nesta problemática (Mattar et al., 2009).

Durante as últimas décadas, as mulheres são mais frequentemente obesas do que os homens com uma prevalência duas vezes superior (Nogueira and Carreiro, 2013). As mulheres obesas em idade fértil têm maior prevalência de infertilidade. Com efeito, a obesidade pode ser desencadeada ou agravada pela gravidez. Tem sido demonstrado que as mulheres grávidas com excesso de peso e obesidade subestimam o seu índice de massa corporal sendo estas últimas as que mais ganham peso durante o desenvolvimento fetal (Shub et al., 2013).
A obesidade materna predispõe a mãe à diabetes gestacional (DMG) e à diabetes tipo 2 (DM2), à hipertensão e até a doenças cardiovasculares, estando definitivamente associada a um risco aumentado de desfechos adversos durante a gravidez (Gomes et al., 2012, Shub et al., 2013). Complicações relacionadas com a obsedidade materna podem ser classificadas em dois grupos (Latifa Mochhoury and Barkat., 2013):

- as que afetam a mãe: diabetes gestacional (Grau de intolerância à glicose diagnosticada durante a gravidez), pré-eclâmpsia, cesariana emergente, entre outras;
- as que afetam o feto e o desenvolvimento do mesmo: macrossomia [2], prematuridade, morte fetal no útero.

O ganho de peso gestacional é definido como sendo a diferença entre o peso materno no momento do nascimento e aquele registado na primeira visita médica (Latifa Mochhoury and Barkat., 2013).
O Instituto de Medicina publicou referências para o ganho de peso tendo em conta o IMC antes da gravidez (Medicine, 1990).

- se o IMC for $< 19,8Kg/m^2$, o ganho de peso deve ser entre 12,5Kg e 18 Kg;
- se o IMC estiver entre $19,8Kg/m^2 e 26Kg/m^2$, o ganho de peso dever ser entre 11,5Kg e 16Kg;
- se o IMC for $> 26 - 29Kg/m^2$, o ganho de peso deve ser entre 7Kg e 11,5Kg;
- se o IMC for $> 29Kg/m^2$, o ganho de peso deve ser inferior a 7kg.

Existe uma associação frequente entre a hipertensão arterial e excesso de ganho de peso, pelo que é necessário valorizar o IMC e a PAM apresentado pela gestante, porque um pode ser determinado pelo outro. Por exemplo, para um valor de IMC elevado, demonstrou-se que a macrossomia fetal é mais frequente quando o ganho de peso materno excede os 8Kg durante toda a gravidez (Edwards et al., 1996).

[2] Doença que se caracteriza, principalmente, pelo excesso de peso do recém-nascido.

Capítulo 3

Componente Teórica

Neste capítulo é apresentado todo o contexto teórico necessário à realização deste estudo.

3.1 Dados Longitudinais

Dados longitudinais são dados em que a variável resposta é avaliada ao longo do tempo; mais precisamente, a resposta é medida no mesmo indivíduo em ocasiões (períodos) diferentes. (Cabral and Gonçalves, 2011, Twisk, 2003).

Estes tipos de dados podem ser obtidos de uma forma prospetiva ou retrospetiva. Se forem obtidos de uma forma prospectiva significa que os indivíduos são seguidos ao longo do tempo; se forem obtidos de uma forma retrospectiva significa que as diversas medições, para cada individuo, foram extraídas do seu historial (Cabral and Gonçalves, 2011).
Em estudos longitudinais as observações de um indívidiuo ao longo do tempo não são independentes umas das outras e, como tal, é necessário aplicar técnicas estatísticas que tenham em conta o facto de que as observações repetidas de cada indivíduo estão correlacionadas (Cabral and Gonçalves, 2011).
Os dados longitudinais têm uma caraterística particular: o facto de poderem ser dados agrupados. Os grupos são constituídos pelas medições repetidas sobre o mesmo indivíduo em diferentes ocasiões. As medições repetidas de cada vetor resposta tendem a ser correlacionadas, como dito anteriormente, e como tal, a estrutura de autocorrelação é fundamental na estimação dos parâmetros do modelo (Cabral and Gonçalves, 2011, Diggle et al., 2002).

Objetivo
Um estudo longitudinal tem como objetivo principal descrever as alterações da variável resposta ao longo do tempo e determinar se as alterações ocorridas dentro do indívidiuo se relacionam, ou não, com um conjunto de covariáveis previamente escolhidas (Cabral and Gonçalves, 2011, Fitzmaurice et al., 2004).

Dados omissos

A existência de dados omissos é um dos problemas deste tipo de estudos e portanto um fator importante na análise de dados longitudinias. O facto de os indivíduos terem um diferente número de observações e de estas terem sido feitas em ocasiões distintas leva à existência de valores omissos. A omissão de dados neste tipo de estudo traz implicações para o mesmo, sendo essas mesmas implicações as seguintes:

- perda de informação e redução da precisão da estimação da resposta ao longo do tempo;
- inferências incorretas para as alterações na resposta (Fitzmaurice et al., 2004).

É importante perceber o porquê de existirem dados omissos. O mecanismo de omissão de dados define-se como um modelo que descreve a probabilidade com que a resposta é ou não observada em determinado momento ou ocasião.

Estes modelos podem ser classificados da seguinte forma (Cabral and Gonçalves, 2011):

- omissão completamente aleatória (MCAR - missing completely at random)
- omissão aleatória (MAR - missing at random)
- omissão não aleatória (NMAR - not missing ar random)

Estamos perante uma omissão completamente aleatória (MCAR) quando o mecanismo de omissão em nada se relaciona com os valores observados da experiência, ou seja, quando por algum motivo externo se omitem valores. Neste tipo de omissão os dados observados podem ser considerados uma amostra aleatória dos dados completos.
Uma omissão aleatória (MAR), existe quando a probabilidade das respostas estarem omissas depende dos valores das respostas observadas, por exemplo, quando num estudo sobre a diminuição do índice de massa corporal, as pessoas que apresentam aumento do mesmo, têm tendência a abandonar o estudo. Neste tipo de omissão certos métodos de análise de dados longitudinais deixam de produzir estimativas válidas se a distribuição conjunta da resposta, não for corretamente especificada, ou mesmo se o mecanismo de omissão não for modelado corretamente. Quando utilizados, os métodos de máxima verosimilhança, levam a melhores estimativas e inferências neste tipo de omissão (Cabral and Gonçalves, 2011).
Quanto à omissão não aleatória, NMAR, esta existe quando a probabilidade da resposta estar omissa se encontra diretamente relacionada com valores que deveriam ter sido obtidos.

3.2 Modelo Linear Misto

Os modelos lineares de efeitos mistos são aplicados a dados agrupados (como, por exemplo, os dados longitudinais) e permitem estudar a relação entre uma variável resposta e uma ou mais covariáveis (Pinheiro and Bates, 2000). Estes modelos denominam-se modelos lineares mistos pois englobam dois tipos de efeitos: os fixos e os aleatórios. Os efeitos aleatórios permitem mostrar as alterações dentro de cada indivíduo e estão

associados aos indivíduos selecionados aleatoriamente da população, por outro lado, os efeitos fixos são parâmetros associados a toda a população.

3.2.1 Estrutura do Modelo

O modelo linear de efeitos mistos para um único nível de agrupamento, descrito por Laird and Ware (1982), é dado por:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{3.1}$$

com $i = 1, ..., n$ (n - número de indivíduos na amostra), onde $Y_i = (Y_{i1}, ..., Y_{iT_i})$ é o vector $T_i \times 1$ de respostas do individuo i, X_i é a matriz de desenho $T_i \times p$ de covariáveis dos efeitos fixos, β é o vector $p \times 1$ dos efeitos fixos, Z_i é a matriz $T_i \times q$ de covariáveis dos efeitos aleatórios, b_i é o vector $q \times 1$ dos efeitos aleatórios e ϵ_i é o vector $T_i \times 1$ dos erros aleatórios dentro do grupo i. As condições do modelo são $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, \Sigma_i)$ e b_i e ϵ_i independentes para os diferentes grupos i e entre si, com $i = 1, ..., n$, onde D é uma matriz $q \times q$ e Σ_i é uma matriz $T_i \times T_i$, ambas definidas positivas.

Com base no modelo (3.1) e nas condições supracitadas conclui-se que a distribuição de Y_i condicionada pelo efeito aleatório b_i é Gaussiana multivariada com valor médio $X_i\beta + Z_i b_i$ e matriz de variância-covariância Σ_i, ou seja,

$$Y_i \mid b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i). \tag{3.2}$$

onde a matriz Σ_i representa a variação intra-grupo (intra-indivíduo), quando o único nível de agrupamento é o indivíduo..

A função densidade de probabilidade (f.d.p.) correspondente é dada por

$$f(y_i|b_i) = (2\pi)^{-T_i/2}|\Sigma_i|^{-1/2} exp\left(-\frac{(Y_i - X_i\beta - Z_i b_i)^\top \Sigma_i^{-1}(Y_i - X_i\beta - Z_i b_i)}{2}\right) \tag{3.3}$$

Dados que a f.d.p. de b_i é:

$$f(b_i) = (2\pi)^{-q/2}|D|^{-1/2} exp\left(-\frac{b_i^\top D^{-1} b_i}{2}\right) \tag{3.4}$$

a f.d.p. marginal de Y_i é dada por:

$$\begin{aligned} f(y_i) &= \int f(y_i, b_i) db_i \\ &= \int f(y_i|b_i) f(b_i) db_i \\ &= (2\pi)^{-T_i/2}|V_i|^{-1/2} exp\left(-\frac{(Y_i - X_i\beta)^\top V_i^{-1}(Y_i - X_i\beta)}{2}\right). \end{aligned} \tag{3.5}$$

Portanto $f(y_i)$ é a função densidade de probabilidade da variável aleatória Gaussiana T_i-dimensional com valor médio $X_i\beta$ e com matriz de variância-covariância $V_i = Z_i D Z_i^\top + \Sigma_i$,

$$Y_i \sim N(X_i\beta, V_i). \tag{3.6}$$

A matriz V_i com dimensão $T_i \times T_i$ e definida positiva, representa a variação entre grupos (inter-indivíduos).

Para cada grupo, obtém-se o seguinte modelo

$$Y = X\beta + Zb + \epsilon, \tag{3.7}$$

onde Y é um vector $N \times 1$, $N = \sum_{i=1}^{n_i} T_i$, X é uma matriz $N \times p$, Z é uma matriz $N \times q$, b e ϵ, são vectores $qn \times 1$ e $N \times 1$, respetivamente.

Para o modelo (3.7), o valor esperado é dado por $X\beta$ e a sua variância por

$$V = Z\tilde{D}Z^\top + \Sigma,$$

onde $Z = diag(Z_1, ..., Z_n)$, $\tilde{D} = diag(D, ..., D)$, $\Sigma = diag(\Sigma_1, ..., \Sigma_n)$ e $V = diag(V_1, ...V_n)$ são matrizes diagonais por blocos de dimensão $N \times qn$, $qn \times qn$, $N \times N$ e $N \times N$ respetivamente. A distribuição da v.a. Y é Gaussiana multivariada dada por

$$Y \sim N(X\beta, V)$$

Portanto, o modelo llinear misto (3.1) pode ser definido através das f.d.p. $f(y_i|b_i)$ e $f(b_i)$. Esta definição é denominada por formulação hierárquica do modelo linear misto.

3.3 Estimação e Inferência no Modelo

Para a estimação dos parâmetros do modelo linear misto são usados dois métodos: o método da máxima verosimilhança (método ML[1]) e o método de máxima verosimilhança restrita (método REML[2]). Os parâmetros a estimar são β, o vector dos efeitos fixos, e as componentes D e Σ da matriz V.
Para cada método, tem-se que:

$$\Sigma_i = \sigma^2 I_{T_i} \; e \; D = \sigma^2 G, \tag{3.8}$$

Ao modelo 3.1 com matriz $\Sigma_i = \sigma^2 I_{T_i}$ dá-se o nome de modelo linear misto básico. A matriz Σ_i pode ainda assumir outras estruturas.
A matriz de variância-covariância de Y_i é da forma $V_i = \sigma^2(Z_i G Z_i^\top + I_{T_i}) = \sigma^2 M_i$, o mesmo se aplica a Y com $V = \sigma^2(Z\tilde{D}Z^\top + I) = \sigma^2 M$, para matrizes M_i e M.

Método da Máxima Verosimilhança

[1] Maximun Likelihood
[2] Restridted Maximum Likelihood

Seja β o vector dos efeitos fixos, θ o vector com todas as componentes da variância em G e $\alpha^\top = (\theta^\top, \sigma^\top)$. Assumindo a independência entre os indivíduos, a função de verosimilhança de uma amostra aleatória $Y_i, ..., Y_n$ é

$$L(y;\beta,\alpha) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{T_i/2}} |V_i(\alpha)|^{-1/2} exp\left(-\frac{(y_i - X_i\beta)^\top V_i^{-1}(\alpha)(y_i - X_i\beta)}{2}\right) \tag{3.9}$$

ou alternativamente,

$$L(y;\beta,\alpha,\sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{T_i/2}} |M_i(\theta)|^{-1/2} exp(-\frac{(y_i - X_i\beta)^\top M_i^{-1}(\theta)(y_i - X_i\beta)}{2\sigma^2}). \tag{3.10}$$

O respectivo logaritmo é dado por

$$l(y;\beta,\alpha) = -\frac{N}{2} log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} log|V_i(\alpha)| - \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^\top V_i^{-1}(\alpha)(y_i - X_i\beta) \tag{3.11}$$

e por

$$l(y;\beta,\alpha,\sigma^2) = -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n} log|M_i(\theta)| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^\top M_i^{-1}(\theta)(y_i - X_i\beta) \tag{3.12}$$

Os estimadores de máxima verosimilhança para os parâmetros do modelo podem ser obtidos da seguinte forma:

1. Para um dado α, ao igualar a zero a derivada parcial de (3.11) em ordem a β e resolvendo-a em ordem a esse mesmo parâmetro, obtém-se:

$$\begin{aligned} \hat{\beta}_{ML}(\alpha) &= \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} Y_i\right) \\ &= \left(X^\top V^{-1} X\right)^{-1} X^\top V^{-1} Y \end{aligned} \tag{3.13}$$

   Ver-se-á mais tarde que $\hat{\beta}_{ML}(\alpha)$ é o estimador dos mínimos quadrados generalizados (GLS) de β, assumindo-se que α é conhecido.

2. Para um dado θ, iguala-se a zero as derivadas parciais de (3.12) em ordem a β e a σ^2, e resolvendo-se o sistema em ordem a esses parâmetros obtém-se:

$$\hat{\beta}_{ML}(\theta) = \left(\sum_{i=1}^{n} X_i^\top M_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^\top M_i^{-1} Y_i\right) \tag{3.14}$$

   e

$$\hat{\sigma}^2_{ML}(\theta) = \frac{\sum_{i=1}^{n}\left(Y_i - X_i\hat{\beta}(\theta)\right)^\top M_i^{-1}\left(Y_i - X_i\hat{\beta}(\theta)\right)}{N} \tag{3.15}$$

3. Quando α é desconhecido, o estimador de máxima verosimilhança para α é obtido maximizando-se (3.11) com respeito a α, após β ter sido substituído por (3.13).

   Os estimadores ML para α e β são designados por $\hat{\alpha}_{ML}$ e $\hat{\beta}_{ML}$, respectivamente, sendo o estimador de $\hat{\beta}_{ML}$ dado por

$$\begin{aligned} \hat{\beta}_{ML} &= (\sum_{i=1}^{n} X_i^\top \hat{V}_i(\hat{\alpha}_{ML}) X_i)^{-1} X_i)^{-1} (\sum_{i=1}^{n} X_i^\top \hat{V}_i(\hat{\alpha}_{ML}) Y_i) \\ &= (X^\top \hat{V}(\hat{\alpha}_{ML}) X)^{-1} X^\top \hat{V}(\hat{\alpha}_{ML}) Y \end{aligned} \tag{3.16}$$

O processo de maximização de (3.9) e (3.10) (ou (3.11) e (3.12)) requer métodos numéricos de optimização.

Método da Máxima Verosimilhança Restrita
No método de máxima verosimilhança, a perda de graus de liberdade devido à estimação dos efeitos fixos não é tida em consideração, e portanto originam-se estimadores enviesados para as componentes da variância. Este problema é resolvido pelo método da máxima verosimilhança restrita, introduzido por Patterson & Thompson (1971) e mais tarde desenvolvido por Harville (1974).

O método de máxima verosimilhança restrita é baseado numa transformação de dados:

$$U = A^\top Y, \tag{3.17}$$

onde A é uma matriz $N \times (N-p)$ de característica completa ortogonal às colunas da matriz X, tal que a distribuição de U não depende de β. Uma outra maneira de se obter esta transformação é fazer-se $A = I - X(X^\top X)^{-1} X^\top$. Tem-se que $U \sim N(0, A^\top V(\alpha) A)$ para qualquer valor de β sendo que aos elementos de U chama-se contrastes de erros.

A função de verosimilhança para os contrastes dos erros (função de verosimilhança restrita) pode ser escrita da seguinte forma (Harville, 1974):

$$\begin{aligned} L_{REML}(y;\alpha) &= (2\pi)^{-(N-p)/2} |\sum_{i=1}^{n} X_i^\top X_i|^{1/2} \\ &\times |\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha) X_i|^{1/2} \prod_{i=1}^{n} |V_i(\alpha)|^{-1/2} \\ &\times exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^\top V_i^{-1}(\alpha)(y_i - X_i\hat{\beta}) \right) \end{aligned} \tag{3.18}$$

onde $\hat{\beta}$ é dado por (3.13) (Cabral and Gonçalves, 2011).

O logaritmo da função de verosimilhança restrita é portanto dado por:

$$
\begin{aligned}
l_R = logL_{REML}(y;\alpha) &= const - \frac{1}{2}log|\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha)X_i| \\
&- \frac{1}{2}\sum_{i=1}^{n} log|V_i(\alpha)| \\
&\times \left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\hat{\beta})^\top V_i^{-1}(\alpha)(y_i - X_i\hat{\beta})\right) \qquad (3.19)
\end{aligned}
$$

Maximizando a equação (3.19) em ordem a α obtêm-se os estimadores do método da máxima verosimilhança restrita de α, $\tilde{\alpha}_{REML}$ (Cabral and Gonçalves, 2011, Fitzmaurice et al., 2004, Pinheiro and Bates, 2000).

