# Aplicação de Modelos Multinível para o Estudo da Sinistralidade no Retalho Alimentar


Catarina Silva Castro
Mestrado em Engenharia Matemática
Departamento de Matemática
2015

**Orientador**
Joaquim Fernando Pinto da Costa
Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

# Agradecimentos

Agradeço ao meu coordenador de Estágio na Sonae - Maia, Engenheiro João Rodrigues, diretor de Segurança e Saúde no Trabalho, a toda a sua equipa fantástica que me acolheu e aos restantes elementos daquela empresa com quem tive o gosto de partilhar esta experiência.

Agradeço ao meu coordenador de dissertação pelo tempo dedicado e aos restantes professores da Faculdade de Ciências da Universidade do Porto, dos diversos departamentos, com quem tive oportunidade de aprender.

Agradeço a todos os meus amigos destes cinco anos de faculdade - alguns para a vida - e aos amigos de sempre e de hoje por estarem por perto, pelo menos do coração. Agradeço muito à minha família pelas oportunidades que abracei.

Agradeço a quem não me deixou ficar para trás, desistir ou fraquejar em algum momento. Agradeço pela força, pelo cuidado, pela ajuda, pelo apoio e pela imensurável coragem de o terem feito por acreditarem em mim.

Muito obrigado.

Catarina Silva Castro

# Resumo

A sinistralidade é um tema transversal a todos os universos laborais. Muito para além dos direitos dos trabalhadores e dos pressupostos de segurança inerentes às instituições, desponta uma cultura de segurança que é preciso cultivar e fazer crescer no seio do mundo laboral.

Ao contactar com o dia-a-dia de uma equipa de Segurança e Saúde no Trabalho adquire-se uma nova visão deste problema. Uma visão mais consciente de que os perigos estão efetivamente por perto e que combater a negligência dos colaboradores se torna numa luta falível, por mais que se invista nas condições de segurança, em novos meios e processos, surgirão sempre novos riscos.

Contudo, cabe-nos sempre fazer um pouco mais. No cenário de uma loja do retalho, os riscos surgem a cada passo, em cada pequena tarefa. Partindo da análise de dados da sinistralidade procurei um meio de ação, procurei outras causas que extrapolassem as imediatas. Reconhecendo o comportamento da variável *número de acidentes de trabalho* numa realidade laboral e cruzando com o seu contexto conseguimos encontrar relações de possível "causa-efeito". O meu objetivo foi apurá-las, averiguá-las e reconhecer a envolvência da variável resposta.

Mesmo considerando certas caraterísticas das lojas, como o volume de negócios e dados dos colaboradores, os resultados obtidos não eram os pretendidos. Por isso, procurei com este trabalho realçar as diferenças entre as tipologias de loja e as disposições geográficas. Esta organização estrutural dos dados permitiu então assumir um padrão hierárquico e uma abordagem diferente ao problema.

Atendendo às limitações dos modelos lineares, mesmo considerando interações entre variáveis, este estudo culmina na aplicação de modelos lineares hierárquicos ou multinível. A sua aplicação não implica os pressupostos da regressão linear e respeita o contexto dos dados, não assumindo erroneamente a independência dos diversos registos. Usando métodos de estimação de máxima verossimilhança restrita, os modelos multinível permitiram avaliar a variabilidade dos registos nos diferentes níveis e ajustar um modelo mais adequado à realidade do problema.

**Palavras-Chave:** acidentes de trabalho, sinistralidade, modelos lineares, modelos lineares hierárquicos, modelos multinível, métodos de estimação, critérios de informação.

# Abstract

Accidents are a transverse theme to all labor universes. Way beyond worker's rights and institution's inherent safety assumptions, there is a growing need for improvement in safety culture in the labor world.

By having daily contact with a Workplace Safety team, one easily acquires a new perspective over this problematic. A more conscious awareness that dangers are, in fact, closer than we assume and that fighting workplace negligence becomes a losable fight. No matter how big the investment in workplace conditions, new means and processes, there will always appear new unexpected risks.

However, it is up to us to try and do a little bit more about it. Picturing a small grocery store, the danger is present in every step, every little task. Starting by analysing accident related data, I tried to look into means of action, into other causes that would surpass the obvious ones. Recognizing the behaviour of the *number of workplace accidents* variable in a workplace environment and matching it with its context we are able to find possible "cause-effect"relations. My objective was finding them, investigating them and getting to know the involvement of the answer variable.

Even by joining multiple store variables, different business volumes and partner's data, the obtained results were not the wanted ones. Taking this into consideration, I decided to focus on the inter store topology differences, as well as geographical disposition, to assume a hierarchical pattern from the data and look into the problem from a different perspective.

Taking into consideration the limitations presented by linear models, even if variable interactions are considered, they still showed tremendous limitations. Therefore, this study ended up using linear or multilevel hierarchical models. Its application does not imply linear regression presumptions and respects the data context, without erroneously assuming its independence. Using restricted maximum verisimilitude estimation methods, multilevel methods allowed for an evaluation of the variability of records at different levels, adjusting the model to one closer to reality.

**Keywords:** workplace accidents, accidents, linear models, hierarchical linear models, multilevel models, estimation methods, information criteria.

# Conteúdo

**Agradecimentos** 3

**1 Introdução** 15
- 1.1 Contexto . . . 15
- 1.2 Conceito . . . 17
- 1.3 Enquadramento . . . 18
- 1.4 Motivação e Objetivos . . . 19
- 1.5 Estrutura da Dissertação . . . 20

**2 Metodologias Usadas** 21
- 2.1 Modelos Lineares . . . 21
  - 2.1.1 O Método de Regressão . . . 21
  - 2.1.2 Descrição do Modelo . . . 22
  - 2.1.3 Os Pressupostos e a Estimação . . . 23
- 2.2 Modelos Multinível . . . 26
  - 2.2.1 Anotações Históricas . . . 26
  - 2.2.2 Da Regressão Clássica à Regressão Multinível . . . 27
  - 2.2.3 Génese da Regressão Multinível . . . 28
  - 2.2.4 Vantagens . . . 28
  - 2.2.5 Áreas de Aplicação e Fundamentos . . . 30
  - 2.2.6 Contrução do Modelo . . . 30
  - 2.2.7 Limitações e Desenvolvimentos Futuros . . . 33
- 2.3 Métodos de Estimação . . . 34
- 2.4 Critérios de Informação . . . 35

**3 Aplicação** 37
- 3.1 Análise Introdutória . . . 37
- 3.2 Análise Exploratória . . . 37
- 3.3 Métodos de Análise Multinível . . . 39
- 3.4 Análise Multinível . . . 45

**4 Conclusões e trabalhos futuros** 55

# Lista de Figuras


1.1 Representação da Pirâmide de Bird (1969) . . . 18
1.2 Sequência da metodologia IPAR . . . 19

2.1 Representação da Reta de Regressão do Modelo (2.5) Campinas . . . 24

3.1 Número total de acidentes de trabalho registados mensalmente nas diversas DOPs entre Janeiro e Junho de 2014 nas insígnias Continente Hipermercado, Continente Modelo e Continente Bom Dia. . . . 38
3.2 Estrutura padrão dos dados, considerando dois níveis: a referência geográfica de operações (DOP) e os registos mensais associados. . . . 39
3.3 Output do comando summary do modelo LMER obtido a partir do pacote estatístico lme4. 40
3.4 Output do comando summary do modelo LME obtido a partir do pacote estatístico nlme. 41
3.5 Valores de output do comando ranef para os efeitos do modelo LMER . . . 42
3.6 Valores de output do comando ranef para os efeitos do modelo LME e concretização gráfica dos efeitos aleatórios por DOP através do comando plot(ranef(LME)) . . . 43
3.7 Outputs dos seguintes comandos aplicados ao modelo LME: (a) plot(residuals(LME)); (b) qqnorm(residuals(LME)) e qqline(residuals(LME)) . . . 43
3.8 Outputs do comando plot(LME) . . . 44
3.9 Padrão hierárquico dos dados, considerando três níveis: a tipologia de loja (INS), a referência geográfica de operações (DOP) e os registos mensais associados . . . 45
3.10 Organograma simplificado da MUC do retalho alimentar Sonae . . . 45
3.11 Nível 1 . . . 46
3.12 Nível 2 . . . 47
3.13 Nível 3 . . . 48
3.14 Sumário do modelo m1 . . . 48
3.15 Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme . . . 49
3.16 Sumário do modelo m7 . . . 50
3.17 Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme . . . 51
3.18 Sumário do modelo m6 . . . 52
3.19 Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme . . . 53

# Abreviaturas

| | |
|---|---|
| ABS | Percentagem de Absentismo |
| AIC | Akaike Information Criterion |
| AT | Acidente de Trabalho |
| BD | Continente Bom Dia |
| BIC | Bayesian Information Criterion |
| CN | Centro Norte |
| CNT | Continente Hipermercado |
| CS | Centro Sul |
| DOP | Direção de Operações |
| DP | Número de Dias Perdidos |
| FTES | Full Time Equivalent |
| HE | Número de Horas Extra |
| ICC | Intraclass Correlation Coefficient |
| IF | Índice de Frequência |
| IG | Índice de Gravidade |
| iid | Independente e Identicamente Dsitribuído |
| INS | Insígnia |
| IPAR | Identificação de Perigos e Avaliação de Riscos |
| IPP | Incapacidade Parcial Permanente |
| ITA | Incapacidade Temporária Absoluta |
| MC | Modelo Continente |
| MDL | Modelo Continente |
| ML | Maximum Likelihood |
| MUC | Marca Única Continente |
| PROD | Produtividade |
| REML | Restricted Maximum Likelihood |
| SST | Saúde e Segurança no Trabalho |
| STK | Volume de Stock em Loja |
| TR | Número de Transações |
| VND | Volume de Vendas |

# Capítulo 1.

# Introdução

Tendo em conta que o termo "sinistro"indicia por si só a ocorrência de um dano e que todo o dano está associado a um risco, podemos deduzir que a sinistralidade releva os aspetos em que estamos mais suscetíveis ao risco.

Segundo a Lei de Murphy, "havendo risco, persistem as possibilidades de danos". É de facto impossível eliminar todos os riscos existentes. Visto que a definição de acidente designa-o como qualquer evento ou facto não desejado que provoque danos a propriedade ou pessoa, reportando esta definição para o meio laboral, constatamos que a sinistralidade laboral é um problema premente. Sendo a segurança definida como a isenção de risco, as medidas de segurança são então um meio para combater os riscos, firmando compromissos para a prevenção dos acidentes.

## 1.1 CONTEXTO

A tendência de globalização a que a atualidade nos obriga faz com que a competitividade seja cada vez mais uma constante a que o universo empresarial está sujeito, atendendo às exigências do mercado e dos consumidores.

Um mercado cada vez mais atento à segurança e qualidade, procurando sempre os melhores produtos, conduz a que as empresas sejam analisadas em vários aspetos, nomeadamente no âmbito financeiro, social e ambiental, nas condições de segurança e trabalho dos seus colaboradores, nas suas referências de mercado, entre outros.

O ano de 1992 foi designado pela Comunidade Europeia como o "Ano Europeu da Segurança, Higiene e Saúde no local de trabalho". Este foi um ponto de partida da consciencialização para a prevenção e melhoria das condições de trabalho. Após a adesão de Portugal à Comunidade Económica Europeia, as evoluções a nível europeu foram implementadas em Portugal, tal como as diversas diretivas legais relativas às condições de higiene, segurança e saúde no trabalho Moreira [2008]. Associada à evolução humana e tecnológica, a regulamentação laboral tornou-se cada vez mais criteriosa. Ao longo das últimas décadas, têm-se verificado melhorias consideráveis em prol da saúde, segurança e bem estar dos colaboradores. As empresas encontraram na implementação de técnicas de trabalho e equipamentos mais sofisticados um meio para aumentar a produtividade e elevar o seu nível de excelência, reduzindo assim os riscos laborais.

Como forma de chegar ao sucesso, as empresas estão sempre disponíveis para a inovação. Graças ao apelo da competitividade e da responsabilidade social, o ambiente de trabalho e a consciencialização para a segurança têm-se tornado divisas de referência.

Surge uma referência a António Vieira de Carvalho num artigo do V Congresso de Excelência em Gestão que vem de encontro ao acima descrito: "A segurança no trabalho, como meio de precaução de acidentes no empreendimento, deve ser considerada como um dos fatores imprescindíveis para o alcance da produtividade", fim de citação Leão et al. [2009].

A sinistralidade laboral é, por isso, um problema a combater. As condicionantes deste problema residem nas tarefas a desempenhar, nas condições em que estas são desenvolvidas e nas técnicas que são utilizadas. Sendo assim, o investimento na melhoria destes aspetos traz inúmeras vantagens.

A prevenção dos acidentes de trabalho, hoje em dia tão relevante para o desempenho e imagem das empresas, tem também repercussões económicas. Atendendo a este último aspeto, surgem diversas teorias ilustrativas associadas. A metáfora mais vulgar da proporção entre os custos diretos e custos indiretos dos acidentes de trabalho surge nos Estados Unidos da América, em 1931, e ficou conhecida como o Iceberg de Heinrich. Heinrich demonstrou, a partir da análise de 5000 casos, que o custo dos acidentes de trabalho se traduzia numa proporção de 1:4, em que as perdas imputadas às empresas eram muito superiores às verbas transferidas pelas seguradoras.

Os sinistros implicam custos diretos para as empresas, como tempo de trabalho perdido, os danos em materiais e equipamentos, custos de assistência e quebra na produtividade. Posteriormente, os custos tendem a aumentar, nomeadamente com o agravamento dos seguros, a reintegração laboral e a perda de distinções e referências ao nível da segurança e qualidade, afetando, por conseguinte, a sua competitividade, desempenho e confiança no âmbito económico e social.

Contudo, a automatização e o parcelamento de tarefas e o stress e a pressão que advêm da competitividade têm também vulgarizado os riscos psicossociais. Embora não tão comuns no passado, estes conduzem agora a um cuidado acrescido com o bem estar dos colaboradores. Num artigo sobre sinistralidade laboral Monteiro et al. [2013] surge uma referência a Corcoran datada de 2002, em que o autor defende que a criação de uma cultura de segurança nas organizações e empresas é necessária para a diminuição da sinistralidade.

Pidgeon Pidgeon [1991], em 1991, definiu o conceito "cultura de segurança" como um conjunto de cuidados, normas, atitudes, regras, práticas sociais e técnicas que têm por objetivo minimizar a exposição dos trabalhadores, chefias, consumidores e público em geral, a condições consideradas perigosas ou prejudiciais.

É certo que uma cultura de segurança se baseia na prevenção, mas é a análise da sinistralidade laboral que nos direciona ao problema e permite identificar meios de ação.

Os índices de sinistralidade laboral são o método de análise mais eficiente, sendo os mais significativos:

- O índice de frequência que dado pelo número médio de acidentes de trabalho que têm por consequência uma ITA por cada mil horas trabalhadas.
- O índice de gravidade que é dado pelo número médio de dias perdidos por cada milhão de horas trabalhadas.
- O índice de incidência que é o número médio de acidentes de trabalho que incorrem em ausência laboral (ITA) por cada mil trabalhadores.

Citando Sylvia Constant Vergana, "se o problema é uma questão a investigar, o objetivo é um resultado a alcançar", referenciada num artigo do V Congresso de Excelência em Gestão Leão et al. [2009]. A redução da sinistralidade e implementação e investimento em medidas de prevenção para a melhoria das condições de trabalho passa pela identificação dos perigos e controlo e avaliação dos riscos.

Assim sendo, a investigação e a análise dos acidentes de trabalho são o caminho a seguir para a sua prevenção no futuro, que passa por detetar falhas, causas e medidas a implementar e pela consciencialização do universo das organizações para esta problemática.

