UNIVERSIDADE ABERTA

Metodologias de *pricing* aplicadas ao ramo dos seguros não vida – uma análise baseada na estatística experimental, análise de risco e *Value Based Pricing*

Filipe Charters de Azevedo

Mestrado em Estatística, Matemática e Computação

Dissertação orientada Profª Doutora Teresa Paula Costa Azinheira Oliveira

Co-orientador: Prof. Doutor Amílcar Manuel do Rosário Oliveira

2015

Resumo

Uma companhia de seguros baseia o seu modelo de produção no valor de uma matéria-prima cujo custo é desconhecido no momento de produção. De uma forma mais directa: uma companhia "compra" sinistros e "vende" segurança – se uma companhia compra os sinistros a um baixo preço então ganha dinheiro; se compra "caro" então perde dinheiro. Na cadeia de valor, uma companhia pode contar com a lei dos grandes números que mitiga a volatilidade e a incerteza de mercado – confere segurança na média.

Pretende-se, com este projecto, fornecer uma base de trabalho geral para desenhar os preços de seguros não-vida que leve em linha de conta: o risco; o comportamento da concorrência; e a elasticidade do clientes.

Os objectivos do projecto foram conseguidos. Assim, é apresentada uma abordagem de análise de risco com Modelos Lineares Generalizados – tendo sido dado especial destaque às regressões Poisson, gama e Tweedie, sendo nesta última estudada a definição do parâmetro p através da maximização da evolução da função verosimilhança. Seguidamente, estuda-se o apuramento das tarifas da concorrência com recurso a desenho ortogonal, Regressões Aparentemente Não Relacionadas com correcção de ponto de massa. A avaliação da elasticidade dos clientes através de Análise Conjunta é estudada em seguida tendo em conta a oferta da concorrência. Por fim, e através de uma abordagem de optimização tendo em conta os princípios de Stackelberg é efectuada usando todos os resultados anteriores. No final, para cada micro mercado (célula tarifária) é apresentado o preço óptimo (*i.e.* maximizador de receita ou de lucro) a definir por uma seguradora, tendo em conta os custos, a oferta das congéneres e a elasticidade dos consumidores.

São apresentados casos práticos reais com bases de contexto segurador.

Palavras-chave: Determinação de preços (seguros não vida), Modelos lineares generalizados; Regressão Box-Cox; Regressões aparentemente não relacionadas; Análise conjunta

Abstract

An insurance company bases its production model in a raw material whose cost is unknown at the time of production. In a more direct way: a company "buys" claims and "sell" security - if a company buys claims at low price then the insurer makes money; buying "expensive" then loses money. In the value chain, a company can rely on the law of large numbers that mitigates volatility and market uncertainty - provides security on average.

It is intended with this project to provide a general approach to draw the prices of non-life insurance that takes into account: risk/cost; competitive performance; and the elasticity of the customers.

The project objectives were achieved. So, a risk analysis approach with Generalized Linear Models is presented – we give special emphasis to Poisson, gamma and Tweedie regressions, where in this last the definition of the parameter p was study and a solution is given by maximizing the evolution of the likelihood function. Next, we study the competition clearance problem using the orthogonal design, Seemingly Unrelated Regressions with mass point correction. An evaluation approach to customer elasticity subject to market environment through Conjoint Analysis is also presented. Finally, and through Stackelberg principals, an optimizing approach was conducted using all the outputs of the previous analysis. At the end, for each micro market (tariff cell) an optimal price (*i.e.* revenue and profit maximizing prices) is presented subjected to cost, competitive set and client elasticity behaviour.

There are actual case studies with insurance context data bases.

Key Words: Pricing (insurance non-life), Generalized linear models; Box-Cox regression; Seemingly unrelated regression, Conjoint analysis

Dedicatória

Para a Carla, por tudo.

Ao David e Luísa, pelo amor e alegria incondicionais.

Agradecimentos

Em primeiro lugar uma palavra à minha mulher e aos meus filhos, pela compreensão, apoio, amizade e alegria. Sem o seu apoio constante nunca teria terminado este projecto. Obrigado por terem compreendido as minhas ausências.

Aos meus pais e irmãos pelas discussões animadas e por permitirem um refúgio nos momentos mais difíceis.

Aos meus sogros um especial obrigado pelo apoio e pela amizade.

Entre aqueles que tornaram possível este trabalho, gostaria de agradecer, muito em especial, à Doutora Carla Castro pela sua paciência, com quem tive oportunidade de discutir as mais diferentes questões, num dialogo sempre estimulante e enriquecedor e para o qual nunca faltaram palavras de apoio e incentivo.

Gostaria também de manifestar o reconhecimento sincero e amigo ao Doutor Nuno Pena por me ter desafiado para a Universidade Aberta e ter aberto a porta à Associação Portuguesa de Seguradores, local onde me permitiu ter uma reflexão e ensaio sobre muitos destes temas.

Um agradecimento é também devido à Associação Portuguesa de Seguradores, na pessoa do doutor Miguel Guimarães e doutora Isabel Salomão e aos meus formandos na Associação Portuguesa de Seguradores e no IPAM que sofreram os meus devaneios e permitiram melhorar (e em muito) os conceitos aqui apresentados.

Devo ainda salientar os contributos amigos e reflexão estratégica do doutor Rui Gonçalves e do doutor Jorge Soeiro Marques.

Queria também recordar as palavras de ânimo da doutora Maria do Carmo Bandeira, Eng. Manuel Leiria, doutor Rogério Dias e do doutor Marcus Martins.

À Professora doutora Conceição Pequito uma palavra de amizade por me ter iniciado nestas andanças académicas.

Aos orientadores, Professora Doutora Teresa Oliveira e Professor Doutor Amílcar Oliveira, uma mensagem de apreço pela sua paciência e disponibilidade para aceitar as minhas liberdades criativas. Importa lembrar que este trabalho deve muito à sua compreensão e diversas formas de apoio.

Por fim, um orbigado a toda a comunidade da Universidade Aberta por terem feito este mestrado possível.

Índice

Índice de tabelas

Índice de figuras

Índice de resultados

Lista de abreviaturas, siglas e acrónimos

ASF – Autoridade de Supervisão de Seguros e Fundos de Pensões

CAPI – *Computer Assisted Personal Interview*

CAWI – *Computer Assisted Web Interview*

ED – *Exponential Dispersion*

EDM – *Exponential Dispersion Model*

G. Controlo – Grupo de Controlo

G. Experimental – Grupo Experimental

GLM – *Generalized Linear Model*

IIA – *Independence from Irrelevant Alternatives*

ISP – Instituto de Seguros de Portugal

MV – Máxima Verosimilhança

OLS – *Ordinary Least Squares*

PAPI – *Paper And Pencil Interview*

SUR – *Seemingly Unrelated Regression*

u.m. – Unidade Monetária

WLS – *Weight Least Squares*

1. Introdução e objectivos

1.1. Descrição da problemática

"Cecil Graham: What is a cynic?
Lord Darlington: A man who knows the price of everything,
and the value of nothing.
Cecil Graham: And a sentimentalist, my dear Darlington,
is a man who sees an absurd value in everything and
doesn't know the market price of any single thing."

— Oscar Wilde, Lady Windermere's Fan

Uma companhia de seguros baseia o seu modelo de produção no valor de uma matéria-prima cujo custo é desconhecido no momento de produção. De uma forma mais directa: uma companhia "compra" sinistros e "vende" segurança – se uma companhia compra os sinistros a um baixo preço então ganha dinheiro; se compra "caro" então perde dinheiro. Na cadeia de valor, uma companhia pode contar com a lei dos grandes números que mitiga a volatilidade e a incerteza de mercado – confere segurança na média.

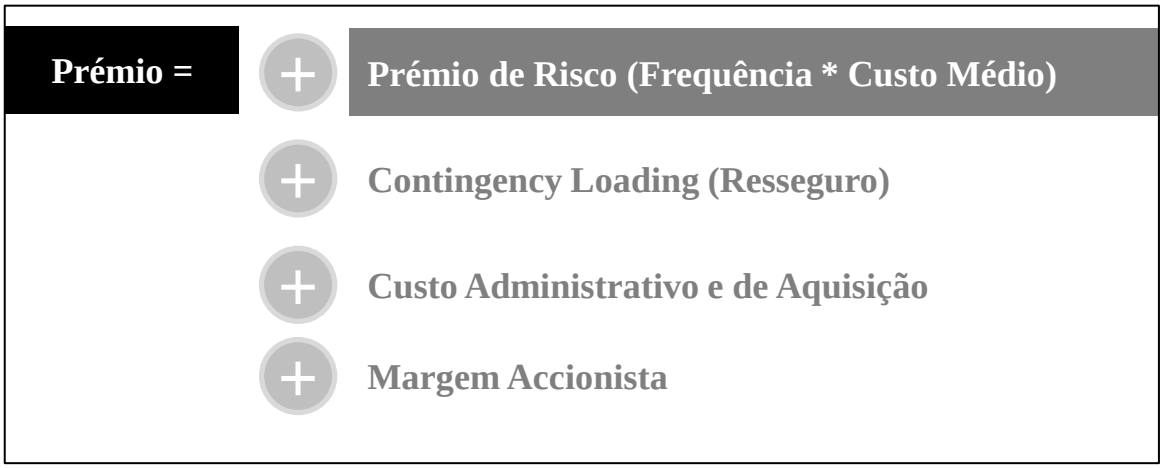
Assim, a principal questão de uma companhia é como avaliar o preço da compra de sinistro[1]: Qual o custo de um risco? Habitualmente uma seguradora conta com seu histórico para estimar esse valor: analisa para o comportamento dos seus clientes e propõe um prémio que é idêntico às responsabilidades assumidas ao qual acrescenta uma parcela para pagar os custos administrativos, de aquisição e de remuneração accionista. Deverá ainda afectar uma parcela de contingência, já que a média garante estabilidade – mas há surpresas.

É necessário determinar o custo desses sinistros sendo que as companhias habitualmente o fazem com modelos GLM – *Generalized Linear Model*. A margem

[1] O problema pode ser visto *ex-post*: depois de ter subscrito um risco, qual o valor que uma seguradora deve provisionar para fazer face às responsabilidades assumidas. De uma maneira geral o problema *ex-post* pertence à categoria de *reserving actuary* e também um desafio contabilístico e de transparência de mercado; o problema exposto, *ex-ant*, pertence à categoria de *pricing actuary* e é de natureza operacional.

é definida pela concorrência do mercado (tendencialmente) e pelo apetite pelo risco dos accionistas; os custos administrativos e de distribuição são dados do problema afectos por economias de escala – e não um desafio na determinação do prémio.

Figura 1.1 - Modelo de determinação do preço normalmente aplicado pelas seguradoras



Fonte: autor

Mas será essa a melhor forma de definir um preço de um produto? De outra forma, o cálculo actuarial[2] apenas permite determinar os custos de produção, mas não o preço. Um qualquer manual de marketing indica que a definição pelos custos não permite saber as tendências futuras e obrigam as organizações a abdicar, a prazo, do seu factor de receita. De facto Mercator, já na ed 1992, afirmava que: "*O preço é uma variável estratégica no marketing-mix, mas é pouco aproveitada*."

1.2. Objectivo

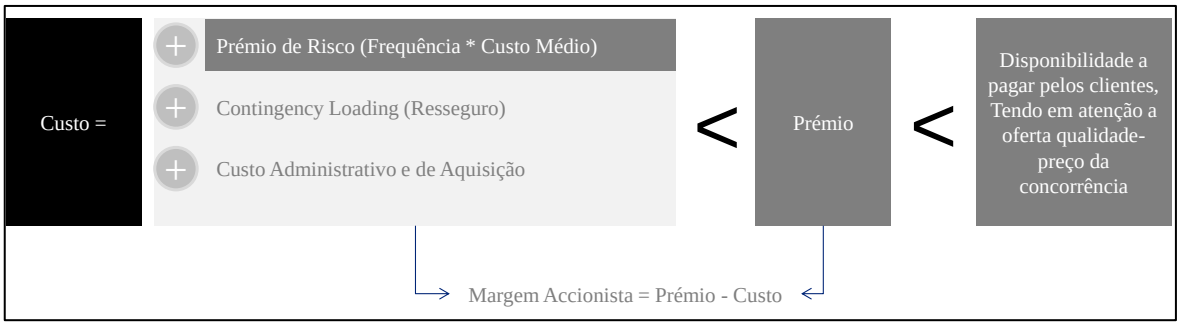
Pretende-se, com este projecto, fornecer uma base de trabalho geral para desenhar os preços de seguros não-vida que leve em linha de conta:

- o risco;

[2] O cálculo actuarial é a disciplina que se aplica métodos de cálculo financeiro e estatística de forma a determinar o risco e retorno em seguros.

- o comportamento da concorrência;
- a elasticidade dos clientes.

Figura 1.2 - Modelo de determinação do preço eficiente



Fonte: autor

Habitualmente, na literatura os problemas são estudados autonomamente (sendo que o desafio de risco é o mais estudado). No mundo real dos seguros, e de acordo com a experiência detida pelo autor, a avaliação do risco pode ser mais ou menos profunda, e o *pricing* final (*i.e.* a tarifa) apenas incorpora a concorrência e a elasticidade dos clientes de forma empírica. Sendo ainda mais directo, no trabalho de divulgação sobre *Value Based Pricing*, Mohamed (2010), afirma *"[c]ritical pricing decisions are often made using arbitrary 'this is the way we've always done it' methods. Companies are shortchanging themselves every day*".
Assim, e de forma sintética, o macro objectivo deste projecto é o de fornecer uma metodologia de análise que permita às companhas de seguro dos ramos reais (seguros não vida) incorporar a elasticidade preço dos seus clientes bem como das estratégias de *pricing* existentes no mercado, afastando-se assim de uma formulação arcaica de definição de preços: custo mais margem.

Os objectivos específicos deste projecto são, para um ramo não vida:

a) Compreender e explicitar, de forma breve, os principais mecanismos de formulação dos custos de seguro (com especial destaque para os modelos GLM).

b) Observar e encontrar formas incluir os preços das companhias congéneres no modelo tarifário de uma companhia.
c) Observar e encontrar formas incluir a reacção dos clientes (elasticidade e função reacção/optimização de resultados) face às melhores ofertas.
d) Encontrar formas possíveis de optimizar (de forma muito linear/simples) os resultados técnicos da companhia no ramo em estudo.

Para concretizar a análise o projecto será baseado no ramo automóvel.

1.3. Apresentação das principais ferramentas estatísticas e abordagem metodológica

Esta dissertação, como se detalhará, baseia-se em quatro grandes ferramentas de estatística que cumpre destacar ainda em contexto de introdução. Assim tem-se:

- Modelos Lineares Generalizados (*GLM – Generalized Linear Models*);
- Desenho experimental e planos ortogonais;
    - *Conjoint analysis;*
- Regressão de Box Cox;
- SUR – *Seemingly Unrelated Regression*.

Nas secções seguintes será feito um breve resumo destas ferramentas, os desafios metedológicos, sendo igualmente indicado onde nesta dissertação pode ser encontrado o detalhe e a respectiva fundamentação.

1.3.1. Modelos Lineares Generalizados

De forma a compreender/caracterizar uma variável escalar é possivel utilizar uma média. De forma explícita ou implícita, transforma-se uma variável numa constante – e esta é pela sua natureza muito mais fácil de compreender: é apenas um número. Esta transformação tem um custo: a perda de variabilidade é

por vezes excessiva. Assim, além da média, um modesto analista apresenta igualmente uma medida de dispersão, como a variância ou o desvio-padrão, que traduz o quanto se perdeu no processo de transformar uma variável numa constante.

No entanto, a quantificação do quanto se perdeu não resolve o problema: estes dois conceitos fazem terraplanagem da informação, limitando por vezes a compreensão dos fenómenos que se pretende estudar. Há por vezes uma heterogeneidade nos dados que tem de ser explicitada se se quiser compreender com rigor um determinado fenómeno.

O conceito de regressão resolve este problema. O modelo de regressão o que faz é avaliar a média em cada situação, apresentando igualmente uma medida de dispersão como o custo dessa regressão. Repare-se, num modelo de regressão há duas componentes:

i. Um vector aleatório $Y = (Y_1, \dots, Y_n)'$ com uma distribuição que envolve um vector de parâmetros desconhecidos $\mu = (\mu_1, \dots, \mu_n)'$.
ii. Uma relação entre μ e um vector de parâmetros $\beta = (\beta_1, \dots, \beta_k)'$, da forma $\mu = f(\beta)$, onde $f(.)$ é função uma contínua e unívoca.

Seguindo a terminologia de Jørgensen (1989), estas duas componentes são designadas, respectivamente, como a componente aleatória e a componente sistemática do modelo. Enquanto a componente aleatória do modelo é compreendida num dado contexto, o ingrediente (ii) é designado como modelo. O vector aleatório Y é designado como a resposta e a distribuição de Y ou de Y_i é designada a distribuição do erro. A componente aleatória pode traduzir qualquer processo estocástico, incluindo os erros de medida. A função $f(.)$ é designada como a função de regressão e os parâmetros β são designados como os parâmetros de regressão. Todo este sistema de vectores e distribuições permite definir a média para cada Y_i sobre as condições de μ; isto é: $E(Y_i|\mu)$. A dispersão associada a $E(Y_i|\mu)$ indicará uma medida de qualidade de ajustamento.
Uma classe importante dos modelos de regressão dá-se quando:

$$g(\mu_i) = \eta_i, \text{ com } i = 1, \dots, n$$
$$\eta_i = \sum_{j=1}^{k} x_{ij}\ \beta_j, \text{ com } i = 1, \dots, n.$$

A função $g(.)$ contínua e unívoca é chamada de função ligação (ou *link function*). Os x_{ij} são constantes designadas por covariáveis ou variáveis explicativas. A matriz $X = \{x_{ij}\}$ é designada como matriz desenho (*design matrix*) do modelo. Um modelo desta forma diz-se linear. Quando $g(.)$ é uma função identidade e a distribuição de Y é homocedástica ou mesmo normal está-se perante o modelo de regressão linear (simples). Habitualmente este caso simples é estimado pela minimização dos mínimos quadrados dos resíduos ou por máxima verosimilhança.

Porém, como será visto no capítulo 3 é possível questionar a tipologia da função $g(.)$ e a distribuição associada a Y. Em relação a este último, uma forma muito conveniente de apurar a heterogeneidade dos dados é assumir que a distribuição Y é definida por uma função de probabilidade da seguinte fórmula:

$$p(y, \theta, \lambda) = \alpha(\lambda, y) \exp[\lambda\{y\theta - k(\theta)\}], \text{ com } y \in \mathbb{R}$$

Sendo que $\alpha(.)$ e $k(.)$ designam funções, $\lambda > 0$ e θ variam num determinado domínio real, a definir. Assim, diz-se que $Y \sim ED(\mu, \sigma^2)$, onde ED designa o modelo exponencial (*exponential distribution*, na designação em inglês), $\mu = k'(\theta)$ é o valor esperado de Y e $\sigma^2 = 1/\lambda$ é o parâmetro de dispersão. Note-se que este modelo ED retrata o modelo de famílias de distribuição exponenciais, onde a distribuição normal, Poisson, gama e outras distribuições comuns fazem parte.

Ao longo do capítulo 3 alguns casos especiais destes modelos serão estudados com detalhe de forma a explicar os custos da actividade seguradora (o contexto de aplicação destes conceitos na tese). Como afirma Jørgensen (1989), o objectivo de uma análise estatística é encontrar o modelo estatístico adequado, avaliar a qualidade do ajustamento, estimar os parâmetros desconhecidos do modelo e testar a hipótese nula associada à relevância dos parâmetros. Neste projecto, o processo de estimação dos parâmetros para cada um dos casos estudados (Poisson, gama e Tweedie) será distinto de forma a trabalhar a

interpretação, fazer pequenos ajustes ao processo de estimação de forma garantir uma melhor capacidade de previsão e de compreensão dos fenómenos e de ajustá-lo à prática seguradora.

O caso geral não deixará de ser estudado (ainda que de forma mais breve) já que a regressão Tweedie é um processo gerador de dados que engloba a distribuição de Poisson e gama. No entanto, é necessário apurar o parâmetro p caracterizador desta distribuição (parâmetro que está ligado à variância como se verá adiante). De facto tem-se que um ED pode (Jørgensen (1989)) ser caracterizado pela sua função de variância:

$$V\ (Y) = \sigma^2 = \phi V(\mu).$$

Sendo que os casos especiais de

$$V\ (Y) = \sigma^2 = \phi\mu^p, \text{ para } p \text{ distintos}$$

assumem uma classe muito importante. Esta classe de modelos de dispersão assume formas muito conhecidas:

- Com $p\ =\ 0$, tem-se um processo gerador de dados normal;
- Com $p\ =\ 1$, tem-se um processo gerador de dados de Poisson;
- Com $p\ =\ 2$, tem-se um processo gerador de dados de gama; e
- Com $p\ =\ 3$, tem-se um processo gerador de dados de Gaussiano inverso.

Para outros valores de p é possível ter outras distribuições:

- Com $p > 2$, tem-se um processo gerador de dados de positivo e estável; e
- Com $p\ = \infty$, tem-se um processo gerador de dados extremamente estável.

Para $0 < p < 1$ não há nenhum modelo Tweedie.

Há, no entanto, uma forte desvantagem num modelo de dispersão exponencial: aparte dos modelos mais conhecidos, com $p = 1, 2\ e\ 3$, a distribuição Tweedie não tem uma função densidade que possa ser escrita de uma forma fechada. Este parâmetro p é assim definido exogenamente antes do processo de estimação. Nesta dissertação será dada uma indicação computacional de como se pode obter o parâmetro p através da maximização da evolução da função verosimilhança.

Os modelos lineares generalizados são analisados com cuidado ao longo do capítulo 3, e em parte ao longo do capítulo 4, no que diz respeito à sua aplicação ao desenho experimental das companhias congéneres.
Fica por estudar o que acontece se aplicarmos um modelo Tweedie num contexto de suposto de processos geradores de Poisson e Gama.

1.3.2. Desenho experimental e planos ortogonais

Experiências são estudos em condições controladas em que uma ou mais variáveis podem ser manipuladas para testar uma hipótese associada a uma variável dependente. As experiências são assim investigações no qual é exigido o envolvimento activo do analista.
O princípio chave do desenho experimental é a manipulação de uma variável de tratamento (habitualmente designada por x), seguida da variável de resposta (habitualmente designada por y). Se uma alteração de x é acompanhada por uma alteração de y; é tentador afirmar que x provoca y. Porém, esta inferência de causalidade é infirmada. A estatística pode afirmar que duas variáveis estão associadas (medidas de correlação, medida de associação, qui-quadrado, etc.), mas não consegue afirmar se duas variáveis estão ligadas com uma relação causal. É uma condição prévia (isto é necessária), mas não suficiente para garantir a causalidade. Além da correlação ou associação, há ainda que considerar a sequência temporal (a causa precede o efeito), a teoria e, sobretudo, não haver outra explicação plausível.

Evidentemente que o que se pretende no desenho experimental é concluir, em ambiente controlado e com o mínimo esforço possível, as relações entre as variáveis de forma a verificar as condições indicadas acima: associação, causalidade temporal, hipóteses que sustentam a teoria. Igualmente tem-se que a experiência maior será a realidade, mas esta poderá não dar todas as relações possíveis e por vezes contem demasiado ruído para uma conclusão fundamentada. Em qualquer caso, quando os dados parecem mostrar que há uma relação linear é possível tentar obter um modelo de regressão que se aproxime da relação apurada. Se ao analisar os dados for evidente que há uma relação 'semelhante' à apresentada abaixo:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \mathbf{erro}.$$

o estimador dos mínimos quadrados ($\hat{\beta}$) é habitualmente obtido através:

$$\underset{\widehat{\boldsymbol{\beta}}}{\text{Min}}\ \mathbf{u}'\mathbf{u} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

onde a notação é a habitual: u os resíduos, y a variável objectivo e $\hat{\mathrm{y}}$ o modelo estimado, X as variáveis explicativas e $\hat{\beta}$ o vector dos coeficientes a estimar. Para resolver este problema é preciso assumir as condições 1 a 4:

- Condição 1: Linearidade nos parâmetros: O modelo tem de ser linear nos parâmetros.
- Condição 2: Valores de X são fixados numa amostragem repetida. (Mais tecnicamente, X é assumido como sendo não estocástico)
- Condição 3: Erro não está correlacionado com as variáveis explicativas: $\mathrm{E}(\mu_i|\mathbf{X}) = 0$. A esta hipótese também se designa por exogenidade condicionada.
- Condição 4: Ausência de perfeita multicolinearidade, isto é que a matriz $\mathrm{X}'\mathrm{X}$ é invertivel.