Comparando a expressão (3.11) com a (3.19) encontra-se a diferença entre ambas: o termo $-\frac{1}{2}log|\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha)X_i|$ existente em (3.19). Como $-\frac{1}{2}log|\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha)X_i|$ não depende de β, os estimadores da máxima
verosimilhança restrita de α e β podem ser obtidos maximizando-se a função de verosimilhança restrita dada por:

$$
L_{REML}(y;\beta,\alpha) = L_{ML}(y;\beta,\alpha)|\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha)X_i|^{-1/2} \qquad (3.20)
$$

com respeito a todos os parâmetros (α e β) (Cabral and Gonçalves, 2011).
Ao utilizar (3.19) ou (3.20) para obter o estimador pelo método da máxima verosimilhança restrita para α, o estimador pelo método da máxima verosimilhança restrita para β, $\hat{\beta}_{REML}$, é obtido substituindo na equação (3.20) V_i por $\hat{V}_i(\hat{\alpha}_{REML})$:

$$
\begin{aligned}
\hat{\beta}_{REML} &= \left(\sum_{i=1}^{n} X_i^\top \hat{V}_i^{-1}(\hat{\alpha}_{REML})X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i^\top \hat{V}_i^{-1}(\hat{\alpha}_{REML})Y_i\right) \\
&= \left(X^\top \hat{V}_{REML}^{-1}X\right)^{-1} X^\top \hat{V}_{REML}^{-1}Y
\end{aligned} \qquad (3.21)
$$

Para qualquer A (3.17), obtém-se os mesmos estimadores das componentes da variância, ou seja, os estimadores não dependem de A e o estimador $\hat{\beta}_{REML}$ não é igual ao estimador $\hat{\beta}_{ML}$ (Cabral and Gonçalves, 2011).

Na ausência de informação sobre β, nenhuma informação sobre α é perdida quando a inferência é baseada em U em vez de Y.

Método da Máxima Verosimilhança *versus* Método da Máxima Verosimilhança Restrita

O método de máxima verosimilhança fornece estimadores para os efeitos fixos, já o método de máxima verosimilhança restrita, por si só, não. Ambos os métodos (ML e

REML) baseiam-se na máxima verosimilhança, logo, produzem estimativas semelhantes. A grande diferença aumenta consoante o aumento do número de termos fixos no modelo. As estimativas pelo método da máxima verosimilhança das componentes da variância são menores do que as obtidas pelo método da máxima verosimilhança restrita (Pinheiro and Bates, 2000).

Tendo em atenção o termo $\frac{1}{2}log|\sum_{i=1}^{n} X_i^\top V_i^{-1}(\alpha)X_i|$ na equação (3.19) nota-se que qualquer alteração na matriz X dos efeitos fixos altera também $logL_{REML}(y;\alpha)$. Enquanto que os estimadores da máxima verosimilhança não variam consoante as reparametrizações dos efeitos fixos (isto é, alterações numa variável explicativa), o mesmo não acontece com os estimadores da máxima verosimilhança restrita (Pinheiro & Bates, 2000). Este aspecto inviabiliza a comparação, com base na função de verosimilhança restrita, de modelo lineares mistos com diferentes estruturas de efeitos fixos.

Os estimadores do método da máxima verosimilhança restrita para os parâmetros da variância e covariância são não enviesados, ao contrário dos obtidos pelo método da máxima verosimilhança (Pinheiro and Bates, 2000).

3.3.1 Efeitos: Fixos e Aleatórios

O verdadeiro interesse numa modelação de dados longitudinais que considerou um modelo linear de efeitos mistos reside na inferência dos efeitos fixos e dos efeitos aleatórios. Antes de se fazer qualquer inferência com base no estimador de um parâmetro deve ter-se em conta a distribuição desse mesmo estimador.

Distribuições Assintóticas

O estimador de β dado por (Cabral and Gonçalves, 2011):

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i^\top V_i^{-1} Y_i$$

tendo em conta a hipótese do modelo marginal dado por (3.6) e condicionado por α, tem distribuição Gaussiana multivariada com valor esperado

$$E(\hat{\beta}(\alpha)) = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} E[Y_i]\right) = \beta$$

e matriz de variância-covariância

$$\begin{aligned} var(\hat{\beta}(\alpha)) &= \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} var(Y_i) V_i^{-1} X_i\right) \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \\ &= \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1}. \end{aligned}$$

Na prática α é desconhecido, e como tal tem de se recorrer a resultados assintóticos.
Segundo Pinheiro (1994, *in* Pinheiro & Bates(2000)), sob certas condições de regularidade, os estimadores de máxima verosimilhança são consistentes e a distribuição assintótica é Gaussiana multivariada. A matriz de variância-covariância aproximada dos estimadores é dada pela inversa da matriz de informação de Fisher correspondente ao logaritmo da função de verosimilhança.

Dado que (Pinheiro and Bates, 2000):

$$E\left[\frac{\partial^2 \ell}{\partial \beta \partial \theta^\top}\right] = 0 \; e \; E\left[\frac{\partial^2 \ell}{\partial \beta \partial \sigma^2}\right] = 0$$

os estimadores de máxima verosimilhança dos efeitos fixos são assintoticamente não correlacionados com os estimadores de θ e σ^2. A distribuição assintótica dos estimadores de máxima verosimilhança é a seguinte:

$$\hat{\beta} \dot{\sim} N(\beta, \sigma^2 (X^\top M^{-1}(\theta) X)^{-1})$$

$$\begin{bmatrix} \hat{\theta} \\ log(\hat{\sigma}) \end{bmatrix} \dot{\sim} N\left(\begin{bmatrix} \theta \\ log(\sigma(\end{bmatrix}, I^{-1}(\theta, \sigma)\right) \tag{3.22}$$

com

$$I(\theta, \sigma) = \begin{bmatrix} \frac{\partial^2 \ell(\theta, \sigma^2)}{\partial \theta \partial \theta^\top} & \frac{\partial^2 \ell(\theta, \sigma^2)}{\partial log\sigma \partial \theta^\top} \\ \frac{\partial^2 \ell(\theta, \sigma^2)}{\partial \theta \partial log\sigma} & \frac{\partial^2 \ell(\theta, \sigma^2)}{\partial^2 log\sigma} \end{bmatrix} \tag{3.23}$$

onde $\ell(\theta, \sigma^2)$ é o logaritmo da função de verosimilhança nos efeitos fixos e $I(\theta, \sigma)$ é a matriz empírica de informação de Fisher.

Os efeitos aleatórios b_i são variáveis aleatórias e refletem o "desvio"na evolução do i-ésimo indivíduo em relação ao valor esperado da população, $X_i\beta$ (Cabral and Gonçalves, 2011, Fitzmaurice et al., 2004).

Melhor Preditor Linear Centrado (BLUP[3])

Sendo os b_i variáveis aleatórias, usam-se os métodos Bayesianos para obter os seus preditores (Fitzmaurice et al., 2004).

Como os dados são observados é possível obter a distribuição *a posteriori* de b_i. Viu-se anteriormente que a distribuição de Y_i condicional a b_i é $Y_i \mid b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i)$ e a distribuição de b_i é $N(0, D)$. De acordo com a abordagem Bayesiana, a distribuição *a priori* para b_i é $N(0, D)$. A distribuição *a posteriori* de b_i, definida como a distribuição de b_i condicional a Y_i, pode ser calculada. Portanto, pelo teorema de Bayes (Fitzmaurice et al., 2004):

[3]Best Linear Unbiased Predictor

$$f(b_i|y_i) = f(b_i|Y_i = y_i) = \frac{f(y_i|b_i)f(b_i)}{f(y_i|b_i)f(b_i)db_i}$$

Tendo em conta as propriedades do modelo Gaussiano (Azzalini, 1996) a distribuição conjunta de b_i e de Y_i é Gaussiana multivariada:

$$\begin{bmatrix} b_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ X_i\beta \end{bmatrix}, \begin{bmatrix} D & DZ_i^\top \\ Z_iD & Z_iDZ_i^\top + \Sigma_i \end{bmatrix} \right)$$

donde se conclui que

$$b_i|y_i \sim N\left(DZ_i^\top(Z_iDZ_i^\top + \Sigma_i)^{-1}(y_i - X_i\beta), D - DZ_i^\top(Z_iDZ_i^\top + \Sigma_i)^{-1}Z_iD\right).$$

Em particular, esta distribuição tem valor esperado dado por:

$$E(b_i|Y_i = y_i) \quad = \quad DZ_i^\top V_i^{-1}(y_i - X_i\beta). \tag{3.24}$$

Este valor esperado é o valor predito.
Assumindo que α é conhecido, o preditor b_i depende da covariância desconhecida entre o vetor resposta para cada indivíduo, como tal os parâmetros da covariância são substituídos pelas suas estimativas. Portanto, o melhor preditor linear centrado (BLUP) de b_i, é obtido substituindo-se, na expressão anterior, β por $\hat{\beta}(\alpha) = (X^\top V^{-1}X)^{-1}X^\top V^{-1}y$ e tem-se $\tilde{b}_i(\alpha) = DZ_i^\top V_i^{-1}(y_i - X_i\hat{\beta}(\alpha)$, ou $\tilde{b}(\alpha) = \tilde{D}Z^\top V^{-1}(y - X\hat{\beta}(\alpha)$. Condicional a α, o preditor da combinação linear $u = a_\beta^\top\beta + a_b^\top b_i$ do vetor dos efeitos fixos β e do vector b_i dos efeitos aleatórios, para os vectores a_β e a_b conhecidos de dimensão $p \times 1$ e $q \times 1$, respectivamente, é dado por:

$$\tilde{u}(\alpha) = a_\beta^\top\hat{\beta}(\alpha) + a_b^\top\tilde{b}_i(\alpha)$$

provando-se que $\tilde{u}(\alpha)$ é o BLUP de u (Cabral and Gonçalves, 2011).

3.3.2 Análise do Modelo

A primeira avaliação dos modelos é feita com base em testes de hipóteses, critérios de informação e análise de resíduos, de forma a que se possa obter o modelo que melhor se ajusta aos dados.

3.3.2.1 Testes de Hipóteses

Teste da Razão de Verosimilhanças

O teste da razão de verosimilhanças é utilizado para comparar modelos encaixados, isto é, modelos que diferem apenas na estrutura dos efeitos fixos e aí o conjunto dos parâmetros de um modelo é um subconjunto do conjunto de parâmetros de outro modelo.

A estatística de teste é dada por:

$$2log\left(\frac{L_1}{L_0}\right) = 2(logL_1 - logL_0) \tag{3.25}$$

onde L_1 é a verosimilhança do modelo mais geral, ou seja, com mais parâmetros, e L_0 é a verosimilhança do modelo encaixado. Este teste apresenta a seguinte hipótese nula: ambos os modelos apresentam igual qualidade de ajustamento aos dados.
A distribuição assintótica da estatística de teste é um qui-quadrado com $k_1 - k_0$ graus de liberdade ($\chi^2_{k_1-k_0}$), onde $k_1 - k_0$ é a diferença entre o número de parâmetros dos dois modelos.
Este teste é válido apenas se os estimadores dos parâmetros fixos nos dois modelos forem estimados pelo método da máxima verosimilhança, uma vez que o logaritmo da função de verosimilhança restrita é alterado se as especificações dos efeitos fixos forem igualmente alteradas.

Este teste realizado nestas circunstâncias tende a ser anti-conservativo, como tal o valor-p do mesmo é inferior ao verdadeiro valor-p do teste. À medida que aumenta a remoção de efeitos fixos do modelo mais pequeno, em comparação com o número total de observações, a imprecisão dos valores-p aumenta. Por este motivo, Pinheiro & Bates recomendam a utilização de testes-t e F aproximados para avaliar a significância dos efeitos fixos.

Teste-t e Teste-F aproximados

O teste-t avalia a significância marginal de cada parâmetro dos efeitos fixos quando todos os outros estão presentes no modelo (Pinheiro and Bates, 2000).
Para testar

$$H_0 : \beta_j = 0 \; vs \; H_1 : \beta_j \neq 0,$$

para algum $j = 1, ..., p$, utiliza-se a estatística de teste dada por:

$$\frac{\hat{\beta}_j}{\hat{\sigma}_{REML}\sqrt{\left[\left(\sum_{i=1}^{n} X_i^\top M_i^{-1}(\hat{\theta}) X_i\right)^{-1}\right]_{jj}}} \tag{3.26}$$

que, sob a hipótese nula, tem distribuição assimptótica t-Student com gl_j graus de liberdade. A estatística de teste está condicionada pelo estimador $\hat{\theta}$ (vector com todas as componentes da variância), sendo σ substituido pelo seu estimador dado pelo método da máxima verosimilhança restrita. Pinheiro & Bates (2000) desginam este teste por teste-t condicional.

O teste-F testa a significância de um ou mais termos dos efeitos fixos do modelo.
Para testar

$$H_0 : L\beta = 0 \; vs \; H_1 : L\beta \neq 0$$

onde L é uma matriz conhecida, é então usada a estatística de teste dada por:

$$F = \frac{\hat{\beta}^\top L^\top \left[L\hat{\sigma}_{REML} \left(\sum_{i=1}^{n} X_i^\top M_i^{-1}(\hat{\theta}) X_i \right)^{-1} L^\top \right] L\hat{\beta}}{r(L)} \tag{3.27}$$

que, sob a hipótese nula, tem distribuição assintótica F de Snedecor com (l, v) graus de liberdade. O número de graus de liberdade, l, do numerador do teste-F é dado pela característica da matriz L, $r(L)$.
Pelo mesmo argumento, Pinheiro & Bates (2000) designam este teste por teste-F condicional.

Para ambos os testes, existem diversos métodos para estimar o número de graus de liberdade do denominador, e os diferentes métodos conduzem a diferentes resultados. Na análise de dados longitudinais, os diferentes indivíduos contribuem com informação independente, o que se traduz num número de graus de liberdade suficientemente grande, qualquer que seja o método utilizado para o estimar, e consequentemente leva a valores-p muito semelhantes.

Em amostras pequenas há alguma incerteza associada à estimativa de θ que precisa ser identificada nas inferências sobre β. Esta fonte adicional de incerteza é reconhecida pelo uso das distibuições *t* e *F* (o que não é fácil considerar) em vez das distribuições normais e da qui-quadrado usuais padrão.

A utilização das distribuições qui-quadrado e normal são válidas quando o Σ ou θ são conhecidos, ou quando, o Σ foi estimado com um grande número de graus de liberdade. Com tamanhos de amostra pequenos, há alguma incerteza na estimativa de θ que devem ser tidas em conta e a utilização das distribuições *t* e *F* com graus de liberdade aproximados pelos métodos de Satterthwaite (1974), ou Kenward e Roger (1997), devem ser tidos em conta (Fitzmaurice et al., 2004).

Intervalos de Confiança

Os intervalos de confiança aproximados para os efeitos fixos são encontrados com base nas estatísticas dos teste-t aproximados.

Seja gl_j o número de graus de liberdade do teste-t correspondente ao j-ésimo efeito fixo. O intervalo de confiança aproximado para β_j com nível de confiança $(1 - \alpha)$ é:

$$\hat{\beta}_j \pm t_{(gl_j, 1-\alpha/2)} \hat{\sigma}_{REML} \sqrt{\left[\left(\sum_{i=1}^{n} X_i^\top M_i^{-1}(\hat{\theta}) X_i \right)^{-1} \right]_{jj}} \tag{3.28}$$

onde $t_{(gl_j, 1-\alpha/2)}$ representa o quantil $(1 - \alpha/2)$ da distribuição t-Student com gl_j graus de liberdade (Pinheiro and Bates, 2000).

Estes mesmos testes são também aplicados, de forma semelhante, na inferência sobre os efeitos aleatórios.

3.3.2.2 Critérios de Informação

Quando queremos comparar modelos não encaixados, o teste da razão de verosimilhanças não é indicado, quer se esteja a testar a significância dos efeitos fixos quer dos efeitos aleatórios. Assim, a comparação de modelos não encaixados é feita com base em critérios de informação.

Os critérios de informação aplicam-se a modelos construídos a partir da maximização do logaritmo da verosimilhança, penalizando os modelos com maior número de parâmetros. Os critérios de informação mais comuns são (Pinheiro and Bates, 2000):

- Critério de Informação de Akaike (AIC [4]), proposto por Akaike (1974), e dado por:

$$AIC = -2l(\hat{\beta}, \hat{\alpha}) + 2n_{par} \tag{3.29}$$

  onde n_{par} é o número de parâmetros do modelo;

- Critério de Informação Bayesiana (BIC [5]), também conhecido por *Schwarz's Bayesian Criterion (SBC)*, proposto por Schwarz (1978), e dado por:

$$BIC = -2l(\hat{\beta}, \hat{\alpha}) + 2n_{par}log(N) \tag{3.30}$$

  onde n_{par} é o número de parâmetros do modelo e N é o número total de observações.

Nestes critérios, quanto menor o valor do critério do modelo melhor é o mesmo. Assim, quando usado o critério AIC para comparação de dois ou mais modelos, escolhemos o modelo com menor valor de AIC, procedendo-se da mesma forma quando usado o critério BIC.
Os dois critérios são muito semelhantes, sendo o BIC mais sensível ao número de parâmetros incluídos no modelo, penalizando o modelo que tem mais parâmetros.

3.3.2.3 Resíduos

A análise dos resíduos é um meio usado para verificar se os pressupostos subjacentes ao modelo ajustado aos dados são válidos, servindo também para avaliar a qualidade do ajustamento do modelo. Os resíduos consistem na diferença entre a resposta observada e o respectivo valor ajustado pelo modelo dentro de cada grupo.
No caso do modelo linear de efeitos mistos as condições a verificar são (Pinheiro & Bates, 2000; Cabral and Gonçalves, 2011):

[4] *Akaike Information Criterion*
[5] *Bayesian Information Criterion*

- os **erros aleatórios** dentro do grupo são independentes e identicamente distribuídos, com distribuição Gaussiana de valor médio nulo e variância constante σ^2 e são independentes dos efeitos aleatórios.

- os **efeitos aleatórios** têm distribuição Gaussiana com valor médio nulo e matriz de variância-covariância D (não dependente do grupo) e são independentes para diferentes grupos.

O uso dos gráficos de diagnóstico são a forma mais usada para a verificação destas condições.

Os gráficos mais usados para avaliar as condições impostas aos erros aleatórios incluem as caixas de bigodes e histogramas dos resíduos por grupo e os gráficos dos resíduos padronizados *versus* os valores ajustados e *versus* as covariáveis de interesse. O gráfico dos resíduos padronizados *versus* os valores ajustados é usado para avaliar a suposição de variância constante. Também os gráficos dos valores observados *versus* os valores estimados e o gráfico da função de autocorrelação empírica são usados neste tipo de análise.

No caso do pressuposto da homoscedasticidade ou da independência dos resíduos dos erros aleatórios ser violada procede-se à modelação da matriz de variância-covariância dos mesmos.