O estudo levado a cabo por Heinrich foi o que mais consciencializou, ao longo dos tempos, para a implementação de um sistema de prevenção de riscos, com vista à redução dos custos às organizações. Torna-se assim imperioso, tomar medidas preventivas para minimizar os riscos, pois os acidentes de trabalho é um dos principais problemas de segurança e saúde prementes.

O olhar preventivo sobre os acidentes passa pela análise de um conjunto de estatísticas confiáveis e são essas que permitem uma visão efetiva da sinistralidade laboral.

## 1.2 CONCEITO

A definição de acidente de trabalho depende dos objetivos de quem a formula e do contexto e implicações a que visa CARMO et al. [1995]. Não existe uma definição que seja massiva e satisfaça em todos os contextos em que é referida.

Lucca e Fávero Lucca and Fávero [1994] apontam o conceito legal que apenas assenta no prejuízo sofrido no trabalho e na indeminização do acidentado e não na prevenção. Zocchio Zocchio [1971] aponta um conceito prevencionista e define-o como "uma ocorrência não programada", inesperada ou não, que interrompe ou interfere no processo normal de uma atividade, ocasionando perda de tempo útil, lesões nos trabalhadores e/ou danos materiais.

Acidente é uma palavra de origem latina, provém de *accidens* que significa acaso. Legalmente, no artigo 6 da lei nº100/97, o acidente de trabalho está definido como aquele que se verifique no local e no tempo de trabalho e produza direta ou indiretamente lesão corporal, perturbação funcional ou doença de que resulte redução na capacidade de trabalho ou de ganho ou a morte". O descrito neste artigo ainda abrange certos acidentes sucedidos fora do local ou do tempo de trabalho ou nas deslocações para o local de trabalho, designados por acidentes de trabalho no itinerário.

Desta definição surge a de quase acidente ou incidente que é tido como qualquer evento com potencial para provocar danos, mas que não chega a causá-los.

Na descretização do conceito acidente de trabalho aparece a distinção entre acidente de trabalho que resulta de um ação súbita e doença profissional que é resultado de uma causa lenta e progressiva. Estudos liderados por Frank Bird permitiram que em 1969 surgisse a teoria da *Insurance Company of North America*. A teoria consistia na reprodução em proporção dos danos pessoais resultantes dos acidentes de trabalho sob a forma de uma pirâmide. A primeira representação, que surgira em 1966 da autoria de Bird, foi reformulada em 1969 e nela foram introduzidos os quase acidentes ou incidentes, na relação representada na figura 1.1 de2.

As situações e cenários de risco a que os trabalhadores estão sujeitos são, por isso, muito superiores ao volume de sinistros que se concretizam. Para a análise da sinistralidade efetiva são consideradas como consequências dos acidentes de trabalho três designações generalizadas:

- AT sem ITA, do qual não se contabilizam dias perdidos


**Figura 1.1:** Representação da Pirâmide de Bird (1969)

- AT com ITA, que dão lugar a uma incapacidade temporária absoluta e ausência laboral e dos quais pode resultar uma percentagem de incapacidade temporária permanente (IPP)
- AT fatal.

## 1.3 ENQUADRAMENTO

> "There are risks and costs to a programme of action, but they are far less than the long range risks and costs of comfortable inaction." **John F. Kennedy**

### SEGURANÇA E SAÚDE NO TRABALHO NO CONTEXTO SONAE

A redução da sinistralidade laboral é um fator favorável às empresas, no que respeita à sua fiabilidade organizacional e como modelo de segurança. É ainda favorável na confiança que potencia aos seus clientes e aos seus colaboradores no contexto pessoal e social que os envolve, sendo ainda profícuo a quem deles depende.

A realidade que presenciamos no que se refere a acidentes de trabalho desagrada, claro está. Este aspeto incita a que sejam tomadas medidas para se controlar esta tendência.

No universo do retalho alimentar as ações de prevenção são diversificadas e apontam para a melhoria das condições de trabalho e do desempenho logístico dos colaboradores. Nos últimos anos, a meta definida na Sonae foi "Zero acidentes"e sob este mote foram tomadas medidas pela Equipa de Segurança e Saúde no Trabalho. Investiu-se, por isso, na formação dos colaboradores com o objetivo de consciencializar para os comportamentos de risco para assim se concretizar uma redução do número de acidentes de trabalho proporcionando mais confiança aos clientes. Foi criada uma entidade designada por "Animador de Segurança"que não são mais do que supervisores/ avaliadores da segurança da loja a que estão alocados, tendo esta medida sido estendida a todas as insígnias MC. A ideia foi criar uma personagem que estivesse no teatro das organizações e pudesse prevenir acidentes e promover melhores condições de trabalho através da identificação de possíveis acidentes, divulgação de campanhas e dinamização de condutas corretas. A sua ação passava por identificar riscos, fazer o registo, avaliação e investigação de ocorrências e orientar a concretização de simulacros.

Na empresa, a análise da sinistralidade passa pelo controlo dos índices de frequência e gravidade, do volume total de acidentes de trabalho e de acidentes de trabalho que incorrem numa ITA, no itinerário e no local de trabalho e do número de dias de ausência laboral.

A gestão dos riscos que tanto são referenciados como causa influenciadora do número de sinistros, segue a metodologia IPAR, que designa a identificação de perigos e avaliação de riscos. Uma equipa de Segurança e Saúde no Trabalho baseia o seu trabalho neste ciclo de ação e prevenção. A este órgão organizacional cabe identificar e caraterizar os perigos, isto é, tudo aquilo que pode causar lesões ou danos, concretização de medidas de prevenção e avaliar os riscos; medindo os riscos para a saúde e segurança dos trabalhadores decorrentes dos perigos do local de trabalho, classificando-os, tendo em conta a adequabilidade dos controlos existentes. Este processo segue a cadeia descrita na figura 1.2.


**Figura 1.2:** Sequência da metodologia IPAR

## 1.4 MOTIVAÇÃO E OBJETIVOS

No âmbito do projeto Sonae *10th Call For Solutions Universities* fiquei alocada a um desafio imbuído na temática dos acidentes de trabalho. Tinha por mote "Caracterizar os acidentes de trabalho e o perfil do acidentado no contexto Sonae MC"e destacava como tarefas:

- caracterizar a sinistralidade;
- caracterizar o perfil do acidentado;
- modelar dados para apurar distribuições e tendências;
- relacionar acidentes trabalho com causas internas e externas.

O objetivo principal baseou-se na análise da sinistralidade em três das insígnias do retalho alimentar, Sonae MC, associadas à MUC, nomeadamente, Continente Hipermercado, Continente Modelo e Continente Bom Dia. Este estudo, desenvolvido para análise de dados em contexto real, conduziu à concretização de um estudo do caso sujeito a diversas limitações e que permitiu contactar com o problema, ter consciência do seu alcance e conhecer todo o processo de tratamento dos sinistros. Através desta abordagem conheci os casos e causas mais comuns e encontrei caraterísticas que me permitiram traçar perfis para o sinistrado loja do retalho alimentar.

O interesse que daqui surge é a prevenção, em que partindo da constatação do risco procuramos um meio para o evitar as consequências. A metodologia dos 3 C's traça um percurso de ação: caso, causa, contramedida. Esta veicula então para a análise do caso, a averiguação da causa e a consequente elaboração da contramedida e baseia-se no triângulo da segurança segundo a interação dos fatores de natureza pessoal, comportamental e ambiental.

Nos fatores de natureza pessoal estão os comportamentos associados à tarefa que o colaborador desenvolve, quanto às suas características, à sua organização e ao seu modo de desempenho. São também considerados os comportamentos inseguros da liderança, nomeadamente no caso de algum dos contornos da tarefas a realizar não estar adequadamente delineado.

Nos fatores de natureza comportamental relacionam-se os comportamentos inseguros individuais, motivados pela cultura, por hábitos, pelo "porque não fazer assim" que parece tornar as tarefas mais rápidas e mais fáceis e que espelham negligência ou mesmo falha na formação para a tarefa que desempenham.

Os fatores de natureza ambiental congregam duas estâncias: as condições inseguras, como o desgaste de equipamentos, deficiências na estrutura, más condições no local de trabalho e as situações em causa, isto é, as circunstâncias exteriores que conduziram à ocorrência do acidente.

Foi este um dos pontos de partida para o estudo mais orientado de acontecimentos concretos e indícios mais precisos deste problema. A investigação de novas causas e novos métodos para a redução do número de sinistros levou a um estudo mais alargado do volume de acidentes nos últimos anos e ao seu cruzamento com outras condicionantes loja, como localização geográfica espectável através da designação da DOP ou a produtividade. Estes aspetos que não estão diretamente ligadas à sinistralidade podem, contudo, ser diferenciadores no volume de sinistros ou na sua variação ao longo do tempo.

A concretização desta tese passa pela análise de um conjunto de dados que nos permita reconhecer razões e averiguar quais os locais para agir em prol da redução da sinistralidade.

Apesar da evolução nos índices ser favorável, observando o comportamento de alguns gráficos, é visível um aumento da sinistralidade em alguns dos meses nas insígnias em estudo. Foi possível verificar que de 2012 em diante através dos índices acumulados a junho, que são calculados considerando todos os sinistros ocorridos nos primeiros seis meses de cada ano, comparativamente aos valores dos anos transatos observa-se um aumento que se parece repercutir no ano de 2014. Partindo deste aspeto, parece conveniente ter em consideração os dados relativos a esse ano com o intuito de indagar o porquê desta tendência evolutiva.

Entre 2010 e 2014, as modificações imputadas a esta entidade, relacionadas com a formação de colaboradores, crescimento e maturação do seu universo público, a nível da produtividade, do volume de vendas, do crescente número de colaboradores, políticas da empresa, número de superfícies comerciais, e tanto outros aspetos foram inúmeras. Entre ações de prevenção e o surgimento de outros fatores problemáticos pretendem-se respostas para problemática da sinistralidade.

## 1.5 ESTRUTURA DA DISSERTAÇÃO

Ao longo deste primeiro capítulo é apresentada a realidade da sinistralidade laboral, tendo em consideração todo o contexto inerente aos acidentes de trabalho, desde as suas causas à forma de prevenção.

No próximo capítulo são descritas as metodologias relativas a modelos, métodos e critérios usadas no terceiro capítulo, onde consta a descrição e análise do tratamento dos dados.

Por fim, no quarto capítulo surgem algumas conclusões e desenvolvimentos futuros que possam conduzir à génese do problema.

# Capítulo 2.

# Metodologias Usadas

## 2.1 MODELOS LINEARES

Os modelos de regressão linear usuais permitem descrever ou prever determinadas variáveis, designadas por variáveis resposta. Estas estão dependentes de outras variáveis, em vulgo variáveis explicativas porque tentam explicar a variação da variável resposta, mas também designadas por covariáveis ou variáveis preditoras.

Partindo das ideias chave apresentadas, a regressão passa pelo ajuste de um conjunto de parâmetros de um modelo que se aproxime do comportamento da variável resposta.

O modelo de regressão tradicional envolve então uma dependência linear entre a variável resposta, considerada como aleatória, e as variáveis explicativas que são não estocásticas, das quais está dependente a variável resposta.

Neste universo, podemos considerar dois grupos de modelos lineares:

- **modelos de regressão**.
- **modelos de análise de variância**.

Enquanto os primeiros têm por objetivo estimar parâmetros para descrever relações e estimar predições, os outros baseiam-se na comparação da influência dos fatores sobre o comportamento da variável dependente.

### O Método de Regressão

O método de regressão está no centro da modelação estatística e é inerente à questão "o que é que acontece a $Y$ quando $x$ varia?" que é constante em diversos estudos. $Y$ designa genericamente a variável resposta e $x$ é uma variável explicativa Bergamo [2002].

A regressão linear não surge apenas para estudar o comportamento da variável resposta mas, por vezes, procura também estimar ou controlar o seu desempenho a longo ou curto prazo. A regressão nasceu da necessidade de relacionar um conjunto de observações de certas variáveis $X_k$, em que $k$ varia de 1 a $p$ com uma variável resposta $Y$.

A regressão linear consiste então num processo de estimação que está inerente aos modelos lineares e que tem subjacente uma reação do tipo:

$$Y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p \tag{2.1}$$

em que $\beta_0, \ldots, \beta_p$ e $X_0 = 1$ são os parâmetros procurados da relação linear.

Pode ter dois objetivos diferentes: explicativo ou preditivo. O explicativo passa por demonstrar uma relação matemática de aparente causa-efeito entre uma variável resposta e uma ou mais variáveis explicativas. O preditivo pretende estimar o comportamento da variável $Y$ por meio de uma combinação linear de parâmetros que dependem do valor das variáveis explicativas designadas por $X_k$, de forma a futuramente, a partir da observação das variáveis $X_k$ prever o valor da variável resposta, sem que seja necessário medi-la Matos [1995].

Tal como neste estudo, a metodologia da regressão linear é vulgarmente usada. Ainda que em muitos desses casos apenas surja em análises primárias, esta metodologia merece um estudo rigoroso, pela sua versatilidade e aplicabilidade em múltiplos contextos.

### Descrição do Modelo

A regressão linear permite modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes a partir de um conjunto de variáveis independentes. Considera-se regressão linear simples se é descrita a relação entre a variável dependente e uma variável independente e regressão linear múltipla se a relação linear envolve uma variável dependente e várias variáveis independentes Rodrigues [2012].

Assumindo, em parte, a notação anteriormente referida e tomando agora a notação matricial, temos que:

$Y$ é da forma $Y = [Y_1, ..., Y_n]^T$
$X$ é da forma $X = [X_0, X_1, ..., X_p]$
$X_0$ é da forma $X_0 = [1, ..., 1]^T$
$X_j$ é da forma $X_j = [X_{1j}, ..., X_{nj}]^T$ em que $j = 1$ , ... , $p$
$\beta$ é da forma $\beta = [\beta_0, \beta_1, ..., \beta_p]^T$
$\epsilon$ é da forma $\epsilon = [\epsilon_1, ..., \epsilon_n]^T$

Deste modo, num modelo da forma:

$$Y = X\beta + \epsilon \tag{2.2}$$

denominamos por:

Y o vetor dos registos da variável resposta
X a matriz das observações
$X_j$ o vetor das observações da $j$-ésima variável independente
$\beta_j$ o coeficiente da regressão associado à $j$-ésima covariável do estudo
$\epsilon_i$ o erro associado ao $i$-ésimo indivíduo da amostra, em que $i = 1, ..., n$.

O interesse do estudo reside em saber como é que $Y$ pode ser descrito pelas covariáveis $X_j$ e o efeito da sua variação no comportamento de $Y$ Matos [1995].

Começamos por considerar o modelo nulo em que:

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \ldots, n \tag{2.3}$$

Trata-se do modelo base em que o valor da variável resposta é dado pela sua média e onde não são consideradas quaisquer variáveis explicativas. A média da variável *Y* definida por $\beta_0$ é chamada a constante do modelo, pois trata-se do seu termo independente. A oscilação do valor de *Y* em torno da sua média é ajustada pelo modelo através da inclusão de variáveis preditoras.