Assim, tem-se:

$$\underset{\widehat{\boldsymbol{\beta}}'}{\text{Min}}\ \mathbf{u}'\mathbf{u} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\mathbf{y}' - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}.$$

Nestas condições, tem-se as habituais condições de primeira ordem:

$$\frac{\partial .}{\partial \widehat{\boldsymbol{\beta}}'} = \mathbf{0} - \mathbf{2X}'\mathbf{y} + \mathbf{2X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$$

que permitem derivar o estimador dos mínimos quadrados:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Este resultado não obriga a nenhuma condição sobre os erros ou resíduos. A única condição sobre o comportamento dos resíduos (exogenidade $\mathrm{E}(\mathrm{u}|\mathrm{X}) = 0$) é apenas necessária para garantir a centragem. Repare-se que:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + \mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}\end{aligned}$$

Aplicando o operador valor esperado, tem-se:

$$\begin{aligned}\mathbf{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{E}(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{E}(\mathbf{u}|\mathbf{X}).\end{aligned}$$

Sob a hipótese clássica de exogenidade condicionada, tem-se que $\mathrm{E}(\mathrm{u}|\mathrm{X}) = 0$:

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}.$$

Em desenho experimental é possível levar ao extremo a condição de exogeneidade condicionada. Tal como já indicado, num desenho experimental pretende-se maximizar a potência dos testes amostrais de forma a concluir com o

máximo de segurança possível e com o menor esforço possível - leia-se a menor amostra. Para tal é necessário garantir que o processo é eficiente, isto é que o desvio padrão associado a cada uma das estimativas dos betas é mínimo:

$$Min_{\mathrm{X}} \operatorname{var}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\sigma}(\mathbf{X'X})^{-1}$$

Ou seja, se se quiser ter uma estimativa eficiente é possível apostar em ter o número de casos suficiente para estimar a regressão e que cada um dos elementos da diagonal da matriz $(\mathrm{X'X})^{-1}$ seja mínimo (assumimos que não há qualquer efeito de interacção).

A melhor forma de o fazer é garantir que a matriz $(\mathrm{X'X})^{-1}$ seja apenas preenchida na diagonal – que não haja qualquer correlação entre os diferentes elementos de X. Nesse caso, diz-se que a matriz $(\mathrm{X'X})^{-1}$ é ortogonal. Ou seja, ao invés de se apostar em ter um estimador:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

com base em X aleatórios, é possível ter:

$$\widehat{\boldsymbol{\beta}} = \mathbf{I}\,\mathbf{X'y}$$

onde X foi desenhado com cuidado de forma a ter $(\mathrm{X'X})^{-1} = (\mathrm{X'X}) = \mathrm{I}$. Desta forma, os valores de $\hat{\beta}$ são muito fáceis de estimar e com a variância mínima.

Nesta tese o desenho ortogonal é aplicado em diferentes contextos:

- Pretende-se recolher o processo gerador de tarifas de várias seguradoras. O número de combinações é demasiado extenso e há demasiadas idiossincrasias para deixar que a compreensão deste processo seja com base a recolhas de preços aleatórios.
- O mesmo princípio de desenho é aplicado para compreender e prever as preferências dos consumidores numa aplicação prática do teorema da preferência revelada desenvolvido por Samuelson em 1938 e que no

marketing research pode ser quantificado por *conjoint analysis* desenvolvido por Green & Wind (1975).

1.3.2.1 *Conjoint analysis*

Por forma a descrever as escolhas dos consumidores quando estes se deparam com diferentes possibilidades de consumo, os economistas desenvolveram o conceito de utilidade. A utilidade significa o proveito/prazer que uma pessoa obtém do consumo de um bem/serviço. Como tal, a utilidade não se trata de um conceito preciso, observável e mensurável, mas sim de uma abstracção teórica que permite explicar a forma como os consumidores racionais tomam as suas decisões. Os economistas e *marketeers* assumem ainda que cada produto ou bem pode ser ainda decomposto em dimensões, cada uma com o seu impacto na utilidade, dependendo do consumidor. Assim sendo, e assumindo que os indivíduos são racionais, estes deverão efectuar as suas escolhas de modo a obter a maior utilidade possível.

A abordagem metodológica de *conjoint analysis* é, na prática, idêntica à abordagem de desenho experimental vista na secção anterior e materializa o conceito de utilidade na escolha dos consumidores: Os inquiridos serão confrontados com um leque de possíveis conceitos. Cada um destes será suficientemente semelhante para ser considerado como um substituto, mas suficientemente distinto, para que os inquiridos possam claramente revelar uma preferência. Este leque de possíveis conceitos é definido com base em desenhos ortogonais.

As respostas dos indivíduos (y) irão permitir através de uma regressão entender a importância de cada dimensão em análise e desenhar a solução (x) que melhor satisfaz as suas preferências.

Para efeitos de modelação, considera-se que a utilidade (dada por y) associada a cada produto segue uma distribuição logística (semelhante à normal) e que a percentagem normalizada desse valor indica a quota de preferência do mesmo produto. No final é possível determinar uma curva de procura, tendo em conta a oferta da concorrência e a elasticidade dos clientes aos diferentes atributos que compõem o produto.

O estudo desta ferramenta é efectuado com detalhe no capítulo 5.

1.3.3. Regressão de Box Cox

A regressão Box Cox ocupa um especial destaque nesta dissertação. De forma a apurar o processo gerador de tarifas das companhias não basta ter um desenho experimental eficiente. A compreensão da forma funcional é fundamental para o sucesso deste desígnio.

Como será estudado no capítulo 3, o processo gerador de tarifas de uma seguradora combina a distribuição de Poisson e gama, e que estes dois 'efeitos' se traduzem numa distribuição Tweedie. E esta distribuição depende de um parâmetro caracterizador p que é difícil de apurar. Esta dificuldade advém do facto de que a distribuição Tweedie não tem uma solução fechada para p. Estando num contexto de desenho experimental, é difícil sustentar que a evolução da verosimilhança ou da soma dos quadrados dos resíduos, possa dar uma indicação deste parâmetro.

Neste projecto optou-se assim para a captura do processo gerador de tarifas das companhias seguradoras congéneres pela regressão de Box-Cox. Este expediente estatístico baseia-se na ideia de que a variável dependente pode ser transformada, de y para $y^{(\lambda)}$:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \text{para } \lambda \neq 0 \\ \log(y) \text{ para } \lambda = 0 \end{cases}$$

de forma a assumir diferentes formulações e nesse sentido acompanhar distribuições menos ortodoxas da variável dependente. Box & Cox (1964) referem que por vezes é preferível ter simplicidade na compreensão de alguns fenómenos à conta de algum possível enviesamento.

Ao longo do capítulo 4 esta solução pragmática é estudada sendo indicado como se pode identificar qual a transformação mais adequada dentro da família Box--Cox.

Uma melhoria a esta transformação é igualmente estudada (Wooldridge (2003)) já que com esta transformação o ponto de massa da distribuição é alterado, podendo causar pequenos enviesamentos na estimativa.
Por fim, é ainda de salientar que um caso prático, com uma base de dados real é apresentada no final do capítulo 4.

1.3.4. SUR – *Seemingly Unrelated Regression*

É razoável assumir que cada companhia de Seguros apresenta o seu modelo gerador de tarifas autónomo – cada uma terá o seu modelo. Mas será que essa autonomia significa de facto independência? Estando a aplicar os seus modelos a uma realidade semelhante é natural que as companhias cheguem a modelos próximos, mas não necessariamente iguais. Da mesma forma, os erros das diferentes companhias deverão estar correlacionados.
O modelo SUR – *Seemingly Unrelated Regression* tenta aproveitar esta informação. Se se considerar apenas duas companhias, o que se está a dizer é que o modelo em causa em vez de ser descrito como:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$$

ou ainda como: $\mathrm{Y} = \mathrm{X}\beta + \mathrm{e}$, onde:

$$\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} = \boldsymbol{e} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1 I_T & 0 \\ 0 & \sigma_2 I_T \end{pmatrix} = \boldsymbol{W} \right] = e \sim ED(\boldsymbol{0}, \boldsymbol{W})$$

pode ser escrito como:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$$

onde

$$\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} = \boldsymbol{e} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\sigma_{11} I_T} & \boldsymbol{\sigma_{12} I_T} \\ \boldsymbol{\sigma_{21} I_T} & \boldsymbol{\sigma_{22} I_T} \end{pmatrix} = \boldsymbol{W} \right] = \boldsymbol{e} \sim ED(\boldsymbol{0}, \boldsymbol{W}).$$

Seguindo Griffiths, Hill & Judge (1993) e Zellner (1962) tem-se por máxima verosimilhança que:

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$= \left[\begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}' \underbrace{\begin{bmatrix}\widehat{\sigma_{11}}\boldsymbol{I_T} & \widehat{\sigma_{12}}\boldsymbol{I_T}\\ \widehat{\sigma_{21}}\boldsymbol{I_T} & \widehat{\sigma_{22}}\boldsymbol{I_T}\end{bmatrix}^{-1}} \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}\right]^{-1} \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}' \begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}$$

$$= \underbrace{\left[\begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}' \begin{bmatrix}\widehat{\boldsymbol{e_1}'\boldsymbol{e_1}}\boldsymbol{I_T} & \widehat{\boldsymbol{e_2}'\boldsymbol{e_1}}\boldsymbol{I_T}\\ \widehat{\boldsymbol{e_1}'\boldsymbol{e_2}}\boldsymbol{I_T} & \widehat{\boldsymbol{e_2}'\boldsymbol{e_2}}\boldsymbol{I_T}\end{bmatrix}^{-1} \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}\right]^{-1}} \begin{bmatrix}\boldsymbol{X_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X_2}\end{bmatrix}' \begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}$$

$$\text{Cov}(\hat{\hat{\boldsymbol{\beta}}}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}$$

Uma apresentação mais tranquila deste estimador é feita no capítulo 4, sendo um caso prático, com uma base de dados real é apresentada no final deste capítulo.

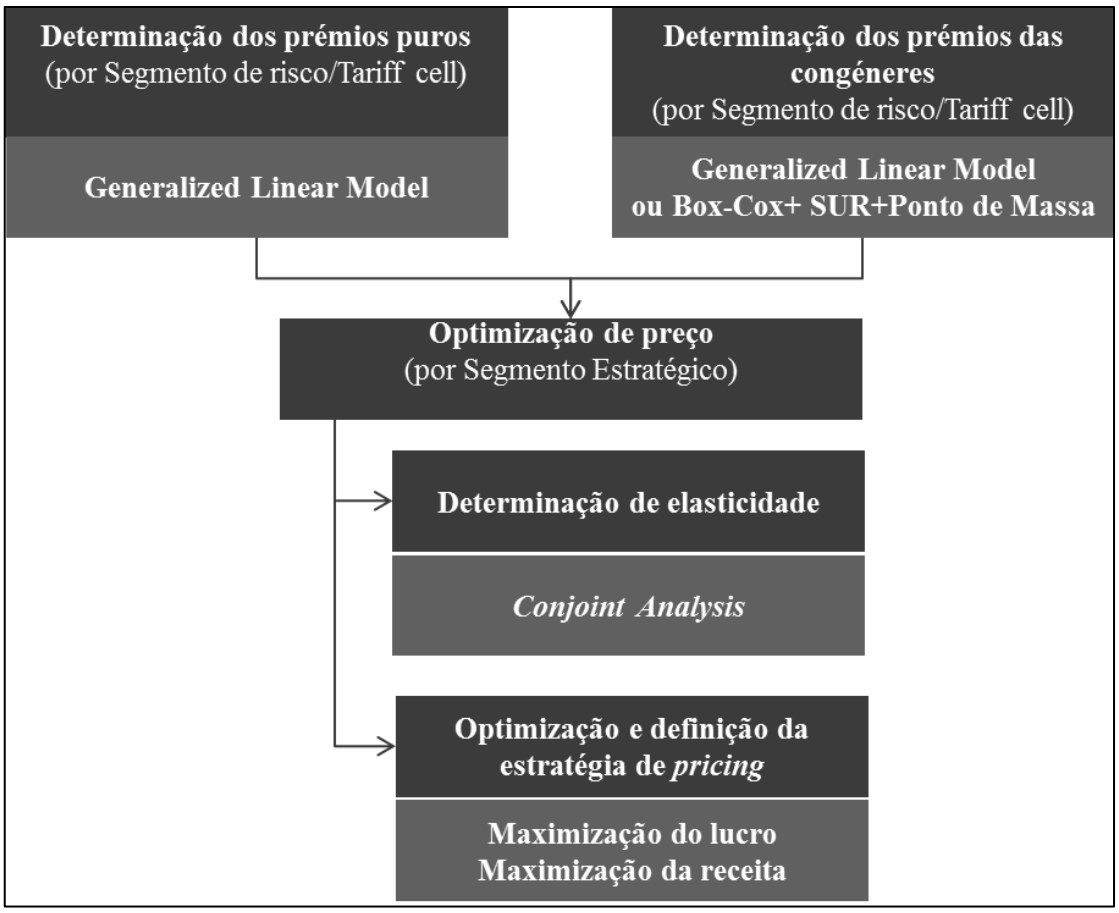
1.3.5. Organização do documento

A abordagem metodológica seguida será a seguinte:

- Determinação dos prémios puros por Segmento de risco/*Tariff cell* (Capítulos 2 e 3), sendo que no capítulo será dado especial destaque aos modelos GLM;
- Determinação dos prémios das congéneres (Capítulo 4), sendo dado especial destaque ao desenho experimental, à regressão Box-Cox, e ao modelo SUR;
- Apuramento das elasticidades (Capítulo 5) com o estudo da metodologia de *conjoint analysis*;
- Optimização e definição da estratégia de *pricing* (Capítulo 6) com estudo das curvas de procura e do modelo de consolidação de todos estes temas;
- Conclusões (Capítulo 7).

A imagem seguinte traduz a abordagem realizada, sendo que os termos técnicos aí referidos serão definidos nos capítulos relevantes.

Figura 1.3 - Abordagem proposta



Fonte: autor

Note-se que a estrutura de trabalho onde se combinam as ferramentas estatísticas e se define o preço baseado no valor a cliente é uma grelha/tabela baseada no conceito de *tariff cell* (combinação das características de seguros, objectos seguros e especificidades regionais), conforme pode ser visto na tabela abaixo.

O processo de optimização/definição do preço a cobrar é feito por linha levando em linha de conta a procura, que é definida pela metodologia de *conjoint analysis* e condicionada à oferta das companhas congéneres, apuradas com a metodologia SUR, e tendo em causa os prémios de risco e custos definidos pelos modelos GLM.

Tabela 1.1 - Tabela síntese das condições de mercado

Segmento de risco/ *Tariff cell*	**Prémio de risco (custos)**	**Prémio Cia A**	**…**	**Prémio Cia Z**	**Quota pref. – P_{High}**	**Quota pref. – P_M**	**Quota pref. – P_{Low}**	***Profit max. price***	***Revenue max. price***
1									
2									
….									
m-ésimo									

Fonte: autor

1.4. *Softwares* utilizados

Sempre que foi necessário concretizar alguns aspectos práticos desta tese recorreu-se ao R (R x64 3.1.3) e ao R Studio (vs 0.98.1103). Os pacotes utlizados foram:

1. DoE.base (vs 0.27): para o cálculo do desenho ortogonal;
2. dummies (vs 1.5.6): para a criação automática de variáveis *dummies*;
3. MASS (vs 7.3-39): para a regressão Box-Cox;
4. Rcmdr (vs 2.1-7): para geração das distribuições Poisson, gama e respectivas regressões;
5. statmod (vs 1.4.21): para a distribuição e regressão Tweedie;
6. systemfit (vs 1.1-16): para aplicação do modelo SUR.

Para verificar os cálculos do modelo SUR recorreu-se pontualmente ao STATA 11.

Em termos de *conjoint analysis* utilizou-se o SPSS vs 20. Infelizmente não há à data uma aplicação no R *cran* com esta metodologia.

Para a determinação da curva de procura e do preço óptimo recorreu-se a folha de cálculo (MS EXCEL), já que uma visualização gráfica é fundamental para a operacionalização da metodologia apresentada.

As bases de dados e a programação em R são disponibilizadas em anexo.

2. Breve indicação de mecanismos de construção dos modelos tarifários

De acordo com o *Guia de Seguros* (2013), do ISP – Instituto de Seguros de Portugal, a entidade reguladora dos seguros em Portugal, actualmente com a designação ASF – Autoridade de Supervisão de Seguros e Fundos de Pensões – "[o] contrato de seguro é um acordo através do qual o segurador assume a cobertura de determinados riscos, comprometendo-se a satisfazer as indemnizações ou a pagar o capital seguro em caso de ocorrência de sinistro, nos termos acordados. Em contrapartida, a pessoa ou entidade que celebra o seguro (o tomador do seguro) fica obrigada a pagar ao segurador o prémio correspondente, ou seja, o custo do seguro".

O conceito de prémio que o regulador apresenta, e tal como já referido, não é mais do que custo do seguro e deve corresponder, *grosso modo*, ao valor esperado das indemnizações. A diferença entre o custo e o prémio reside fundamentalmente na necessidade da seguradora fazer reflectir, no prémio, os custos de gestão e a margem exigida/desejada pelo accionista. Estes *loadings* (custos de gestão e margem) não serão analisados nesta tese, serão apenas parcelas a ser acrescentadas à tarifa final comercial.

Neste capítulo, só será analisado o prémio puro, aquele que representa o custo da indemnização associado ao risco subscrito. Desta forma, ao olhar para o mercado uma companhia tem de assumir a responsabilidade por um risco pelo qual não sabe se o seu verdadeiro custo. Para estimar esse custo, a Companhia segue as seguintes etapas: (i) analisa os factores de risco, (ii) define uma variável-chave que traduz o risco envolvido e (iii) desenvolve uma metodologia de cálculo que associa os factores de risco à variável-chave.

2.1. Factores de risco

Em cada apólice – contrato que formaliza a aceitação do risco sob determinadas condições – o prémio é determinado pelo valor de determinadas variáveis que se designam: factores de risco ou *rating factors*. Para estimar a relação entre estes factores de risco e o prémio, os factores de risco são habitualmente classificados em três categoriais:

- Características dos segurados;

- Características dos riscos seguros;
- Especificidades regionais[3] ou de contexto genérico.

Uma análise elementar poderia sugerir que estas características e especificidades entrassem no modelo de acordo com a sua natureza – sendo mais específico, se fossem variáveis contínuas deveriam entrar como variáveis contínuas; se forem binárias, ou ordinais, deveriam entrar com essa natureza.

Porém, de uma maneira geral, todas as variáveis entram de forma categórica. Ohlsson & Johansson (2010) defendem que esta metodologia de traduzir os factores de risco reside no facto de nos "*old days the tariff was a price list on paper*"; onde a cada combinação de factor-nível estavam associados uns pontos cuja soma total se traduzia num preço.

Mas há outras vantagens na discretização dos factores, por exemplo é mais fácil compreender/retractar formas não lineares e sobretudo, compreender o problema de selecção adversa. Como exemplifica Siddiqi (2005), no contexto de modelos de *scoring* para a actividade bancária, com a categorização é possível compreender por que motivo um segmento de risco ou *tariff cell* (combinação das características de seguros, objectos seguros e especificidades regionais) não pode suportar os riscos de outro segmento. Ou seja, em mercados liberalizados, como são, de uma maneira geral, os seguros, não é possível ter um prémio puro médio idêntico para vários segmentos; já que podendo discriminar bem os clientes por segmento de risco, uma outra seguradora poderia oferecer uma tarifa mais adequada para um segmento específico e ter lucro, ao mesmo tempo que deixava os clientes com risco acima da média na companhia inicial. Este conceito é muito importante e explica por que motivo não é possível ter um prémio médio em vários segmentos ou permitir subsidiação cruzada entre segmentos em mercados competitivos (*i.e.* onde há liberdade comercial de atrair e apostar em diferentes segmentos). Assim, as companhias de seguro tentam ajustar os

[3] Este projecto centra-se sobretudo em ramos particulares não-vida, e sempre que necessário no ramo automóvel. Para ramos empresariais deverão considerar-se os ramos regionais.

segmentos de risco a planos de segmentação/*marketing* e a diferentes planos tarifários.

Sistematizando, por haver alguma racionalidade e alguma tradição de mercado os factores de risco de uma maneira geral são categorizados ou alvo de uma discretização.

Tabela 2.1 - Exemplo de construção de tariff cell

Segmento de risco/*Tariff cell*	***Covariates***			**Duração (Exposição)**	**Frequência**
	Relação peso potência	**Idade**	**Zona N1/N2/N3**		
1	*Nível 11*	*Nível 21*	*Nível 31*	A definir	A definir
2	*Nível 12*	*Nível 22*	*Nível 32*	A definir	A definir
3	*Nível 13*	*Nível 23*	*Nível 33*	A definir	A definir
…	…	…	…	…	…

Fonte: autor

2.2. Variável chave

Já foi analisado, ainda de que de forma não exaustiva, quais as variáveis que poderiam traduzir o risco de uma apólice de seguro. Mas não foi, porém, explicitada qual a variável objectivo que traduzirá o risco.

Assim, é necessário definir alguns conceitos básicos (Ohlsson & Johansson (2010)):

- Duração da apólice (*duration of a policy*): é o tempo que uma apólice está em vigor, sendo habitualmente contabilizada em anos. A duração total de um grupo de apólices é conseguida somando as durações das apólices individuais;
- Sinistro (*claim*): De acordo com o Glossário do ISP (consultado em Setembro de 2014) “trata-se de um evento ou série de eventos resultantes de uma mesma causa susceptível de fazer funcionar as garantias de um ou mais contractos de seguro”;

- Frequência de sinistros (*claim frequency*): é o número de sinistros sobre duração de um conjunto de apólices. De uma maneira geral este número é dado em percentagem ou permilagem e em termos médios: média de número de sinistros num dado ano;
- Severidade de sinistros (*claim severity*): Trata-se do montante despendido sobre o número de sinistros – *i.e.* trata-se de um custo médio. De uma maneira geral este número está expurgado dos sinistros que não apresentam qualquer custo;
- Prémio puro (*pure premium*): Traduz o valor médio do sinistro sobre a duração, *i.e.*, o custo médio anual (habitualmente é esta a medida de tempo). Consequentemente, o prémio puro é o produto da frequência e da severidade de sinistros;
- Prémios recebidos (*earned premium*): Trata-se do valor recebido por unidade de tempo; ou seja, é a duração (medido em anos) multiplicado pelo prémio anual;
- Rácio de sinistralidade (*claims ratio*): Trata-se do valor total de sinistros sobre o total dos prémios recebidos.

Frequência de sinistros (*claim frequency*), severidade de sinistros (*claim severity*), prémio puro (*pure premium*) e rácio de sinistralidade (*claims ratio*) são designados por rácios chave (*key ratios*).

Tabela 2.2 - Rácios chave (*key ratios*)

Exposição $(\boldsymbol{w})$	**Resposta** $(\boldsymbol{X})$	**Rácio chave** $(\boldsymbol{Y = X/w})$
Duração	Número de sinistros	Frequência de sinistros
Duração	Custo de sinistros	Prémio Puro
Número de sinistros	Custo de sinistros	Severidade (média) de sinistros
Prémios recebidos	Custo de sinistros	Rácio de sinistralidade (Claims ratio)

Fonte: autor

2.3 Métodos de apuramento de risco

Há agora que perceber como relacionar estes factores categorizados e o risco subjacente, traduzido no prémio puro (ou em qualquer um dos rácios chave). Há duas grandes alternativas:

- Tabelas de duas entradas: um método em desuso;
- Modelos GLM: o método mais robusto;

que serão, nas secções seguintes, explicitadas.

2.3.1. Tabelas de duas entradas – um método em desuso

Os métodos mais tradicionais, nos idos anos 60, funcionavam como tabelas de duas entradas, testes e muitas iterações. Uma análise preliminar pode ser encontrada logo nas primeiras páginas de Feldblum & Brosius (2002).

A ideia era fazer tabelas de duplas entradas e "apanhar" as relações relativas dos factores. Essas relações relativas podem ser escritas em forma de equação, assumindo um caso base. O objectivo deste método é de forma iterativa resolver todas as equações subjacentes a cada um dos factores. Este método tem vários problemas (ver Donlan & Turnacioglu (2006) e Anderson, Modlin, Schirmacher, Schirmacher, Thandi (2007)):

- Não haverá uma solução única, mas apenas aproximada de resolver a questão;
- Assume-se que os efeitos entre os factores são independentes e não se considera a correlação ou interacção entre variáveis;
- A forma funcional da equação que gera a tarifa depende dos dados disponíveis (o habitual); mas sobretudo da experiência dos analistas;
- Não há uma forma objectiva de validar os resultados.

De uma forma simples, não se trata de um processo preciso de desenhar os preços, mas de um ajustamento numérico.

Ainda hoje esta técnica é usada, mas com relativa penetração no mercado. Uma análise *one-way*, como este método é também designado, centra a análise nos rácios chave de frequência ou de rácios de sinistralidade (*claims ratio*).

2.3.2. Modelos GLM: o método mais robusto

Tal como indicado acima, a análise e construção da tarifa por *tariff cell* é uma opção muito comum em seguros e que representa uma forma de combater a selecção adversa. Assim, uma forma expedita de estimar o modelo seria a de olhar para as despesas técnicas (ou outro *key ratio*) associado a cada uma das *tariff cells* (tendo o cuidado de ajustar os efeitos de inflação).