3.4 Matriz de Variância-Covariância dos Erros Aleatórios

A modelação da matriz das componentes da variância, relativamente ao número de efeitos aleatórios e à estrutura de correlação entre eles, é fundamental para a interpretação da variação dos dados, e importante para a obtenção de inferências válidas para os parâmetros do modelo.
Para um dado β, a comparação de modelos encaixados com diferentes estruturas da matriz de variância-covariância dos efeitos aleatórios, corresponde a testar a hipótese nula de ter q efeitos aleatórios contra a hipótese alternativa de ter $q+k$ efeitos aleatórios ou testar a hipótese nula de ter q efeitos aleatórios independentes contra a hipótese alternativa de não serem independentes.
Como já foi referido os parâmetros do modelo devem ser estimados pelo método da máxima verosimilhança restrita; para testar as hipóteses referidas é usado o teste de razão de verosimilhanças.
Uma das condições exigidas para que a estatística de teste tenha, assintoticamente, uma distribuição qui-quadrado com número de graus de liberdade igual à diferença entre as dimensões dos espaços de parâmetros especificados em H_0 e H_1, é a de que a hipótese nula não esteja na fronteira do espaço de parâmetros.

Pinheiro & Bates (2000) aconselham a utilização *naive* da distribuição assintótica de um qui-quadrado, com um número de graus de liberdade dado pela diferença entre os

parâmetros estimados pelos modelos especificados nas hipóteses alternativa e nula, respectivamente [6].

Intervalos de Confiança

Os intervalos de confiança aproximados para as componentes da matriz de variância-covariância são obtidos através da distribuição assintótica dos estimadores dados pelos métodos da máxima verosimilhança e máxima verosimilhança restrita e dos teste-t aproximados (Pinheiro and Bates, 2000).

Designando-se por $[I^{-1}]_{\sigma\sigma}$ o último elemento da diagonal da inversa da matriz de informação de Fisher, um intervalo de confiança aproximado com nível de confiança $(1-\alpha)$ para o desvio padrão σ é:

$$\left[\hat{\sigma}exp\left(-z_{(1-\alpha/2)}\sqrt{[I^{-1}]_{\sigma\sigma}}\right), \hat{\sigma}exp\left(z_{(1-\alpha/2)}\sqrt{[I^{-1}]_{\sigma\sigma}}\right)\right] \tag{3.31}$$

em que $z_{(1-\alpha/2)}$ representa o quantil $(1-\alpha/2)$ da distribuição Gaussiana padrão.

O intervalo de confiança pode ser usado para o estimador do método da máxima verosimilhança e para o da máxima verosimilhança restrita, com as devidas alterações. Os intervalos de confiança para as componentes da matriz de variância-covariância dos efeitos aleatórios são mais difíceis de construir e são estimados com menor precisão do que os intervalos para os efeitos fixos e do que para o desvio padrão dentro dos grupos. O aumento da precisão na estimação dos intervalos de confiança para as componentes da variância só é possível com o aumento do número de grupos estudados (Pinheiro and Bates, 2000).

O modelo linear de efeitos mistos permite uma certa flexibilidade em relação aos efeitos aleatórios mas impõe a seguinte condição $\Sigma_i = \sigma^2 I_{T_i}$ para a estrutura dos erros aleatórios. Quando considerada no contexto dos dados longitudinais é pouco realista, pois neste tipo de dados as medições sobre o mesmo indivíduo estão geralmente correlacionadas pelo que esta estrutura não é adequada. Assim, serão apresentadas seguidamente estruturas de correlação e variância para a modelação da matriz dos erros aleatórios. (Cabral and Gonçalves, 2011).

Considere-se o modelo:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{3.32}$$

com a generalização $\epsilon_i \sim N(0, \sigma^2\Lambda_i)$, $i = 1, ..., n$, onde Λ_i é uma matriz $T_i \times T_i$ definida positiva parametrizada por um número de parâmetros que se designa por λ (Pinheiro and Bates, 2000).

A matriz Λ_i admite raiz quadrada invertível (Thisted, 1988) $\Lambda_i^{1/2}$, com inversa $\Lambda_i^{-1/2}$, de modo que:

[6] Solução existente na biblioteca *nlme* do R

$$\Lambda_i = (\Lambda_i^{1/2})^\top \Lambda_i^{1/2}$$

e

$$\Lambda_i^{-1} = \Lambda_i^{-1/2}(\Lambda_i^{-1/2})^\top.$$

Considere-se a reparametrização do modelo:

$$\begin{aligned} Y_i^* &= (\Lambda_i^{-1/2})^\top Y_i \\ X_i^* &= (\Lambda_i^{-1/2})^\top X_i \\ Z_i^* &= (\Lambda_i^{-1/2})^\top Z_i \\ \epsilon_i^* &= (\Lambda_i^{-1/2})^\top \epsilon_i. \end{aligned} \tag{3.33}$$

Tendo em conta que:

$$\epsilon_i^* \sim N\left[(\Lambda_i^{-1/2})^\top 0, \sigma^2(\Lambda_i^{-1/2})^\top \Lambda_i \Lambda_i^{-1/2}\right] = N(0, \sigma^2 I), \tag{3.34}$$

pode-se reescrever a equação 3.36 como

$$Y_i^* = X_i^*\beta + Z_i^* b_i + \epsilon_i^*, \tag{3.35}$$

onde $b_i \sim N(0, D)$ e $\epsilon_i^* \sim N(0, \sigma^2 I)$, $i = 1, ..., n$, isto é, Y_i^* é descrito através de um modelo linear misto básico (Pinheiro and Bates, 2000).

Atendendo a que $dy_i^* = |\Lambda_i^{-1/2}|$, a função de verosimilhança para o modelo (3.32) tendo em conta uma amostra aleatória $y = (y_1, ...y_n$ é dada por:

$$\begin{aligned} L(y; \beta, \theta, \sigma^2, \lambda) &= \prod_{i=1}^{n} f(y_i; \beta, \theta, \sigma^2, \lambda) \\ &= \prod_{i=1}^{n} f(y_i^*; \beta, \theta, \sigma^2, \lambda)|\Lambda_i^{-1/2}| \\ &= L(y^*; \beta, \theta, \sigma^2, \lambda) \prod_{i=1}^{n} |\Lambda_i^{-1/2}| \end{aligned} \tag{3.36}$$

onde f(.) é a função densidade de probabilidade de Y_i.
A função de verosimilhança do modelo (3.32) é a função de verosimilhança do modelo linear de efeitos fixos básico, logo os resultados apresentados nas secções anteriores são válidos. O mesmo se pode dizer para a função de verosimilhança restrita do modelo (3.35) que é dada por (Pinheiro and Bates, 2000):

$$L_{REML}(y; \theta, \sigma^2, \lambda) = \int L(y; \beta, \theta, \sigma^2, \lambda) d\beta = L_{REML}(y^*; \theta, \sigma^2, \lambda) \prod_{i=1}^{n} |\Lambda_i^{-1/2}| \tag{3.37}$$

3.4.1 Decomposição da Matriz

As matrizes Λ_i podem ser decompostas num produto de matrizes mais simples (Pinheiro and Bates, 2000):

$$\Lambda_i = W_i C_i W_i \tag{3.38}$$

onde W_i é uma matriz diagonal e C_i é uma matriz de correlação, isto é, uma matriz definida positiva com todos os elementos da diagonal iguais a 1. A matriz W_i não é única pois, podemos multiplicar cada uma das linhas por -1 e obter a mesma decomposição. Para garantir a sua unicidade impõe-se que W_i tenha todos os elementos da diagonal principal positivos. Por outro lado,

$$\begin{aligned} var(\epsilon_{it}) &= \sigma^2 \left[W_i\right]^2_{tt} \\ corr(\epsilon_{it}, \epsilon_{it'}) &= \left[C_i\right]_{tt'} , \end{aligned}$$

portanto W_i descreve a variância dos erros ϵ_i dentro do grupo e C_i descreve a correlação. Esta decomposição da matriz Λ_i em duas componentes, uma de estrutura de variância e outra de estrutura de correlação, permite a modelação destas estruturas separadamente dando ao modelo linear de efeitos mistos uma grande flexibilidade (Cabral and Gonçalves, 2011)

3.4.1.1 Heterocedasticidade

As funções de variância são usadas para modelar a estrutura de variância dos erros dentro de cada grupo.
A variância dos erros dentro do grupo associada ao modelo (3.32) pode escrever-se na forma (Cabral and Gonçalves, 2011):

$$var(\epsilon_{it}|b_i) = \sigma^2 g(\mu_{it}, \nu_{it}, \delta),\ i = 1, ..., n;\ t = 1, ..., T_i, \tag{3.39}$$

onde $\mu_{it} = E[y_{it}|b_i]$, ν_{it} é o vector de covariáveis, δ é o vector dos parâmetros da variância e $g(.)$ é a função de variância, contínua em δ. Esta função é escolhida de modo a refletir a variabilidade, por exemplo: função exponencial, logarítmica, potência ou uma combinação destas funções, descritas na tabela 3.1.

Classe	Variância ($var(\epsilon_{it})$)
VarFixed - Variância com uma única covariável	$\sigma^2 \nu_{it}$
VarIdent - Variâncias diferentes para cada categoria da covariável	$\sigma^2 \delta^2_{s_{it}}$
VarPower - Potência de uma covariável	$\sigma^2 \lvert\nu_{it}\rvert^{2\delta}$
VarExp - Exponencial de uma covariável	$\sigma^2 exp(2\delta\nu_{it})$
VarConstPower - Constante + Potência de uma covariável	$\sigma^2(\delta_1 + \lvert\nu_{it}\rvert^{\delta_2})^2$

ν_{it} - covariável; s_{it} - variável de estratificação; $\delta_1 > 0$

Tabela 3.1: Funções de variância para a modelação da heterocedasticidade

A formulação da função de variância permite que a variância por indivíduo dependa dos efeitos fixos β e dos efeitos aleatórios b_i, através dos valores esperados μ_{it}. Porém, coloca alguns problemas teóricos e computacionais, pois os erros dentro do grupo e os efeitos aleatórios deixam de ser independentes (Pinheiro and Bates, 2000). Assumindo que $E[\epsilon_{it}|b_i] = 0$ então $var(\epsilon_{it}) = E[var(\epsilon_{it}|b_i)]$, a dependência dos erros dentro do individuo, em relação aos efeitos aleatórios, pode ser evitada integrando-se em relação aos efeitos aleatórios.

Pelo facto de a função de variância não ser linear em b_i, a integração da mesma dada em (3.39) em relação aos efeitos aleatórios, é geralmente complicada do ponto de vista computacional. Assim, Davidian and Giltinian (1995) sugerem que se use um modelo aproximado em que os valores esperados μ_{it} são substituídos pelos seus BLUP $\hat{\mu}_{it}$
Os erros e os efeitos aleatórios deixam de estar correlacionados e portanto os resultados obtidos anteriormente continuam válidos.

3.4.1.2 Dependência

No contexto do modelo linear de efeitos mistos, as estruturas de correlação são usadas para modelar a dependência entre os erros dentro do grupo. É assumido que as estruturas de correlação são isotrópicas, isto é, a correlação entre dois erros ϵ_{it} e $\epsilon_{it'}$ dependem dos vectores de posição p_{it} e $p_{it'}$ através da distância entre os mesmos e não dos valores particulares que assumem (Pinheiro and Bates, 2000).

A expressão geral para a estrutura de correlação dentro do grupo é expressa, para $i = 1, ..., M$ e $j, j^{'} = 1, ..., T_i$, da seguinte forma:

$$corr(\epsilon_{it}, \epsilon_{it'}) = h[d(p_{it}, p_{it'}), \rho], \tag{3.40}$$

onde ρ é um vector de parâmetros de correlação e $h(.)$ é uma função de correlação que assume valores entre -1 e 1, contínua em ρ e tal que $h(0, \rho) = 1$. Em particular, quanto mais próximos, no espaço ou no tempo, estiverem dois erros aleatórios, maior será a sua dependência (Cabral and Gonçalves, 2011).

Estrutura de Correlação Serial Este tipo de estrutura é usado para modelar a dependência em dados de séries temporais, ou seja, em observações feitas sequencialmente ao longo do tempo. Simplificando o pressuposto de isotropia, o modelo de correlação serial é dado por:

$$corr(\epsilon_{it}, \epsilon_{it'}) = h[|p_{it} - p_{it'}|, \rho]$$

Sejam

$$r_{it} = (y_{it} - \hat{y}_{it})/\hat{\sigma}_{it}, \tag{3.41}$$

onde $\hat{\sigma}_{it}$ é o estimador da variância de ϵ_{it}, os resíduos padronizados do modelo ajustado. A função de autocorrelação no espaçamento (lag) l é dada por:

$$\hat{\rho}(l) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i - l} r_{it} r_{i(t+l)} / N(l)}{\sum_{i=1}^{n} \sum_{t=1}^{T_i} r_{it}^2 / N(0)} \tag{3.42}$$

onde $N(l)$ representa o número de pares de resíduos utilizados no somatório do numerador da função.

Quando as observações são igualmente espaçadas, o gráfico da função de autocorrelação empírica é usado para identificar o processo:

- se os valores se aproximam de zero gradualmente então o processo pode ser identificado como auto-regressivo,
- caso a função de autocorrelação seja consistente dentro de

  $$\pm z_{(1-\alpha/2)} / \sqrt{N(l)} z_{(1-\alpha/2)}$$

  após o lag 2 ou 3 então o modelo pode ser identificado como um processo de médias móveis de ordem 1 ou 2.

As estruturas de correlação serial mais usadas são:

Geral Cada correlação é dada por um parâmetro diferente. A função de correlação é:

$$h(k, \rho) = \rho_k, \ k = 1, 2, ... \tag{3.43}$$

Pelo facto do número de parâmetros em (3.43) aumentar quadraticamente com o número máximo de observações dentro do grupo, esta estrutura leva a modelos sobre-parametrizados, sendo útil apenas quando existem poucas observações por grupo (Cabral and Gonçalves, 2011, Pinheiro and Bates, 2000).

Simetria Composta Assume-se uma correlação igual entre todos os erros aleatórios dentro do mesmo grupo; isto é, para o mesmo indivíduo os erros correspondentes a diferentes tempos estão todos igualmente correlacionados. A função de correlação é:

$$corr(\epsilon_{it}, \epsilon_{it'}) = \rho, \forall t \neq t', \ h(k, \rho) = \rho, \ k = 1, 2, ...$$

O único parâmetro de correlação ρ é designado por coeficiente de correlação intraclasse.

É bastante útil quando todas a observações dentro do grupo são recolhidas ao mesmo tempo (Pinheiro and Bates, 2000).

Auto-regressivo - Médias Móveis É uma família de estruturas de correlação que inclui diferentes classes de modelos lineares estacionários: modelos auto regressivos (AR), modelos de médias móveis (MA) e modelos auto-regressivos de médias móveis (ARMA).

Assume-se que as observações são feitas em intervalos de tempo inteiros. Para simplificar omite-se o índice referente ao individuo, pelo que, ϵ_t designa a observação que ocorreu no instante de tempo t. A distância (lag), entre duas observações ϵ_t e ϵ_s é dada por $|t-s|$, logo $lag1$ refere-se a observações feitas com uma unidade de distância, $lag2$ a observações feitas com duas unidades de distância, e assim sucessivamente (Pinheiro and Bates, 2000).

- **Modelos Auto-regressivos - AR** Modelos que exprimem uma observação como combinação linear das observações anteriores acrescida de um ruído homocedástico, a_t, centrado em zero $E[a_t] = 0$ e independente das observações anteriores

$$\epsilon_t = \phi_1 \epsilon_{t-1} + ... + \phi_p \epsilon_{t-p} + a_t,$$

  onde p é o número das observações anteriores incluídas no modelo linear e designado por ordem do modelo auto-regressivo. Escreve-se $AR(p)$. Assim, existem p parâmetros de autocorrelação num modelo $AR(p)$ dados por $\Phi = (\phi_1, ..., \phi_p)$.

- **Modelos de Médias Móveis - MA** Assume-se que qualquer observação é uma combinação linear de termos de ruído, ou seja,

$$\epsilon_t = \theta_1 a_{t-1} + ... + \theta_q a_{t-q} + a_t, \tag{3.44}$$

  onde q é número de termos de ruído incluídos no modelo linear. O valor de q é designado a ordem do modelo de médias móveis, $MA(q)$. Existem q parâmetros de correlação dados por $\theta = (\theta_1, ..., \theta_q)$ e a função de correlação para um modelo $MA(q)$ é a seguinte:

$$h(k,\theta) = \begin{cases} \dfrac{\theta_k + \theta_1\theta_{k-1} + ... + \theta_{k-q}\theta_q}{1 + \theta_1^2 + ... + \theta_q^2}, & k = 1, ..., q \\ 0, & k = q+1, q+2, ... \end{cases} \tag{3.45}$$

  As observações separadas por mais do que q unidades de tempo não estão correlacionadas, uma vez que não partilham qualquer termo de ruído.

  O modelo mais simples é o modelo de ordem 1, $MA(1)$:

$$h(1,\theta) = \rho_1 = \frac{\theta_1}{1+\theta_1^2},\ |\rho_1| < 0.5.$$

- **Modelo Auto-regressivos de Médias Móveis - ARMA** Estes modelos como o próprio nome indica, são obtidos combinando os modelos auto-regressivos e os modelos de médias móveis . Um modelo $ARMA(p,q)$ é dado por:

$$\epsilon_t = \sum_{i=1}^{p} \phi_i \epsilon_{t-i} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t.$$

Este modelo tem $p+q$ parâmetros de correlação, que correspondem à combinação dos p parâmetros auto-regressivos $\Phi = (\phi_1, ..., \phi_p)$ e dos q parâmetros de médias móveis $\theta = (\theta_1, ..., \theta_q)$.

A função de correlação de um modelo $ARMA(p, q)$ é a seguinte:

$$h(k,\rho) = \begin{cases} \phi_1 h(|k-1|,\rho) + ... + \phi_p h(|k-p|,\rho) + \theta_1 \psi(k-1,\rho) + ... + \theta_q \psi(k-q,\rho), & k = 1, ..., q \\ \phi_1 h(|k-1|,\rho) + ... + \phi_p h(|k-p|,\rho), & k = q+1, q+2, ... \end{cases} \tag{3.46}$$

onde $\psi(k, \phi, \theta) = \dfrac{E[\epsilon_{t-k} a_t]}{var(\epsilon_t)}$. Note-se que $\psi(k, \phi, \theta) = 0$ para $k = 1, 2, ...$ dado que, neste caso ϵ_{t-k} e a_t são independentes e $E[a_t] = 0$.

O modelo $ARMA(1, 1)$ é um modelo "intermédio"entre os modelos $AR(1)$ e $MA(2)$ que apresenta um decaimento exponencial da função de autocorrelação para $lags \geq 2$ mas que permite maior flexibilidade na primeira autocorrelação (Pinheiro and Bates, 2000):

$$\begin{aligned} \epsilon_t &= \phi \epsilon_{t-1} + \theta a_{t-j} + a_t \\ h(1,\rho) &= \rho_1 = \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 + \theta_1^2 + 2\phi_1\theta_1} \\ h(k,\rho) &= \rho_k = \phi_1^{k-1}\rho_1, k \geq 2. \end{aligned}$$

3.5 Método dos Mínimos Quadrados Generalizados

Os método dos mínimos quadrados generalizados (*Generalized Least Squares*) permite ajustar modelos com erros heterocedásticos e correlacionados dentro do grupo, mas com a seguinte particularidade: não inclui efeitos aleatórios. O objetivo é encontrar o melhor modelo que se ajusta a um determinado conjunto de dados tentando minimizar a soma dos quadrados dos resíduos.