Considerando a adição de uma variável explicativa ao modelo temos que:

$$Y_i = \beta_0 + \beta_1 \, X_{i1} + \epsilon_i, \quad i = 1, \ldots, n \tag{2.4}$$

em que temos *n* observações e:

| | |
|---|---|
| $Y_i$ | trata-se do valor da resposta correspondente ao i-ésimo indivíduo |
| $\beta_0$ | é a constante do modelo, ou seja, representa a ordenada na origem |
| $\beta_1$ | é a variação esperada em $Y_i$ quando $X_{1i}$ aumenta uma unidade e representa o declive da reta |
| $X_{i1}$ | é o valor da variável explicativa para o *i*-ésimo indivíduo, que não é mais do que um fator explicativo da variável resposta para esse indivíduo |
| $\epsilon_i$ | é o erro associado ao i-ésimo indivíduo, em que $\epsilon_i \sim N(0,\sigma^2)$ com distribuição i.i.d. |

No caso geral, a equação do modelo de regressão linear considerando *j* variáveis explicativas fica definida por:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \tag{2.5}$$

em que são descritas *n* observações segundo *p* variáveis explicativas e:

| | |
|---|---|
| Y | trata-se do vetor da variável resposta |
| $\beta_0$ | é a constante do modelo |
| $\beta_j$ | o valor para o parâmetro de regressão associado à *j*-ésima variável explicativa |
| $X_j$ | o vetor de valores da *j*-ésima variável explicativa para os *n* indivíduos |
| $\epsilon$ | o vetor dos erros associado aos n indivíduos, em que $\epsilon \sim N(0, \sigma^2)$ com distribuição i.i.d. |

### Os Pressupostos e a Estimação

O destaque dos modelos lineares nos mais diversos estudos deve-se ao facto destes dependerem de forma linear dos seus parâmetros. O primeiro passo do método de regressão é a estimação. O processo de estimação, por sua vez, consta da atribuição de um valor a esses parâmetros, para os quais não se conhece o valor absoluto. O termo "estimação" é então usado para designar a produção de estimativas para os parâmetros de um modelo a partir de uma amostra PINTO [2007]. O ajuste de um modelo linear conduz à procura de uma combinação dos elementos da amostra, que é mais fácil nestes modelos comparativamente

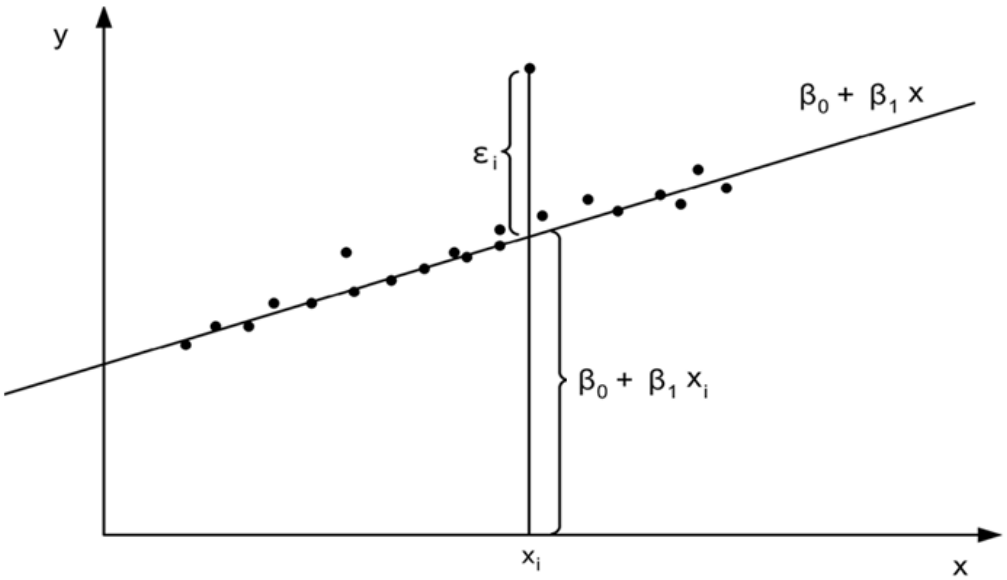


**Figura 2.1:** Representação da Reta de Regressão do Modelo (2.5) Campinas

aos modelos não lineares. Esse conjunto de estimativas para os parâmetros é denominado por estimador, pois pretende-se que a sua utilização no modelo nos permita estimar corretamente a variável reposta. A abordagem mais vulgar e menos tendenciosa para essa estimação passa pelo método dos mínimos quadrados.

Este método de estimação pode ser descrito como o método que se baseia na procura do modelo que se ajuste aos dados e minimize a soma dos quadrados das distâncias das diversas observações à função de regressão.

Pretende-se obter as melhores estimativas dos coeficientes de regressão $\beta_j$ de modo a minimizar os desvios $\epsilon_i$ entre os valores observados e estimados. O anteriormente referido equivale então a minimizar o quadrado da norma do vetor dos erros, $\epsilon$=($\epsilon_1$, ..., $\epsilon_n$). Os coeficientes estimados possuem duas importantes propriedades: são centrados e apresentam uma variância mínima entre os estimadores dos restantes métodos que são centrados e combinação linear de $Y_i$.

Os modelos lineares têm por base alguns pressupostos. Consideramos a independência entre as observações e a hipótese da ausência de multicolinearidade entre as variáveis explicativas. A hipótese da multicolinearidade advém da existência de uma linearidade quase exata entre duas variáveis independentes e surge quando estas possuem uma forte correlação. A verificação desta hipótese influencia os resultados da regressão introduzindo erros padrão elevados, prejudicando, por isso, a modelação do problema. Para ser um bom conjunto de estimadores para determinada variável tem de ter um elevado valor "explicativo". É o seu nível de informação que permite uma melhor predição/estimação da variável resposta, daí que a multicolinearidade influencie negativamente o modelo linear, pois revela redundância na informação Matos [1995] Kasznar and GOLÇAVES [2011]. Baseada na não correlação das observações, pressupõe-se a independência e a aleatoriedade dos erros e, constituindo a necessária hipótese de homocedasticidade, os erros tâm igual distribuição em 0 e a variância dos erros é constante. A não verificação destes pressupostos vem condicionar inúmeras hipóteses deduzidas para estes modelos, daí que também sejam vistos como fragilidades inerentes à regressão usual.

O pressuposto da independência das observações é o ponto fraco mais vulgarmente apontado a estes modelos, uma vez que o contexto dos dados é completamente descartado e todas as análises que daí advêm podem ser postas em causa, caso se refute a premissa da independência das observações. É sem

dúvida o maior problema da regressão linear múltipla ou simples, pois trata-se de um pressuposto central, que em muitas ocasiões, é violado. Na regressão, como consequência da dependência entre as observações sobrevém a subestimação dos erros padrão dos coeficientes de regressão Laros and Marciano [2013].

Uma forma de contornar a dependência das variáveis é incluir interações no modelo linear, isto é, efeitos conjuntos de variáveis. Sucede-se quando o efeito de uma variável preditora depende dos níveis de outra ou outras variáveis preditoras e é definido pelo produto cruzado dessas variáveis. A função de regressão para um modelo com interação é descrita por:

$$Y_i = \beta_{0j} + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i \tag{2.6}$$

Considerando que $X_{i3} = X_{i1} X_{i2}$ obtemos o modelo linear usual em que $X_{i3}$ possui o valor da interação:

$$Y_i = \beta_{0j} + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \tag{2.7}$$

De outro modo, podemos optar pela aplicação da modelação adequada aos dados em estudo, evitando assim o risco da violação dos seus pressupostos. Um dado relevante para o uso da modelação adequada é o agrupamento de uma população que é definido pelo coeficiente de correlação intraclasse. O coeficiente de correlação intraclasse (ICC) representa a homogeneidade entre os indivíduos associados a um mesmo grupo e, em simultâneo, a heterogeneidade entre grupos distintos, baseando-se na dependência entre as observações nos diferentes níveis de agregação Laros and Marciano [2013]. Esta estatística toma valores entre 0 e 1 em que:

- **se o seu valor é próximo de zero**, não existe estrutura nos dados;
- **se o seu valor é próximo de um**, pode-se inferir que existe uma estrutura de agrupamento forte.

Conclui-se então que quanto maior a correlação entre os indivíduos, maior é a inadequação do modelo de regressão usual. O descrito traduz uma maior dependência entre indivíduos do mesmo grupo e, por conseguinte, uma maior necessidade de um método de regressão que respeite a estrutura de agregação dos dados.

A cada um destes casos adequam-se, por isso, diferentes modelos de regressão. No primeiro, a regressão linear tradicional e para o segundo, a regressão multinível é uma alternativa apropriada. Atendendo a que a modelagem deve respeitar a estrutura dos dados (Goldstein, 1995) Goldstein [2011], este método incorpora efeitos aleatórios que permitem considerar o contexto dos dados e respeitar a sua estrutura hierárquica inerente, melhorando assim as estimativas.

Os modelos de análise multinível têm também atribuídas outras designações, tal como "modelos hierárquicos"por Bryk e Raudenbush (1992) Bryk and Raudenbush [1992] ou "modelos de coeficientes aleatórios"por Longford (1993) Aitkin and Longford [1986]. São ainda conhecidos como modelo hierárquico linear ou modelo de regressão hierárquica, uma vez que, são considerados uma extensão dos modelos de regressão usual para a análise de variáveis dispostas em vários níveis de agregação.

Vistos como uma alternativa adequada ao estudo de diversas estruturas de dados, graças à sua versatilidade e qualidade das estimações, foram conduzidas investigações teóricas sobre esta temática que são abordadas no capítulo seguinte.

## 2.2 MODELOS MULTINÍVEL

Os modelos de análise multinível são descritos como modelos em que o tipo de análise de regressão considera numa só estrutura dados organizados em diferentes níveis de agregação. Essa organização dos dados, designada por estrutura hierárquica, é descrita pela conglomeração de unidades segundo determinadas características que as diferenciam, atribuindo-as a diferentes grupos de nível mais baixo. Estes, por sua vez, pertencem a unidades de um nível mais alto, considerando os seus aspetos em comum, e assim sucessivamente.

Estes modelos são vistos como uma evolução do modelo de regressão usual em que as variáveis são organizadas e analisada em múltiplos níveis, tornando assim mais correta a estimação dos valores para erros padrão, intervalos de confiança e testes de hipóteses Laros and Marciano [2013]. Dado que têm em conta os contextos em que os indivíduos estão inseridos e os aspetos acima descritos, foram reconhecidos como um dos métodos mais interessantes nas análises quantitativas que nos últimos anos foram desenvolvidas.

Estes modelos permitem ao investigador verificar se um hipotético modelo explicativo se adequa ao estudo a desenvolver, pois são vistos como uma resposta à necessidade premente de analisar a relação entre as unidades experimentais e o meio que as rodeia. São indicados para estudo de fenómenos cuja compreensão está dependente quer das características dos elementos em estudo, quer da sua própria organização. Ao nível da sociedade, podem ser basilares, uma vez que nos permitem separar o papel de cada uma das características de um contexto, permitindo assim melhorar o conhecimento da realidade e intervir de forma mais eficiente em problemas sociais.

### Anotações Históricas

Os estudos com análise de estruturas hierárquicas de dados surgem inicialmente na área da Educação e das Ciências Sociais, em que se destacam autores como Goldstein Goldstein [2011], Lewis e Cullingford Torrecilla and Javier [2012] Gray et al. [1995]. A sua utilização está precisamente relacionada com um estudo conduzido nos anos 70, por Bennett, com crianças do ensino básico, em Inglaterra Hunt [1976].

Nos anos 80, a literatura da metodologia multinível, fez convencer de que a análise de um único nível de efeitos contextuais corresponderia a uma aproximação concetualmente fraca para avaliar e estimar o papel do contexto social e geográfico na modelação de atitudes individuais e comportamentos Gadd et al. [2011]. Como, anos mais tarde, surgiria em documentação de Bryk e Raudenbush Bryk and Raudenbush [1992], as observações em unidades socialmente ou geograficamente definidas são suscetíveis de ter mais coisas em comum entre elas do que com observações de outra unidade. Daí que, o pressuposto da independência entre os erros assumida na regressão usual para níveis individuais é violada sempre que os dados apresentem agrupamentos significativos.

Na verdade, se considerarmos com alguma atenção e cuidado a estrutura de dados recolhidos no seio de uma qualquer pesquisa, será fácil identificar padrões hierárquicos ou multiníveis, que Heck e Thomas Heck and Thomas [2015] denominam de estrutura organizacional da informação. A não consideração desta estrutura hierárquica conduz a uma leitura demasiado parcial e truncada da informação disponível proporcionando uma visão distorcida e fragmentada do que se pretendia obter uma resposta mais clara e

abrangente Maia et al. [2005].

Entre os vários trabalhos de Aitkin e Longford Aitkin and Longford [1986] consta um artigo de 1986 vastamente referenciado como tendo revolucionado a investigação educativa. Este demonstra que o uso da regressão linear estaria condicionada a conjuntos de unidades experimentais independentes, o que viria a restringir a sua aplicação e daria destaque aos modelos multinível.

No início dos anos 90, a regressão multinível emergiu nas ciências sociais e humanas, voltada para as mais diversas áreas de dados, em vulgo com respostas discretas, com aplicações na epidemiologia, criminologia, crescimento animal e vegetal, gestão organizacional e muitas outras.

### Da Regressão Clássica à Regressão Multinível

A regressão clássica e os seus pressupostos vieram levantar muitas dúvidas quanto à sua aplicação e à validade das suas predições. Ignorar o real não cumprimento dos pressupostos por ela exigidos é um problema que pode ser inflacionado pelos erros associados à estimação dos coeficientes da regressão. Nomeadamente, na sua aceitação nos testes de hipóteses inerentes à construção do modelo, onde se verifica se a adição de determinada variável explicativa ao modelo contribui ou não para o melhor conhecimento da variável resposta.

Os modelos multinível, isentos dos pressupostos necessários à aplicação dos modelos lineares, oferecem estimativas mais robustas dos erros, considerando a não-independência das observações. Estes têm por base o contexto dos dados na sua estrutura o que permite variar as formas como se desenvolvem hipóteses de contextualização dos efeitos.

A regressão multinível aponta ser solução para três problemas Quené and Van den Bergh [2004]:

- respeitar a não homocedasticidade dos dados
- adaptar-se à análise de dados de amostras com estrutura multinível
- ser robusto contra a ausência de dados, isto é, à composição não-equilibrada da amostra.

Os modelos multinível surgem assim como uma alternativa aos modelos lineares na análise de dados com contexto, em que uma estrutura hierárquica lhe é inerente. O uso de métodos inadequados na análise dos dados, que por sinal, conduzam à desagregação da informação, ignorando o seu padrão natural, levam à perda de informação. Tudo porque a leitura feita aos dados é limitativa e há um nível de informação do qual não se usufruiu. Inferir sobre dados com esta estrutura, de um modo pouco claro, confundindo aspetos de nível micro e macro, descarta questões como:

- A **heterogeneidade entre as retas de regressão** dos diferentes grupos, reduzindo a qualidade de ajuste do modelo.
- A **dependência entre as observações** cujos indivíduos pertencem ao mesmo contexto e que, por isso, tendem partilhar caraterísticas semelhantes. De facto, dificilmente unidades pertencentes ao mesmo nível e à mesma unidade de nível superior são independentes e o uso de modelos de dependência linear pressupõe a independência entre as observações.
- A **agregação ou contexto**, na medida em que, ou os dados são analisados a nível de grupos e não a nível individual ou são considerados a nível individual e não se tem em consideração a variação intergrupo, visto não existir a análise diferenciada por níveis. O estudo dos dados desagregados

não permite averiguar a variação intra e intergrupo, advindo deste facto a possibilidade de erros de "efeitos de delineamento", designação dada por Kish (1965) Kish [1965] às estimativas erradas dos erros padrão.

### Génese da Regressão Multinível

Segundo Raudenbush e Bryk Bryk and Raudenbush [1992], muitas das estruturas naturais envolvem hierarquias e, nesses casos, pode haver algo a ganhar em não descartar a análise dessas estruturas. Pela sua própria natureza, a abordagem multinível impele-nos a considerar a possibilidade de que o que conhecemos pode ser diferente entre grupos, comunidades ou períodos temporais. Os modelos multinível são apropriados a dados nessas condições. A dados que têm algum tipo de estrutura hierárquica ou encaixada na sua génese, permitindo chegar à análise de outros aspetos de questões sob investigação. Isto porque têm potencial para gerar novo conhecimento, visto que desafia as assunções, estimula um pensamento adicional que contempla o sentido com que o estudo é conduzido, atendendo aos vários estratos e contextos complexos inerentes aos dados Gadd et al. [2011].