Contudo, na vida real facilmente há *tariff cells* não preenchidas com o volume de dados necessários. A título ilustrativo repare-se que se se tiver três factores tarifários, cada um dos quais com três níveis, haverá 27 combinações/*tariff cells*. Cada um destes deverá ter no mínimo 30 casos para se estimar uma média relevante o que significa que são necessários 810 casos, distribuídos uniformemente, para conseguir retirar qualquer conclusão. Porém, basta que um dos níveis, de um dos factores tarifários, tenha uma taxa de penetração de 5%, para que a amostra global necessária para preencher essa *tariff cell* seja de 5.045. Tal acontece pois, com 5% de penetração, haverá nove *tariff cells* com 0,58% de possibilidade de preenchimento. Sabendo que 5% de penetração é apesar de tudo um aspecto optimista, que há mais factores e níveis a considerar, que há *outliers* e que há efeitos cruzados, facilmente se compreende que este efeito a amostra necessária tomará uma dimensão muito pouco razoável. É ainda de referir que no exemplo dado por Smyth (2014) baseado em Hallin & Ingenbleek (1983) foram analisados 2.383.170 apólices automóveis na Suécia em 1977 e nem todas as 2205 *tariff cells* estão preenchidas. Assim, não se pode apenas construir tabelas com *tariff cells* e observar o custo técnico – este pode não existir em todas as *tariff cells*.

Tendo em conta a dificuldade amostral, o desafio será como apurar o custo técnico para as *tariff cells* com poucos ou nenhuns casos preenchidos, garantindo porém a estabilidade de estimação. Será necessário usar um modelo que estime os resultados numas classes e se obtenha outras.

Mesmo supondo que é possível ter uma amostra razoável em cada uma das *tariff cells*, e observar directamente a frequência e os custos directos, há vantagens em usar um modelo de estimação. Uma companhia de seguros deve compreender a 'força' que cada factor tarifário tem na frequência e no custo médio do seu processo de estimação.

A análise que a seguir se apresenta, responde às necessidades acima expressas, e centra-se apenas na estimação da frequência de um sinistro e o seu custo médio. Serão excluídos da análise os outros *key ratios*. Esta opção prende-se apenas com o objectivo da tese e está alinhada com a prática de mercado – porém, é de salientar que, como já visto, o racional de apuramento dos *key ratios* é o mesmo, e o processo de estimação, apresentado de seguida, pode ser generalizado.

A equação (1) indica como habitualmente se estima o prémio puro:

$$\text{Prémio puro}_i = \text{constante} + \text{frequência sinistro}_i \times \text{custo médio}_i + \text{erro}_i \qquad (1)$$

com i correspondendo a uma *tariff cell*.

Para apurar os termos da equação (1), o mecanismo mais tradicional é através de modelos GLM compósitos, *i.e.* misturando a frequência de sinistro (geralmente assumindo um modelo de contagem/Poisson) com o custo médio assente na distribuição gama (ver Anderson, Modlin, Schirmacher, Schirmacher, Thandi (2007) e Maccullagh & Nelder (1989)). Os dados são ponderados por forma a dar mais peso aos sinistros já tratados (onde o custo médio está todo apurado) ou todos os sinistros foram comunicados[4]. A

Figura 2.1 indica a forma como a equação (1) pode ser desdobrada nos dois modelos mais simples mencionados:

[4] A questão da ponderação dos dados, que garante uma maior eficiência no processo de estimação GLM, será analisado nas secções seguintes.

$$\text{Frequência sinistro}_{\,i} = \text{constante} + \ \text{f}_{linear}\,(\text{variáveis explicativas}) + \text{erro}_i \qquad (2)$$

com i correspondendo a uma *tariff cell*.

$$\text{Custo médio}_{\,i} = \text{constante} + \ \text{f}_{linear}\,(\text{variáveis explicativas}) + \text{erro}_i \qquad (3)$$

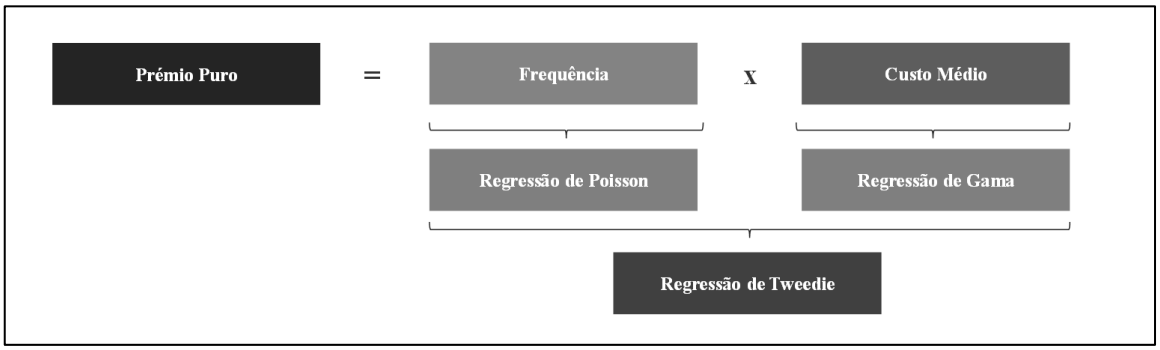
com i correspondendo a uma *tariff cell*.

Sendo que $\text{f}_{linear}\,(.)$ é uma função diferente em cada uma das equações apresentadas.

Evidentemente que se poderia estudar a equação (1) directamente. Porém nem sempre é fácil. Meyers (2009) explica porquê: os modelos tradicionais de regressão não lidam bem com a mistura de perdas nulas (discretas) com perdas contínuas positivas e não simétricas.

Smyth & Jørgensen forneceram em 2002 uma solução. Estes autores argumentam que o desafio está em estimar a dispersão associada dos custos que tem de ser efectuada simultaneamente e que tal é possível na designada regressão Tweedie. A Figura 2.1 retracta esta abordagem. Note-se que por vezes na literatura a regressão de Tweedie também é designada por *compound Poisson* e *compound gamma*.

Figura 2.1 - Esquema de aplicação do modelo GLM à tarifação



Fonte: autor

Porém, subsiste ainda um outro problema: quando se combina os dois fenómenos

(de frequência e custo médio) implicitamente está-se a assumir que os factores que afectam os fenómenos de frequência e custo médio são os mesmos e que têm o mesmo modelo de impacto. Em qualquer caso, voltaremos a estimação destes efeitos conjuntos mais adiante.

Habitualmente o processo de estimação das equações (2) e (3), acima mencionadas, têm por base o modelo GLM. De uma forma breve, estes modelos (GLM) generalizam a regressão habitual (OLS) para distribuições associadas ao erros não normais. As distribuições têm de ser da família exponencial: normal, Bernoulli, binomial, Poisson, gama, etc. Ou seja, a formulação de um modelo GLM inclui três componentes:

- Componente aleatória – identificação da distribuição associada a Y.
- Componente sistemática – forma de traduzir a heterogeneidade dos dados como preditores lineares: ($X\beta$).
- *Link function* – função invertível ($g(.)$) que liga a média da resposta ($E[\mathrm{y}_i] = \mu_i$) à componente sistemática (habitualmente através dos parâmetros da componente aleatória).

Note-se que estes métodos são de facto *standard* no mercado, quando se pretende a tarifação. Anderson, Modlin, Schirmacher, Schirmacher, Thandi (2007) referem mesmo que “os modelos GLM são largamente reconhecidos na indústria como o método *standard* para tarifar automóveis ligeiros de passageiros e outras linhas de pequenos ramos comerciais”. Estes autores indicam ainda que têm evidências de que estes modelos são usados por companhias no RU, Irlanda, França, Itália, Holanda, países da Escandinávia, Espanha, Portugal, Bélgica, Suíça, África do Sul, Israel, Austrália. O documento refere ainda que este modelo tem ganho popularidade no Canadá, Japão, Coreia, Brasil, Singapura, Malásia e países do leste europeu.

É ainda de referir que para estes processos de estimação, Anderson; Feldblum; Modlin; Schirmacher; Schirmacher; & Thandi (2007) afirmam que dependendo da frequência e do número de factores analisados pode ser necessário 100 mil

exposições. Ou seja, há uma desvantagem na aplicação destes métodos GLM: um elevado volume de informação.

3. Estudo dos Modelos GLM

O objectivo deste capítulo é o de indicar uma abordagem que permita apurar os custos de risco (o prémio puro) de um determinado ramo de uma companhia de seguros.

Nas secções seguintes serão explicados os casos das regressões de Poisson, gama e Tweedie. Para cada uma destas regressões, será estudado (de forma sintética) o (i) comportamento da função distribuição; (ii) as razões de heterogeneidade subjacente; (iii) modelo de estimação; (iv) aplicação para o tema em estudo (frequência, custo médio e efeito combinado). A demonstração destes modelos de regressão será adiante formalizada, mas dando um cariz interpretativo aos objectivos da tese e à indústria seguradora em geral.
É de salientar que a escolha de cada um destes métodos de estimação está associado a um dos rácios chave, apresentados na Tabela 2.2.
Em anexo são apresentados alguns casos práticos para cada uma das técnicas apresentadas com recurso a uma base de dados clássica.

3.1. Regressão de Poisson

Para estimar a frequência de uma apólice habitualmente exige-se o cumprimento dos seguintes pressupostos (Ohlsson & Johansson (2010)):

a) Independência das apólices: a resposta das apólices deverão ser independentes – é fácil imaginar algumas situações em que tal não acontece – por exemplo, catástrofes naturais, ou acidentes em veículos de frotas.

b) Independência temporal: A probabilidade de frequência de um sinistro é constante ao longo do tempo – é fácil imaginar algumas situações em que tal não acontece: depois de uma casa roubada, colocam-se trancas à porta; um indivíduo com um sinistro adopta um comportamento mais prudente.

c) Homogeneidade: As apólices que constam da mesma *tariff cell* deverão apresentar o mesmo comportamento. Mas como se define o mesmo

comportamento? Qual o grau de homogeneidade desejada? Há por vezes um reduzido acesso ao conhecimento, pelo que os subscritores optam por agravar o modelo de risco estimado e adoptam políticas de *Bonus-Malus* que tentam incluir experiência mais recente no risco estimado.

Com base nestas condições é possível concluir que o modelo de frequência é independente. E acima de tudo que se está nas condições de uma distribuição de Poisson: contagens independentes. Note-se que para obter uma frequência, como anteriormente definida, é necessário dividir o modelo estimado pelas apólices presentes[5].

3.1.1.Poisson: Distribuição

Diz-se que y tem uma distribuição de Poisson com parâmetro $\mu > 0$ se:

$$P(y|\,\mu) = \frac{exp(-\mu)\mu^{y}}{y!}, \text{ para } y = 0,1,2 \ldots. \quad (4)$$

Esta distribuição traduz a probabilidade ($\mathrm{P}(\mathrm{y}|\,\mu)$) de uma série de eventos independentes (y) ocorrerem num dado período de tempo. O parâmetro μ irá traduzir o número esperado de ocorrências que ocorrem num dado intervalo de tempo. Assim, por exemplo:

- Se y=0, temos $P(0|\,\mu) = \frac{exp(-\mu)\mu^{0}}{0!} = exp(-\mu)$;
- Se y=1, temos $P(1|\,\mu) = \frac{exp(-\mu)\mu^{1}}{1!} = exp(-\mu)\mu$;
- (...)

[5] Note-se poderia ser usado uma regressão logística. Porém, além de esta não verificar em pleno as condições acima enunciadas, o resultado da regressão logística será a probabilidade de apenas um sinistro por apólice.

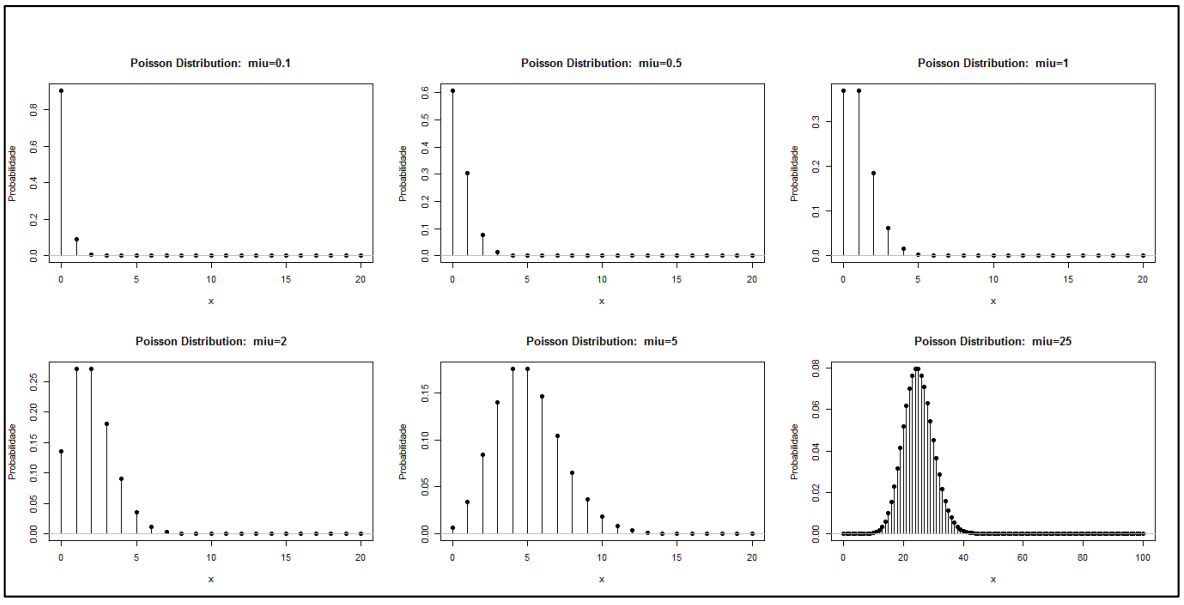
- Se y=3, temos $P(3|\ \mu) = \frac{exp(-\mu)\mu^3}{3!} = exp(-\mu)\mu^3\ /6.$
- (...)

Vale a pena estudar um pouco esta distribuição. Algumas considerações:

1. Pode-se definir μ como a taxa de ocorrência de um dado acontecimento num dado período de tempo. Trata-se de uma média;
2. A variância, numa distribuição de Poisson, é igual à média;
3. À medida que μ aumenta, a probabilidade de ter contagens próximas de zero são reduzidas (ver como ilustração: Figura 3.1);
4. À medida que μ aumenta, a distribuição "chega-se para a direita". Abaixo segue uma ilustração no "R".

À medida que μ aumenta a distribuição de probabilidade aproxima-se de uma normal (ver Figura 3.1).

Figura 3.1 - Número esperado de ocorrências que ocorrem num dado intervalo de tempo



Fonte: autor

3.1.2. Poisson: Heterogeneidade

Imagine-se (ver Hassett & Stewart (2006)) que se tem uma companhia de seguros onde os seus clientes apresentam um sinistro a uma taxa média de $\mu = 0{,}45$ ao ano e que tem 500 clientes por ano. Qual a probabilidade de um cliente apresentar exactamente um sinistro?

Se se considerar que Y representa o número de sinistros, então:

$$P(y = 1|\mu = 0{,}45) = \frac{exp(-0{,}45)0{,}45^1}{1!} \approx 0.2869.$$

A questão que se coloca, com o modelo simples de Poisson, é a seguinte: será que a média de contagem representa completamente o modelo em causa? Ou melhor ainda, será que a taxa de apresentação de sinistros é idêntica entre todos os indivíduos?

Por exemplo, e adaptando o caso de Long (1997) na apresentação da equação de Poisson, considere-se que a taxa de sinistros (média e variância) dos homens é de $\mu+\delta$; e que a das mulheres é de $\mu-\delta$; e que o processo gerador de sinistros segue uma Poisson. A taxa de sinistralidade conjunta será a média de homens e mulheres: $\mu = [(\mu+\delta) + (\mu-\delta)]/2$; mas a variância, evidentemente, irá exceder μ.

Uma solução possível seria a de calcular um modelo tarifário de seguros para homens e outro para mulheres; ou, generalizando, um modelo para cada *tariff cell*, onde esta era apenas explicada pelo sexo. Porém, tal exigiria um ainda maior volume de casos para cada célula em estudo, ou uma reduzida estabilidade do modelo e resultados que não contemplam a totalidade do modelo estimado. Adicionalmente, é de referir que a possibilidade de unir *tariff cells* (por questões de insuficiência de dados, ou por sensibilidade comercial, por exemplo) seria fortemente desaconselhada, já que o modelo agregado poderia não ser a soma dos restantes.

Este exemplo e estas considerações permitem chegar à regressão de Poisson. Esta introduz a heterogeneidade dos indivíduos, através das características

observadas e estabelece condições de suficiência para a centragem e para a eficiência (modelo linear nos parâmetros, amostra aleatória, erros não correlacionados com as variáveis explicativas, ausência de multocolinearidade perfeita) do modelo ao usar todos os dados disponíveis. Infelizmente, mostra que dificilmente a média será idêntica à variância[6] em modelos de contagem.

Ou seja, pretende-se, de alguma forma, que a média μ, do número de eventos Y, que tem uma distribuição de Poisson, dependa das características do indivíduo e que a transição entre as características dos produtos seja "suave" e estável ao longo das *tariff cells*. Assim tem-se:

$$\mu_i = E(y_i|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i\boldsymbol{\beta}), \quad (5)$$

Ou seja, a equação (5) indica que:

- Há uma relação entre as variáveis observadas e a taxa de apresentação de sinistros;
- A relação funcional entre a taxa de apresentação de sinistros e as variáveis observadas é exponencial.

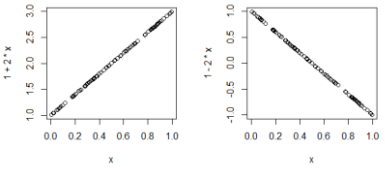
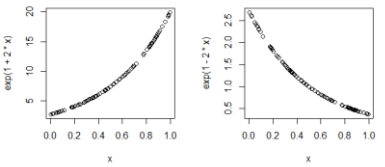
Note-se que a forma exponencial força μ_i a ser positivo; o que é exigido pela distribuição de Poisson. Outras relações são possíveis entre μ_i e $\boldsymbol{x\beta}$ (como $E(y|x) = \boldsymbol{x\beta}$**)**; mas na verdade estas opções raramente são usadas. A forma funcional tem de ser tal, que (i) minimize a eventual heterocedasticidade dos dados ("*nice to have*"), (ii) obrigue a um comportamento de "μ_i" estritamente positivo, e (iii) que μ_i tenha um comportamento pouco sensível a *outliers*. Mas, acima de tudo, a forma funcional tem de ser tal (iv) que faça sentido quando

[6] Felizmente (ver Wooldridge (2003)) prova-se que o estimador de Poisson é consistente, mesmo quando a variância não é idêntica a média prevista. Ainda assim se se quisesse corrigir esta situação poder-se-ia utilizar a distribuição binomial negativa – Wooldridge (2003) e Long (1997). A demonstração destes resultados ou análise destas alternativas não será efectuada.

analisada.

Juntando estas quatro condições, com especial ênfase à condição (iv), é habitual optar pelo modelo exponencial (log-lin) tanto mais que os coeficientes têm uma leitura quase imediata: traduzem o incremento ou a elasticidade. A tabela abaixo ilustra esta questão.

Tabela 3.1 - Algumas formas funcionais da função link no contexto da regressão de Poisson

Formas funcionais	Função base	Forma não linear	Imagem gráfica	Modelo estatístico	Declive
Modelo linear	y=a+bx	-----		$y_i=\beta_1+\beta_2x_i+e_i$	β_2
Modelo Exponen-cial (log-lin)	y=exp(a+bx)	$y=\exp(\beta_1+$ $+\beta_2x_i+e_i)$		$\ln(y_i)=\beta_1+$ $+\beta_2x_i+e_i$ com y >0	$\beta_2\ y_i$

Fonte: autor

Com a equação (5) tenta-se que a variância dependa da característica do indivíduo. Há agora que ligar (*link*) esse comportamento com a distribuição de Poisson.

3.1.3. Poisson: Processo de estimação[7]

A distribuição da contagem de sinistralidade será dada por:

[7] Para resolver este processo de estimação será seguido de muito perto a análise apresentada por Maddala (1983), pp52 e seguintes. Wooldridge (2003) apresenta também um processo de estimação, mas minimizando a soma dos quadrados dos resíduos.

$$P(y_i|\, x_i) = \frac{exp(-\mu_i)\mu_i{}^{y_i}}{y_i!},\ \text{para } y = 0,1,2\ \ldots. \tag{6}$$

Tendo em conta a equação (3) já apresentada:

$$\mu_i = E(y_i|x_i) = exp(\boldsymbol{x}_i\boldsymbol{\beta}), \tag{3}$$

tem-se,

$$P(y_i|\, x_i) = \frac{exp(-\mu_i)\mu_i{}^{y_i}}{y_i!} = \frac{exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))\exp(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i}}{y_i!} \tag{7}$$

Esta fórmula (7) é a base do modelo de regressão linear de Poisson. Tendo obtido os $\boldsymbol{\beta}$, bastaria colocar os valores na fórmula para se obter a média ou a distribuição estimada. A secção seguinte indica como se obtém uma estimativa para $\boldsymbol{\beta}$.

Com base na equação (6) é possível construir a função de verosimilhança da seguinte forma:

$$L(\mu_i|\boldsymbol{y},\boldsymbol{X}) = \prod_{i=1}^{n} P(y_i|\, x_i) = \prod_{i=1}^{n} \frac{exp(-\mu_i)\mu_i{}^{y_i}}{y_i!} \tag{8}$$

Esta forma pode ser simplificada aplicando o operador logaritmo e a função *link*:

$$L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{\exp\left(-\sum_i \mu_i + \sum_{j=1}^{p} \beta_j \sum_i x_{ij} Y_i\right)}{\prod_i Y_i!} \Rightarrow$$

$$\Rightarrow \log(L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})) = -\sum_i \mu_i + \sum_{j=1}^{p} \beta_j \sum_i x_{ij} Y_i - \sum_i \log(Y_i!)$$

$$\Rightarrow \log(L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})) = -\sum_i \mu_i + \beta_0 \sum_i Y_i + \sum_{j=1}^{p} \beta_j \sum_i x_{ij} Y_i - \sum_i \log(Y_i!) \tag{9}$$

Este modelo de estimação é o primeiro passo para os dados de contagem. As equações de primeira ordem são:

$$\frac{\partial .}{\partial \hat{\beta}_0} = 0 <=> \sum_i Y_i = \sum_i \widehat{\mu_i}$$

$$\frac{\partial .}{\partial \hat{\beta}_i} = 0 <=> \hat{\beta}_i = \sum_i x_{ij}\mu_i, \text{ com } i = 1,2,\dots,p$$

Onde, $\hat{\mu_i} = \exp(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta_j} \sum_i x_{ij})$

Maddala (1983) afirma que estas equações podem ser resolvidas de forma numérica pelo método Newton-Rapshon. A matriz das segundas derivadas é dada por:

$$\frac{\partial^2 .}{\partial \hat{\beta}_i \partial \hat{\beta}_k} = -\sum_i x_{ij} x_{ik} \mu_i \text{ , com } j,k = 0,1,2,\dots p \text{ e } x_{i0} = 1.$$

3.1.4. Poisson: *Offset* e ponderação

Se houver poucos casos observados em cada umas das *tariff cells* o modelo de contagem estimará, evidentemente, também poucos sinistros. Se houver mais casos observados, haverá também mais sinistros. Assim, independente dos factores tarifários que compõem a *tariff cell*, a sinistralidade estimada dependerá do número de casos presentes.

Há duas soluções habituais para este tipo de problema:

1. Ponderar os dados;
2. Incluir um *offset*.

De acordo com Siddiqi (2005) a ponderação é muito útil para gerar resultados baseados em *scorecards.* Mais, defende este autor, este ajustamento dá uma melhor interpretação nas variáveis explicativas categóricas, como é, em certa medida, o caso.

Para obter o ponderador para cada *tariff cell* i basta aplicar a fórmula abaixo:

$$Ponderador_i = \left(\frac{N}{\#\ tariff\ cells}\right) / (\text{número de casos presentes na } tariff\ cells_i)$$

Onde, como habitualmente, $N = N\text{úmero de observações } total$.

Uma outra solução seria somar um *offset* na equação a estimar. Ou seja, se se considerar (ver Ooi (2013)) o *offset* como $\log \varepsilon$, sendo que ε é exposição, tem-se que o modelo a estimar é dado por:

$$\log E(Y) = \beta' X + \log \varepsilon$$

que pode ser simplificado em:

$$\log E(Y/\varepsilon) = \beta' X.$$

A variável a estimar continua a ser Y, mas ao dividir por εestá-se a estimar uma equação por rácio de sinistros por unidade de risco. Esta divisão altera a variância da resposta; pelo que a inclusão de uma *offset* obriga a uma ponderação dos dados, por ε. Para complicar, note-se que neste caso o algoritmo de estimação não pode ser o de Poisson, já que não se está a falar de contagens, mas da estimativa de um rácio. Neste casos, habitualmente usa-se o método quasipoisson (ver Zeileis, Kleiber e Jackman (2008)).

Há contudo uma dificuldade neste método: como transpor as regras de classificação de dados para o modelo final, quando temos *tariff cells* sem ser observadas. A solução mais expedita é construir uma tabela com todas as *tariff cells* (será sempre uma tabela finita) e aplicar o modelo de previsão.