3.5.1 Descrição do Método

Considere-se o modelo dado por (Fox and Weisberg, 2010, Kariya and Kurata, 2004, Pinheiro and Bates, 2000):

$$Y_i = X_i\beta + \epsilon_i, \tag{3.47}$$

onde $i = 1, ..., n$ (n - número de indivíduos), onde $Y_i = (Y_{i1}, ..., Y_{iT_i})$ é o vector resposta para o individuo i, X_i é a matriz de covariáveis dos efeitos fixos, β é o vector dos

coeficientes estimados pela regressão dos efeitos fixos, ϵ_i é o vector dos erros dentro do grupo i e onde $\epsilon_i \sim N(0, \sigma^2\Lambda_i)$ onde Λ_i é uma matriz definida positiva parametrizada por um número de parâmetros que se designa por λ.

Uma vez que não se consideram efeitos aleatórios este modelo é uma simplificação do modelo linear de efeitos mistos. Usando a mesma transformação que foi usada na secção da modelação da matriz de variância-covariância (3.4) podemos reescrever o modelo (3.47) como um modelo de regressão linear com erros independentes e homocedásticos (Fox and Weisberg, 2010):

$$Y_i^* = X_i^*\beta + \epsilon_i^*, \tag{3.48}$$

com $i = 1, ..., n$ (n - número de indivíduos) e onde $\epsilon_i \sim N(0, \sigma^2 I)$.

3.5.2 Estimação dos Parâmetros

Atendendo à expressão da função de máxima verosimilhança, o logaritmo da mesma é uma função dependente apenas de λ (parâmetro da matriz de variância-covariância dos erros) dada por (Kariya and Kurata, 2004, Pinheiro and Bates, 2000):

$$l(\lambda|y) \quad = \quad const - Nlog\|y^* - X^*\hat{\beta}(\lambda)\| - \frac{1}{2}\sum_{i=1}^{n} log|\lambda_i| \tag{3.49}$$

O logaritmo da função de máxima verosimilhança restrita é dado por (Harville, 1974):

$$l_R(\lambda|y) \quad = \quad const - (N-p)log\|y^* - X^*\hat{\beta}(\lambda)\| - \frac{1}{2}|(X^*)^\top X^*| - \frac{1}{2}\sum_{i=1}^{n} log|\lambda_i| \tag{3.50}$$

onde *p* representa a dimensão do vetor β.

Para um dado λ fixo os estimadores de máxima verosimilhança de β e σ^2 são obtidos através de um problema de mínimos quadrados generalizados que resultam de uma redução do problema a uma aplicação do método de mínimos quadrados ordinários e consequente otimização da função de verosimilhança. Denotando X^* a matriz de empilhamento de todas as matrizes X_i correspondentes aos diferentes individuos, os estimadores de máxima verosimilhança de β e de σ^2 são:

$$\begin{aligned} \hat{\beta}(\lambda) &= \left((X^*)^\top X^*\right)^{-1}(X^*)^\top y^* \\ \hat{\sigma^2}(\lambda) &= \frac{\|y^* - X^*\hat{\beta}(\lambda)\|^2}{N} \end{aligned} \tag{3.51}$$

O estimador de máxima verosimilhança restrita de σ^2 é o seguinte:

$$\hat{\sigma^2}(\alpha) = \frac{\|y^* - X^*\hat{\beta}(\alpha)\|^2}{N-p}$$

3.5.3 Análise do Modelo

Dado que o método dos mínimos quadrados generalizados é aplicado em modelos que incluem apenas efeitos fixos, os testes de hipóteses utilizados para avaliar estes modelos são os mesmos que os realizados nos modelos de efeitos mistos: teste de razão de verosimilhanças, teste t e F (secção 3.3.2.1), assim como os critérios de informação e a análise dos resíduos.

3.5.4 Matriz de Variância-Covariância

A matriz de variância-covariância do vetor resposta Y_i no modelo linear misto é dada por $V_i = \sigma^2(Z_i G Z_i^\top + \Lambda_i)$. Esta decomposição permite modelar a heterocedasticidade e a correlação através de duas componentes: a componente dos efeitos aleatórios dada por $\sigma^2(Z_i G Z_i^\top)$, e uma outra componente referente ao grupo Λ_i.

No método dos minimos quadrados generalizados como se pretende não incorporar efeitos aleatórios escolhe-se a segunda componente da matriz de variância-covariância, Λ_i, do vetor resposta Y_i para modelar diretamente a estrutura de variância-covariância da resposta (Pinheiro and Bates, 2000).

Como visto na secção 3.4.1 tem-se ainda que a matriz Λ_i pode ser decomposta no produto:

$$\Lambda_i = W_i C_i W_i \tag{3.52}$$

onde W_i é uma matriz diagonal e C_i é uma matriz de correlação.
Sabe-se que:

$$var(\epsilon_{it}) = \sigma^2 \left[W_i\right]^2_{tt}$$
$$corr(\epsilon_{it}, \epsilon_{it'}) = \left[C_i\right]_{tt'},$$

onde W_i descreve a variância dos erros ϵ_i dentro do grupo e C_i descreve a correlação. Esta decomposição permite estudar separadamente a estrutura de variância e a estrutura de correlação.

Análogo ao que foi visto sobre a heterocedasticidade na secção 3.4.1.1, e uma vez que no estudo deste modelo não se incorpora efeitos aleatórios, a função de variância para os erros dentro do grupo difere um pouco da do modelo linear de efeitos mistos. Assim, no caso do modelo de regressão com estimação pelo método dos mínimos quadrados generalizados a função de variância é dada por:

$$var(\epsilon_{it}) = \sigma^2 g^2(\mu_{it}, \nu_{it}, \delta),\ i = 1, ..., n;\ t = 1, ..., T_i, \tag{3.53}$$

onde $\mu_{it} = E[y_{it}]$, ν_{it} é o vector de covariáveis, δ é o vector dos parâmetros da variância e $g(.)$ é a função de variância, contínua em δ.
Quanto à função de correlação é igual à vista no modelo linear de efeitos mistos.
Relativamente ao estudo e utilização das estruturas de variância para modelar a heterocedasticidade dos erros, podem ser aplicadas as estruturas vistas na secção 3.4.1.1. O mesmo acontece com as estruturas de correlação utilizadas para modelar a dependência entre os erros, secção 3.4.1.2.

Capítulo 4

Análise dos Dados

4.1 Objetivo do Estudo

O principal objetivo deste trabalho foi estudar a variação do IMC e a variação da PAM ao longo da gravidez em gestantes normotensas e hipertensas, perceber quais as diferenças entre estes dois grupos de gestantes. Foi ainda realizada uma análise sobre a relação entre a PAM e o IMC.

Os dados utilizados neste trabalho foram recolhidos em pacientes do Centro Hospitalar Materno Infantil do Porto. Neste estudo, foram analisadas 461 mulheres grávidas com uma gestação não complicada, sendo que 93% são normotensas e 7% são hipertensas. A orientação médica deste trabalho foi realizada pelo Dr. Luís Guedes-Martins.

De entre todas as variáveis da base de dados do estudo constam as seguintes: peso, altura, idade, paridade, menarca, sexo fetal, peso fetal, pressão arterial sistólica, pressão arterial diastólica e o índice de apgar. A variável paridade indica o número de gestações de uma mulher: neste estudo toma o valor 0 se a mulher se encontra na sua primeira gravidez, e toma o valor 1 se a gestante já teve um ou mais filhos.

4.2 Base de Dados

Os dados em estudo foram recolhidos pelo Dr. Luís Guedes. A base encontrava-se em formato largo (*wide format*) e foi necessário converter a base em formato longo (*long format*) de modo a prosseguir-se com o estudo de forma correta. Utilizou-se no software *R* o comando *reshape*, que se encontra na biblioteca *Reshape2*, que permite realizar esta transformação da base.

As variáveis peso, pressão arterial sistólica e pressão arterial diastólica foram avaliadas em determinados momentos específicos da gravidez. O peso foi registado em cinco períodos diferentes: início da gravidez, 12-14 semanas, 18-22 semanas, 29-33 semanas e no momento do parto. As pressões arteriais, sistólica e diastólica, foram apenas

avaliadas desde as 12-14 semanas até ao parto, tendo em conta os intervalos de tempo anteriores.

Os cinco períodos de avaliação foram codificados da seguinte forma: o momento inicial foi considerado o tempo zero e o momento do parto o tempo um. Em relação aos restantes três períodos procedeu-se de outra forma: encontrou-se o ponto médio do período em semanas e dividiu-se pelo número médio de semanas de uma gestação (38 semanas). Assim, por exemplo, o período referente ás 12-14 semanas de registo passou a ser o momento 0.3, pois o ponto médio entre 12 e 14 é o 13, que dividido por 38 é 0.3.

Momento do registo dos dados	Respetivo tempo
Ínicio	tempo 0
12-14 semanas	tempo 0.3
18-22 semanas	tempo 0.5
29-33 semanas	tempo 0.8
Parto	tempo 1

Tabela 4.1: Codificação dos momentos do registo dos dados

As características clínicas da base de dados em estudo são apresentadas nas tabelas abaixo representadas. Os testes de hipótese utilizados foram os seguintes: teste do Qui-Quadrado ou teste de Fisher para comparar as frequências de uma variável categórica ou para estudar a independência entre dois fatores, e também o teste-t para avaliar a significância estatística da diferença entre as médias de duas populações.

		n(%)	p-value	NT n=429	HT n=32	p-value
Idade (anos)	16-24	102(22%)	< 0.001	102(24%)	0	< 0.001
	25-35	303(66%)		285(66%)	18(56%)	
	36-43	56(12%)		42(10%)	14(44%)	
Menarca	11.84(1.24)	-	NA	11.76(1.16)	12.84(1.74)	0.002
Paridade	0	238(52%)	0.485	226(53%)	12(38%)	0.140
	≥ 1	223(48%)		203(47%)	20(62%)	
Idade gestacional (semanas)	39.22(1.20)	-	NA	39.24(1.17)	38.93(1.68)	0.308
Sexo fetal	1	245(53%)	0.177	229(53%)	16(50%)	0.852
	2	216(47%)		200(47%)	16(50%)	
Peso fetal (gramas)	3128(334)	-	NA	3136(329)	3007(379)	0.070
Índice de Apgar após 5minutos o nascimento	< 7	0	NA	0(0%)	0(0%)	< 0.001
	7-10	461(100%)		429(100%)	32(100%)	

Tabela 4.2: Normotensas *versus* Hipertensas

A tabela 4.2 mostra que a maioria das grávidas têm idade compreendida ente os 25 e 35 anos (66%), 22% e 12% têm entre 16-24 e 36-43 anos, respetivamente. Em mulheres grávidas normotensas, 24% pertencem ao grupo mais jovem (idade entre 16 e 24), 66%

têm idade entre 25 e 35 anos e 10% correspondem às mais velhas, com idade entre 36 e 43. A maioria das mulheres grávidas com hipertensão arterial crónica forma o grupo com idade intermédia, 56%. De salientar que não existem mulheres com hipertensão no grupo etário mais novo. Nesta amostra, 52% das mulheres estão na sua primeira gravidez. A maioria das mulheres normotensas não tem filhos (53%) e a maioria das hipertensas já tinha dado à luz um ou mais filhos (62%). No entanto, esta diferença não é estatisticamente significativa (p = 0.140). A idade média da menarca é 11.76 anos em normotensas e 12.84 anos nas hipertensas com diferenças significativas (p = 0.002). A distribuição dos recém-nascidos por sexo é similar (p = 0.852). O peso fetal médio e a idade gestacional média ao nascer é de 3.128 gramas e 39.22 semanas, respetivamente. Entre os grupos NT e HT não há diferenças estatisticamente significativas relativamente a estas variáveis (p = 0.070, p = 0.308). Quanto ao índice de apgar o resultado reflete o bem-estar geral e o grau de asfixia do recém-nascido . Uma pontuação superior a 7 é normal e menor pode indicar asfixia leve. Cinco minutos após o nascimento, todos os recém-nascidos tiveram uma pontuação acima de 7.

	Períodos	Normotensas	Hipertensas
Peso (kg)	Inicial	64.09±12.65	75.17±17.06
	12-14 semanas	65.94±13.03	76.83±17.67
	18-22 semanas	69.82±13.69	78.69±16.96
	29-33 semanas	75.92±13.99	85.03±16.34
	Parto	80.96±14.12	92.95±16.93
Pressão arterial sistólica (mmHg)	Inicial	-	-
	12-14 semanas	119.79±10.62	136.22±9.36
	18-22 semanas	114.37±10.28	123.65±9.29
	29-33 semanas	119.91±11.03	140.81±8.52
	Parto	121.50±11.66	143.75±8.84
Pressão arterial diastólica (mmHg)	Inicial	-	-
	12-14 semanas	63.05±8.15	76.09±6.89
	18-22 semanas	64.09±11.44	72.62±6.84
	29-33 semanas	62.93±10.85	78.94±8.05
	Parto	66.13±11.54	76.88±9.21
Índice de massa corporal (kg/m^2)	Inicial	25.10±5.18	28.76 ±6.54
	12-14 semanas	25.83±5.38	29.40±6.79
	18-22 semanas	27.35±5.66	30.12±6.52
	29-33 semanas	29.75±5.86	32.56±6.39
	Parto	31.73±5.98	35.59±6.59
Pressão arterial média (mmHg)	Inicial	-	-
	12-14 semanas	81.96±6.75	96.14±6.01
	18-22 semanas	80.85±8.52	86.64±4.45
	29-33 semanas	81.93±8.28	99.56±6.45
	Parto	84.59±8.82	99.16±6.91

Tabela 4.3: Descrição dos diferentes períodos

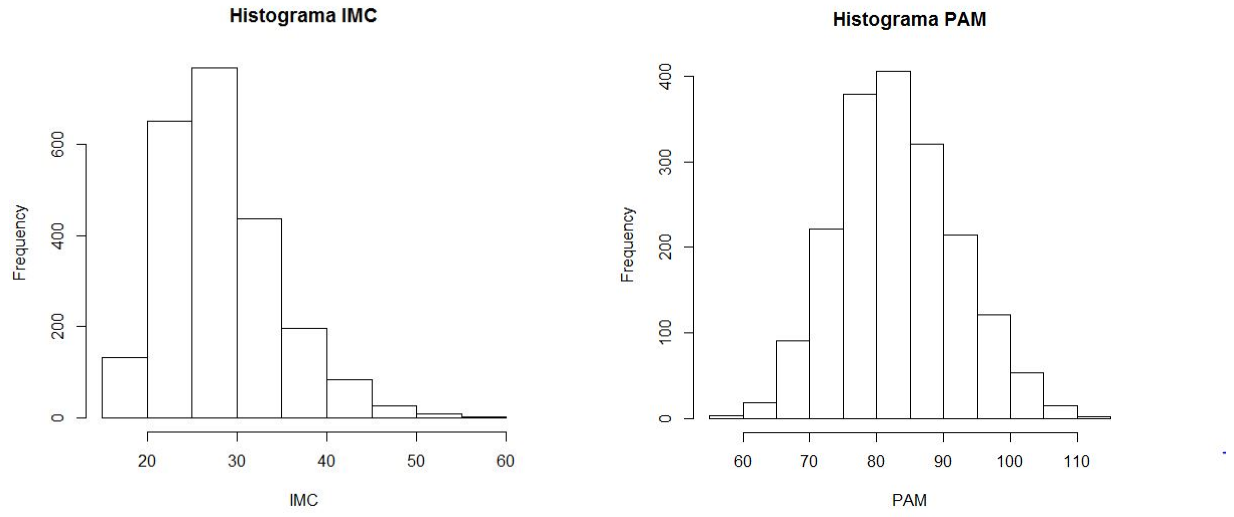
A tabela 4.3 descreve os valores do peso e da pressão arterial nos diferentes momentos da gestação. Para o cálculo da pressão arterial média (PAM) foi usada a seguinte expressão: PAM $= \frac{PAS + 2 \times PAD}{3}$; para o cálculo do índice de massa corporal foi encontrado através da seguinte expressão: $IMC = \frac{Peso}{Altura^2}$

As mulheres normotensas iniciam a sua gravidez com um peso médio de 64.09 (DP:

12.65)Kg e terminam com 80.96 (DP: 14,12)Kg. Por outro lado, as mulheres hipertensas iniciam a gravidez, com um peso mais alto: 75.17 (DP: 17.06)Kg e como tal terminam a gestação também com um peso mais elevado: 92.95 (DP: 16.93)Kg. Em relação à pressão arterial, a pressão arterial sistólica e diastólica diminuem durante o segundo trimestre e aumentam até ao parto, para ambos os grupos, o mesmo se verifica para a pressão arterial média, não sendo uma descida acentuada. Como esperado, a pressão sanguínea, quer sistólica ou diastólica, são maiores para as mulheres hipertensas. Relativamente ao índice de massa corporal, o mesmo está relacionado com o peso e a altura da gestante. Assim,as gestantes hipertensas apresentam um valor de IMC mais alto que as normotensas, 28.76 $\pm 6.54\ kg/m^2$ e 25.10$\pm 5.18\ kg/m^2$.

4.3 Análise Exploratória

Antes de se iniciar o estudo sobre o modelo que melhor se ajusta aos dados, é necessário analisar previamente os mesmos de forma a perceber como se comportam. Para a construção dos gráficos apresentados neste capítulo foram utilizadas as seguintes bibliotecas do R *lattice* e *ggplot2* (Sarkar, 2008, Wickham, 2009).
Começando pelos histogramas da PAM e do IMC observa-se o seguinte:

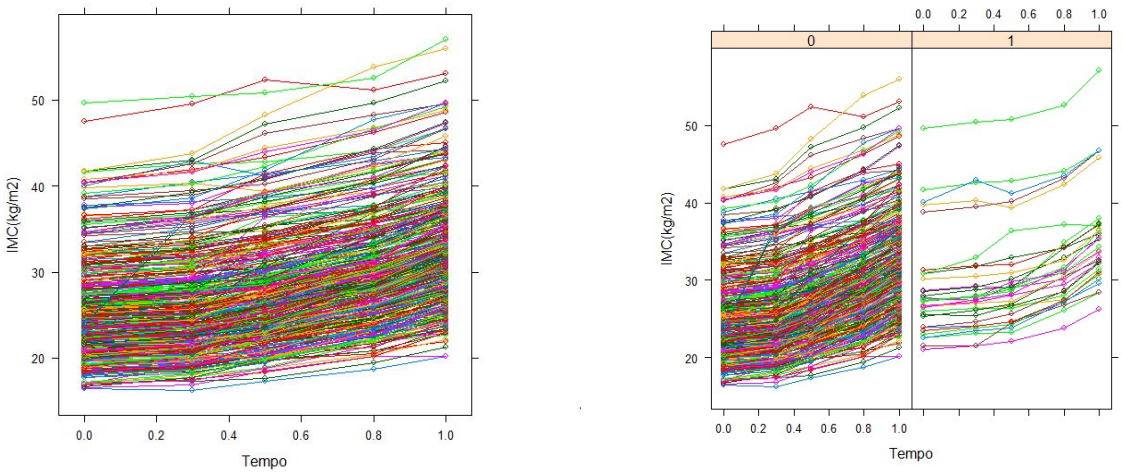


(a) (b)

Figura 4.1: (a) Histograma do IMC; (b) Histograma da PAM

O histograma do IMC mostra uma assimetria enviesada à direita que aliás conduzirá posteriormente a uma transformação da variável.