Os modelos multinível conduzem a uma análise estatística mais rigorosa. Permitem uma visão mais consciente das relações hipotéticas entre as variáveis e têm provado ser úteis num amplo horizonte de problemas em que o estudo dos contextos temporal, espacial e relacional são vistos como aspetos fundamentais.

Estes aspetos fundamentam o facto de, ao longo das últimas décadas, a análise multinível ser aplicada nos mais diversos campos do conhecimento. Devido ao reconhecimento crescente das suas propriedades, passou a ser uma referência corrente nos estudos de génese organizacional.

Nestes moldes, os modelos hierárquicos são reconhecidos pela sua abrangente e eficiente aplicação e, consequentemente, obtenção de resultados, realçando-se diversas características vantajosas no seu uso, algumas delas já referidas ao longo desta secção e que são descritas na secção seguinte.

### Vantagens

As análises estatísticas que consideram a estrutura de agrupamento dos dados têm à partida algumas vantagens Soares [2013]:

- apresentam modelos mais flexíveis e estruturados que utilizam um maior nível de informação dos dados da amostra e, ainda permitem, estabelecer uma equação para cada unidade de agrupamento, o que permite análises individuais para cada grupo
- possibilitam a formulação e o teste de hipóteses relativas a efeitos entre os níveis a partir do uso da informação de cada agrupamento dos dados
- graças ao alcance da análise dos dados nestes modelos é possível a partição da variabilidade da variável resposta entre os diferentes níveis, de acordo com a proporção explicada por cada um deles.

Os modelos de regressão multinível têm por objetivo descrever, através de um modelo matemático, a relação entre variáveis explicativas e independentes tendo em conta vários níveis de agregação. Os aspetos base supradescritos desdobram-se em vantagens técnicas e de análise que ao longo deste capítulo são referenciados.

As qualidades estatísticas deste tipo de análise proporciona a obtenção de estimativas dos coeficientes da regressão e dos valores da sua variação, como erros padrão e intervalos de confiança, mais corretos.

O facto de considerar a existência de correlação intraclasse permite imediatamente estimativas mais conservadoras. Apresenta uma enorme flexibilidade na modelação da estrutura de variância dos dados em função das covariáveis, permitindo analisar dados cuja variância não é homogénea. Possibilita a análise, com detalhe, das caraterísticas dentro de cada nível, da correlação entre as variáveis, das dissemelhanças entre níveis, do comportamento da variância e do encaixe ou cruzamento entre níveis.

Torna-se num modelo muito mais isento de erros devido à presença dos efeitos aleatórios considerados para cada nível, que são os responsáveis pelos coeficientes aleatórios que descrevem a variabilidade entre as unidades de agregação. Os parâmetros aleatórios destes modelos variam de forma simples, no caso da variabilidade da constante, de forma complexa nos coeficientes de regressão ou até de ambos os modos, tornando as unidades experimentais de cada grupo como amostras aleatórias de uma população. Este aspeto permite contornar os riscos associados à amostragem, que é uma das limitações referidas na secção 2.2.7.

Outra especificidade benéfica destes modelos é a possibilidade de modelar efeitos que contemplam a classificação cruzada, referenciada por Raudenbush e Bryk (2002) Raudenbush and Bryk [2002] e Snijders e Bosker (1999) Snijders and Bosker [1999]. A versatilidade desta valência permite a esta metodologia adaptar-se a casos mais complicados expectando melhores estimativas dos efeitos das variáveis explicativas, identificar com maior precisão as componentes da variância e realizar um estudo diferencial dos efeitos.

Constata-se que a análise multinível se revelou muito útil para a investigação de interações entre variáveis de diferentes níveis, permitindo compreender melhor a decomposição da variância do erro nos vários níveis. Uma vez que cada unidade individual difere da média total num determinado valor, que se designa por erro aleatório, e esses erros são independentes, isto permite determinar as componentes da variância. O papel do contexto, nível macro, na análise de dados foi aí reconhecido para a compreensão dos comportamentos do nível micro. Surge assim a resposta à necessidade de analisar a relação entre os indivíduos e o meio que os rodeia, permitindo uma intervenção mais eficiente, através de um melhor conhecimento da realidade. Partindo do facto de que não se assume erroneamente o pressuposto da independência entre as observações, como nas análises usuais, este método permite, à partida, estimativas mais fidedignas.

A sua versatilidade de estudo permite que a resposta possa ser avaliada ao nível individual, populacional ou comunitário. A análise da variabilidade relativa entre os níveis, unidades experimentais, grupos ou outros conglomerados existentes que estes modelos proporcionam, permitem saber onde se aloja a estocasticidade do sistema. Conduz-nos assim a melhores estimativas dos efeitos fixos que são designados por coeficientes de regressão no sistema linear e que, por sinal, têm estimativas muito semelhantes. São estimados a nível individual, permitindo predições mais informadas e ajustadas à realidade devido aos efeitos aleatórios presentes no modelo.

Ao possibilitar que a variabilidade da resposta seja explicada pelas covariáveis incluídas nos diferentes níveis e quantificada de forma a que a proporção da variabilidade explicada em cada nível possa ser diretamente comparada. Permitindo um melhor conhecimento dos dados, e como tem em atenção a sua estrutura, é salvaguardado o caso da introdução de erros nos estudos.

A abordagem multinível ajusta-se às necessidades inerentes ao desenho amostral, uma vez que pode existir correlação entre os níveis da hierarquia, a qual, através da análise simples não é detetada. Ao permitir o estudo de covariáveis no nível mais baixo, pode averiguar influência dessas variáveis nas observações e com a inclusão dessas noutros níveis superiores, permite apurar a influência destas nos grupos. Ao basear-se na procura dos pontos que caraterizam cada nível, torna-se fácil entender as variáveis e a forma como fazem variar a resposta.

## Áreas de Aplicação e Fundamentos

Segundo Bryk e Raudenbush Bryk and Raudenbush [1992], os modelos lineares multinível não são apontados como a solução para todos os problemas, mas foram considerados um grande avanço para a obtenção de resultados estatisticamente melhores e mais concretos, pois o desperdício de informação é evitado ao máximo. Na verdade, a maioria dos dados recolhidos de forma empírica estão sujeitos a um determinado contexto e, por conseguinte, têm um padrão hierárquico inerente. Ao não ter em conta essa estrutura, é de esperar que um forte enviesamento seja imputado aos resultados.

Visto como um dos métodos mais interessantes na investigação quantitativa, é referido como uma extensão dos modelos tradicionais, considerando as variáveis dispostas em diferentes níveis de agregação, respeitando assim a sua natureza.

Os modelos hierárquicos, definidos como uma ferramenta flexível e potente para análise de grupos de dados, são usados na investigação da interação entre variáveis de diferentes níveis nas mais diversas áreas. Não descartando a relevância dos estudos pioneiros na área das Ciências Sociais onde esta metodologia foi aplicada, ao longo das últimas décadas a sua aplicação foi-se disseminando. A análise multinível, vulgar em estudos comportamentais e populacionais, foi também aplicada por Hox na metanálise Hox and Leeuw [2003]. A metanálise designa uma metodologia estatística em que os objetivos de análise são agrupados. Trata-se na prática de uma técnica que integra resultados de dois ou mais estudos sobre a mesma questão num único estudo, concretizando uma única resposta.

Hoje em dia, a modelação multinível está presente em estudos nas áreas:

- da **saúde**, como é o caso das doenças epidemiológicas ou outras como o cancro e o HIV, com tratamentos a serem desenvolvidos e testados
- da **demografia**, em estudos de crescimentos populacionais ou tendências de desenvolvimentos económico
- da **sociologia**, em gestão organizacional ou criminologia, para o estudo de comportamentos e atitudes a nível da comunidade e individual e a envolvência social desta realidade
- da **zoologia**, para a análise e previsão do crescimento de colónias e adaptação a diferentes meios
- da **educação**, seguindo o mote dos primórdios deste método, e que continua a ser muito aplicado para estudos de âmbito escolar, principalmente para a análise de dados com correlação temporal
- entre outras.

## Contrução do Modelo

O modelo multinível leva em consideração a estrutura de agrupamento dos dados. Isso reflete-se concretamente na especificação do modelo multinível. Nos modelos de regressão usual, simples ou múltipla, o valor da constante e os coeficientes que permitem o ajuste do modelo são parâmetros fixos, enquanto que, para o modelo multinível esses parâmetros são considerados aleatórios, pois encontram-se sob influência dos níveis hierárquicos superiores Soares [2013].

A estrutura fundamental dos modelos multinível em estudos de natureza hierárquica ou contextual implicam a especificação de, pelo menos, duas equações, uma por cada nível em estudo. Uma relativa ao nível micro e outra para o nível macro. A equação ao nível micro modela a relação entre as diferentes

caraterísticas das unidades experimentais, como variáveis explicativas interindividuais ou variáveis que ajudam a discretizar as diferenças entre as observações. A equação de nível macro tem um valor constante que é transversal a todos os grupos desse nível e o declive varia entre eles. Esse declive define os coeficientes de regressão, que não são mais do que variáveis aleatórias que seguem determinada distribuição. Como a modelação das unidades de nível superior, implica a modelação das unidades de nível inferior, o conteúdo dessas equações perfaz uma equação geral do modelo, em que ambos os níveis são considerados.

Assuma-se dois níveis de agregação dos dados de uma amostra e a existência de *J* unidades de nível 2 em que *j* = 1, ..., *J* e *I* unidades experimentais de nível 1 em cada uma delas em que $i = 1, ..., I$. A equação de nível micro é descrita por:

$$Y_{ij} = \beta_{0j} + \sum \beta_{qj} X_{qij} + \epsilon_{ij} \tag{2.8}$$

em que:
*i* é o índice das observações
*j* é o índice do grupo
*q* é o índice da variável explicativa de nível micro
$\beta_{0j}$ representa a média da variável resposta da amostra no grupo j
$\beta_{qj}$ representa a média da variável resposta no *j*-ésimo grupo
$X_{qij}$ representa o valor da *q*-ésima variável de nível micro para o indivíduo *i* do grupo *j* em que $q = 1, ..., Q$
$\epsilon_{ij}$ é o erro no nível micro, que contempla a variabilidade entre os indivíduos do mesmo grupo.

A equação de nível macro tem em considerando as diferenças entre os grupos descreve a variação do valor médio da resposta entre os grupos:

$$\beta_{qj} = b_{q0} + \sum b_{qj} X_{sj} + r_{qj} \tag{2.9}$$

em que:
*j* é o índice do grupo
*s* é o índice da variável explicativa de nível macro
$b_{q0}$ representa a média da *q*-ésima variável de nível micro na amostra amostra
$b_{qj}$ representa a média da *q*-ésima variável de nível micro no *j*-ésimo grupo
$X_{sj}$ representa o valor da *s*-ésima variável de nível macro no grupo j em que $s = 1, ..., S$

Os modelos hierárquicos incorporam naturalmente a estrutura de agrupamento da população em estudo Fernandes [2005]. Formalmente, os estudos de natureza hierárquica, também designados por análise contextual ou multinível, implicam a já referida especificação de uma equação para cada nível em estudo. Reconhecendo a sua complementaridade perfazem uma só equação geral para o modelo Hox [1998]. Temos que tanto a constante como os coeficientes que são responsáveis pelo declive das retas de regressão, podem ser considerados variáveis aleatórias que variam entre os grupos.

Suponhamos então que $Y_{ij}$ é uma variável resposta para um qualquer registo *i* de um sujeito *j*. A resposta pode ser descrita como o desvio da média $\beta_{0j}$ do sujeito *j*, isto é:

$$Y_{ij} = \beta_{0j} + \mathrm{e}_{ij}, \quad i = 1, ..., I, \quad j = 1, ..., J \tag{2.10}$$

Assumindo que os resíduos $e_{ij}$ têm distribuição normal, com média zero e $\sigma^2_{e_{ij}}$, podemos também considerar que a resposta média $\beta_{0j}$ para o sujeito *j* como um desvio da grande média $\gamma_{00}$. Isto é:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2.11}$$

em que consideramos a média de um sujeito em função da grande média mais um erro aleatório.

Assim como os resíduos $\epsilon_{ij}$ no nível macro (das observações), os resíduos $u_{0j}$ no nível micro (dos sujeitos), assumem-se como normalmente distribuidos, com média zero e variância e $\sigma^2_u$. Considera-se ainda que os resíduos de nível superior $u_{0j}$ são não correlacionados com os resíduos do nível inferior $e_{ij}$, ou seja, que as variâncias se comportam de forma independente.

Substituindo eq. 2.12 na eq. 2.11, temos o modelo multinível base, também designado por modelo nulo, uma vez que não contempla variáveis explicativas:

$$Y_{ij} = \gamma_{00} + (\mathrm{u}_{0j} + e_{ij}), \quad i = 1, ..., I, \quad j = 1, ..., J \tag{2.12}$$

em que *i* designa as unidades de nível 1, as observações, e *j* designa as unidades de nível 2, os sujeitos Barbosa and Fernandes [2013].

Este modelo é então composto por duas partes: a parte fixa e a parte aleatória. A parte fixa passa por modela o efeito da média total sobre todas as observações. A parte aleatória decompõe a variância total de *Y* pelos níveis, neste caso, em variância entre sujeitos $\sigma^2_{u_{ij}}$ e variância entre as observações de cada sujeito $\sigma^2_{e_{ij}}$ Fernandes [2005].

Com a inclusão de variáveis explicativas, o modelo nulo pode ser facilmente extendido, basta considerarmos uma variável e começarmos por inclui-la na parte fixa. Consideramos uma variável explicativa *X* de nível 2, tomamos o modelo desse nível e obtemos:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,\mathrm{X} + \mathrm{e}_{ij} \quad i = 1, ..., I, \quad j = 1, ..., J \tag{2.13}$$

considerando apenas a parte fixa $\beta_{1j} = \gamma_{01}$, a equação de nível 1 fica:

$$Y_{ij} = \gamma_{00} + \gamma_{01} + (u_{0j} + e_{ij}), i = 1, ..., I, j = 1, ..., J \tag{2.14}$$

Este modelo, inclui agora os efeitos da variável *X* apenas na parte fixa. Assim, o efeito da variável

adicionada é o mesmo para todas unidades experimentais, independentemente do valor que tenham de $X$. Com a assunção desta afirmação o modelo poderia ser considerado inválido. Por isso, ele tem de ser melhorado para que a variância seja explicada também pela variável explicativa, tal como a grande média já é. Considerando agora a variável adicionada em ambas as parte do modelo temos que:

$$\begin{aligned} Y_{ij} &= \gamma_{00}+\gamma_{01}X+(u_{0j}+u_{1j}+e_{ij}) \\ &= (\gamma_{00}+u_{0j})+(\gamma_{01}+u_{1j})X+e_{ij}, \quad i=1,...,I, \quad j=1,...,J \end{aligned} \tag{2.15}$$

Para este modelo, já se pode considerar que a variância depende da variável explicativa e que certamente será uma melhor aproximação da variável resposta. Dado que mais nos aproximamos do contexto do problema e assim melhor o conseguimos explicar Hox [2010].

A estrutura dos dados é muito importante para os modelos multinível e a versatilidade destes permite a concretização de modelos mais complexos, com três ou mais níveis e considerando variáveis explicativas discretas e contínuas O'Connell and McCoach [2008].