3.1.5. Poisson: Interpretação

Há várias formas de interpretar os resultados um modelo de contagem, dependendo se o que interessa é o valor esperado da contagem ou a distribuição associada à mesma. Estas formas de interpretação são agora apresentadas:

- Para o valor esperado $E(y_i|\, \boldsymbol{x_i})$, para um dado x_k, tem-se:

$$\frac{\partial E(y_i|\, \boldsymbol{x_i})}{\partial x_{ki}} = \frac{\partial \exp(\boldsymbol{x_i\beta})}{\partial \boldsymbol{x_i\beta}} \frac{\partial \boldsymbol{x_i\beta}}{\partial x_{ki}} = \exp(\boldsymbol{x_i\beta})\, \beta_k = E(y_i|\, \boldsymbol{x_i})\, \beta_k \qquad (10)$$

Uma vez que o modelo é não linear, o efeito do valor marginal depende do coeficiente associado a x_k e do valor esperado de y para um dado $\boldsymbol{x_i}$**.** Ou seja, o efeito marginal sobre $E(y_i|\,\boldsymbol{x_i})$ depende do nível de todas as restantes variávies. Habitualmente, para efeitos de compreensão é habitual colocar o vector $\boldsymbol{x_i}$**,** excepto x_{ki}, idêntico à média (ou a moda no caso de variáveis categóricas).

- Para o valor esperado $E(y_i|\,\boldsymbol{x_i})$, para uma variação de x_k, é necessário compreender que:

$$E(y_i|\,\boldsymbol{x_i}) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_K x_K)$$

$$= \exp(\beta_0)\exp(\beta_1 x_1) \ldots \exp(\beta_k x_k) \ldots \exp(\beta_K x_K)$$

Então:

$$E(y_i|\boldsymbol{x_i}, x_{ki} + \delta) = \exp(\beta_0)\exp(\beta_1 x_1) \ldots \exp(\beta_k x_k)\exp(\beta_k \delta) \ldots \exp(\beta_K x_K)$$

O que significa que a variação de $\boldsymbol{x_i}$, pode ser definida como:

$$\frac{E(y_i|\boldsymbol{x_i}, x_{ki}+\delta)}{E(y_i|\boldsymbol{x_i})} = \exp(\beta_k \delta) \qquad (11)$$

Mais uma vez, o efeito do valor marginal depende do coeficiente associado a x_k e do valor esperado de y para um dado $\boldsymbol{x_i}$**.** Ou seja, o efeito marginal sobre $E(y_i|\,\boldsymbol{x_i})$ depende do nível de todas as restantes variáveis. Habitualmente, para efeitos de compreensão é habitual colocar o vector $\boldsymbol{x_i}$**,**

excepto x_{ki}, idêntico à média (ou a moda no caso de variáveis categóricas).

- Para o valor esperado $E(y_i|\,\boldsymbol{x_i})$, para uma variação discreta de x_k, ou assumindo que x_k é uma variável categórica tem-se:

$$\frac{\Delta E(y_i|\boldsymbol{x_i})}{\Delta x_k} = E(y_i|\boldsymbol{x_i},\ x_k = nivel\ j) - E(y_i|\boldsymbol{x_i},\ x_k = nivel\ i) \quad (12)$$

Mais uma vez, o efeito do valor marginal depende do coeficiente associado a x_k e do valor esperado de y para um dado $\boldsymbol{x_i}$**.** Ou seja, o efeito marginal sobre $E(y_i|\,\boldsymbol{x_i})$ depende do nível de todas as restantes variáveis. Habitualmente, para efeitos de compreensão é habitual colocar o vector $\boldsymbol{x_i}$**,** excepto x_{ki}, idêntico à média (ou a moda no caso de variáveis categóricas).

- Para compreender a distribuição de y_i, para uma dada contagem específica (m) considerando que o vector $\boldsymbol{x_i}$ é conhecido, tem-se:

$$P(y_i|\,x_i) = \frac{\exp(-\hat{\mu})\hat{\mu}^m}{m!} \quad (13)$$

onde $\hat{\mu} = \exp(\boldsymbol{x_i}\widehat{\boldsymbol{\beta}})$**.** Assim, para apurar a distribuição para um dado $\hat{\mu} = \exp(\boldsymbol{x_i}\widehat{\boldsymbol{\beta}})$ bastará ir escolhendo valores específicos de m distintos.

Este aspecto, de compreender a distribuição de y_i, é especialmente relevante no contexto de regressão de Poisson. A distribuição subjacente não é simétrica e a sua "cauda" tem um espaçamento que é difícil de compreender. A média subjacente de y_i será obtida pela seguinte fórmula:

$$\overline{P(y_i = m)} = \frac{1}{N}\sum_{i=1}^{N} P(y_i = m|\, x_i) \; = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(-\hat{\mu}_i)\hat{\mu}_i^m}{m!} \qquad (14)$$

- *Partial Residual Plots*: Efeito de distribuição de uma variável, mantendo todos os restantes constantes.

  O *Partial Residual Plots* é uma forma gráfica e eficiente de avaliar o efeito de uma variável no modelo, mantendo as restantes constantes. Do ponto de vista teórico o algoritmo é bastante simples e pode ser executado da seguinte forma:

  $$X_{j\in K,i} = \widehat{\beta_J} X_{j\in K,i} + erro_i \; .$$

3.2. Regressão gama

Da análise de uma qualquer curva de custos médios de sinistros é evidente a observação dos seguintes factos comportamentais:

a) Custos não negativos: Os custos de um sinistro são sempre positivos. Um sinistro de valor nulo ou negativo deve ser considerado como pouco comum, senão mesmo impossível, caso contrário não daria aso à participação de sinistro;

b) Dados assimétricos à direita: A curva de distribuição de probabilidade dos custos é assimétrica à direita, ou seja, a maior parte da área limitada pela função densidade localiza-se nas proximidades da origem. A direita a mesma curva de probabilidade decresce gradualmente, conforme os custos crescem;

c) Variância está positivamente correlacionada média: Quanto maior a média, maior a variância – sendo ainda mais preciso o coeficiente de variação é constante: a variância é proporcional à média ao quadrado.

Com base nestas observações é possível concluir que o modelo de custos tem

uma “bossa” que nasce na origem e se vai prolongando. Há várias distribuições que se assemelham a este comportamento sendo as distribuições a log-normal e a gama as mais conhecidas. A distribuição gama é a mais usada, por uma questão de flexibilidade de forma funcional (pertence à família exponencial), pelo facto de os erros estarem numa escala razoável (por oposição a logaritimizada) (McCullagh & Nelderm (1989)) e também por tradição (Ohlsson & Johansson (2010)). McCullagh & Nelderm (1989), no seu trabalho seminal, indicam que a distribuição gama é também mais estatisticamente eficiente para um processo de variância constante[8].

Note-se que o custo de uma apólice deverá cumprir os pressupostos, já discutidos no contexto da regressão de Poisson, de independência das apólices, independência temporal e homogeneidade. Também não é difícil, tal como no caso da Poisson, de encontrar casos na indústria seguradora em que estes pressupostos possam ser questionados.

3.2.1. Gama: Distribuição

Diz-se que y tem uma distribuição gama se a sua função densidade de probabilidade for da forma:

$$P(y|p,\lambda) = \begin{cases} 0; & y \leq 0 \\ \frac{\lambda^p}{\Gamma(p)} y^{p-1} \exp(-\lambda y); & y > 0; p, \lambda > 0 \end{cases} \quad (15)$$

Onde $\Gamma(p) = \int_0^{+\infty} y^{p-1} \exp(-y)\, dy$

O integral converge para $p > 0$ e define a função conhecida por função gama. Note-se que para esta função se tem:

[8] Para tal este autores referem que a transformação $\log(y)$ tem aproximadamente os seguintes momentos:
$E(\log(Y)) = \log(\mu) - \frac{\sigma^2}{2}$ e $Var(\log(Y)) \approx \sigma^2$, o que permite uma aproximação eficiente apenas para desvios-padrões pequenos.

$$\Gamma(p+1) = \int_0^{+\infty} x^p \exp(-y)\, dy = (-y^p exp(-y)]_0^{+\infty} + \int_0^{+\infty} y^{p-1} \exp(-y)\, dy = p\, \Gamma(p)$$

O integral é uniformemente convergente relativamente a p e, consequentemente, $\Gamma(p)$ é uma função contínua. De facto:

- $\Gamma(1) = \int_0^{+\infty} 1^{p-1} \exp(-1)\, dx = (-\exp(-1)]_0^{+\infty} = 1 = 0!$;
- $\Gamma(2) = \Gamma(1+1) = 1 \times \Gamma(1) = 1 \times 0! = 1!$
- $\Gamma(3) = \Gamma(2+1) = 2 \times \Gamma(2) = 2 \times 1! = 2!$
- $\Gamma(4) = \Gamma(3+1) = 3 \times \Gamma(3) = 3 \times 2! = 3!$
- $\Gamma(5) = \Gamma(4+1) = 4 \times \Gamma(4) = 4 \times 3! = 4!$
- $(\ldots)$
- $\Gamma(n+1) = n\Gamma(n) = n \times (n-1)! = n!$

A função distribuição de uma variável aleatória $G(p, \lambda)$ é dada por:

$$G(y|p,\lambda) = \begin{cases} 0; & y \leq 0 \\ \frac{\lambda^p}{\Gamma(p)} y^{p-1} \int_0^y t^{p-1} \exp(-t)\ dt; & y > 0; p, \lambda > 0 \end{cases} \quad (16)$$

Onde $\Gamma(p) = \int_0^{+\infty} y^{p-1} \exp(-y)\, dy$

Vale a pena estudar um pouco esta distribuição. Algumas considerações:

1. Sendo Y uma variável aleatória com distribuição $G(y|p,\lambda)$ tem se:

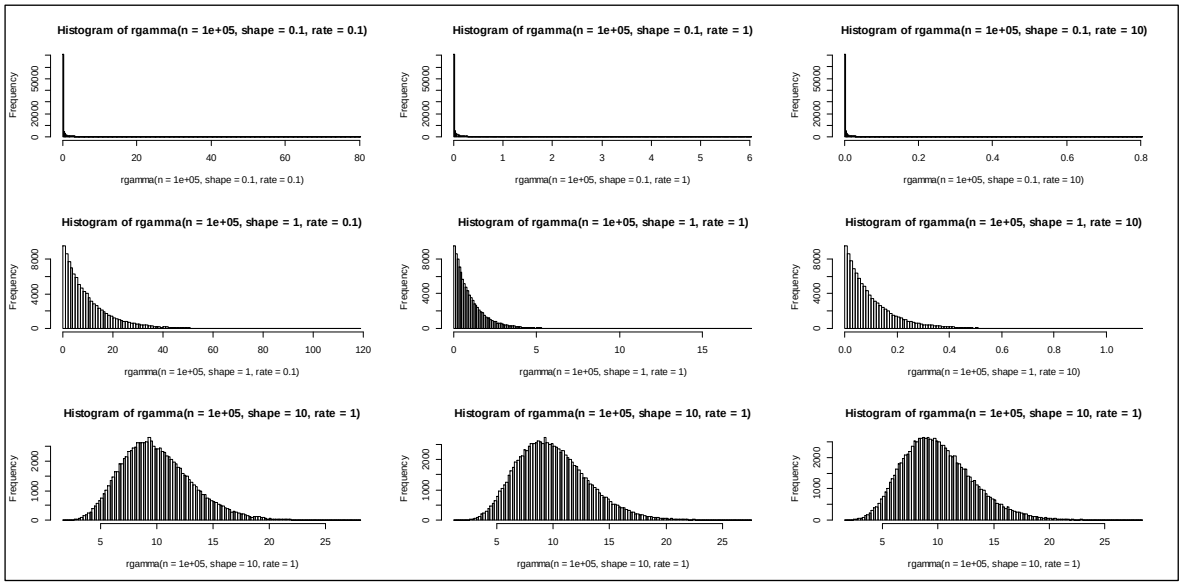
   $E(Y) = \frac{p}{\lambda}$; $V(Y) = \frac{p}{\lambda^2}$

2. p é um parâmetro de forma da função de probabilidade. Com $p > 1$ as distribuições gama são unimodais; sendo que com $p < 1$ a curva apresenta um «J» deitado, ver figura abaixo. E, com $p = 1$, esta distribuição atinge o seu máximo com $Y = 0$.
3. λ é um parâmetro de escala da função de probabilidade.

4. A distribuição gama tem um domínio não-negativo.
5. Alguns membros especiais desta família obtém-se:
    a. Com $p = 1$, trata-se de uma distribuição exponencial com parâmetro λ.
    b. Com $p = \frac{n}{2}$ e $\lambda = \frac{1}{2}$, trata-se de uma distribuição do qui-quadrado com n graus de liberdade.
    c. Se p for suficientemente grande, então $Y \sim N(\frac{p}{\lambda}; \sqrt{\frac{p}{\lambda^2}})$.

A figura abaixo retracta alguns membros da família gama. Face ao indicado até agora, é de supor que numa distribuição de custos $p > 1$. Já o coeficiente de escala, "λ", poderá assumir qualquer valor.

Figura 3.2 - Alguns membros da família gama



Fonte: autor

3.2.2. Gama: Heterogeneidade

Imagine-se que se tem uma companhia de seguros onde os seus clientes apresentam um custo médio de 1.000 unidades monetárias (u.m.), e que o

desvio-padrão associado ao custo é de 316 u.m.. Qual a probabilidade de um cliente apresentar um sinistro com um custo inferior a 800 u.m.?

Se se considerar que Y representa o custo associado a um sinistro, então, usando as equações caracterizadoras dos parâmetros da distribuição gama:

$$E(Y) = \frac{p}{\lambda} = 1.000$$

$$V(Y) = \frac{p}{\lambda^2} = 100.000$$

Tem-se:

$$P(y \leq 800 | p = 1000; \lambda = 0{,}01) = \frac{0{,}01^{1000}}{\Gamma(1000)} 800^{1000-1} \int_0^{800} t^{1000-1} \exp(-t)\, dt \approx 28\%.$$

A questão que se coloca com o modelo de impacto, é a seguinte: será que o custo médio representa completamente o modelo em causa? Ou melhor ainda, será que o custo médio é idêntico entre todos os indivíduos? Trata-se da mesma questão de heterogeneidade apresentada em secções anteriores e que motiva o conceito de regressão.

Seguindo o racional visto anteriormente, considere-se, por exemplo, que os custos médios associados aos sinistros dos homens são de $\mu+\delta$, com variância σ^2; que o custo dos sinistros associado ao das mulheres é de $\mu-\delta$, com variância σ^2; e que o processo gerador de sinistros segue uma gama. Os custos médios conjuntos serão a média de homens e mulheres: $\mu = [(\mu+\delta) + (\mu-\delta)]/2$. Mas, evidentemente, o modelo não terá um coeficiente de variação constante para cada segmento/*tariff cell*.

3.2.3. Gama: Processo de estimação

Para explicitar o modelo de estimação gama será utilizado uma redução do processo gerador de custos com sinistros ao caso clássico de estimação por mínimos quadrados (OLS – *Ordinary Least Square*), de forma a obter um estimador centrado. De forma a garantir um estimador estatisticamente eficiente, será feito um paralelismo às condições do teorema Gauss–Markov, de forma a

garantir que se está em presença de o melhor estimador linear não enviesado.[9]

O estimador dos mínimos quadrados ($\widehat{\boldsymbol{\beta}}$) é habitualmente obtido através:

$$\underset{\widehat{\boldsymbol{\beta}}}{\text{Min}}\ \mathbf{u}'\mathbf{u} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

onde a notação é a habitual: a negrito os vectores e as matrizes, $\mathbf{u}$ os resíduos, $\mathbf{y}$ a variável objectivo e $\hat{\mathbf{y}}$ o modelo estimado, $\mathbf{X}$ as variáveis explicativas e $\widehat{\boldsymbol{\beta}}$ o vector dos coeficientes a estimar.

Para resolver este problema algébrico é preciso assumir as condições 1 a 3 abaixo indicadas:

- Condição 1: Linearidade nos parâmetros: O modelo tem de ser linear nos parâmetros[10].
- Condição 2: Valores de $\boldsymbol{X}$ são fixados numa amostragem repetida. (Mais tecnicamente, $\boldsymbol{X}$ é assumido como sendo não estocástico)
- Condição 3: Erro não está correlacionado com as variáveis explicativas: $E(\mu_i|X) = 0$.[11] A esta hipótese também se designa por exogenidade condicionada.

Assim, tem-se:

$$\underset{\widehat{\boldsymbol{\beta}}'}{\text{Min}}\ \mathbf{u}'\mathbf{u} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\mathbf{y}' - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$

[9] Esta secção é baseada McCullagh & Nelderm (1989) pp 285 e seguintes, embora o detalhe aqui apresentado seja bem mais extenso do que o seguido por estes autores.

[10] Será esta condição dramática? Qualquer função pode ser aproximada por uma recta num ponto; ou por um polinómio (uma função linear nos parâmetros), sendo por isso possível calibrar as variáveis de forma a ter formas funcionais mais ajustadas. Ou seja, se o modelo não for linear, é possível ter uma forma aproximada.

[11] Outra forma de ver esta condição é a de avaliar se não se exclui do modelo qualquer variável relevante que porventura esteja relacionada com as variáveis presentes. Aliás, quanto maior for a relação de X excluído com o erro, maior o enviesamento no estimador de $\widehat{\boldsymbol{\beta}}$'.

Nestas condições, tem-se as habituais condições de primeira ordem:

$$\frac{\partial .}{\partial \widehat{\boldsymbol{\beta}}'} = 0 - \mathbf{2}\boldsymbol{X}'\boldsymbol{y} + \mathbf{2}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0$$

Para resolver este sistema de equações é necessária aplicar mais uma condição:

- Condição 4: Ausência de perfeita multicolinearidade, isto é que a matriz $\boldsymbol{X}'\boldsymbol{X}$ é invertível[12].

Posto isto, e para o problema em concreto, tem-se que:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \tag{17}$$

Este resultado advêm da minimização dos resíduos e apenas exige que a $(\mathbf{X}'\mathbf{X})$ tenha inversa (ausência de multicolinearidade), e que o modelo seja linear nos parâmetros. Não se obriga a nenhuma condição sobre os erros ou resíduos. A única hipótese sobre o comportamento dos resíduos (exogenidade $\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$) é apenas necessária para garantir a centragem. Repare-se que:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + \mathbf{u} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{u}
\end{aligned}$$

Aplicando o operador valor esperado, tem-se:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{E}(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{u}|\mathbf{X})$$

[12] Ou seja, $(\mathrm{X}'\mathrm{X})$ não se pode ter informação redundante.

$$= \boldsymbol{\beta} + \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}|\mathbf{X})$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{E}(\mathbf{u}|\mathbf{X}) \quad (18)$$

Sob a hipótese clássica de exogenidade condicionada, tem-se que $\mathrm{E}(\mathrm{u}|\mathrm{X}) = 0$:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$$

Ou seja, o estimador é centrado em presença de heterocedasticidade ou de homocedasticidade.

A importância da dispersão dos resíduos advém do facto de que se os resíduos apresentarem um comportamento homocedástico é possível concluir sobre a sua eficiência e, assumindo a normalidade é possível fazer inferência sobre os parâmetros.

Mais do que isso, um modelo não homocedástico não é estatisticamente eficiente. Interessa, para o objecto da tese, demonstrar a eficiência do estimador dos mínimos quadrados.

Eficiência do estimador dos mínimos quadrados

Com base na equação 17 sabe-se que:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

E que:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}.$$

Imagine-se agora que há outro estimador $\widehat{\boldsymbol{\beta}}^{*}$ linear nos parâmetros que

pode ser definido com a inclusão de mais variáveis, pelo que o estimador $\widehat{\boldsymbol{\beta}}^*$ pode ser escrito da seguinte forma:

$$\widehat{\boldsymbol{\beta}}^* = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\,\mathbf{y}.$$

Então, sabendo que a definição do modelo de $\mathbf{y} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{u})$, tem-se:

$$\widehat{\boldsymbol{\beta}}^* = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\,(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}).$$

O que simplificando, tem-se:

$$\widehat{\boldsymbol{\beta}}^* = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{C}\mathbf{u})$$

$$\widehat{\boldsymbol{\beta}}^* = \boldsymbol{\beta} + \mathbf{C}\mathbf{X}\boldsymbol{\beta} + \ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{C}\mathbf{u}$$

Ora, se $\widehat{\boldsymbol{\beta}}^*$ é não enviesado, tem-se que $\mathbf{C}\mathbf{X} = \mathbf{0}$, o que significa:

$$\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta} = \ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{C}\mathbf{u}.$$

Aplicando a definição da matriz variância e covariância ao estimador $\widehat{\boldsymbol{\beta}}^*$ tem-se:

$$\text{Var}\,(\widehat{\boldsymbol{\beta}}^*) = \mathbf{E}\big(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\big)\big(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\big)' = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{C}\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} + \mathbf{C}\mathbf{u}]'$$

$$= \mathbf{E}\big[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1'} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{C}' + \mathbf{C}\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1'} + \mathbf{C}\mathbf{u}\mathbf{u}'\mathbf{C}'\big] \quad (19)$$

Usando o conceito de homocedasticidade ($\mathrm{E}(\mathbf{u}\mathbf{u}') = \boldsymbol{\sigma}^2\mathbf{I}$) tem-se:

$$\text{Var}\,(\widehat{\boldsymbol{\beta}}^{*}) = \boldsymbol{\sigma^2}(\mathbf{X'X})^{\mathbf{-1}} + \boldsymbol{\sigma^2}\mathbf{CC'} \tag{20}$$

Esta equação permite concluir que a matriz variância e covariância de um qualquer outro estimador não enviesado é idêntica a matriz de variância e covariância do estimador OLS ($\boldsymbol{\sigma^2}(\mathbf{X'X})^{\mathbf{-1}}$), acrescido da parcela: $\boldsymbol{\sigma^2}\mathbf{CC'}$ que é semi-definida positiva (por construção). Assim, $\widehat{\boldsymbol{\beta}}$ tem a menor variância na classe dos estimadores lineares, já que $\text{Var}(\widehat{\boldsymbol{\beta}}^{*})$ é sempre maior que $\text{Var}(\widehat{\boldsymbol{\beta}})$**,** a não ser que $\mathbf{C} = \mathbf{0}$, o que tornará $\widehat{\boldsymbol{\beta}}^{*} = \widehat{\boldsymbol{\beta}}$.

Assim, é possível concluir o estimador linear nos parâmetros é estatisticamente eficiente, se e só se tiver em presença de homocedasticidade.

Vale a pena recordar que para esta demonstração foi necessário apenas garantir a homocedasticidade, além das condições dos mínimos quadrados – não tendo sido necessário impor qualquer condição em relação à normalidade dos resíduos.

No caso em apreço, *i.e.* quando o processo gerador de custos de sinistros segue uma distribuição gama, os erros apresentam um padrão de dispersão não constante (quadrático) – não se está em presença de homocedasticidade. Assim, o modelo de regressão clássico aplicado a uma estimativa de custos será centrado, mas não será estatisticamente eficiente. De uma forma mais matematizada e continuando o racional visto à pouco, a variância do estimador $\widehat{\boldsymbol{\beta}}^{*}$ centrado, mas agora não homocedástico, será:

$$\begin{aligned}
&\text{Var}\left(\widehat{\boldsymbol{\beta}}^{*}\right) = \text{E}\left(\widehat{\boldsymbol{\beta}}^{*} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}^{*} - \boldsymbol{\beta}\right)' \\
&= \text{E}[(\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'u}][(\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'u}]' \\
&= \text{E}[(\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'uu'X}(\mathbf{X'X})^{\mathbf{-1}}] \\
&= [(\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'}\,\mathbf{W}\,\mathbf{X}(\mathbf{X'X})^{\mathbf{-1}}]
\end{aligned} \tag{21}$$

$$\text{onde } \mathbf{W} = \mathrm{E}\,(\mathbf{u}\mathbf{u}').$$

No caso em concreto, num processo gerador de dados assente numa distribuição gama, a hipótese de homocedasticidade dificilmente se mantem: o coeficiente de variação é constante, logo a variância das estimativas cresce com a média. Logo, não se está nas condições do teorema do limite central, logo não se tem um estimador de variância mínima.

Intuitivamente, é fácil de compreender a razão desta não eficiência: se houver um padrão evidente para a heterocedasticidade ($\mathbf{W}$), o modelo OLS não a usa. De facto o modelo OLS trata cada unidade de análise de igual forma, não incorporando qualquer padrão de heterocedasticidade. Sendo ainda mais concreto, e relembrando mais uma vez que se está com um padrão de dispersão gama: se se souber que a heterocedasticidade aumenta com X – isto é, se há mais incerteza à medida X aumenta – deveríamos ser capazes de incluir essa informação no modelo e tratar cada unidade de análise de forma diferente e ter um desvio padrão associado à estimativa de $\boldsymbol{\beta}$.

No processo gerador de custos de sinistros o padrão de dispersão é conhecido (o coeficiente de variação é constante), pelo que é possível forçar o modelo de estimação OLS a reconhecer o padrão de heterocedasticidade. A título ilustrativo, note-se o caso com apenas duas varáveis (sendo que a primeira, como habitualmente, é uma constante), onde se sabe o padrão de heterocedasticidade (coeficiente de variação constante):

$$\mathrm{y_i} = \beta_1 + \beta_2 \mathrm{x_{2i}} + \mathrm{u_i}$$

Como se sabe qual o padrão de variabilidade dos dados, é possível fazer:

$$\frac{\mathrm{y_i}}{\sigma_i^2} = \beta_1 \frac{1}{\sigma_i^2} + \beta_2 \frac{\mathrm{x_{2i}}}{\sigma_i^2} + \frac{\mathrm{u_i}}{\sigma_i^2},$$

desta forma, a variância associada ao resíduo $\left(\frac{u_i}{\sigma_i^2}\right)$ será constante.