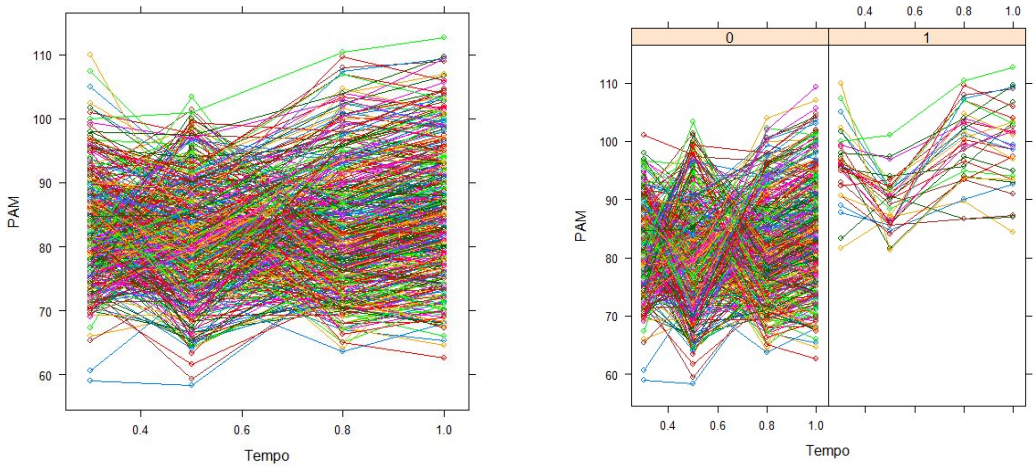
Através do comando *groupedData()*, disponível na library *nlme* do software *R* pode agrupar-se os dados por indivíduo e assim obter para cada paciente o seu perfil, isto é, o seu comportamento ao longo da gestação (Pinheiro et al., 2013).



(a) (b)

Figura 4.2: Perfil individual: (a) Evolução do IMC ao longo da gravidez; (b) Evolução do IMC ao longo da gravidez para gestantes normotensas (0) e hipertensas (1)

Na figura 4.2 é apresentada a evolução do IMC ao longo da gestação para as gestantes hipertensas (1) e normotensas (0) (perfil individual). O IMC é avaliado em cinco momentos distintos: no inicio da gravidez (0), das 12-14 semanas (0.3), das 18-22 semanas (0.5), das 29-33 semanas(0.8) e no momento do parto (1). Pela análise gráfica verifica-se que ao longo da gestação o IMC vai aumentando quer nas gestantes hipertensas (1), quer nas normotensas (0). É também possível observar que existem gestantes hipertensas, assim como gestantes normotensas, com elevados valores de IMC. O contrário também é observado, ou seja, em gestantes normotensas é visivel valores de IMC entre 15 kg/m^2 e os 18 kg/m^2. Os gráficos sugerem uma diferença de observações entre as gestantes hipertensas e as gestantes normotensas.

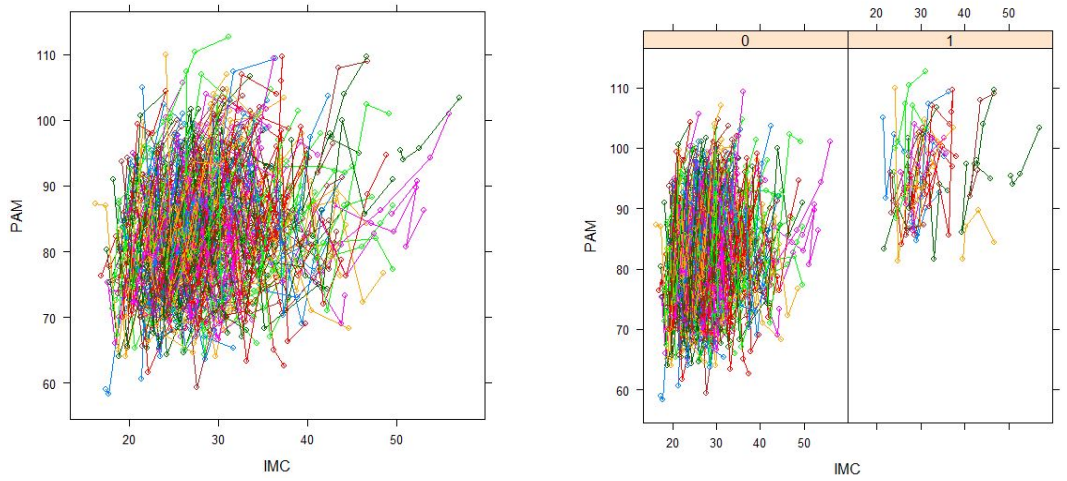


(a) (b)

Figura 4.3: Perfil individual: (a) Evolução da PAM ao longo da gestação; (b) Evolução da PAM ao longo da gravidez em gestantes normotensas (0) e hipertensas (1)

A PAM não é avaliada no inicio da gravidez, apenas a partir das 12-14 semanas (tempo 0.3) e até ao momento do parto. A figura 4.3 mostra a evolução da PAM ao longo da gravidez nas hipertensas (1) e nas normotensas (0). Verifica-se que nas gestantes hipertensas a PAM apresenta genericamente valores mais altos do que nas gestantes normotensas, o que seria de esperar uma vez que as primeiras apresentam hipertensão arterial crónica. É ainda de realçar o facto de a PAM diminuir entre as 12-14 semanas e as 18-22 semanas de gestação, o que também é de esperar quando não se revelam complicações na gravidez. No gráficos dos perfis individuais das gestantes normotensas detetam-se alguns perfis em que a PAM não diminui, no período indicado anteriormente, pelo contrário, aumenta. Isto pode efetivamente acontecer apesar de mais frequentemente diminuir. Após as 18-22 semanas (tempo 0.5), a PAM sobe até ao final da gravidez.
Entre as 12-14 semanas e as 18-22 semanas, as gestantes hipertensas parecem apresentar uma descida maior do valor da PAM do que as gestantes normotensas.

Em ambas as figuras, 4.2 e 4.3, o comportamento das gestantes hipertensas e normotensas, ao longo da gestação, parecem ser semelhantes.



(a) (b)

Figura 4.4: Perfis individuais (por pessoa): (a) Evolução da PAM em função do IMC; (b) Evolução da PAM em função do IMC em gestantes normotensas (0) e hipertensas (1)

Dado que o IMC é avaliado em cinco momentos distintos ao longo da gestação como já foi mencionado, e a PAM nos últimos quatro, para a obtenção destes gráficos teve-se em conta a interseção dos momentos de avaliação das gestantes. Como tal, no estudo desta relação os momentos de avaliação são os seguintes: 12-14 semanas, 18-22 semanas, 29-33 semanas e o momento do parto.
Quanto ao gráfico da direita são apresentados os perfis individuais das gestantes tendo em conta o estado hipertensivo da mesma. Como seria de esperar as gestantes hipertensas apresentam valores de PAM mais altos, acima dos $80mmHg$, do que as gestantes normotensas. Apesar deste facto, é de realçar que os valores para o IMC variam entre baixos e altos, o que significa que não há apenas gestantes com IMC alto.

Capítulo 5

Resultados

Neste capítulo serão apresentados todos os resultados obtidos da aplicação dos modelos mencionados no capítulo três: modelo linear de efeitos mistos e modelo de regressão com estimação pelo método dos mìnimos quadrados generalizados, na avaliação da evolução da pressão arterial média e do índice de massa corporal. A aplicação dos mesmos foi feita no software livre R, versão 3.0.3 (R Development, Core Team, 2012), para um nível de significância fixado em 0.05. Todas as bibliotecas utilizadas na aplicação dos modelos serão mencionadas quando necessário.

Relativamente às estruturas de variância, tendo em conta a tabela 3.1 apresentada no capítulo 3 e as estruturas de correlação também apresentandas na secção 3.4.1.2, apenas algumas foram utilizadas:

Estruturas de Variância	**Estruturas de Correlação**
VarIdent ($form = 1\|HT$)	corAR1($form = 1\|ID$)
VarPower ($form = tempo\|HT$)	corSymmm ($form = 1\|ID$)
VarExp ($form = tempo\|HT$)	corCompSymm ($form = 1\|ID$)

Tabela 5.1: Estruturas de variância e correlação utilizadas

O facto de se estudar o efeito da hipertensão na PAM e no IMC conduziu-nos à escolha da variável hipertensão como fator de agregação na estrutura de variância. Uma vez que as diferenças de dispersão por estado hipertensivo poderão ser relevantes no estudo (VarIdent ($form = 1|HT$)), assim como o facto de a variabilidade aumentar linearmente ou exponencialmente ao longo da gestação (VarPower ($form = tempo|HT$), VarExp ($form = tempo|HT$)), levou-nos a considerar no presente estudo as estruturas de variância representadas na tabela 5.1. Quanto às estruturas de correlação considerou-se o fator de agregação o índividuo (representado pelo seu ID). A primeira estrutura apresentada na tabela, corAR1($form = 1|ID$), mostra que a observação atual de uma gestante depende da observação avaliada no momento anterior dessa mesma gestante. Estudou-se o facto de a correlação entre os erros de observações de uma mesma gestante serem iguais (corCompSymm ($form = 1|ID$)), ou precisamente o contrário, em que as correlações diferem (corSymm ($form = 1|ID$)).

5.1 Pressão Arterial Média

Procedeu-se à aplicação dos modelos apresentados no capítulo três: o modelo linear de efeitos mistos, MLEM, e do modelo de regressão com estimação pelo método dos mínimos quadrados generalizados, MMQG. Para ambas as aplicações, foram desenvolvidos diversos modelos para o estudo da PAM, modelos esses que incluiram algumas variáveis explicativas: a idade, a menarca, o tempo, e nomeadamente a variável paridade, onde se esperaria, tendo em conta o contexto clínico, que fosse significativa no modelo. Contudo, isto não se verificou. Para se perceber melhor o porquê, procedeu-se à análise gráfica da mesma:

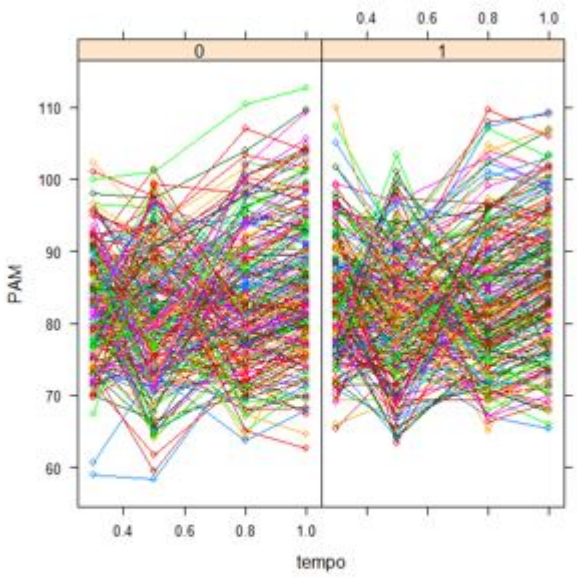


Figura 5.1: Perfil individual: Variação da PAM tendo em conta a variável paridade

A variável paridade está dividida em duas categorias: primíparas (categoria 0), que significa que a gestante não tem filhos e esta é a sua primeira gravidez com sucesso; e multíparas (categoria 1) que nos indica que a gestante tem um ou mais filhos. O gráfico acima sugere uma idêntica dispersão dos valores da PAM em ambas as categorias da paridade, poderá ter sido este o motivo pela qual a variável não se mostrou estatísticamente significativa.

5.1.1 Modelo LEM- Aplicação

Numa primeira abordagem, começou-se pela aplicação do modelo LEM, (Kirchkamp, 2014). Antes de se iniciar o ajustamento dos dados a um modelo linear de efeitos mistos deve proceder-se a uma avaliação gráfica da existência dos efeitos aleatórios. Através

da função *lmList()* da biblioteca *nlme* do R é possível ter uma primeira informação sobre a existência de efeitos aleatórios que poderão ser relevantes para o modelo.

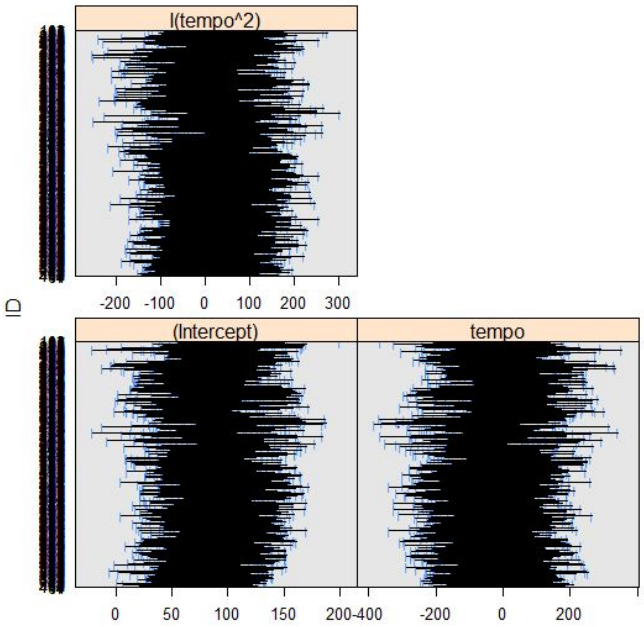


Figura 5.2: Estimativas dos intervalos de confiança a 95% para os parâmetros do modelo - PAM

A figura 5.2 apresenta as estimativas dos intervalos de confiança para os parâmetros do modelo ajustado a cada indivíduo. Da análise da figura será de esperar que a inclusão de qualquer efeito não seja o mais apropriado para o estudo da PAM, pois a variabilidade na constante e até mesmo no declive parece não ser adequada. Ainda assim, estudou-se o modelo com efeito aleatório na constante e a melhor estrutura de correlação encontrada, tendo sempre em consideração o menor valor do critério BIC, foi a auto-regressiva de ordem 1 (AR1). Através da função *lme()* da biblioteca *nlme()* do R é possível proceder à aplicação do modelo.

O modelo LEM para a evolução da PAM ao longo da gestação representa-se da seguinte forma:

- **Modelo 1**: $PAM_{it} = (\beta_0 + b_{0i}) + \beta_1 tempo_{it} + \beta_2 tempo_{it}^2 + \beta_3 tempo_{it}^3 + \beta_4 HT_i + \epsilon_{it}$

O modelo acima representa a evolução da PAM para o indivíduo i no tempo t, em que $i = 1, ..., 461$ e $t = 0.3, 0.5, 0.8, 1$. A variável HT é uma variável binária que representa o estado hipertensivo da gestante: 0 em gestantes normotensas e 1 para gestantes hipertensas. A classe de referência foi tida como sendo as gestantes normotensas. A variável aleatória b_{0i} representa o efeito aleatório na constante e ϵ_{it} representa os erros aleatórios. Assume-se que $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, \sigma^2 \Lambda_i)$ e b_i e ϵ_i independentes para os diferentes grupos, onde D é uma matriz diagonal e Λ_i é a matriz de variância-covariância dos erros. De salientar que a PAM só começa a ser avaliada apartir das 12-14 semanas

de gestação. As estimativas obtidas pelo modelo LEM acima mencionado encontram-se na tabela abaixo, tabela (5.2):

Efeitos Aleatórios				
	Intercept	Residual		
StdDev	1.587	7.867		
Efeitos Fixos				
Variável		**Coef.**	**Erro Padrão**	**valor-p**
constante		90.403	2.957	0.000
tempo		-43.210	16.024	0.007
$tempo^2$		56.105	26.303	0.033
$tempo^3$		-18.665	13.396	0.164
HT		13.925	0.980	0.000
Correlação: $\rho = 0.37$				
BIC: 12718				

Tabela 5.2: Estimativas obtidas pelo modelo LEM para a evolução da PAM (modelo 1)

Pela análise da tabela 5.2 observa-se o $tempo^3$ apresenta um valor-p de 0.164 (superior a 0.05). A variância da variável aleatória foi estimada em 2.509, e os erros aleatórios seguem uma distribuição $N(0, \sigma^2\Lambda_i)$, a matriz Λ_i tem uma estrutura de correlação AR1 com um parâmetro estimado de 0.37. Tendo em consideração o efeito aleatório na constante, o seu desvio padrão apresenta um valor de 1.587 e portanto a percentagem de variância explicada pelo mesmo é de 2%, um valor bastante baixo, que seria de esperar tendo em conta a primeira análise gráfica dos intervalos de confiança que revelava não ser necessário a inclusão dos efeitos aleatórios.

Quando estudados os modelos de efeitos aleatórios no tempo, $tempo^2$ e no $tempo^3$, os valores percentuais de variância explicada pela inserção dos efeitos aleatórios foram mais uma vez baixos, e desta forma entende-se que este tipo de modelo não é o mais adequado para o estudo da PAM.

5.1.2 Modelo MMQG - Aplicação

Face à insatisfação dos resultados obtidos no modelo LEM, procedeu-se ao estudo da PAM nas gestantes hipertensas e normotensas através do modelo de regressão com estimação pelo método dos mínimos quadrados generalizados.
Para modelar a heterocedasticidade e dependência dos erros aleatórios foram consideradas várias estruturas de correlação e variância como visto anteriormente (tabela 5.1). Tendo em conta o menor valor de BIC apresentado, quando comparadas as estruturas de correlação, a escolha recaíu sobre a estutura de correlação auto-regressiva de ordem 1, AR1. Relativamente à estrutura de variância foram utilizadas algumas (tabela 5.1), sendo que a escolhida foi a varIdent (form $=\sim 1 \mid$ HT) tendo em conta o menor valor de BIC apresentado e o valor-p apresentado no teste de razão de verosimilhanças que nos indicou que esta estrutura seria a melhor a ter em conta. Esta escolha também é sustentada pelo gráfico abaixo que nos mostra, claramente, a diferença de dispersão da PAM tendo em conta o estado hipertensivo da gestante.