### Limitações e Desenvolvimentos Futuros

O processo de obtenção de uma amostra da população é denominado de amostragem. Nos modelos multinível descreve-se como uma escolha aleatória entre as unidades de nível macro, seguida de uma seleção igualmente aleatória das unidades de nível micro. Daqui advém no imediato uma limitação: os riscos da amostragem. O processo de amostragem envolve riscos, pois toma-se decisões sobre toda a população com base em apenas uma parte dela. Mais acresce quando por motivos práticos, de âmbito financeiro ou logístico, a amostragem trata-se da recolha das unidades experimentais disponíveis, uma vez escolhidas as unidades de nível macro Laros and Marciano [2013].

Uma das barreiras à análise multinível que o desenvolvimento computacional tem conseguido colmatar é o problema da interpretação de dados com padrão hierárquico que residia na inexistência de *softwares* adequados e de fácil manuseamento.

Nos anos 80, surgem as primeiras possibilidades de modelação computorizada que permitiriam interrogar os dados e chegar a respostas que respeitassem um delineamento de estudo cuidado e criterioso, respeitando a metodologia da análise multinível.

Na área das Ciência da Educação destacou-se o Hierarchical Linear Models – HLM – da coautoria de Raudenbush e mais tarde, do trabalho de Goldstein Goldstein [2011], surge o Multilevel Modeling – MLn/MlwiN – mais interativo e versátil do que o anterior. Softwares estatísticos como R, SPSS e outros, desenvolveram implementações que permitiam o uso destes métodos de forma intuitiva e prática Pinheiro and Bates [2006].

No desenvolvimento prático deste estudo o software estatístico usado é o R e em específico são analisados dois pacotes de implementações, nomeadamente o nlme e o lme4.

A futura direção da modelação multinível é a sua extensão a situações em que existem estruturas mais complexas, como *cross-classified data*, com dados encaixados ou estruturas de erro, como erros auto regressivos. Contudo, a integração das técnicas que potenciam e flexibilizam a gestão dos dados em pacotes de análise, aumentaram as capacidades dos *softwares*, como anteriormente foi referido, permitindo um avanço na implementação destas metodologias.

## 2.3 MÉTODOS DE ESTIMAÇÃO

Um modelo estatístico é uma representação simplificada da realidade, sendo composto por equações que descrevem as relações entre determinadas quantidades aleatórias. Essas equações contêm um conjunto de parâmetros aleatórios, associados aos parâmetros fixos nos diferentes níveis, que são estimados para que o modelo se ajuste o melhor possível à realidade.

Um método de estimação é uma técnica estatística que, tendo em consideração os valores observados, permite estimar os valores para os parâmetros desconhecidos do modelo. De entre os métodos de estimação clássicos, o método de máxima verosimilhança é, sem dúvida, o mais comummente utilizado.

A estimação por máxima verossimilhança é dos métodos mais utilizados no campo da estatística. Requer, para o seu cálculo, um processo iterativo e tem por principal propriedade ser um estimador assintoticamente não enviesado. Contudo, existe um viés associado à estimação. A sua origem está no número de estimações realizadas no modelo e que, em alguns casos, pode ser importante para averiguar a sua qualidade, como é referido no próximo capítulo.

A precisão das estimativas reside na dimensão das amostras e depende do que se está a estimar. Se são parâmetros fixos e respetivos erros padrão ou parâmetros aleatórios e respetivos erros padrão. A estimação em cada destes casos visa uma robustez e um poder de ajuste aos dados bem diferente.

De acordo com os problemas em estudo, são vulgarmente usadas duas versões do método de estimação por máxima verossimilhança:

- **máxima verossimilhança** (ML) em que se assume a normalidade dos erros
- **máxima verossimilhança restrita** (REML).

Em qualquer uma das versões, as estimativas de parâmetros fixos nunca são enviesadas. Porém, na estimação dos parâmetros aleatórios do método de máxima verossimilhança provêm estimadores enviesados. Este enviesamento deve-se à perda de graus de liberdade da estimação que resulta da estimação dos parâmetros fixos e que o método de máxima verosimilhança não considera Júnior. Para colmatar este problema surge a máxima verossimilhança restrita que, além de ponderar o ajuste do número de graus de liberdade, é o mais indicado no estudo de dados não equilibrados.

O método de estimação de REML, proposto como solução para o problema das estimativas enviesadas do método de ML, maximiza separadamente as funções de verosimilhança dos efeitos fixos e aleatórios. Aquando da estimação dos efeitos aleatórios retira o viés introduzido pela perda de graus de liberdade que a estimação dos efeitos fixos introduz nos dados Torman [2011].

Esta técnica é amplamente utilizada para a estimação de componentes da variância, com referências em estudos de Patterson e Thompson datados de 1971 Patterson and Thompson [1971], oferece estimativas consistentes. Dado que se trata de um estimador assintoticamente normal, em que a sua distribuição em torno do valor real do parâmetro a estimar se aproxima de uma distribuição normal com variância na ordem de grandeza de $1/n$, em que $n$ designa o número de elementos da amostra Lehmann and Casella [1998]. Em contraposição à sua eficiência apresenta por vezes algumas limitações devido à convergência do seu algoritmo numérico para valores negativos.

## 2.4 CRITÉRIOS DE INFORMAÇÃO

A existência de vários modelos que se ajustam à ilustração do mesmo caso ou situação real torna inevitável a escolha do "melhor". Isto obriga à concretização de um critério de seleção para, de alguma forma, justificar a escolha.

A criação dos critérios de informação partiu dessa necessidade de se compararem modelos cuja estimação dos seus parâmetros é realizada através do logaritmo de máxima verossimilhança.

Essa seleção teria de se basear, claro está, em princípios científicos. Partindo do facto de que nenhum modelo é verdadeiro, isto é, nenhum é completamente fiel à realidade, a escolha passaria pela procura do que mais se ajuste aos dados e que, por consequência, cause a menor perda de informação.

O objetivo é encontrar o modelo que melhor explica o fenómeno em estudo, atendendo a que quantos mais parâmetros forem estimados mais erros têm de ser considerados. Os critérios de informação primam pela simplicidade e eficiência, o que implica que penalizam os modelos com muitos parâmetros.

Vulgarmente, são considerados dois critérios de informação, AIC e BIC que são posteriormente referenciados e descritos. Tratam-se de critérios muito semelhantes, apesar do BIC ser mais sensível ao número de parâmetros estimados pelo modelo, concretizando sobre esse facto uma maior penalização.

Em 1974, Akaike proporia um critério que usa a divergência de Kullback-Leibler Kul, para testar o ajuste do modelo à distribuição "real". A distância de Kullback-Leibler mede a similaridade entre o modelo estatístico e a distribuição "real" e é definida por:

$$I(g;f) = E_Y log\langle g(Y)/f(Y)\rangle = \int_{-\infty}^{\infty} log\langle g(y)/f(y)\rangle \mathrm{g(y)dy} \tag{2.16}$$

em que *g(y)* é a distribuição "real" e o modelo estatístico que aproxima *g(y)* é *f(y)*. As suas propriedades são:

1) I(g;f)$\geq 0$
2) I(g;f)=0 $\iff g(y) = f(y)$

Contudo, *g(y)* é desconhecida, pois depende do modelo "real", o que limita o seu uso.

Partindo de que o melhor modelo será o que apresenta as melhores estimativas dos parâmetros, o critério de seleção teria de passar pela comparação das estimativas. Esta comparação, por sua vez, implica a utilização da máxima verossimilhança como medida de ajustamento. O número de estimativas realizadas em cada modelo é igual ao número de parâmetros. Dado que cada estimação introduz um valor de viés na função de máxima verossimilhança, o cálculo da qualidade do modelo deve considerar uma penalização por cada parâmetro que é estimado.

Assim, também em 1974, Akaike propõe o Critério de Informação de Akaike – AIC – que é definido por:

$$AIC = -2log(L_{MV}) + 2p \tag{2.17}$$

em que $p$ é o número de parâmetros estimados no modelo.

Em 1978, Schwarz seguindo o pressuposto da existência de um modelo "verdadeiro", aquele que descreve corretamente a relação entre a variável resposta e as respetivas covariáveis, propõe o Critério Bayesiano (BIC), definido por:

$$BIC = -2log(L_{MV}) + 2p * log(N) \tag{2.18}$$

em que $p$ é o número de parâmetros estimados e $N$ o número de observações. Este trata-se do método mais adequado para comparar modelos não aninhados EMILIANO et al. [2010].

Os critérios de informação enquanto meio de seleção indicam a qualidade do ajustamento no valor inverso. Ou seja, o melhor modelo é o que apresenta menor valor do critério de informação. Posto isto, desta seleção deve resultar o modelo mais parcimonioso, isto é, que envolva o mínimo de parâmetros e que melhor explique a variável resposta.

Note-se que, se a estimação dos parâmetros usar a máxima verossimilhança restrita no cálculo dos critérios de informação, como é o caso dos modelos multinível, basta considerar o logaritmo da verossimilhança restrita, na vez do logaritmo da função de verossimilhança.

# Capítulo 3.

# Aplicação

## 3.1 ANÁLISE INTRODUTÓRIA

O universo das lojas Sonae MC do setor alimentar, sob a marca de referência Continente, empreende três tipologias de loja, designadas por insígnias. Nomeadamente, Continente Hipermercado, Continente Modelo e Continente Bom Dia. Cada uma destas tipologias está dividida em direções de operações, DOPs, que agrupam as lojas do mesmo tipo por regiões, tendo em consideração a localização e questões como a produtividade e número de colaboradores agregados a essas lojas. Assim sendo, a insígnia Continente Hipermercado tem DOPs Norte, Centro e Sul, a insígnia Continente Modelo tem Norte, Centro Norte, Centro, Centro Sul, Sul e Madeira e, por fim, a insígnia Continente Bom Dia possui Norte e Sul.

Este estudo principiou-se com a análise dos registos mensais do volume de sinistros ocorridos em cada direção de operações de cada uma das insígnias da MUC. O conjunto de dados incluía os valores obtidos entre janeiro e dezembro dos anos 2010 a 2014, recolhidos a partir dos registos de sinistralidade da equipa de Saúde e Segurança no Trabalho da Sonae Maia. A partilha e busca de informação em diversos ramos de trabalho da estrutura empresarial permitiu a análise de outras variáveis referentes ao mesmo período temporal que foram consideradas no estudo, na expetativa de fornecerem informações relevantes. Refiro-me então à produtividade, ao volume de vendas, ao número de transações, ao registos de stock, ao número de FTEs, às horas extra, à percentagem de absentismo e aos índices de sinistralidade, em particular, do índice de frequência e do índice de gravidade.

Cada variável dispõe de 60 registos mensais. Pretendo com este estudo averiguar a influência de variáveis não diretamente relacionadas com a sinistralidade sobre o volume de sinistros mensais.

Para a primeira abordagem foi aplicado um modelo linear com o objetivo de reconhecer as covariáveis potencialmente relacionadas com a variável resposta. Perante a insuficiência dos resultados deste estudo preliminar, foi aplicada uma modelagem multinível na procura de melhores resultados.

## 3.2 ANÁLISE EXPLORATÓRIA

A amostra tem um total de 1052 sinistros, que ocorreram entre janeiro e junho de 2014. Partindo da análise gráfica dos dados é possível verificar diferenças no volume de sinistros entre as três insígnias da MUC. O mesmo se deve ao diferente âmbito e volume comercial, dimensão, número de colaboradores, natureza das tarefas entre outros aspetos.

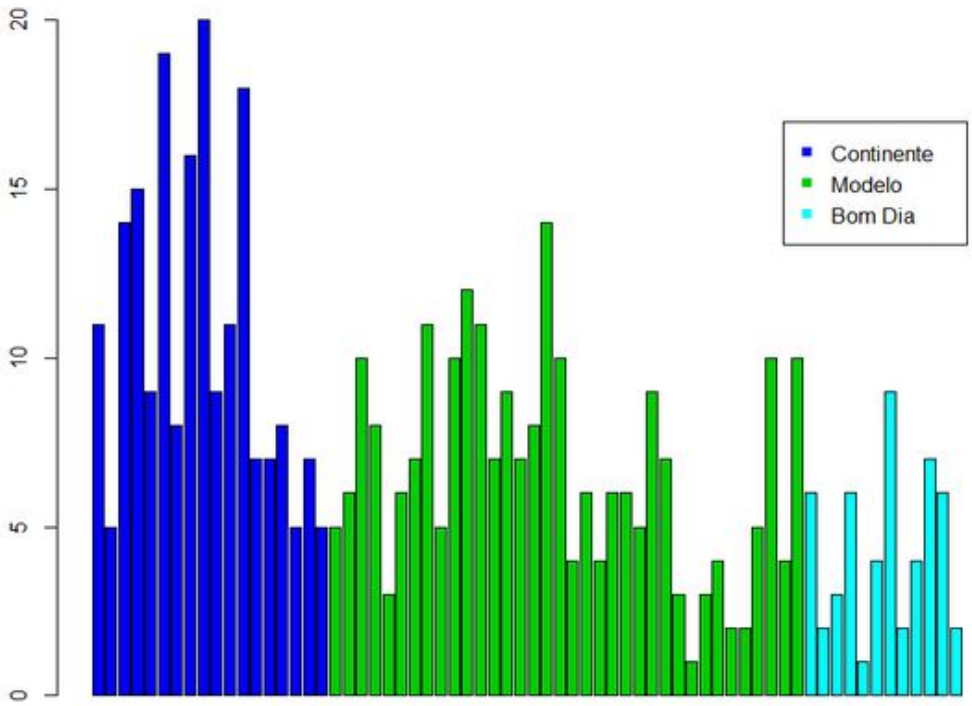


**Figura 3.1:** Número total de acidentes de trabalho registados mensalmente nas diversas DOPs entre Janeiro e Junho de 2014 nas insígnias Continente Hipermercado, Continente Modelo e Continente Bom Dia.

Além dessa diferenciação é também possível detetar dissemelhanças entre os dados das diversas direções de operações (DOP) de cada insígnia (INS). Existe uma clara tendência de um maior número de acidentes nas direções de operações localizadas a norte, principalmente em insígnias mais estáveis como Continente Hipermercado e Continente Modelo. Quando me refiro a estabilidade, tenho em consideração os valores de sinistralidade e as alterações a que a insígnia foi sujeita nos últimos anos. Atendendo à realidade da insígnia Continente Bom Dia, uma insígnia mais jovem, cujos colaboradores apresentam a menor média de idades e que não tem uma estrutura organizacional tão rígida devido ao seu âmbito comercial e ao seu crescimento acentuado a nível do número de superfícies nos últimos anos. Os factos referidos fazem com que os seus dados comerciais e de sinistralidade apresentem uma maior variação.

Com o intuito de uma análise preliminar, partimos para o ajuste de um modelo linear. O principal interesse é reconhecer variáveis explicativas e respetivas interações que se mostrem pertinentes na descrição da variável resposta, isto é, que justifiquem no contexto do problema o número de sinistros e a sua variação. Posto isto e atendendo ao já referido, tomei em consideração os dados entre janeiro e junho de 2014 relativos apenas a uma insígnia, nomeadamente a Continente Hipermercado, pela sua antiguidade, volume de lojas constantes, trabalhadores mais experientes e características e registos similares nos anos transatos. Nestes moldes, a estrutura padrão dos dados consta apenas de dois níveis: o nível micro em que constam todos os registos e o nível macro, no qual os registos são agrupados de acordo com a direção de operações a que pertencem.