Quando o padrão de σ_i^2 é conhecido, ou facilmente identificado/estimado, é possível garantir a variância mínima dos estimadores. A solução mais habitual é (como visto anteriormente) que numa distribuição de custos o coeficiente de variação seja constante, *i.e.*, a variância é proporcional à média ao quadrado, que: $E(u_i^2) = \sigma^2[E(y_i)]^2$, o que faz com que o modelo a estimar seja:

$$\frac{y_i}{E(y_i)} = \beta_1 \frac{1}{E(y_i)} + \beta_2 \frac{x_{2i}}{E(y_i)} + \frac{u_i}{E(y_i)}.$$

Contudo, $E(y_i)$ não é conhecido pelo que este ajustamento é inoperacional. Mas se nos recordarmos que a estimativa $\hat{y}_i$ por OLS é centrada para y_i,tem-se:

$$\frac{y_i}{\hat{y}_i} = \beta_1 \frac{1}{\hat{y}_i} + \beta_2 \frac{x_{2i}}{\hat{y}_i} + \frac{u_i}{\hat{y}_i}.$$

Assim, foi possível reduzir o modelo com heterocedasticidade a um modelo clássico de regressão que irá produzir estimativas centradas e eficientes. Assim, no caso de heterocedasticidade há que estimar o modelo em duas etapas, ponderado as observações pelo padrão de heterocedasticidade – tal procedimento é habitualmente designado por Mínimos Quadrados Ponderados (WLS – *Weight Least Squares*).

Contudo, este método de estimação gera um efeito de escala que é evidente no coeficiente associado à ordenada na origem. Este coeficiente terá um enviesamento aproximado de $\sigma^2/2$.

Assim, tendo em conta que se pretende manter o efeito de escala e não transformar a resposta prevista, é habitual antes estimar o modelo com a seguinte forma:

$$\frac{y_i}{\exp(\hat{y}_i)} = \beta_1 \frac{1}{\exp(\hat{y}_i)} + \beta_2 \frac{x_{2i}}{\exp(\hat{y}_i)} + \frac{u_i}{\exp(\hat{y}_i)}.$$

Ou seja, assume-se que há uma função *link* entre o valor esperado e as variáveis explicativas:

$$E(\boldsymbol{Y}) = \exp(\mathbf{X}'\boldsymbol{\beta}).$$

Esta transformação mantem a escala e a relação quadrática entre a variância e o valor esperado. De um ponto de vista prático o modelo pode ser estimado com recurso às condições de mínimos quadrados vistas acima, percebendo que as variáveis explicativas padecem de uma simples mudança de variável. A este processo designa-se como regressão gama com uma função *link* exponencial. Como indicado no início desta secção, o processo de estimação de uma regressão gama pode ser explicado através de uma generalização do modelo de mínimos quadrados clássico.

Note-se que a modelação dos custos de sinistros como um modelo de correcção de heterocedasticidade associado ao modelo de Mínimos Quadrados permite formulações mais vastas e, caso se revele apropriado, assumir outros padrões 'fechados' entre a média e a variância. Harrison & McCabe (1979) indicam um conjunto de testes para avaliar o padrão de heterocedasticidade que poderão complementar o desenho de expansão dos resíduos.

Aliás, e assumindo que a regressão gama é uma solução para um problema de heterocedasticidade em que há um relação entre a média e a variância, uma outra alternativa a estudar seria aplicar um estimador linear robusto, que levasse em conta um qualquer padrão de dispersão dos resíduos, e não apenas o de coeficiente de variação constante. Hayashi (2001) ou Wooldridge (2003) oferecem uma visão genérica sobre o problema e respectiva solução.

3.2.4. Gama: *Offset* e ponderação

Se houver poucos sinistros em cada umas das *tariff cells*, o modelo de custos médios estimará, evidentemente, também pouca despesa. Se houver mais casos observados, haverá também uma maior despesa. Assim, independente dos factores tarifários que compõem a *tariff cell*, a despesa estimada dependerá do

número de sinistros presentes.

Tal como visto no caso de Poisson há duas soluções habituais para este tipo de problema:

1. Ponderar os dados;
2. Incluir um *offset*.

Sendo que o *offset* a solução mais usada com recurso ao número de sinistros.

3.2.5. Gama: Interpretação

A interpretação dos coeficientes de uma regressão gama, ou de uma regressão robusta, é idêntica a uma regressão clássica e sai pela simplicidade fora do âmbito desta tese. Para uma apreciação inicial dos métodos de interpretação consultar Griffiths, Hill, & Judge (1993), Gujarati (1995) ou Wooldridge (2003) e confrontar com a secção “3.1.5. Poisson: Interpretação”.

3.3. Regressão Tweedie

Muitas vezes a única alternativa viável para compreender os custos técnicos, como veremos adiante na secção 4.3, é ter uma modelização combinada da frequência e dos custos médios. Assim, de um ponto de vista prático, há que garantir que se conhece uma distribuição que combine a frequência e os custos médios.

Aliás, esta poderia ser o ponto de partida de um processo de tarifação. Porém nem sempre é fácil. Meyers (2009) explica porquê: os modelos tradicionais de regressão não lidam bem com a mistura de perdas nulas (discretas) com perdas contínuas positivas e não simétricas.

A distribuição Tweedie permite combinar os modelos de frequência de Poisson e de severidade de gama (ver Anderson, Modlin, Schirmacher, Schirmacher, Thandi (2007), Maccullagh, Nelder & Ashworth (1989) Meyers (2009) e Ohlsson, Johansson (2010)), dando origem a um modelo de dispersão exponencial

(*Exponential Dispersion Model* – EDM).

3.3.1. Tweedie: Distribuição

Nesta secção será estudado, em primeiro lugar, o modelo de dispersão exponencial em sentido lato, sendo que apenas no final será feita uma concretização para a distribuição Tweedie e respectivo modelo de regressão.

Um modelo de dispersão exponencial (EDM) é importante em si mesmo, já que traduz uma generalização dos modelos lineares (GLM) acima estudados: normal, Poisson, gama etc. De facto, o que se pretende com um EDM (Dunn & Smyth (2005) e Jørgensen (1989)) é obter uma função de probabilidade baseada numa formulação da variância baseada numa relação média-variância.

Assim tem-se que um EDM pode ser caracterizado pela sua função de variância:

$$V\,(Y) = \phi V(\mu).$$

Sendo que os casos especiais de

$$V\,(Y) = \phi\mu^{p}, \text{ para } p \text{ distintos} \qquad (22)$$

assumem uma classe muito importante.

Recorde-se que o que se procura é ter uma função distribuição de probabilidade que apresente uma relação EDM. Jørgensen (1992) chama a esta classe modelos de Poisson compostos ou de Tweedie, salientando o trabalho pioneiro de Tweedie produzido na década de 1980. Esta classe de modelos de dispersão assume formas muito conhecidas[13]:

[13] Sendo que para uma questão de simplicidade se vai assumir que $\phi = 1$.

- Com $p = 0$, tem-se um processo gerador de dados normal;
- Com $p = 1$, tem-se um processo gerador de dados de Poisson;
- Com $p = 2$, tem-se um processo gerador de dados de gama; e
- Com $p = 3$, tem-se um processo gerador de dados de Gaussiano inverso.

Para outros valores de p é possível ter outras distribuições:

- Com $p > 2$, tem-se um processo gerador de dados de positivo e estável; e
- Com $p = \infty$, tem-se um processo gerador de dados extremamente estável.

Para $0 < p < 1$ não há nenhum modelo Tweedie.

Há, no entanto, uma forte desvantagem num modelo de dispersão exponencial: aparte dos modelos mais conhecidos, com $p = 1, 2 \; e \; 3$, a distribuição Tweedie não tem uma função densidade que possa ser escrita de uma forma fechada.

Com base na variação dos parâmetros, conclui-se que a distribuição Tweedie tem a capacidade de explicitar a distribuição de perdas acumuladas (*aggregate loss distribution*): já que tem um ponto de massa em zero que é acompanhado por uma distribuição positiva e contínua, assumindo formas muito diversas conforme os seus parâmetros.

Vale a pena estudar os casos em que $1 < \mathrm{p} < 2$ e compreender a sua função de distribuição. Uma vez que não se trata de uma distribuição muito conhecida e muito complexa vale a pena confrontar Jørgensen (1989) ou Jørgensen e de Souza (1994). Em qualquer caso, a função de distribuição para $1 < \mathrm{p} < 2$ é definida como:

$$P(y, Y > 0|\theta, \lambda, \alpha) = \sum_{n=1}^{\infty} \frac{\left\{(\lambda\omega)^{1-\alpha} \; k_\alpha\left(\frac{-1}{y}\right)\right\}^n}{\Gamma(-n\alpha)n!y} . exp\{\lambda\omega[\theta_0 y \; - \; k_\alpha(\theta_0)]\}, \quad com \; y > 0 \tag{23}$$

e $P(Y=0) = \exp\{-\lambda\omega \ k_\alpha(\theta_0)\}$

Onde

$$k_\alpha(\theta) = \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$$

$$\theta_0 = \theta\ \lambda^{1/1-\alpha}$$

ω é o peso correspondente à exposição (de custos) em causa.

Pode ser demonstrado que a variância da distribuição acima é dada por:

$$V(Y) = \frac{1}{\lambda}\mu^p, \qquad (24)$$

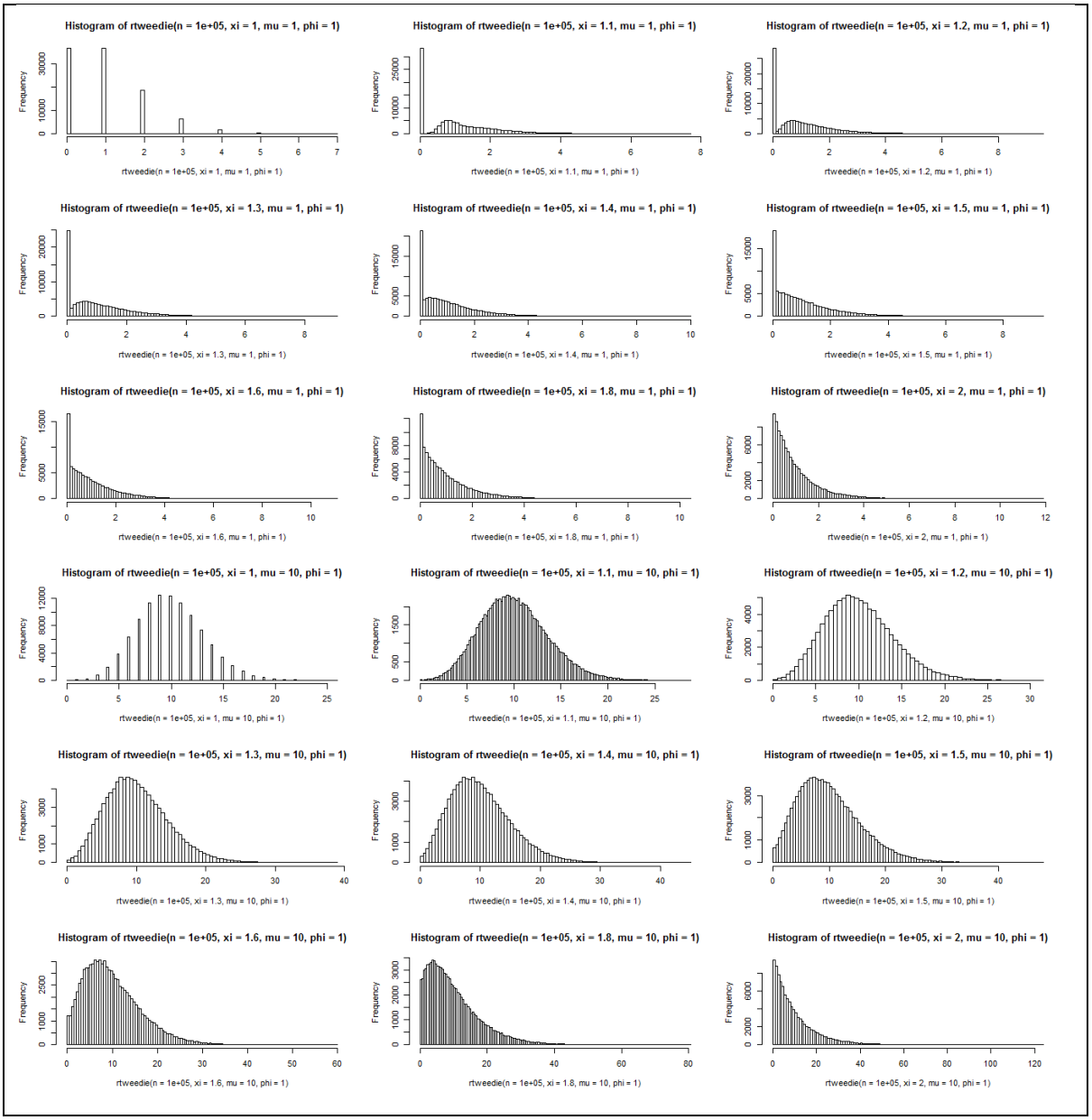
ou

$V(Y) = \phi\mu^p$, com $\phi = \frac{1}{\lambda}$ em linha com a formulação dos modelos da família exponencial.

A figura abaixo retracta alguns membros da família Tweedie. Face ao indicado até agora, é de supor que numa distribuição de custos técnicos $1 < p < 2$ já o coeficiente de escala, ϕ, poderá assumir qualquer valor. Note-se que:

- A distribuição para *scales* elevados, a distribuição parece acompanhar uma distribuição log-normal ou gama.
- Para frequências de sinistro elevadas a distribuição parece acompanhar uma distribuição normal.

Figura 3.3 - Alguns membros da família Tweedie



Fonte: autor

3.3.2. Tweedie: Heterogeneidade

A ideia de heterogeneidade é semelhante aos modelos gama e Poisson acima descritos: Será que em cada classe de risco ou *tariff cell* têm o mesmo mecanismo gerador de dados? Esta ideia faz com que não seja necessário apenas compreender a distribuição, mas compreender as variáveis que podem gerar diferentes comportamentos. Tal como anteriormente, não fosse a regressão

de Poisson e gama um caso especial da Tweedie, a heterogeneidade dos dados entra na distribuição Tweedie através da função *link* designada, habitualmente, *canonical link*. O que esta função tem de assumir é a seguinte:

$$\boldsymbol{\theta} = (b')^{-1}(g^{-1}(\boldsymbol{\eta})) = \boldsymbol{\eta},$$

ou seja, implica que o inverso da função *link*, $\mathrm{g}^{-1}(.)$, é o inverso de b'. Se, no caso da regressão de Poisson e de gama, há um racional (já discutido) para a escolha do *link canónico*, na regressão Tweedie não há qualquer imperativo para a escolha de uma função (Anderson, Feldblum, Modlin, Schirmacher, Schirmacher, & Thandi (2004)). No entanto, e ainda, de acordo com os mesmos autores, para os diferentes tipos de distribuição são usados os links canónicos como apresentado abaixo.

Tabela 3.2 - Tweedie – links canónicos

Distribuição	Link Canónico
Normal	μ
Poisson	$log\ (\mu)$
Gama	$1/\mu$
Binomial	$log\ (\mu/(1-\mu))$
Gausssiano Inverso	$1/\mu^2$

Fonte: adaptado de Anderson, Feldblum, Modlin, Schirmacher, Schirmacher, & Thandi (2004)

3.3.3. Tweedie: Processo de estimação

Por forma a estimar este modelo, é necessário construir a distribuição log-verosimilhança. Uma vez que não se trata de uma tarefa imediata, a construção "passo a passo" pode ser vista de perto em Peters; Shevchenko & Wüthrich (2009) ou Anderson, Feldblum, Modlin, Schirmacher, Schirmacher, & Thandi

(2004).

No caso de distribuições da família exponencial, como é o caso, a função log de verosimilhança pode assumir a seguinte forma simplificada:

$$l = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \qquad (25)$$

Neste modelo de máxima verosimilhança a calculatória é difícil e a precisão muito discutível. Dunn & Smyth (2005) explicam porquê: as famílias Tweedie são aqueles EDM onde há uma relação entre a média e a variância. A distribuição normal, Poisson, gama e inversa guassiana são distribuições da família Tweedie. Aparte destes casos especiais as distribuições Tweedie não têm uma função densidade que possa ser escrita da forma fechada. Podem ser no entanto escritas como um conjunto de somas infinitas. Sendo ainda mais directo afirmam estes autores, para $p \neq 0, 1, 2, 3$ "*there are generally no closed forms for the Tweedie densities, full likelihood analysis is very difficult.*"

Veja-se essa afirmação na prática. As condições de primeira ordem da função de verosimilhança acima descrita são:

$$\frac{\partial l}{\partial \beta_j} = 0 \text{ com j=1,…m}$$

Usando a função canónica, descrita acima, que estabelece uma relação entre $\boldsymbol{\beta}$ e $\boldsymbol{\theta}_i$ e aplicando a regra da derivação em cadeia três vezes (é mais fácil do que derivar directamente na função objectivo), tem-se:

$$0 = \frac{\partial l}{\partial \beta_j} = \sum_i \frac{\partial l}{\partial \theta_i}(\ldots).\frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \eta_i}.\frac{\partial \eta_i}{\partial \beta_i}$$

Sabendo que:

$$\mu_i = b'(\theta_i) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} ,$$

$$\eta_i = g(\mu_i) \Rightarrow \frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)} ,$$

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_m x_{im} \Rightarrow \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \ ,$$

tem-se:

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)}{a(\phi)} . \frac{1}{b''(\theta_i)} . \frac{1}{g'(\mu_i)} . x_{ij}$$

$$= \sum_i \frac{\omega_i (y_i - \mu_i) x_{ij}}{\phi V(\mu_i) g'(\mu_i)} \ .$$

Apesar da resposta teórica ser fácil de obter, a resposta prática é mais difícil de obter. E é fácil de compreender porquê:

- Temos três parâmetros a estimar: $\phi; \boldsymbol{\theta}_i$ e… p. As condições de primeira ordem apenas permitem apurar $\boldsymbol{\theta}_i = \boldsymbol{\eta}_i$.
- É preciso definir a função canónica, de forma a conhecer: g'. Como visto acima, há algumas funções habituais que se usam conforme a distribuição conhecida.
- $V(\mu_i)$ depende de p; logo a determinação de p é fundamental. Jørgensen & De Souza (1994) resolvem esta questão apurando o custo de sinistro por unidade segurada (μ_i), colocando o problema no contexto dos modelos lineares generalizados. Smyth & Jørgensen (2002) constatam que é necessário também estimar a dispersão, além da média. Para tal usam um

método de "*double generalized linear models*".[14]

- Porém, de uma maneira geral, na prática, assume-se que p é conhecido e fixado pelo analista com base numa boa teoria. Uma outra alternativa, testada pelo autor, poderá ser o uso de um método de tentativa e erro de forma a descobrir o melhor p. Para tal, será necessário definir uma função objectivo (função verosimilhança), sendo que o ambiente de regressões lineares convida ao uso do mínimo da soma dos quadrados dos resíduos.

Resultados 3.1 - Tweedie – Programação em R para achar o p associado à função Tweedie

```
n= 8000 #podemos por n até m-1; mas por vezes o modelo não converge quando se está tão próximo
do limite
        #pelo que por vezes é difícil obter o mínimo
m=10000
yy=rep(NA, n)

for (i in 1:n) {
  xx=1+i/m

link, which is 1-var.power.
  GLM.aux <- glm(y ~ x1 +x2+ x3 + x4, family=tweedie (var.power=xx,link.power=0), data=ADefinir
)
  aux<- (exp(predict(GLM.aux)-Adefinir$y)^2)
  yy[i]<-sum(aux,na.rm = any(!is.na(x)))
}

cat("Descoberta da posição do p.")
which (yy ==min(yy))

cat("Definição do p =")
p=1+which (yy ==min(yy))/m
p
plot(yy)# para ver se a função é bem comportada
```

3.3.4. Tweedie: *Offset* e ponderação

Se houver poucos sinistros em cada umas das *tariff cells*, o modelo de custos totais estimará também pouca despesa. Se houver mais casos observados, haverá também uma maior despesa. Assim, independente dos factores tarifários

[14] De uma forma intuitiva, o que aqui se afirma é que o padrão de heterogeneidade é necessário para obter estimativas mais precisas, em linha com o que foi afirmado no contexto da regressão Gama com correcção da heterogeneidade (regressão robusta).

que compõem a *tariff cell*, a despesa estimada dependerá do número de casos (apólises e sinistros) presentes.

Tal como visto no caso de Poisson e gama há duas soluções habituais para este tipo de problema:

1. Ponderar os dados;
2. Incluir um *offset*.

Sendo que o *offset* a solução mais usada com recurso aos sinistros.

3.3.5. Tweedie: Interpretação

A interpretação dos coeficientes de uma regressão Tweedie não é complexa já que acompanha a de um a modelo linear clássico. Em qualquer caso, como a função não é fechada, o ideal será analisar o impacto das respostas finais por *tariff cell* e desconstruir o efeito canónico para melhor compreender a força de cada variável.

4. Apuramento dos prémios das seguradoras congéneres

4.1. Introdução

O objectivo deste capítulo é o de indicar uma abordagem que permita compreender e prever o modelo tarifário das companhias congéneres. De forma muito concreta, o resultado final deste capítulo será a construção de conjunto de equações (o processo gerador de tarifas) que permita determinar o preço apresentado por cada seguradora congénere em cada uma das *tariff cells*.

Ao contrário do que se passa com outros bens de consumo, os preços (mais formalmente, prémios) de seguros não são fáceis de apurar já que, como visto anteriormente, os valores a cobrar dependem do segmento de risco (*tariff cell*). Mais, ao contrário de outros produtos de retalho, não há claramente um marca de 'primeiro preço', e muito menos em todos os segmentos de risco o diferencial de preço entre as seguradoras se mantem: isto é, uma seguradora não é sempre x% mais barata face à empresa líder.

Habitualmente, e como visto nos capítulos anteriores, as seguradoras contam com seu histórico para estimar o prémio: analisam para o comportamento dos seus clientes e propõem um prémio que é idêntico às responsabilidades assumidas (ao qual acrescenta uma parcela para pagar os custos administrativos, de distribuição e de remuneração accionista).

Assim, quando o objectivo é estimar o prémio das seguradoras congéneres em cada uma das *tariff cells*, há que obter um método exequível e eficiente (em termos de trabalho de campo) para apurar os prémios das diferentes companhias de forma a integrá-los no modelo global de optimização de preço – objectivo final desta tese.

Neste desígnio, as palavras-chave são 'exequível' e 'eficiente'. Em teoria, uma seguradora poderia recolher todas as cotações possíveis de todas as seguradoras no mercado e assim compreender, sem erro amostral, o modelo de preços existente no mercado. Porém esta recolha total seria extremamente onerosa e por vezes não exequível.

A abordagem metodológica para responder a este desafio de determinar qual o

processo gerador de tarifas de cada uma das seguradoras congéneres segue a abordagem associada a um desenho experimental clássico:

- Fase 1: Identificar os factores e níveis que definem o preço em cada uma das *tariff cells*;
- Fase 2: Desenho óptimo;
- Fase 3: Recolha de informação;
- Fase 4: Análise.

Neste capítulo será indicado:

- o que se entende por desenho óptimo (a Fase 2 corresponde à secção *2,* deste capítulo);
- qual o mecanismo de análise (Fase 4 corresponde às secções *4* e *5,* deste capítulo);
- quais as variáveis, factores e níveis (ver secções 4.3 e 4.4) relevantes para o objectivo geral do estudo (Fase 1).

É de notar que a Fase 3 do processo clássico de desenho experimental não terá grande destaque, já que sai fora do âmbito desta tese. Em qualquer caso, vale a pena mencionar que a recolha de informação pode ser feita com recurso a Mediadores com acesso aos sistemas tarifários das companhias ou, em casos extremos, a Cliente Mistério.