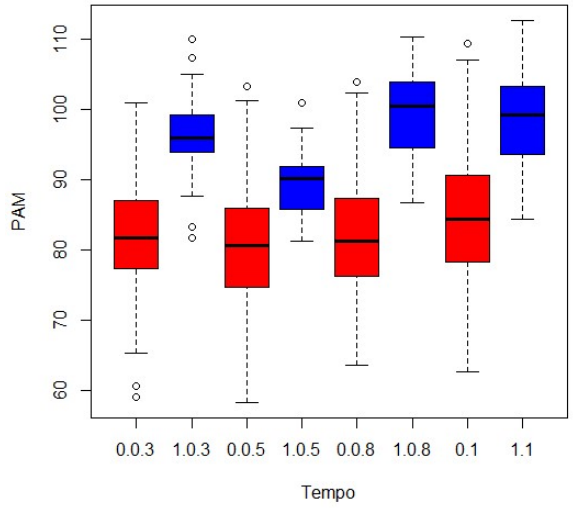


Figura 5.3: Dispersão da PAM em gestantes normotensas (vermelho) e em gestantes hipertensas (azul)

O modelo escolhido é apresentado da seguinte forma:

- **Modelo 2**: $PAM_{it} = \beta_0 + \beta_1 tempo_{it} + \beta_2 tempo_{it}^2 + \beta_3 tempo_{it}^3 + \beta_4 HT_i + \epsilon_{it}$

A aplicação deste modelo no R foi executada através da função *gls()*, da biblioteca *nlme()*. O modelo representa a evolução da PAM ao longo da gestação para o indivíduo i no tempo t, em que $i = 1, ..., 461$ e $t = 0.3, 0.5, 0.8, 1$, ϵ_{it} representa os erros aleatórios, $\beta_0, \beta_1, \beta_2, \beta_3 e \beta_4$ são os parâmetros de regressão. Como no modelo LEM, a variável HT é uma variável binária que representa o estado hipertensivo da gestante: 0 em gestantes normotensas e 1 para gestantes hipertensas. A classe de referência foi tida como sendo as gestantes normotensas.

A tabela 5.3 apresenta as estimativas obtidas pelo modelo MMQG para os dados em análise.

Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	91.992	2.936	0.000
tempo	-52.373	15.903	0.001
$tempo^2$	71.362	26.097	0.006
$tempo^3$	-26.332	13.290	0.048
HT	13.934	16.762	0.000
Variância: $varIdent(form = \sim 1 \mid HT)$			
Parâmetros estimados			
	0	1	
	1.000	0.838	
Correlação: $\rho = 0.39$			
BIC: 112726			

Tabela 5.3: Estimativas obtidas pelo modelo MMQG para a evolução da PAM (modelo 2)

O modelo selecionado é um modelo com progressão temporal cúbica. Pela análise da tabela anterior verifica-se que todas as variáveis do modelo são estatísticamente significativas e o valor do BIC do modelo é de 12726, o mais baixo entre todos os modelos estudados. Quanto à matriz dos erros esta apresenta um valor estimado para a estrutura de variância de 0.84, e um valor estimado para o parâmetro de correlação de 0.39.

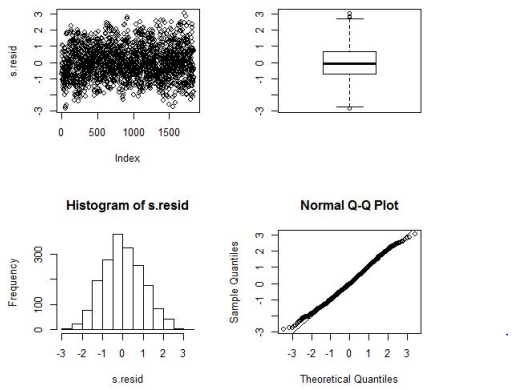


Figura 5.4: Gráficos de diagnóstico do modelo MMQG - PAM

Na figura 5.4 são apresentados gráficos de diagnóstico do modelo. Os mesmos revelam a existência de três outliers, sendo que estes não interferem na normalidade dos resíduos do modelo, como tal não foram retirados. Os gráficos sugerem bastante homocedasticidade nos resíduos acompanhada de uma simetria revelada pelo histograma e pelo gráfico de probabilidades normal apresentado.

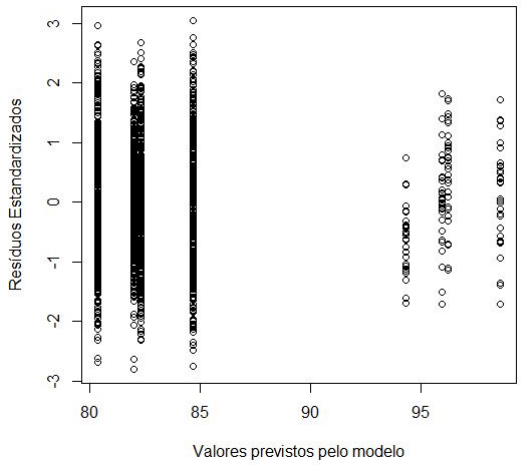
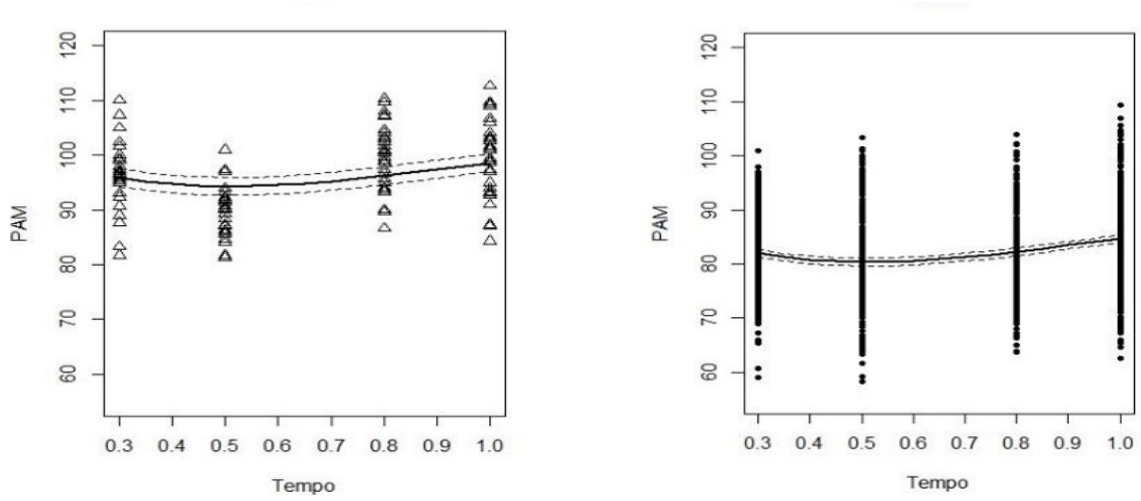


Figura 5.5: Gráfico dos resíduos estandardizados *versus* valores previstos pelo modelo MMQG (modelo 2) - PAM

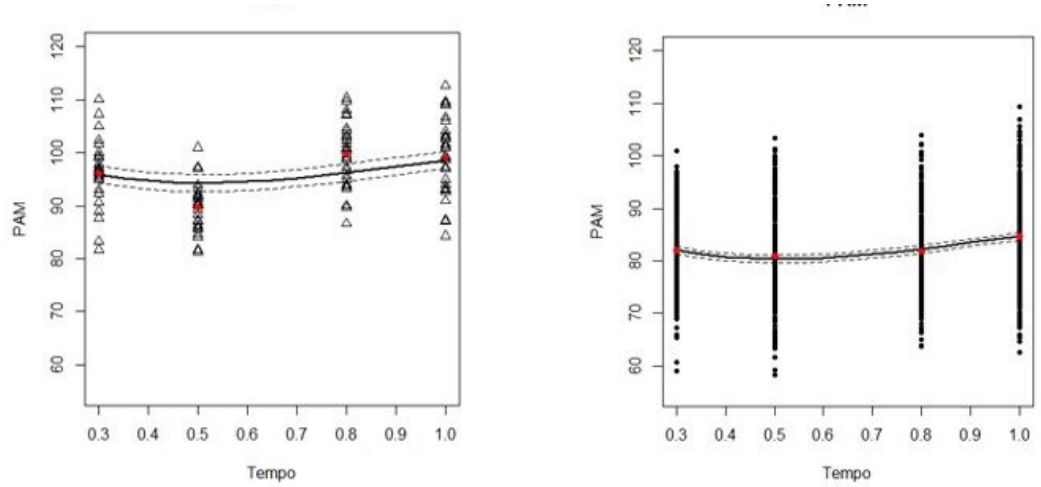
Na figura 5.5 observa-se que os resíduos estandardizados em gestantes normotensas são mais elevados por comparação com os resíduos em gestantes hipertensas. A inserção de interação no modelo poderia melhorar as previsões para gestantes hipertensas, contudo devido ao reduzido tamanho amostral este procedimento não foi considerado razoável. Assim, são apresentados na figura 5.6 as previsões do modelo MMQG para o estudo da evolução da PAM com respetivas bandas de confiança a 95%. Duas curvas cúbicas são esperadas, uma para cada estado hipertensivo, sendo que nas gestantes hipertensas esperam-se valores significativamente superiores em relação ás gestantes normotensas.



(a) (b)

Figura 5.6: Curvas dos valores previstos do modelo MMQG para a evolução da PAM ao longo da gestação com respetivo intervalo de confiança: (a) em gestantes hipertensas; (b) em gestantes normotensas

Na figura 5.6 são apresentadas as curvas dos valores previstos, obtidas pelo modelo MMQG para cada estado hipertensivo. Prevê-se um valor médio de pressão arterial média de $95.9mmHg$ em gestantes hipertensas, entre as 12-14 semanas (tempo igual a 0.3), que vai diminuindo até às 18-22 semanas (tempo igual a 0.5) atingindo um valor de $94.3mmHg$, que de seguida aumenta até ao momento do parto (tempo igual a 1) onde atinge uma valor máximo de $98.6mmHg$. Em gestantes normotensas, o comportamento da pressão arterial média é idêntico, apresentando valores de PAM mais baixos. Entre as 12-14 semanas, o valor médio da PAM nestas gestantes é de $82.0mmHg$, atingindo um valor máximo de $84.7mmHg$ também no momento do parto.



(a) (b)

Figura 5.7: Curvas do modelo MMQG para a evolução da PAM ao longo da gestação com pontos médios em cada tempo: (a) em gestantes hipertensas; (b) em gestantes normotensas

O gráfico da figura 5.7 sugere alguma imprecisão na predição dos valores da PAM ao longo da gestação, no caso das gestantes hipertensas. Entre as 18-22 semanas (tempo igual a 0.5) e entre as 29-33 semanas (tempo igual a 0.8), observa-se que a curva do modelo se afasta do ponto médio encontrado naqueles tempos. Quanto às extremidades, isto é, entre as 12-14 semanas (tempo igual a 0.3) e no momento do parto (tempo igual a 1) a curva do modelo situa-se junto do ponto médio. A redução da predição nos tempos intermédios de avaliação da PAM nas gestantes hipertensas parece estar a sugerir o ajustamento de um polinómio cúbico (em interação com o estado hipertensivo), mas como referido atrás, optou-se por não considerar pela baixa representatividade amostral do grupo de mulheres hipertensas.

5.2 Índice de Massa Corporal

O IMC é dado pelo seguinte cálculo: $IMC = \frac{Peso}{Altura^2}$ em que o peso se apresenta em Kg e a altura em $metros$, como visto no capítulo anterior. Realizou-se então o estudo da evolução do IMC ao longo da gestação. Esta variável tem em conta mais um tempo de avaliação, assim na aplicação que se segue o tempo varia de 0 a 1: 0, 0.3, 0.5, 0.8, 1, ao contrário do que aconteceu com a PAM. A variável IMC foi então submetida a uma *transformação logarítmica*:

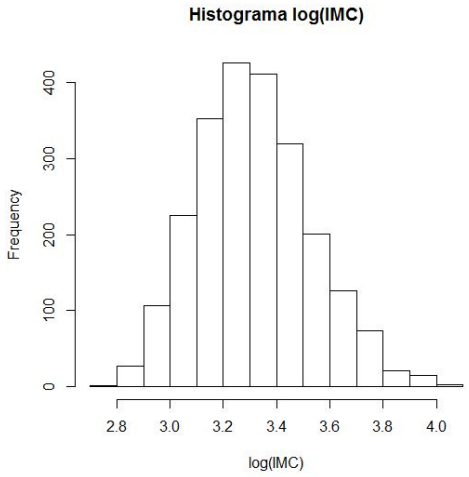


Figura 5.8: Histograma do logaritmo do IMC

Foi necessária esta transformação para que posteriormente os erros do modelo pudessem seguir uma distribuição normal. Para o estudo utilizou-se o logaritmo da variável IMC na aplicação dos modelos e respetiva análise.

5.2.1 Modelo LEM - Aplicação

Para ter uma primeira ideia sobre quais os efeitos aleatórios a introduzir no modelo LEM aplicou-se a função *lmList* da biblioteca *nlme* do software R. Esta função traduz graficamente as estimativas dos intervalos de confiança para os parâmetros do modelo ajustado para cada indivíduo e ignorando a estrutura longitudinal dos dados (Cabral and Gonçalves, 2011).

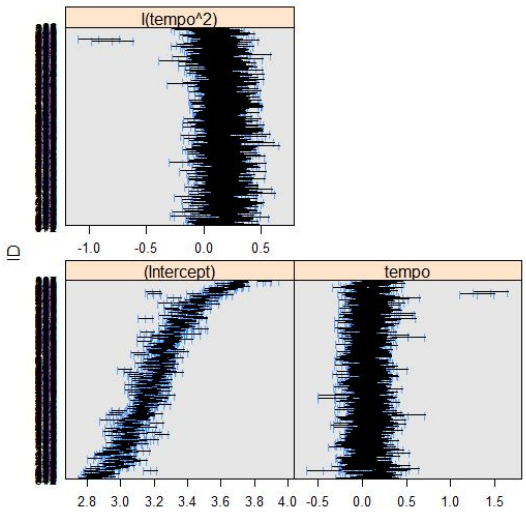


Figura 5.9: Estimativas dos intervalos de confiança a 95% para os parâmetros do modelo - IMC

Pela análise da figura 5.9 podemos esperar obter efeitos aleatórios na constante, pois é onde é visível uma maior variabilidade nos intervalos, o que leva à suposição de que a variabilidade inter-individual existe e que a incorporação deste mesmo efeito deverá ser tida em conta.
Após a escolha de quais os efeitos aleatórios a inserir no modelo linear misto, mais uma vez, através da função *lme* da biblioteca *nlme* do R foram ajustados vários modelos aos dados com diversas estruturas de correlação e variância, tendo em conta o menor valor do critério BIC e do teste de hipóteses da razão de verosimilhanças. A estrutura de correlação encontrada que melhor se adaptou aos dados foi a correlação auto-regressiva de ordem 1. Quanto às estruturas de variância nenhuma das selecionadas para este trabalho (tabela 5.1) foi inserida no modelo uma vez que, os intervalos obtidos para a variância do modelo não mostraram ser significativos. O gráfico abaixo apoia esta decisão.

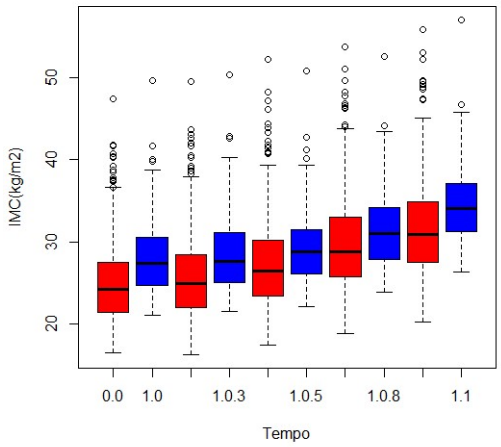


Figura 5.10: Dispersão do IMC em gestantes normotensas (vermelho) e em gestantes hipertensas (azul)

Verifica-se no gráfico da figura 5.10 que a dispersão do IMC em gestantes normotensas não difere muito em comparação com as gestantes hipertensas, portanto não se verifica a violação de homocedasticidade o que mostra que a estrutura de variância não se tornou tão importante como verificado no estudo da PAM.

O modelo LEM eleito para o estudo da evolução do IMC ao longo da gestação foi o seguinte:

- **Modelo 1**: $log(IMC_{it}) = (\beta_0 + b_{0i}) + \beta_1 tempo_{it} + \beta_2 HT_i + \beta_3 tempo_{it}^2 + \epsilon_{it}$

O modelo acima representa a evolução do IMC para o indivíduo i no tempo t, em que $i = 1, ..., 461$ e $t = 0, 0.3, 0.5, 0.8, 1$; a variável HT é uma variável binária que representa o estado hipertensivo da gestante: 0 em gestantes normotensas e 1 para gestantes hipertensas, a classe de referência são as gestantes normotensas; b_{0i} representa os

efeitos aleatórios e ϵ_{it} são os erros aleatórios. Assume-se que $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, \sigma^2\Lambda_i)$ e b_i e ϵ_i independentes para os diferentes grupos, onde D é uma matriz diagonal e Λ_i é a matriz de variância-covariância dos erros.