Contudo, principiando com uma análise linear considero erroneamente a independência entre todos os registos, dispensando o contexto dos dados. Assim, a localização geográfica passa a ser tomada como uma caraterística do registo, enquanto variável discreta. Ajustamos primeiramente um modelo em que constam todas as variáveis disponibilizadas para este estudo: produtividade (PROD), volume de transações (TR), volume de vendas (VND), números de trabalhadores a horário completo (FTES) calculado pelo número

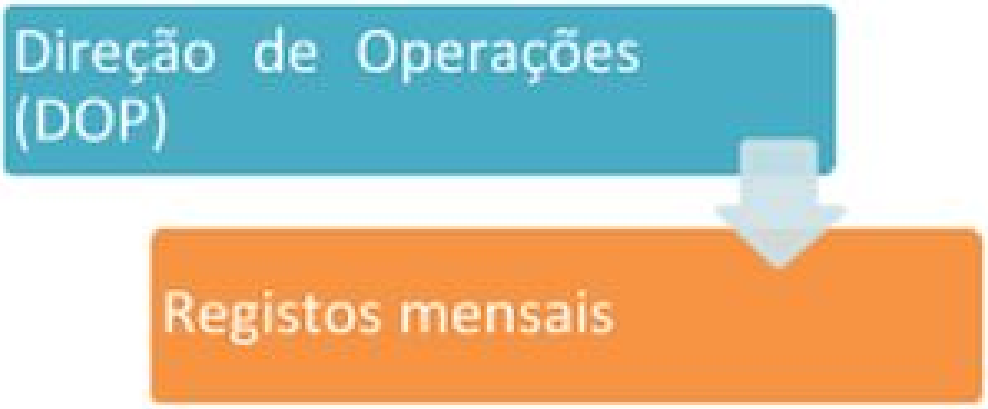


**Figura 3.2:** Estrutura padrão dos dados, considerando dois níveis: a referência geográfica de operações (DOP) e os registos mensais associados.

total de horas trabalhadas, índice de frequência (IF), índice de gravidade (IG), volume de stock em loja (STK), número de dias perdidos (DP) devido a ausência por acidente de trabalho, número de horas extra (HE), percentagem de absentismo (ABS) e a direção de operações a que cada observação está associada através de uma variável *dummy*, da qual considero duas classes de referência, a DOP norte e a DOP sul, respetivamente.

```
Modelo1 <- lm(AT~PROD+FTES+STK+HE+ABS+VND+TR+DP+IG+DOPS+DOPC, data=DD)
Modelo2 <- lm(AT~PROD+FTES+STK+HE+ABS+VND+TR+DP+IG+DOPN+DOPC, data=DD)
```

Através da aplicação do comando *step* aos modelos iniciais, que consta de um algoritmo que seleciona a conjugação que variáveis explicativas que concretizam o modelo com o menor valor do logaritmo da verosimilhança, obtenho os modelos a seguir.

```
Modelo1 <- lm(AT ~ PROD + STK + VND + TR + IG + DOPC ,data=DD)
Modelo2 <- lm(AT ~ PROD + FTES + STK + VND + TR + DP + IG + DOPN, data = DD)
```

Dos dois modelos acima faço nota das variáveis explicativas que neles se incluem: produtividade, dias perdidos, número de transações, volume de vendas, índice de gravidade, volume de stock e FTEs. Verificamos ainda que a variável *dummy* considerada é significativa em ambos o modelos apresentados, o que se pode argumentar a favor do agrupamento dos registos por DOP.

Relativamente às interações, em ambos os modelos iniciais foram testadas as diversas interações possíveis, revelando-se significativas as seguintes: STK:VND, VND:TR, DOPcentro:HE, PROD:VND com as duas classes de referência e DOPcentro:PROD e DOPnorte:STK com a DOP norte como classe de referência.

Identificadas as variáveis explicativas para a parte fixa do modelo, o principal objetivo desta análise preliminar, podemos ainda apontar, devido a significância da variável *dummy*, a variância que reside na localização geográfica das lojas, identificadas pelas direções de operações.

## 3.3 MÉTODOS DE ANÁLISE MULTINÍVEL

No software R existem dois pacotes estatísticos muito similares para modelos lineares mistos que são vulgarmente utilizados; nomeadamente, lme4 desenvolvido por Bates, Maechler e Bolker Bates [2010] e o nlme por Pinheiro e Bates. Estes integram técnicas estatísticas tratam-se de funções que consideram

os dados e produzem estimativas para os parâmetros desconhecidos. Têm como funções principais lmer e lme, respetivamente. Os algoritmos implementados nestas são usados nos cálculos e têm em conta os princípios estatísticos adequados ao modelo, otimizando a sua solução Lorenzo [2013].

```
LMER <- lmer(AT ~ 1 + (1|INS/DOP), DD)
LME<-lme( AT ~ 1, data = DD, random =~ 1 | INS/DOP )
```

Os testes de significância usados em ambos são o teste F e o teste t para avaliar a significância das estimativas para os termos fixos, ambos baseados na máxima verosimilhança restrita que é o método de estimação que está implementado por defeito nestes modelos, pois permite a estimativa condicional da variância. Contudo, o facto de o p-valor estar explícito através do sumário do modelo lme, torna este de análise mais imediata, para além de que apresenta uma precisão superior ao modelo lmer na estimação dos parâmetros. Especificações dos modelos e dados de significância explícitos pelas instruções nas figuras 3.3 e 3.4.

```
> summary(LMER)
Linear mixed model fit by REML
Formula: AT ~ 1 + (1 | INS/DOP)
   Data: DD
   AIC BIC logLik deviance REMLdev
 361.2 370 -176.6    356.2   353.2
Random effects:
 Groups   Name        Variance Std.Dev.
 DOP:INS  (Intercept) 4.8594   2.2044
 INS      (Intercept) 7.6696   2.7694
 Residual             9.6788   3.1111
Number of obs: 66, groups: DOP:INS, 11; INS, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)    7.331      1.802   4.067
```

**Figura 3.3:** Output do comando summary do modelo LMER obtido a partir do pacote estatístico lme4.

Observando os valores retornados acima, vem que os efeitos fixos são iguais considerando que os valores do output do sumário do modelo lme apresentam maior precisão. O mesmo acontece com a variabilidade quantificada em cada nível, apesar de no sumário do modelo lmer esta estar explícita, o valor calculado pelo quadrado do erro com o modelo lme trata-se de um valor semelhante e mais uma vez a diferença incide apenas na precisão Pinheiro and Bates [2006].

Os efeitos aleatórios designados como aninhados, que ilustram o encaixe entre os vários níveis da hierarquia são fácil e explicitamente implementados nos dois modelos. No modelo lme considerando:

**AT~1 | INS/DOP em que a DOP está subordinada à insígnia (INS)**

Já os efeitos aleatórios cruzados obrigam a outros argumentos como o pdBlocked e o pdIdent, que não constam no nosso estudo. No modelo lmer, os efeitos aleatórios aninhados são modelados por:

**AT~1 + (1 | INS/DOP) em que a DOP está subordinada à insígnia (INS)**

```
> summary(LME)
Linear mixed-effects model fit by REML
 Data: DD
       AIC      BIC    logLik
  361.2068 369.9043 -176.6034

Random effects:
 Formula: ~1 | INS
        (Intercept)
StdDev:    2.769384

 Formula: ~1 | DOP %in% INS
        (Intercept) Residual
StdDev:    2.204429 3.111071

Fixed effects: AT ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 7.330794  1.802569 55 4.066859   2e-04

Standardized within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.1317139 -0.6408297 -0.1404075  0.6252642  2.3683435

Number of Observations: 66
Number of Groups:
         INS DOP %in% INS
           3           11
```

**Figura 3.4:** Output do comando summary do modelo LME obtido a partir do pacote estatístico nlme.

E para os efeitos aleatórios cruzados é apenas necessário:

$$\mathbf{AT\tilde{}1 + (1 \mid INS) + (1 \mid DOP)}$$

Na representação gráfica dos efeitos aleatórios correspondem aos pacotes estatísticos diferentes funções, respetivamente nas figuras 3.5 e 3.6.

Os outputs dessas figuras vêm mais uma vez atestar a variabilidade do volume de sinistros entre as diversas DOPs e insígnias, onde os valores dos efeitos aleatórios atribuídos são mais distintos.

Para a representação gráfica dos resíduos os comandos são similares para os modelos lme e lmer e os resultados também o são, pois variam na precisão, o que não é visível graficamente. O último dos gráficos a seguir descritos o lmer não o incorpora na função.

Tal como seria de esperar, uma vez que valores apenas se diferenciarem apenas na precisão, mais elevada através do pacote estatístico nlme, os gráficos obtidos para cada modelo são "aparentemente" iguais. Contudo, o lme4 não permite a representação gráfica do modelo lmer, que nos permitiria analisar tendências mais intuitivamente.

O modelo lmer não permite incorporar facilmente estruturas de correlação, enquanto o modelo lme tem argumento para essa inclusão e ainda este último permite modelar a heteroscedasticidade usando varFunc.

Atendendo ao acima descrito optei pelo pacote estatístico nlme, devido às suas maiores funcionalidades, por ser mais intuitivo e mais preciso.

```
> ranef(LMER)
$`DOP:INS`
                (Intercept)
CENTRO1:CNT       2.7371759
CENTRO2:MDL       1.8154312
CN:MDL            1.9405598
CS:MDL           -1.1876560
MADEIRA:MDL      -0.9373987
NORTE1:CNT        1.6110182
NORTE2:MDL       -0.3117556
NORTE3:BD        -1.1684390
SUL1:CNT         -2.6433553
SUL2:MDL         -1.6881705
SUL3:BD          -0.1674100

$INS
    (Intercept)
BD   -2.1081032
CNT   2.6904059
MDL  -0.5823027
```

**Figura 3.5:** Valores de output do comando ranef para os efeitos do modelo LMER

```
> ranef(LME)
Level: INS
    (Intercept)
BD   -2.1079590
CNT   2.6901912
MDL  -0.5822322

Level: DOP %in% INS
             (Intercept)
BD/NORTE3     -1.1683361
BD/SUL3       -0.1672997
CNT/CENTRO1    2.7370946
CNT/NORTE1     1.6109287
CNT/SUL1      -2.6434759
MDL/CENTRO2    1.8154580
MDL/CN         1.9405875
MDL/CS        -1.1876512
MDL/MADEIRA   -0.9373921
MDL/NORTE2    -0.3117443
MDL/SUL2      -1.6881694
```

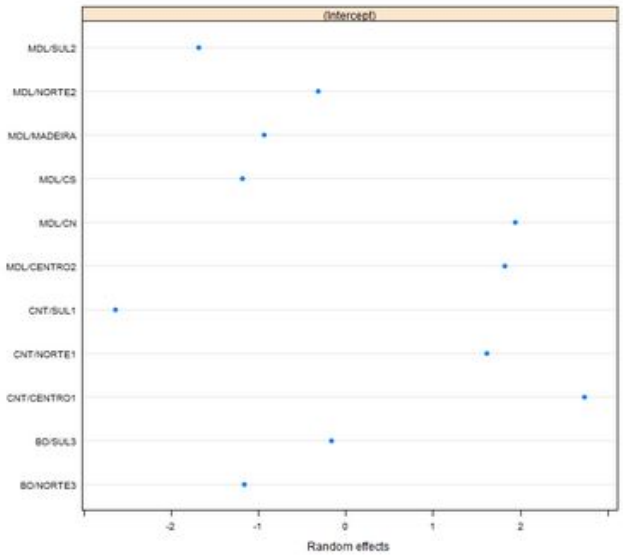


**Figura 3.6:** Valores de output do comando ranef para os efeitos do modelo LME e concretização gráfica dos efeitos aleatórios por DOP através do comando plot(ranef(LME))

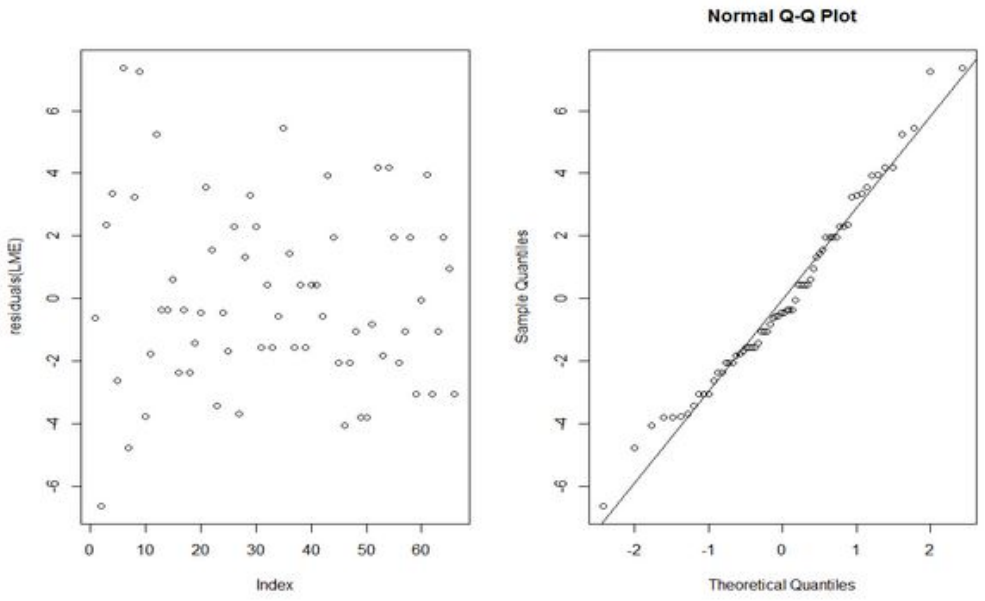


**Figura 3.7:** Outputs dos seguintes comandos aplicados ao modelo LME: (a) plot(residuals(LME)); (b) qqnorm(residuals(LME)) e qqline(residuals(LME))

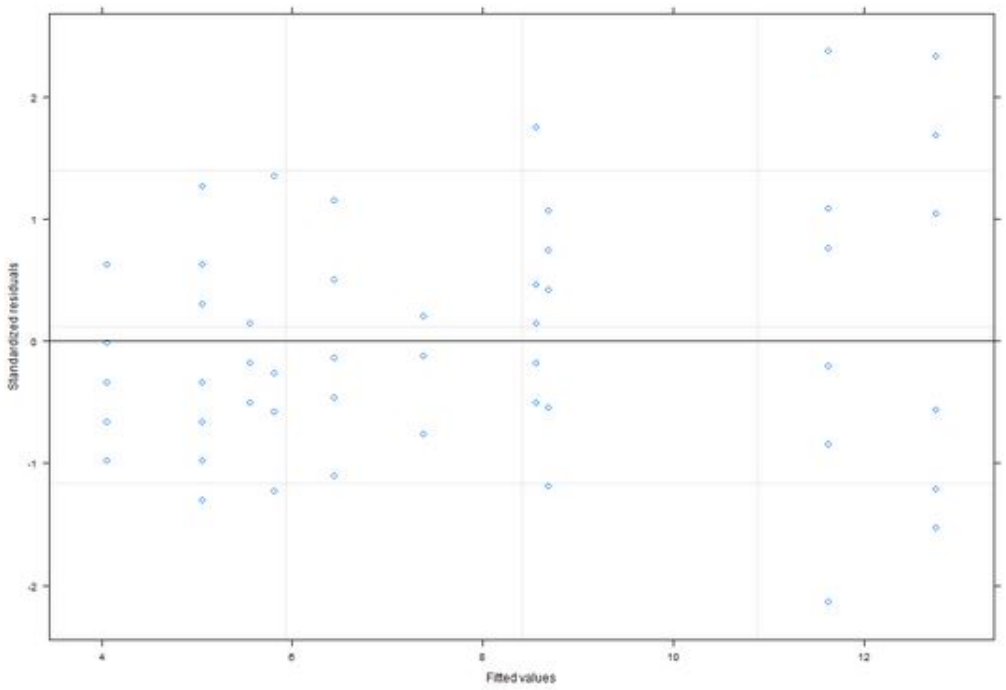


**Figura 3.8:** Outputs do comando plot(LME)

## 3.4 ANÁLISE MULTINÍVEL

A estrutura fundamental da análise multinível reside num estudo de natureza hierárquica. Considerando agora de uma forma mais generalizada o conjunto dos 1052 registos entre janeiro e junho de 2014 das três insígnias, surge um novo padrão hierárquico a que acrescemos um nível (figura 3.9).


