U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Estudo de Satisfação de Clientes

Nuno Daniel Monteiro Bastos Marinho

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Engenharia Matemática

2018

MSc

2.º CICLO

FCUP
2018

U.PORTO

Estudo de Satisfação de Clientes

Nuno Daniel Monteiro Bastos Marinho

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

NORS

We Know How

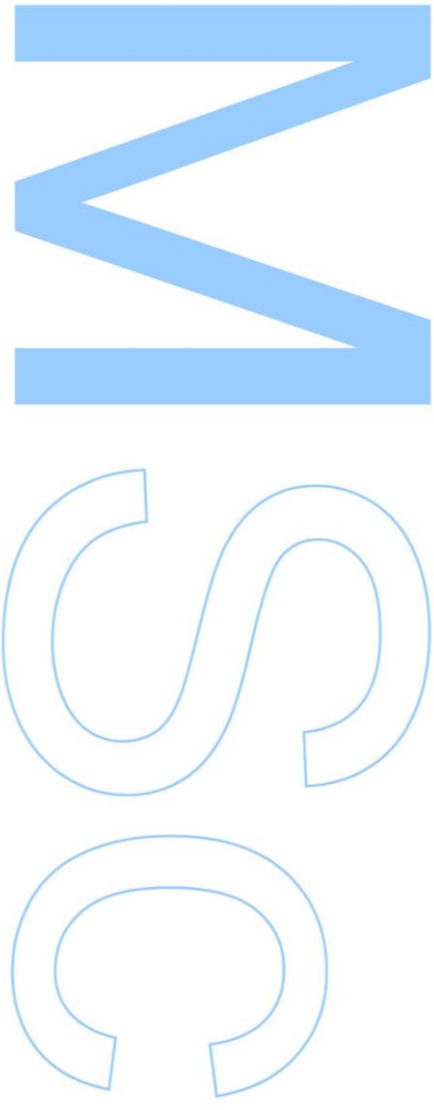
Estudo de Satisfação de Clientes

Nuno Daniel Monteiro Bastos Marinho
Mestrado em Engenharia Matemática
Departamento de Matemática
2018

Orientador Interno
Sandra Ramos, Professor Adjunto
Departamento de Matemática
Instituto Superior de Engenharia do Porto – IPP

Orientador FCUP
Teresa Mendonça, Professor Auxiliar
Departamento de Matemática
Faculdade de Ciências da Universidade do Porto

Supervisor de Estágio
Patrícia Araújo, Gestão Operacional,
Market Intelligence & Customer Experience
(MI & CE)



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do júri,

__/__/_____.

Dedicatória

À minha mãe,
pela coragem, determinação e compreensão

À minha avó, Ana,
pelo amor, preocupação e dedicação

A toda a minha restante família,
pelo apoio, orientação e motivação

E a todos aqueles que*, assim como eu,*
tem gosto e paixão pela Matemática Aplicada e pela Estatística

Agradecimentos

À minha família, especialmente, à minha mãe, avó e tios,
por todo o apoio, incentivo e compreensão ao longo de todos os anos da minha formação

Aos professores da FCUP,
pela sabedoria transmitida ao longo da minha formação

Entre estes, em particular, à professora doutora Rita Gaio,
a quem devo todos os conhecimentos adquiridos na área da estatística

À comissão de mestrado, ao grupo NORS e à FCUP
pelo proporcionamento da realização deste estágio curricular

À Dra. Patrícia Araújo, elo com o grupo NORS,
pela simpatia, disponibilidade e prestabilidade que sempre evidenciou

À minha excelente orientadora, professora doutora Sandra Ramos,
pelo aconselhamento, por toda a ajuda, simpatia, prontidão, amizade, disponibilidade e compreensão demonstradas

Resumo

A avaliação dos níveis de satisfação de clientes tornou-se uma tarefa crucial para a manutenção da competitividade e para as decisões comerciais das empresas. No presente trabalho apresentam-se resultados de um estudo de avaliação dos níveis de satisfação de clientes de 5 empresas do Grupo NORS.

Iniciaram-se os trabalhos, após um extenso pré-processamento, com uma análise descritiva exploratória dos dados existentes.

Seguidamente, após definição da variável resposta, ajustaram-se modelos de regressão logística de forma a identificar, entre os vários aspetos avaliados (por exemplo, a satisfação com os preços, com a execução do trabalho realizado, com a diversidade de oferta, etc.), aqueles que, de forma significativa, melhor explicam a satisfação dos clientes - chamados de *drivers* de excelência. Os modelos obtidos para as várias empresas do grupo apresentaram um poder descriminante bom e uma boa exatidão.

Finalmente, este estudo investigou a existência de potenciais relações entre os níveis de satisfação do cliente e as vendas. Nesta avaliação consideraram-se duas abordagens. A primeira usa métodos de regressão linear onde se utiliza, como variável objetivo, o declive das vendas. A segunda usa modelos longitudinais (modelos lineares gerais e modelos lineares mistos) onde se utilizam diretamente as vendas como variável objetivo. Desta análise resultou, para uma das empresas do grupo, apenas um modelo que estimou que, em média, um cliente passar de neutro a promotor (na variável recomendar) resulta num aumento de 3.9 pontos percentuais na margem percentual das vendas. Nos restantes casos e empresas, não foi encontrada relação estatisticamente significativa entre os níveis de satisfação do cliente e as vendas.

Palavras-chave: Análise de dados, análise de dados longitudinais, estatística, R, modelação, programação, regressão logística e linear, satisfação de clientes, vendas.

Abstract

Assessing customer satisfaction levels has become a crucial task for maintaining competitiveness and for business decision-making. This paper presents the results of a study evaluating customer satisfaction levels of 5 NORS Group companies.

The work was started after an extensive pre-processing, with an exploratory descriptive analysis of the existing data.

Then, after defining the response variable, logistic regression models were adjusted in order to identify, among the various aspects evaluated (for example, satisfaction with prices, execution of work performed, diversity of supply, etc.).), those who, in a significant way, better explain customer satisfaction - called drivers of excellence. The models obtained for the various companies of the group presented good descending power and good accuracy.

Finally, this study investigated the existence of potential relationships between levels of customer satisfaction and sales. Two approaches were considered in this evaluation. The first uses linear regression methods where the sales slope is used as the objective variable. The second uses longitudinal models (general linear models and mixed linear models) where sales are directly used as objective variables. From this analysis, of the companies in the group, only a model that estimated that, on average, a customer goes from neutral to promoter (in the recommend variable) results in an increase of 3.9 percentage points in the percentage margin of sales. In the other cases and companies, no statistically significant relationship was found between customer satisfaction levels and sales.

Keywords: Data analysis, longitudinal data analysis, statistics, R, modeling, programming, logistic and linear regression, customer satisfaction, sales.

Conteúdo

Glossário

MLG/GLM .. Modelo(s) Lineare(s) Generalizado(s)/Generalized Linear Model(s)

f.d.p ..... Função densidade de probabilidade

f.m.p ..... Função massa de probabilidade

i.i.d ..... Independentes e identicamente distribuídos

NPS ..... Net Promotor Score®

a.a ..... Amostra aleatória

MMV ..... Método da Máxima Verosimilhança

EMV ..... Estimador(es)/Estimação de/por Máxima Verosimilhança

GLS ..... Generalized Least Squares

MMQG ..... Método dos mínimos quadrados generalizado

MVN ..... Multivariate Normal Distribution (Distribuição Normal Multivariada)

$IC_{100(1-\alpha)\%}$ ... Intervalo de confiança a $100(1-\alpha)\%$

ML/MV ..... *Maximum Likelihood*/Máxima Verosimilhança

REML ..... *Restricted Maximum Likelihood Residual* (Máxima Verosimilhança Residual Restrita)

ACF ..... *Autocorrelation function* (função de auto-correlação)

EB ..... *Empirical Bayes*

EM ..... Expectation-Maximization

NR ..... Newton-Raphson

BLUP ..... *Best Linear Unbiased Predictor*

OR ..... *Odds Ratio*

Curva ROC .. *Receiver Operating Characteristic Curve*

AUC ..... Área sob a curva ROC

ACC ..... Acurácia/Exatidão

NA ..... *Missing values*

BD ..... Base de Dados

EP ..... Erro Padrão

Est ..... Estimativa

$\mathbf{se}(\widehat{\beta}_{\mathbf{j}})$ ..... Desvio padrão de $\widehat{\beta}_j$

$\overset{\mathbf{a}}{\sim}$ ..... Segue assimptoticamente

Lista de Figuras

Lista de Tabelas

Capítulo 1.
Introdução

"Por mares nunca dantes navegados"

Luís de Camões - Canto I «Os Lusíadas»

1.1 Enquadramento do problema

O conteúdo presente nesta secção é proveniente das seguintes fontes da *web*: *Conceito de satisfação do cliente* (2013), Oliveira (2018) e Duarte (2012).

A noção de satisfação do cliente diz respeito ao nível de conformidade da pessoa quando realiza uma compra ou utiliza um serviço. O senso comum indica que, quanto maior for a satisfação, maior é a possibilidade de o cliente voltar a comprar ou a contratar serviços no mesmo estabelecimento. É possível definir a satisfação do cliente como sendo o nível do estado de espírito de um indivíduo, que resulta da comparação entre o rendimento auferido do produto ou serviço com as suas expectativas. Isto significa que o objectivo de manter o cliente satisfeito é primordial para qualquer empresa. Os especialistas de marketing afirmam que é mais fácil e barato voltar a vender algo a um cliente habitual do que conquistar um novo cliente.

Os benefícios da satisfação do cliente são numerosos. Um cliente satisfeito

- terá maior probabilidade de repetir a compra, comprando mais e com maior frequência;
- comunica as suas experiências positivas a quem o rodeia (conhecidos, familiares, etc.), ou seja, a propaganda "boca a boca" promovida por clientes satisfeitos é tão eficaz quanto outras ferramentas de comunicação e muito mais barata;
- é fiel à empresa e, nestas condições, demonstra menor sensibilidade ao preço;

- gera menos reclamações, portanto, resulta num menor custo operacional.

É importante, por conseguinte, controlar as expectativas do cliente de forma periódica para que a empresa esteja atualizada na sua oferta e proporcione aquilo que o comprador procura.

Por outro lado, os clientes insatisfeitos atuam de forma negativa. Num estudo realizado com a Dell, uma das maiores empresas de tecnologia do mundo, publicado em Reichheld e Markey (2011), chega-se à conclusão que são necessários 5 comentários positivos para neutralizar 1 comentário negativo. Note-se assim a importância da avaliação da satisfação de clientes. Para além disso, a avaliação do nível de satisfação permitirá obter informações como as seguintes:

- O que está certo ou errado no atendimento/serviços prestados;
- Qual empresa se destaca no atendimento e que merece reconhecimento;
- Qual a empresa que tem tendência a ser líder em vendas;
- Quais são os meios de se alcançar a excelência no atendimento.

No grupo NORS, a existência de inquéritos de satisfação não é uniforme, ou seja, algumas empresas do grupo fazem inquéritos há mais de 40 anos, outras só desde 2013, etc. Para além disso, segundo o *Relatório e contas consolidadas* (2017) (página 26), em linha com a estratégia definida para a região Ibéria, foi criada em 2017 uma área de *Market Intelligence and Customer Experience*, que dedica a sua atividade ao suporte das empresas e negócios, no âmbito do conhecimento de mercado e da gestão do relacionamento com clientes. Esta área visa dotar a Região e as empresas de um melhor domínio do seu mercado alvo, assegurando dessa forma a melhoria no prospeção de mercado, o aumento da penetração de vendas e o aumento da rentabilidade associada. Tudo isto assente numa visão de relacionamento de longo prazo com todos os clientes.

1.2 Objetivos

Esta dissertação tem como objetivo estudar os níveis de satisfação de clientes[1] do Grupo NORS, ou seja, compreender a satisfação de cliente nas variáveis mais relevantes para o Grupo

[1] Mercado objeto de estudo. Podem ser clientes provenientes de vários setores (venda, pós venda, pesados, ligeiros).

e antecipar o comportamento do cliente na compra de produtos e/ou serviços. Pretende-se que a análise seja executada sobre dados recolhidos em várias empresas do grupo e dividida em 3 fases.

Numa **fase inicial**, após um extenso pré-processamento e cruzamento de bases de dados, pretende-se fazer uma análise dos dados (análise exploratória univariada e multivariada) com o objetivo de identificar qual é o nível médio de satisfação para estas marcas. Seguidamente, o objetivo é obter a distribuição destes resultados dentro dos vários *clusters* (ex: nível de satisfação por distritos, por concelhos, etc.) e comparar os resultados possíveis.

Numa **segunda fase**, o objetivo será identificar quais são os aspetos que melhor influenciam (positiva ou negativamente) a satisfação dos clientes de forma a serem tidos em conta em decisões administrativas.

Na **fase final**, pretende-se avaliar em que medida a satisfação dos clientes afeta a quantidade e o volume das suas vendas[2].

Note-se que todas as fases têm uma interligação entre si quase que criando um efeito catapulta pois primeiramente identificam-se os níveis onde a satisfação média assume valores satisfatórios/insatisfatórios[3], depois identificam-se os aspetos em que uma dada marca deve investir de forma a melhorar a satisfação dos clientes e, por fim, de que forma esta afeta o volume e valor líquido das vendas.

Os resultados desta análise deverão ser utilizados como *inputs* dos modelos existentes para a segmentação de clientes e para definição da política comercial.

1.3 Metodologia

Iniciar-se-ão os trabalhos com a limpeza e preparação das bases de dados disponíveis nas várias empresas em estudo, ou seja, será feito um extenso pré-processamento dos dados que tem como objetivo limpar os dados, categorizar variáveis, cruzar bases de dados, remover colunas desnecessárias/redundantes, etc. Após isso, serão consideradas técnicas de análise descritiva e exploratória para dar resposta à fase inicial deste projeto.

[2] Obtidas ao longo do tempo para os clientes que deram resposta ao questionário feito.

[3] A definir posteriormente.

A procura de indicadores com efeito significativo nos níveis de satisfação dos clientes será baseada na regressão logística. Para isso, será definida uma variável objetivo (sucesso/insucesso) à custa da variável de satisfação selecionada a partir dos questionários[4]. Após isso, recorre-se a testes de hipóteses[5], usando todas as perguntas/variáveis pretendidas, para avaliar quais devem pertencer ao modelo (inicial) de regressão logística. Concluída esta fase, procede-se ao ajustamento do modelo de regressão logística que melhor explica[6] a satisfação dos clientes identificando, assim, quais as variáveis → perguntas do questionário → aspetos da empresa se devem ter em conta nas decisões de gestão a tomar em tempos futuros.

A regressão linear multivariada e a análise de dados longitudinais foram os métodos usados para dar resposta à fase final. Numa primeira fase, a variável objetivo será o declive das vendas. Recorrendo a uma técnica mais elaborada, usou-se diretamente o valor das vendas ao longo do tempo. Desta forma, dá-se por terminada a metodologia utilizada.

1.4 Linguagem e Software usados

A componente prática desta dissertação foi realizada utilizando a versão 3.4.4. da linguagem *R*, por meio do *software RStudio* (versão 1.1.419).

Figura 1.1: *R* e *RStudio*.

O `R` (Cordeiro (2017)) é uma linguagem gratuita (*software* livre) e um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos que está disponível para Windows, para as mais diversas variantes de Unix/Linux e Mac OS X. É também uma das ferramentas de análise de dados mais usada tanto no meio académico como no meio empresarial. Esta linguagem é uma versão da linguagem e ambiente S, sendo por isso semelhante a este (Torgo (2017)). A linguagem S foi criada na Universidade de Auckland (Nova Zelândia) tendo sido desenvolvida

[4] Que irá ser comum a todas as empresas estudadas.

[5] Teste de χ^2/Fisher ou Mann-Whitney (a serem definidos posteriormente).

[6] Tendo em conta um certo conjunto de métricas de qualidade do ajustamento e testes de hipóteses (apresentados mais à frente).

na Bell Laboratories[7] por John Chambers e outros colegas sendo, atualmente, desenvolvida pelos mais diversos investigados espalhados pelo mundo. Embora existam algumas diferenças importantes, grande parte do código escrito S é executado inalterado sob `R`. A linguagem `R` fornece uma grande variedade de técnicas estatísticas e gráficas e é altamente extensível. O ambiente S é muitas vezes a primeira escolha para pesquisa em metodologia estatística e `R` fornece, neste sentido, uma via de código aberto. Um dos pontos fortes do `R` é a facilidade com que se podem produzir representações gráficas de grande qualidade[8], incluindo símbolos matemáticos e fórmulas, sempre que necessário.

O RStudio é um ambiente de desenvolvimento integrado (IDE) para R. Inclui uma consola, um editor de texto com realce de sintaxe que suporta a execução direta de código, bem como ferramentas para representação gráfica, importação de dados, histórico, depuração e gestão do ambiente de trabalho. Para além disso, simplifica bastante o uso de certas ferramentas úteis, como, por exemplo, o *R Markdown*.

1.5 Estrutura da dissertação

Esta dissertação está organizada em 5 capítulos, sendo eles, por ordem, introdução, o grupo NORS, conceitos fundamentais, resultados e conclusões. Inclui também as referências usadas e anexos.

No primeiro capítulo é feita uma introdução que engloba o enquadramento do problema, os objetivos, a metodologia usada, o *software* usado e a presente descrição. No capítulo seguinte, é feita a apresentação do grupo NORS onde se pode ficar a conhecer a distribuição do grupo pelo mundo, o seu modelo organizacional, as suas áreas de negócio e marcas abordadas, a sua visão, missão e valores, perceber a posição que o grupo tem perante os clientes e, por fim, perceber a origem e a história do grupo. No terceiro capítulo, são introduzidos todos os conceitos fundamentais[9] que foram usados para se obterem os resultados apresentados no Capítulo 4. Para terminar, segue-se o capítulo das conclusões provenientes dos resultados

[7] Anteriormente AT&T, agora Lucent Technologies.

[8] Por exemplo, usando a *package* ggplot2.

[9] Para melhor entendimento dos métodos utilizados.

obtidos, seguindo-se as referências (consultas feitas para auxiliar a realização deste trabalho) e os anexos que, devido a questões de confidencialidade, não incluirão qualquer código usado na realização deste projeto.

Capítulo 2.
O Grupo NORS

"So the problem is not so much to see what nobody has yet seen, as to think what nobody has yet thought concerning that which everybody sees."

Arthur Schopenhauer (1788-1860)

O que agora se apresenta foi obtido utilizando o site oficial do *Grupo Nors* (2014) e um *Documento interno: Company profile* (2017) fornecido pelo grupo.

O grupo NORS possui o *slogan: We know how* e é um grupo português cuja visão é ser um dos líderes mundiais em soluções de transporte, equipamentos de construção e equipamentos agrícolas.

Tem na sua génese 84 anos de história e atividade em Portugal, iniciada por Luís Óscar Jervell, com a representação da marca Volvo, em 1933.

2.1 O Grupo no Mundo

Assumindo integralmente a sua vocação multinacional, o Grupo NORS rege-se hoje por uma estratégia assente em princípios e políticas transversais e por uma cultura de Grupo global, direcionada para um crescimento sustentado, suportado por produtos, serviços e Recursos Humanos de excelência.

Atualmente, o Grupo NORS está presente em 17 países distribuídos por 4 continentes (conforme se pode ver na Figura 2.1): Portugal, Espanha, Brasil, Angola, Botswana, Namíbia, Moçambique, Cuba, México, EUA, Turquia, Áustria, Republica Checa, Eslováquia, Roménia, Hungria e Croácia com cerca de 3.749 colaboradores e um volume de negócios superior a 1.4

mil milhões de euros.

Figura 2.1: O grupo NORS no Mundo. Fonte: *Documento interno: Company profile* (2017).

2.2 Modelo Organizacional

O Grupo NORS é constituído por seis áreas operacionais:

- **NORS Ibéria**[1]**:** reúne todas as operações do Grupo em Portugal e Espanha e incorpora as empresas Auto Sueco, Galius, Civiparts Portugal, Civiparts España, AS Parts e ONEDRIVE;
- **NORS Angola:** com operações em todas as principais cidades Angolanas, que incorpora as empresas Auto Sueco Angola, Civiparts Angola, Auto-Maquinaria, ONEDRIVE Angola e Vitrum;
- **NORS Brasil:** com as empresas Auto Sueco São Paulo e Auto Sueco Centro Oeste, com operações nos estados de Mato Grosso, Rondônia e Acre, e a Agro New em Catanduva e Votupuranga, no interior do Estado de São Paulo;
- **NORS África:** que reúne a Auto Sueco Botswana, Auto Sueco Namíbia e Auto Sueco Moçambique;

[1] Nesta dissertação, serão apenas abordadas empresas desta área operacional.

- **NORS Ventures** onde se agruparam as empresas Auto Sueco Automóveis, Biosafe, Amplitude Seguros e Sotkon;

- **O Grupo Ascendum:** que a NORS detém em 50%, é um importante ativo do Grupo, e um dos maiores fornecedores mundiais de equipamentos industriais para construção e infraestruturas.

Em Outubro de 2013 o Grupo Ascendum expandiu as suas atividades para a Europa Central, onde 282 colaboradores e 33 unidades de negócio asseguram a operação do Grupo na Áustria, Hungria, República Checa, Eslováquia, Roménia e Croácia - e desenvolvem o negócio de Máquinas e Equipamentos de Construção Volvo nestes países.

A NORS tem ainda uma participação na Unevol empresa com sede em Havana, Cuba, que exerce atividade de importação dos produtos Volvo Penta e presta assistência após venda a produtos Volvo *Construction Equipment*.

2.3 Áreas de negócio e marcas abordadas

Historicamente associado à sua liderança no setor automóvel, o Grupo NORS é hoje uma multinacional com um âmbito de atuação alargado, que desenvolve as suas atividades em quatro grandes áreas de negócio: *Original Equipment Solutions, Integrated Aftermarket Solutions, Recycling Solutions* e *Safekeeping Solutions* (conforme mostra a Figura 2.2). Nesta dissertação serão apenas abordadas empresas das duas primeiras áreas referidas anteriormente. Segue-se uma breve explicação das mesmas e das empresas (que lhe estão associadas) usadas na realização deste trabalho.

- ***Original Equipment Solutions:*** Materializa a atividade histórica do Grupo, fruto da relação que mantém com a Volvo desde 1933 e inclui a venda e após-venda de camiões, autocarros, máquinas de construção, equipamentos agrícolas, automóveis, motores marítimos e industriais, geradores e componentes originais. O desenvolvimento desta área de negócio é impulsionado pela presença das empresas que a constituem em diversos continentes. Engloba as empresas Auto Sueco, Galius, Auto Sueco Automóveis, Auto Sueco

Angola, Auto-Maquinaria, Auto Sueco Botswana, Auto Sueco Namíbia, Auto Sueco São Paulo, Auto Sueco Centro Oeste, Agro New, no interior de São Paulo, Auto Sueco Moçambique e o Grupo Ascendum.

Destacam-se as empresas[2]:

- **Auto Sueco:** Empresa-mãe do Grupo, importadora distribuidora exclusiva de camiões, autocarros e motores marítimos Volvo para Portugal. A Auto Sueco é também importadora exclusiva de grupos geradores SDMO para Portugal.
- **Galius:** Importador e distribuidor exclusivo de camiões Renault Trucks, para Portugal, com atividades de venda e após-venda.

Figura 2.2: Marcas do Grupo NORS. Fonte: *Documento interno: Company profile* (2017).

- ***Integrated Aftermarket Solutions:*** Na área de *Integrated Aftermarket Solutions*, o Grupo reúne o conjunto de empresas da sua estrutura de após venda, que inclui a importação e distribuição de peças O.E.M. (*Original Equipment Manufacturer*) multimarca para camiões e automóveis e vidro de construção e decoração através das marcas Civiparts, AS Parts, ONEDRIVE e Vitrum.

  Destacam-se as seguintes marcas:

[2]Divididas em dois grupos: *Workshop* (parte da oficina/após-venda) e *Sales* (parte comercial/vendas).

- **Civiparts:** Importa e distribui peças e equipamentos oficinais para veículos pesados multimarca. A Civiparts foi adquirida pelo Grupo em 2003 e está presente em Portugal, Espanha e Angola;
- **AS Parts:** Dedica-se à distribuição de peças e acessórios multimarca para veículos ligeiros em Portugal;
- **ONEDRIVE:** Retalhista de peças para automóveis ligeiros em Portugal e Angola.

2.4 Visão, Missão e Valores

- **Visão:** ser um dos líderes mundiais em soluções de transporte e equipamentos de construção.
- **Missão:** gerar prosperidade para clientes e fornecedores, de forma a desenvolver os seus colaboradores e criar valor para os acionistas, através de relações de confiança, construídas por uma atitude de exigência e entrega das melhores soluções.
- **Valores:**
  i) **Ambição:** ambicionam crescer e liderar, antecipando as oportunidades e procurando a excelência e a melhoria contínua nas soluções que entregam.
  ii) **Confiança:** pretendem estabelecer relações de confiança e parceria com os *stakeholders*, assentes na competência e na qualidade do seu trabalho e numa atitude pautada pela responsabilidade, exigência, transparência e rigor.
  iii) **Talento:** procuram ativamente atrair, reter e desenvolver talento, valorizando o mérito e desempenho dos seus colaboradores e das suas equipas.

2.5 Clientes

No método de trabalho ágil, é fundamental obter *feedback* constante dos clientes, antes mesmo até de uma tecnologia ou de um produto estarem totalmente desenvolvidos. Uma das

principais tendências é uma colaboração cada vez mais estreita e mais cedo. Assim sendo, o grupo definiu como objetivo estratégico colocar o cliente no centro da organização. Para isso, pretende-se:

- Garantir uma melhoria consistente do índice de satisfação de cliente;
- Incentivar um *mindset* criativo e colaborativo.

2.6 Marcos Históricos

A génese do Grupo NORS remonta a 1933, quando Luiz Óscar Jervell se tornou representante da marca Volvo em Portugal. A 1 de abril de 1949 surge pela primeira vez a denominação Auto Sueco, quando, com Yngvar Poppe Jensen decidem autonomizar o negócio.

Desde 1959 até 1979, a empresa foi inaugurando instalações por todo o país, criando empresas associadas e alargando a rede de concessionários.

O ano de 1991 marca o início da internacionalização do Grupo com a atividade Volvo, com a sua instalação na República de Angola. Essa expansão internacional continuou até 2002, onde o grupo se instalou em Cuba. Depois, essa expansão terminou em 2003 pois o Grupo Auto Sueco adquiriu o Grupo Civiparts (comercialização de peças para veículos pesados) e retornou em 2004 com a criação da ASC Construction Equipment USA – hoje Ascendum USA – e com o início das operações em Windhoek, na Namíbia.

2007 marca o início das operações do Grupo no Brasil e na Angola. Já o ano de 2008 ficou marcado pela comemoração dos 75 Anos do Grupo. Em 2009, Tomás Jervell é nomeado Presidente do Grupo Auto Sueco e o Grupo entra no mercado de Cabo Verde com a inauguração da AS Parts Cabo Verde.

2010 é para o Grupo Auto Sueco um ano de investimento com a aquisição da Vocal – hoje Auto Sueco São Paulo –,do Grupo Express Glass e da Volvo Otomotiv Turk. 2 anos depois, em 2012, o Grupo Auto Sueco ultrapassa a fasquia dos 1.000 milhões de Euros e a Auto Sueco Coimbra passa a designar-se Grupo Ascendum. É ainda em 2012 que o Grupo entra no negócio das inspeções obrigatórias a veículos com a aquisição da MasterTest em Portugal.

Em 2013, o Grupo assume a sua vocação multinacional, com presença em 4

continentes e 24 mercados, o que dá origem a uma nova identidade – NORS. 2 anos mais tarde, em Março de 2015, o Grupo NORS assumiu a presença da Renault Trucks em Portugal através da sua nova empresa Galius, que integrou as operações de venda e após-venda, colaboradores, instalações e outros ativos, assim como a relação com a rede privada de serviços da Renault Trucks Portugal Lda.

O último marco histórico do grupo, de conhecimento público, é referente ao início de 2017 onde a Axial Angola passa a ter a denominação de Vitrum, uma marca que se posiciona no mercado angolano como a primeira empresa especializada em vidro de construção e decoração (*building glass*) e películas de segurança, controlo térmico/solar e decoração para viaturas e construção.

Capítulo 3.
Conceitos Fundamentais

"He invented a game that allowed players to predict the outcome?"

Susanna Gregory - To Kill or Cure

3.1 Modelos Lineares Generalizados

A exposição teórica presente nesta secção segue de perto Amaral Turkman (2000) (e as referências nele incluídas) e os documentos Gaio (2012), Gaio (2016a) e Gaio (2016b), fornecidos pela professora doutora Ana Rita Pires Gaio (FCUP/CMUP) após frequência, na FCUP, da unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia (M4015) no ano letivo 2017/2018.

3.1.1 Contextualização histórica

Os Modelos Lineares Generalizados (MLG/GLM[1]) foram introduzidos por Nelder e Wedderburn (1972), como uma família unificadora de modelos para análise de regressão não-standard com respostas não-normais. Estes modelos tiveram um impacto muito grande no desenvolvimento da estatística aplicada, contudo foram necessários cerca de 20 anos para que a sua aplicação passasse para o conhecimento geral. Nos dias de hoje, devido ao facto de o *software* desenvolvido na altura ter ficado mais fácil de manusear, existem vários pacotes estatísticos que contém ferramentas para lidar com este tipo de modelação.

Os MLG constituem uma generalização da regressão linear clássica para respostas (aproxi-

[1] Generalized Linear Models, na literatura inglesa.

madamente) normais à regressão para respostas não-normais, incluindo proporções (respostas binárias), contagens e outras, como as que provêm das distribuições gama ou binomial negativa. Mais geralmente, os modelos envolvem variáveis resposta cuja distribuição pertença a uma família de distribuições com propriedades muito específicas: a família exponencial. Assim sendo, o modelo de regressão linear clássico e o modelo de regressão logística são casos particulares de MLG[2] e são os únicos modelos lineares generalizados a serem abordados nesta dissertação.

Apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo facto da distribuição da variável resposta se restringir à família exponencial e por exigirem a independência das respostas, os MLG são cada vez mais usados devido à facilidade na interpretação e análise dos resultados obtidos e pela sua rápida resposta computacional. Do ponto de vista teórico, estes modelos são igualmente importantes pois a sua metodologia permite obter uma abordagem unificada de muitos procedimentos estatísticos usualmente usados nas aplicações e promover o papel central da verosimilhança na teoria da inferência.

3.1.2 Notação, Terminologia e Tipo de Dados

Considere-se a situação em que há uma variável aleatória Y, variável resposta ou variável dependente, e um vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ de p variáveis explicativas[3], que se acredita explicar parte da variabilidade inerente a Y. A variável resposta Y pode ser contínua, discreta ou dicotómica. As covariáveis podem ser também de qualquer natureza.

Considerem-se dados da forma

$$(y_i, \mathbf{x}_i), \quad i = 1, \ldots, n, \tag{3.1}$$

resultantes da realização de $(Y, \mathbf{x})$ em n indivíduos ou unidades experimentais, sendo as componentes Y_i do vetor aleatório $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ independentes, que em forma matricial podem

[2] Modelos de análise de variância e covariância, o modelo de regressão de Poisson, modelos log-lineares para tabelas de contingência multidimensionais e o modelo *probit* para estudos de proporções são outros casos particulares dos modelos lineares generalizados. Alguns deles podem ser consultados em Amaral Turkman (2000).

[3] Também designadas por covariáveis ou variáveis independentes.

ser representados da seguinte forma

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}. \tag{3.2}$$

3.1.3 A Família Exponencial

Os modelos lineares generalizados pressupõem que a variável resposta tenha uma distribuição pertencente à família exponencial.

Definição 3.1.1 (Família Exponencial). *Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função densidade de probabilidade ou função massa de probabilidade se puder escrever na forma:*

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3.3}$$

onde θ e ϕ são parâmetros escalares, $a(\cdot)$, $b(\cdot)$ e $c(\cdot,\cdot)$ são funções reais conhecidas.

Na definição anterior, θ é a forma canónica do parâmetro de localização e ϕ é um parâmetro de dispersão suposto, em geral, conhecido. Neste caso a distribuição descrita em (3.3) faz parte da família exponencial uniparamétrica (diz-se que distribuição de Y está na forma canónica). Quando o parâmetro ϕ é desconhecido a distribuição pode ou não fazer parte da família exponencial bi-paramétrica, tal como é geralmente definida (veja-se, por exemplo, Cox e Hinkley (1974)). Admite-se, ainda, que a função $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros. Neste caso prova-se que a família em consideração obedece às condições habituais de regularidade (para um estudo de condições de regularidade necessárias no desenvolvimento do estudo que se vai fazer, deve consultar-se um livro avançado de Estatística. Aconselha-se, por exemplo, Sen e Singer (1993)).