4.2. Fundamentos do Delineamento Experimental e a generalidade das suas técnicas

Uma experiência é um teste (aliás é um conjunto de testes) onde variações são impostas às variáveis de *input*, de um processo ou sistema, de forma a que se possa observar e identificar as razões da mudança no *output*. Nesta secção 2 serão apresentados os fundamentos do delineamento experimental de forma genérica. Na secção *2* será dado o detalhe ao desenho experimental no contexto

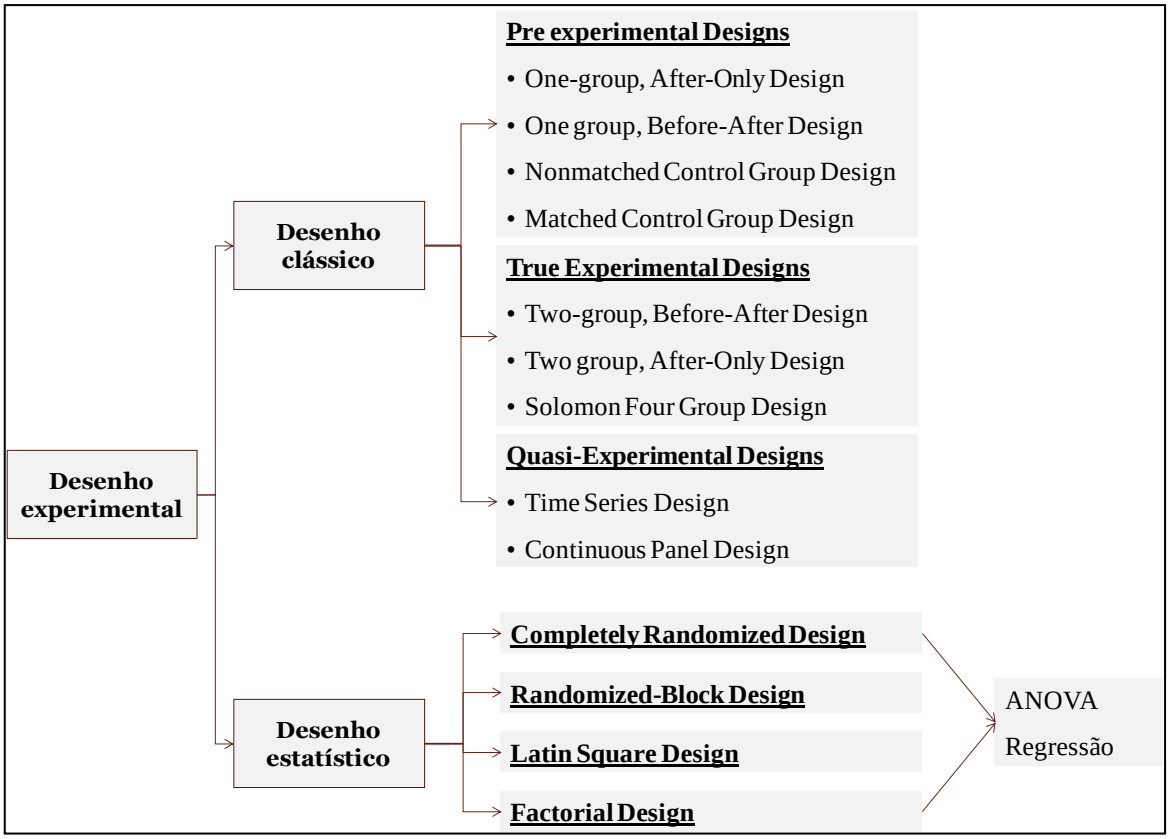
de modelo gerador de dados Tweedie. Na secção *2* será apresentada uma solução pragmática para o desenho experimental (transformação Box-Cox com *Seemingly Unrelated Regression*).

Há diferentes tipos de desenho experimental:

- Desenho clássico: considera apenas uma mudança de estado na variável de tratamento.
- Desenho estatístico: permite o exame de diferentes variáveis de tratamento e da sua alteração dos níveis no tratamento numa dada variável objectivo.

O esquema abaixo apresenta os diferentes tipos de desenho experimental.

Figura 4.1 - Tipos de desenho experimental



Fonte: adaptado de Aaker; Kumar; Day (1998).

Desenho clássico: *Pre experimental Designs*

O desenho pré-experimental é um estudo exploratório onde não há um controlo forte de todos os factores que influenciam os resultados. Tem a vantagem de estabelecer desenhos experimentais simples e explorar a intuição.

Desenho clássico: *Quasi-Experimental Designs*

Estas técnicas fornecem algum controlo ao analista, mas ainda de forma limitada. Tipicamente dão ao analista mais informação e mais métricas que o desenho pré-experimental.

Desenho clássico: *True Experimental Designs*

A maior parte dos problemas do desenho experimental podem ser controlados ao adoptar um procedimento amostral aleatório, onde todos os indivíduos têm a mesma probabilidade de ser escolhidos. A vantagem deste método é 'espalhar' as características que podem influenciar os resultados pelos diferentes indivíduos de forma a minimizar os efeitos exógenos.

Este tipo de desenhos tem duas vantagens adicionais:

- Habitualmente permitem a existência de grupo(s) de controlo;
- A colocação de unidades amostrais de forma aleatória no(s) grupo(s) experimental(ais) e de controlo.

Porém, este tipo de desenhos não é uma panaceia: é apenas um procedimento de minimizar os efeitos adversos no início de uma experiência, sabendo que, quanto maior a amostra, maior a capacidade de estabelecer uma teia de causa consequência. Além disso, há também algumas desvantagens evidentes: não é prático garantir a colocação aleatória de unidades nos diferentes grupos.

Desenho clássico: *Quasi experimental design*

Trata-se de analisar para a evolução de uma determinada variável objectivo ao longo do tempo e ver quais os factores temporais (ou não) que melhor justifiquem o seu resultado.

Desenho estatístico:

O desenho estatístico distingue-se do desenho experimental já que permite avaliar o efeito de variáveis e de diferentes níveis. De uma maneira geral, os desenhos estatísticos são (usando a designação das secções anteriores) "*after-only*" que requerem uma análise computacional um pouco mais complexa.

Tal como visto na figura anterior há vários tipos de desenho estatístico: *Completely Randomized Design*, *Randomized-Block Design*, *Latin Square Design* e *Factorial Design*.

Todos estes tipos de desenho são apresentados de seguida com algum detalhe.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
One-group, After-Only Design *(Pre experimental Designs)* **Desenho clássico**	A forma mais simples de experimentação é simples: aplicar o tratamento experimental a um grupo e medir os resultados. De forma infográfica: **Grupo Ex: X->O** (“O” Significa *outcome*)	• Fácil de aplicar • Cria intuição (leia-se hipóteses)	• Os resultados podem estar influenciados a outros factores (normalmente a questão da história e da maturação). • Erro de selecção. • Não permite o teste de hipóteses. • Não permite a antecipação sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos. • Efeito do instrumento: A forma de colocação do instrumento pode condicionar a forma como se observa os dados.
Nonmatched Control Group *(Pre experimental Designs)* **Desenho clássico**	Uma forma de controlar a história e a maturação é introduzir o grupo de controlo. Muitas vezes esse grupo de controlo é separado com uma linha temporal.	• Fácil de aplicar. • Permite controlar a história e a maturação. • Barato (são registos históricos, habitualmente)	• Erro de selecção (sobretudo selecção adversa ou auto selecção no G. Experimental). • Dificuldade de encontrar um grupo de controlo apropriado.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
	De forma infográfica: **G. Experimental: X ->O_1** **G. Controlo: ->O_2**		• Não permite o teste de hipóteses. • Não permite a antecipação sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos. • Efeito do instrumento: A forma de colocação do instrumento pode condicionar a forma como se observa os dados.
Matched Control Group *(Pre experimental Designs)* **Desenho clássico**	Uma forma de controlar o erro de selecção é fazer um "casamento" entre grupos de controlo e experimental para uma (ou várias) chave. De forma infográfica: **G. Experimental: M->X ->O_1** **G. Controlo: M->X ->O_2**	• Fácil de aplicar. • Permite controlar a história, maturação e o erro de selecção. • Barato (muito adequado quando há pouco dinheiro e impossibilidade de fazer grandes amostras).	• Dificuldade de encontrar um grupo de controlo apropriado (cf. amostragem por quotas). • Permeável ao julgamento do analista na constituição das variáveis-chave e de controlo. • Não permite o teste de hipóteses. • Não permite a antecipação

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
	("M" quer dizer que os grupos são escolhidos de forma a fazer um casamento (*match*) com as variáveis-chave.)		sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos.
One Group, Before-After Design *(Pre experimental Designs)* **Desenho clássico**	A forma mais simples de experimentação é usar o próprio grupo de tratamento como grupo de controlo. Ou seja, avaliar antes quais os comportamentos e avaliar depois quais os comportamentos justificados. De forma infográfica: **G Experimental: O_1 ->X->O2**	• Relativamente fácil de aplicar. • Permite controlar a história, maturação.	• Possibilidade de colocar os respondentes sob alerta. • Mortalidade das unidades de análise. • Efeito do instrumento: A forma de colocação do instrumento pode condicionar a forma como se observa os dados. • Dificuldade de encontrar um grupo de controlo apropriado (cf. amostragem por quotas). • Permeável ao julgamento do analista na constituição das variáveis-chave e de controlo. • Não permite o teste de hipóteses. • Não permite a antecipação

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
			sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos.
Two group, After-Only Design *(True Experimental Designs)* **Desenho clássico**	Uma forma de controlar o erro de selecção é fazer um "casamento" entre grupos de controlo e experimental para uma (ou várias) chave. Adicionalmente: Os membros do grupo experimental e de controlo são escolhidos em (i) largo número e (ii) de forma aleatória. Com estes casos consegue-se controlar mais efeitos do grupo de controlo. De forma infográfica: **G. Experimental: R->X ->O_1** **G. Controlo: R->X ->O_2** ("R" quer dizer que as unidades de	• Não há medidas de pré-teste para apurar as variáveis chave. • Permite conclusões sólidas.	• Tamanho amostral largo. • Não permite o teste de hipóteses. • Não permite a antecipação sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
	teste são aleatoriamente (*random*) escolhidas.)		
Two-group, Before-After Design *(True Experimental Designs)* **Desenho clássico**	Escolher os elementos de dois grupos aleatoriamente, sendo que ao primeiro é controlado antes e depois do tratamento. O segundo é também alvo de duas avaliações; embora não sofra qualquer intervenção. A vantagem é controlar o efeito temporal e as externalidades das experiências. De forma infográfica: **G. Experimental:** R O_1-> X -> O_2 **G. Controlo:** R O_3-> -> O_4	• Evita erros de história e de maturação. • Permite conclusões sólidas.	• Obriga a medidas de pré-teste para apurar as variáveis chave. • Possibilidade de colocar os respondentes sob alerta. • Mortalidade das unidades de análise. • Possibilidade de colocar os respondentes sob alerta. • Não permite o teste de hipóteses. • Não permite a antecipação sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
Solomon Four Group Design *(True Experimental Designs)* **Desenho clássico**	Uma possível solução para combater a possibilidade de colocar os respondentes sob alerta e juntar os tipos de teste anteriores. Assim, e de forma infográfica, temos: **G. Experimental:** R O_1-> X -> O_2 **G. Controlo:** R O_3-> -> O_4 **G. Experimental:** R -> X -> O_2 **G. Controlo:** R -> X -> O_4	• Permite controlar efeitos de instrumentalização e de alerta. Os efeitos de instrumentalização podem ser controlados da seguinte forma: (O2- O4) (O2- O1)- (O4- O3) (O6- O5)	• Obriga a medidas de pré-teste para apurar as variáveis chave. • Mortalidade das unidades de análise. • Possibilidade de colocar os respondentes sob alerta. • Possibilidade de colocar os respondentes sob alerta. • Não permite o teste de hipóteses. • Não permite a antecipação sustentada de comportamentos. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Não permite a elaboração de multitratamentos.
Time Series Design *(Quasi-Experimental Designs)* **Desenho clássico**	Uma forma simples de experimentação simples: apurar a tendência de resultados para sucessivos tratamentos.	• Fácil de aplicar	• Os resultados podem estar influenciados a outros factores (normalmente a questão da **história** e da **maturação**). • Erro de selecção.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
			• Possibilidade de colocar os respondentes sob alerta. • Mortalidade das unidades de análise. • Não permite a avaliação de termos de interacção entre variáveis de causa.
Continuous Panel Design *(Quasi-Experimental Designs)* **Desenho clássico**	Permite controlar a dimensão pessoal (*cross-sections*) e a dimensão temporal. Permite apurar a tendência ao mesmo tempo que as características chave da população.	• Evita erros de história e de maturação. • Permite conclusões sólidas.	• Tamanho amostral largo. • Mortalidade das unidades de análise. • Desenho relativamente caro. • Erro de selecção. • Possibilidade de colocar os respondentes sob alerta. • Mortalidade das unidades de análise. • Não permite a avaliação de termos de interacção entre variáveis de causa.

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
Completely Randomized Design **Desenho estatístico**	É a forma mais simples de desenho experimental estatístico. Os tratamentos a aplicar são atribuídos aleatoriamente às análises. Muito usado para prever *scorings* de *default* de crédito, e em quase todos os estudos de natureza *cross-section*. As formas mais comuns de análise nestes casos são a ANOVA e o modelo de regressão. De forma infográfica, temos: Experiência 1: R -> X1 -> O_1 Experiência 2: R -> X2 -> O_2 Experiência 3: R -> X3 -> O_3	• Não há grande erro de selecção. • Permite o teste de hipóteses. • Permite a antecipação sustentada de comportamentos. • Pode permitir a avaliação de termos de interacção entre variáveis de causa. • Permite a elaboração de multitratamentos.	• Não permite avaliar com grande rigor o impacto das variáveis de controlo: todos os comportamentos são admissíveis. • Mede apenas o impacto numa variável de output (não mais).

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
Quadrado Latino **Desenho estatístico**	É um método que reduz os grupos envolvidos quando as interacções entre os níveis das variáveis e o controlo das variáveis não é importante (ver desenho ortogonal)	• Permite avaliar com maior rigor o impacto das variáveis de controlo. • Os resultados não podem estar influenciados a outros factores se o analista tiver cuidado. • Não há grande erro de selecção. • Permite o teste de hipóteses. • Permite a antecipação sustentada de comportamentos. • Permite a avaliação de termos de interacção entre variáveis de causa. • Permite a elaboração de multitratamentos. • Permite o controlo de duas variáveis independentes sem expandir a amostra.	• Obriga a que todas as variáveis de controlo tenham o mesmo número de níveis. • Não permite apurar os temos de interacção. • Não permite a avaliação de termos de interacção entre variáveis de causa. • Mede apenas o impacto numa variável de output (não mais)

Tipo de Experiência	Definição	Síntese das Vantagens	Síntese das Desvantagens
Factorial Design **Desenho estatístico**	Quando se deseja medir o impacto de medidas de interacção. No desenho estatístico poderá desejar-se avaliar o impacto em teias de causalidade de duas variáveis de output.	• Não há grande erro de selecção. • Permite o teste de hipóteses. • Permite a antecipação sustentada de comportamentos. • Permite a avaliação de termos de interacção entre variáveis de causa. • Permite a elaboração de multitratamentos. • Permite avaliar com grande rigor o impacto das variáveis de controlo: todos os comportamentos são admissíveis – desde que antecipados	• Dificuldade algébrica • Eventualmente maior dimensão amostral.

Fonte: adaptado de Aaker; Kumar; Day (1998)

Para a construção do processo gerador de tarifas vai-se optar pelo *Completely Randomized Design*), já que permite testar em simultâneo multitratamentos.

Note-se que quando no modelo estatístico desenhado (e posteriormente estimado por máxima verossimilhança) há um vector de parâmetros da distribuição e uma matriz variância-covariância associado aos parâmetros estimados. O inverso dessa matriz é designada por matriz de informação de Fisher. O objectivo desta matriz de informação é prever como é que uma experiência irá conter os parâmetros do modelo, antes de executar a experiência ou fazer qualquer simulação. É possível assim prever sobre a variabilidade dos resultados de experiências e fazer uma escolha entre a precisão e o custo de recolha de informação.

A estimativa de verosimilhança fornece uma estimativa do grau de incerteza, já que o próprio método e apuramento dos parâmetros baseia-se na minimização dessa incerteza. De uma forma mais formal: A Matriz de Informação de Fisher é uma forma de medir a quantidade de informação que uma variável aleatória possui sobre um parâmetro θ, parâmetro esse que caracteriza a sua distribuição de probabilidade[15].

---

[15] Construção da matriz de Fisher

1. **Construir a função verosimilhança:**

$$L(\theta) = L(\theta \mid w) = f_y(w_1 \mid \theta) \,.\, f_y(w_2 \mid \theta) \ldots f_y(w_n \mid \theta)$$

Onde θ são os parâmetros a obter e w as observações recolhidas. Nota: É importante observar que os valores de y são conhecidos por meio da amostra aleatória, de modo que L(θ|w) é função apenas de θ.

Por uma questão e facilidade no processo de derivação usa-se antes esta fórmula:

$$\log L(\theta) = \log(L(\theta \mid w))$$

Note-se que como o logaritmo natural de L é uma função crescente, então LnL=ln(L(θ)) tem o mesmo máximo de L(θ))

Porém, quando o modelo estatístico tem vários parâmetros a minimização da sua variância é complicada. É assim habitual usar critérios mais simples de optimização, mas relacionados com a minimização da variância dos parâmetros a estimar. Abaixo seguem as diferentes estratégias utilizadas:

- ***A-optimality* ("*average*" ou traço)**: Este critério pretende minimizar o traço (soma dos elementos da diagonal da matriz) da matriz de informação inversa. Este critério resulta numa minimização da variância média (A-Average) dos seus parâmetros.

---

2. Maximizar a função verosimilhança:

$$\max_{\theta} L(\theta) \quad \Rightarrow \quad \max_{\theta} \log(L(\theta \mid w))$$

3. Usar as derivadas como um indicador da variabilidade.

Note-se que se o que se deseja encontrar é um indicador de variabilidade (ou precisão) do estimador, nada melhor do que as derivadas primeira ordem do processo de estimação. Num processo de otimização, estas condições devem ser idênticas a zero se se desejar identificar um possível do máximo/mínimo.

Ou seja:

Se **I(θ)=**Informação de Fisher é uma forma de medir a quantidade de informação que uma variável aleatória possui sobre um parâmetro θ, que caracteriza a sua distribuição de probabilidade; então um sério candidato é a variabilidade do score=$\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)$.

Ou seja:

$$I(\theta) = E\left(\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2 \middle| \theta\right)$$

já que a média do score (por definição do processo de verosimilhança) tem de ser zero.

Sob as condições de regularidade e se a verossimilhança é duas vezes diferenciável então a informação de Fisher também pode ser obtida como:

$I(\theta) = E\left(\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2 \middle| \theta\right) = -E\left(\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right) \middle| \theta\right)$.

- ***C-optimality***: Este critério pretende minimizar a variância de um modelo linear.
- ***D-optimality* (*determinant*)**: Um critério popular é o da optimalidade-D que procura minimizar apenas $|(X'X)^{-1}|$, ou de forma equivalente, maximizar o determinante da matriz informação X'X.
- **E-optimality (eigenvalue)**: Este critério pretende maximizar o mínimo valor próprio da matriz de informação.
- **T-optimality**: Este critério pretende maximizar o traço da matriz informação.
- **G-optimality** (para previsão): Este critério (também muito popular) pretende maximizar o traço da matriz $(X'X)^{-1}X'$. Este critério minimizará assim a variância máxima dos valores previstos.

No contexto dos modelos GLM, habitualmente opta-se pelo critério *D-optimality* por uma questão de facilidade algébrica – segue-se assim a opção indicada por Dror & Steinberg (2006) e Dror & Steinberg (2008). Nos modelos OLS ou de MV o critério de *D-optimality* também é muito usado, sobretudo quando há efeitos de heterocedasticidade, ou quando o desenho experimental assume factores/níveis categorizados.

4.2.1. Desenho experimental: GLM com Tweedie

Foi visto nos capítulos anteriores que as companhias de seguro apostam, habitualmente, numa forma de tarifação por GLM onde combinam um modelo de Poisson+gama; ou num modelo compósito de Tweedie. Pretende-se com uma amostra recolhida no mercado, apurar o modelo Poisson+gama.

Porém, a questão que fica é assim de escolher qual a melhor ferramenta de observação do modelo já que não é possível fazer um ajustamento de

Poisson+gama[16], pois não é observada separadamente a frequência e o custo médio.

Há várias alternativas. Alguns autores, em diferentes contextos da indústria seguradora, têm-se focado no apuramento de modelos GLM com Tweedie no contexto de Desenho experimental (uma resposta possível ao problema aqui analisado). Como visto anteriormente, a Tweedie acompanha os modelos de frequência e custo médio. Da bibliografia analisada para o desenho amostral na regressão Tweedie é de referir as seguintes abordagens:

- Desenho sequencial: O desenho sequencial para respostas binárias tem uma história rica, que vem desde 1948 e na tentativa de encontrar desenhos cujos resultados apresentassem propriedades assimptóticas.

  Tem havido outros trabalhos um pouco mais recentes nesta área, desenvolvidos por Haines, Perevozskaya, & Rosenberger (2003), Ivanova, Wang (2004); Karvanen, Vartiainen, Timofeev, & Pekola (2007). Sendo trabalhos relevantes, não cabe porém neste trabalho o estudo aprofundado destes documentos basilares, tanto mais que estes autores concentram o trabalho em desenhos unifatoriais; e o desafio aqui é multifactorial.

  Woods, Lewis, Eccleston, Russell, (2006) e Dror & Steinberg (2008) apresentam algumas soluções para esta problemática de multifactorialidade; mas computacionalmente complexas, e cuja metodologia se baseia em "avanços e recuos" e tentativa e erro tornando o processo complexo e muito pouco intuitivo.

- Desenho baseado em *clusters*: Note-se o desafio da regressão de Tweedie prende-se em estimar três vectores de parâmetros: $\phi; p$ e θ; sendo que o p condiciona a calculatória dos restantes. No desenho por *clusters* procura-se encontrar grupos homogéneos de observações de forma a apurar p. Ou seja, p é de alguma forma estimado exogenamente. Esta ideia é

[16] Adicionalmente, é de referir que se observa apenas o prémio comercial final, que inclui, como já indicado, a frequência, o custo médio, as cargas administrativas e a carga para desvios de sinistralidade (além do lucro).

conceptualmente interessante, e computacionalmente é fácil de executar.

Dror & Steinberg (2006) sugerem uma abordagem baseada essencialmente no *cluster K-means* já que este processo, afirmam os seus autores, permite uma exploração rápida dos vários desenhos "*outperform the existing alternatives*." Assim, continuam estes autores, "*[g]iven a set of local D-optimal designs, the core of the method proposed is to combine them into a set of location vectors and use K-means clustering to derive a robust design*".

A possibilidade de encontrar um óptimo local com este método tem, no entanto, um sério problema no que diz respeito aos restantes coeficientes do modelo: como avaliar a estimativa do seu grau de precisão? Já que os *clusters* foram definidos de forma exógena e num desenho experimental a amostra é sempre reduzida para fazer grandes experiências. Uma alternativa a estudar oportunamente, em outro trabalho, será o uso de 'força bruta' computacional de forma a garantir o melhor modelo assente em diferentes níveis de p. Algoritmos baseados na ideia de *random forest* poderão ser uma pista relevante a estudar num outro trabalho.

4.2.2. Desenho experimental: Uma solução pragmática em populações Tweedie – Box-Cox

De acordo com a revisão vista acima, os modelos GLM com Tweedie não são facilmente aplicáveis no contexto experimental. Há que encontrar uma solução pragmática

Note-se que o maior problema é ter um modelo de análise experimental em condições de heterocedasticidade, ou em modelos de dispersão, onde a Tweedie se adapta bem.

Box & Cox (1964) afirmam que para a estabilizar a variância o método habitual é determinar empiricamente ou teoricamente a relação entre a variância e a média. Sendo que a relação empírica pode ser encontrada pelo gráfico do logarítimo com a média. Adicionalmente, referem estes autores, um outro método poderá ser a

escolha de uma transformação. É neste contexto de regressão que se centra o pragmatismo anunciado.

Transformações de dados sempre foram usadas nos modelos de regressão/ANOVA de forma a conseguir essencialmente (Box & Cox (1964)): (i) simplicidade na estrutura de E(y); (ii) homocedasticidade, (iii) normalidade, (iv) independência das observações e (Gujarati (1995)) (iv) linearidade dos parâmetros, (v) uma melhor interpretação da forma funcional adequada à teoria.

Para valores esperados positivos[17], a transformação de Box–Cox:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \text{para } \lambda \neq 0 \\ \log(y) \text{ para } \lambda = 0 \end{cases} \quad (26)$$

é muito usada, já que: "*is continuous as λ tends to zero and monotone increasing with respect to x for any λ*". Há alguns valores de λ fáceis de interpretar, conforme indicado na tabela abaixo.

Tabela 4.1 - Interpretação para possíveis formulações de Box-Cox

λ	**Interpretação**
1,00	Não é necessária qualquer transformação; os resultados produzidos são idênticos aos dados originais
0,50	Transformação de raiz quadrada
0,33	Transformação de raiz cúbica
0,25	Transformação de raiz à quarta
0,00	Log natural

Fonte: autor

A escolha de λ no entanto é habitualmente feita de forma automática. Osborne

[17] Note-se que aqui se está a estimar o prémio comercial, e não o prémio puro, pelo que o valor esperado será sempre positivo.