A seguinte tabela (tabela 5.4) descreve as estimativas obtidas pelo modelo LEM na evolução do IMC, modelo 1:

Efeitos Aleatórios				
	Intercept	Residual		
StdDev	0.158	0.103		
Efeitos Fixos				
Variável		**Coef.**	**Erro Padrão**	**valor-p**
constante		3.205	0.009	0.000
tempo		0.067	0.006	0.000
HT		0.124	0.034	3e-04
$tempo^2$		0.167	0.006	0.000
Correlação: $\rho = 0.94$				
BIC: -7333				

Tabela 5.4: Estimativas obtidas pelo modelo LEM para a evolução do IMC (modelo 1)

Da análise da tabela acima ressalta o facto de as covariáveis presentes no modelo serem estatísticamente significativas, com valores-p inferiores a 0.05 e o valor do critério BIC ser -7333. A variância da variável aleatória foi estimada em 0.025, e os erros aleatórios seguem uma distribuição $N(0, \sigma^2\Lambda_i)$, a matriz Λ_i tem uma estrutura de correlação AR1 com um parâmetro estimado de 0.94. Quando analisado o efeito aleatório, o mesmo apresenta um desvio padrão de 0.158 e portanto a percentagem de variância explicada pelo efeito aleatório é de aproximadamente 70% ($0.158^2/(0.158^2 + 0.103^2)$). Isto sugere que o presente efeito aleatório deverá contar do modelo, pois a variabilidade inter-individual é elevada.
Após a análise das estimativas obtidas pelo modelo, avaliou-se o comportamento dos resíduos do mesmo, através dos seguintes gráficos de diagnóstico (figura 5.11):

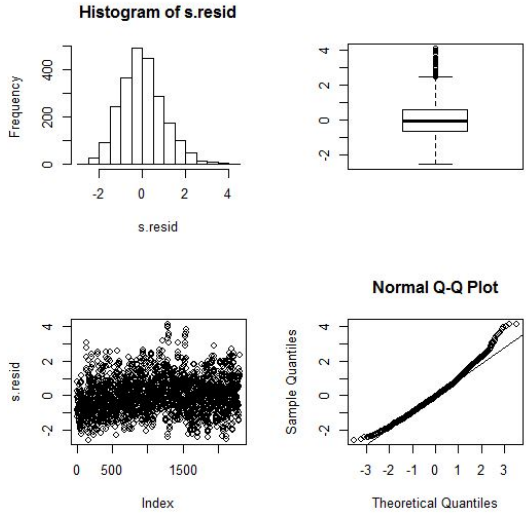


Figura 5.11: Gráficos de diagnóstico do modelo 1 - IMC

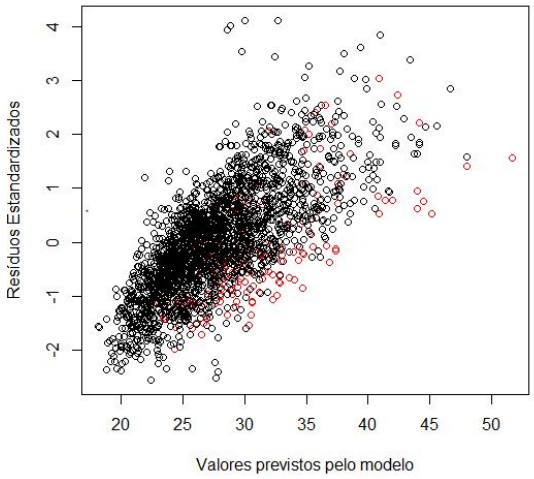
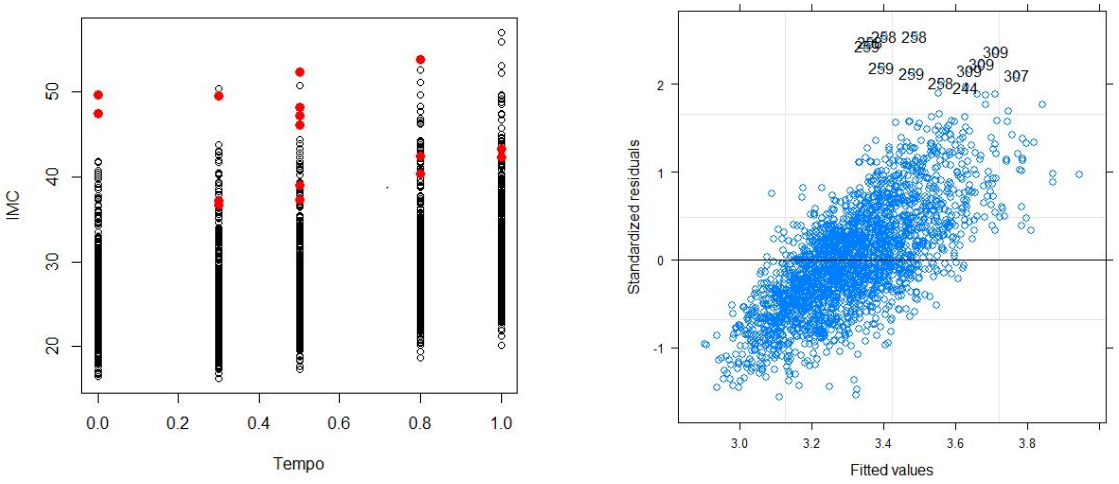


Figura 5.12: Gráfico de diagnóstico dos resíduos estandardizados *versus* valores previstos pelo modelo 1 por estado hipertensivo - IMC

A análise gráfica dos resíduos aponta para a existência de outliers no modelo. Estes pontos influenciam a assimetria visivel no histograma e o desvio na cauda do gráfico dos quantis. Na figura seguinte, 5.12, verifica-se também a existência de observações com valores altos. As mesmas dizem respeito a gestantes normotensas. As observações a cor vermelha referem-se a gestantes hipertensas que como se oberva apresenta resíduos mais baixos. Assim os gráficos das figuras 5.11 e 5.12 sugerem não existir normalidade dos resíduos e como tal, decidiu-se identificar os outliers graficamente (figura 5.13):



(a) (b)

Figura 5.13: Identificação dos outliers do modelo 1 - IMC

Os pontos assinalados no gráfico (a) e identificados pelo seu ID no gráfico (b) dizem respeito a sete observações da amostra. De entre as sete, apenas uma gestante apresenta

hipertensão arterial crónica com um valor de IMC no ínicio da gravidez de $49kg/m^2$, sendo as restantes seis observações gestantes normotensas também com valores de IMC altos, acima dos $35kg/m^2$. Esta descrição dos *outliers* mostra que os mesmos, por comparação com o resto da amostra, apresentam valores de IMC muito díspares, mais precisamente valores bastante altos.

Estas observações foram, de seguida, retiradas da base de dados e voltou a ajustar-se o modelo 1, obtendo-se o modelo 2. A tabela 5.5 apresentada mostra as estimativas desse modelo.

Efeitos Aleatórios				
	Intercept	Residual		
StdDev	0.155	0.104		
Efeitos Fixos				
Variável		**Coef.**	**Erro Padrão**	**valor-p**
constante		3.205	0.009	0.000
tempo		0.060	0.006	0.000
HT		0.125	0.034	3e-04
$tempo^2$		0.172	0.005	0.000
Correlação: $\rho = 0.95$				
BIC: -7564				

Tabela 5.5: Estimativas obtidas pelo modelo LEM para a evolução do IMC sem outliers (modelo 2)

Muito idêntico às estimativas obtidas no modelo 1, este novo modelo sem os outliers, mostra que as variáveis explicativas mantêm a sua significância, e o valor dos coeficientes não sofre uma grande alteração.
O valor do critério BIC diminuiu, passando de -7333 para -7564, sugerindo que este modelo poderá ser melhor que o anterior, pois quanto menor o valor deste critério melhor é o ajustamento do modelo aos dados. A estimação do parâmetro de correlação pouco aumentou. Quanto ao efeito aleatório, mantém o valor alto de variância explicada pelo mesmo (desvio padrão igual a 0.155).
Como última análise deste modelo, falta apresentar os seus resíduos estandardizados.

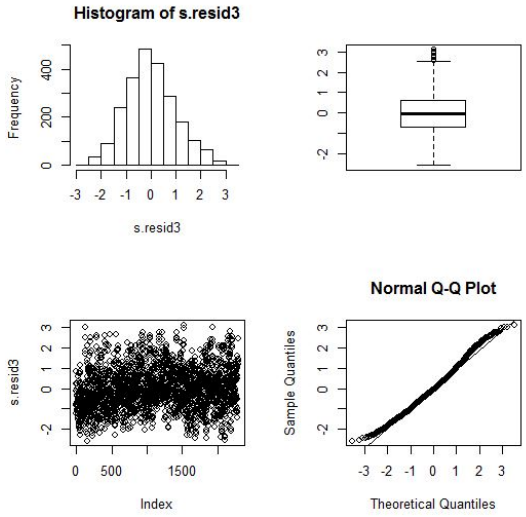


Figura 5.14: Gráficos de diagnóstico do modelo LEM sem outliers(modelo 2) - IMC

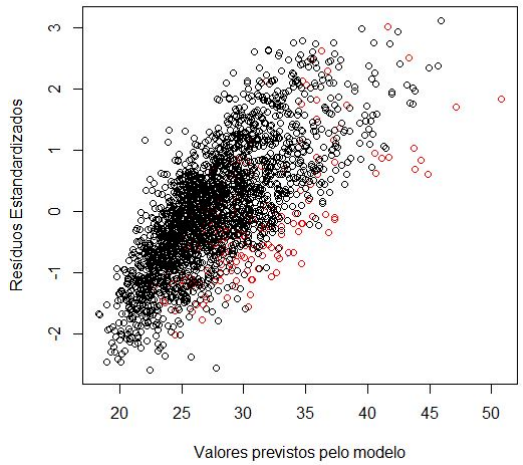
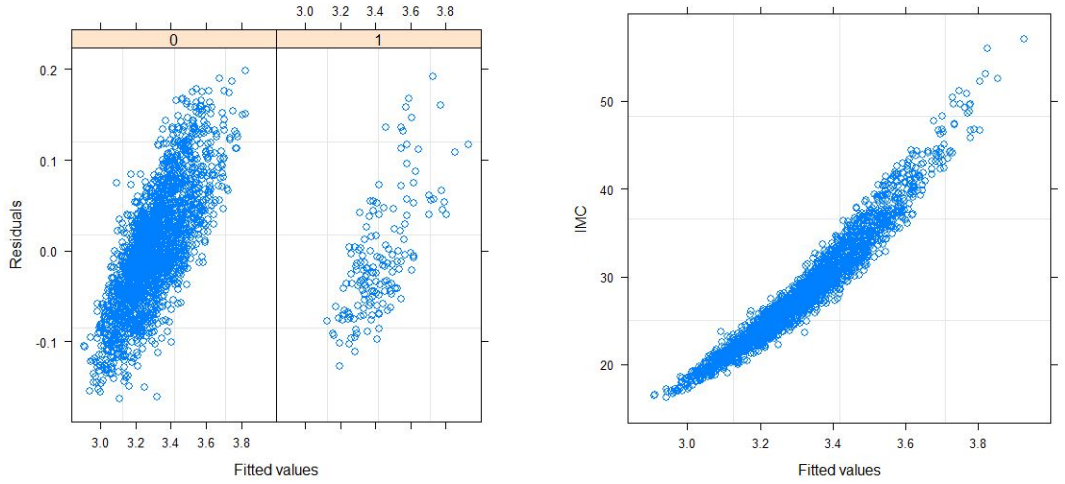


Figura 5.15: Gráfico de diagnóstico dos resíduos estandardizados *versus* valores previstos pelo modelo 2 - IMC

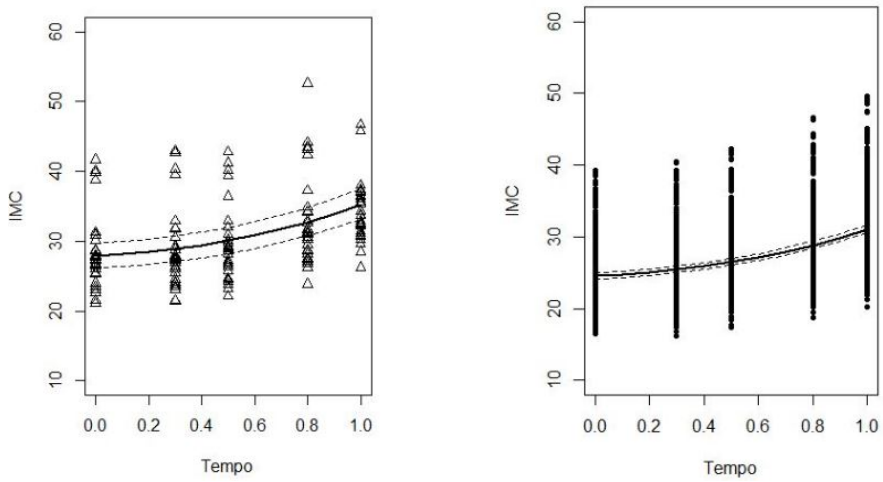
As figuras acima sugerem que os resíduos sofrem melhorias em termos de normalidade, por comparação com os do modelo 1, bem como em termos de simetria, revelada pelo histograma e pelo *boxplot*. O gráfico de quantis da figura 5.14 apoia novamente a ideia de normalidade encontrada por este novo modelo. A homocedasticidade dos resíduos é observável no gráfico de pontos da figura 5.14. Pela análise do gráfico da figura 5.15 referente aos resíduos estandardizados *versus* valores previstos pelo modelo, verifica-se que os valores dos resíduos diminuíram com a eliminação dos outliers. O gráfico sugere que as gestantes hipertensas continuam com valores baixos de resíduos.



(a) (b)

Figura 5.16: (a) Gráfico dos resíduos *versus* valores ajustados (modelo 2) em gestantes normotensas (0) e em gestantes hipertensas (1) - IMC; (b)Gráfico dos valores obervados *versus* valores estimados(modelo 2) - IMC

No gráfico da figura 5.16, a variabilidade parece bastante homogénea entre cada estado hipertensivo, mostrando que a inserção de uma estrutura de variância no modelo parece ser desnecessária, pois a homocedasticidade não é violada. Quanto ao gráfico dos valores observados *versus* valores estimados mostra que este modelo parece adequado para modelar o comportamento do IMC ao longo da gestação. De seguida são apresentadas as curvas dos valores previstos do modelo 2 de acordo com o estado hipertensivo da gestante.



(a) (b)

Figura 5.17: Curvas dos valores previstos do modelo LEM (modelo 2) para a evolução da IMC ao longo da gestação com respetivo intervalo de confiança: (a) em gestantes hipertensas; (b) em gestantes normotensas

As curvas da figura 5.17 mostram a evolução do IMC ao longo da gravidez de acordo com o estado hipertensivo da gestante. É de fácil inferência que as mulheres com hipertensão arterial crónica apresentam um valor médio de IMC superior ao valor médio de IMC em gestantes normotensas. Este modelo prevê uma progressão quadrática ao longo da gravidez de acordo com o estado hipertensivo, sendo portanto duas curvas significativamente diferentes uma da outra. Para algum momento da gravidez o modelo prevê um IMC médio mais alto para as gestantes hipertensas do que para as gestantes normotensas. No ínicio da gravidez o modelo prediz um valor médio de IMC de $24.8kg/m^2$ para gestantes normotensas e no final de $31.3kg/m^2$. Em gestantes hipertensas o modelo prevê inicialmente um valor médio de IMC de $28.1kg/m^2$ e no final de $35.4kg/m^2$.

5.3 Relação entre a PAM e o IMC

Após o estudo individual da evolução da PAM e do IMC ao longo do tempo, ajustado ao estado hipertensivo da gestante, realizou-se o estudo da relação entre ambas as variáveis. Pelos motivos já referidos no início deste capítulo, os modelos que se seguem, usaram apenas as estruturas de correlação e variância mencionadas na tabela 5.1.

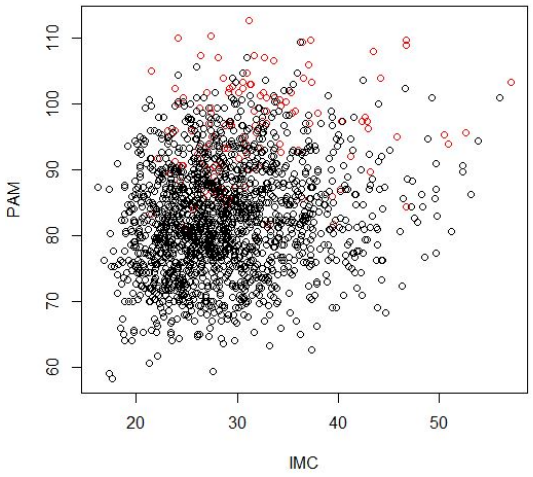


Figura 5.18: Dispersão das observações na relação entre a PAM e o IMC em gestantes normotensas (cor preta) e em gestantes hipertensas (cor vermelha)

Pela análise da figura acima, 5.18, observa-se que as gestantes hipertensas apresentam valores mais altos de PAM do que as gestantes normotensas. Assim, de acordo com o que foi feito anteriormente, iniciou-se o estudo pela aplicação de um modelo LEM.

5.3.1 Modelo LEM - Aplicação

Começou-se então por considerar o modelo LEM, em que a variável resposta é a PAM e as variáveis explicativas são o IMC e o estado hipertensivo das gestantes. Como visto

nas secções anteriores, a aplicação de um modelo LEM requer uma primeira análise sobre quais os efeitos aleatórios a introduzir no modelo. Para isso, utilizou-se, mais uma vez, a função *lmList* da biblioteca *nlme*, obtendo o seguinte gráfico:

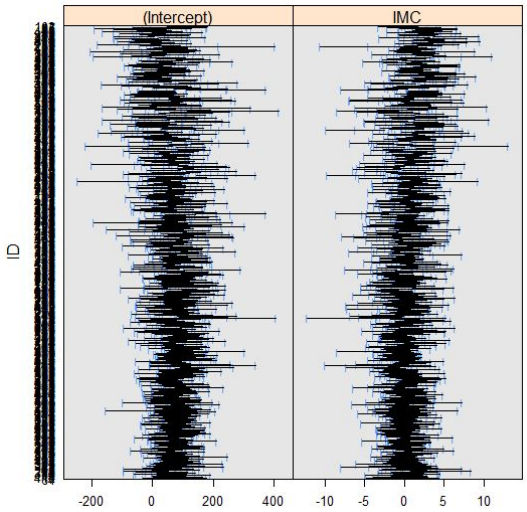


Figura 5.19: Estimativas dos intervalos de confiança a 95% para os parâmetros do modelo - Relação da PAM com IMC

O gráfico acima apresenta as estimativas dos intervalos de confiança para os parâmetros do modelo ajustado a cada indivíduo. Analisando a figura, será de esperar que a inclusão de qualquer efeito não seja o mais adequado para o estudo da relação entre a PAM e o IMC, sendo que não parece existir grande variabilidade na constante nem no declive. Contudo, estudou-se o modelo LEM com efeito aleatório na constante.
Assim, utilizando a função *lme()* da biblioteca *nlme()* do R o modelo encontrado foi o seguinte:

- **Modelo 1**: $PAM_{it} = (\beta_0 + b_{0i}) + \beta_1 IMC_{it} + \beta_2 HT_i + \epsilon_{it}$

O modelo acima representa a relação entre a PAM e o IMC para o indivíduo i no tempo t, em que $i = 1, ..., 461$ e $t = 0.3, 0.5, 0.8, 1$; a variável HT é uma variável binária que representa o estado hipertensivo da gestante: 0 em gestantes normotensas e 1 para gestantes hipertensas, sendo que a classe de referência são as gestantes normotensas; b_{0i} representa o efeitos aleatório e ϵ_{it} são os erros aleatórios. Assume-se que $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, \sigma^2\Lambda_i)$ e b_i e ϵ_i independentes para os diferentes grupos, onde D é uma matriz diagonal e Λ_i é a matriz de variância-covariância dos erros. A melhor estrutura de correlação encontrada, tendo em consideração o menor valor do critério BIC, foi a auto-regressiva de ordem 1 (AR1). Quanto à estrutura de variância, optou-se pela não inclusão da mesma no modelo, pois a estimativa do intervalo de confiança para os parâmetros da variância mostrou não haver diferenças significativas entre os valores. A tabela abaixo apresenta as estimativas obtidas para os coeficientes do modelo:

Efeitos Aleatórios				
	Intercept	Residual		
StdDev	0.059	7.850		
Efeitos Fixos				
Variável		**Coef.**	**Erro Padrão**	**valor-p**
constante		75.098	1.131	< 0.001
IMC		0.259	0.038	< 0.001
HT		13.074	0.993	< 0.001
Correlação: $\rho = 0.35$				
BIC: 12746				

Tabela 5.6: Estimativas obtidas para os parâmetros do modelo LEM para a relação entre a PAM e o IMC na gestação (modelo 1)

Pela análise da tabela, verifica-se que todas as variáveis explicativas são estatísticamente significativas. Neste modelo não foi incorporada a estrutura de variância como referido anteriormente. A variância da variável aleatória foi estimada em 0.003, e os erros aleatórios seguem uma distribuição $N(0, \sigma^2\Lambda_i)$, a matriz Λ_i tem uma estrutura de correlação AR1 com um parâmetro estimado de 0.35. Tendo em consideração o efeito aleatório na constante, o seu desvio padrão toma o valor de 0.059, e portanto a percentagem de variância explicada pelo mesmo no modelo é inferior a 1% ($0.059^2/(0.059^2 + 7.850^2)$). Sendo um valor bastante baixo, sugere que a variabilidade do termo constante entre indivíduos não existe, o que vai de encontro à primeira análise gráfica relativa às estimativas para os parâmetros do modelo para cada indivíduo. O valor estimado para o parâmetro de correlação foi de 0.35. Numa tentativa de melhoramento do modelo, aplicou-se a interação entre o IMC e a hipertensão, mas a mesma não se revelou estatísticamente significativa (valor-p=0.994).