3.1.3.1 Valor médio e variância

Seja $\ell(\theta;\phi,y) = ln(f(y|\theta,\phi))$. Defina-se a função *score*

$$S(\theta) = \frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}. \tag{3.4}$$

Nestas condições, tem-se

$$\begin{aligned} &E(S(\theta)) = 0 \\ &E(S^2(\theta)) = E\left[\left(\frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ell(\theta;\phi,Y)}{\partial \theta^2}\right] \end{aligned} \tag{3.5}$$

e portanto como, no caso em que $f(y|\theta,\phi)$ é dado por (3.3),

$$\ell(\theta;\phi,y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi),$$

obtém-se

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)} \qquad \frac{\partial S(\theta)}{\partial \theta} = -\frac{b''(\theta)}{a(\phi)}, \tag{3.6}$$

onde $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ e $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$.

Assim de (3.5) e (3.6)

$$E(Y) = \mu = a(\phi)E(S(\theta)) + b'(\theta) = b'(\theta) \tag{3.7}$$

$$var(Y) = a^2(\phi)var(S(\theta)) = a^2(\phi)\frac{b''(\theta)}{a(\phi)} = a(\phi)b''(\theta). \tag{3.8}$$

Vê-se assim que a variância de Y é o produto de duas funções: $b''(\theta)$ e $a(\phi)$. Onde

- $b''(\theta)$, que depende apenas do parâmetro canónico θ (e portanto do valor médio μ), é designada por **função de variância** e representada por $V(\mu)$;
- $a(\phi)$ depende apenas do parâmetro de dispersão ϕ.

Observa-se que a função $a(\phi)$ toma a forma $a(\phi) = \dfrac{\phi}{\omega}$, onde ω é uma constante conhecida, obtendo-se portanto a variância de Y como o produto do parâmetro de dispersão por uma função apenas do valor médio.

Neste caso a função definida em (3.3) escreve-se na forma

$$f(y|\theta,\phi,\omega) = \exp\left(\frac{\omega}{\phi}(y\theta - b(\theta)) + c(y,\phi,\omega)\right). \tag{3.9}$$

3.1.3.2 Exemplos

São várias as distribuições frequentemente usadas que pertencem à família exponencial. Destacam-se as distribuições: normal, Bernoulli (para variáveis dicotómicas), binomial (para proporções de êxitos em n provas de Bernoulli), Gama (distribuição contínua assimétrica), Poisson (para variáveis de contagem) e gaussiana invertida (distribuição contínua assimétrica).

Detalhe do caso da distribuições binomial e normal.[4]

Se Y for tal que mY tem uma distribuição binomial com parâmetros m e π ($Y \sim B(m, pi)/m$), a sua f.m.p é dada por

$$\begin{aligned} f(y|\pi) &= \binom{m}{ym}\pi^{ym}(1-\pi)^{m-ym} \\ &= \exp\{ym\ln\pi + m(1-y)\ln(1-\pi) + ln\binom{m}{ym}\} \\ &= \exp\{m(y\theta - \ln(1+e^{\theta})) + \ln\binom{m}{ym}\} \end{aligned}$$

com $y \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$ e $\theta = \ln(\frac{\pi}{1-\pi})$.

Vê-se assim que esta f.m.p é da forma (3.9) com

$$\begin{aligned} &\theta = \ln\left(\frac{\pi}{1-\pi}\right), \\ &b(\theta) = \ln(1+e^{\theta}), \qquad c(y,\phi) = \binom{m}{ym}, \\ &b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi, \qquad b''(\theta) = V(\mu) = \frac{e^{\theta}}{(+e^{\theta})^2} = \pi(1-\pi), \\ &a(\phi) = \frac{\phi}{\omega}, \quad \phi = 1, \quad \omega = m. \end{aligned}$$

De (3.7) e (3.8) obtém-se diretamente

$$E(Y) = b'(\theta) = \pi, \quad var(Y) = b''(\theta)a(\phi) = \frac{\pi(1-\pi)}{m}.$$

O parâmetro canónico é a função *logit*, $\ln\left(\frac{\pi}{1-\pi}\right)$.

No caso da distribuição normal, $Y \sim N(\mu, \sigma^2)$ e

$$\begin{aligned} f(y,\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\left(\frac{\mu y - \mu^2/2}{\sigma^2}\right) - \left(\frac{1}{2}\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]. \end{aligned}$$

[4] Pois os resultados aqui obtidos serão usados mais à frente.

Aqui

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2}, \quad a(y,\phi) = \sigma^2 = \phi, \quad c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + 2\log(2\pi\sigma^2)\right).$$

A distribuição está escrita na forma canónica (3.3), o parâmetro canónico é a média $\theta = \mu$ e o parâmetro de dispersão é a variância $\phi = \sigma^2$. De (3.7) e (3.8) obtém-se diretamente

$$E(Y) = b'(\theta) = \theta = \mu, \quad var(Y) = b''(\theta)a(\phi) = \sigma^2.$$

3.1.4 Descrição do Modelo Linear Generalizado

Os modelos lineares generalizados são uma extensão do modelo linear clássico

$$\mathbf{Y} = Z\boldsymbol{\beta} + \varepsilon, \tag{3.10}$$

onde:

- Z é uma matriz de dimensão $n \times k$ de especificação do modelo ou matriz do desenho, em geral, da forma

$$Z = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix},$$

  ou seja, a matriz de covariáveis X com um primeiro vetor unitário. Daqui resulta que $k = p + 1$ e que $\mathbf{z}_i = (1, x_{i1}, \dots, x_{ip})^T$, onde p é o número de variáveis explicativas;

- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$ é o vetor de parâmetros - coeficientes ou parâmetros de regressão - a estimar. Este vetor também pode ser escrito, em função do número de variáveis explicativas, da seguinte forma: $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$. Nesta caso, β_0 diz-se o coeficiente independente ou termo constante (*intercept*). No designação anterior, esse coeficiente é o β_1;

- ε é um vetor de erros aleatórios com distribuição que se supõe $N_n(\mathbf{0}, \sigma^2\mathbf{I})$, onde $\mathbf{I}$ é a matriz identidade de dimensão $n \times n$;

- $\hat{\varepsilon} = \mathbf{Y} - \widehat{\mathbf{Y}}$ é a diferença entre os valores reais e os valores previstos, ou seja, é o vetor dos resíduos.

Estas hipóteses implicam que o valor esperado da variável resposta seja uma função linear das variáveis explicativas, isto é, $E(\mathbf{Y}|Z) = \boldsymbol{\mu} = Z\boldsymbol{\beta}$.

Os MLG são compostos por três componentes:

1. **Componente aleatória**

   As variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial e com $E(Y_i|\mathbf{x}_i) = \mu_i$ (para a i-ésima de n observações independentes).

2. **Componente estrutural ou sistemática**

   Define-se um preditor linear (com carácter preditivo), η_i, como combinação das variáveis explicativas dado por

   $$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{z}_i^T \boldsymbol{\beta}$$

3. **Função de ligação**

   As componentes aleatória e sistemática estão relacionadas por uma função de ligação $g(\cdot)$, monótona e diferenciável, da seguinte forma

   $$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{z}_i^T \boldsymbol{\beta}.$$

   Dado que esta função é invertível, tem-se ainda que

   $$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) = g^{-1}(\mathbf{z}_i^T \boldsymbol{\beta}).$$

   A escolha da função de ligação a utilizar depende do tipo de resposta e do problema com o qual se esteja a lidar. Por exemplo, no caso da regressão linear e da regressão logística, as funções de ligação que se usaram foram a função identidade e a função *logit*, respetivamente. Estas e outras funções de ligação consideradas nos MLG podem ser consultadas na Tabela 3.1.

Note-se que a existência de variáveis explicativas qualitativas leva à necessidade da sua codificação à custa de variáveis binárias mudas (designadas por *dummies*). Por exemplo, caso uma variável qualitativa (ou factor) tenha q categorias, então são necessárias $q - 1$ variáveis binárias para a representar, sendo que todas estas variáveis devem ser incluídas no vetor $\mathbf{z}$.

Tabela 3.1: Funções de ligação consideradas nos MLG.

identidade	recíproca	quadrática inversa
μ	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
raiz quadrada	expoente	logarítmica
$\sqrt{\mu}$	$(\mu + c_1)^{c_2}$	$\ln(\mu)$
logit	complementar log-log	*probit*
$\ln(\frac{\mu}{1-\mu})$	$\ln[-ln(1-\mu)]$	$\Phi^{-1}(\mu)$

$\Phi(\cdot)$ é a função de distribuição da distribuição normal *standard*.

3.1.5 Metodologia dos Modelos Lineares Generalizados

Perante a tarefa de modelar dados através de um MLG, há 3 passos essenciais que se devem ter em conta:

- Formulação dos modelos;
- Ajustamento dos modelos;
- Seleção e validação dos modelos.

Na **formulação do modelo** é necessário ter em linha de conta

1. escolha da distribuição para a variável resposta. A verificação da distribuição da variável resposta é fundamental para a escolha do MLG a aplicar. Por exemplo, caso se esteja a lidar com uma variável contínua, fazer o seu histograma e gráficos que verifiquem a adesão à normalidade podem levar a que se opte por uma regressão linear, caso estes gráficos mostrem que se está na presença de uma variável resposta aproximadamente normal. Por outro lado, este tipo de análise pode levar à necessidade de se transformar previamente os dados. Assim, uma análise preliminar dos dados, é fundamental para que se possa fazer uma escolha adequada da família de distribuições a considerar. Para além disso, também é possível verificar se não houve erros na obtenção dos dados e efetuar a sua devida correção.

2. escolha das covariáveis e formulação apropriada da matriz de especificação. Aqui há que entrar em linha de conta com o problema específico em estudo e, muito particularmente,

ter em conta a codificação apropriada das variáveis de natureza categórica de modo a facilitar a sua identificação. Por outro lado, muitas das vezes é-se confrontado com bases de dados com muitas variáveis e, muitas vezes, apenas se tem interesse em resolver o problema com algumas delas daí ser necessário efetuar uma escolha das covariáveis.

3. escolha da função de ligação. A escolha de uma função de ligação adequada aos dados com que se está a trabalhar deve resultar de uma combinação de considerações sobre o problema em causa, análise intensiva dos dados e facilidade de interpretação do modelo pois uma escolha errada da função de ligação conduz a uma modelação errada do problema em causa.

A fase do **ajustamento do modelo** (ou modelos) passa pela estimação dos coeficientes β's associados aos preditores, e do parâmetro de dispersão ϕ caso este esteja presente. Para além disso, é importante verificar a significância dos parâmetros, obter intervalos de confiança e realizar testes de qualidade do ajustamento.

A fase de **seleção e validação dos modelos** tem por objetivo modelos que sejam candidatos a modelo final, ou seja, modelos que com um número razoável de parâmetros que sejam adequados aos dados e que tenham possuam todos os efeitos significativos[5]. Detetar discrepâncias entre os dados e os valores previstos, averiguar a existência e eventual remoção de *outliers* e/ou observação frequentes são procedimentos que se podem e devem adotar para encontrar candidatos a melhor modelo. Para além disso, adequabilidade, parcimónia[6] e interpretação são três aspetos bastante importantes na seleção do modelo final. Se este atingir um equilíbrio entre esses três fatores, encontra-se na presença de um bom modelo.

3.1.6 Inferência

De modo a se poder aplicar a metodologia dos modelos lineares generalizados a um conjunto de dados há a necessidade de, após a formulação do modelo que se pensa adequado, proceder à realização de inferências sobre esse modelo. A inferência com MLG é, essencialmente, baseada na verosimilhança[7]. Consequentemente, a estimação dos parâmetros de regressão, os testes

[5] No caso de variáveis categóricas, basta apenas uma das *dummy* ser significativa para que a variável permaneça no modelo.

[6] Sempre que possível, optar pelo modelos mais simples, ou seja, com menos variáveis explicativas.

[7] Probabilidade de ocorrência dos dados naquela distribuição (modelo), ou seja, probabilidade de o modelo se ajustar aos dados.

de hipóteses sobre os coeficientes do modelo e o testes de qualidade do ajustamento são, em geral, baseados na verosimilhança. Com efeito, não só o método da máxima verosimilhança é o método de eleição para estimar os parâmetros de regressão, como também os testes de hipóteses sobre os coeficientes do modelo e de qualidade do ajustamento são, em geral, métodos baseados na verosimilhança.

Os métodos inferenciais a serem discutidos nesta secção pressupõem que o modelo está completa e corretamente especificado, de acordo com a formulação apresentada em 3.1.4. Para mais detalhes sobre este tópico, consulte-se Amaral Turkman (2000).

3.1.6.1 Estimação dos parâmetros do modelo

Considere-se que se tem os dados na forma $(y_i, \mathbf{x}_i)$, $i = 1, \ \ldots, n$, como em (3.1), onde y_i é o valor observado da variável resposta para a i-ésima unidade experimental e $\mathbf{x}_i$ é o correspondente vetor de covariáveis. Designe-se ainda por $\mathbf{z}_i = \mathbf{z}_i(\mathbf{x})$ o vetor de especificação de dimensão k, associado ao vetor de preditores definido anteriormente. Para evitar complexidades no desenrolar da exposição teórica, admita-se ainda que a matriz de especificação Z dada por:

$$Z = (\mathbf{z}_1, \mathbf{z}_2, \ \ldots, \mathbf{z}_n)^T = \begin{pmatrix} z_{11} & z_{12} & \ldots & z_{1k} \\ z_{21} & z_{22} & \ldots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \ldots & z_{nk} \end{pmatrix} \tag{3.11}$$

é de característica completa[8], isto é, tem característica igual à ordem k (mínimo entre n e k já que se assume $n > k$). Isto implica que a matriz $Z^T Z$ tem característica k.

A metodologia de estimação a ser aqui apresentada não é alterada quer se tenham dados desagrupados[9] ou agrupados. De modo a facilitar a exposição aqui descrita, e sem perda de generalidade, suponha-se sempre que a dimensão é n.

Num modelo linear generalizado o parâmetro $\boldsymbol{\beta}$ é aquele que interessa estimar. O parâmetro de dispersão ϕ, quando existe, é considerado um parâmetro perturbador, sendo a sua estimação feita pelo método dos momentos. Os parâmetros de um MLG podem ser estimados usando o **método da máxima verosimilhança**(**MMV**).

[8] *Full rank.*

[9] Amostra de dimensão n.

O estimador $\widehat{\boldsymbol{\beta}}$ de máxima verosimilhança do vetor de parâmetros β é obtido maximizando a *log-verosimilhança*

$$\ell(\theta, \phi|y) = \sum_{i=1}^{n} \ell_i(\theta_i, \phi_i|y_i)$$

onde a contribuição da observação i de uma a.a. $y_1, \ldots, y_n$ é, a menos de uma constante relacionada com $c(y_i, \phi)$ e que portanto não contém $\theta(x)$,

$$\ell_i(\theta_i, \phi_i|y_i) = \omega_i \left(\frac{y_i\theta_i(x) - b(\theta_i(x))}{\phi} \right)$$

(usa-se a distribuição da família exponencial na forma canónica com $a_i(\phi) = \phi/\omega_i$, onde $\boldsymbol{\omega}$ é um vetor de pesos conhecidos).

Para a obtenção de solução, o mais simples é iniciar-se um processo de otimização numérica para o estimador de máxima verosimilhança (por exemplo o método de Newton-Raphson com Fisher *scoring*) e observar se existe divergência ou convergência das soluções. Sendo certo que o processo pode parar numa solução local, o recomendado é que se efetuem vários ciclos de iterações com diferentes escolhas dos valores iniciais.

Note-se ainda que a não-existência de soluções pode ser ultrapassada usando um tamanho amostral maior, já que a teoria garante existência assimptótica de estimadores de máxima verosimilhança, ou por estimação de Bayes escolhendo à priori uma distribuição estritamente côncava para $\boldsymbol{\beta}$.

No caso dos modelos lineares generalizados, mostra-se (McCullagh e Nelder (1989)) que o método de Newton-Raphson combinado com o *scoring* de Fisher é assimptoticamente equivalente a um método de mínimos quadrados iterativamente pesados (*iteratively re-weighted least squares*), na medida em que, à medida que n cresce, as propriedades dos estimadores, em termos de distribuição, tornam-se idênticas.

No caso do modelo gaussiano, os estimadores de máxima verosimilhança conseguem ser exatamente determinados, sem recorrer a métodos de aproximação numérica. Esses estimadores de máxima verosimilhança coincidem com os valores obtidos pelo método dos mínimos quadrados e existem em forma fechada porque a média estimada da resposta pode ser um número real qualquer, sem restrições, enquanto que, por exemplo, no modelo binomial essa média (que é uma proporção) tem de estar entre 0 e 1.

Note-se que o parâmetro de dispersão também pode ser estimado usando o método da máxima verosimilhança.

As propriedades assimptóticas dos EMV, assim como considerações sobre a existência e unicidade dos mesmos, podem ser encontradas em Amaral Turkman (2000).

3.1.6.2 Testes de Hipóteses: Teste de Wald

De acordo com a hipótese nula a testar, escreva-se o vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ na forma

$$\boldsymbol{\beta} = (\boldsymbol{\beta}', \boldsymbol{\beta}'')$$

com $\boldsymbol{\beta}' \in \mathbb{R}^{r+1}$ onde $r \geq 0$ e $\boldsymbol{\beta}'' \in \mathbb{R}^{p-r}$ com $p > r$ (as restrições garantem que nenhum dos subvetores seja vazio).

A hipótese nula que incide sobre um ou mais parâmetros de regressão pode então escrever-se como:

$$H_0 : \boldsymbol{\beta}'' = \mathbf{0} \qquad versus^{10} \qquad H_1 : \boldsymbol{\beta}'' \neq \mathbf{0}$$

Há três testes de hipóteses, estatisticamente equivalentes, com versões univariadas ou multivariadas, que podem ser usados para testar H_0: o teste da razão de log-verosimilhanças, o teste de Wald e o teste de Score cujas estatísticas, são deduzidas das distribuições assintóticas dos estimadores de máxima verosimilhança e de funções adequadas desses estimadores.

I A *Estatística de Wilks* ou *Estatística de razão de log-verosimilhanças*, baseada na distribuição assintótica da razão do máximo das verosimilhanças sob as hipóteses H_0 e $H_0 \cup H_1$.

II A *Estatística de Wald*, baseada na normalidade assintótica do estimador de máxima verosimilhança $\widehat{\boldsymbol{\beta}}$.

III A *Estatística de Rao* ou *Estatística score*, baseada nas propriedades assintóticas da função *score*.

Dos testes acima referidos, apresenta-se aquele que foi usado e que está implementado por defeito no `R`: o **teste de Wald**.

[10] Mais geralmente, pode-se considerar $\boldsymbol{\beta}'' = \boldsymbol{\beta}''^0$, para um qualquer valor $\boldsymbol{\beta}''^0$.

O teste de Wald é uma aproximação quadrática da estatística da razão de verosimilhanças, e portanto computacionalmente mais atrativo. É um teste simples que usa a distribuição normal assimptótica do estimador de máxima verosimilhança $\widehat{\boldsymbol{\beta}}$, de $\boldsymbol{\beta}$.

A **estatística de Wald** é uma forma quadrática que corresponde à distância pesada entre a estimativa $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\beta}}'')$ de $\boldsymbol{\beta}$ e o valor $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}', 0)$ de $\boldsymbol{\beta}$ sob H_0:

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^t \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\right)^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$$

onde $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ é a matriz de covariância de $\widehat{\boldsymbol{\beta}}$.

A versão multivariada deste teste é muito pouco usada, por requerer operações entre vetores e matrizes, e, portanto, não trazer ganhos computacionais sobre o teste da razão de verosimilhança.

Na versão univariada, isto é, no caso em que a hipótese nula afirma que apenas um coeficiente é nulo, diga-se

$$H_0 : \beta_j = 0$$

tem-se $r = p-1$ e todas as $p-1$ coordenadas de $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ devem ser eliminadas, bem como todas as linhas e colunas da matriz $\left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\right)^{-1}$ correspondente. Sob H_0, obtém-se

$$W = \left(\frac{\widehat{\beta}_j}{\widehat{se}(\widehat{\beta}_j)}\right)^2 \sim \chi^2(1), \qquad j = 0, 1, \ldots, p.$$

Na verdade, a normalidade assimptótica do estimador de máxima verosimilhança $\widehat{\boldsymbol{\beta}}$, diz que, sob $H_0 : \beta_j = 0$,

$$W = \frac{\widehat{\beta}_j}{\widehat{se}(\widehat{\beta}_j)} \sim N(0,1), \qquad j = 0, 1, \ldots, p.$$

Valores elevados de W indicam uma distância grande entre $\widehat{\boldsymbol{\beta}}$ e $\boldsymbol{\beta}^0$, pelo que corresponderão à rejeição de H_0, ou seja, a hipótese nula é rejeitada, a um nível de significância α, se o valor observado de W for superior ao quantil de probabilidade $1-\alpha$ de uma $N(0,1)$ (ou de um $\chi^2(1)$, consoante a estatística W usada).

O método de Wald permite obter um intervalo de confiança para os coeficientes do modelo, individualmente. Para β_j, $j = 1, 2, \ldots, p$, o intervalo com $100(1-\alpha)\%$ de confiança correspondente é

$$IC_{100(1-\alpha)\%} = \left(\widehat{\beta}_j - N_{1-\alpha/2}\widehat{se}(\widehat{\beta}_j), \widehat{\beta}_j + N_{1-\alpha/2}\widehat{se}(\widehat{\beta}_j)\right)$$

onde $N_{1-\alpha/2}$ representa o $(1-\alpha/2)$-quantil da distribuição normal reduzida, $N(0,1)$.

3.1.7 Seleção e Validação de Modelos

Na exposição feita até aqui admitiu-se que o modelo proposto, em termos da combinação - distribuição da variável resposta e função de ligação - era um modelo adequado. No entanto, quando se trabalha com muitas covariáveis, tem interesse saber qual o modelo mais parcimonioso, isto é, com o menor número de variáveis explicativas, que ofereça uma boa interpretação do problema posto e que ainda se ajuste bem aos dados. O problema da seleção do modelo corresponde à procura do "melhor modelo", no sentido de ser um modelo que atinge um bom equilíbrio entre os três factores "bom ajustamento", "parcimónia" e "interpretação". Dado que no processo de seleção há uma série de modelos em consideração, convém descrever vários que são comummente referidos durante o processo:

- o **modelo saturado** com n parâmetros $(\mu_1, \ldots, \mu_n)$ linearmente independentes, um por cada observação, e que coincidindo com os dados explica-os exatamente (pois as estimativas de máxima verosimilhança dos μ_i são as próprias observações, isto é, $\widehat{\mu}_i = y_i$). A matriz do modelo é a identidade de dimensão $n \times n$. Os n parâmetros são geralmente obtidos usando modelos de regressão polinomial de ordem suficientemente alta ou tratando as covariáveis como fatores, e depois considerando várias interações. Não oferece qualquer simplificação e, como tal, não tem interesse na interpretação do problema, já que não faz sobressair características importantes transmitidas pelos dados. Além disso, tem pouca hipótese de ser um modelo adequado em réplicas do estudo. Este modelo é pouco informativo porque traduz exatamente o mesmo que os dados. Para além disso, atribui toda a variação dos dados à componente sistemática dada por: $\mu(x_i) = y_i$;

- o **modelo nulo**, que representa a situação em que não existe qualquer relação entre as variáveis explicativas e a resposta; geralmente este modelo tem um único parâmetro μ comum a todas as observações (pense-se, por ex., num modelo linear gaussiano). É um modelo, de interpretação sem dúvida simples, mas que raramente captura a estrutura inerente aos dados. A matriz do modelo é, neste caso, um vetor coluna unitário. Contrariamente ao modelo anterior, este modelo atribui toda a variação nos dados à componente

aleatória; a componente sistemática é $\mu(x_i) = \mu$.

A existência de um número elevado de modelos a considerar traz problemas de ordem combinatória - o número de combinações possíveis torna-se rapidamente não manejável - e de ordem estatística - como decidir sobre o equilíbrio entre o efeito da inclusão ou exclusão de um termo na discrepância entre y e $\widehat{\mu}$ e a complexidade de um modelo maior? Há pois necessidade de estabelecer uma estratégia para a seleção do melhor, ou dos melhores modelos, já que raramente se pode falar na existência de um único "melhor modelo".

3.1.7.1 Método de seleção de variáveis

Mesmo apresentando um bom ajustamento aos dados, um modelo linear geral com muitas variáveis explicativas traz sempre associados problemas de interpretação. O que se gostaria de ter era um modelo mais simples, com menos variáveis, que evidenciasse os efeitos mais fortes.

Existem vários métodos de seleção de variáveis, destaca-se aquele que foi usado no âmbito desta dissertação designado por *stepwise*. É o procedimento mais sofisticado e mais utilizado (Gaio (2016a)). Envolve inclusão e exclusão de variáveis e pode partir do modelo nulo ou do modelo completo. Em cada passo, avalia-se a adição de uma nova variável ao modelo. Se essa variável contribuir para um modelo melhor (segundo um critério definido a priori), a variável é retirada e todas as variáveis explicativas do novo modelo são testadas para avaliar esse novo modelo. Aquelas que já não contribuírem de forma significativa são excluídas. Idealmente, o método identifica o menor conjunto de variáveis explicativas a considerar na regressão.

3.1.7.2 Critérios de Informação

Se os modelos M_1 e M_2 não forem encaixados, é aconselhável utilizar-se um critério que meça a **quantidade de informação** que o modelo não consegue explicar. Uma medida da quantidade de informação é o logaritmo da verosimilhança do modelo. Como a log-verosimilhança cresce com o número de parâmetros, um tal critério tem de ser parcimonioso quanto ao número de parâmetros para evitar o ajustamento de modelos saturados. Desse modo, um tal critério incorpora um termo que penaliza um número grande de parâmetros e é também função do tamanho da amostra.

Dois critérios de informação usuais são:

- **critério de informação de Akaike**

$$AIC = -2LL + 2p$$

- **critério de informação Bayesiana**

$$BIC = -2LL + p\log(n)$$

onde LL é o logaritmo da verosimilhança do modelo, p é o número total de parâmetros, e n é o número total de observações.

Para qualquer destes dois critérios, quanto menor for o seu valor preferível é o modelo.

Observe-se que o critério BIC penaliza mais o número de parâmetros do que o critério AIC porque o número de parâmetros é multiplicado por $log(n)$ que é maior do que 1 quando $n > 3$. O facto da penalização também crescer com o número de observações serve para penalizar os modelos ajustados a muitas observações e que sejam "pouco verosímeis", i.e., que tenham uma log-verosimilhança pequena. Entre os dois critérios de informação, é preferível usar-se o critério BIC (mas nunca se usam os dois em simultâneo).

3.1.8 Regressão Linear

Conforme referido anteriormente, os MLG correspondem a uma generalização do modelo de regressão linear. Com efeito, se se tiverem n respostas independentes $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ onde

$$\mu_i = \mathbf{z}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} z_{ij}\beta_j,$$

o modelo considerado é um modelo linear generalizado, dado que:

- as variáveis resposta são independentes;
- a distribuição é da forma (3.9), com $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta}$, $\phi = \sigma^2$ e $\omega_i = 1$;
- o valor esperado μ_i está relacionado com o *preditor linear* $\eta_i = \mathbf{z}_i^T\boldsymbol{\beta}$ através da relação $\mu_i = \eta_i$;

- a *função de ligação* é a função identidade, que é, neste caso a função de ligação canónica.

Para este modelo pode-se escrever, para a observação i, $Y_i = \mathbf{z}_i^T\boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$ onde os ε_i são i.i.d e $\varepsilon_i \sim N(0, \sigma^2)$.

Note-se ainda que a formulação apresentada inclui facilmente o caso especial em que $Y_i \sim N(\mu_i, \sigma_i^2)$, com $\sigma_i^2 = \frac{\sigma^2}{\omega_i}$, onde ω_i é um peso conhecido associado à i-ésima observação.

3.1.9 Regressão Logística

Nesta secção, que segue de perto Cordeiro (2017), dar-se-á especial importância ao caso em que a componente aleatória do GLM segue uma distribuição binomial e o conjunto de variáveis explicativas pode ser de qualquer natureza. Esta-se perante um modelo de regressão logística.

Para além da descrição deste modelo, serão também descritas outras ferramentas usadas para a obtenção do modelo final (aquele que será considerado o "melhor modelo").

3.1.9.1 Descrição do modelo

Uma das distribuições de probabilidade que pertence à classe de distribuições da família exponencial é a distribuição binomial. Quando a variável resposta é binária, deve utilizar-se a regressão logística para modelar a probabilidade de ocorrência de uma das suas realizações das classes desta variável.

Considere-se um vetor de p variáveis explicativas $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, onde $i = 1, \ldots, n$ e a variável resposta Y_i que assume o valor 1 (sucesso - presença de determinada característica) com probabilidade π_i e o valor 0 (insucesso - ausência dessa mesma característica) com probabilidade $1 - \pi_i$. Assim, Y_i $(i = 1, \ldots, n)$ é uma variável aleatória com distribuição de Bernoulli onde $E(Y_i|\mathbf{x}_i) = \mu_i = \pi_i = P(Y_i = 1|\mathbf{x}_i)$.[11]

A sua função de probabilidade pode ser escrita na forma

$$\begin{aligned} f(y_i|\pi_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad y_i = 0, 1, \quad i = 1, \ldots, n \\ &= \exp\left(y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) - (-\ln(1-\pi_i))\right) \end{aligned}$$

o que prova a sua pertença à família exponencial.

[11] Para simplificar a escrita, costuma escrever-se $E(Y_i)$ em vez de $E(Y_i|\mathbf{x}_i)$.

Neste caso, a função de ligação é a função *logit*

$$\theta_i = logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} = \eta_i$$

em que a probabilidade do sucesso é dada por

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}} \tag{3.12}$$

Facilmente se verifica que a função $F : \mathbb{R} \to [0,1]$, definida por $F(x) = \dfrac{e^x}{1+e^x}$ é uma função de distribuição. A esta função chama-se função de distribuição logística e, por isso mesmo, o modelo binomial com função de ligação *logit* é conhecido por modelo de regressão logística.

Para obter estimativas para o vetor $\boldsymbol{\beta}$ dos parâmetros do modelo pode ser utilizado o método de máxima verosimilhança. Pretende-se encontrar valores para $(\beta_0, \beta_1, \ldots, \beta_p)$ que maximizem a log-verosimilhança[12].

A função de verosimilhança é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{3.13}$$

e a log-verosimilhança, obtida aplicando o logaritmo a (3.13), é dada por

$$\ell(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i).$$

Usando a igualdade (3.12) obtém-se

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i\left(\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}\right) + \ln\left(1+\exp\left(\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}\right)\right)\right) \tag{3.14}$$

Teoricamente, derivar-se-ia parcialmente em ordem a cada parâmetro a expressão em (3.14) e igualando cada uma das equações a 0 obtinham-se as estimativas de máxima verosimilhança, $\widehat{\boldsymbol{\beta}}$, dos parâmetros do modelo. Na prática, como já foi referido, estas estimativas são obtidas com recurso a métodos numéricos já que, geralmente, as equações não têm solução analítica.

3.1.9.2 *Odds Ratio*

O *odds ratio* consiste numa medida de associação utilizada em regressão logística para completar o teste à significância das covariáveis. O *odds* para o sucesso é definidas pelo quociente

[12] Maximizar a log-verosimilhança facilita bastante o processo, dadas as propriedades do logaritmo. Para além disso, é equivalente a maximizar a verosimilhança pois o logaritmo é uma função crescente.

entre as probabilidades de sucesso e insucesso, sendo que, para uma probabilidade de sucesso π,

$$odds = \frac{\pi}{1-\pi}.$$

A própria probabilidade de sucesso é função do *odds*, já que se tem $\pi = \dfrac{odds}{odds+1}$.

O *odds ratio*, como o próprio nome indica, consiste na razão entre dois *odds*:

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)} \tag{3.15}$$

O *odds ratio* pode tomar qualquer valor não negativo. Para dois eventos independentes com $\pi_1 = \pi_2$, $odds_1 = odds_2$ e $\theta = 1$. Este valor θ serve assim como base de comparação. Quando $\theta > 1$, o *odds* para o sucesso é maior para o evento 1. Desta forma, os integrantes deste evento são mais propícios a sucessos que os do evento 2, isto é, $\pi_1 > \pi_2$. Quando $\theta < 1$, acontece exatamente o contrário, ou seja, $\pi_1 < \pi_2$.

Substituindo as expressões na equação (3.15) pelas probabilidades obtidas a partir do modelo de regressão logística (consultar equação (3.12)), pode estimar-se o valor de θ a partir dos coeficientes do modelo de regressão. Para o caso em que se tem uma variável independente dicotómica que assume os valores 0 ou 1,

$$\widehat{\theta} = \frac{\left(\frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1}}\right) / \left(\frac{1}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1}}\right)}{\left(\frac{e^{\widehat{\beta}_0}}{1+e^{\widehat{\beta}_0}}\right) / \left(\frac{1}{1+e^{\widehat{\beta}_0}}\right)} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1}}{e^{\widehat{\beta}_0}} = e^{\widehat{\beta}_1}.$$

Um intervalo de confiança para o *odds ratio* pode ser obtido calculando os extremos de um intervalo de confiança para β_1 e procedendo à exponenciação dos seus extremos. Os extremos deste intervalo terão assim forma

$$\exp\left[\widehat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{SE}\left(\widehat{\beta}_1\right)\right]$$

em que $\widehat{SE}\left(\widehat{\beta}_1\right) = \left[\widehat{Var}\left(\widehat{\beta}_1\right)\right]^{1/2}$, relação entre erro padrão e variância da estimativa para o parâmetro β_1.