**Figura 3.9:** Padrão hierárquico dos dados, considerando três níveis: a tipologia de loja (INS), a referência geográfica de operações (DOP) e os registos mensais associados

Agora considero a variabilidade a nível individual explicada pelos preditores de nível micro anteriormente identificados e pelas características dos níveis superiores. A direção de operações informa-nos sobre a referência geográfica do sinistro e a insígnia a que cada registo pertence identifica a tipologia de loja. Esta estrutura vai de acordo com o representado na figura 3.10.

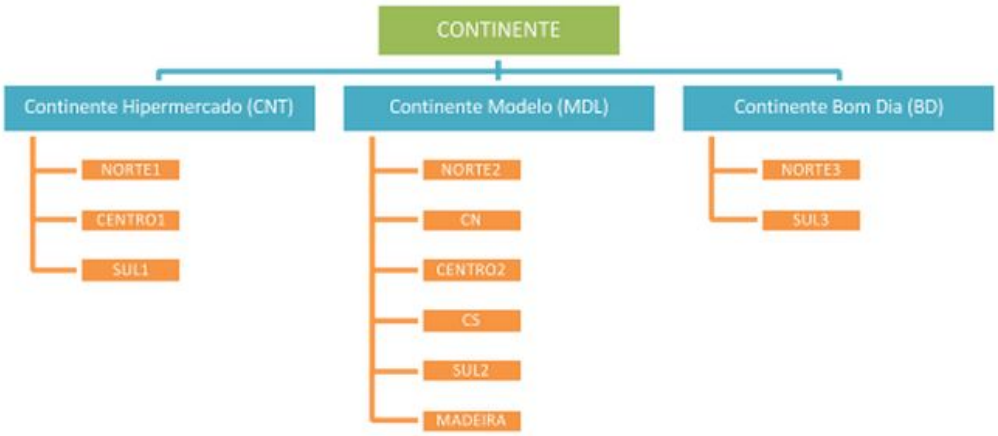


**Figura 3.10:** Organograma simplificado da MUC do retalho alimentar Sonae

Ajustando um modelo a cada grupo através do comando lmList, poderemos averiguar se os efeitos aleatórios se justificam, isto é, se existe efetivamente variabilidade entre os registos das diferentes insígnias e das diferentes direções de operações. Nas figuras 3.11, 3.12 e 3.13 são apresentados os coeficientes estimados para as retas de regressão associadas a cada grupo, tanto para os do nível 2 ( 3.11), as direções

de operações, como para os do nível 3 ( 3.12), entre as insígnias, verificando assim a sua variabilidade.

```
> d1<-lm(AT ~ 1, data=DD)
> summary(d1)

Call:
lm(formula = AT ~ 1, data = DD)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3636 -3.1136 -0.3636  2.3864 12.6364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3636     0.5239   14.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.256 on 65 degrees of freedom

> coef(d1)
(Intercept)
   7.363636
```

**Figura 3.11:** Nível 1

Atendendo aos valores apresentados para o nível 2, podemos realçar a discrepância entre os valores próximos das DOPs Norte e Centro (NORTE1 E CENTRO1) relativamente à DOP Sul (SUL1) da insígnia Continente Hipermercado. Na insígnia Continente Modelo destacam-se das restantes as DOPs Centro Norte (CN) e Centro (CENTRO2). Quanto à insígnia Continente Bom Dia, a amplitude entre os coeficientes é menor, revelando menor diferenciação entre as DOPs.

No que se refere aos coeficientes do terceiro nível, a disparidade entre os valores não é extremista, mas balanceia bastante bem as diferenças entre as três insígnias. Analisando o valor do erro verificamos que a maior instabilidade relativamente a um número médio de sinistros ocorre na insígnia Continente Bom Dia. Este facto justifica-se em aspetos como a baixa média de idades e antiguidade dos colaboradores, visto que se trata da insígnia mais recente e com maior crescimento nos últimos anos, com emprega mais trabalhadores temporários, entre outras.

Partindo do modelo constituído, considerando o contexto fornecido pela análise preliminar, as variáveis explicativas indicadas na análise exploratória e os efeitos aleatórios indicados pela estrutura hierárquica dos dados, procedemos à modelação da sua estrutura. Através dos critérios de informação e da significância das estimativas dos parâmetros é revista a seleção dos efeitos fixos e aleatórios para a constituição do modelo que melhor se ajusta ao problema.

$$\text{m1<-lme(AT}\tilde{}\text{1, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m2<-lme(AT}\tilde{}\text{PROD, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m3<-lme(AT}\tilde{}\text{PROD+IG, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m4<-lme(AT}\tilde{}\text{PROD+IG+TR, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m5<-lme(AT}\tilde{}\text{PROD+IG+TR+VND, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m6<-lme(AT}\tilde{}\text{PROD+IG+TR+VND+STK, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$
$$\text{m7<-lme(AT}\tilde{}\text{PROD+IG+TR+VND+STK+FTES, data = DD, random =}\tilde{}\text{ 1 | INS/DOP)}$$

```
> d2<-lmList(AT ~ 1 | DOP, data=DD)
> summary(d2)
Call:
  Model: AT ~ 1 | NULL
   Data: DD

Coefficients:
   (Intercept)
         Estimate Std. Error   t value     Pr(>|t|)
CENTRO1 13.666667   1.270091 10.760383 3.774758e-15
CENTRO2  9.166667   1.270091  7.217330 1.661355e-09
CN       9.333333   1.270091  7.348554 1.012639e-09
CS       5.166667   1.270091  4.067950 1.526630e-04
MADEIRA  5.500000   1.270091  4.330398 6.356216e-05
NORTE1  12.166667   1.270091  9.579366 2.542411e-13
NORTE2   6.333333   1.270091  4.986519 6.497878e-06
NORTE3   3.666667   1.270091  2.886932 5.550196e-03
SUL1     6.500000   1.270091  5.117743 4.065780e-06
SUL2     4.500000   1.270091  3.543053 8.151202e-04
SUL3     5.000000   1.270091  3.936726 2.344457e-04

Residual standard error: 3.111075 on 55 degrees of freedom

> coef(d2)
         (Intercept)
CENTRO1   13.666667
CENTRO2    9.166667
CN         9.333333
CS         5.166667
MADEIRA    5.500000
NORTE1    12.166667
NORTE2     6.333333
NORTE3     3.666667
SUL1       6.500000
SUL2       4.500000
SUL3       5.000000
```

**Figura 3.12:** Nível 2

Verificamos que considerando o modelo m1, no qual não foram incluídas variáveis explicativas, a variância calculada para o nível 3 é de 7,67 e para o nível 2 é de 4,86. Quando aos critérios de informação o valor do AIC é de 361,2068 e do BIC é de 369,9043 e o logaritmo da verosimilhança é -176,6034.

Considerando o modelo obtido pela adição das variáveis que potencialmente descrevem a variável resposta, verificamos que o valor da variância calculado no nível 3 é de 7,53 e no nível 2 é de 0,20. Além disso, os critérios de informação o valor do AIC é 449,3017 e o valor do BIC é 470,0771 e o logaritmo da verosimilhança é -214,6509. Numa comparação imediata é possível verificar a clara discrepância entre os valores do logaritmo de verosimilhança e dos critérios de informação. A diminuição progressiva do valor do logaritmo de verosimilhança verifica-se entre o modelo m1 e o modelo m7. Embora o modelo m7 seja apontado como o que melhor se ajusta ao problema, atendendo ao valor do logaritmo de verosimilhança, os critérios de informação aumentaram. Contudo, a variância que reside nos níveis macro diminuiu com a inclusão das variáveis explicativas, uma vez que aproximam os valores ajustados dos valores reais. Esta diferença é mais relevante na variância do nível 2 que desce de 4,55 para 0,22. É, por isso, necessário calibrar a informação contida em ambos os modelos para averiguar o seu interesse e viabilidade para a

```
> d3<-lmList(AT ~ 1 | INS, data=DD)
> summary(d3)
Call:
  Model: AT ~ 1 | NULL
   Data: DD

Coefficients:
   (Intercept)
     Estimate Std. Error   t value     Pr(>|t|)
BD   4.333333  1.0539531  4.111505 1.158603e-04
CNT 10.777778  0.8605491 12.524303 0.000000e+00
MDL  6.666667  0.6085001 10.955900 4.440892e-16

Residual standard error: 3.651001 on 63 degrees of freedom

> coef(d3)
    (Intercept)
BD     4.333333
CNT   10.777778
MDL    6.666667
```

**Figura 3.13:** Nível 3

```
Linear mixed-effects model fit by REML
 Data: DD
       AIC      BIC    logLik
  361.2068 369.9043 -176.6034

Random effects:
 Formula: ~1 | INS
        (Intercept)
StdDev:    2.769384

 Formula: ~1 | DOP %in% INS
        (Intercept) Residual
StdDev:    2.204429 3.111071

Fixed effects: AT ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 7.330794  1.802569 55 4.066859   2e-04

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.1317139 -0.6408297 -0.1404075  0.6252642  2.3683435

Number of Observations: 66
Number of Groups:
         INS DOP %in% INS
           3           11
```

**Figura 3.14:** Sumário do modelo m1

descrição da variável resposta através do modelo.

Através do nível de significância das variáveis explicativas do modelo m7, apenas o índice de gravidade e o número de transações passariam o teste de hipóteses e teriam um coeficiente diferente de zero associado. Perante estes factos, o modelo m7 trata-se de um modelo com conteúdo pobre. Posto isto,

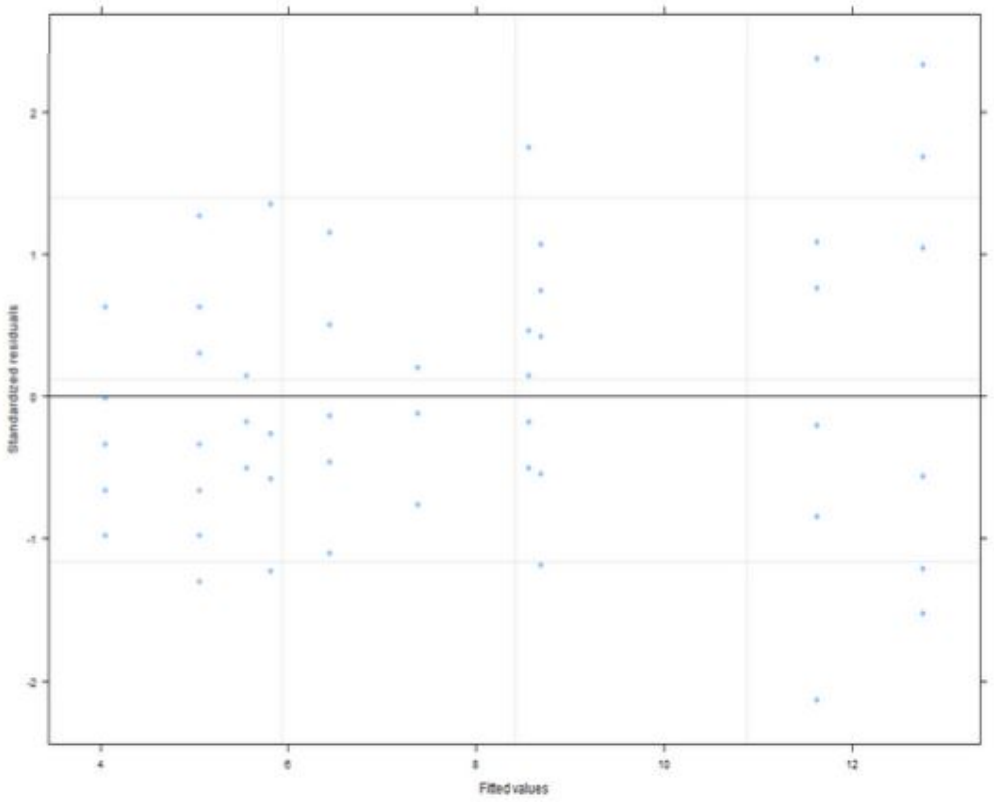


**Figura 3.15:** Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme

será conveniente rever os modelos intermédios entre o m1 e o m7. Atendendo aos *gráficos dos valores ajustados vs resíduos estandardizados* de cada um dos modelos, verifica-se que o gráfico visualmente menos tendencioso é apresentado pelo modelo m6. Este modelo apresenta como variáveis significativas o índice de gravidade, as transações e as vendas e ainda apresenta um valor de variância do segundo nível inferior, de apenas 0,03.

```
Linear mixed-effects model fit by REML
 Data: DD
       AIC      BIC    logLik
  449.3017 470.0771 -214.6509

Random effects:
 Formula: ~1 | INS
        (Intercept)
StdDev:    2.744753

 Formula: ~1 | DOP %in% INS
        (Intercept) Residual
StdDev:   0.4447122 2.986615

Fixed effects: AT ~ PROD + IG + TR + VND + STK + FTES
                Value Std.Error DF   t-value p-value
(Intercept)  6.480169 12.788923 49  0.506702  0.6146
PROD         0.000627  0.001013 49  0.618470  0.5391
IG           7.453902  2.015468 49  3.698349  0.0005
TR          -0.000022  0.000007 49 -3.158468  0.0027
VND          0.000001  0.000001 49  1.350271  0.1831
STK         -0.000001  0.000001 49 -1.045245  0.3010
FTES         0.002726  0.012167 49  0.224061  0.8236
 Correlation:
     (Intr) PROD   IG     TR     VND    STK
PROD -0.927
IG    0.111 -0.172
TR    0.200 -0.472  0.027
VND   0.896 -0.817  0.096  0.083
STK   0.310 -0.519  0.115  0.679  0.161
FTES -0.841  0.886 -0.103 -0.539 -0.840 -0.609

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.00888847 -0.73974927 -0.04105737  0.59122844  1.97855250

Number of Observations: 66
Number of Groups:
         INS DOP %in% INS
           3           11
```

**Figura 3.16:** Sumário do modelo m7

Num balancear de hipóteses, as variáveis explicativas apontadas na análise exploratória são jogadas nos modelos num enlace de valores de se vão ajustando à nossa variável resposta. A constância verificada nos resultados é reduzida e, por isso, a segurança na sua associação a "causa" ou "mote" de influência é vaga. Perante isto, indico como melhor modelo o m1, pois apresenta o menor valor de AIC e BIC e explicita de forma coerente a variabilidade que reside em cada nível da hierarquia dos dados, que vem reforçar o desfasamento entre as realidades vividas nas diferentes insígnias e nas várias direções de operações.

Deste último aspeto ainda advêm variadas questões para se reconhecer em que âmbito a referência geográfica influencia o volume de sinistros, considerando-se a mesma tipologia de loja e sendo administrada a mesma formação e as mesmas metodologias. Questões essas que ainda não foram respondidas e que este trabalho, não encontrando os indícios pretendidos, veio reforçar.