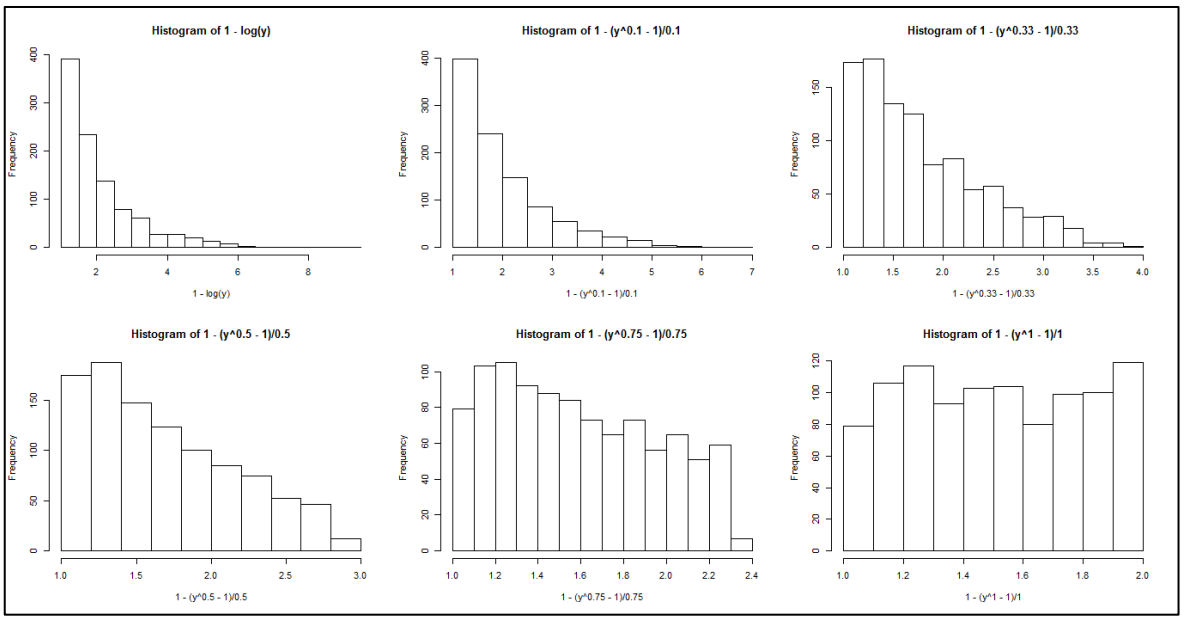
(2010) propõe o seguinte algoritmo:

1. Dividir a variável chave em 10 (ou mais) intervalos;
2. Calcular a média e o desvio-padrão para cada intervalo;
3. Fazer um gráfico com $\log(\sigma)$ *vs* $\log(\mu)$ para cada uma das regiões;
4. Estimar o declive médio do gráfico, e usar o $1-\lambda$ como valor inicial de λ.

Apesar deste autor não mencionar, na prática trata-se do clássico método (já referido) por Box & Cox (1964): "*An adequate empirical relation may often be found by plotting log of the within-cell variance against log of the cell mean"* e que pode ainda ser mais estudado em Harrison & McCabe (1979).

Note-se que este algoritmo não é unanime. Charpentier (2010), Ripley, Venables, Bates, Hornik, Gebhardt & Firth (2015) assumem que a melhor forma de estimar λ será aquele que garante a maior verosimilhança. Desenhar a evolução da função máxima verosimilhança pode ser útil neste caso. Contudo este método não é prudente, já que qualquer $|\lambda|$ elevado irá reduzir excessivamente a variabilidade da variável objectivo. E quando se re-construir a variável objectivo de $y^{(\lambda)}$ para y o resultado poderá ser, no limite, uma variável estimada "tendencialmente horizontal" – isto é sem qualquer variabilidade. Alguns *softwares* e pacotes estatísticos (MASS no R, STATA) mantêm esta abordagem, mas impõem limites a $|\lambda|$, habitualmente em $|\lambda| < 1$, $|\lambda| < 2$.

Figura 4.2 - Exemplos das distribuições associadas à transformação Box-Cox



Fonte: autor

Por fim, uma outra alternativa, é optar por um valor de λ que faça sentido para o analista. Uma leitura cuidada de Box&Cox (1964) aponta nesse sentido. Assim é de salientar que a forma $loglin <=> \lambda = 0$ teoricamente é a faz mais sentido usar no modelo de prémios, e é mais fácil de interpretar:

1. A distribuição de custos totais/prémios puros (isto é a Tweedie com parâmetros credíveis) é, visualmente, próxima de uma log normal ou gama (ver figura 14).
2. O modelo log-lin tem a vantagem dos coeficientes representarem elasticidades; um conceito com muito significado em termos de prémios.

Gujarati (1995)[18] apresenta uma forma muito simples de encontrar e testar a

[18] Na verdade trata-se de um teste desenvolvido por: MacKinnon, H. White, and R. Davidson, "Tests for Model Specification in the Presence of Alternative Hypothesis; Some Further Results." Journal of Econometrics, vol. 21, 1983, pp. 53–70. Gujarati refere ainda que este teste é similar ao proposto em A. K. Bera and C. M. Jarque, "Model Specification Tests: A Simultaneous Approach," Journal of Econometrics, vol. 20, 1982, pp. 59–82.

transformação dos dados – pelo menos entre a forma linear e log. Embora o processo de estimação pareça um pouco complexo; a verdade é que a lógica do teste é muito simples: se o modelo linear é de facto o correto, a fórmula $\exp(\log(\widehat{y_{com\ base\ no\ modelo\ log}})$ estará relacionada com o modelo em estudo (bastará fazer uma regressão e aplicar um teste-t). Da mesma forma, que se o modelo for logarítmico, $\log(\widehat{y_{com\ base\ no\ modelo\ lin}})$ não estiver relacionado com o log da variável objectivo então esta fórmula funcional não será a mais ajustada. Neste último caso, se o teste der negativo para a forma log-linear, a questão centra-se em estimar λ.

Resumindo, para apurar o modelo tarifário da concorrência em contexto de desenho experimental, a opção Box-Cox será a opção seguida em detrimento da regressão Tweedie,

Além disso, após a análise da melhor forma funcional, numa segunda fase poderá ser aplicada uma regressão de Tweedie; obtendo com Box-Cox uma indicação grosseira do que poderá ser p (isto é a forma da Tweedie), fugindo das dificuldades apontadas na secção 2.1, deste capítulo. Tal como mencionado, a opção de 'força bruta' poderá ser uma alternativa a estudar.

4.2.2.1. Desenho experimental: Correção à regressão de Box-Cox

A variável chave a estimar é $y = \text{prémio comercial}$ e não $y^{(\lambda)}$. Após o processo de estimação será necessário decompor a formulação de Box-Cox na sua formulação correcta:

$$\hat{y} = \begin{cases} \widehat{y^{(\lambda)}}\lambda + 1 \text{ quando } \lambda \neq 0 \\ \exp\left(\widehat{y^{(\lambda)}}\right) \text{ quando } \lambda = 0 \end{cases} \quad (27)$$

Mas esta formulação não é a mais estatisticamente eficiente… De facto, a aplicação da fórmula Box-Cox subestima sistematicamente o valor esperado de y. De uma forma intuitiva, o ponto de massa da reta de regressão Box-Cox (na média de x e na média de y) não é idêntico nas duas equações em ordem a $\widehat{y^{(\lambda)}}$ e

em ordem a $\hat{y}$.

Uma solução expedita (Wooldridge (2003), para quando $\lambda = 0$, mas que pode ser generalizado para a qualquer variante da regressão Box-Cox) é possível de obter regredindo y em $\hat{y}$, sem constante. O coeficiente associado a $\hat{y}$ dará o factor de correcção dos pontos de massa. Assim a previsão final de y será a de:

$$\hat{\hat{y}} = \hat{y} \times coeficiente\ de\ correção.$$

4.3. Identificar os factores que definem o produto

Em relação ao apuramento dos factores, o trabalho experimental está facilitado: Cada cliente tem de preencher uma folha de simulação (ou são-lhe pedidos dados para o mesmo efeito) para que uma cotação lhe seja entrega. Habitualmente, não há mais dados a trabalhar além dos solicitados (embora seja conhecido que algumas seguradoras em regime de *bancassurance* usem os dados de comportamento bancário; e afirma-se que noutros países o perfil nas redes sociais possa ser utilizado). Ou seja, o trabalho de validação dos factores em sessões de *brainstorming* e com entrevistas em profundidade com diferentes especialistas (ver Barker (1994)[19]) é grosso modo dispensado, já que as companhias indicam na simulação quais os factores recolher.

No caso de ramo automóvel, os factores habitualmente considerados são:

1. Características dos segurados

    - Género: O racional desta variável está associado essencialmente a uma diferente frequência de sinistralidade de acordo com os géneros.

      É de salientar, porém, que a 1 Março de 2011 o Tribunal Europeu de

[19] Este autor refere que é necessário falar com:
1. The experts
2. The peripheral experts
3. The operators or technicians
4. The customers [neste caso não aplicável],

Justiça decidiu que as companhias de seguros que usassem o género como um factor de risco estavam a desrespeitar as leis de igualdade da UE. E que em Fevereiro de 2015, (Lei n.º 9/2015) passaram a ser "*admitidas diferenciações nos prémios e prestações individuais desde que proporcionadas e decorrentes de uma avaliação do risco baseada em dados actuariais e estatísticos relevantes e rigorosos*". A possibilidade de utilização desta variável a partir de 2015 passou assim a ser uma realidade.

- Idade: O racional desta variável é medir a inexperiência e o apetite ao risco dos segurados. Trata-se de uma variável com impacto na frequência de sinistro e, dependendo das coberturas, no custo médio.
- Histórico de sinistralidade: Para a inadaptação do condutor usa-se cada vez mais e de forma automática o histórico de sinistralidade que pode ser consultado pela seguradoras portuguesas na SegurNet (uma plataforma gerida pela associação do sector com o histórico de sinistralidade).
- Idade do condutor quando tirou a carta: Quando combinado com a idade, obtém-se uma variável instrumental da experiência do condutor.
- Estado civil: Pouco usado em Portugal, embora haja alguma sensibilidade que aponte para o facto de que condutores casados terem menos acidentes que o resto da população.
- Percurso habitual: Variável que indica a frequência de sinistro – quanto maior a distância casa-trabalho, maior a probabilidade de acidente.
- Forma de pagamento: A forma de pagamento revela a pressão financeira que o condutor está sujeito, factor que está correlacionado com o perfil de condução – frequência. Além disso, a forma de pagamento está correlacionada com o consumo de capital da

seguradora e, por isso, com efeito no prémio comercial via cargas administrativas e coeficiente de lucro.

2. Características dos riscos seguros

   - Classificação do veículo: A relação peso potência aumenta a frequência de sinistro. Se um carro for leve para a cilindrada que tem, haverá uma frequência elevada de sinistro e provavelmente um maior custo médio.
   - Marca e classificação do veículo: Há marcas cujas peças custam mais, do que outras, pelo que esta variável tem um impacto no custo médio. O tipo de construção, os dispositivos de segurança também tem a sua influência no custo médio.
   - Uso do objecto seguro: Se o objecto é fundamental para o dia-a-dia, ou para uso profissional, a frequência de sinistro aumenta – (mas a frequência por unidade de exposição (medida em km condução) diminui. Assim, ter a profissão e qualquer variável instrumental do uso do objecto seguro é relevante.

3. Especificidades regionais ou de contexto genérico

   - Meteorologia: O contexto de sinistralidade tem sido dos temas menos considerados na construção de uma tarifa. Por exemplo, se chove mais, há mais acidentes, mas as companhias fazem as tarifas para um dado país e onde implicitamente as taxas de pluviosidade não variam). Se há uma crise, as pessoas andam menos de automóvel, logo há menos acidentes. Estas variáveis de contexto são ligadas à evolução do tempo; e deveriam ser analisadas.
   - Região: Há muitas vezes variáveis regionais que são muitas vezes descuradas, já que tudo é colocado em grandes regiões comerciais e não com a granularidade suficiente. Tipicamente o poder de compra concelhio (por exemplo) está relacionado com variáveis de custo médio e apetite pelo risco dos condutores.

- Sectoriais: Se a utilização for profissional e em determinados sectores, por exemplo transportes e distribuição, há uma maior exposição, logo, uma maior probabilidade de maior frequência.

4. Companhia

- Aos factores apresentados deve ser considerado um factor extra: a companhia. É de supor que este factor altere a relação entre todos os outros – é se se quiser um elemento chave de interacção com todos os factores: cada companhia terá o seu modelo de tarifação. Essa informação deve ser aproveitada para apurar o verdadeiro risco de mercado, já que em teoria todas as companhias estão a medir o mesmo risco: frequência e custo médio. De outra forma, uma vez que (e é razoável que assim seja) não há troca de modelos tarifários entre entidades seguradoras, haverá tantos modelos quanto o número de companhias existentes – não se está a dizer que o efeito das companhias é um efeito de ruído ou causador de *nuances* nos dados; está-se antes a afirmar que o modelo é estatisticamente distinto de companhia para companhia e que tal não se controla com uma simples aleatorização. (A inclusão desta informação no processo de estimação será detalhada na secção seguinte). De uma forma matematizada, e considerando que x tem a leitura de variável exógena habitual, tem-se que o modelo terá a seguinte forma:

$$y_i = f_{companhia\ j}(x_i) \tag{28}$$

e não:

$$y_i = f\ \ (\text{Companhia}_\text{i}, x_i) \tag{29}$$

onde x_i designa o conjunto de variáveis acima identificados.

4.4. Identificar os níveis que definem os factores

Em relação aos níveis a questão é distinta, já que a recolha é feita de forma contínua em algumas variáveis chave; mas o tratamento pode ser discreto. Mais, pode haver alguns rácios e valores derivados (o apuramento do peso potência é talvez o caso mais evidente).

É de notar, no entanto, que a opção pela escolha dos níveis tem de ser de tal forma que se minimize o volume de informação a recolher e torne o projecto de campo eficiente e exequível. Aqui neste ponto terá de se assumir que com recurso a um painel de especialistas (ver Barker (1994)) é possível minimizar/agregar o número de níveis. Este especialista, Barker, revela, com especial ênfase, a forma de organizar a informação neste sentido: "*It is best to organize an experiment as a team effort and use the brainstorming technique to scope the entire problem."*

4.5. Desenho óptimo

Quando os dados parecem mostrar que há uma relação linear é possível tentar obter um modelo de regressão que se aproxime da relação apurada – está-se no contexto de *Completely Randomized Design*.

Para que esse modelo seja verdadeiro centrado é necessário verificar um conjunto de hipóteses, muitas vezes chamadas hipóteses clássicas, e estimar o processo gerador de tarifas por máxima verosimilhança. Com condições clássicas (linearidade nos parâmetros, amostra aleatória, ausência de perfeita multicolinearidade) é possível garantir a centragem dos estimadores de máxima verosimilhança.

Na análise na análise do potencial viés não foi necessário assumir qualquer distribuição. Para tal basta recordar a equação 18, que por uma questão de conveniência aqui se repete:

$$\begin{aligned}\mathbf{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{E}(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{E}(\mathbf{u}|\mathbf{X}) \qquad (18)\end{aligned}$$

com u sendo o vetor de erros, e o modelo a estimar ser da forma: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$.

O que esta equação permite concluir é: se porventura os dados recolhidos estiverem correlacionados com alguma variável presente no resíduo (*i.e.* $\mathrm{E}(\mathbf{u}|\mathbf{X}) \neq \mathbf{0}$) este poderá perder precisão. Esta perca de precisão será tanto maior quanto maior for a correlação entre as variáveis explicativas (*i.e.* quanto maior for $(\mathbf{X}'\mathbf{X})^{-1}$).

Assim, numa amostra aleatória, como é habitual trabalhar num contexto de regressão, há um mecanismo aleatório que "escolhe" a amostra dentro de todas as amostras possíveis.

No contexto de desenho experimental é possível ver o problema de outra forma. Se houver um processo gerador de dados (e não um mecanismo aleatório de selecção das amostras), a estrutura de dados mantem-se e a fórmula do estimador será a mesma e a centragem está garantida (condição essencial). É assim possível recolher quaisquer dados para garantir a centragem do estimador $\hat{\beta}$. Mas é possível ir mais longe, a centragem não é a única condição essencial. Com este pressuposto do mecanismo gerador de dados é, também, possível ir mais além em termos de variância dos testes.

A segunda condição para um bom desenho amostral é maximizar a potência dos testes. Para garantir que o processo é eficiente, isto é que o desvio padrão associado a cada uma das estimativas das betas é mínimo, à que analisar a formulação do estimador de beta:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (30)$$

E o desvio padrão de $\hat{\beta}$ será dado por:

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\sigma}(\mathbf{X'X})^{-\mathbf{1}} \qquad (31).$$

Assim, se se quiser ter uma estimativa eficiente, pode-se apostar em ter o número de casos suficiente para estimar a regressão e que cada um dos elementos da diagonal da matriz $(\mathrm{X'X})^{-1}$ seja mínimo (assume-se que não há qualquer efeito de interacção).

A melhor forma de o fazer (cf. Cramer-Rao) é garantir que a matriz $(\mathbf{X'X})^{-1}$ seja apenas preenchida na diagonal – que não haja qualquer correlação entre os diferentes x. Nesse caso, diz-se que a matriz $(\mathbf{X'X})^{-1}$ é ortogonal.

Ou seja,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

$$\widehat{\boldsymbol{\beta}} = \mathbf{I}\,\mathbf{X}'\mathbf{y} \qquad (32).$$

Com recurso à $\widehat{\boldsymbol{\beta}} = \mathbf{I}\,\mathbf{X}'\mathbf{y}$ os valores do estimador são muito fáceis de estimar e terão sempre a variância mínima. Note-se que se $(\mathbf{X}'\mathbf{X})^{-1} \neq \mathrm{I}$ haverá um fenómeno de confusão (*confounding*) não sendo possível estimar sem um erro elevado associado os coeficientes estimados. A questão é que a variância é reduzida apenas se se estiver a analisar o verdadeiro mecanismo gerador de dados – o que implica uma boa dose de presunção.

O desenho ortogonal tem ainda uma vantagem adicional em termos de centragem. Se de facto tiver sido esquecida uma variável relevante, (*i.e.* $\mathrm{E}(\mathbf{u}|\mathbf{X}) \neq \mathbf{0}$) o seu impacto será mínimo, já que $(\mathbf{X}'\mathbf{X})^{-1} = \mathrm{I}$ (ver equação 18).

4.5.1. Desenho óptimo – ajustes funcionais

Após o desenho amostral e análise de dados, é possível fazer um simples modelo de regressão linear para apurar a importância dos factores (através de um teste t) e o grau de criticidade dos seus níveis (novamente com o teste t, e assumindo os níveis como variáveis *dummy*), ajustando eventualmente, a melhor forma funcional.

Vale porém a pena explorar o que significa a equação (28) e a necessidade de dizer que se tem de ter uma função por companhia vista na secção anterior. A equação (28) foi vista na secção anterior, mas por uma questão de conveniência aqui se repete:

$$y_i = f_{companhia\ j}(x_i) \qquad (28).$$

Esta equação indica que se opta por recolher e modelar os dados para uma única companhia, assumindo que cada uma das companhias deverá ter os modelos de tarifação autónomos. Se se considerar apenas duas companhias, o que se está a dizer é que o modelo em causa pode ser descrito como:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix}$$

Ou ainda: $y = \boldsymbol{X\beta} + \mathbf{e}$ (33)

onde:

$$\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} = \boldsymbol{e} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1 I_T & 0 \\ 0 & \sigma_2 I_T \end{pmatrix} = \boldsymbol{W} \right] = e \sim N(\mathbf{0}, \boldsymbol{W})$$

Mas uma questão surge ao analisar para estes dados: “As companhias estão a operar no mesmo mercado, será que se está a usar toda a informação disponível para capturar o modelo tarifário no mercado?”. Ou de forma ainda mais directa: “Os modelos tarifários são autónomos, mas será que são independentes?” De

uma forma mais matematizada (aplicando o mesmo racional de Griffiths, Hill & Judge, (1993)): "E se os erros das diferentes equações, $\boldsymbol{e_1}$ e $\boldsymbol{e_2}$, estiverem correlacionados?" De tal forma que (33) possa ser escrito da seguinte forma:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} \tag{34}$$

com

$$\begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \end{bmatrix} = \boldsymbol{e} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{pmatrix} = \boldsymbol{W} \right] = e \sim N(\boldsymbol{0}, \boldsymbol{W}) \ .$$

A ideia é que podemos estimar a equação (32) por blocos, de forma a assumir diferentes formas funcionais, e quem sabe diferentes variáveis explicativas; mas com (33) podemos ter maior precisão na previsão e maior potência nos testes. A demonstração destas afirmações, segue abaixo, apenas para o caso de duas companhias, embora a generalização seja directa (e pode ser confirmada no Anexo do cap. 17 de Griffiths, Hill & Judge, (1993) e junto de Zellner (1962)). Por máxima verosimilhança, tem-se que:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'W^{-1}X})^{-1}\boldsymbol{X'y}$$

$$= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \underbrace{\begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix}^{-1}} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$E(\boldsymbol{ee'}) \neq \boldsymbol{\sigma I}$, o caso habitual não se aplica, pelo que: $E(\boldsymbol{e}\, \boldsymbol{e'}) = \boldsymbol{W}$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'W^{-1}X})^{-1}$$

Porém, este estimador não passível de ser calculado, já que a matriz $\boldsymbol{W}$ não é conhecida; então tem de ser estimada. Ou seja, é necessário apurar a seguinte relação: $\hat{\sigma} = \boldsymbol{\hat{e}\widehat{e'}}$.

Então tem-se:

$$\widehat{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{35}$$

$$= \left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{\sigma_{11}} I_T & \widehat{\sigma_{12}} I_T \\ \widehat{\sigma_{21}} I_T & \widehat{\sigma_{22}} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}\right]^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$= \underbrace{\left[\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} \widehat{\boldsymbol{e_1}'\boldsymbol{e_1}} I_T & \widehat{\boldsymbol{e_2}'\boldsymbol{e_1}} I_T \\ \widehat{\boldsymbol{e_1}'\boldsymbol{e_2}} I_T & \widehat{\boldsymbol{e_2}'\boldsymbol{e_2}} I_T \end{bmatrix}^{-1} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}\right]^{-1}}_{\text{Cov}(\widehat{\widehat{\boldsymbol{\beta}}}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}} \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

Uma vez que o modelo sem inclusão das restantes seguradoras não é homocedástico (ou seja, tem-se $E(\boldsymbol{ee}') \neq \boldsymbol{\sigma I}$), os diferentes estimadores $\widehat{\boldsymbol{\beta}}$ já não são centrados e de eficiência mínima. Porém, $\widehat{\widehat{\boldsymbol{\beta}}}$ é, em caso de boa parametrização, (tendencialmente) homocedástico, pois está-se a incluir o padrão de dispersão dos erros, e por isso mais eficiente[20]. A análise dos testes t e F será assim mais precisa e o desenho experimental ganha mais consistência. E desta forma consegue-se capturar quais os mecanismos geradores de dados/prémios puros das diferentes companhias.

Evidentemente que o ajustamento SUR deve ser aplicado antes da desconstrução da variável objectiva e nesse sentido da correcção do ponto de massa apurado em b2.i).

4.6. Caso prático

Para melhor compreender a aplicação deste conceito é usada uma Base de

[20] Vale a pena referir que qualquer um dos estimadores ($\widehat{\boldsymbol{\beta}}$ ou $\widehat{\widehat{\boldsymbol{\beta}}}$) apresentados, desde que bem especificado, é assintoticamente eficiente (Griffiths, Hill & Judge, (1993)).

Dados que retracta as condições de mercado segurador em 2011 do automóvel português para as coberturas de responsabilidade civil e assistência em viagem de forma agregada.

Os dados apresentados foram ligeiramente retocados de forma a garantir o anonimato das companhias sob análise. A variável 'forma de pagamento' foi excluída da análise de forma a criar mais um elemento de não-identificação das seguradoras.

A recolha de informação foi com recurso a Mediadores Mistério cuja ficha técnica sai fora do âmbito desta tese (além de permitir a identificação das seguradoras).

As subsecções seguintes acompanham as fases clássicas de desenho experimental, já referidas na secção a deste capítulo.

Fase 1: Identificar os factores e níveis que definem o produto;

As variáveis consideradas não foram definidas no âmbito desta tese, nem pelo seu autor. A discussão das variáveis presentes é indicada na tabela abaixo. A simulação deve corresponder a um seguro com capital mínimo obrigatório de Responsabilidade Civil (RC), mais uma cobertura mínima de Assistência em Viagem.