Para maior detalhe pode consultar-se Hosmer e Lemeshow (2013).

3.1.9.3 Teste de independência do χ^2

Quando se está na presença de dados que resultam de contagens, é recorrente a utilização de tabelas de frequências, em que estão discriminadas todas as classificações dos dados, segundo as

suas várias características. Estas classificações são exaustivas e mutuamente exclusivas. A estas tabelas, no caso de se terem duas ou mais variáveis, dá-se o nome de tabelas de contingência.

Numa primeira fase, interessa averiguar se as variáveis aleatórias segundo as quais foi feita a classificação cruzada são ou não independentes, ou seja, se é possível que exista alguma associação entre elas. As hipóteses em teste são:

H_0 : as variáveis são independentes vs H_1 : as variáveis não são independentes.

No caso de duas variáveis com r e c categorias, a estatística de teste terá distribuição χ^2 om $(r-1)(c-1)$ graus de liberdade. O valor observado da estatística de teste é dado por

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}},$$

onde o_{ij} e e_{ij} representam, respetivamente, as frequências observada e esperada na célula correspondente ao cruzamento na tabela das categorias i e j.

Sob a validade da hipótese nula, rejeita-se H_0 se $\chi^2_{obs} \geq \chi^2_{1-\alpha;(r-1)(c-1)}$, ou seja, quando a estatística de teste é maior ou igual que o quantil de probabilidade $1-\alpha$ de uma distribuição χ^2 com $(r-1)(c-1)$ graus de liberdade.

Note-se que o teste não deve ser utilizado se mais de 20% das frequências esperadas, sob hipótese de independência, forem inferiores a 5 ou se alguma delas for mesmo igual a 0. Para além disso, a rejeição da hipótese nula implica que, para o nível de significância considerado, os dois factores em causa não são independentes e portanto estão relacionados. Dado que apenas se sabe que variáveis independentes têm correlação nula (e que o contrário não é necessariamente verdadeiro), dois factores que não sejam independentes podem ou não ter correlação não nula.

Sempre que os dados não se adaptam bem ao teste do qui-quadrado (por terem frequências esperadas muito baixas), o `R` dá essa indicação sob a forma de aviso a seguir ao resultado do teste. Nesta situação, deverá recorrer-se ao teste exato de Fisher que se aplica na situação em que a amostra é pequena ou existem contagens esperadas inferiores a 5. O teste supõe que as margens da tabela são fixas e, sob a hipótese nula de independência entre os factores, isso conduz à distribuição hipergeométrica dos números das células da tabela.

3.1.9.4 Utilidade do teste na regressão logística

O teste acima descrito foi usado neste projeto antes de se começar a efetuar modelos de regressão logística. Testou-se, individualmente, a hipótese de cada uma das variáveis explicativas ser independente da variável resposta. Desta forma, todas as variáveis para as quais não se rejeitou a hipótese nula foram excluídas da modelação, ou seja, não se consideram (na modelação) as variáveis que apresentaram valor-p do teste de independência do χ^2 ou Fisher maior ou igual a α.[13] Por exemplo, suponha-se que, numa base de dados com as variáveis Y (binária), X_1, X_2 e X_3 (categóricas), não se rejeita a hipótese nula de as variáveis Y e X_2 serem independentes, pois o valor-p do teste é 0.5. Então, a variável X_2 não é considerada na modelação. O procedimento prossegue até se efetuar o teste com todas as variáveis.

3.1.9.5 Teste de Hosmer e Lemeshow

O **teste de Hosmer-Lemeshow** (Hosmer e Lemeshow (2013)) é muito utilizado em regressão logística com a finalidade de testar a qualidade do ajustamento, ou seja, o teste comprova se o modelo proposto pode explicar bem o que se observa. Assim sendo, tem-se:[14]

H_0 : o modelo ajusta-se bem aos dados vs H_1 : o modelo não se ajusta bem aos dados.

O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas.

A qualidade do teste é baseada na divisão da amostra segundo as probabilidades ajustadas com base nos valores dos parâmetros estimados pela regressão logística. Os valores ajustados são dispostos de forma crescente e, de seguida, separados em g grupos de tamanho aproximadamente igual. Hosmer e Lemeshow propõe que seja utilizado $g = 10$.

Na literatura há pouca orientação sobre como escolher o número de grupos. As simulações mostradas por Hosmer e Lemeshow foram baseadas no uso de $g > p + 1$, em que p é o número de covariáveis do modelo ajustado. Se as frequências esperadas em alguns dos grupos forem muito pequenas, a estatística do teste de Hosmer-Lemeshow calculada entretanto pode não ser

[13] Em alguns casos é frequente usar-se $\alpha = 0.15$ em vez do habitual $\alpha = 0.05$, para o nível se significância. Neste trabalho, foi considerado $\alpha = 0.05$. Para variáveis contínuas, aplica-se o teste de Wald em vez do teste de independência do χ^2.

[14] Observe-se que, para este teste, não se pretende rejeitar H_0.

confiável. Neste caso, deve-se especificar um número menor de grupos, não se podendo utilizar menos de 3 grupos, pois com $g < 3$ é impossível calcular a estatística do teste.

Antes do cálculo da estatística do teste, é necessário estimar a frequência esperada dentro de cada grupo. Para isso divide-se a variável resposta, que é dicotómica. Para $Y = 1$, a frequência esperada estimada é dada pela soma das probabilidades estimadas de todas as unidades experimentais dentro daquele grupo. Para $Y = 0$, a frequência esperada estimada é dada pela soma dos complementares das probabilidades estimadas de todas as unidades experimentais dentro daquele grupo.

Tendo as frequências esperadas, calcula-se a estatística de Hosmer e Lemeshow, $\widehat{C}$, que é obtida da seguinte forma:

$$\widehat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k\overline{\pi}_k)^2}{n_k\overline{\pi}_k(1-\overline{\pi}_k)},$$

em que:

- n_k é o número de indivíduos no k-ésimo grupo;
- $\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j\overline{\pi}_j}{n_k}$;
- c_k é o número total de combinações de níveis dentro do k-ésimo decil;
- $o_k = \sum_{j=1}^{c_k} y_j$ é o número total de respostas dentro do grupo k.

Hosmer e Lemeshow mostraram por simulação que a estatística do teste segue, aproximadamente, uma distribuição χ^2 com $g-2$ graus de liberdade, quando o modelo está especificado corretamente.

3.1.9.6 Matriz de Confusão e Curva ROC

Quando se desenvolvem modelos de previsão de resultados, é importante validar os resultados de forma a quantificar o seu poder discriminativo e identificar um procedimento ou método como bom ou não para determinada análise. No entanto, deve ter-se presente que a simples quantificação de acertos num conjunto de teste não reflete necessariamente o quão eficiente um modelo é, pois essa quantificação dependerá fundamentalmente da qualidade e distribuição dos dados no conjunto de teste.

A chamada **matriz de confusão**, como se pode verificar na Tabela 3.2, trata-se de uma tabela de contingência em que se apresentam, por linhas, os valores previstos e, por colunas, os valores verdadeiros. Consideram-se valores positivos que o modelo previu positivos como verdadeiros positivos (acerto), valores positivos que o modelo previu negativos como falsos negativos (erro), valores negativos que o modelo previu como negativos como verdadeiros negativos (acerto) e valores negativos que o modelo previu positivos como falsos positivos (erro).

Tabela 3.2: Matriz de confusão

		Valor Verdadeiro	
		positivos (**P**)	negativos (**N**)
Valor Previsto (predito pelo modelo)	positivos	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	negativos	**FN** Falso Negativo	**VN** Verdadeiro Negativo

Esta matriz serve de base para as seguintes medidas:

- **Sensibilidade/*Recall*** (proporção de verdadeiros positivos): capacidade do modelo para predizer corretamente a condição para casos que realmente a têm.

$$P(\widehat{Y} = 1|Y = 1) = \frac{\text{Acertos Positivos}}{\text{Total de Positivos}} = \frac{VP}{VP + FN}$$

- **Especificidade** (proporção de verdadeiros negativos): capacidade do modelo para predizer corretamente a ausência da condição para casos que realmente não a têm.

$$P(\widehat{Y} = 0|Y = 0) = \frac{\text{Acertos Negativos}}{\text{Total de Negativos}} = \frac{VN}{VN + FP}$$

- **Acurácia/Exatidão** (ACC): proporção de predições corretas, sem considerar o que é positivo e o que é negativo mas sim o acerto global. É altamente suscetível a desequilíbrios no conjunto de dados, pelo que pode facilmente induzir a uma conclusão errada sobre o desempenho do modelo.

$$P\left(\left[\widehat{Y} = 1|Y = 1\right] \cup \left[\widehat{Y} = 0|Y = 0\right]\right) = \frac{VP + VN}{P + N}$$

Quando se esta na presença de uma variável resposta binária (1 se o evento se verifica e 0 caso contrário) é necessário escolher uma regra de predição ($\widehat{Y} = 0$ ou 1), já que $\widehat{\pi}$ está entre 0 e 1.

É intuitivo pensar que se o valor de $\widehat{\pi}_i$ for grande, $\widehat{Y}_i = 1$ e se $\widehat{\pi}_i$ for pequeno, $\widehat{Y}_i = 0$. Torna-se assim evidente a necessidade de definir um ponto de corte, ou um limiar de decisão, para se classificar e contabilizar o número de predições positivas e negativas. Dada a arbitrariedade que existe para a sua seleção, uma boa prática consiste na comparação do desempenho dos modelos, sob o efeito de diferentes pontos de corte. Contudo, neste trabalho, considerou-se apenas um único ponto de corte dado que apenas se pretendia uma medida de orientação. Esse ponto de corte foi 0.45.

A **Curva ROC** (*Receiver Operating Characteristic Curve*) foi desenvolvida por engenheiros elétricos e engenheiros de sistemas de radar durante a Segundo Guerra Mundial para detetar objetos inimigos em campos de batalha, também conhecido como teoria de detecção de sinais. Há várias décadas que a análise ROC tem sido utilizada em medicina, radiologia, psicologia e outras áreas. Mais recentemente, foi introduzida em áreas como a aprendizagem automática e a prospecção de dados.

Para cada ponto de corte são calculados valores de sensibilidade e especificidade, que podem ser dispostos num gráfico denominado por curva ROC, que apresenta no eixo das ordenadas os valores de sensibilidade e nas abcissas o complementar da especificidade, ou seja, o valor (1-especificidade). Acaba por se tratar de encontrar uma combinação ótima entre sensibilidade e especificidade.

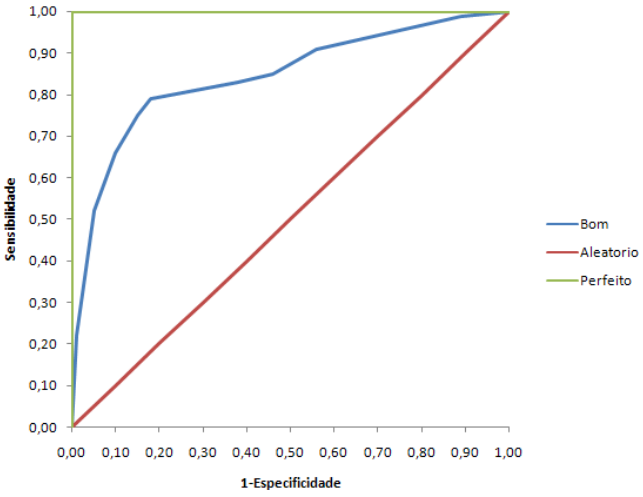


Figura 3.1: Exemplos de curvas ROC.
Fonte: `http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/`.

Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, o que dificilmente é alcançado. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita e quanto maior a distância à linha diagonal, melhor o modelo. Exemplos destes

casos são apresentados na Figura 3.1. A linha diagonal indica uma classificação aleatória, um modelo que seleciona aleatoriamente *outputs* positivos ou negativos. Mesmo que a curva esteja localizada abaixo da diagonal, pode ser convertida, bastanto para isso inverter os seus *outputs*, o que faz com que a curva também seja invertida.

Uma medida padrão para a comparação de modelos é a área sob a curva (AUC), que pode ser obtida por métodos de integração numérica, como por exemplo, o método dos trapézios.

Teoricamente, quanto maior a AUC, melhor o modelo. Hosmer e Lemeshow (Hosmer e Lemeshow (2013)) apresentam valores indicativos da AUC que podem ser utilizados para classificar o poder discriminante do modelo de regressão e que são dados pela Tabela 3.3.

Tabela 3.3: AUC e poder discriminante do modelo.

Área sob a curva ROC (AUC)	Poder discriminante do modelo
< 0.5	Sem poder discriminante
$[0.5, 0.7[$	Discriminação fraca
$[0.7, 0.8[$	Discriminação aceitável
$[0.8, 0.9[$	Discriminação boa
≥ 0.9	Discriminação excecional[15]

Pode ainda dizer-se que as curvas ROC descrevem a capacidade discriminativa de um teste diagnóstico para um determinado número de pontos de corte. Isto permite evidenciar os valores para os quais existe maior otimização da sensibilidade em função da especificidade.

[15] Estes valores deve podem ser indicadores de sobreajustamento.

3.2 Análise de Dados Longitudinais

A exposição teórica apresentada nesta secção é baseada nos documentos Gaio (2018a), Gaio (2018b), Gaio (2018c), Gaio (2018d), e nas referências neles incluídas, e nos apontamentos obtidos após frequência, na FCUP, na unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia (M4015) no ano letivo 2017/2018, lecionada pela professora doutora Ana Rita Pires Gaio (FCUP/CMUP).

Mais detalhes sobre este assunto podem ser encontrados em várias exposições existentes na literatura, por exemplo, Pinheiro e Bates (2000), Diggle et al (1994), Verbeke e Molenberghs (2000) e Davidian e Giltinan (1995).

Encontra-se no Anexo 1 uma exposição teórica que complementa e pode auxiliar a compreensão desta secção, como tal, se necessário, esta deve ser consultada.

Na exposição feita anteriormente, tinha-se apenas uma única observação por unidade experimental. Aqui, tem-se que uma unidade experimental, i, pode ter mais do que uma medição ao longo do tempo, ou seja, t_i medições. Os **dados longitudinais** (designação para este tipo de dados) são assim caracterizados por terem poucas medições por unidade experimental (número finito pequeno). Note-se que estes são diferentes de séries temporais pois estas apresentam um número de medições, por unidade experimental, elevado e os dados longitudinais apresentam especificidades que as séries temporais não apresentam.

Muitas das vezes esta-se perante estudos onde todas as unidades experimentais possuem o mesmo número de medições e estas são obtidas nos mesmos instantes de tempo - **dados equilibrados**. Nestas condições $t_i = t_j = t \ \forall \ i, \ j$. Para além disso, os instantes de tempo podem ou não ser igualmente espaçados.

Por outro lado, há situações onde as observações longitudinais das unidades experimentais não são todas obtidas nos mesmos instantes de tempo (acontece se houver *missings*, por exemplo). Neste caso, esta-se perante **dados não equilibrados**. Nestas condições, $t_i = t_j$ se a unidade experimental i tiver o mesmo número de medições da unidade experimental j ou $t_i \neq t_j$ se o número de medições de i e j for diferente. Tal como anteriormente, os instantes de tempo podem ou não ser igualmente espaçados.

Para simplificação de escrita e sem perda de generalidade, considere-se que se está perante o caso onde os dados são equilibrados onde i designa a unidade experimental i e t representa o tempo. Assim sendo,

- $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})$ é o vetor com todas as observações da unidade experimental i correspondentes à variável aleatória Y;

- $\mathbf{y}_i$ independente de $\mathbf{y}_{i'}$, $\forall\ i \neq i'$;

- $\mathbf{y}_i \sim MVN(X_i\boldsymbol{\beta}, \Sigma)$[16], onde MVN designa a distribuição normal multivariada, Σ a matriz de var-cov de dimensão $t \times t$ e X_i é a matriz que possui a primeira coluna unitária e nas restantes todas as observações de i ao longo do tempo para todas as variáveis explicativas, ou seja,

$$X_i = \begin{pmatrix} 1 & x_{i11} & \ldots & x_{ip1} \\ 1 & x_{i12} & \ldots & x_{ip2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1t} & \ldots & x_{ipt} \end{pmatrix} \in \mathbb{R}^{t\times(p+1)},$$

  onde t é o número de medições, p o número de variáveis explicativas e x_{ipt} representa, para a unidade experimental i, o valor da variável explicativa X_p na medição t.

3.2.1 Estruturas dos dados longitudinais

Os dados longitudinais podem ser apresentados segundo dois formatos: o formato *wide* e o formato *long*. Usualmente, trabalha-se com os dados em formato *long* (pelo que se estes forem fornecidos em formato *wide*[17] ter-se-á de os transformar para a estrutura pretendida).

Pretende-se então que os dados sejam dispostos da seguinte forma:

[16] Ter-se-ia Σ_i em vez de Σ no caso de os dados não serem equilibrados. Nesta situação Σ_i teria dimensão $t_i \times t_i$. Apesar disso, estas matrizes teriam todas a mesma estrutura.

[17] Ou noutro formato qualquer.

Tabela 3.4: Dados Longitudinais em formato *long*.

Unidade Experimental	Observação	Tempo	Variável Resposta	Covariáveis		
1	1	t_{11}	y_{11}	x_{111}	$\dots$	x_{1p1}
1	2	t_{12}	y_{12}	x_{112}	$\dots$	x_{1p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	t_1	t_{1t_1}	y_{1t_1}	x_{11t_1}	$\dots$	x_{1pt_1}
2	1	t_{21}	y_{21}	x_{211}	$\dots$	x_{2p1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	t_2	t_{2t_2}	y_{2t_2}	x_{21t_2}	$\dots$	x_{2pt_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	t_{n1}	y_{n1}	x_{n11}	$\dots$	x_{np1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	t_n	t_{nt_n}	y_{nt_n}	x_{n1t_n}	$\dots$	x_{npt_n}

Note-se que há casos em que a coluna "Observação" é igual à coluna "Tempo". Desta forma, pode-se suprimir uma delas.

3.2.2 Modelo Linear Geral

Uma das formas de modelar dados este tipo é usando um modelo linear geral que, conforme o nome indica, corresponde à obtenção de uma relação linear entre a resposta Y e as variáveis explicativas tendo em linha de conta a estrutura longitudinal. Apresentar-se-á um modelo linear geral do tipo GLS. Salienta-se que este tipo de modelos é usado para fazer inferência a nível populacional e que as conclusões são obtidas em média.

O modelo linear geral, para a unidade experimental i, é então descrito da seguinte forma

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{u}_i, \quad \mathbf{u}_i \sim MVN(0, \Sigma_i), \quad i = 1, \dots, n \tag{3.16}$$

Note-se que, se os dados forem equilibrados, $\Sigma_i = \Sigma$, $\forall i$.

3.2.2.1 Estimação dos parâmetros do modelo

A estimação dos parâmetros do modelo pode ser feita de formas diferentes. Apresentam-se brevemente 2 formas de o fazer, contudo ambas obtém a solução através do método dos mínimos quadrados generalizado.

Uma delas é recorrendo ao **método da máxima verosimilhança**. A função de verosimilhança é dada por $L(\boldsymbol{\beta}, \Sigma|(\mathbf{y}_i, x_i)) = f((y_1, \ldots, y_n)|\boldsymbol{\beta}, \Sigma) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\beta}, \Sigma)$. Consequentemente, a log-verosimilhança é dada por $\ell(\boldsymbol{\beta}, \Sigma|(\mathbf{y}_i, x_i)_i) = -\frac{nt}{2}\log(2\pi) - \frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta})$, onde $S_\Sigma(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta})$. Pretende-se maximizar ℓ que é equivalente a minimizar $S_\Sigma(\boldsymbol{\beta})$. Contudo, opta-se por outro procedimento mais simples.

Outra forma de obter o pretendido é **considerar o modelo básico** (3.10) onde os erros são homocedásticos ($\Sigma = \sigma^2\mathbf{I}$) e obter o estimador pretendido usado o estimador do modelo básico recorrendo à decomposição de Cholesky. No final obtém-se o seguinte vetor de dimensão p+1, $\widehat{\boldsymbol{\beta}}_{EMV} = \widehat{\boldsymbol{\beta}}_{MMQG} = \left(\sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i\right)\sum_{i=1}^{n} X_i^T\Sigma^{-1}\mathbf{y}_i$ que assume Σ conhecida.

Propriedades: Supondo Σ conhecida e fazendo $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{EMV} = \widehat{\boldsymbol{\beta}}_{MMQG}$

a) $E(\widehat{\boldsymbol{\beta}}|\Sigma) = \boldsymbol{\beta}$

b) $Cov(\widehat{\boldsymbol{\beta}}|\Sigma) = \left(\sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i\right)^{-1}$

c) Se $\mathbf{y}_i|X_i \sim MVN$ então $\boldsymbol{\beta}|\Sigma \sim MVN(\boldsymbol{\beta}, Cov(\widehat{\boldsymbol{\beta}}|\Sigma))$ - convergência exata.
Se $\mathbf{y}_i|X_i \nsim MVN$ então $\widehat{\boldsymbol{\beta}}_{MMQG}|\Sigma \overset{a}{\sim} MVN(\boldsymbol{\beta}, Cov(\widehat{\boldsymbol{\beta}}|\Sigma))$

d) O estimador de $\boldsymbol{\beta}$ mais eficiente (menor variância) é o que usa o valor verdadeiro de Σ.

Problema: Em geral, Σ é desconhecida!

Estratégia:

1. Estimar $\boldsymbol{\beta}$, supondo Σ conhecida. Obtém-se $\widehat{\boldsymbol{\beta}}_{EMV} = \widehat{\boldsymbol{\beta}}(\Sigma)$;
2. Determinar $\widehat{\Sigma}_{EMV}$ fazendo $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{EMV}$;
3. Tomar $\widehat{\boldsymbol{\beta}}(\widehat{\Sigma}) = \left(\sum_{i=1}^{n} X_i^T\widehat{\Sigma}^{-1}X_i\right)\sum_{i=1}^{n} X_i^T\widehat{\Sigma}^{-1}\mathbf{y}_i$.

Propriedades: Em amostras grandes ($n >> 0$), os estimadores $\widehat{\boldsymbol{\beta}}_{EMV}(\widehat{\Sigma})$ e $\widehat{\boldsymbol{\beta}}_{MMQG}(\widehat{\Sigma})$ verificam:

1. Supondo $\mathbf{y}_i|X_i \sim MVN$:

   a) $E(\widehat{\boldsymbol{\beta}}(\widehat{\Sigma})) = \boldsymbol{\beta}$

b) $Cov(\widehat{\boldsymbol{\beta}}(\widehat{\Sigma})) = \left(\sum_{i=1}^{n} X_i^T \widehat{\Sigma}^{-1} X_i\right)^{-1}$

c) $\widehat{\boldsymbol{\beta}}(\widehat{\Sigma}) \overset{a}{\sim} MVN(\boldsymbol{\beta}, Cov(\widehat{\boldsymbol{\beta}}(\widehat{\Sigma}))$

2. Se $\mathbf{y}_i|X_i \nsim MVN$ mas os dados não tem *missings* ou são incompletos (têm *missings*), pelo mecanismo MCAR (a ser apresentado em 3.2.2.2), as propriedades acima mantêm-se.

3.2.2.2 Mecanismo de valores em falta

É frequente ter de se trabalhar com dados que apresentam *missings*. Estes podem ser provenientes de vários mecanismos.

- **MCAR - *Missing completely at random***: a probabilidade de uma observação estar em falta não depende:

  i) do valor que deveria ter sido lido;

  ii) dos valores observados.

  Neste caso, $\widehat{\boldsymbol{\beta}}_{EMV}$ e $\widehat{\boldsymbol{\beta}}_{MMQG}$ têm as propriedades referidas em 3.2.2.1.

- **MAR - *Missing at random***: a probabilidade de uma observação estar em falta:

  i) depende do conjunto de dados observado;

  ii) não depende do valor que deveria ter sido lido.

  Neste caso, $\widehat{\boldsymbol{\beta}}_{EMV}$ e $\widehat{\boldsymbol{\beta}}_{MMQG}$ permitem fazer inferências válidas apenas se $\mathbf{y}_i|X_i \sim MVN$.

- **MNAR - *Missing not at random***: o facto de a observação estar em falta está relacionado com a razão pela qual está a faltar. Este é o pior dos cenários e aquele com o qual se deve ter bastante cuidado pois as propriedades dos estimadores não se verificam.

3.2.2.3 Restricted Maximum Likelihood Residual

Na estatística, a abordagem por REML (Wikipedia (2017)) é uma forma particular de estimação por máxima verosimilhança que fornece estimativas através do uso de uma função de verosimilhança que é obtida a partir de um conjunto de dados que é transformado de tal forma

que os parâmetros perturbadores[18] (como por exemplo a variância ou a média) não tenham qualquer efeito.

No caso da estimação da componente da variância, o conjunto de dados original é substituído por um conjunto de contrastes calculado a partir dos dados e a função de verosimilhança é calculada através da distribuição de probabilidade desses contrastes tendo em conta o modelo obtido usando o conjunto de dados completo. Em oposição à estimação por máxima verosimilhança, REML fornece estimativas centradas/não-enviesadas para os parâmetros da variância e covariância.

Considere-se $\mathbf{Y} = \widehat{\mathbf{Y}} + \widehat{\mathbf{u}} = X\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}$, onde $\widehat{\mathbf{u}}$ representa os resíduos. A projeção sobre $< X >$ é: $X(X^TX)^{-1}X^T\mathbf{Y} = X\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$. Pretende-se escolher $K \in \mathbb{R}^{n\times(n-(p+1))}$ tal que $K^TX = 0$. Por exemplo, $K^T = \mathbf{I} - X(X^TX)^{-1}X^T$ (o que resta da projeção sobre $< X >$) pois $K^TX = X - X(X^TX)^{-1}X^TX = 0$.

Tem-se que $K^T\mathbf{Y} = K^T(X\boldsymbol{\beta} + \mathbf{u}) = K^T\mathbf{u} = \widehat{\mathbf{u}} \sim N(0, K^T\Sigma K)$, onde $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma)$. Destaca-se então o seguinte

1. $K^T\mathbf{Y}$ tem uma distribuição que não depende dos β's;

2. Usando MV para os dados $K^TY_1, \ldots, K^TY_n$ obtém uma estimativa para Σ designada por $\widehat{\Sigma}_{REML}$;

3. Usando MMQG=EMV obtém-se uma estimativa para $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}}_{REML} = (X^T\widehat{\Sigma}^{-1}_{REML}X)^{-1}(X^T\widehat{\Sigma}^{-1}_{REML}\mathbf{Y})$$

Nota: Escrever $\widehat{\boldsymbol{\beta}}_{REML}$ é abuso de notação porque $\widehat{\boldsymbol{\beta}}_{REML}$ não é obtido de REML.

Assim sendo, neste tipo de modelos, pode-se usar estimação por ML e por REML sendo que, neste último caso, o desvio-padrão para as estimativas é maior. Neste caso, é mais difícil obter a significância dos parâmetros uma vez que $\boldsymbol{\beta}$ é tanto mais significativo quanto menor for o seu desvio-padrão.

[18] Nuisance parameters.

3.2.2.4 Decomposição da Matriz de Var-Cov dos Erros

Considere-se o modelo linear geral, (3.16), após decomposição da estrutura aleatória:

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{u}_i, \quad \mathbf{u}_i \sim MVN(0, \sigma^2 M_i), \quad i = 1, \ldots, n.$$

Note-se que os modelos são iguais, apenas se decompôs a matriz $\Sigma_i = \sigma^2 M_i$.

A matriz M_i pode ser decomposta num produto de matrizes mais simples: $M_i = W_i C_i W_i$ onde se mostra que

1. $Var(u_{it}) = \sigma^2 (W_i)^2_{tt}$;
2. $Corr(u_{it}, u_{it'}) = (C_i)_{tt'}$.

e, portanto,

- W_i é uma matriz diagonal com desvios-padrão na diagonal (e zeros nas restantes entradas) que descreve a variância dos erros da unidade experimental i;
- C_i é uma matriz de correlação (semi-definida positiva) com diagonal unitária que descreve a correlação entre os erros da unidade experimental i.

Esta decomposição permite a modelação separada da estrutura de variância e da estrutura de correlação dos erros de uma unidade experimental que é bastante útil na aplicação prática.

Mais detalhes sobre este assunto podem ser encontrados no Anexo 2.

3.2.2.5 Testes de hipóteses e intervalos de confiança sobre β_k's

Depois de obtidas as estimativas para o vetor $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ é usual fazerem-se testes de hipóteses e determinarem-se intervalos de confiança sobre os β_k's. Pode-se fazer isso das seguintes formas

a) **Sobre $\boldsymbol{\beta_k}$**: aplica-se o teste de Wald. As hipóteses de aplicação são: $\widehat{\beta}_k \overset{\text{aprox.}}{\sim}$ Normal e testa-se o seguinte $\begin{cases} H_0\text{: } \beta_k = \beta_k^{(0)} \\ H_1\text{: } \beta_k \neq \beta_k^{(0)} \end{cases}$ onde, normalmente, $\beta_k^{(0)} = 0$. A estatística de teste usada é $Z = \dfrac{\widehat{\beta}_k}{se(\widehat{\beta}_k)} = \dfrac{\widehat{\beta}_k}{\sqrt{\widehat{Cov(\widehat{\boldsymbol{\beta}})_{kk}}}} \sim N(0,1)$, onde *se* corresponde ao desvio padrão.

Assim, rejeita-se H_0 se o valor da estatística de teste for igual ou superior ao quantil $1-\frac{\alpha}{2}$ da distribuição normal reduzida. Por fim, o intervalo de confiança a $100(1-\alpha)\%$ para β_k é dado por $\widehat{\beta}_k \pm N_{1-\frac{\alpha}{2}}(0,1) \times se(\widehat{\beta}_k)$.

b) **Sobre combinações lineares de $\boldsymbol{\beta_k}$'s**: Consoante a combinação linear usada, define-se L que pode ser uma matriz ou um vetor de pesos conhecidos (contrastes)[19]. Contudo, em ambos os casos pretende-se testar o seguinte $\begin{cases} H_0\colon\ L\boldsymbol{\beta} = \mathbf{0} \\ H_1\colon\ L\boldsymbol{\beta} \neq \mathbf{0} \end{cases}$.

Caso 1: L é um vetor (linha).

Nesta situação, tem-se que $L\widehat{\boldsymbol{\beta}} \overset{a}{\sim} MVN(L\boldsymbol{\beta}, L\widehat{Cov(\widehat{\boldsymbol{\beta}})}L^T)$ o que dá origem a estatística de teste de Wald usada $Z = \dfrac{L\widehat{\boldsymbol{\beta}}}{se(L\widehat{\boldsymbol{\beta}})} \sim N(0,1) \Rightarrow Z^2 \sim \chi^2(1)$. Mais uma vez, rejeita-se H_0 se o valor da estatística de teste for igual ou superior ao quantil $1-\frac{\alpha}{2}$ da distribuição normal reduzida.

Caso 2: L é uma matriz.

Nesta situação, a estatística de teste de Wald usada é $Z^2 = (L\widehat{\boldsymbol{\beta}})^T(L\widehat{Cov(\widehat{\beta})}L^T)^{-1}(L\widehat{\boldsymbol{\beta}}) \sim \chi^2(r)$, onde $L \in \mathbb{R}^{r\times(p+1)}$, ou seja, r é o número de linhas da matriz L. Rejeita-se H_0 se o valor da estatística de teste for igual ou superior ao quantil $1-\frac{\alpha}{2}$ da distribuição $\chi^2(r)$.

Exemplos sobre assunto podem ser encontrados no Anexo 3. Conforme se pode ver através de um desses exemplos, a formulação do caso 1 coincide com a formulação feita em a) se se considerar um único β_k.

c) **Teste da razão de verosimilhanças**[20]: usado para amostras pequenas (coincide com o teste de Wald para amostras grandes) pois, nesta situação, os testes de Wald não têm convergência definida, daí ser preferível usar o teste da razão de verosimilhanças. Considerem-se dois modelos M_1 e M_2, que usam exatamente o mesmo conjunto de dados. Para além disso, considere-se que estes modelos são encaixados/aninhados[21], ou seja, $M_1 \subset M_2 \Leftrightarrow \{\text{parâmetros } M_1\} \subset \{\text{parâmetros } M_2\}$.