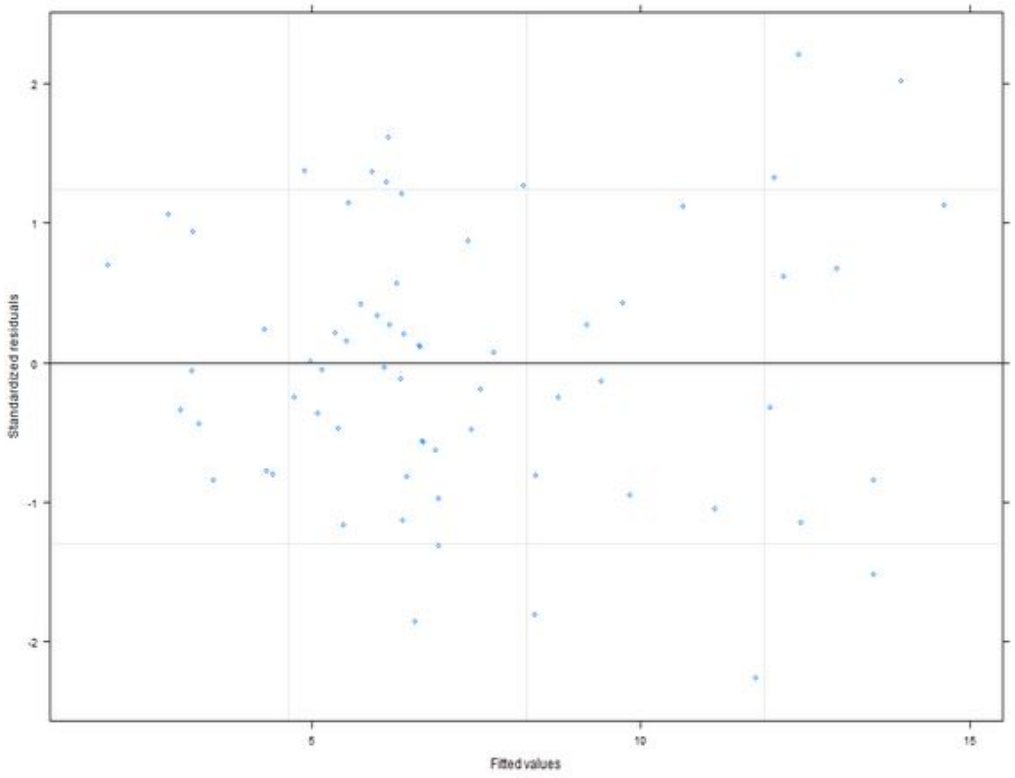


**Figura 3.17:** Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme

```
> summary(m6)
Linear mixed-effects model fit by REML
 Data: DD
       AIC      BIC    logLik
  440.3323 459.1814 -211.1662

Random effects:
 Formula: ~1 | INS
        (Intercept)
StdDev:    2.562438

 Formula: ~1 | DOP %in% INS
        (Intercept) Residual
StdDev:   0.1583129 2.980879

Fixed effects: AT ~ PROD + IG + TR + VND + STK
                Value Std.Error DF   t-value p-value
(Intercept)  8.301680  6.672693 50  1.244127  0.2193
PROD         0.000458  0.000456 50  1.003070  0.3207
IG           7.487775  1.950043 50  3.839799  0.0003
TR          -0.000021  0.000005 50 -3.753435  0.0005
VND          0.000001  0.000000 50  2.914453  0.0053
STK         -0.000001  0.000000 50 -1.159547  0.2517
 Correlation:
     (Intr) PROD   IG     TR     VND
PROD -0.735
IG    0.046 -0.182
TR   -0.556  0.021 -0.028
VND   0.639 -0.278  0.013 -0.811
STK  -0.451  0.035  0.069  0.529 -0.818

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.94069876 -0.76236505 -0.02865462  0.58018028  2.01607589

Number of Observations: 66
Number of Groups:
         INS DOP %in% INS
           3           11
```

**Figura 3.18:** Sumário do modelo m6

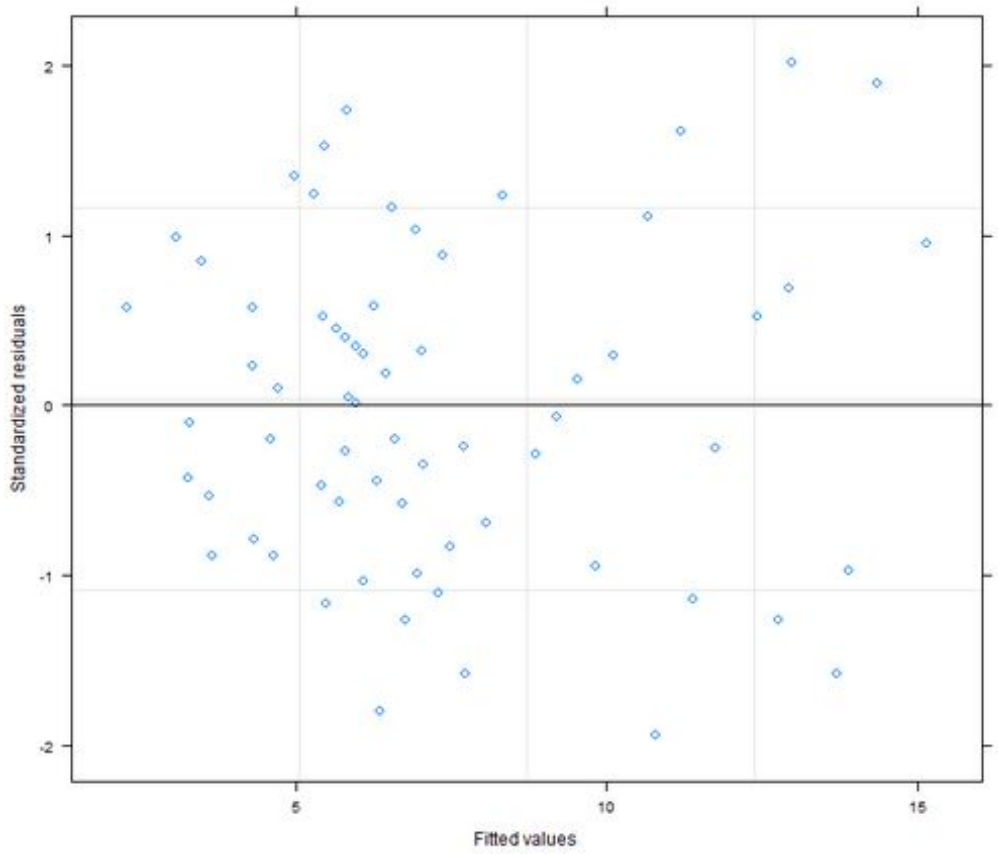


**Figura 3.19:** Gráfico dos valores ajustados vs erros estandardizados obtido pelo comando plot, usando o pacote estatístico nlme

# Capítulo 4.

# Conclusões e trabalhos futuros

Procurei causas e medidas de ação, mas os resultados não nos permitiram chegar ao génesis da questão. Reconheci cenários no panorama da sinistralidade do retalho alimentar Sonae. Deparei-me com discrepâncias e semelhanças, seguidas de hipóteses, de motivos e de razões. Contudo, as variáveis que tive oportunidade de indagar e cruzar com os dados desta realidade não me deram segurança a quaisquer conclusões.

Perante a investigação que desenvolvi em torno deste tema e considerando o que hoje conheço da sinistralidade em loja e dos constrangimentos e benefícios a que os colaboradores estão sujeitos, tomaria em atenção alguns casos de estudo que seria conveniente analisar. Alguns estiveram prementes ao longo de toda a pesquisa, mas nunca tiveram uma base adequada de trabalho. Quanto mais refinado for o nosso universo de estudo, mas escassos são os dados e menos representativos se tornam.

Refiro-me a dois casos em específico. O primeiro é o estudo específico de lojas com histórico de sinistros extensos, que apesar de serem muitos para quem considera a meta dos "Zero Acidentes"não têm representatividade suficiente de valor que nos permita avaliar evoluções e o acaso. Neste caso, joga em nosso desfavor, pois uma única dessas ocorrências pode alterar toda a análise. O segundo é o estudo de uma direção de operações de uma insígnia, por exemplo, a DOP sul da insígnia Continente Hipermercado em contraste com a DOP norte da mesma insígnia e daí, apurar diferenças em metodologias de loja, comportamentos ou ações. São, apesar de tudo, questões facilmente abaláveis, pois tratam-se de direções de operações da mesma insígnia, com o mesmo âmbito de negócio, colaboradores com a mesma formação e metodologia. E inúmeros casos semelhantes poderia referir.

Tratando-se de amostras de uma variável de contagem com número de ocorrências inconstante, se descermos até às direções de operações os registos da variável escasseiam, fragilizam-se as conclusões e torna-se difícil especificar soluções. Atendendo à proporção das observações comparativamente a variáveis como a produtividade, o volume de vendas, as FTES, entre outras, a oscilação em uma unidade da variável resposta causa logo um abalo na estrutura de análise e, por conseguinte, nos resultados. Esta oscilação tem como porta de entrada o acaso. Porque o colaborador escorrega, cai, tropeça, é atingido por objetos, e nem sempre é por falha técnica, ambiental ou pessoal.

Uma solução seria olhar os dados através de um filtro, considerando apenas os que efetivamente resultaram de falha humana, de falha técnica ou por falha nas condições de trabalho. Mais uma vez, refinar, diminuir o número de sinistros considerados no estudo, mas estarem conscientemente isentos do fator

aleatório "acaso".

Conhecendo a realidade das lojas, ir até as seções mais sinistradas, como a peixaria, a padaria, o talho e procurar indícios e razões para se poder agir. Por aqui caminhamos novamente para a génese da questão e começamos a perdem amostra, a levantar outras dúvidas, mas continuo a considerar que seria o meio mais conveniente para se chegar a novas conclusões. Entre os vários despistes que fui realizando aos dados, julgo que este seja um bom mote para trabalhos futuros, com um igual background e uma boa ideia de análise a aplicar.

A sinistralidade e as suas consequências humanas e financeiras parecem-se razão mais que suficiente para a procura de causas e novos veículos de ação.

# Bibliografia


Kullback-leibler information number and log-likelihood. http://www-ssc.igpp.ucla.edu/personnel/russell/ESS265/Ch9/autoreg/node14.html. Accessed: 2015-08-01.

Curso de especialização em engenharia de segurança no trabalho. http://academico.escolasatelite.net/system/application/materials/uploads/25/m1d1-guia-de-estudo-parte-i---a-evolucao-historica-da-engenharia-de-seguranca-do-trabalho.pdf. Accessed: 2015-06-01.

M. Aitkin and N. Longford. Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, pages 1–43, 1986.

M. E. F. Barbosa and C. Fernandes. Modelo multinível: uma aplicação a dados de avaliação educacional. *Estudos em Avaliação Educacional*, (22):135–154, 2013.

D. M. Bates. lme4: Mixed-effects modeling with r. *URL http://lme4. r-forge. r-project. org/book*, 2010.

G. C. Bergamo. *Aplicação de modelos multiníveis na análise de dados de medidas repetidas no tempo*. PhD thesis, Universidade de São Paulo, 2002.

A. S. Bryk and S. W. Raudenbush. *Hierarchical linear models: applications and data analysis methods*. Sage Publications, Inc, 1992.

U. Campinas. Análise de regressão. http://www.portalaction.com.br/analise-de-regressao/12-estimacao-dos-parametros-do-modelo. Accessed: 2015-10-21.

J. C. d. CARMO, I. Almeida, M. Binder, M. Settimi, and R. Mendes. Acidentes do trabalho. *Patologia do trabalho*, pages 431–455, 1995.

P. C. EMILIANO, E. P. VEIGA, M. J. VIVANCO, and F. S. MENEZES. Critérios de informação de akaike versus bayesiano: Análise comparativa. *19º Simpósio Nacional de Probabilidade e Estatística*, 2010.

C. A. C. Fernandes. *Estimação das escalas dos construtos capital social, capital cultural e capital econômico e análise do efeito escola nos dados de Peru-PISA 2000*. PhD thesis, PUC-Rio, 2005.

D. Gadd, S. Karstedt, and S. F. Messner. *The Sage Handbook of Criminological Research Methods*. Sage, 2011.

H. Goldstein. *Multilevel statistical models*, volume 922. John Wiley & Sons, 2011.

J. Gray, D. Jesson, H. Goldstein, K. Hedger, and J. Rasbash. A multi-level analysis of school improvement: Changes in schools' performance over time. *School Effectiveness and School Improvement*, 6(2):97–114, 1995.

R. H. Heck and S. L. Thomas. *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus*. Routledge, 2015.

J. Hox. Multilevel modeling: When and why. In *Classification, data analysis, and data highways*, pages 147–154. Springer, 1998.

J. Hox. *Multilevel analysis: Techniques and applications*. Routledge, 2010.

J. J. Hox and E. D. d. Leeuw. Multilevel models for meta-analysis. 2003.

D. E. Hunt. Teaching styles and pupil progress. *Interchange*, 7(4):39–45, 1976.

W. A. d. S. Júnior, Oliveira e Cruz. Correção de alta ordem de estimadores de máxima verossimilhança. http://repositorio.ufpe.br:8080/handle/123456789/12153. Accessed: 2015-08-01.

I. K. Kasznar and B. GOLÇAVES. Regressão múltipla: uma digressão sobre seus usos, 2011.

L. Kish. Survey sampling. 1965.

J. A. Laros and J. L. P. Marciano. Análise multinível aplicada aos dados do nels: 88. *Estudos em avaliação educacional*, 19(40):263–278, 2013.

A. L. D. B. C. Leão, M. R. Prata Júnior, W. C. Centurion, and D. E. P. d. Silva. Segurança e saúde ocupacional: o caso de uma instituição pública de sergipe. 2009.

E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

A. Lorenzo. Mixed models in r: lme4, nlme, or both? http://freshbiostats.wordpress.com/2013/07/28/mixed-models-in-r-lme4-nlme-both/, 2013. Accessed: 2015-07-21.

S. R. d. Lucca and M. Fávero. Os acidentes do trabalho no brasil-algumas implicações de ordem econômica, social e legal. *Rev. bras. saúde ocup*, 22(81):7–14, 1994.

J. A. Maia, V. P. Lopes, R. Silva, A. Seabra, J. Ferreira, and V. Cardoso. Modelação hierárquica ou multinível. uma metodologia estatística e um instrumento útil de pensamento na investigação em ciências do desporto. 2003.

J. A. Maia, R. Silva, A. Seabra, V. P. Lopes, J. Vinagre, D. Freitas, and A. Prista. Dados longitudinais e modelação hierárquica. um tutorial para investigadores das ciências do desporto. 2005.

M. A. Matos. Manual operacional para a regressão linear. *FEUP, Porto, Portugal*, 1995.

P. Monteiro, F. A. Santos, and G. Santos. Costs of safety at work vs. costs of "no" safety at work—building sector. *Occupational Safety and Hygiene*, page 149, 2013.

A. M. Moreira. Introdução à segurança no trabalho. *Gestão e Segurança de Obras e Estaleiros*, 2008.

A. A. O'Connell and D. B. McCoach. *Multilevel modeling of educational data*. IAP, 2008.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

N. F. Pidgeon. Safety culture and risk management in organizations. *Journal of cross-cultural psychology*, 22(1):129–140, 1991.

J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

E. d. A. PINTO. *Hidrologia estatística*. CPRM, 2007.

H. Quené and H. Van den Bergh. On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1):103–121, 2004.

S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

S. C. A. Rodrigues. Modelo de regressão linear e suas aplicações. 2012.

T. A. Snijders and R. J. Bosker. Introduction to multilevel analysis, 1999.

T. M. Soares. Influência do professor e do ambiente em sala de aula sobre a proficiência alcançada pelos alunos avaliados no simave-2002. *Estudos em Avaliação Educacional*, (28):103–124, 2013.

M. V. B. L. Torman. *Coeficiente de Correlação Intraclasse: Comparação entre métodos de estimação clássico e bayesianos.* PhD thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2011.

M. Torrecilla and F. Javier. Los modelos multinivel como herramienta para la investigación educativa. *Magis. Revista Internacional de Investigación en Educación*, 1(1), 2012.

Á. Zocchio. *Prática da prevenção de acidentes: ABC da segurança do trabalho*. Editôra Atlas, 1971.