Quando estudado o modelo LEM com efeito aleatório no IMC e em ambos, constante e IMC, a percentagem de variância no modelo explicada por estes mesmos efeitos foi mais uma vez baixa (inferior a 1%), sugerindo que o modelo LEM não é o mais adequado para o estudo da relação entre a PAM e o IMC.

5.3.2 Modelo MMQG - Aplicação

Dado que a inserção do efeito aleatório no modelo não se revelou satisfatória na modelação da relação entre a PAM e o IMC, aplicou-se uma outra metodologia. Seguidamente é apresentada a aplicação do modelo MMQG para o estudo referido. A aplicação deste modelo no R foi executada através da função *gls()*, da biblioteca *nlme()*. Foram consideradas diversas estruturas de correlação e variância para incluir no modelo (tabela 5.1). A estrutura de correlação que melhor se ajustou aos dados foi a correlação autoregressiva de ordem 1, o que significa que a correlação de uma dada observação depende linearmente da observação passada. Quanto à estrutura de variância, nenhuma foi incluída, os intervalos de confiança para a estrutura de variância não revelaram diferenças significativas entre os valores estimados.

Assim, o modelo escolhido foi o seguinte:

- **Modelo 2**: $PAM_{it} = \beta_0 + \beta_1 IMC_{it} + \beta_2 HT_i + \epsilon_{it}$

O modelo representa a relação entre a PAM e o IMC para o indivíduo i no tempo t, em que $i = 1, ..., 461$ e $t = 0.3, 0.5, 0.8, 1$, a variável ϵ_{it} representa os erros aleatórios e $\beta_0, \beta_1 e \beta_2$ são os parâmetros (escalares) de regressão. Tal como no modelo LEM, a variável HT é uma variável binária que representa o estado hipertensivo da gestante: 0 em gestantes normotensas e 1 em gestantes hipertensas. A classe de referência foi tida como sendo as gestantes normotensas.
A tabela 5.7 apresenta as estimativas obtidas pelo modelo MMQG:

Variável	**Coef.**	**Erro Padrão**	**valor-p**
constante	75.356	1.109	< 0.001
IMC	0.251	0.038	< 0.001
HT	13.085	0.969	0.001
Correlação: $\rho = 0.38$			
BIC: 12739			

Tabela 5.7: Estimativas obtidas para os parâmetros do modelo MMQG para a relação entre a PAM e o IMC na gestação (modelo 2

Pela análise da tabela acima verifica-se que todas as variáveis explicativas são estatisticamente significativas. No que diz respeito à matriz dos erros, o parâmetro de correlação foi estimado em 0.38. Pode inferir-se, a partir da tabela acima, que a valores altos de PAM estão associados valores altos de IMC. Verifica-se ainda que, em gestantes hipertensas a PAM toma valores mais altos (superiores em $13.09 mmHg$) do que nas gestantes normotensas, o que é de esperar pois as gestantes hipertensas sofrem de hipertensão arterial crónica. Por cada aumento de uma unidade no IMC, espera-se que a PAM aumente em $0.25 mmHg$ o seu valor.
De seguida, é apresentada a análise gráfica dos resíduos do modelo tendo em conta os resíduos estandardizados.

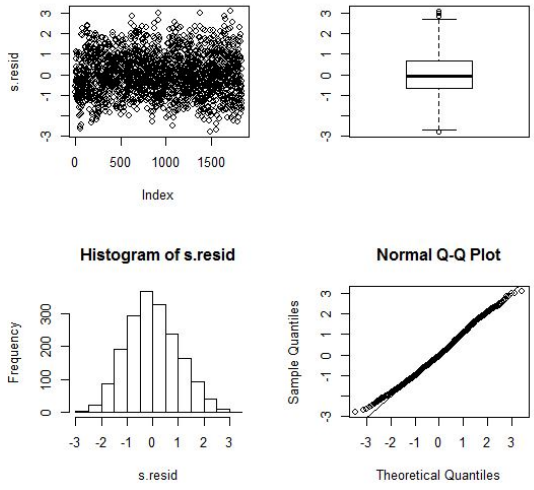


Figura 5.20: Alguns gráficos de diagnóstico do modelo MMQG (modelo2)

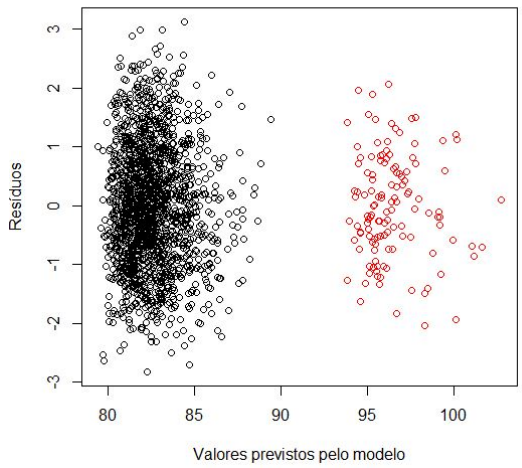
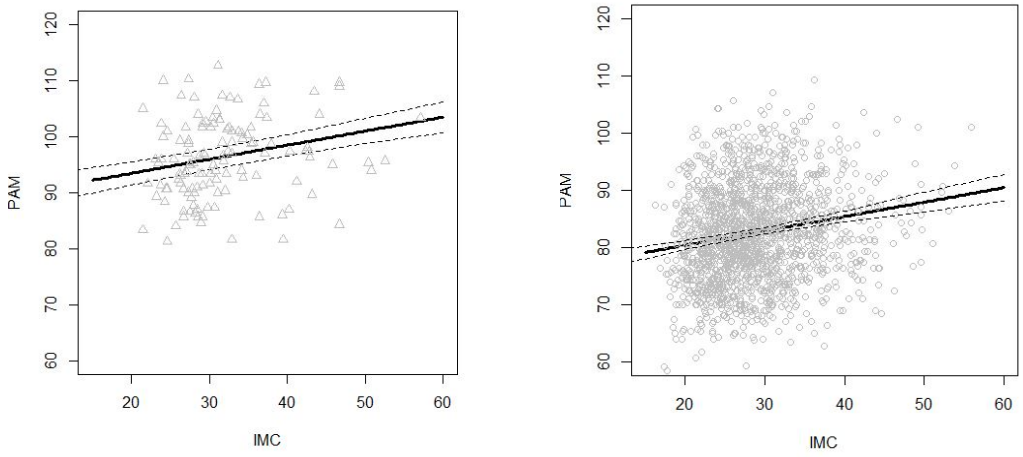


Figura 5.21: Gráficos dos resíduos estandardizados *versus* valores previstos para o modelo MMQG (modelo2)

Os gráficos de diagnóstico apresentados na figura 5.20 e 5.21, sugerem normalidade dos resíduos. Apesar da existência de outliers visíveis através do boxplot, estes aparentemente não interferem em desvios da normalidade dos resíduos (afirmação sustentada pela análise dos restantes gráficos), como tal não foram retirados. A figura 5.21 sugere algumas diferenças entre a variabilidade das mulheres hipertensas e a das normotensas. Contudo, a inclusão duma estrutura no modelo não se revelou estatísticamente significativa.
Após a análise completa ao modelo, são apresentados os gráficos dos valores previstos pelo modelo MMQG no estudo da relação entre a PAM e o IMC.



(a) (b)

Figura 5.22: Reta dos valores previstos pelo modelo MMQG na relação da PAM com o IMC na gestação com respetivo intervalo de confiança (modelo2): (a) em gestantes hipertensas; (b) em gestantes normotensas

Pelo gráfico da figura acima,figura 5.22, observa-se que estamos perante um modelo de regressão com progressão linear. O modelo prevê que a PAM média em gestantes hipertensas sofre um acréscimo de $13.09mmHg$ em relação à PAM média em gestantes normotensas. Apesar disto, quando relacionada com o IMC, o modelo prevê que um aumento de uma unidade no mesmo provoca um aumento de $0.25mmHg$ na PAM média para ambas as gestantes. Por estes motivos, tendo em conta o estado hipertensivo das gestantes, estamos perante duas retas paralelas, em que uma assume valores mais altos do que a outra.

Capítulo 6

Conclusões

Este trabalho teve como objetivo o estudo de modelos de regressão para dados longitudinais e a sua aplicação na área da saúde. Mais concretamente, o Dr. Luís Guedes, do Centro Hospitalar Materno Infantil do Porto, propôs-nos o estudo do comportamento do índice de massa corporal e da pressão arterial média ao longo da gravidez em gestantes normotensas e hipertensas, e a identificação das diferenças entre os seus comportamentos. Sugeriu ainda que se estudasse a possível relação entre a pressão arterial média e o índice de massa corporal. As variáveis em estudo foram analisadas separadamente e tendo em conta duas metodologias, sendo que a estrutura longitudinal, como muito importante que é neste tipo de estudo, esteve sempre presente.

Os dados recolhidos para as variáveis em estudo foram tratados através de duas metodologias: o modelo de regressão com estimação pelo método dos mínimos quadrados e o modelo linear de efeitos mistos. É importante referir que ao longo do estudo dos modelos foram introduzidas várias variáveis explicativas, contudo, nem todas se revelaram estatísticamente significativas.

Eventualmente teríamos diferenças significativas entre os comportamentos das gestantes de acordo com o seu grupo etário (existe sugestão disto por parte dos dados) mas como não foram identificadas gestantes hipertensas pertencentes ao grupo etário mais novo, não foi possível considerar esta hipótese no estudo.

No estudo da evolução da pressão arterial média o modelo que se melhor se ajustou aos dados foi o modelo MMQG com progressão cúbica no tempo. Neste modelo todas as variáveis explicativas, $tempo$, $tempo^2$, $tempo^3$ e hipertensão apresentaram-se estatísticamente significativas. O facto da variável hipertensão ser significativa permite inferir que tendo em conta o estado hipertensivo da gestante, a pressão arterial média, apresenta valores distintos. Mais precisamente, para grávidas com hipertensão arterial crónica, a pressão arterial média ao longo da gravidez tem valores significativamente mais altos em todos os tempos. Este facto vai de encontro a resultados referidos em alguns artigos documentados na bibliografia. Assim, o modelo apresenta duas curvas cúbicas com comportamento idêntico, a pressão arterial média vai diminuindo até às 18-22 semanas onde atinge o seu valor mínimo aumentando novamente até às 29-33 semanas e estabilizando este crescimento até ao momento do parto, onde atinge

o seu valor máximo. A curva representativa das gestantes hipertensas encontra-se acima da curva representativa das gestantes normotensas, o modelo prevê, ao longo da gravidez, um valor médio de pressão arterial média mais alto para a população hipertensa. O modelo linear de efeitos mistos aplicado ao estudo desta variável, mostrou não ser o melhor. A inserção de efeitos aleatórios no modelo não revelou ser adequado, a percentagem de variância explicada pelos mesmos no modelo era bastante baixa, rondando os 2%, o que indica que a variabilidade entre as gestantes não é significativa.

Relativamente à evolução do índice de massa corporal foi ajustado um modelo linear de efeitos mistos, com progressão temporal quadrática. Neste modelo foi visível a presença de *outliers* que influenciavam a normalidade dos resíduos, como tal foi necessário retirá-los. As variáveis do modelo, $tempo$, $tempo^2$ e hipertensão, apresentaram-se estatísticamente significativas. Mais uma vez a variável explicativa hipertensão mostra que de acordo com o estado hipertensivo da gestante os valores do índice de massa corporal são díspares. O efeito aleatório assumido no modelo foi um efeito na constante, a variância explicada pelo mesmo foi aproximadamente 70%, o que significa que a variabilidade inter-individual para o índice de massa corporal durante a gravidez é bastante elevada, e, de facto, era necessário um modelo com efeito aleatório para prever a evolução do índice de massa corporal ao longo da gestação. Este modelo apresenta duas curvas quadráticas, sendo que a curva representante da população hipertensa apresenta um valor médio de IMC superior ao da curva representante da população normotensa. Apesar disto, o comportamento das curvas é idêntico, obervando-se um aumento do índice de massa corporal até ao momento do parto, onde é atingido o seu máximo.

Quando realizado o estudo sobre a relação entre a pressão arterial média e o índice de massa corporal, deparámo-nos com o mesmo que aconteceu com o estudo sobre a evolução da pressão arterial média: o modelo linear de efeitos mistos com efeito aleatório, quer na constante ou no índice de massa corporal, revelou uma percentagem de variância explicada pelo mesmo inferior a 1%, daí que não se tenha considerado este modelo como o melhor. A tentativa de aplicação de uma interação entre o IMC e a hipertensão não foi bem sucedida, uma vez que esta não se revelou estatisticamente significativa (valor-p=0.994). O melhor modelo que se ajustou aos dados foi o modelo de regressão com estimação pelo método dos mínimos quadrados generalizados, sendo as variáveis explicativas, estatisticamente significativas, as seguintes: o índice de massa corporal e o estado hipertensivo da gestante. O modelo prevê que um aumento de uma unidade no IMC provoca um aumento de $0.25mmHg$ na PAM média em ambas as gestantes. Apesar disto, este modelo prevê que, independentemente do valor de IMC, a PAM média em gestantes hipertensas é superior em $13.09mmHg$ em relação às gestantes normotensas. Assim, espera-se duas retas paralelas, embora a que represente os valores previstos pelo modelo nas gestantes hipertensas esteja acima da que representa os valores previstos pelo modelo em gestantes normotensas.

A principal limitação deste estudo foi claramente o reduzido tamanho amostral de gestantes hipertensas. Num total de 461 mulheres apenas 32 mulheres apresentavam hipertensão arterial crónica, e este facto conduziu a algumas limitações no estudo dos

modelos, pois teve-se sempre em conta o número de parâmetros que seria razoável estimar de acordo com este número. A inserção de mais interações nos modelos poderia melhorar, certamente, as previsões da população hipertensa, contudo devido à razão apresentada não foi possível prosseguir o estudo nesse sentido. De qualquer forma, a prevalência da hipertensão arterial crónica em grávidas é bastante baixa e as mulheres foram seguidas durante praticamente toda a sua gravidez.

Após a conclusão deste trabalho, os resultados obtidos no mesmo foram apresentados e discutidos com a equipa médica envolvida e não foram detetadas incongruências com a sensibilidade clinica desses elementos. Como tal, o trabalho foi entretanto submetido para publicação num jornal internacional com arbitragem científica.

Referências

Barra, S., do Carmo Cachulo, M., Providência, R., and Leitão-Marques, A. (2012). Hypertension in pregnancy: The current state of the art. *Revista Portuguesa de Cardiologia*, (31(6)).

Cabral, M. S. and Gonçalves, M. H. (2011). *Análise de Dados Longitudinais.* Sociedade Portuguesa de Estatística.

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2002). *Analisys of Longitudinal Data.* Oxford University Press, Oxford.

Edwards, L. E., Hellerstedt, W. L., Alton, I. R., Story, M., and Himes., J. H. (1996). Pregnancy complications and birth outcomes in obese and normal-weight women: effects of gestational weight change. *Obstetrics and Gynecology*, 87(3):389–394.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis.* John Wiley and Sons, Inc., New York.

Fox, J. and Weisberg, S. (2010). *An Appendix to An R Companion to Applied Regression, Second Edition.*

Gomes, E., Soares, A. L., and Campos, R. (2012). Obesidade e gravidez: conhecer para atuar precocemente? a realidade numa unidade de saúde familiar. *Revista Portuguesa de Endocrinologia, Diabetes e Metabolismo*.

Harville, D. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, (61):383–385.

Hermida, R. C., Ayala, D. E., and Iglesias, M. (2001). Predictable blood pressure variability in healthy and complicated pregnancies. *American Heart Association*.

Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*. John Wiley and Sons, Ltd.

Kirchkamp, O. (2014). *Mixed effects models*.

Laird, N. and Ware, J. (1982). *Random-Efects models for longitudnal data.*, volume 38. Biometrics.

Latifa Mochhoury, Rachid Razine, J. K. M. K. and Barkat., A. (2013). Body mass index, gestational weight gain, and obstetric complications in moroccan population. *Journal of Pregnancy*.

Macdonald-Wallis, C., Lawlor, D. A., Fraser, A., May, M., Nelson, S. M., and Tilling, K. (2012). Blood pressure change in normotensive, gestational hypertensive, preeclamptic, and essential hypertensive pregnancies. *American Heart Association*, (59:1241-1248).

Macedo, M. E., Lima, M. J., Silva, A. O., Alcântara, P., Ramalhinho, V., and Carmona., J. (2007). Prevalência, conhecimento, tratamento e controlo da hipertensão em portugal. estudo pap. *Revista Portuguesa de Cardiologia*, (26(1)).

Mattar, R., Torloni, M. R., Betrán, A. P., and Merialdi4, M. (2009). Obesity and pregnancy.

Medicine, I. (1990). Nutrition during pregnancy. part i, weight gain. *Comittee on Nutritional Status During Pregnancy and Lactation*.

Nogueira, A. I. and Carreiro, M. P. (2013). Obesidade e gravidez. 23(1).

Obstetricians, T. A. C. and Gynecologists (2012). Chronic hypertension in pregnancy.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer-Verlag, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-113.

R Development Core Team (2012). R: A language and environment for statistical computing. *ISBN 3-900051-07-0, retrieved from http://www.R-project.org.*

S. Paiva, L. R., M.Campos, Melo, M., Santos, J., Lobo, A., Sobral, E., Marta, E., Moura, P., and Carvalheiro., M. (1998). Obesidade e gravidez. *Revista Portuguesa de Endocrinologia, Diabetes e Metabolismo*.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Seely, E. W. and Maxwell., C. (2007). Chronic hypertension in pregnancy.

Shub, A., Huning, E. Y.-S., Campbell, K. J., and McCarthy, E. A. (2013). Pregnant women's knowledge of weight, weight gain, complications of obesity and weight management strategies in pregnancy. (6:278).

Twisk, J. (2003). *Applied Longitudinal Data Analysis for Epidemiology.* Cambridge University Press.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.