[19] *Constrast matrix.*

[20] *Likelihood-ratio test.*

[21] *Nested models.*

Pretende-se testar

$$\begin{cases} H_0\text{: qualidade do ajustamento de } M_1 = \text{ qualidade do ajustamento de } M_2 \\ H_1\text{: qualidade do ajustamento de } M_1 \neq \text{ qualidade do ajustamento de } M_2 \end{cases}.$$

A estatística de teste é $2\ln\left(\frac{L_2}{L_1}\right) = 2(\ell_2 - \ell_1) \sim \chi^2(n_2 - n_1)$, onde n_1 e n_2 representam o número de parâmetros de M_1 e M_2, respetivamente, e L_1, L_2, ℓ_1, ℓ_2 as funções de verosimilhança e log-verosimilhança associadas a estes modelos. Assim sendo, rejeita-se H_0 se o valor desta estatística de teste for igual ou superior ao quantil $1-\frac{\alpha}{2}$ da distribuição $\chi^2(n_2 - n_1)$ e, neste caso, retém-se M_2. Caso não se rejeite H_0, retém-se M_1, usando-se, assim, o princípio da parcimónia.

Nota: Para amostras consideradas grandes, o teste também pode ser aplicado. Nesta situação tem-se $\begin{cases} H_0\text{: } \beta_k = 0 \\ H_1\text{: } \beta_k \neq 0 \end{cases}$ " $\Leftrightarrow$ " $\begin{cases} H_0\text{: [Modelo onde } \beta_k = 0] = \text{[Modelo com todos os } \beta_k's] \\ H_1\text{: } \ldots \end{cases}$

Veja-se agora como **obter intervalos de confiança** para esta situação. São intervalos de confiança baseados na verosimilhança de perfil[22]. Suponha-se que se pretende encontrar um intervalos de confiança a $100(1-\alpha)\%$ para β_k. Deve adotar-se o seguinte procedimento

1. Fixar β_k;
2. Determinar expressões para $(\widehat{\beta}_0, \ldots, \widehat{\beta}_{k-1}, \widehat{\beta}_{k+1}, \ldots, \widehat{\beta}_p)$ que, claramente, virão em função de β_k;
3. A verosimilhança de perfil[23] é uma função de β_k dada por

$$\ell_p(\beta_k) = \ell\left(\widehat{\beta}_0(\beta_k), \ldots, \widehat{\beta}_{k-1}(\beta_k), \beta_k, \widehat{\beta}_{k+1}(\beta_k), \ldots, \widehat{\beta}_p(\beta_k)\right)$$

Assim sendo, o $\text{IC}_{100(1-\alpha)\%}$ para β_k baseado na verosimilhança de perfil é

$$\{\beta_k \in \mathbb{R} : 2\left(\ell_p(\widehat{\beta}_k) - \ell_p(\beta_k)\right) \leq \chi^2_{1-\alpha}(1)\}.$$

Observação: Usar em amostras pequenas/respostas discretas.

[22] *Profile-likelihood confidence intervals.*

[23] *Profile log-likelihood.*

3.2.2.6 Diagnóstico e Análise de Resíduos

Há várias condições a avaliar:

1. Os erros aleatórios dentro do grupo, $\mathbf{u}_i$, são identicamente distribuídos com uma distribuição normal multivariada de média 0 (e independentes e com variância constante, caso se esteja no modelo linear misto básico);
2. Os erros aleatórios são independentes dos efeitos aleatórios;
3. Em grupos diferentes, os resíduos são independentes.

A verificação dos pressupostos é feita graficamente[24]. Para além disso, destaca-se o seguinte:

- a avaliação qualitativa das condições sobre os erros $\mathbf{u}_i$ usa os resíduos $\mathbf{r}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$ por constituírem boas estimativas dos erros. Os vetores $\mathbf{u}_i$ e $\mathbf{r}_i$ têm ambos média 0 mas as matrizes de var-cov são diferentes (mas semelhantes);
- Verbeke e Molenberghs (2000):
  - a avaliação deve ser feita num momento intermédio do processo de escolha do modelo e, eventualmente, repetida na fase final;
  - na avaliação intermédia, é conveniente considerar um modelo linear misto básico ($\mathbf{u}_i \sim MVN(0, \sigma^2\mathbf{I})$) de forma a extrair informações sobre a estrutura da matriz var-cov dos erros.

Condições sobre os erros

Apresentam-se os gráficos a considerar[25] juntamente com algumas considerações relativamente aos mesmos.

1. **Gráfico dos resíduos estandardizados contra os valores ajustados**
   - Quaisquer tendências devem ser tomadas em consideração na modelação;
   - No modelo final, não devem existir tendências;
   - Permite a identificação de *outliers*.

[24] Designados por gráficos de diagnóstico.

[25] A análise gráfica a seguir descrita foi realizada na âmbito desta dissertação apesar de não ser aqui apresentada.

2. **Boxplot dos resíduos estandardizados por grupo**

   - Avaliar se os resíduos dentro do grupo são centrados em 0, independentes dos grupos e homocedásticos;
   - Havendo violação da homocedasticidade, modelar a variância dos erros usando estruturas de variâncias adequadas. Uma vez corrigida essa estrutura, o gráfico deve deixar de apresentar problemas;
   - Permite a identificação de *outliers*.

3. **Gráfico dos resíduos estandardizados contra variáveis explicativas**

   - Não deve exibir tendências. Se necessário, corrigir introduzindo termo quadrático ou transformação de variável.

4. **Gráfico dos valores observados contra os valores ajustados**

5. **qq-plot ou pp-plot dos resíduos**

   - Avalia a normalidade dos resíduos;
   - Simetria na distribuição indica que as estimativas para os efeitos fixos não devem mudar muito caso a distribuição seja alterada (por exemplo para uma mistura de normais ou uma t);
   - Se necessário, transformar a resposta;
   - Se necessário, complementar a análise com histograma e/ou boxplot.

3.2.2.7 Comparação entre modelos

Para além de ter de se usar o mesmo conjunto de dados, salientam-se outros aspetos relativamente à comparação de modelos.

1. ML fornece $\widehat{\beta}_{ML}$ (efeitos fixos). REML não (lembrar que escrever $\widehat{\beta}_{REML}$ é abuso de notação);

2. Na situação em que $M_1 \subset M_2$ (modelos encaixados) com diferentes β's deve-se usar ML para comparar estes dois modelos e não REML;

3. REML pode ser usado para comparar modelos com os mesmo β's mas com diferentes estruturas de var-cov[26];

4. Habitualmente, ML e REML produzem estimativas semelhantes. A diferença diminui quando o tamanho amostral é substancialmente maior do que p (número de variáveis explicativas). Na regressão linear ordinária ($\Sigma = \sigma^2\mathbf{I}$), $\widehat{\sigma}^2_{REML} = \dfrac{RSS}{n-(p+1)}$ e $\widehat{\sigma}^2_{ML} = \dfrac{RSS}{n}$;

5. Não comparar modelos estimados por ML com modelos estimados por REML, isto é, não usar um teste da razão de verosimilhanças nessa situação;

6. **Teste da razão de verosimilhanças para comparar modelos com estrutura de var-cov encaixadas** (e mesma estrutura fixa):
   Regra de Fitzmaurice: usar $\alpha = 0.1$ em vez de $\alpha = 0.05$ no teste da razão de verosimilhanças usual (definido em 3.2.2.5).

7. Nas restantes situações devem usar-se critérios de informação.

[26] Este tópico será abordado com mais detalhe mais à frente.

3.3 Net Promotor Score® (NPS)

Segue-se agora a apresentação de um método de análise de satisfação de clientes e, portanto, usado na realização deste trabalho. O que vai ser exposto nesta secção é baseado num documento feito e fornecido por Patrícia Araújo (supervisora de estágio), obtido após a consulta de Reichheld e Markey (2011).

3.3.1 O que é?

Segundo Frederick F. Reichhedl (Reichheld (2003)), o *NPS* (2017), é um indicador utilizado para medir o grau de satisfação e lealdade dos clientes com a organização, através de uma pergunta única e que permite a comparabilidade do nível da satisfação global com outras empresas do setor.

Parece simples, mas permite *insights* muito poderosos sobre o comportamento do cliente e quais as melhores ações a tomar pelo negócio, com vista à sua transformação.

3.3.2 Como se calcula?

Através dos inquéritos de satisfação, é possível obter a avaliação e opinião dos clientes sobre vários aspetos da empresa. Na utilização do NPS, o grau de satisfação, para dada empresa do grupo, é obtido com base na resposta à pergunta: **de 1 a 10, recomendaria a um/a amigo/a...?**.

Em função da resposta[27] obtida, é possível agrupar os clientes em 3 categorias:

Figura 3.2: Categorização do NPS. Fonte: *Documento interno: Net Promoter Score* (2018).

[27] Valores inteiros entre 1 e 10, onde 1 é o nível mais baixo e 10 o nível mais alto (escala ordinal - Likert).

Esta categorização permite a recolha de dados racionais e emocionais de forma simples e imediata através de um pequeno calculo efetuado entre a percentagem de promotores e detratores.

O resultado global[28] obtém-se calculando:

$$\boxed{\%Promotores} - \boxed{\%Detratores} = NPS$$

Os resultados obtidos permitem aferir o nível de satisfação dos clientes, quer global, quer em detalhe, em função da sua categorização (por exemplo, por área geográfica). Para compreender os motivos que justificam os resultados obtidos com o NPS, é possível a sua utilização conjunta com outras questões.

3.3.3 Propriedades

Observando a forma como o NPS é obtido, destacam-se os seguintes aspetos:

- NPS$\in [-100, 100]$, ou seja, varia de -100 a 100;
- Nos casos em que se tem NPS$= -100$ ou NPS$= 100$ isso significa que todos os clientes são detratores (o que não se pretende, pois é o pior cenário possível) ou promotores (o que seria o ideal, pois é o melhor cenário possível), respetivamente.
- No caso de se ter NPS ≤ 0 isso significa que a percentagem de detratores (consequentemente, o seu número) é maior ou igual à percentagem/número de promotores. Este é um dos piores cenários possíveis e não é de todo o pretendido pelo grupo;
- No caso de se ter NPS > 0 isso significa que a percentagem de promotores (consequentemente, o seu número) é maior do que a percentagem/número de detratores. De facto, este é o panorama que se pretende obter no cálculo do NPS. Contudo, **pretende-se NPS$\gg 0$**, ou seja, **o mais próximo possível de 100.**

Nota:

Não existe nenhum valor a partir do qual se considere que o valor do NPS é bom. Define-se que este valor é aceitável por comparação com um resultado de NPS "mais global".

[28] Para um dado conjunto de dados.

Por exemplo, se o NPS calculado tendo em conta todas as empresas do grupo - NPS global - é de 50 e uma dada empresa apresenta NPS= 60, considera-se que este resultado é bom, por comparação. Se se obtivesse NPS= 40 em vez de NPS= 60 aí o resultado já não seria satisfatório.

Quando se compara o NPS de duas empresas distintas, claramente que a que tiver maior NPS será a que apresenta melhor resultado.

Para além disso, existem plataformas onde é possível consultar o NPS de várias empresas (*NPS Benchmarks*). Com isto, é possível obter o NPS de uma empresa concorrente a uma empresa do grupo e efetuar a respetiva comparação. Permite também saber que valor de NPS atingir para superar a concorrência. Nesta situação é preciso ter especial atenção à plataforma onde se consultam estes resultados. Não está garantido que as empresas que fornecem os seus dados às plataformas usam todas o mesmo esquema de amostragem nem está garantido que estas não enviesem os dados de forma a terem valores altos de NPS. Por isso, antes de se consultar o NPS de empresas concorrentes neste tipo de plataformas convém ler os termos e as condições das mesmas de forma a verificar se os resultados apresentados são viáveis ou não. Por exemplo, a plataforma *NPS Benchmarks*, apesar de fornecer o NPS de várias empresas, refere nos seus termos e condições que não garante a confiabilidade dos resultados apresentados.

Duma forma geral, o NPS permite a comparação entre empresas do grupo e entre empresas concorrentes (ter em conta o que se disse anteriormente sobre este tema).

Capítulo 4.
Resultados

"Why does he insist that we must have a diagnosis? Some things are not meant to be known by man."

Susanna Gregory - An Unholy Alliance

O presente capítulo apresenta as principais conclusões e resultados "inferidos" de vários inquéritos realizados a clientes de 5 empresas do grupo NORS no decurso dos anos de 2013 a 2017. Por questões de confidencialidade, as empresas estudadas serão referidas como Empresa λ, onde λ =A, B, C, D e E.

4.1 Dados e análises desenvolvidas

Nesta secção apresentam-se os dados recolhidos para análise da satisfação dos clientes do grupo NORS assim como a metodologia usada. Os dados disponíveis foram recolhidos através de 3 inquéritos diferentes. Em todos os momentos de recolha dos dados:

- o Universo em estudo foi composto pelos clientes com compras no período em estudo;
- foi considerado um esquema de amostragem aleatória simples;
- as entrevistas foram realizadas via telefone, apoiadas em questionário estruturado de perguntas abertas e fechadas, inserido num programa informático (C.A.T.I. - *Computer-Assisted Telephone Interviewing*) gestor das entrevistas.

Ilustra-se na Figura 4.1 um esquema que descreve resumidamente a estrutura do questionário, o período de recolha e o tamanho amostral.

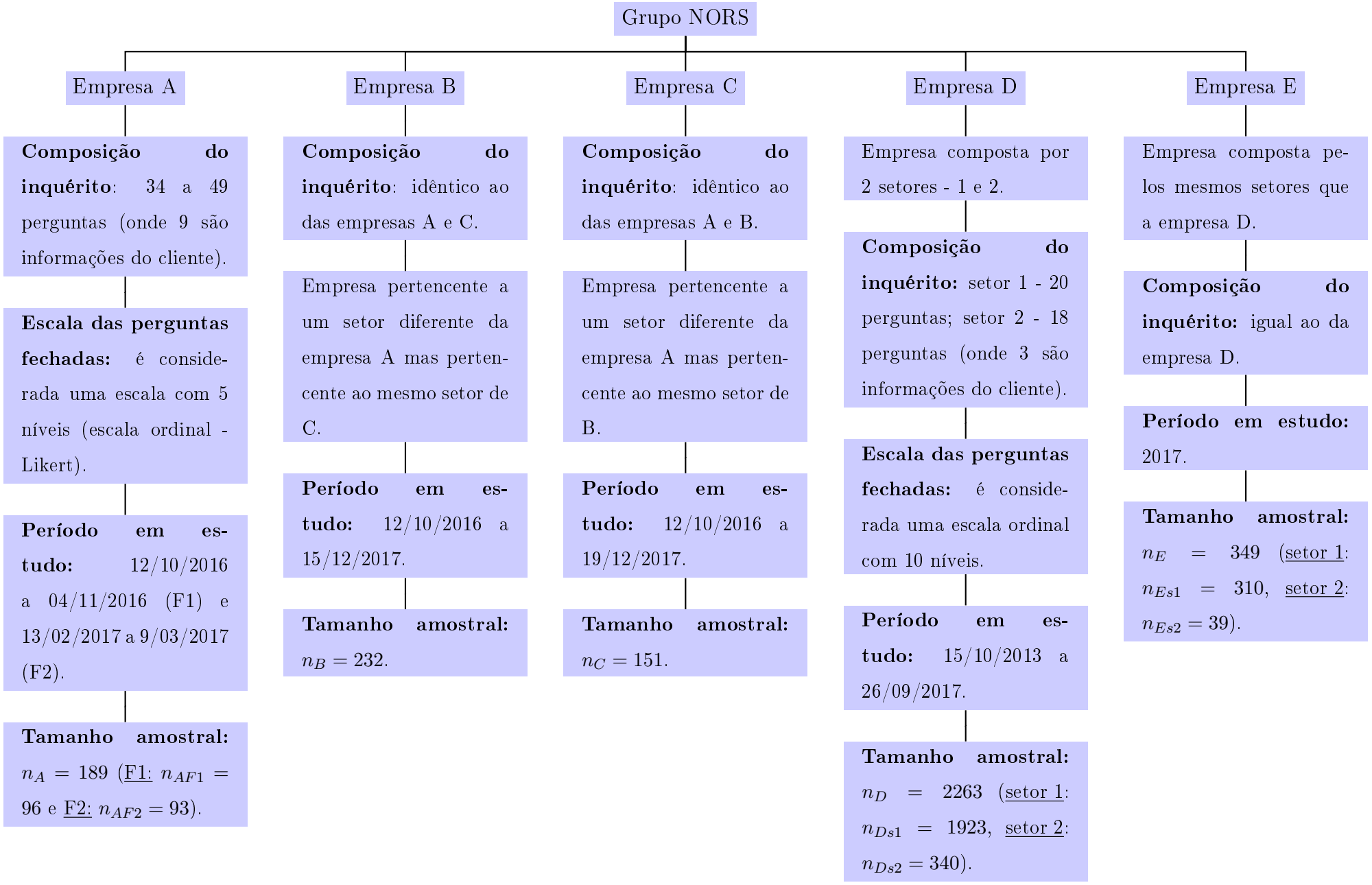


Figura 4.1: Sobre os dados.

Na empresa A (consequentemente nas empresas B e C), o número de questões respondidas não é exato visto que há dependências entre elas, ou seja, só é feita uma dada pergunta ao cliente se este tiver respondido afirmativamente numa questão colocada anteriormente.

Conforme se pode observar, pela Figura 4.1, as empresas E e F apresentam 2 setores distintos: setor 1 e setor 2. Desta forma, os resultados dos inquéritos têm de ser analisados separadamente.

Note-se, também, que o período de recolha não é igual em todas as empresas. Para além disso, as análises desenvolvidas não foram as mesmas. De forma a esquematizar as análises executadas, para cada caso específico, foi construído o esquema apresentado na Figura 4.2.

4.2 Apresentação dos resultados obtidos

Por questões de confidencialidade e a pedido do grupo NORS, algumas das análises feitas e referidas na Figura 4.2 não serão aqui apresentadas. As análises não apresentadas são aquelas que envolvem a divulgação da localização das lojas, nome dos concessionários e nome dos revendedores.

4.2.1 Apontamento metodológico

Em contexto empresarial é frequente sinalizar, como indicadores com prioridade de intervenção para efeitos de melhoria, indicadores com um nível de satisfação médio inferior a 85 (numa escala de 0 a 100).

Tal como já referido, a escala das questões não é a mesma em todos os questionários. No sentido de uniformizar os resultados, juntamente com a escala original dos dados, apresenta-se a conversão para a escala de 100 pontos.

Neste estudo foi dado particular destaque à questão *Utilizando uma escala de 1 a m, em que 1 significa "Não recomendaria com toda a certeza" e m "Recomendaria com toda a certeza", em que medida diria que recomenda os produtos e serviços da empresa (...) a um amigo/colega?* ($m = 5$ ou $m = 10$, consoante a empresa em questão). De forma a simplificar a referência a esta variável, considere-se a designação: Recomendar/Recomendação.

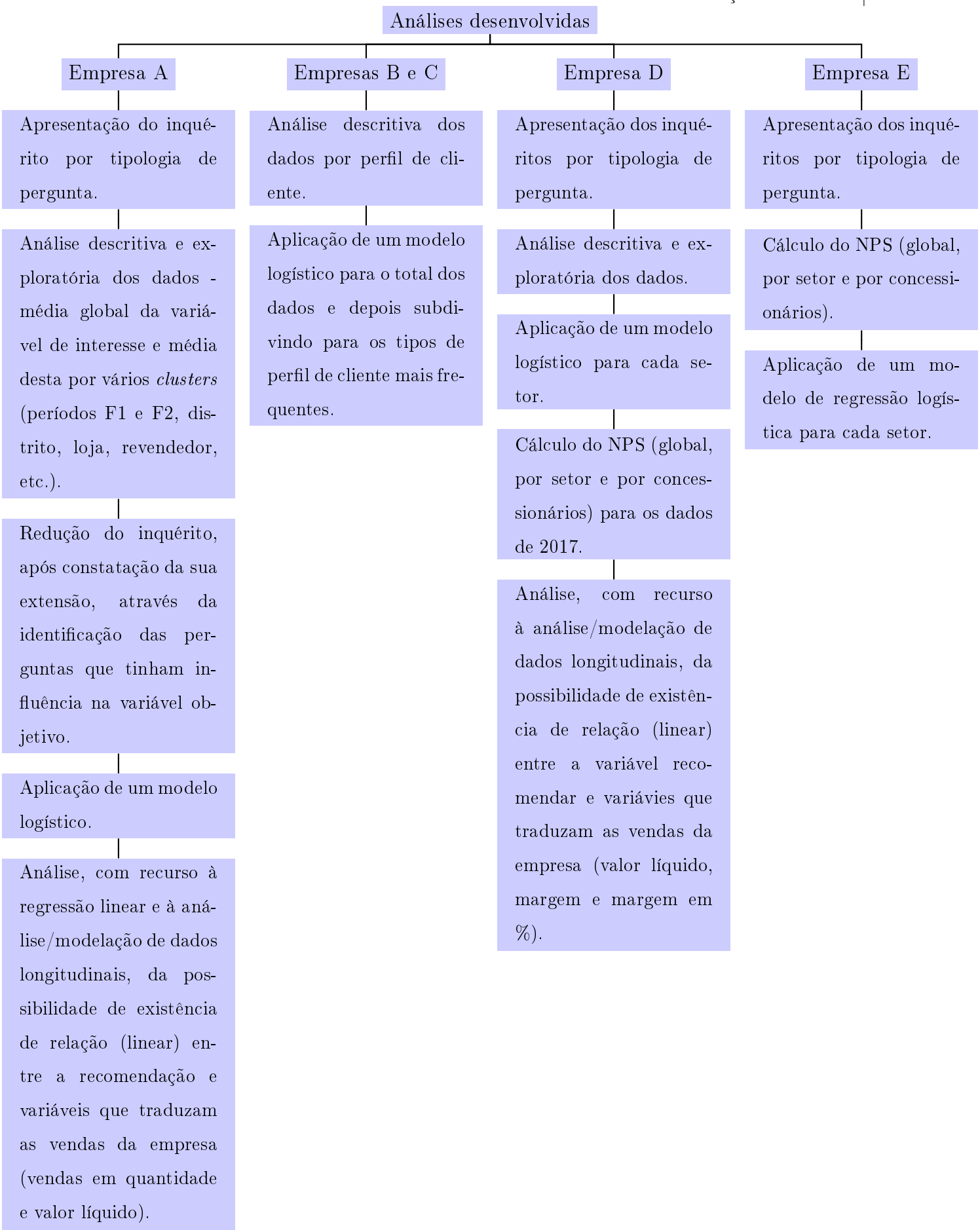


Figura 4.2: Análises desenvolvidas para cada empresa do grupo NORS.

Para a aplicação da regressão logística, é necessário uma variável (resposta) binária.

Para as empresas A, B e C, para ir de encontro ao objetivo pretendido, categorizou-se a variável inicial (Recomendação) da seguinte forma: Recomendação= $\begin{cases} 1, \text{ se Recomendação=5 (promotor)} \\ 0, \text{ caso contrário (não promotor)} \end{cases}$.
Para além disso, nas questões fechadas do inquérito, usou-se a mesma categorização onde a categoria 0 será a classe de referência. Em algumas variáveis teve de se introduzir a categoria "Não se aplica" pois a pergunta feita não se aplicava a estes, por exemplo, se estes nunca efetuaram reclamações não faria sentido perguntar a sua satisfação na resolução com as reclamações.

Para as empresas D e E a variável inicial, Recomendar, será transformada da seguinte forma: Recomendar = $\begin{cases} 1, \text{ se Recomendar=9 ou 10 (promotor)} \\ 0, \text{ caso contrário (não promotor)} \end{cases}$. Para além disso, nas questões fechadas do inquérito, usou-se a categorização do NPS (ver Figura 3.2), onde a categoria 1 (detrator) é a classe de referência. **Nota:** Sentiu-se a necessidade, após análise das mesmas, de agrupar algumas das variáveis em 2 categorias em vez de 3. Agregou-se a categoria 1 com a 2 (detratores e neutros, respetivamente), ficando assim com uma variável com 2 categorias (não promotor(2)/promotor(3)).

Em todas as empresas, relativamente à regressão logística, para facilitar a leitura/interpretação dos resultados do modelo escolhido, considere-se que o sucesso se refere ao facto dos clientes serem promotores na variável objetivo (recomendação/recomendar).

4.2.2 Empresa A

O questionário encontra-se estruturado em 13 grupos. Nesta secção, apresentam-se os resultados obtidos através do mesmo.

4.2.2.1 Apresentação do inquérito por tipologia de pergunta

Nota: Na Tabela 4.1, a coluna: **Número de perguntas/campos** não é, em alguns casos, um valor único, mas sim um intervalo de valores. Isso deve-se ao facto de que certas questões só são colocadas, se se obtiver determinada resposta na questão anterior.

Tabela 4.1: Apresentação do inquérito da empresa A por tipologia de pergunta.

Tipologia	**Número de perguntas/campos**
Dados para o contacto	9
Serviço e qualidade do mesmo	5
Meios de contato	2 a 8
Logística	3 ou 4
Devoluções	1 ou 2
Produto	3 ou 4
Condições comerciais	1 ou 2
Serviço online	2 ou 3
Marca própria de peças do Grupo NORS	2 ou 3
Campanhas	1 ou 3
Conhecimento das novidades	1
Outros fornecedores	3 ou 4
Confidencialidade	1

4.2.2.2 Análise descritiva e exploratória dos dados

Veja-se agora a distribuição da variável resposta (antes de ser categorizada em não promotor/promotor) pelos vários *clusters*.

- **Distribuição por períodos de tempo**

Tabela 4.2: Período F1.

Classificação	1	2	3	4	5	*NA*
Frequência Absoluta	0	1	11	32	49	3
Frequência Relativa	0%	11.1%	11.5%	33.3%	51.0%	3.1%

Tabela 4.3: Período F2.

Classificação	1	2	3	4	5
Frequência Absoluta	6	4	11	34	38
Frequência Relativa	6.4%	4.3%	11.8%	36.6%	40.9%

Perante estes resultados, observa-se que, que 3 clientes não deram classificação à pergunta em questão e que, em ambos os períodos, a classificação mais frequente é 5, o que é algo bom. Contudo o ideal era esta frequência ser 100%.

- **Média da variável resposta**

  Para se ter uma melhorar ideia se estes resultados são bons ou não observam-se os valores médios[1] desta variável.

Tabela 4.4: Média da variável resposta da empresa A.

Escala	F1	F2	Global
5 pontos	4.4	4.0	4.2
100 pontos	87.7	80.2	84

Conclusão: É possível constatar que no 1ª trimestre de 2017 (F2) e globalmente, a pontuação dada pelos clientes, em média, não satisfaz os requisitos estabelecidos. Isto é indicativo de que há algo que não está a correr bem e que é necessário efetuar uma análise mais detalhada, de forma a se conseguir melhorar estes resultados aumentando a satisfação do cliente.

De forma a tentar obter alguma informação sobre a origem de uma classificação média inferior ao pretendido, na variável de estudo, serão feitas mais análises a seguir descritas.

- **Análise por código do segmento**

  Na BD está presente uma variável designada por: **Código do Segmento** que pode tomar um dos seguintes níveis: C0, C1, C2, C3, C4 ou C5 e que indica o nível de importância de um dado cliente. Assim, a categoria C0 é um código temporário sujeito a alteração após avaliação comercial do cliente, a C1 representa os clientes menos importantes e a categoria C5 os mais importantes pois são os que mais investem/compram, ao contrário dos outros.

  Obteve-se a seguinte distribuição do código do segmento nos vários períodos[2]:

Tabela 4.5: Período F1.

Código	C0	C1	C2	C3	C4	C5	*NA*
Frequência Absoluta	7	19	19	22	18	5	6
Frequência Relativa	7.3%	19.8%	19.8%	22.9%	18.8%	5.2%	6.2%

[1] Estes valores foram calculados usando todos os valores da amostra (como existiam 3 *missing values*, o tamanho amostral é 186 (93 em cada período)).

[2] A última coluna representa os *missing values*.

Tabela 4.6: Período F2.

Código	C0	C1	C2	C3	C4	C5	*NA*
Frequência Absoluta	4	14	27	24	18	5	1
Frequência Relativa	4.3%	15.1%	29%	25.8%	19.4%	5.4%	1.1%

O resultado global é o seguinte

Tabela 4.7: Frequência absoluta e relativa do código do segmento da empresa A.

Código	C0	C1	C2	C3	C4	C5	*NA*
Frequência Absoluta	11	33	46	46	36	10	7
Frequência Relativa	5.8%	17.5%	24.3%	24.3%	19.1%	5.3%	3.7%

Conclusão: Pela tabela anterior, é possível constatar que não foi atribuído nenhum código de segmento a 7 clientes. Por outro lado, é possível observar que os resultados centram-se nas categorias C1,C2,C3 e C4, havendo poucos clientes com os códigos C0 e C5 (cerca de 11% da amostra total).

Com o objetivo de tentar perceber o comportamento dos clientes dentro de cada código, observa-se as classificações dadas dentro de cada grupo:

Tabela 4.8: Tabela de contingência entre o código do segmento (CS) e a classificação dada na variável de estudo, para a empresa A.

		Níveis da resposta					
		1	2	3	4	5	*NA*
CS	C0			1	2	8	
	C1	1		6	9	17	
	C2	3	3	4	15	20	1
	C3	1	2	4	16	22	1
	C4	1		6	15	14	
	C5				7	3	
	NA			1	2	3	1

É possível constatar, entre várias coisas, que:

- 2 clientes, cujo código do segmento é C2 e C3 respetivamente, não responderam à questão relativa à variável de estudo;

– 6 clientes que não possuem código do segmento, responderam à pergunta da variável de estudo;

– 1 cliente não possui código do segmento nem respondeu à pergunta em questão;

– Os clientes cujo código é C5, são praticamente todos não promotores;

– Os clientes dos códigos C0 a C4, praticamente, respondem 4 ou 5 na recomendação.

Em média, os resultados obtidos foram os seguintes:

Tabela 4.9: Cálculo da média, numa escala de 100 pontos, da variável de estudo da empresa A por código do segmento, globalmente e por período de tempo.

Código	**Tamanho amostral (n)**	**F1**	**F2**	**Global**
C0	7+4=11	91.4	95	92.7
C1	19+14=33	88.4	80	84.8
C2	19+27=46	87.8	75.6	80.4
C3	22+24=36	88.6	81.7	84.9
C4	18+18=36	84.4	81.1	82.8
C5	5+5=10	88	84	86
NA	6+1+7	88	80	86.7

Conclusões:

– em média, no código C0 obtiveram-se resultados satisfatórios em ambos os períodos e, consequentemente, no geral;

– em média, nos restantes códigos, houve, pelo menos um período (e globalmente também) em que os resultados não foram satisfatórios.

– o código do segmento onde os resultados foram mais e menos satisfatórios por período (e globalmente também), são mostrados na tabela seguinte.[3]

Resultados...	**F1**	**F2**	**Global**
...menos satisfatórios	C4	C2	C2
...mais satisfatórios	C3	[4]	C5

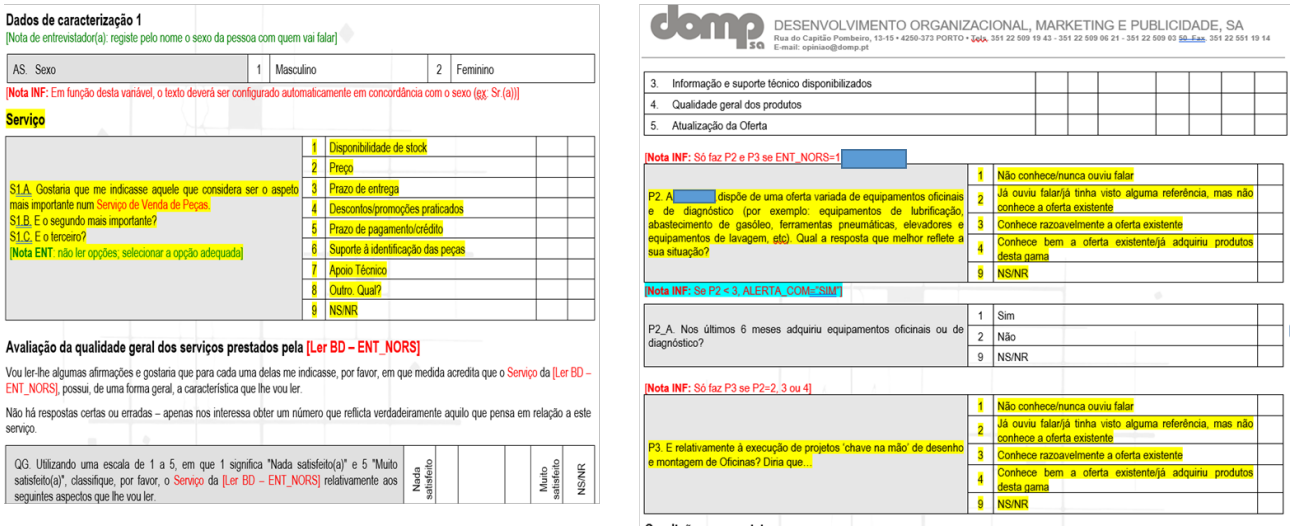
[3] Visto que C0 é um código temporário, este não foi considerado.

[4] Para além dos clientes com código C0, nenhum outro apresenta resultados satisfatórios, neste período.

4.2.2.3 Redução do inquérito

Conforme já se verificou, o inquérito desta empresa é bastante extenso. Sentiu-se a necessidade de o tentar reduzir baseado num critério científico. Para isso, efetuaram-se testes de independência (descritos em 3.1.9.3), onde se testou, individualmente, a hipótese de cada uma das variáveis explicativas ser independente da variável resposta. Desta forma, todas as variáveis para as quais não se rejeite a hipótese nula serão assinaladas para que o grupo NORS reflita (consoante as suas intenções com o questionário) sobre a sua retirada ou não do referido inquérito.

A Figura 4.3 exemplifica o que foi aqui descrito. Apresentou-se ao grupo NORS o questionário da Empresa A onde se assinalou a amarelo as questões eventualmente a ser retiradas do inquérito por não existir evidência estatística suficiente para concluir que a referida questão é não independente da resposta.



Dados de caracterização 1

[Nota de entrevistador(a): registe pelo nome o sexo da pessoa com quem vai falar]

AS. Sexo	1	Masculino	2	Feminino

[Nota INF: Em função desta variável, o texto deverá ser configurado automaticamente em concordância com o sexo (eg. Sr (a))]

Serviço

S1.A. Gostaria que me indicasse aquele que considera ser o aspeto mais importante num Serviço de Venda de Peças. S1.B. E o segundo mais importante? S1.C. E o terceiro? [Nota ENT: não ler opções, selecionar a opção adequada]	1	Disponibilidade de stock					
	2	Preço					
	3	Prazo de entrega					
	4	Descontos/promoções praticados					
	5	Prazo de pagamento/crédito					
	6	Suporte à identificação das peças					
	7	Apoio Técnico					
	8	Outro. Qual?					
	9	NS/NR					

Avaliação da qualidade geral dos serviços prestados pela [Ler BD – ENT_NORS]

Vou ler-lhe algumas afirmações e gostaria que para cada uma delas me indicasse, por favor, em que medida acredita que o Serviço da [Ler BD – ENT_NORS], possui, de uma forma geral, a característica que lhe vou ler.

Não há respostas certas ou erradas – apenas nos interessa obter um número que reflicta verdadeiramente aquilo que pensa em relação a este serviço.

	Nada satisfeito				Muito satisfeito	NS/NR
QG. Utilizando uma escala de 1 a 5, em que 1 significa "Nada satisfeito(a)" e 5 "Muito satisfeito(a)", classifique, por favor, o Serviço da [Ler BD – ENT_NORS] relativamente aos seguintes aspectos que lhe vou ler.						

dompsa DESENVOLVIMENTO ORGANIZACIONAL, MARKETING E PUBLICIDADE, SA

Rua do Capitão Pombeiro, 13-15 • 4250-373 PORTO • ~~Tel~~. 351 22 509 19 43 - 351 22 509 06 21 - 351 22 509 03 ~~50 Fax~~ 351 22 551 19 14

E-mail: opiniao@domp.pt

3. Informação e suporte técnico disponibilizados							
4. Qualidade geral dos produtos							
5. Atualização da Oferta							

[Nota INF: Só faz P2 e P3 se ENT_NORS=1]

P2. A [redacted] dispõe de uma oferta variada de equipamentos oficinais e de diagnóstico (por exemplo: equipamentos de lubrificação, abastecimento de gasóleo, ferramentas pneumáticas, elevadores e equipamentos de lavagem, etc). Qual a resposta que melhor reflete a sua situação?	1	Não conhece/nunca ouviu falar	
	2	Já ouviu falar/já tinha visto alguma referência, mas não conhece a oferta existente	
	3	Conhece razoavelmente a oferta existente	
	4	Conhece bem a oferta existente/já adquiriu produtos desta gama	
	9	NS/NR	

[Nota INF: Se P2 < 3, ALERTA_COM="SIM"]

P2_A. Nos últimos 6 meses adquiriu equipamentos oficinais ou de diagnóstico?	1	Sim	
	2	Não	
	9	NS/NR	

[Nota INF: Só faz P3 se P2=2, 3 ou 4]

P3. E relativamente à execução de projetos 'chave na mão' de desenho e montagem de Oficinas? Diria que...	1	Não conhece/nunca ouviu falar	
	2	Já ouviu falar/já tinha visto alguma referência, mas não conhece a oferta existente	
	3	Conhece razoavelmente a oferta existente	
	4	Conhece bem a oferta existente/já adquiriu produtos desta gama	
	9	NS/NR	

Figura 4.3: Exemplo da proposta apresentada para redução do inquérito da Empresa A.

De facto, uns meses mais tarde, após a apresentação desta proposta, o inquérito da Empresa A foi mesmo reduzido. Para além disso, a escala das perguntas fechadas também foi alterada para uma escala de 10 níveis.

4.2.2.4 Modelo de Regressão Logística

Pretende-se agora proceder ao ajustamento de um modelo de regressão logística para esta empresa.

Após consideração e análise do extenso inquérito, o grupo NORS decidiu que para a realização desta tarefa se devia trabalhar com uma base de dados composta por 22 variáveis, sendo elas:

- Logística: Boas condições
- Logística: Cumprimento prazos
- Logística: Eficácia entrega
- Logística: Qualidade geral
- Logística: Rapidez entrega
- Recomendação (variável resposta)
- Satisfação (Capacidade resposta)
- Fornecedor preferencial?[a]
- Resolução das reclamações
- Satisfação (Tangíveis)
- Satisfação (Confiança)
- Satisfação (Empatia)
- Satisfação (Fiabilidade)
- Satisfação Produto: Adequação produtos
- Satisfação Produto: Atualização da Oferta
- Satisfação Produto: Diversidade da oferta
- Satisfação Produto: Informação/suporte técnico
- Satisfação Produto: Qualidade geral produtos
- Satisfação Serviço: Cond. pagamento
- Satisfação Serviço: Descontos praticados
- Satisfação Serviço: Prazos pagamento
- Satisfação Serviço: Preços

[a]Esta variável é diferente das referidas anteriormente pois apenas se pergunta ao cliente se a Empresa A é o seu fornecedor preferencial ou não. É, portanto, uma variável binária (Não/Sim).

Após se aplicar a metodologia referida em 3.1.9.4 não se excluíram quaisquer variáveis da base de dados.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), na Empresa A, segue a seguinte tabela com a frequência absoluta e relativa. Observando a tabela, nota-se

Tabela 4.10: Distribuição da variável resposta na Empresa A.

0 - Não Promotor	**1 - Promotor**	**Total**[5]
99 (53.2%)	87 (46.8%)	186

uma maior percentagem de não promotores (algo negativo). Contudo, as percentagens de não promotores/promotores são próximas.

O modelo obtido foi então o seguinte:

[5]Relembra-se a existência de 3 *missing values* na variável resposta.

Tabela 4.11: Sumário do modelo de regressão logística para a Empresa A juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−2.000	0.474	-4.218	**< 0.001**	-
Satisfação (Tangíveis)1	1.005	0.386	2.602	**0.009**	2.731 (1.281-5.821)
Satisfação (Capacidade de resposta)1	0.948	0.420	2.257	**0.024**	2.582 (1.133-5.883)
Sat. Prod.: Qualidade geral produtos1	1.465	0.621	2.361	**0.018**	4.328 (1.283-14.607)
Fornecedor preferencial?Sim	1.011	0.506	2.000	**0.045**	2.748 (1.021-7.403)
Número de observações usadas: 165 (21 excluídas devido a *missing values*)					AIC: 185.63
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.928
Desempenho Preditivo: AUC (IC 95%)					0.794 (0.724-0.856)
Desempenho Preditivo: ACC					0.709

Pela Tabela 4.11 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é aceitável (70.9%) e tem de se classificar o seu poder discriminativo como aceitável (aceitável a bom julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);
- O *odds* para o sucesso nos clientes promotores na satisfação (tangíveis) é 2.731 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (tangíveis);
- O *odds* para o sucesso nos clientes promotores na satisfação (capacidade de resposta) é 2.582 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (capacidade de resposta);
- O *odds* para o sucesso nos clientes promotores na satisfação com a qualidade geral dos produtos é 4.328 vezes o *odds* para o sucesso nos clientes não promotores na satisfação com a qualidade geral dos produtos;
- O *odds* para o sucesso nos clientes que consideram a Empresa A como o seu fornecedor preferencial é 2.748 vezes o *odds* para o sucesso nos clientes que não consideram a Empresa A como o seu fornecedor preferencial.

4.2.2.5 Análise das variáveis que traduzam as vendas da empresa e eventual relação com a recomendação

Pretendem-se verificar se a variável recomendação (que neste caso é constante ao longo do tempo) tem relação (linear) com as vendas da empresa, ou seja, para esta tarefa, a variável objetivo são as vendas da empresa (ou uma medida que as traduza) e o preditor linear será a recomendação. Para tal, estudou-se o registo de vendas mensal (quantidade e valor líquido), relativamente a cada cliente (no período de 1 de Maio de 2012 a 15 de Dezembro de 2017), obtendo-se uma amostra de 160 clientes.

Pretende-se identificar a medida ideal a selecionar (por exemplo, o declive das vendas) para depois se averiguar em que medida esta influência a variável de estudo. A identificação dessa medida, foi feita através de observação gráfica das respetivas séries temporais.

- **Gráficos das séries temporais das vendas**

  Apresenta-se a seguinte análise gráfica das vendas (em quantidade e valor líquido) para que no final se escolha a medida a usar.

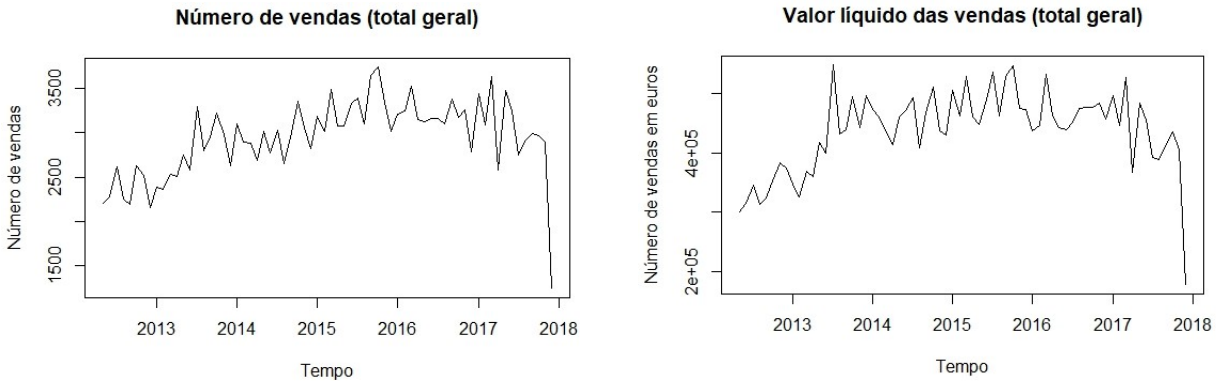


Figura 4.4: Representação da média das vendas da Empresa A ao longo do tempo.

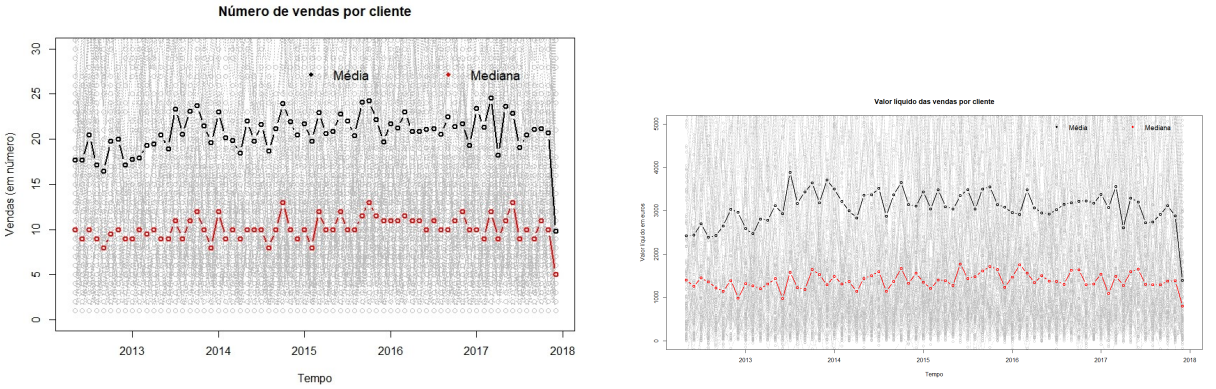


Figura 4.5: Zoom da representação, por cliente, das vendas da Empresa A ao longo do tempo, juntamente com a média e com a mediana.

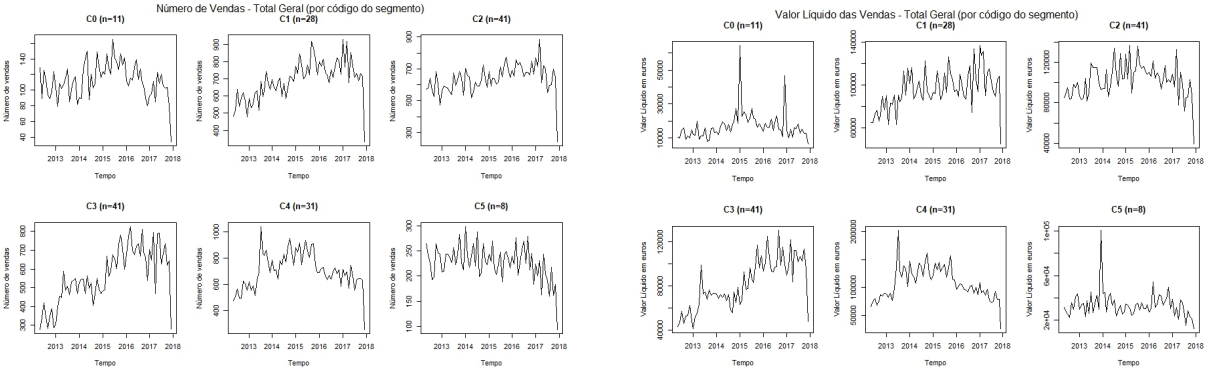


Figura 4.6: Representação da média das vendas da Empresa A ao longo do tempo, por código do segmento.

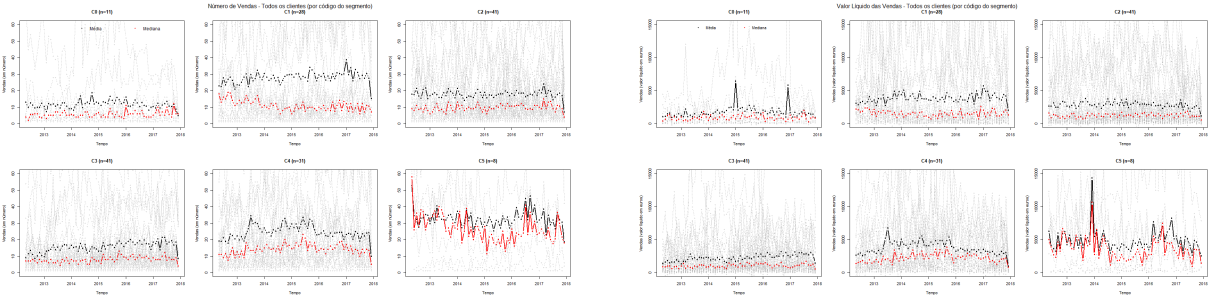


Figura 4.7: Zoom da representação, por cliente, das vendas da Empresa A ao longo do tempo, juntamente com a média e com a mediana, por código do segmento.

Conclusão: A média e mediana apresentam algumas variações, contudo não se observa nenhuma tendência. Perante estes gráficos, opta-se por usar o que já se havia referido anteriormente: o declive das vendas.

Com o objetivo de verificar se a resposta tem influência nas vendas (tanto em valor líquido como em quantidade) irão utilizar-se duas abordagens:

- Regressão linear;
- Modelação de dados longitudinais.

- **Regressão Linear**

Inicialmente, **extraiu-se**, para cada cliente, o **declive** das suas vendas[6] e aplicou-se regressão linear. Contudo, esta metodologia não pareceu adequada. Como tal, filtrou-se o período de vendas para cada cliente, ou seja, para um dado cliente, só se consideraram as

[6] Não se lida diretamente com as vendas.

vendas deste a partir da data da sua resposta ao inquérito, extraindo-se o declive apenas dessas vendas.

Para além disso, eliminou-se o mês de Dezembro de 2017 pois este não se encontrava completo[7] e isso causaria um decréscimo acentuado nas vendas (conforme se pode ver nos gráficos apresentados anteriormente). Para além da recomendação inclui-se o código do segmento, na análise, para se evitar a modelação para cada um dos códigos.

Conclusão: Desta análise resultou que não faz sentido considerar este tipo de modelo para resolver este problema, pois (em nenhum dos casos) se rejeita a hipótese de os coeficientes do modelo serem nulos, ou seja, não se rejeita a hipótese nula de o modelo não existir (*Análise de Variância (Teste F) - Medidas de associação*). Ou seja, por esta análise, verifica-se que não existe relação linear entre as vendas e a resposta.

- **Modelação de dados longitudinais**

  Para dar resposta ao problema colocado recorreu-se à modelação de um modelo linear geral pois pretendia-se inferência a nível populacional.

  Filtrou-se, nova e analogamente, o período de vendas para cada cliente só se considerando as vendas a partir da data da resposta ao inquérito e o mês de Dezembro de 2017 foi também eliminado. Selecionaram-se apenas os indivíduos com 4 ou mais observações para aplicar este método. Para além disso, também se incluiu o código do segmento pelas razões apontadas anteriormente.

  Antes de se efetuar qualquer modelação deve-se fazer uma análise gráfica dos dados usados.

  Observando a Figura 4.8 verifica-se que não existe nenhuma tendência nas vendas ao longo do tempo nem grande variabilidade. Pelo que, o uso de um GLS, parece efetivamente adequado. Pela Figura 4.9, é possível averiguar que os grupos de clientes com os códigos C3, C4 e C5 apresentam vendas superiores aos clientes dos grupos com os outros códigos.

  Pela Figura 4.10, não se notam grandes diferenças na evolução das vendas ao longo do tempo tendo em conta se estes são não promotores (detrator/neutro) ou promotores. Isto pode ser indicativo de que na modelação esta variável seja não significativa. Pela Figura

[7] Continha apenas 15 dias de vendas.

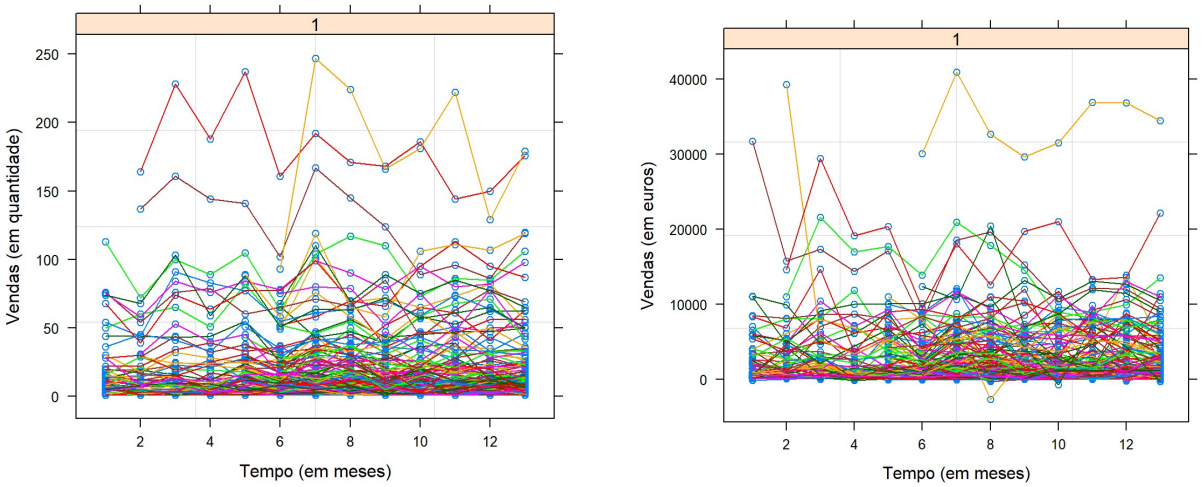


Figura 4.8: Representação das vendas de todos os clientes da Empresa A ao longo do tempo.

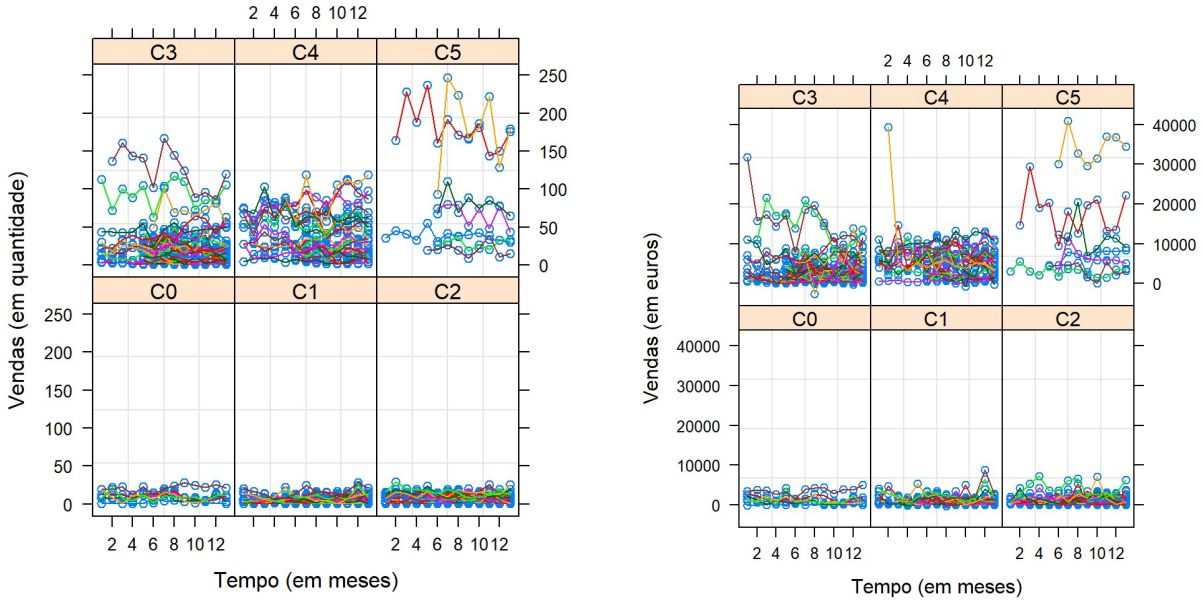


Figura 4.9: Representação das vendas de todos os clientes da Empresa A ao longo do tempo, por código do segmento.

4.11, reforça-se o que se disse anteriormente sobre a Figura 4.9 pois, em média, ao longo do tempo o grupo de clientes com os códigos C3, C4 e C5 apresentam vendas superiores aos outros, sendo este último o que apresenta vendas maiores.

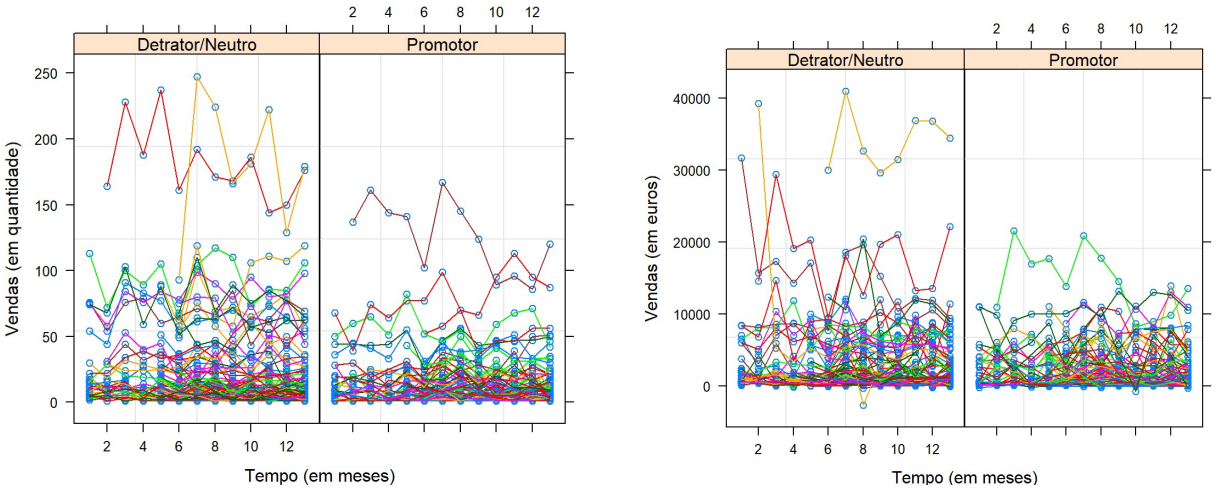


Figura 4.10: Representação das vendas de todos os clientes da Empresa A ao longo do tempo, pelos níveis da variável recomendação (detrator/neutro - não promotor/-promotor).

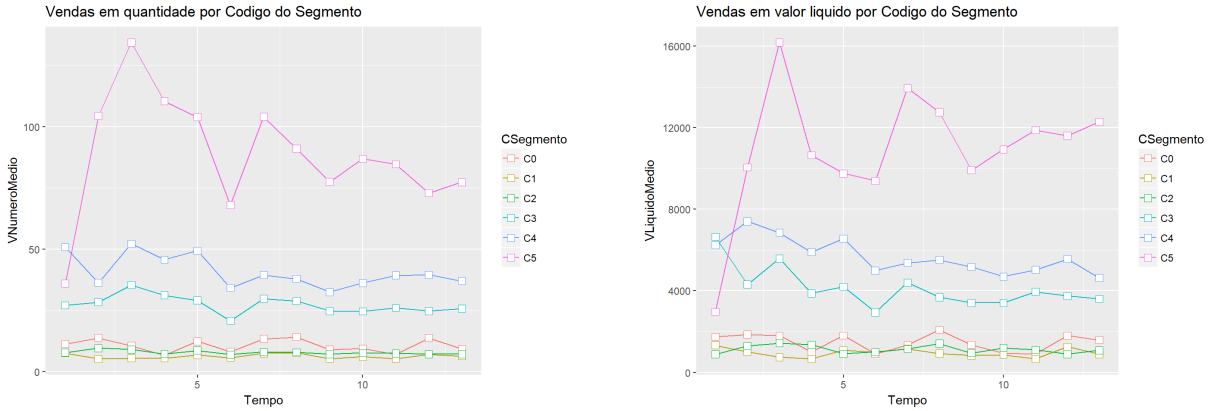


Figura 4.11: Representação da média das vendas da Empresa A ao longo do tempo, por código do segmento.

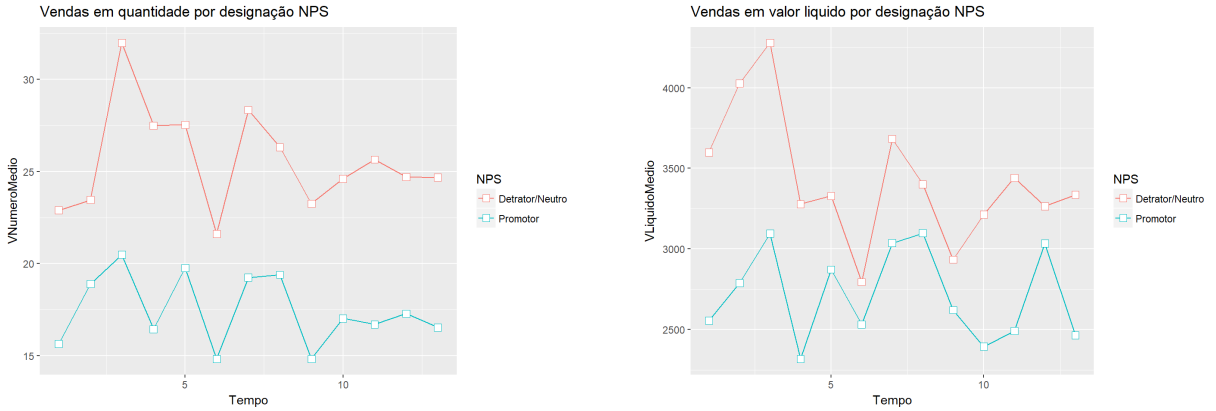


Figura 4.12: Representação da média das vendas da Empresa A ao longo do tempo, pelos níveis da variável recomendação (não promotor/promotor).

Curiosamente, ao contrário do que seria de esperar, em média e ao longo do tempo, os clientes não promotores apresentam vendas superiores ao promotores (Figura 4.12).

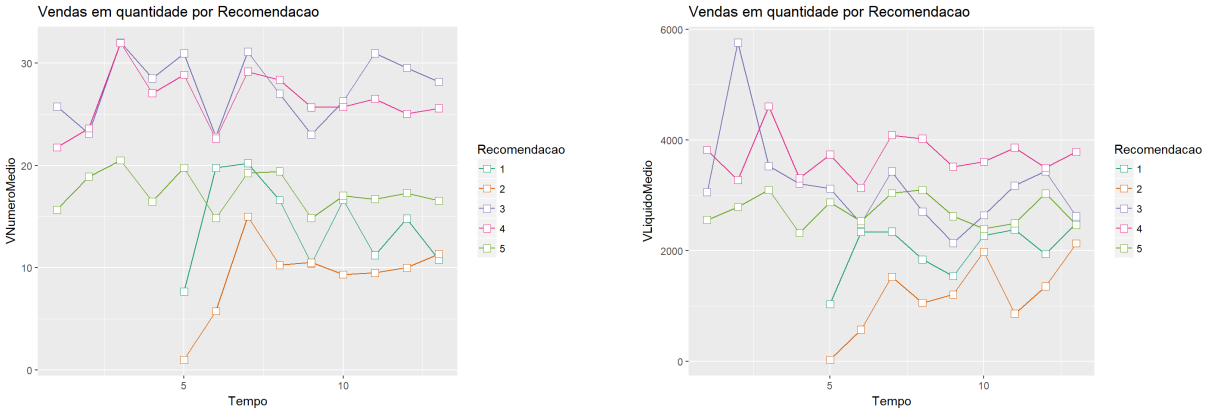


Figura 4.13: Representação da média das vendas da Empresa A ao longo do tempo, pelos níveis da variável recomendação (1 a 5).

Usando a variável recomendação com os níveis de 1 a 5, verifica-se que os clientes que dão 5 não são de todos os que apresentam vendas, em média e ao longo do tempo, mais elevadas. As Figuras 4.12 e 4.13, reforçam a sugestão dada pela Figura 4.10.

Resultados: Usando (separadamente) as duas variáveis que traduziam as vendas da Empresa A, a variável recomendação (categorização: não promotor/promotor ou categorização: 1 a 5) não foi estatisticamente significativa (tanto individualmente como em conjunto com o código do segmento). Desta forma, verifica-se que não existe relação (linear) entre a recomendação e as vendas da Empresa A.

Quanto ao código do segmento, esta variável foi estatisticamente significativa. Contudo, após exclusão da recomendação por esta ser estatisticamente não significativa, os modelos que continham apenas o código do segmento no preditor linear não tinham interesse para o grupo e, como tal, não foram explorados e, por esse motivo, não serão aqui apresentados.

4.2.3 Empresas B e C

4.2.3.1 Análise do perfil de cliente

Um cliente faz 2 tipos de compras: compras de mecânica e compras de carroçaria. Estas compras podem ser feitas através do portal online ou podem ser feitas em loja. Pretende-se traçar o perfil de um dado cliente consoante o tipo de compra e o meio que usa para as fazer.

Desta forma, foram traçados os seguintes perfis:

- Para o tipo de compra: diz-se que um cliente é do perfil

  - **Mecânica:** se 60% ou mais das suas compras pertencem a esta categoria;
  - **Carroçaria:** se 60% ou mais das suas compras pertencem a esta categoria;
  - **Misto:** caso contrário.

- Para o tipo de canal que usa: diz-se que um cliente é do perfil
  - **Portal online:** se 60% ou mais das suas compras pertencem a esta categoria;
  - **Loja:** se 60% ou mais das suas compras pertencem a esta categoria;
  - **Misto:** caso contrário.

Analisando todos os clientes, obtiveram-se os seguintes resultados:

Tabela 4.12: Frequência absoluta (relativa) dos resultados por tipo de compra.

	Empresa B	**Empresa C**
Carroçaria	15 (6.5%)	22 (14.5%)
Mecânica	197 (84.8%)	115 (76.2%)
Misto	12 (5.2%)	8 (5.3%)
Missing values	8 (3.5%)	6 (4%)

Tabela 4.13: Frequência absoluta (relativa) dos resultados por canal de compra.

	Empresa B	**Empresa C**
Loja	33 (14.3%)	129 (85.4%)
Portal Online	173 (74.5%)	5 (3.3%)
Misto	18 (7.8%)	12 (8%)
Missing values	8 (3.5%)	5 (3.3%)

Assim, para estas amostras, verifica-se que, em ambas as empresas, a esmagadora maioria dos clientes efetua compras de mecânica. Quanto ao meio que estes utilizam para fazer as suas compras, na Empresa B, a grande maioria dos clientes recorre ao portal online e, na Empresa C, a grande maioria dos clientes recorre ao serviço prestado em loja.

4.2.3.2 Modelos de Regressão Logística

Pretende-se agora proceder ao ajustamento de modelos de regressão logística, para ambas as empresas, usando toda a base de dados e filtrando essa mesma base de dados consoante o perfil de cliente mais comum detetado anteriormente.

Para ambas as empresas trabalhou-se com uma base de dados composta por 28 variáveis, sendo elas:

- Logística: Boas condições
- Logística: Cumprimento prazos
- Logística: Eficácia entrega
- Logística: Qualidade geral
- Logística: Rapidez entrega
- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado
- Recomendação (variável resposta)
- Resolução das reclamações
- Satisfação (Capacidade resposta)
- Tipo de compra
- Satisfação (Confiança)
- Satisfação (Empatia)
- Satisfação (Fiabilidade)
- Satisfação Produto: Adequação produtos
- Satisfação Produto: Atualização da Oferta
- Satisfação Produto: Diversidade da oferta
- Satisfação Produto: Informação/suporte técnico
- Satisfação Produto: Qualidade geral produtos
- Satisfação Serviço: Cond. pagamento
- Satisfação Serviço: Descontos praticados
- Satisfação Serviço: Prazos pagamento
- Satisfação Serviço: Preços
- Satisfação (Tangíveis)
- Tratamento das devoluções
- Canal de compra

Resultados globais

Veja-se que resultados se obtiveram usando as bases de dados completas.

Após se aplicar a metodologia referida em 3.1.9.4 excluíram-se as seguintes variáveis das bases de dados:

Empresa B

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado
- Tipo de compra
- Canal de compra

Empresa C

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado
- Canal de compra

Ficou-se, assim, com 22 variáveis na Empresa B (retiraram-se 6) e 23 na Empresa C (retiraram-se 5) que foram consideradas na modelação.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), pelas 2 empresas, segue a seguinte tabela com a frequência absoluta e relativa.

Tabela 4.14: Distribuição da variável resposta nas Empresa B e C.

	Empresa B	**Empresa C**
0 - Não Promotor	146 (62.9%)	76 (50.3%)
1 - Promotor	86 (37.1%)	75 (49.7%)
Total	232	151

Observando a tabela, nota-se um equilíbrio entre a distribuição dos resultados na Empresa C e uma diferença notária na Empresa B.

Os modelos obtidos foram então os seguintes:

Tabela 4.15: Sumário do modelo de regressão logística para a Empresa B juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−1.941	0.257	-7.556	**< 0.001**	-
Cumprimento prazos1	1.679	0.433	3.881	**< 0.001**	5.360 (2.296-12.514)
Satisfação (Confiança)1	1.032	0.497	2.076	**0.038**	2.808 (1.059-7.442)
Satisfação (Capacidade resposta)1	1.184	0.515	2.298	**0.022**	3.269 (1.190-8.975)
Sat. Prod.: Atualização da Oferta1	1.230	0.525	2.470	**0.014**	3.653 (1.307-10.212)
Número de observações usadas: 205 (27 excluídas devido a *missing values*)					AIC: 185.42
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.974
Desempenho Preditivo: AUC (IC 95%)					0.832 (0.769-0.892)
Desempenho Preditivo: ACC					0.829

Tabela 4.16: Sumário do modelo de regressão logística para a Empresa C juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−0.495	0.534	-0.928	0.353	-
Resolução das reclamações0	−1.346	0.582	-2.313	**0.021**	0.260 (0.083-0.814)
Resolução das reclamações1	0.089	0.833	0.106	0.915	1.093 (0.214-5.591)
Satisfação (Empatia)1	1.100	0.540	2.037	**0.042**	3.004 (1.043-8.653)
Satisfação (Fiabilidade)1	1.754	0.634	2.769	**0.006**	5.778 (1.669-20.003)
Sat. Prod.: Diversidade da Oferta1	2.155	0.647	3.329	**0.001**	8.627 (2.426-30.673)
Número de observações usadas: 132 (19 excluídas devido a *missing values*)					AIC: 125.64
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.845
Desempenho Preditivo: AUC (IC 95%)					0.871 (0.801-0.926)
Desempenho Preditivo: ACC					0.841

Empresa B: Pela Tabela 4.15 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (82.9%) e tem de se classificar o seu poder discriminativo como bom (aceitável a bom julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);
- O *odds* para o sucesso nos clientes promotores na satisfação com o cumprimento de prazos é 5.360 vezes o *odds* para o sucesso nos clientes não promotores na satisfação com o cumprimento de prazos;
- O *odds* para o sucesso nos clientes promotores na satisfação (confiança) é 2.808 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (confiança);
- O *odds* para o sucesso nos clientes promotores na satisfação (capacidade de resposta) é 3.269 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (capacidade de resposta);
- O *odds* para o sucesso nos clientes promotores na satisfação com a atualização das ofertas o cumprimento de prazos é 3.653 vezes o *odds* para o sucesso nos clientes não promotores na satisfação com a atualização das ofertas.

Empresa C: Pela Tabela 4.16 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (84.1%) e tem de se classificar o seu poder discriminativo como bom (bom a excecional julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Excluindo o *Intercept*, há apenas um fator de risco (*Resolução das reclamações*0). Todas as outras variáveis são fatores de proteção;
- O *odds* para o sucesso nos clientes não promotores na satisfação com a resolução das reclamações é 74% inferior ao *odds* para o sucesso nos clientes que não efetuaram reclamações;

- O *odds* para o sucesso nos clientes promotores na satisfação (empatia) é 3.004 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (empatia);

- O *odds* para o sucesso nos clientes promotores na satisfação (fiabilidade) é 5.778 vezes o *odds* para o sucesso nos clientes não promotores na satisfação (fiabilidade);

- O *odds* para o sucesso nos clientes promotores na satisfação com a diversidade da oferta é 8.627 vezes o *odds* para o sucesso nos clientes não promotores na satisfação com a diversidade da oferta.

Os resultados obtidos por tipo e canal de compra mais frequentes podem ser consultados no Anexo 4.

4.2.4 Empresa D

O questionário encontra-se estruturado em 3 grupos. Nesta secção, apresentam-se os resultados obtidos através do mesmo.

4.2.4.1 Apresentação dos inquéritos por tipologia de pergunta

Tabela 4.17: Tipologia de pergunta do inquérito do setor 1 da Empresa D.

Tipologia	Número de perguntas/campos
Dados do cliente	3
Perguntas gerais	3
Perguntas específicas sobre o serviço do setor	11

Tabela 4.18: Tipologia de pergunta do inquérito do setor 2 da Empresa D.

Tipologia	Número de perguntas/campos
Dados do cliente	3
Perguntas gerais	3
Perguntas específicas sobre o serviço do setor	10
Ofertas feitas ao cliente	2

Nota: Nas perguntas gerais é onde está incluída a variável objetivo.

4.2.4.2 Análise descritiva e exploratória dos dados

Após análise da BD fornecida, chegou-se à conclusão que, neste caso, seria interessante não considerar apenas uma, mas sim duas variáveis para esta análise (embora que estudadas em separado). Elas são designadas por: Satisfação Global[8] e Recomendar. Como consequência, foi criada uma nova variável, designada por I_N, que é a média das duas anteriores.

Obteve-se a seguinte distribuição dos resultados conforme a variável em causa:

- Satisfação Global:

Tabela 4.19: Setor 1.

Classificação	Frequência Absoluta	Frequência Relativa
1	30	1.6%
2	14	0.7%
3	10	0.5%
4	23	1.2%
5	79	4.1%
6	82	4.3%
7	222	11.5%
8	613	31.9%
9	297	15.4%
10	550	28.6%
NA	3	0.2%

Tabela 4.20: Setor 1.

Classificação	Frequência Absoluta	Frequência Relativa
1	8	2.4%
2	3	0.9%
3	1	0.3%
4	2	0.6%
5	9	2.6%
6	9	2.6%
7	25	7.4%
8	100	29.4%
9	68	20%
10	113	33.2%
NA	2	0.6%

- Recomendar:

[8] É colocada ao cliente a seguinte questão: numa escala de 1 a 10 qual é a sua satisfação global com a Empresa D?

Tabela 4.21: Setor 1.

Classificação	Frequência Absoluta	Frequência Relativa
1	34	1.7%
2	9	0.5%
3	9	0.5%
4	11	0.6%
5	71	3.7%
6	66	3.4%
7	161	8.4%
8	478	24.9%
9	256	13.3%
10	814	42.3%
NA	14	0.7%

Tabela 4.22: Setor 2.

Classificação	Frequência Absoluta	Frequência Relativa
1	10	2.9%
2	1	0.3%
3	0	0%
4	2	0.6%
5	13	3.8%
6	4	1.2%
7	19	5.6%
8	81	23.8%
9	48	14.1%
10	157	46.2%
NA	5	1.5%

- Variável I_N:

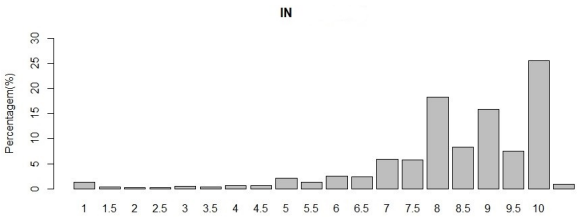


Figura 4.14: Setor 1.

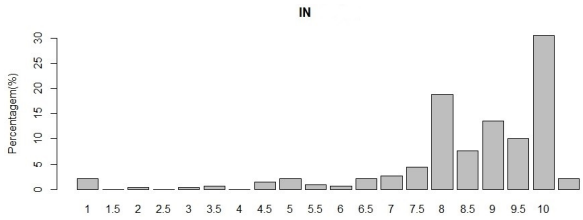


Figura 4.15: Setor 2.

Pelas tabelas e gráficos anteriores observamos que existem vários *missing values*. Isto significa que:

- 17 clientes não responderam a pelo menos uma das perguntas em causa nos questionário do setor 1;
- 7 clientes não responderam a pelo menos uma das perguntas em causa nos questionário do setor 2;
- no total, 24 clientes não responderam a pelo menos uma das perguntas, em causa, num dos inquéritos.

Para se ter uma melhorar ideia se estes resultados são bons ou não observam-se os valores médios destas variáveis.

Tabela 4.23: Média da satisfação global e da variável recomendar na Empresa D, por setor, juntamente com a correlação de *Spearman*.

Escala	Setor 1		Setor 2	
	Satisfação Global	Recomendar	Satisfação Global	Recomendar
10pontos	8.2	8.5	8.4	8.6
100pontos	82.1	85.2	84.1	86.3
Correlação de *Spearman*				
	0.72		0.77	

Tabela 4.24: Média da variável I_N da Empresa D, por setor.

Escala	Setor 1	Setor 2
10pontos	8.4	8.4
100pontos	83.5	84.1

Analisando as tabelas anteriores, é possível averiguar que, em média, os resultados da satisfação global são inferiores aos da variável recomendar, nos dois setores. Para além disso, os resultados médios são satisfatórios[9] sendo ligeiramente superiores no setor 2. Pelos resultados da correlação apresentados anteriormente podemos concluir que existe uma relação moderada entre estes dois objetos de estudo[10].

Veja-se agora a classificação dada pelos clientes nas duas variáveis (satisfação global e recomendar) em simultâneo em cada um dos setores. Para o efeito, apresenta-se a matriz de confusão e o plot entre estas duas variáveis.

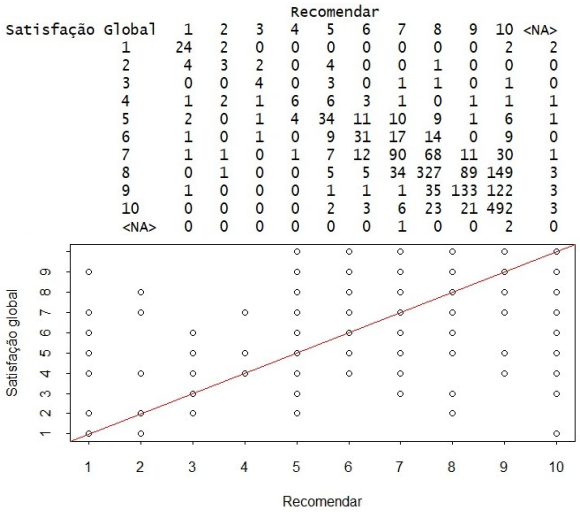


Figura 4.16: Setor 1.

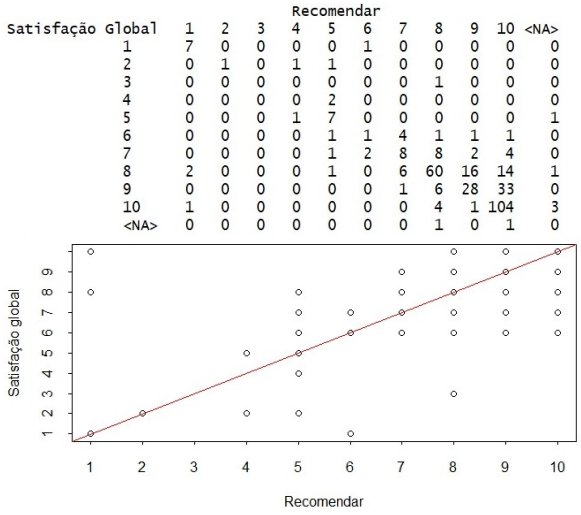


Figura 4.17: Setor 2.

[9] Focando na variável recomendar.

[10] Satisfação global e recomendar.

As conclusões que se podem retirar pela análise das figuras anteriores são as seguintes:

Setor 1

- 1144 clientes deram a mesma classificação nas duas perguntas, ou seja, quase 60%;
- 187 clientes deram uma classificação maior na satisfação global do que na variável recomendar, ou seja, quase 10%;
- 575 clientes deram uma classificação maior na variável recomendar do que na satisfação global, ou seja, quase 30%;
- todos os clientes responderam a, pelo menos, uma das questões.

Setor 2

- 216 clientes deram a mesma classificação nas duas perguntas, ou seja, quase 64%;
- 27 clientes deram uma classificação maior na satisfação global do que na variável recomendar, ou seja, quase 8%;
- 90 clientes deram uma classificação maior na variável recomendar do que na satisfação global, ou seja, quase 27%;
- todos os clientes responderam a, pelo menos, uma das questões.

Para além disso, notam-se pontuações bastante incongruentes nos cantos superior esquerdo e inferior direito dos gráficos apresentados anteriormente. Por exemplo, em ambos os setores, há clientes que na recomendação dão a pontuação máxima mas depois na satisfação global dão as pontuações mais baixas e vice-versa.

Por fim, efetuou-se também uma análise por distrito. Contudo, será dados mais destaque a esta análise mais a frente. Assim sendo, apresentam-se apenas os resultados obtidos para o setor 2 pois os resultados do outro setor serão apresentados em 4.3.1.

Observando a Figura 4.18, poderão tirar-se várias conclusões (em média), consoante o distrito que o leitor queira analisar. Devido à extensão que teria a escrita dessas conclusões, estas não serão descritas.

Distrito	Freq.Abs	Freq.Relativa(em %)	Satisfação Global	Recomendar	IN
AVEIRO	28	8.2	86.1	90.7	88.4
BEJA	2	0.6	75.0	85.0	80.0
BRAGA	43	12.6	85.1	85.8	85.5
BRAGANÇA	4	1.2	80.0	80.0	80.0
CASTELO BRANCO	1	0.3	80.0	80.0	80.0
COIMBRA	12	3.5	70.0	74.2	72.1
ÉVORA	7	2.1	78.6	81.4	80.0
FARO	9	2.6	91.1	95.6	93.3
GUARDA	9	2.6	91.1	92.2	91.7
ILHA DA MADEIRA	2	0.6	100.0	100.0	100.0
ILHA DE SÃO MIGUEL	1	0.3	90.0	100.0	95.0
ILHA TERCEIRA	1	0.3	10.0	10.0	10.0
LEIRIA	23	6.8	86.5	90.9	88.9
LISBOA	36	10.6	84.0	82.9	83.2
PORTALEGRE	2	0.6	85.0	90.0	87.5
PORTO	46	13.5	89.6	90.2	90.3
SANTARÉM	28	8.2	79.3	85.4	82.0
SETÚBAL	11	3.2	81.8	82.7	82.3
VIANA DO CASTELO	14	4.1	90.0	88.6	89.3
VILA REAL	8	2.4	91.2	88.8	90.0
VISEU	13	3.8	79.2	85.8	81.7
NA	40	11.8	79.5	83.1	81.0

Figura 4.18: Médias da satisfação global, recomendar e I_N por distrito da Empresa D, no setor 2.

4.2.4.3 Modelos de regressão logística

Pretende-se agora proceder ao ajustamento de modelos de regressão logística, para ambos os setores da empresa.

Trabalhou-se então com as seguintes bases de dados (uma para cada setor):

Setor 1 (14 variáveis)

- Satisfação global
- Recomendar (variável objetivo)
- Comprar novamente
- W1: Cortesia e prestabilidade
- W2: Disponibilidade da oficina
- W3: Informação/orçamento sobre a reparação/manutenção
- W4: Conclusão da reparação/manutenção acordada
- W5: Explicação do trabalho executado
- W6: Informação sobre trabalhos adicionais
- W7: Relações a longo prazo
- W8: Qualidade da manutenção/reparação
- W9: Resolução do problema à primeira visita
- W10: Disponibilidade das peças
- W11: Contato com a pessoa certa

Setor 2 (15 variáveis)

- Satisfação global
- Recomendar (variável objetivo)
- Comprar novamente
- S1: Cortesia a prestabilidade
- S2: Conhecimentos e competências
- S3: Relações a longo prazo
- S4: Solução geral
- S5: Cumprimento do prazo de entrega acordado
- S6: Comunicação sobre o prazo de entrega
- S7: Estado do camião aquando da entrega
- S8: Informações fornecidas durante a entrega
- S9: Contato/acompanhamento após a entrega
- S10: Contato com a pessoa certa
- SO1: Foi-lhe fornecido algum servido adicional?
- SO2: Comprou algum serviço adicional?

Após se aplicar a metodologia referida em 3.1.9.4, excluiu-se apenas a variável SO2 da base de dados do setor 2. Ficou-se, assim, com 14 variáveis na base de dados do setor 1 e 14 na do setor 2 (retirou-se 1) que foram consideradas na modelação.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), pelas 2 empresas, segue a seguinte tabela com a frequência absoluta e relativa.

Tabela 4.25: Distribuição da variável resposta na Empresa D dividindo pelos 2 setores.

	Setor 1	**Setor 2**
0 - Não Promotor	839 (43.9%)	130 (38.2%)
1 - Promotor	1070 (56.1%)	210 (61.8%)
Total	1909	340

Relembra-se a existência de *missing values* na variável resposta. Essas observações foram removidas. Trabalha-se então com o número de observações indicado na tabela anterior.

Observando a tabela, em ambos os setores, há uma percentagem maior de promotores, sendo maior no setor 2.

Os modelos obtidos foram então os seguintes:

Tabela 4.26: Sumário do modelo de regressão logística para o setor 1 da Empresa D juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−5.152	0.639	-8.066	**< 0.001**	-
Satisfação Global$_2$	0.701	0.286	2.253	**0.014**	2.016 (1.151-3.530)
Satisfação Global$_3$	2.880	0.306	9.421	**< 0.001**	17.813 (9.784-32.430)
Comprar Novamente$_2$	0.133	0.380	0.351	0.725	1.143 (0.543-2.404)
Comprar Novamente$_3$	1.930	0.355	5.433	**< 0.001**	6.890 (3.434-13.822)
W5$_2$	0.717	0.354	2.025	**0.043**	2.049 (1.023-4.102)
W5$_3$	1.268	0.355	3.567	**< 0.001**	3.552 (1.770-7.128)
W7$_2$	0.309	0.552	0.560	0.575	1.363 (0.462-4.020)
W7$_3$	1.077	0.549	1.962	**0.049**	2.937 (1.001-8.617)
W11$_2$	0.586	0.360	1.627	0.104	1.796 (0.887-3.637)
W11$_3$	0.981	0.349	2.809	**0.005**	2.666 (1.345-5.286)
Número de observações usadas: 1848 (61 excluídas devido a *missing values*)					AIC: 1314.5
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.501
Desempenho Preditivo: AUC (IC 95%)					0.922 (0.909-0.933)
Desempenho Preditivo: ACC					0.838

Setor 1: Pela Tabela 4.26 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (83.8%) e tem de se classificar o seu poder discriminativo como excecional.

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos os factores, são fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);
- O *odds* para o sucesso nos clientes neutros na satisfação global é 2.016 vezes o *odds* para o sucesso nos clientes detratores na satisfação global;
- O *odds* para o sucesso nos clientes promotores na satisfação global é 17.813 vezes o *odds* para o sucesso nos clientes detratores na satisfação global;

Tabela 4.27: Sumário do modelo de regressão logística para o setor 2 da Empresa D juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−3.754	1.072	-3.502	**< 0.001**	-
Satisfação Global$_2$	0.719	0.891	0.807	0.420	2.053 (0.358-11.778)
Satisfação Global$_3$	3.280	0.896	3.663	**< 0.001**	26.576 (4.594-153.740)
Comprar Novamente$_2$	0.619	0.581	1.066	0.286	1.858 (0.595-5.800)
Comprar Novamente$_3$	1.994	0.587	3.397	**< 0.001**	7.346 (2.325-23.214)
S4$_2$	0.744	0.913	0.814	0.415	2.104 (0.351-12.596)
S4$_3$	2.150	0.923	2.329	**0.020**	8.586 (1.406-52.423)
Número de observações usadas: 321 (19 excluídas devido a *missing values*)					AIC: 218.45
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.977
Desempenho Preditivo: AUC (IC 95%)					0.927 (0.895-0.955)
Desempenho Preditivo: ACC					0.866

- O *odds* para o sucesso nos clientes promotores na variável comprar novamente[11] é 6.890 o *odds* para o sucesso nos clientes detratores na variável comprar novamente;
- O *odds* para o sucesso nos clientes neutros na questão W5 é 2.049 vezes o *odds* para o sucesso nos clientes detratores na questão W5;
- O *odds* para o sucesso nos clientes promotores na questão W5 é 3.552 vezes o *odds* para o sucesso nos clientes detratores na questão W5;
- O *odds* para o sucesso nos clientes promotores na questão W7 é 2.937 vezes o *odds* para o sucesso nos clientes detratores na questão W7;
- O *odds* para o sucesso nos clientes promotores na questão W11 é 2.666 vezes o *odds* para o sucesso nos clientes detratores na questão W11.

Setor 2: Pela Tabela 4.27 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (86.6%) e tem de se classificar o seu poder discriminativo como excecional (bom a excecional julgando pelo IC).

[11] Nesta questão é pedido aos clientes que numa escala de 1 a 10 indiquem qual a possibilidade de comprarem novamente.

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos os factores, são fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);

- O *odds* para o sucesso nos clientes promotores na satisfação global é 26.576 vezes o *odds* para o sucesso nos clientes detratores na satisfação global;

- O *odds* para o sucesso nos clientes promotores na variável comprar novamente é 7.346 vezes o *odds* para o sucesso nos clientes detratores na variável comprar novamente;

- O *odds* para o sucesso nos clientes promotores na questão S4 é 8.586 vezes o *odds* para o sucesso nos clientes detratores na questão S4.

4.2.4.4 Cálculo do NPS para os dados de 2017

Os resultados que aqui se apresentam foram obtidos usando apenas os dados de 2017.

O *layout* com que se apresentam os resultados do NPS é da autoria do grupo NORS.

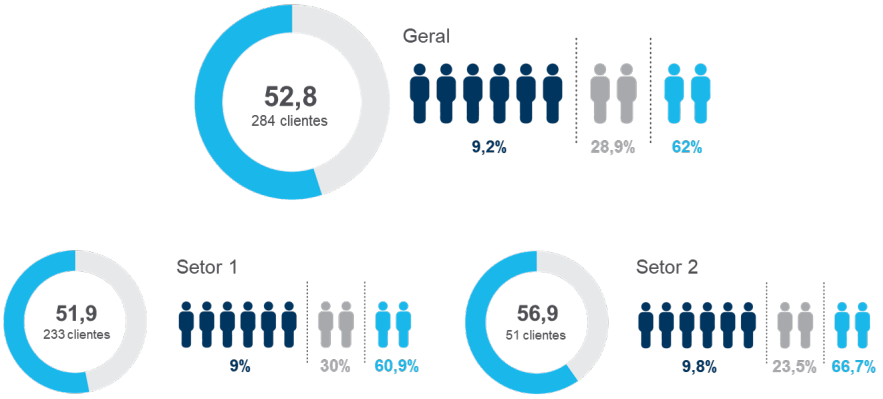


Figura 4.19: Resultados do NPS global e por setor da Empresa D.

Figura 4.20: Resultados do NPS por concessionário e por setor da Empresa D.

Nota: Nos casos em que não se apresentam os resultados subdivididos por setores quer dizer que o concessionário em questão só pertence a um dos setores.

Como já se referiu, não existe um valor a partir do qual se considere que o resultado do NPS é bom. Como tal, comparam-se os resultados (por setor e concessionário) com o resultado geral. Os resultados acima do resultado geral são considerados satisfatórios. Os outros, não. Por exemplo, considera-se que os resultados obtidos no setor 2 são bons, ao contrário dos do setor 1.

4.2.4.5 Análise das variáveis que traduzam as vendas da empresa e eventual relação com a recomendação

Pretendem-se verificar se a variável recomendação tem relação (linear) com as vendas da empresa, ou seja, para esta tarefa, a variável objetivo são as vendas da empresa e o preditor linear será a recomendação.

Para tal efetuou-se uma modelação para cada uma das variáveis que traduzam as vendas da empresa. Sendo elas

- Valor Líquido;
- Margem;
- Margem percentual$= \dfrac{\text{Margem}}{\text{Valor Líquido}} \times 100\%$.

Para tal, consoante os dados disponíveis, estudou-se o registo anual de cada uma destas variáveis, relativamente a cada cliente, no período de 2013 a 2017, no setor 1, e no período de 2014 a 2017 no setor 2.

A base de dados, para ambos os setores, ficou estruturada consoante os anos a que o cliente respondeu ao inquérito, ou seja, se um dado cliente respondeu ao inquérito em 2015 e depois em 2017, apenas se selecionam as suas vendas nesses anos o que dá origem a uma estrutura à apresentada em 3.2.1. Onde a variável resposta é uma das variáveis apresentadas anteriormente e as covariáveis são a variável recomendar numa escala de 1 a 10 e a variável recomendar usando a categorização do NPS (claro que não serão usadas em simultâneo no preditor linear, serão usadas em separado). Desta forma, todas as variáveis presentes nas bases de dados podem variar com o tempo (ano). Note-se que, os instantes temporais não são iguais para todos os clientes, nem igualmente espaçados, pois, por exemplo, um cliente pode ter respondido ao inquérito em 2015 e depois em 2017 e outro cliente pode ter respondido em 2014 e depois em 2017.

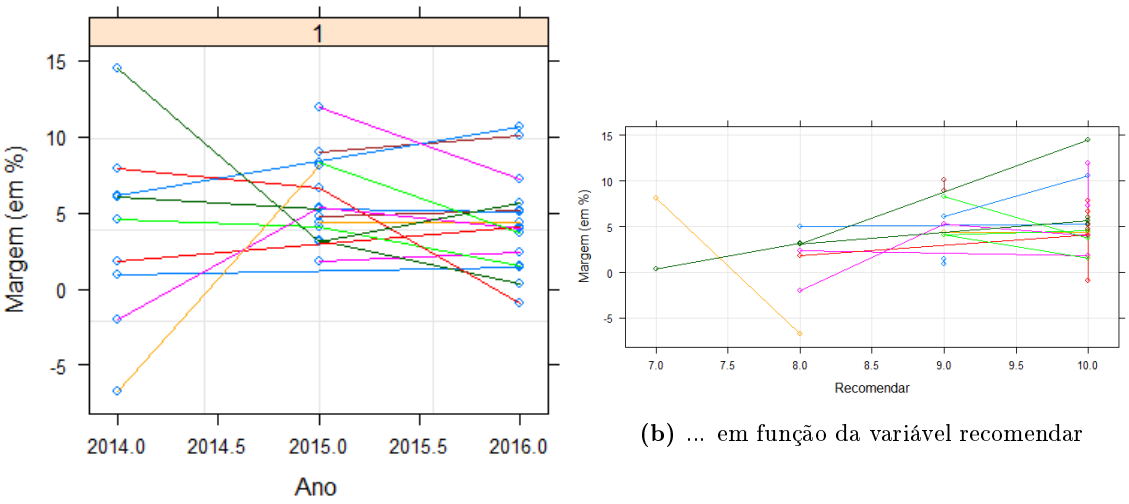
Para o setor 1, verificou-se que não existe relação (linear) entre a variável recomendar e as vendas, pois não se obteve nenhum modelo significativo usando modelos lineares gerais.

Para o setor 2, verificou-se que não existe relação (linear) entre a variável recomendar e o valor líquido das vendas nem existe relação (linear) entre a variável recomendar e a margem das vendas, pois não se obteve nenhum modelo significativo usando modelos lineares gerais. Quanto à relação entre a variável recomendar e a margem percentual das vendas, a situação já é diferente. Como tal, este caso será o único caso a ser explorado neste secção pois é diferente dos apresentados anteriormente.

Relação entre a variável recomendar e a margem percentual das vendas para o setor 2 da Empresa D

Os resultados que agora se apresentam foram obtidos usando uma amostra de 17 clientes o que deu origem a 38 observações.

Comece-se pela análise gráfica dos dados. Nota-se que a Figura 4.21 não apresenta qualquer erro. A margem percentual pode mesmo ser negativa por razões que não podem ser referidas por questões de confidencialidade. Para além disso, não se observa nenhuma tendência evidente e a variabilidade (Figura 4.21a) praticamente parece não existir. Este último facto juntamente

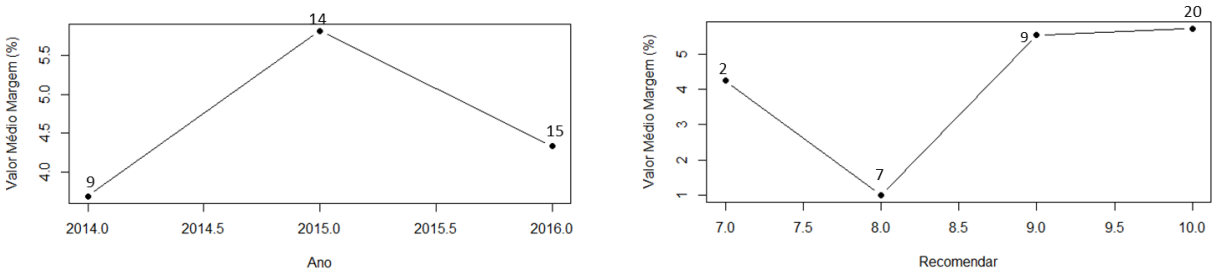


(a) ...ao longo do tempo.

(b) ... em função da variável recomendar

Figura 4.21: Representação da margem percentual de todos os clientes...

com o facto de o grupo pretender uma inferência a nível populacional é um indicativo para se aplicarem modelos lineares generalizados.



(a) ...ao longo do tempo.

(b) ... em função da variável recomendar

Figura 4.22: Representação da margem percentual média, com o número de observações usadas para o cálculo, ...

Pela Figura 4.22 observa-se que a margem percentual média apresenta valor mais alto em 2015 em relação aos outros anos (onde o valor se encontra muito próximo). Também se pode verificar (Figura 4.22b) que na base de dados não existem clientes detratores e que, em média, valores mais altos de satisfação traduzem uma margem percentual maior.

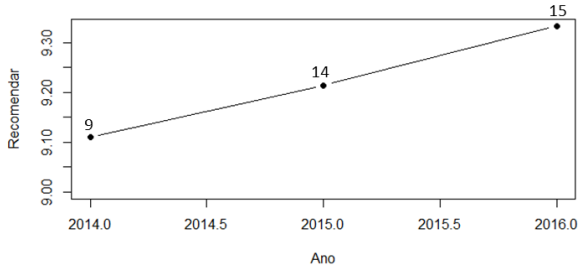


Figura 4.23: Média da variável recomendar ao longo do tempo.

Pela Figura 4.23 verifica-se que em média, a variável recomendar aumenta ao longo tempo. Pela Figura 4.24 verifica-se uma diferença acentuada na média da margem percentual entre os

Recomendar_NPS	n	`Margem Percentual Média`
<fct>	<int>	<dbl>
Neutro	9	1.72
Promotor	29	5.66

Figura 4.24: Margem percentual média em função da variável recomendar (categorização do NPS).

neutros e os promotores. Isto pode ser indicativo que a variável recomendar (com a categorização segundo o NPS) seja significativa nos modelos a estimar.

O modelo linear generalizado obtido, após se ter aplicado toda a metodologia descrita sobre este assunto em 3.2.2, apresenta o seguinte sumário

Tabela 4.28: Sumário do modelo linear generalizado para o setor 2 da Empresa D cuja variável resposta é a margem percentual.

Variável	Est	EP	t	valor-p
$(Intercept)$	1.720	1.195	1.439	0.159
Recomendar_$\text{NPS}_{\text{Promotor}}$	3.938	1.367	2.880	**0.007**
A matriz Σ_i tem a estrutura *default* do R.				AIC: 205.630

Assim, o modelo estima que, em média, um cliente passar de neutro a promotor resulta num aumento de 3.938 pontos percentuais na margem percentual das vendas.

4.2.5 Empresa E

4.2.5.1 Apresentação dos inquéritos por tipologia de pergunta

Igual ao que foi apresentado em 4.2.4.1.

4.2.5.2 Cálculo do NPS

O *layout* com que se apresentam estes resultados é da autoria do grupo NORS.

Nota: Nos casos em que não se apresentam os resultados subdivididos por setores quer dizer que o concessionário em questão só pertence a um dos setores.

Como já se referiu, não existe um valor a partir do qual se considere que o resultado do NPS é bom. Como tal, comparam-se os resultados (por setor e concessionário) com o resultado geral. Os resultados acima do resultado geral são considerados satisfatórios. Os outros, não. Por exemplo, considera-se que os resultados obtidos no setor 1 são bons, ao contrário dos do setor 2.

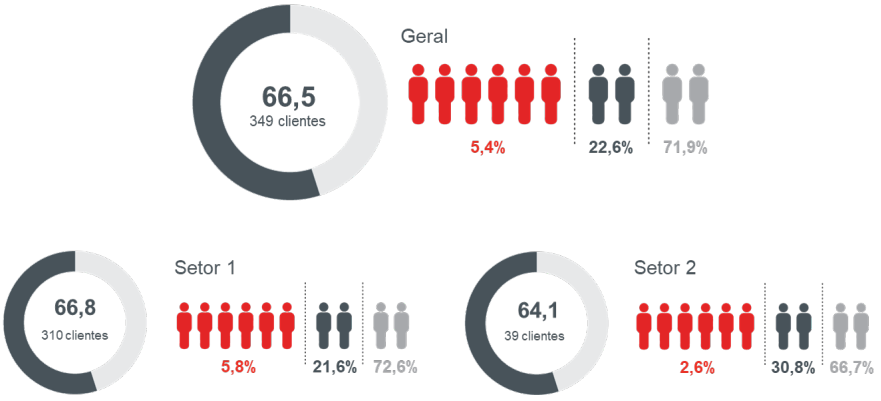


Figura 4.25: Resultados do NPS global e por setor da Empresa E.

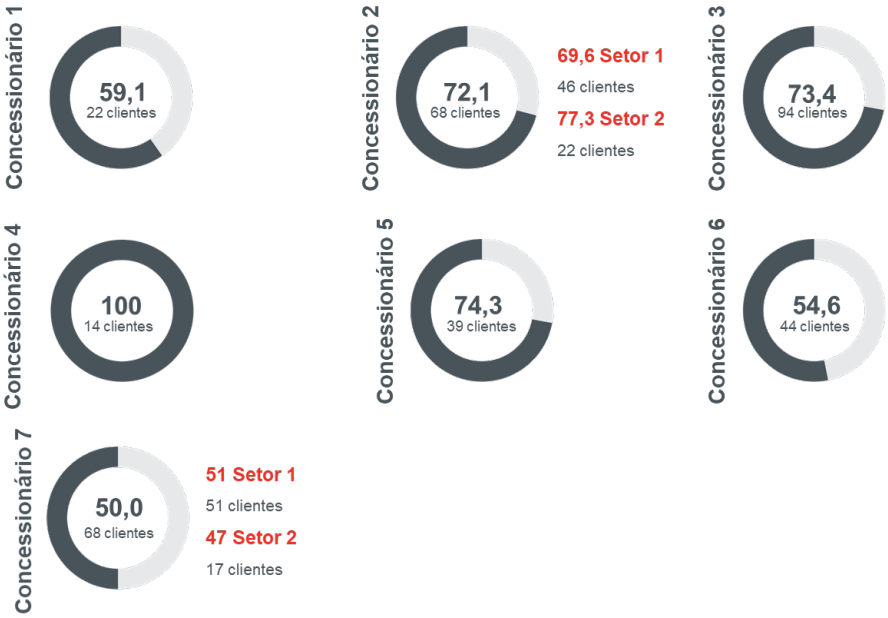


Figura 4.26: Resultados do NPS por concessionário e por setor da Empresa E.

4.2.5.3 Modelos de Regressão Logística

Pretende-se agora proceder ao ajustamento de modelos de regressão logística, para ambos os setores da empresa.

Trabalhou-se com as bases de dados apresentadas em 4.2.4.3.

Após se aplicar a metodologia referida em 3.1.9.4, excluíram-se apenas as variáveis SO1 e SO2 da base de dados do setor 2. Ficou-se, assim, com 14 variáveis na base de dados do setor 1 e 13 na do setor 2 (retiraram-se 2) que foram consideradas na modelação.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), pelas 2 empresas, segue a seguinte tabela com a frequência absoluta e relativa.

Tabela 4.29: Distribuição da variável resposta na Empresa E dividindo pelos 2 setores.

	Setor 1	**Setor 2**
0 - Não Promotor	85 (27.4%)	13 (33.3%)
1 - Promotor	225 (72.6%)	26 (66.7%)
Total	310	39

Observando a tabela, em ambos os setores, há uma percentagem (muito) maior de promotores, sendo esta maior no setor 1.

Os modelos obtidos foram então os seguintes:

Tabela 4.30: Sumário do modelo de regressão logística para o setor 1 da Empresa E juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	-3.166	1.086	-2.914	**0.004**	-
Satisfação Global_3	3.199	0.555	5.758	$<$ **0.001**	24.497 (8.258-72.733)
W1_3	1.172	0.470	2.510	**0.012**	3.228 (1.285-8.111)
W6_2	1.925	1.097	1.755	0.079	6.852 (0.798-58.858)
W6_3	2.149	1.086	1.980	**0.048**	8.576 (1.021-72.064)
W10_2	-0.675	0.652	-1.035	0.301	0.509 (0.142-1.827)
W10_3	1.313	0.665	1.974	**0.048**	3.717 (1.010-13.686)
Número de observações usadas: 276 (34 excluídas devido a *missing values*)					AIC: 153.97
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.506
Desempenho Preditivo: AUC (IC 95%)					0.942 (0.904-0.972)
Desempenho Preditivo: ACC					0.891

Tabela 4.31: Sumário do modelo de regressão logística para o setor 2 da Empresa E juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	-2.197	1.054	-2.085	**0.037**	-
S1_3	4.030	1.184	3.405	**0.001**	56.250 (5.530-572.280)
Número de observações usadas: 39					AIC: 33.771
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=1
Desempenho Preditivo: AUC (IC 95%)					0.827 (0.692-0.962)
Desempenho Preditivo: ACC					0.872

Setor 1: Pela Tabela 4.30 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (89.1%) e tem de se classificar o seu poder discriminativo como excecional.

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);
- O *odds* para o sucesso nos clientes promotores na satisfação global é 24.497 vezes o *odds* para o sucesso nos clientes não promotores na satisfação global;
- O *odds* para o sucesso nos clientes promotores na questão W1 é 3.228 vezes o *odds* para o sucesso nos clientes não promotores na questão W1;
- O *odds* para o sucesso nos clientes promotores na pergunta W6 é 8.576 vezes o *odds* para o sucesso nos clientes detratores na pergunta W6;
- O *odds* para o sucesso nos clientes promotores na pergunta W10 é 3.717 vezes o *odds* para o sucesso nos clientes detratores na pergunta W10.

Setor 2: Pela Tabela 4.31 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (87.2%) e tem de se classificar o seu poder discriminativo como bom (fraco a excecional julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Excluindo o *Intercept*, a variável presente no modelo é um fator de proteção;
- O *odds* para o sucesso nos clientes promotores na questão S1 é 56.250 vezes o *odds* para o sucesso nos clientes não promotores na questão S1.

4.3 Comparação dos resultados obtidos

4.3.1 Empresa A *versus* Empresa D

4.3.1.1 Comparação dos inquéritos por tipologia e número de perguntas

Comece-se por comparar os inquéritos por tipologia e número de pergunta. De forma a se ter uma melhor comparação, vai-se adotar a tipologia do inquérito da Empresa A e ver onde as perguntas do da Empresa D (setor 1) se enquadram.

Nota: Os inquéritos da Empresa E são iguais aos da Empresa D. Por isso, a comparação dos inquéritos fica reduzida a Empresa A *versus* Empresa D/E.

Tabela 4.32: Comparação dos inquéritos das Empresa A e D por tipologia de pergunta.

Tipologia	**Empresa A**	**Empresa D/E**
Dados para o contacto	9	3
Serviço e qualidade do mesmo	5	6
Meios de contato	2 a 8	2
Logística	3 ou 4	3
Devoluções	1 ou 2	0
Produto	3 ou 4	3
Condições comerciais	1 ou 2	0
Serviço online	2 ou 3	0
Alea	2 ou 3	0
Campanhas	1 ou 3	0
Conhecimento das novidades	1	0
Outros fornecedores	3 ou 4	0
Confidencialidade	1	0

Pela Tabela 4.32, pode-se observar que o inquérito da Empresa A é bem mais extenso do que o da Empresa D. A Tabela 4.33 evidencia isso mesmo.

4.3.1.2 Comparação de resultados

Ambas as BD referentes a estes inquéritos possuem a mesma variável, **Recomendar/-Recomendação**. O objetivo agora é comparar os resultados que estas empresas obtiveram

Tabela 4.33: Comparação dos inquéritos das Empresa A e D por número de perguntas.

Empresa	Total de perguntas/campos
A	$9 + (25 \text{ a } 40) = 34 \text{ a } 49$
D	$3 + 14 = 17$

onde 9 e 3 são os dados do cliente.

na mesma variável, embora que em períodos de tempo diferentes. Começando pela média[12] obtêm-se os resultados apresentados na Tabela 4.34.

Tabela 4.34: Comparação da média da variável recomendar/recomendação.

Empresa	Média
A	84
D	85.2

Pela Tabela 4.34 pode-se concluir que, em média, os resultados da Empresa D (satisfatórios) são superiores aos da Empresa A (insatisfatórios).

Será feita uma análise comparativa onde se considera o distrito a que pertence o cliente.

Nos distritos onde é possível fazer-se comparações (Tabela 4.35), em média:

- os resultados foram satisfatórios na Empresa D e insatisfatórios na Empresa A em: Aveiro, Braga, Lisboa, Portalegre, Setúbal, Vila Real e Viseu;
- os resultados foram satisfatórios na Empresa A e insatisfatórios na Empresa D em: Coimbra e Leiria;
- Nas restantes regiões os resultados foram igualmente satisfatórios/insatisfatórios, embora com valores diferentes.

Existem 33 NIF's de clientes que aparecem nas bases de dados duas empresas[13]. Compare-se a média as suas respostas.

Tabela 4.36: Média da variável recomendar nos clientes comuns nas 2 empresas.

Empresa A (n=34)	Empresa D (n=41)
82.4	82.2

Daqui resulta que os resultados foram insatisfatórios nas duas empresas.

[12] Numa escala de 100 pontos.

[13] O mesmo NIF pode aparecer mais do que uma vez. Pessoas diferentes respondem ao inquérito embora que sejam codificadas com o mesmo NIF.

Tabela 4.35: Frequência absoluta e relativa e média da variável recomendar nos distritos das 2 empresas.

Distrito	Frequência Absoluta (Relativa)		Média variável Recomendar	
	Empresa A	Empresa D	Empresa A	Empresa D
Aveiro	12 (6.3%)	199 (10.3%)	81.7	88.7
Beja	-	13 (0.7%)	-	90.0
Braga	5 (2.6%)	171 (8.9%)	80.0	87.0
Bragança	-	55 (2.9%)	-	90.0
Castelo Branco	11 (5.8%)	18 (0.9%)	82.0	81.7
Coimbra	13 (6.9%)	81 (4.2%)	88.3	84.4
Évora	-	17 (0.9%)	-	84.7
Faro	5 (2.6%)	78 (4.1%)	92.0	89.6
Guarda	1 (0.5%)	20 (1.0%)	100.0	90.5
Madeira	-	55 (2.9%)	-	87.6
Porto Santo	-	1 (0.1%)	-	80.0
São Jorge	-	3 (0.2%)	-	86.7
São Miguel	1 (0.5%)	22 (1.1%)	80.0	83.6
Corvo	-	2 (0.1%)	-	45.0
Pico	-	4 (0.2%)	-	97.5
Leiria	50 (26.5%)	119 (6.2%)	85.6	84.4
Lisboa	27 (14.3%)	208 (10.8%)	78.5	85.1
Portalegre	3 (1.6%)	21 (1.1%)	73.3	85.2
Porto	24 (12.7%)	170 (8.8%)	86.7	86.5
Santarém	20 (10.6%)	90 (4.7%)	86.0	86.8
Setúbal	3 (1.6%)	86 (4.5%)	80.0	85.2
Viana do Castelo	1 (0.5%)	27 (1.4%)	100.0	88.1
Vila Real	2 (1.1%)	27 (1.4%)	60.0	88.5
Viseu	4 (2.1%)	101 (5.3%)	80.0	88.0
NA	7 (3.7%)	335 (17.4%)	86.7	78.1

4.3.1.3 Comparação dos resultados obtidos no mesmo período de tempo

A comparação anterior foi feita em períodos de tempo diferentes. Como tal, selecionou-se o mesmo período de tempo nas duas empresas. Esse período de tempo comum é: 13/02/2017 a 14/02/2017 (dois dias). Desta forma, os resultados obtidos serão mais conclusivos que os anteriores. Obteve-se então o seguinte:

Tabela 4.37: Média e tamanho amostral da variável recomendar selecionando o mesmo período de tempo nas duas empresas.

Empresa	Tamanho amostral	Média
Empresa A	9	73.3
Empresa D	12	84.2

Donde se concluir que, em média, os resultados não atingiram o objetivo pretendido nas duas empresas.

Analogamente à Empresa A, a Empresa D também dispõe, na BD, uma variável designada por: Código do Segmento, mas com designações diferentes. Valores que pode tomar são: "a", "b", "c", "d", "e" e "f". Esta variável indica o nível de importância de um dado cliente: a categoria "f" é um código temporário sujeito a alteração após avaliação comercial do cliente, a "e" representa os clientes menos importantes e a categoria "a" os mais importantes pois são os que mais investem/compram, ao contrário dos outros. Fazendo a comparação por código do segmento, obtém-se:

Tabela 4.38: Média e tamanho amostral da variável recomendar selecionando o mesmo período de tempo nas duas empresas, por código do segmento.

Código do segmento	Empresa A ($n = 9$)	Empresa D ($n = 12$)
C0 / "f"	-	-
C1 / "e"	40 ($n = 2$)	100 ($n = 1$)
C2 / "d"	80 ($n = 3$)	80 ($n = 3$)
C3 / "c"	100 ($n = 1$)	80 ($n = 2$)
C4 / "b"	80 ($n = 3$)	85 ($n = 2$)
C5 / "a"	-	90 ($n = 1$)
NA	-	83.3 ($n = 3$)

Fazendo uma comparação por distritos obtêm-se os resultados apresentados na Tabela 4.39. Tal como anteriormente, devido ao baixo tamanho amostral em ambas as empresas, não é possível observar conclusões estatisticamente significativas. Contudo, no único caso onde é possível comparar (Santarém), observa-se que os resultados não atingiram o objetivo nas duas empresas.

Após cruzamento das duas bases de dados, é possível observar que não existe um grupo de clientes que tenha respondido ao inquérito das duas empresas no período de tempo selecionado

Tabela 4.39: Frequência absoluta e relativa e média da variável recomendar nos distritos selecionando o mesmo período de tempo nas duas empresas.

Distrito	Frequência Absoluta (Relativa)		Média da variável recomendar	
	Empresa A ($n = 9$)	**Empresa D** ($n = 12$)	**Empresa A**	**Empresa D**
Aveiro	1 (11.1%)	-	20	-
Beja	-	1 (8.3%)	-	90
Braga	-	1 (8.3%)	-	90
Castelo Branco	1 (11.1%)	-	40	-
Faro	-	2 (16.7%)	-	100
Madeira	-	1 (8.3%)	-	90
Leiria	5 (55.6%)	-	88	-
Lisboa	-	1 (8.3%)	-	80
Portalegre	-	1 (8.3%)	-	100
Porto	1 (11.1%)	-	80	-
Santarém	1 (11.1%)	4 (33.3%)	80	75
NA	-	1 (8.3%)	-	60

(período comum).

4.3.1.4 Comparação dos resultados obtidos num período de tempo equivalente

O período analisado anteriormente é bastante curto. Assim sendo, selecionou-se um período equivalente (não necessariamente igual) nas duas empresas. Considerou-se:

- **Empresa A:** 12/10/2016 a 04/11/2016 e 13/02/2017 a 09/03/2017;
- **Empresa D:** 01/10/2016 a 31/12/2016 e 01/01/2017 a 31/03/2017.

Desta forma, espera-se obter resultados mais significativos (estatiscamente) do que os apresentados anteriormente.

Obteve-se então o seguinte:

Tabela 4.40: Média e tamanho amostral da variável recomendar selecionando um período de tempo equivalente nas duas empresas.

Empresa	Tamanho amostral	Média
A	189	84.0
D	228	85.6

Donde se concluir que, em média, os resultados apenas atingiram o objetivo pretendido na Empresa D.

Comparando por código do segmento:

Tabela 4.41: Média e tamanho amostral da variável recomendar selecionando um período de tempo equivalente nas duas empresas, por código do segmento.

Código do segmento	**Empresa A** ($n = 189$)	**Empresa D** ($n = 228$)
C0 / "f"	92.7 ($n = 11$)	-
C1 / "e"	84.8 ($n = 33$)	86.0 ($n = 20$)
C2 / "d"	80.4 ($n = 46$)	87.1 ($n = 91$)
C3 / "c"	84.9 ($n = 46$)	85.7 ($n = 37$)
C4 / "b"	82.8 ($n = 36$)	89.2 ($n = 13$)
C5 / "a"	86.0 ($n = 10$)	87.5 ($n = 8$)
NA	86.7 ($n = 7$)	81.7 ($n = 59$)

Observa-se que a Empresa D, em média, obtém resultados satisfatórios e superiores aos da Empresa A.

Comparando por distritos, obtêm-se os resultados apresentados na Tabela 4.42 onde se verifica que, em média, ambas as empresas apresentam resultados satisfatórios e insatisfatórios.

Existem 6 NIF's de clientes que aparecem nas duas bases de dados após a seleção dos períodos em causa. Compara-se a média da variável recomendar neste caso:

Tabela 4.43: Média da variável recomendar nas duas empresas selecionando os mesmos clientes no período de tempo equivalente.

Empresa A (n=7)	**Empresa D (n=6)**
77.1	76.7

Onde se verifica que em ambas as empresas, os resultados não atingiram o objetivo que, após visualização da Figura 4.27, se pode constatar que se deve ao facto de, na Empresa A, um cliente que atribuiu 10 na variável recomendar e de, na Empresa A, a maioria das classificações estarem abaixo do pretendido (maioritariamente 60 e 80).

Tabela 4.42: Frequência absoluta e relativa e média da variável recomendar nos distritos selecionando um período de tempo equivalente nas duas empresas.

Distrito	**Frequência Absoluta (Relativa)**		**Média da variável recomendar**	
	Empresa A ($n = 189$)	**Empresa D** ($n = 228$)	**Empresa A**	**Empresa D**
Aveiro	12 (6.3%)	23 (10.1%)	81.7	94.3
Beja	-	4 (1.8%)	-	90.0
Braga	5 (2.6%)	21 (9.2%)	80.0	82.9
Bragança	-	7 (3.1%)	-	94.3
Castelo Branco	11 (5.8%)	1 (0.4%)	82.0	10.0
Coimbra	13 (6.9%)	10 (4.4%)	88.3	88.0
Évora	-	3 (1.3%)	-	76.7
Faro	5 (2.6%)	15 (6.6%)	92.0	97.1
Guarda	1 (0.5%)	6 (2.6%)	100.0	90.0
Madeira	-	7 (3.1%)	-	85.7
Porto Santo	-	1 (0.4%)	-	80.0
São Miguel	1 (0.5%)	3 (1.3%)	80.0	70.0
Corvo	-	1 (0.4%)	-	50.0
Leiria	50 (26.5%)	16 (7.0%)	85.6	90.0
Lisboa	27 (14.3%)	24 (10.5%)	78.5	80.4
Portalegre	3 (1.6%)	6 (2.6%)	73.3	90.0
Porto	24 (12.7%)	15 (6.6%)	86.7	79.3
Santarém	20 (10.6%)	8 (3.5%)	86.0	78.8
Setúbal	3 (1.6%)	10 (4.4%)	80.0	86.0
Viana do Castelo	1 (0.5%)	3 (1.3%)	100.0	86.7
Vila Real	2 (1.1%)	5 (2.2%)	60.0	92.0
Viseu	4 (2.1%)	11 (4.8%)	80.0	95.5
NA	7 (3.7%)	28 (12.3%)	86.7	77.5

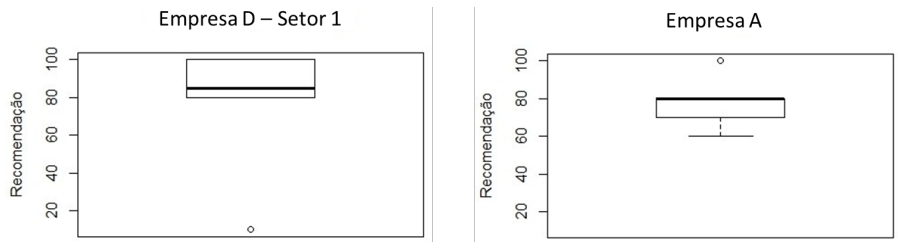


Figura 4.27: Boxplot da variável recomendar para os mesmos clientes no período de tempo equivalente.

4.3.2 Restantes empresas

Relativamente à comparação dos resultados da análise descritiva e exploratória, não se sentiu necessidade (nem houve tempo útil para tal, pois a base de dados da Empresa E foi fornecida numa fase bastante adiantada do estágio) de se efetuar uma comparação entre os resultados da Empresa E com as restantes (em específico com a Empresa A e D).

Quanto à comparação dos resultados da análise descritiva dos clientes comuns em ambas as empresas, não foi possível fazer essa análise[14] entre Empresa A/Empresa E e Empresa D/Empresa E pois a base de dados da Empresa E não possui qualquer elemento de comparação[15] que se possa usar para identificar os mesmos clientes em ambas as bases de dados.

[14] Idêntica à apresentada anteriormente.

[15] Por exemplo, NIF, número de cliente, etc.

Capítulo 5.

Conclusão

"How absurdly simple!", I cried.

"Quite so!", said he, a little nettled. "Every problem becomes very childish when once it is explained to you."

Arthur Conan Doyle - The Adventure of the Dancing Men

5.1 Conclusões gerais

Com este trabalho, começando pela análise descritiva e exploratória dos dados das várias empresas do grupo NORS, foi possível identificar o grupo de clientes (por exemplo, clientes com um dado revendedor local/código loja) das mesmas que mereciam especial atenção devido ao facto de os seus resultados (médios) estarem abaixo do pretendido. Dentro daquilo que foi possível, os resultados das várias empresas foram comparados uns com os outros onde as conclusões podem ser diversas (devido à distribuição dos resultados) e já foram apresentadas anteriormente. Para além disso, a aplicação do NPS a todas as empresas da região Ibéria, permitindo a sua comparação, motivou a futura implementação deste em todo o grupo NORS (enquanto indicador único de sucesso/satisfação).

Numa mais avançada deste projeto, o uso da regressão logística, após definição da variável resposta, permitiu identificar os aspetos (com 95% de confiança) que, ao serem melhorados, iriam fazer com que o grupo obtive-se melhores resultados na variável objetivo. Na generalidade, todos os modelos obtidos tiveram um bom poder discriminante e uma boa exatidão. Este procedimento foi designado como **identificação dos *drivers* de excelência**. As variáveis pre-

sentes nos modelos obtidos são, então, **os *drivers* de excelência** que, dada a sua diversidade, têm de ser consultados individualmente por empresa (resultados apresentados anteriormente) não sendo possível obter uma conclusão geral sobre estes, para além daquilo que já foi referido. De facto, esta metodologia impressionou tanto o grupo que foi usada como ferramenta de apoio às decisões comerciais e pretende-se que esta seja também aplicada em todo o grupo NORS.

Já na fase final, pretendeu-se verificar se um cliente mais satisfeito iria proporcionar um aumento das vendas. Apenas se verificou relação na Empresa D, para a margem percentual das vendas, onde o modelo obtido (usando uma amostra de 17 clientes e com AIC=205.630), com recurso a modelos lineares gerais, estimou (com 95% de confiança) que, em média, um cliente passar de neutro a promotor (na variável recomendar) resulta num aumento de 3.938 pontos percentuais na margem percentual das vendas.

5.2 Limitações

Dada a forma como o grupo dispunha dos seus dados, sentiram-se algumas dificuldades no cruzamento de bases de dados, nomeadamente no cruzamento da base de dados dos inquéritos com outras bases de dados do grupo.

Para além disso, relativamente às bases de dados dos inquéritos, não se conseguiu garantir que quem responde aos inquéritos é de facto quem compra, e isso pode de certa forma influenciar os resultados obtidos.

5.3 Trabalho futuro

Não se pretende que este trabalho constitua um ponto final na abordagem ao problema colocado. Visto nunca nada semelhante ter sido feito anteriormente no grupo, este foi o passo inicial no sentido da compreensão de um fenómeno e espera-se que seja uma ajuda para quem o decida abraçar noutro momento.

Este projeto suscitou o surgimento das seguintes ideias:

- perante a existência de dados, efetuar uma análise (longitudinal) de dados binários, ou

seja, analisar todos os inquéritos e ficar apenas com os dados dos clientes que responderam ao mesmo em vários instantes temporais. O objetivo será aplicar um modelo marginal GEE ou um modelo linear generalizado misto (por exemplo) para dados binários. Isto iria resultar numa melhoração da identificação dos *drivers* de excelência;

- Verificar se as vendas tem alguma relação com alguma das outras questões efetuadas nos diferentes inquéritos (por exemplo: satisfação com os prazos de entrega). A abordagem poderá ou não ser idêntica com a que se apresentou aqui;

- Estudar se o facto de os clientes serem Promotores (na variável recomendação/recomendar) leva ao surgimento de uma quantidade significativa de novos clientes.

Estas ideias despertaram grande interesse no grupo e seriam ideias para um novo projeto no futuro.

Para além disso, seria uma boa ideia replicar o trabalho aqui desenvolvido internacionalmente a todas as empresas do grupo espalhadas pelo mundo, algo que foi solicitado pelos administradores pois estes ficaram bastante fascinados com o trabalho desenvolvido no âmbito deste projeto.

Bibliografia

Action, P. *Análise de Variância (Teste F) - Medidas de associação.* URL: http://www.portalaction.com.br/analise-de-regressao/24-analise-de-variancia-teste-f-medidas-de-associacao (acedido em 25/07/2018).

Amaral Turkman M.A. e Silva, G. (2000). *Modelos Lineares Generalizados - da teoria à prática.* Sociedade Portuguesa de estatística.

Barnes, R. (2006). *Matrix Differentiation.* Department of Civil Engineering, University of Minnesota Minneapolis, Minnesota, USA.

Conceito de satisfação do cliente (2013). URL: https://conceito.de/satisfacao-do-cliente (acedido em 02/07/2018).

Cordeiro, V. (out. de 2017). «Modelação do cancelamento de apólices de seguro automóvel, por parte do cliente». Instituto Superior de Engenharia do Porto (ISEP).

Cox, D. e D. Hinkley (1974). *Theoretical Statistics.* Chapman e Hall, London.

Davidian, M. e D. M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data.*

Diggle et al, P. (1994). *Applied Longitudinal Analysis.* Oxford Statistical Science Series.

Documento interno: Company profile (2017). NORS.

Documento interno: Net Promoter Score (2018). NORS.

Duarte, T. (2012). *São necessários 05 comentários positivos para neutralizar 01 negativo.* URL: https://satisfacaodeclientes.com/sao-necessarios-05-comentarios-positivos-para-neutralizar-01-negativo/ (acedido em 03/09/2018).

Gaio, R. (2012). *Apontamentos da unidade curricular Estatística Aplicada às Ciências e Engenharia. Ficheiro: Modelos Lineares Generalizados.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2016a). *Apontamentos da unidade curricular Estatística Aplicada às Ciências e Engenharia. Ficheiro: Regressão Linear.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2016b). *Apontamentos da unidade curricular Estatística Aplicada às Ciências e Engenharia. Ficheiro: Regressão Logística.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2018a). *Apontamentos da unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia. Ficheiro: Decomposição da Matriz Var-Cov dos Erros.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2018b). *Apontamentos da unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia. Ficheiro: Decomposição da Matriz Var-Cov dos Erros.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2018c). *Apontamentos da unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia. Ficheiro: Estruturas de Correlação Serial.* Faculdade de Ciências da Universidade do Porto.

Gaio, R. (2018d). *Apontamentos da unidade curricular Modelos Estatísticos Avançados em Ciências e Engenharia. Ficheiro: Matrizes Var-Cov dos Efeitos Aleatórios.* Faculdade de Ciências da Universidade do Porto.

Gauge, C. *NPS Benchmarks.* URL: `https://npsbenchmarks.com/` (acedido em 04/04/2018).

Grupo Nors (2014). URL: `http://www.nors.com/pt.aspx` (acedido em 03/01/2018).

Hosmer, D. e S. Lemeshow (2013). *Applied Logistic Regression.* 3ª ed. John Wiley & Sons, Inc. ISBN: 0470582472.

McCullagh, P. e J. Nelder (1989). *Generalized Linear Models.* 2ª ed. Chapman e Hall, London.

Nelder, J. e R. Wedderburn (1972). *Generalized linearmodels.* Journal of the Royal Statistical Society, A 135, 370-384.

NPS (2017). URL: `https://www.netpromoter.com/know/` (acedido em 16/10/2017).

Oliveira, A. (2018). *Vantagens da satisfação dos clientes para a empresa.* URL: `https://www.cpt.com.br/dicas-cursos-cpt/vantagens-da-satisfacao-dos-clientes-para-a-empresa` (acedido em 02/07/2018).

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

R. URL: `https://www.r-project.org/about.html`.

R Markdown. URL: `https://rmarkdown.rstudio.com/rmarkdown_websites.html`.

Reichheld, F. (2003). *The One Number You Need to Grow*. URL: https://hbr.org/2003/12/the-one-number-you-need-to-grow (acedido em 03/04/2018).

Reichheld, F. e R. Markey (2011). *The Ultimate Question 2.0 (Revised and Expanded Edition): How Net Promoter Companies Thrive in a Customer-Driven World*. Brain & Company.

Relatório e contas consolidadas (2017). URL: http://www.nors.com/media/431276/nors_pt_2017_web.pdf (acedido em 19/07/2018).

RStudio. URL: https://www.rstudio.com/.

Sen, P. e J. Singer (1993). *Large Sample Methods in Statistics*. An Introduction with Applications. Chapman e Hall, New York.

Torgo, L. (2017). *Apontamentos da unidade curricular Data Mining I. Ficheiro: The R Software Environment - a (very) short introduction*. Faculdade de Ciências da Universidade do Porto.

Verbeke, G. e G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Wikipedia (2017). *Restricted maximum likelihood*. URL: https://en.wikipedia.org/wiki/Restricted_maximum_likelihood (acedido em 05/07/2018).

Anexos

Anexo 1

Cálculo diferencial (real) matricial

O assunto aqui descrito pode ser visto com mais detalhe em Barnes (2006).

Sejam $\mathbf{x}$, $\mathbf{y}$ e $\mathbf{w}$ vetores, em $\mathbb{R}^n$, e A uma matriz, em $\mathbb{R}^{n\times n}$. Destacam-se as seguintes propriedades

1. $\dfrac{\partial}{\partial \mathbf{w}}(A\mathbf{x}) = A\dfrac{\partial \mathbf{x}}{\partial \mathbf{w}}$
2. $\dfrac{\partial}{\partial \mathbf{x}}(\mathbf{y}^T A\mathbf{x}) = \mathbf{y}^T A$
3. $\dfrac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = \mathbf{x}^T(A + A^T)$
   A simétrica $(A = A^T) \Rightarrow \dfrac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = 2\mathbf{x}^T A$
4. $\dfrac{\partial}{\partial \mathbf{w}}(\mathbf{y}^T\mathbf{x}) = \mathbf{x}^T\dfrac{\partial \mathbf{y}}{\partial \mathbf{w}} + \mathbf{y}^T\dfrac{\partial \mathbf{x}}{\partial \mathbf{w}}$
   $\dfrac{\partial}{\partial \mathbf{w}}(\mathbf{x}^T\mathbf{x}) = 2\mathbf{x}^T\dfrac{\partial \mathbf{x}}{\partial \mathbf{w}}$

Propriedades sobre matrizes e vetores

Seja $\mathbf{v}$ um vetor em $\mathbb{R}^n$ e A uma matriz simétrica em $\mathbb{R}^{n\times n}$. Destaca-se o seguinte

a) A é **definida positiva** se $\mathbf{v}^T A\mathbf{v} > 0$, $\forall\ \mathbf{v} \neq \mathbf{0}$, isto é, se todos os valores próprios de A são maiores do que zero, ou seja, se $\det(A) \neq 0$.

b) A é **semi-definida positiva** se $\mathbf{v}^T A\mathbf{v} \geq 0$, $\forall\ \mathbf{v} \neq \mathbf{0}$, isto é, se todos os valores próprios de A são maiores ou iguais a zero.

c) $||\mathbf{v}||^2 = \mathbf{v}^T\mathbf{v}$.

Matriz de variância-covariância

Seja $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ um vetor de n variáveis aleatórias. A sua matriz de variância-covariância (var-cov) é dada por

$$Cov(\mathbf{Y}) \equiv Var(\mathbf{Y}) = E((\mathbf{Y} - E(\mathbf{Y})) \centerdot (\mathbf{Y} - E(\mathbf{Y}))^T) = \Sigma =$$

$$\begin{pmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \ldots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & \ldots & Cov(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \ldots & Var(Y_n) \end{pmatrix}$$

Destaca-se que esta matriz é simétrica pois $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$, $\forall\ i, j$.

Proposição: Qualquer matriz var-cov é semi-definida positiva.

Demonstração. Seja $\mathbf{v} \in \mathbb{R}^n \backslash \{0\}$ um vetor, Σ uma matriz de var-cov e $\mu = E(\mathbf{Y})$. Então

$$\begin{aligned} \mathbf{v}^T \Sigma \mathbf{v} =& \mathbf{v}^T E((\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T)\mathbf{v} = E(\mathbf{v}^T(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T\mathbf{v}) \\ =& E(((\mathbf{Y} - \mu)^T\mathbf{v})^T(\mathbf{Y} - \mu)^T\mathbf{v}) = E(||(\mathbf{Y} - \mu)^T\mathbf{v}||^2) \\ =& E(|(\mathbf{Y} - \mu)^T\mathbf{v}|^2) \geq 0 \end{aligned}$$

□

Anexo 2

Modelação da Heteroscedasticidade

Apresentam-se de seguida várias funções de variância para a modelação da **variabilidade intra-indivíduo**[1], ou seja, para a modelação da matriz W_i.

Davidian e Giltinan (1995): pode-se escrever

$$Var(u_{it}|\mathbf{b}_i) = \sigma^2 g^2(\mu_{it}, \mathbf{v}_{it}, \boldsymbol{\delta}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, t_i$$

onde

- $\mu_{it} = E(y_{it}|\mathbf{b}_i)$. Como esta quantidade depende dos efeitos fixos e aleatórios, também a função de variância pode depender desses efeitos;
- $\mathbf{v}_{it}$ é o vetor de **covariáveis da variância**;
- $\boldsymbol{\delta}$ é o vetor de **parâmetros da variância**;
- g é a uma **função de variância**, contínua em $\boldsymbol{\delta}$ (exponencial, logaritmo, potência, etc.).

Funções de variância disponíveis no pacote nlme

Veja-se primeiramente a estrutura geral:

`varFunc(value,form)`

onde

- `value`: especifica os valores dos parâmetros da variância δ
- `form`: formula $1-$ *sided* do tipo

(covariáveis da variância)|(variável de estratificação).

Havendo variável de estratificação, são usados parâmetros diferentes por estrato.

[1] De forma mais geral, a variabilidade dentro de cada unidade experimental.

Podem-se usar, com o pacote `nlme`, as seguintes funções de variância:

- **varFixed:** $Var(u_{it}) = \sigma^2 \mathbf{v}_{it}$. Neste caso, a variância é fixa, não existem parâmetros $\boldsymbol{\delta}$ a estimar, há apenas 1 covariável $\mathbf{v}_{it}$, não existem variáveis de estratificação e só existe um argumento, `value`, que é uma fórmula $1 - sided$;

- **varIdent:** $Var(u_{it}) = \sigma^2 \delta_{s_{it}}$. Neste caso, existe uma variável de estratificação (categórica) s com S estratos onde $s \in \{1, 2, \ldots, S\}$ (mais do que uma: juntar numa única), as variâncias são diferentes por estrato da variável de estratificação, usam-se $S-1$ parâmetros para representar as variáveis dentro dos estratos (δ_1 : d.padrão no 1^{o} estrato (faz-se $\delta_1 = 1$), δ_2 =(d.padrão 2^{o} estrato)/(d.padrão 1^{o} estrato), δ_3 =(d.padrão 3^{o} estrato)/(d.padrão 1^{o} estrato), etc.) e o argumento `form` é da forma $\sim 1|s$;

- **varPower:** $Var(u_{it}) = \sigma^2 |\mathbf{v}_{it}|^{2\boldsymbol{\delta}}$. Neste caso, a função de variância é uma potência do valor absoluto de uma covariável $\mathbf{v}$, ou seja, $g(\mathbf{v}_{it}, \boldsymbol{\delta}) = |\mathbf{v}_{it}|^{\boldsymbol{\delta}}$, existe um parâmetros ($\boldsymbol{\delta}$) que varia em $\mathbb{R}$ e a variância pode aumentar ou diminuir com o valor absoluto da covariável (parâmetros positivo ou negativo, respetivamente). Para além disso, este modelo de variância não pode ser usado com covariáveis que tomem o valor 0 pois origina problemas na estimação;

- **varExp:** $Var(u_{it}) = \sigma^2 \exp(2\boldsymbol{\delta}\mathbf{v}_{it})$. Neste caso, a função de variância é uma função exponencial da covariável, ou seja, $g(\mathbf{v}_{it}, \boldsymbol{\delta}) = \exp(\sigma \mathbf{v}_{it})$, existe um parâmetros ($\boldsymbol{\delta}$) que varia em $\mathbb{R}$, a variância pode aumentar ou diminuir com o valor da covariável (parâmetros positivo ou negativo, respetivamente) e não há restrições quantos aos valores que a covariável pode tomar;

- **varConstPower:** $Var(u_{it}) = \sigma^2(\delta_1 + |\mathbf{v}_{it}|^{\delta_2})^2$. Neste caso, a função de variância é $g(\mathbf{v}_{it}, \boldsymbol{\delta}) = \delta_1 + |\mathbf{v}_{it}|^{\delta_2}$ que corresponde à soma de uma constante com uma potência de uma covariável, δ_1 tem de ser positivo, δ_2 varia em $\mathbb{R}$ e podem ser usadas variáveis de estratificação. Para além disso, este modelo pode ser usado em função de varPower quando a covariável toma valor 0. Assim, quando a covariável toma valores perto de 0, a função de variância é essencialmente constante e igual a δ_1 e, quando a covariável se

afasta de 0, a função de variância aumenta ou diminuiu com o valor absoluto da covariável conforme $\delta_2 > 0$ ou $\delta_2 < 0$, respetivamente;

- **varComb:** permite a combinação de dois ou mais modelos de variância, apresentados anteriormente, através do produto das correspondentes funções de variância;
- **Outras estruturas de variância:** outras estruturas de variância (sempre de classe `varFunc`) podem ser definidas e construídas pelo utilizador.

Estruturas de Correlação Serial[2]

Veja-se agora como modelar a matriz C_i, ou seja, como modelar a dependência entre os erros de cada unidade experimental[3].

Estruturas de Correlação

Sobre estruturas de correlação pode dizer-se o seguinte:

- historicamente, foram desenvolvidas para séries temporais e para dados espaciais;
- assume-se a mesma estrutura de correlação para os erros de todas as unidades experimentais;
- assume-se que os erros aleatórios de uma unidade experimental, u_{it}, estão associados a vetores posição, $\mathbf{p}_{it}$ (usualmente escalares inteiros, no caso de séries temporais);
- só se assumem estruturas de correlação isotrópicas: a correlação entre os erros u_{it} e $u_{it'}$ depende apenas da posição relativa entre os vetores posição $\mathbf{p}_{it}$ e $\mathbf{p}_{it'}$ e não dos valores desses tempos;
- a expressão geral para a estrutura de correlação dentro da unidade experimental é: $Corr(u_{it}, u_{it'}) = h(d(\mathbf{p}_{it}, \mathbf{p}_{it'}), \boldsymbol{\rho})$, para $i = 1, \ldots, n$ e $t, t' = 1, \ldots, t_i$, onde $\boldsymbol{\rho}$ é um vetor de parâmetros de correlação e $h(\cdot)$ é uma função de correlação tomando valores entre -1 e 1, contínua em $\boldsymbol{\rho}$ e tal que $h(0, \boldsymbol{\rho}) = 1$;

Notar que quanto mais próximo (no tempo ou espaço) estiverem dois erros de uma mesma unidade experimental, maior será a sua dependência.

[2] *Correlation structure of the within-group erros.*

[3] *within-group erros.*

Estruturas de Correlação Serial

Sobre estruturas de correlação serial pode dizer-se o seguinte:

- são usadas para modelar dependências em dados de séries temporais;
- simplifica-se a condição de isotropia para: $Corr(u_{it}, u_{it'}) = h(|\mathbf{p}_{it} - \mathbf{p}_{it'}|, \boldsymbol{\rho})$ onde a função h é designada por **função de auto-correlação**;
- usualmente, exigem que os dados sejam observados em instantes de tempo indexados por números inteiros.

Estruturas mais frequentes

De seguida apresentam-se as estruturas de correlação mais frequentes. Estão disponíveis na biblioteca `nlme` e correspondem a classes de objetos do tipo `corStruct`. Por defeito, os modelos estimados pela instrução `lme` têm `correlation=NULL` o que corresponde a erros intra-indivíduo não correlacionados (e com variâncias iguais).

Podem usar-se, então, as seguintes estruturas:

- **Estrutura Geral**: corresponde ao modelo de correlação $Corr(u_{it}, u_{it'}) = \rho_{tt'}$ o que dá origem à matriz de correlações $\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1t_i} \\ \rho_{12} & 1 & \rho_{23} & \dots & \rho_{2t_i} \\ \rho_{13} & \rho_{23} & 1 & \dots & \rho_{3t_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1t_i} & \rho_{2t_i} & \rho_{3t_i} & \dots & 1 \end{pmatrix}$. Neste caso, cada correlação entre os erros é representada por um parâmetro diferente e o número de parâmetros depende de forma quadrática do número de observações da unidade experimental (havendo t_i observações para uma unidade experimental, o número de parâmetros é $t_i(t_i-1)/2$). Para além disso, este modelo de correlação é útil apenas quando existem poucas observações por unidade experimental;
- **Estrutura de Simetria Composta:** corresponde à estrutura de correlação serial mais simples cujo modelo de correlação é $Corr(u_{it}, u_{it'}) = \rho$, $\forall t \neq t'$, e que dá origem à matriz

de correlações $\begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}$. Neste caso, assume-se igual correlação entre todos os erros de uma mesma unidade experimental e é estimado um único parâmetro $\rho \in [-1, 1]$ que é por vezes designado por **coeficiente de correlação intra-classe**[4];

- **Estrutura Auto-Regressiva de Ordem 1 - AR(1):** corresponde ao modelo de correlação $Corr(u_{it}, u_{it'}) = \rho^{t-t'}$ o que dá origem à matriz de correlações $\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{t_i-1} \\ \rho & 1 & \rho & \dots & \rho^{t_i-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{t_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t_i-1} & \rho^{t_i-2} & \rho^{t_i-3} & \dots & 1 \end{pmatrix}$. Neste caso, a correlação decresce exponencialmente com o valor absoluto do espaçamento e é estimado um único parâmetro $\rho \in [-1, 1]$. Para além disso, o modelo anterior corresponde a ter $u_{it} = \rho u_{i,t-1} + a_{it}$, onde a_{it} é um ruído, $E(a_{it}) = 0$, a_{it} independente de $u_{i,t-1}$ e a_{it} homocedásticos para t;

- **Estrutura Auto-Regressiva de Ordem p - AR(p):** corresponde ao modelo linear $u_{it} = \rho_1 u_{i,t-1} + \rho_2 u_{i,t-2} + \cdots + \rho_p u_{i,t-p} + a_{it}$ com $E(a_{it}) = 0$, a_{it} independente das observações anteriores a u_{it} e a_{it} homocedásticos para t. A ordem de um modelo autoregressivo diz respeito ao número de observações anteriores incluídas no modelo linear anteriormente descrito. Neste caso, a distância, *lag*, entre duas observações u_{it} e $u_{it'}$ é dada por $|t - t'|$ (o modelo AR(1) corresponde a ter *lag* 1) e assume-se que os dados são observados em instantes de tempo inteiros;

- **Estrutura autoregressiva para tempo contínuo - CAR(p):** corresponde a uma generalização do modelo AR(1) a tempos de medição contínuos (portanto a situações em que os espaçamentos temporais entre os dados não têm de ser inteiros). O modelo CAR(1) - *Continuous AutoRegressive model of order 1* - tem uma função de autocorrelação definida

[4] *Intra-class correlation coefficient.*

por $h(s,\phi)=\phi^s$, $s\geq 0$ e $\phi\geq 0$ que é o seu único parâmetro. Salienta-se que a função de autocorrelação de modelos autoregressivos contínuos de ordem superior a 1 não tem uma fórmula exata sendo definida recursivamente por uma equação às diferenças (Pinheiro e Bates (2000));

- **Modelos de médias móveis - MA(q):** são usados para observações realizadas em tempos inteiros. Cada observação é uma combinação linear de termos de ruído i.i.d. $u_{it}=\theta_1 a_{i,t-1}+\cdots+\theta_q a_{i,t-p}+a_{it}$, onde q é o número de termos de ruído, ou seja, a **ordem do modelo**. A função de correlação é $h(k,(\theta_1,\ldots,\theta_q))=\begin{cases}\frac{\theta_k+\theta_1\theta_{k-1}+\cdots+\theta_{k-q}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2}, & k=1,\ldots,q\\ 0, & k>q\end{cases}$, onde se observa que os parâmetros a estimar são $(\theta_1,\ldots,\theta_q)$ e que as observações separadas por mais de q unidades de tempo são não correlacionadas. No caso específico de $q=1$ obtem-se o modelo MA(1) que corresponde ao seguinte modelo de correlação $Corr(u_{it},u_{it'})=\begin{cases}\frac{\theta}{1+\theta^2}, & |t-t'|=1\\ 0, & |t-t'|>1\end{cases}$.
- **Modelos Autoregressivos - Médias Móveis - ARMA(p,q):** são usados para observações realizadas em tempos inteiros. Para além disso, combinam modelos autoregressivos com modelos de médias móveis da seguinte forma: $u_{it}=AR(p)+MA(q)=\sum_{s=1}^{p}\rho_s u_{i,t-s}+\sum_{k=1}^{q}\theta_k u_{i,t-k}+a_{it}$, onde os parâmetros a estimar são $(\rho_1,\ldots,\rho_p)$ e $(\theta_1,\ldots,\theta_q)$. Desta forma, salienta-se que (convenção) $ARMA(p,0)=AR(p)$ e $ARMA(0,q)=MA(q)$. A expressão (geral) para a função de correlação pode ser encontrada em Pinheiro e Bates (2000). No caso particular de $p=q=1$ obtém-se o modelo ARMA(1,1) que corresponde ao seguinte modelo de correlação $Corr(u_{it},u_{it'})=\begin{cases}\frac{(1+\rho\theta)(\rho+\theta)}{1+\theta^2+2\rho\theta}=:\phi_1, & |t-t'|=1\\ \rho^{|t-t'-1|}\phi_1, & |t-t'|>1\end{cases}$.

Função de autocorrelação empírica

Seguem-se algumas consideração sobre a função de autocorrelação empírica:

- é uma estimativa não paramétrica da função de auto-correlação;
- avalia a semelhança entre os resíduos do modelo em função do espaçamento temporal entre eles;

- para remover os efeitos das variáveis explicativas, usa os resíduos padronizados (média 0 e variância 1) do modelo linear misto: $r_{it} = \dfrac{y_{it} - \widehat{y}_{it}}{\widehat{\sigma}^2_{it}}$, $\widehat{\sigma}^2_{it}$ estimador de $Var(u_{it})$;

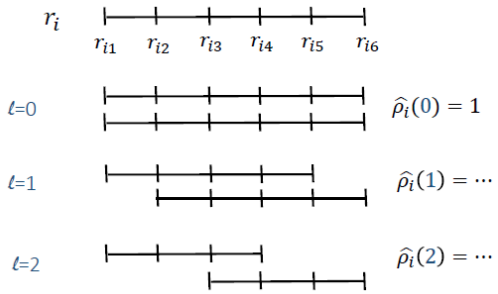


Figura .1: Exemplificação da obtenção da função de autocorrelação empírica. Fonte: Gaio (2018c).

- a **função de autocorrelação no espaçamento (lag)** ℓ é $\widehat{\rho}(\ell) = \dfrac{\sum_{i=1}^{n}\sum_{t=1}^{t_i-\ell} r_{it}r_{i,t+\ell}/N(\ell)}{\sum_{i=1}^{n}\sum_{t=1}^{t_i} r_{it}^2/N(0)}$ sendo $N(\ell)$ o número total de pares de resíduos com espaçamento ℓ;
- recorde-se que $\widehat{\rho}_{XY} = \dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$. Se X e Y forem centrados $\widehat{\rho}_{XY} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$;
- o gráfico da função de auto-correlação permite identificar a estrutura de correlação serial a usar se as observações forem igualmente espaçadas. Caso contrário, não pode ser estabelecido um espaçamento constante entre os tempos;

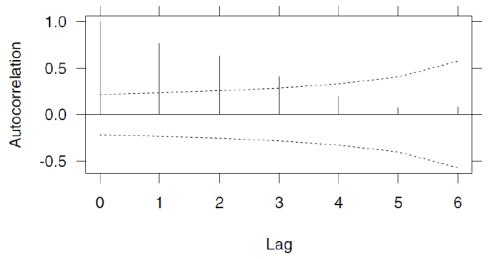


Figura .2: Exemplo do gráfico da função de auto-correlação. Fonte: Gaio (2018c).

- sob a condição de não correlação, um coeficiente de correlação tem um desvio padrão aproximadamente igual a $1/\sqrt{N}$ onde N é o número de pares independentes usados no cálculo;
- usando a regra anterior, definem-se limites no gráfico da função de autocorrelação para ter ACF=0. Para um espaçamento ℓ, os limites são $\left(-\dfrac{2}{N(\ell)}, \dfrac{2}{N(\ell)}\right)$ onde $N(\ell)$ representa o número de pares de observações com espaçamento ℓ;

- a título de exemplo, na Figura .2, a função de autocorrelação empírica indica que os resíduos de cada unidade experimental são correlacionados e que a correlação diminui com o espaçamento. Isto sugere que uma estrutura do tipo AR(1) poderia ser adequada. Nota-se assim, a utilidade do que foi aqui descrito.

Variograma

A presente exposição teórica segue de perto apontamentos fornecidos pela professora doutora Sandra Ramos.

O variograma, usado tradicionalmente para estudar estruturas de correlação espaciais (ver Diggle et al (1994)), fornece uma ferramenta alternativa para a exploração da estrutura de correlação quando as observações longitudinais das unidades experimentais não são todas obtidas nos mesmos instantes de tempo[5]. Suponha-se também que esses instantes não são igualmente espaçados.

A função do variograma é definida como

$$\gamma(\ell) = \frac{1}{2}E[\{Y(t) - Y(t-\ell)\}^2], \tag{.1}$$

onde ℓ é um espaçamento (*lag*) positivo e t é um tempo qualquer.

O variograma é aplicável a processos fracamente estacionários[6]. Note-se que não existe t no lado esquerdo da equação (.1).

Estacionaridade fraca[7] significa que, fixando um ℓ (qualquer), para todo o t tem-se o seguinte:

- **Média constante:**

$$E\{Y(t)\} = E\{Y(t-\ell)\} = \mu$$

- **Variância constante:**

$$\text{var}\{Y(t)\} = \text{var}\{Y(t-\ell)\} = \sigma^2$$

- **A correlação apenas depende de ℓ:**

$$\text{corr}\{Y(t), Y(t-\ell)\} = \rho(\ell)$$

[5] Dados espaçados no tempo de forma irregular.

[6] *Weakly stationary processes.*

[7] *Weak stationarity.*

Sobre estas condições resulta e demonstra-se que o variograma ($\gamma(\ell)$) relaciona-se com a função de autocorrelação ($\rho(\ell)$) da seguinte forma:

$$\gamma(\ell) = \sigma^2\{1 - \rho(\ell)\} \tag{.2}$$

Para que a média seja constante, os variogramas devem ser obtidos com base nos resíduos, removendo os efeitos do tempos e de outras variáveis explicativas importantes.

Interpretação do variograma

A equação (.2) pode ser reescrita da seguinte forma:

$$\rho(\ell) = 1 - \gamma(\ell)/\sigma^2.$$

Com base nesta expressão, é possível aferir que:

- Quando $\widehat{\gamma}(\ell)$ tende para 0, a autocorrelação tende para 1;
- Quando $\widehat{\gamma}(\ell)$ aumenta, a autocorrelação diminui;
- Quando $\widehat{\gamma}(\ell)$ tende para $\widehat{\sigma}^2$, a autocorrelação tende para 0;
- Uma lacuna entre $\widehat{\gamma}(\ell)$ e $\widehat{\sigma}^2$ para valores elevados de ℓ indica uma autocorrelação positiva para espaçamentos (*lag's*) ainda maiores.

Anexo 3

Exemplos de testes de hipóteses para combinações lineares de $\boldsymbol{\beta_k}$'s

No que se segue, considere-se que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$.

Caso 1: L é um vetor linha.

Exemplo 1: Suponha-se que se pretende testar $\begin{cases} H_0\text{: } \beta_1 = \beta_2 \Leftrightarrow L\boldsymbol{\beta} = 0 \\ H_1\text{: } \beta_1 \neq \beta_2 \Leftrightarrow L\boldsymbol{\beta} \neq 0 \end{cases}$. Nesta situação, $L = (0, 1, -1, 0)$.

Exemplo 2: Suponha-se que se pretende testar $\begin{cases} H_0\text{: } \beta_2 = 0 \Leftrightarrow L\boldsymbol{\beta} = 0 \\ H_1\text{: } \beta_2 \neq 0 \Leftrightarrow L\boldsymbol{\beta} \neq 0 \end{cases}$. Nesta situação, $L = (0, 0, 1, 0)$.

Caso 2: L é uma matriz.

Exemplo: Suponha-se que se pretende testar $\begin{cases} H_0\text{: } \beta_0 = \beta_1 = \beta_2 \Leftrightarrow L\boldsymbol{\beta} = 0 \\ H_1\text{: } \cdots \Leftrightarrow L\boldsymbol{\beta} \neq 0 \end{cases}$. Nesta situação,

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix} \Leftrightarrow \begin{cases} \beta_0 = \beta_1 \\ \beta_1 = \beta_2 \end{cases} \Leftrightarrow \beta_0 = \beta_1 = \beta_2.$$

Anexo 4

Seguem-se os resultados da regressão logística para as empresas B e C por tipo e canal de compra mais frequente.

Resultados por tipo de compra mais frequente

Veja-se que resultados se obtiveram filtrando as bases de dados para o tipo de compra mais frequente. Relembra-se que, em ambas as empresas, o tipo de compra mais frequente é do tipo mecânica.

Após se aplicar a metodologia referida em 3.1.9.4 excluíram-se as seguintes variáveis das bases de dados:

Empresa B

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado

Empresa C

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado
- Canal de compra.

Ficou-se, assim, com 23 variáveis na Empresa B (retiraram-se 4) e 22 na Empresa C (retiraram-se 5) que foram consideradas na modelação.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), pelas 2 empresas, segue a seguinte tabela com a frequência absoluta e relativa.

Tabela .1: Distribuição da variável resposta nas Empresa B e C, filtrando por tipo de compra mais frequente.

	Empresa B	**Empresa C**
0 - Não Promotor	125 (63.5%)	64 (55.7%)
1 - Promotor	72 (36.5%)	51 (44.3%)
Total	197	115

Os modelos obtidos foram então os seguintes:

Tabela .2: Sumário do modelo de regressão logística para o tipo de compra mecânica da Empresa B juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−1.717	0.325	-5.283	**< 0.001**	-
Logística: Rapidez entrega1	1.332	0.551	2.416	**0.016**	3.789 (1.286-11.164)
Satisfação (Empatia)1	1.367	0.541	2.526	**0.012**	3.922 (1.359-11.323)
Satisfação Serviço: Preços1	1.749	0.741	2.360	**0.018**	5.751 (1.345-24.597)
Satisfação (Tangíveis)1	1.280	0.642	1.996	**0.046**	3.598 (1.023-12.652)
Número de observações usadas: 117 (80 excluídas devido a *missing values*)					AIC: 120.96
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.964
Desempenho Preditivo: AUC (IC 95%)					0.811 (0.728-0.891)
Desempenho Preditivo: ACC					0.786

Tabela .3: Sumário do modelo de regressão logística para para o tipo de compra mecânica da Empresa C juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−2.061	0.430	-4.793	**< 0.001**	-
Satisfação (Fiabilidade)1	1.796	0.587	3.061	**0.002**	6.025 (1.908-19.025)
Sat. Prod.: Diversidade da Oferta1	2.252	0.788	2.856	**0.004**	9.506 (2.027-44.569)
Satisfação (Tangíveis)1	1.992	0.599	3.323	**0.001**	7.327 (2.264-23.719)
Número de observações usadas: 99 (16 excluídas devido a *missing values*)					AIC: 88.759
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.999
Desempenho Preditivo: AUC (IC 95%)					0.885 (0.814-0.948)
Desempenho Preditivo: ACC					0.808

Empresa B: Pela Tabela .2 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (78.6%) e tem de se classificar o seu poder discriminativo como bom (aceitável a bom julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos os factores, são fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);

- O *odds* para o sucesso nos clientes promotores na satisfação com a rapidez de entrega é 3.789 o *odds* para o sucesso nos clientes não promotores na satisfação com a rapidez da entrega;
- O *odds* para o sucesso nos clientes promotores na satisfação (empatia) é 3.922 o *odds* para o sucesso nos clientes não promotores na satisfação (empatia);
- O *odds* para o sucesso nos clientes promotores na satisfação com os preços é 5.751 o *odds* para o sucesso nos clientes não promotores na satisfação com os preços;
- O *odds* para o sucesso nos clientes promotores na satisfação (tangíveis) prazos é 3.598 o *odds* para o sucesso nos clientes não promotores na satisfação (tangíveis).

Empresa C: Pela Tabela .3 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (80.8%) e tem de se classificar o seu poder discriminativo como bom (bom a excecional julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Excluindo o *Intercept*, todas variáveis são fatores de proteção;
- O *odds* para o sucesso nos clientes promotores na satisfação (fiabiliade) é 6.025 o *odds* para o sucesso nos clientes não promotores na satisfação (fiabilidade);
- O *odds* para o sucesso nos clientes promotores na satisfação com a diversidade da oferta é 9.506 o *odds* para o sucesso nos clientes não promotores na satisfação com a diversidade da ofertas;
- O *odds* para o sucesso nos clientes promotores na satisfação (tangíveis) é 7.327 o *odds* para o sucesso nos clientes não promotores na satisfação (tangíveis).

Resultados por canal de compra mais frequente

Veja-se que resultados se obtiveram filtrando as bases de dados para o canal de compra mais frequente. Relembra-se que, para a Empresa B, o canal de compra mais usado é o portal online e, para a Empresa C, o canal de compra mais usado é o serviço prestado em loja.

Após se aplicar a metodologia referida em 3.1.9.4 excluíram-se as seguintes variáveis das bases de dados:

Empresa B

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado
- Tipo de compra

Empresa C

- Não comprar à Marca: Comparar prazo de entrega
- Não comprar à Marca: Comparar valor do orçamento
- Não comprar à Marca: Não existem Campanhas
- Não comprar à Marca: Tempo para procurar mercado

Ficou-se, assim, com 22 variáveis na Empresa B (retiraram-se 5) e 23 na Empresa C (retiraram-se 4) que foram consideradas na modelação.

Para se ter uma ideia da distribuição da variável resposta (definida em 4.2.1), pelas 2 empresas, segue a seguinte tabela com a frequência absoluta e relativa.

Tabela .4: Distribuição da variável resposta nas Empresa B e C, filtrando por canal de compra mais frequente.

	Empresa B	**Empresa C**
0 - Não Promotor	112 (64.7%)	63 (48.8%)
1 - Promotor	61 (35.3%)	66 (51.2%)
Total	173	129

Pela tabela anterior, verifica-se um distribuição praticamente equilibrada da Empresa C e um desfasamento na Empresa B.

Os modelos obtidos foram então os seguintes:

Tabela .5: Sumário do modelo de regressão logística para o canal de compra Portal Online da Empresa B juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−2.097	0.331	-6.333	**< 0.001**	-
Logística: Cumprimento prazos1	1.515	0.514	2.946	**0.003**	4.551 (1.661-12.470)
Sat. Serv.: Cond. Pagamento1	1.335	0.480	2.781	**0.005**	3.799 (1.483-9.734)
Satisfação (Fiabilidade)1	1.464	0.641	2.285	**0.022**	4.323 (1.232-15.171)
Sat. Prod.: Diversidade da Oferta1	1.309	0.504	2.596	**0.009**	3.701 (1.378-9.940)
Número de observações usadas: 149 (24 excluídas devido a *missing values*)					AIC: 138.88
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.782
Desempenho Preditivo: AUC (IC 95%)					0.851 (0.784-0.909)
Desempenho Preditivo: ACC					0.779

Tabela .6: Sumário do modelo de regressão logística para para o canal de compra Loja da Empresa C juntamente com OR e respetivo IC a 95% de confiança.

Variável	Est	EP	z	valor-p	OR (IC 95%)
(*Intercept*)	−1.871	0.429	-4.365	**< 0.001**	-
Satisfação (Capacidade resposta)1	1.630	0.624	2.611	**0.009**	5.103 (1.502-17.343)
Satisfação (Fiabilidade)1	1.826	0.778	2.346	**0.019**	6.210 (1.351-28.551)
Sat. Prod.: Diversidade da Oferta1	2.282	0.799	2.856	**0.004**	9.795 (2.046-46.890)
Sat. Serv.: Descontos Praticados1	1.855	0.870	2.132	**0.033**	6.393 (1.161-35.198)
Número de observações usadas: 102 (27 excluídas devido a *missing values*)					AIC: 88.412
Avaliação do modelo					
Qualidade do ajuste: Teste χ^2 de Hosmer e Lemeshow:					valor-p=0.995
Desempenho Preditivo: AUC (IC 95%)					0.895 (0.830-0.951)
Desempenho Preditivo: ACC					0.814

Empresa B: Pela Tabela .5 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (77.9%) e tem de se classificar o seu poder discriminativo como bom (aceitável a excecional julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Todos fatores de proteção (excluindo o *Intercept* todas as outras estimativas são positivas);
- O *odds* para o sucesso nos clientes promotores na satisfação com o cumprimento de prazos

é 4.551 o *odds* para o sucesso nos clientes não promotores na satisfação com o cumprimento de prazos;

- O *odds* para o sucesso nos clientes promotores na satisfação com as condições de pagamento é 3.799 o *odds* para o sucesso nos clientes não promotores na satisfação com as condições de pagamento;
- O *odds* para o sucesso nos clientes promotores na satisfação (fiabilidade) é 4.323 o *odds* para o sucesso nos clientes não promotores na satisfação (fiabilidade);
- O *odds* para o sucesso nos clientes promotores na satisfação com a diversidade da oferta é 3.701 o *odds* para o sucesso nos clientes não promotores na satisfação com a diversidade da oferta.

Empresa C: Pela Tabela .6 é possível averiguar que não se rejeita a hipótese de se ter um bom ajustamento do modelo aos dados. Para além disso, analisando o desempenho preditivo, verifica-se que a exatidão do modelo é boa (81.4%) e tem de se classificar o seu poder discriminativo como bom (bom a excecional julgando pelo IC).

Os efeitos estatisticamente significativos estimados pelo modelo foram:

- Excluindo o *Intercept*, todas variáveis são fatores de proteção;
- O *odds* para o sucesso nos clientes promotores na satisfação (capacidade de resposta) é 5.103 o *odds* para o sucesso nos clientes não promotores na satisfação (capacidade de resposta);
- O *odds* para o sucesso nos clientes promotores na satisfação (fiabilidade) é 6.210 o *odds* para o sucesso nos clientes não promotores na satisfação (fiabilidade);
- O *odds* para o sucesso nos clientes promotores na satisfação com a diversidade da oferta é 9.795 o *odds* para o sucesso nos clientes não promotores na satisfação com a diversidade da ofertas;
- O *odds* para o sucesso nos clientes promotores na satisfação com os descontos praticados é 6.393 o *odds* para o sucesso nos clientes não promotores na satisfação com os descontos praticados.