Tabela 4.2 - Discussão de factores e níveis

Variáveis	**Número de níveis**	**Identificação dos níveis**
Género	2	Homem Mulher
Idade	7	19 anos 23 anos 28 anos 35 anos 45 anos 57 anos 67 anos

Variáveis	Número de níveis	Identificação dos níveis
Histórico de sinistralidade	21	0 sin. / 10 anos / 15 anos 0 sin. / 15 anos / 15 anos 0 sin. / 2 anos / 2 anos 0 sin. / 4 anos / 4 anos 0 sin. / 5 anos / 10 anos 0 sin. / 5 anos / 12 anos 0 sin. / 5 anos / 5 anos 0 sin. / 5 anos / 6 anos 0 sin. / 5 anos / 7 anos 0 sin. / 5 anos / 8 anos 0 sin. / 7 anos / 9 anos 1 sin. / 0 anos / 1 anos 1 sin. / 0 anos / 2 anos 1 sin. / 0 anos / 3 anos 1 sin. / 0 anos / 4 anos 1 sin. / 0 anos / 5 anos 1 sin. / 1 anos / 5 anos 1 sin. / 2 anos / 9 anos 1 sin. / 3 anos / 9 anos 1 sin. / 4 anos / 10 anos sem experiência
Idade do condutor quando tirou a carta	Excluido	
Estado civil	Excluido	
Percurso habitual	Excluido	
Classificação do veículo	6	Caminheta Ligeiro Particular Misto Monovolume Pickup Todo o Terreno
Idade do veículo	13	0 1 2 3 4 5 6 7 8 9 10 15 20
Marca do veículo	Excluido (associado ao tipo de veículo e idade do veículo)	
Uso do objecto seguro	Excluido	
Metereologia	Excluido	
Região	58	Açores - ANGRA DO HEROISMO Açores - PONTA DELGADA Açores - VILA DO PORTO Aveiro - AVEIRO Aveiro - OLIVEIRA DE AZEMEIS Aveiro - VALE DE CAMBRA Beja - BEJA Beja - MÉRTOLA Beja - ODEMIRA Braga - BARCELOS Braga - BRAGA

Variáveis	Número de níveis	Identificação dos níveis
		Braga - CELORICO DE BASTO Braga - GUIMARÃES Bragança - BRAGANÇA Bragança - MONCORVO Castelo Branco - BELMONTE Castelo Branco - COVILHÃ Castelo Branco - SERTÃ Coimbra - COIMBRA Coimbra - OLIVEIRA DO HOSPITAL Évora - ESTREMOZ Évora - ÉVORA Faro - CASTRO MARIM Faro - LOULÉ Faro - SILVES Funchal - FUNCHAL Funchal - PORTO SANTO Funchal - SANTANA Guarda - GUARDA Guarda - TRANCOSO Guarda - VILA NOVA FOZ COA Leiria - LEIRIA Leiria - PENICHE Lisboa - CADAVAL Lisboa - LISBOA Lisboa - LOURES Lisboa - SINTRA Lisboa - TORRES VEDRAS Portalegre - AVIS Portalegre - ELVAS Portalegre - PORTALEGRE Porto - AMARANTE Porto - PAREDES Porto - PORTO Santarém - ABRANTES Santarém - OURÉM Santarém - SANTARÉM Setúbal - MONTIJO Setúbal - SANTIAGO CACÉM Setúbal - SEIXAL Setúbal - SETÚBAL Viana do Castelo - MELGAÇO Viana do Castelo - VIANA DO CASTELO Vila Real - SABROSA Vila Real - VILA REAL Viseu - OLIVEIRA DE FRADES Viseu - SÃO PEDRO DO SUL Viseu – VISEU
Companhias	7	Confidenciais

Fonte: autor

Fase 2: Desenho óptimo

A amostra a recolher foi desenhada por variável, mantendo o número de observações mínimas. A ortogonalidade não foi garantida. O modelo estimado é assim centrado, embora não seja necessariamente estatisticamente eficiente.

Assim, definiu-se um caso padrão e por variável foram alterados os níveis, sendo desenhadas 5 subamostras.

Apesar da construção da amostra total ser ortogonal, se o modelo gerador de dados seguir uma distribuição cujo resultado final dependa do nível em que se está poderá ser perdido alguma capacidade de previsão do modelo.

O caso padrão não pode ser revelado por questões de confidencialidade do projecto.

Fase 3: Recolha de informação

Esta base de dados foi recolhida pelo autor no início da década para um projecto específico de consultoria. Uma ficha técnica detalhada poderia permitir a identificação do cliente, dos objectivos do projecto e das condições de mercado. Assim, todos os dados foram cuidadosamente calibrados de forma a não permitir determinar as companhias alvo, nem os resultados operacionais.

No entanto, é de referir que para a realização da recolha de informação recorreu-se a mediadores não exclusivos. Esta não exclusividade é uma garantia para avaliar, de forma imparcial e objetiva, a estrutura de custos dos serviços prestados. Estes mediadores recolheram simulações, de acordo com um cenário/perfil previamente estruturado para ser simulado e com posterior registo de informação em guiões de observação.

Houve o cuidado de garantir que os mediadores faziam a recolha em concelhos pertencentes às regiões de trabalho.

Fase 4: Análise

Antes de mais, é de salientar que o modelo a estimar será executado com os seguintes passos:

1. Construção de um modelo por seguradora e apuramento do λ;
2. Estimação do modelo por SUR;
3. Correção do ponto de massa;
4. Apuramento do melhor estimador de β;

5. Análise crítica de resultados e, eventualmente, repetição do ciclo.

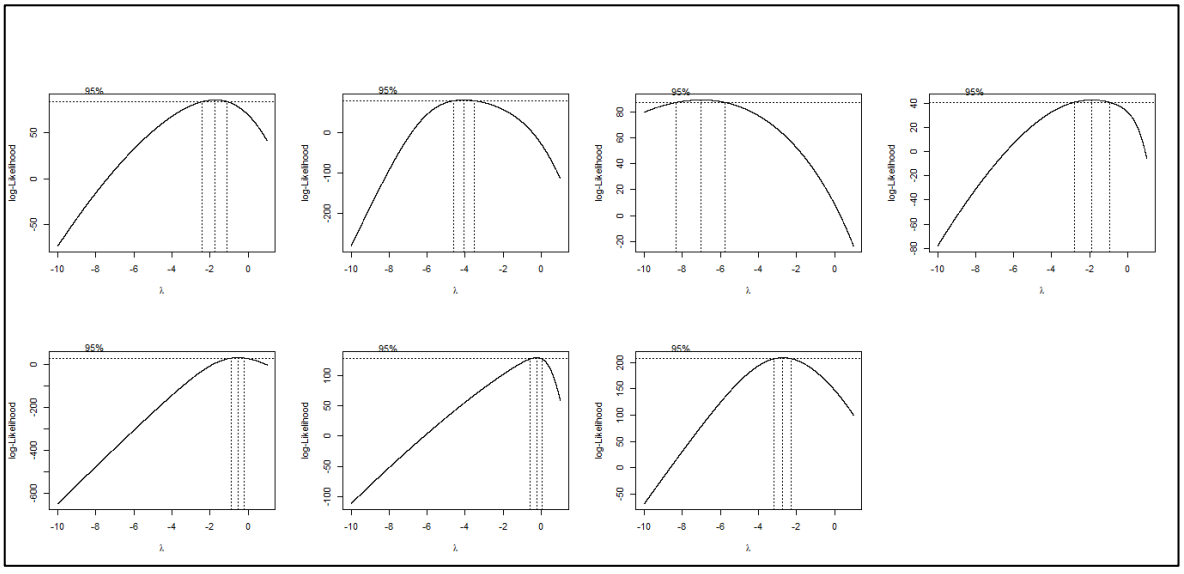
F4.1. Construção de um modelo por seguradora e apuramento do λ.

Conforme discutido, o modelo deverá incluir todas as variáveis recolhidas para cada seguradora.

Para a escolha de λ optou-se inicialmente por fazer um gráfico com a evolução da função verosimilhança entre -10 e 1, sendo os limites definidos em torno de zero e de forma a incluírem os máximos de cada uma das funções. Os resultados podem ser vistos na figura abaixo e indicam que os λ que maximizam as funções objectivo são: -3,504; -1,723; -4,055; -3,469; -1.380; -3,276; e -3,024.

Todos os pontos óptimos se afastam dos intervalos fáceis de interpretar, e indicados na tabela 9. É ainda de salientar que aparentemente as companhias usam de facto diferentes modelos tarifários reforçando a ideia de estimar cada modelo separadamente, de acordo com a equação 28.

Figura 4.3 - Evolução das funções verosimilhança para cada uma das companhias e apuramento do λ

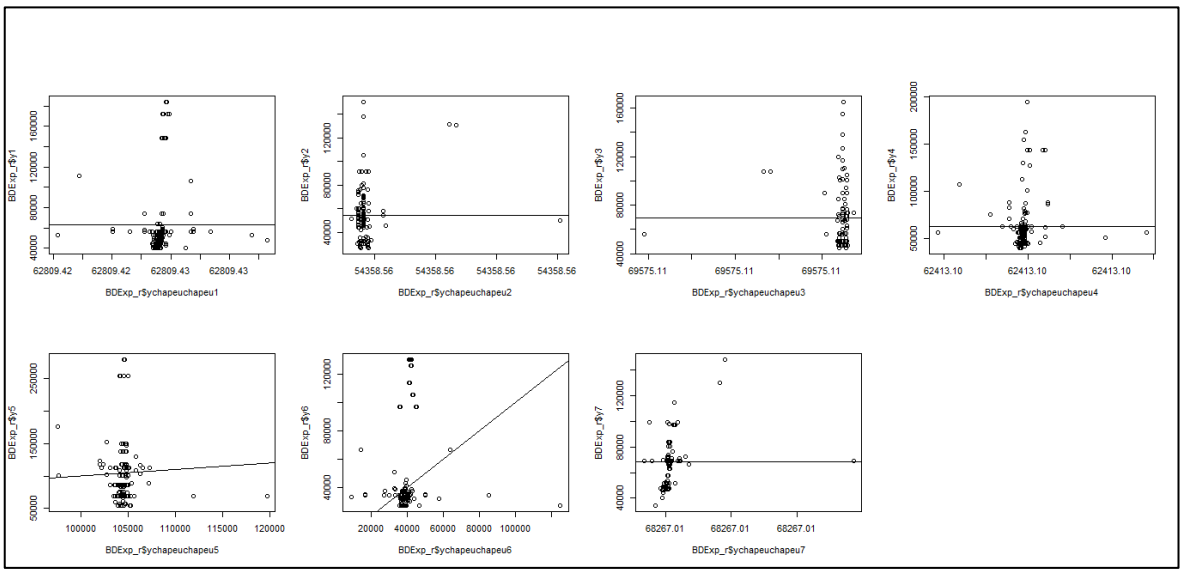


Fonte: autor

Quando se usam estes valores de lambda, e se estima o processo gerador de

dados por OLS e se reconstrói a variável objectivo final, o resultado é desolador. A escolha tão rigorosa de lambda acaba por eliminar toda a variabilidade do modelo. A figura abaixo salienta os resultados, com $\lambda < \sim -3{,}2$ a variável objectivo final fica uma recta paralela ao eixo das abcessias. Se se alargar a precisão do software, de forma a que este trabalhe com mais casas decimais, o resultado melhora um pouco, mas não de forma significativa.

Figura 4.4 - função de $\hat{y}$ com transformação de Box Cox que maximiza a função verosimilhança



Fonte: autor

O modelo será assim estimado com recurso a $\lambda = 0$.

F4.3 Estimação com modelo SUR

O modelo estimado por SUR, ainda com a transformação Box Cox, apresenta uma matriz de correlação[21] forte (apreciação qualitativa) entre os modelos das diferentes seguradoras. De facto, a correlação entre as tarifas estimadas varia

[21] A matriz de correlação (W') é mais fácil de interpretar que a matriz de covariância (W). Evidentemente que W foi a usada no processo de estimação.

entre 49% e os 93%.

Tabela 4.3 - Matriz de correlação dos resíduos dos modelos estimados por companhia

The correlations of the residuals

	eq1	eq2	eq3	eq4	eq5	eq6	eq7
eq1	1.00	0.28	0.40	0.53	0.64	0.51	0.34
eq2	0.28	1.00	0.60	0.52	0.40	0.32	0.56
eq3	0.40	0.60	1.00	0.56	0.48	0.54	0.38
eq4	0.53	0.52	0.56	1.00	0.52	0.29	0.58
eq5	0.64	0.40	0.48	0.52	1.00	0.49	0.24
eq6	0.51	0.32	0.54	0.29	0.49	1.00	0.01
eq7	0.34	0.56	0.38	0.58	0.24	0.01	1.00

Fonte: autor

F4.3 Correcção do ponto de massa

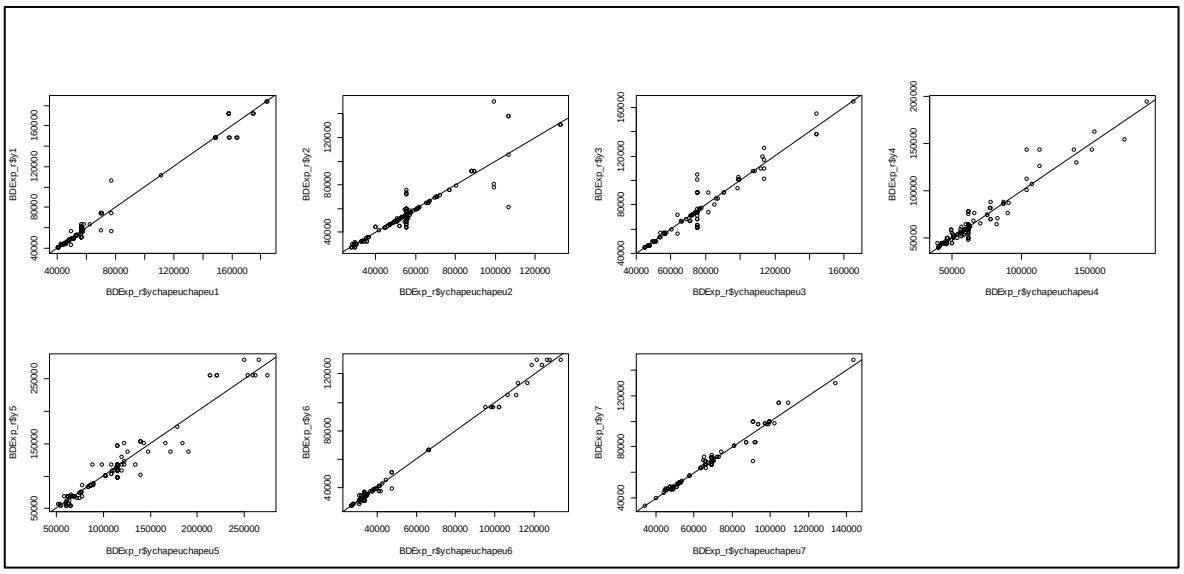
A correcção do ponto de massa tem impacto variado: numas é quase insignificante, noutras pode implicar uma correcção de >4%. De facto, por companhia é possível encontrar os seguintes factores de correcção: 1,0016, 1,0101, 1,0034, 1,0059, 1,0100, 1,0022 e 1,0015.

F4.4. Análise crítica de resultados e, eventualmente, repetição do ciclo.

Para o desenvolvimento desta tese apenas será analisado o R2 entre a variável final estimada e a variável objectivo – de facto para o objectivo geral deste projecto interessa sobretudo avaliar a capacidade de previsão dos modelos. Assim tem-se os seguintes coeficientes R2: 0,98069; 0,86545; 0,92327; 0,93014; 0,93484; 0,9941; e 0,96843 – valores muito elevados e que indicam uma excelente capacidade de ajustamento.

Os restantes indicadores de qualidade habitualmente calculados numa análise de regressão poderão ser aplicados, mas estão fora de âmbito. Em qualquer caso é interessante comparar o modelo estimado com o modelo observado.

Figura 4.5 - Comparação do modelo SUR com os dados originais



Fonte: autor

O apuramento do modelo final será efectuado com recurso à matriz *W* apurada em F4.3 e à equação 35, que por uma questão de conveniência abaixo se repete:

$$\widehat{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad (35)$$

5. Apuramento das preferências dos clientes a diferentes ofertas

5.1. Introdução

O objectivo deste capítulo é o de desenvolver uma abordagem que indique a elasticidade-preço dos clientes a diferentes conceitos de seguro, de forma a observar e encontrar formas de incluir a reacção dos clientes (elasticidade e função reacção/optimização de resultados) face às melhores ofertas.

Subjacente a este estudo de elasticidade está uma estratégia de auscultação directa ao consumidor, através da aplicação de um questionário. Neste capítulo será apenas tratado a metodologia estatística de análise de informação, deixando para trás questões de amostragem e de organização de questionários. A parte estatística onde se modeliza a elasticidade terá por base *conjoint analysis*.

Uma outra possibilidade de estudo das elasticidades poderia ser analisada com recurso à metodologia de Guven & McPhail (2013). Porém, a estratégia destes autores implica um custo excessivo de recolha de informação – de "pagar para ver" e apenas modeliza o factor preço, considerando tudo o resto um bem indiferenciado apenas corrigido com um ou dos parâmetros.

5.1.1. Apuramento das preferências: visão da microeconomia

Em economia, de forma a compreender as escolhas dos consumidores, é usado o conceito de 'utilidade'. Este conceito traduz assim uma representação de preferências de bens e serviços e que traduz de que forma o cliente racional faz as suas escolhas[22]. Dá-se assim um pendor racional à decisão do cliente, já que este escolhe sempre a opção que lhe traz maior 'utilidade'. Neste paradigma de racionalidade perfeita, se porventura o cliente não opta pelo produto que lhe confere maior utilidade ou bem-estar, assume-se que não está na posse de toda a informação. É ainda habitual assumir ainda que a utilidade pode ser decomposta

[22] Com algum abuso de linguagem, neste documento usar-se-ão os termos 'preferência' ou 'utilidade' de forma indiferente.

em pequenos atributos (ver Neumann e Morgenstern[23]) que compõem o próprio produto.

Uma forma eficiente de apurar quais as utilidades dos clientes é aplicar o princípio da preferência revelada. Este método, desenvolvido por Paul Samuelson em 1938, assume que as preferências dos consumidores podem ser reveladas pelos seus hábitos. Samuelson num dos seus muitos trabalhos seminais enunciou o *Weak Axiom of Revealed Preference* ao dizer: "*if an individual selects batch one over batch two, he does not at the same time select two over one*." Ou seja, o produto 'um' é preferido ao produto 'dois'; ou no mínimo 'um' é tão bom quanto o produto 'dois'. Este teorema revolucionou a forma de analisar o comportamento do consumidor: deixou-se de estudar apenas qual a escolha de produtos a um dado orçamento. Porém, e só em 1967, Sydney Afriat com um conjunto finito de preços e de escolhas de consumidores tentou construir uma função de utilidade que fosse consistente com as escolhas observadas. O processo de investigação foi algo heterodoxo, contra-intuitivo, afirma Varian (2006) e por isso a maior parte dos analistas não reconheceu o seu valor. Além disso, continua este autor: "*Afriat's exposition was not entirely transparent*." Ainda assim, alguns dos mais importantes microeconomistas da segunda metade do século XX foram-se dedicando a este tema (p.ex: Sydney Afriat, Andreu Mas-Colell, Walter Diewert, Paul Samuelson (Prémio Nobel)) com o objectivo de responder a (Varian 2006):

- Consistência: Quando é que o comportamento observado é consistente com a maximização de utilidade?
- Forma: Quando é que o comportamento observado é consistente com uma forma particular de utilidade?
- Recuperação: Como recuperar um conjunto de funções de utilidade que são consistentes com um conjunto dado de escolhas?
- Previsão: Como prever qual a procura face um novo orçamento?

[23] Neumann, John von e Morgenstern, Oskar Theory of Games and Economic Behavior. Princeton, NJ. Princeton University Press. 1944, sec.ed. 1947, th.ed. 1953.

5.1.2. Apuramento das preferências: visão do marketing

Do lado dos *marketeers* um desafio semelhante surgiu, mas seguindo uma linha diferente de análise. Neste campo de saber, não se pretendia saber qual o cabaz escolhido de um determinado consumidor face a um determinado orçamento, nem determinar a utilidade subjacente a um produto. Tratava-se de apurar as preferências de um produto com base na decomposição dos seus atributos (preço, marca, forma, por exemplo).

Dificilmente se consegue perguntar e ter uma resposta válida sobre quais as preferências de um cliente alvo, tendo em conta a avaliação individual dos atributos que compõem o produto. Sem um sistema de penalização entre atributos, sem uma análise conjunta de atributos, inevitavelmente, os clientes irão indicar os melhores níveis – o 'nirvana'.

Hauser & Rao (2003) indicam que "*[h]e [Paul Green] sought a means to decompose consumer preferences into the partial contribution (partworth) of product features. In this manner, researchers could not only explain the preferences of existing products, but could simulate preferences for entirely new products that were defined by feature combinations*".

Apesar de ser um caminho alternativo à microeconomia, há aqui um princípio de preferência revelada, mais uma vez. "*Conjoint measurement has psychometric origins as a theory to decompose an ordinal scale of holistic judgment into interval scales for each component attributes. The theory details how the transformation depends on the satisfaction of various axioms such as additivity and independence*". Apesar de seguir uma corrente autónoma, Green seguiu uma abordagem semelhante ao da microeconomia explicada no início deste capítulo.

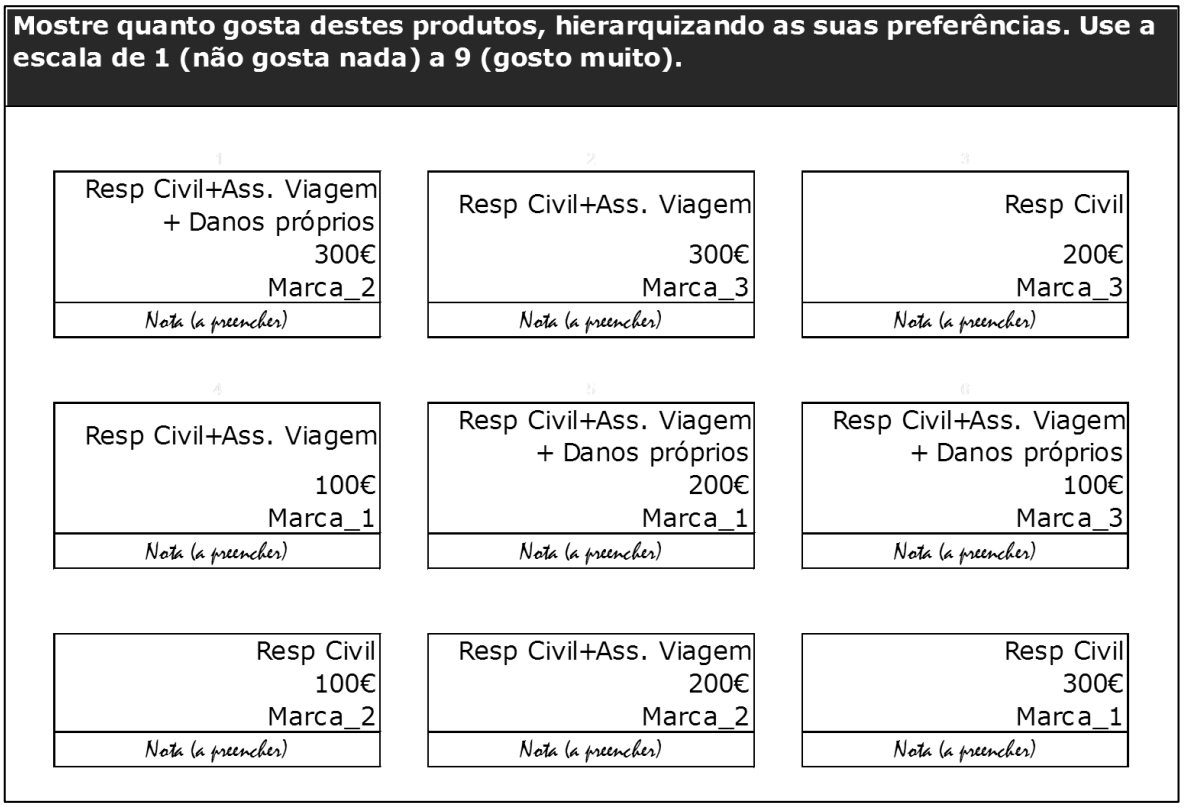
Em qualquer caso, no marketing e na microeconomia, o princípio de preferência revelada é uma forma prática de suportar o conceito de utilidade, que é uma assunção extraordinariamente forte e difícil de justificar.

Assim, e de uma forma simples, a análise conjunta (*conjoint analysis*), baseia-se nos pressupostos de utilidade, escolhas racionais, utilidade por atributos e preferências reveladas. De facto, Green & Wind (1975) no seu trabalho seminal, baseado em *Carpet Cleaners* (!), assumem que a preferência (y) é modelada em

função dos atributos (x) que o definem, dando origem a uma relação geral do tipo: $y = f(x)$.

Recorde-se que a *conjoint analysis* é um sistema é muito útil por forma a avaliar a sensibilidade ao preço e a diferentes características do produto. A ideia é obrigar os indivíduos a fazerem escolhas entre diferentes produtos criteriosamente escolhidos, numa situação em tudo semelhante à existente na vida real. A imagem seguinte apresenta uma possível formulação dos produtos que os clientes têm de avaliar, atribuindo uma classificação (*score*) ou ordenando os possíveis produtos (*rank*).

Figura 5.1 - Possível questionário de conjoint analysis aplicado ao contexto segurador (auto)



Fonte: autor

Com isto, e de uma forma simples, os indivíduos revelam as suas preferências e as características do produto que mais valorizam.

5.2. Necessidade de aplicar desenhos experimentais

A forma mais acessível de recolha de dados no contexto de análise conjunta, processa-se usando o método de *full profile* – ou seja, combinando todos os factores-níveis possíveis. No trabalho pioneiro, Green & Wind (1975), um *Carpet Cleaner*, o produto pelo qual se queria apurar as elasticidades, foi decomposto em *Package design, Brand name, Price, Good house keeping seal* e *Money-back guarantee*, sendo que foi possível construir 18 produtos distintos, que os consumidores tiveram de valorizar.

No caso de seguros, se se tiver 4 factores, cada um deles com 4 níveis é possível ter 256 ($4 * 4 * 4 * 4 = 256$) produtos que os clientes têm de avaliar. Porém, é difícil acreditar que qualquer entrevistado (*i.e.* potencial consumidor), consiga classificar ou ordenar 256 produtos, sobretudo num questionário onde não há contrapartida ou verdadeira venda.

Contudo, é necessário ver que esta comparação completa, de 256 produtos, só tem interesse quando se acredita que além dos efeitos aditivos na construção da utilidade, há igualmente efeitos interactivos de impacto significativo. De outra forma, e seguindo Gonçalves e Reis (2000), há que avaliar se a utilidade global resulta da adição da utilidade dos atributos (efeitos principais); ou se a utilidade global pode ser superior (ou inferior) à soma da utilidade dos atributos.

Em *conjoint analysis* assume-se, sem grandes dificuldades, que apenas os efeitos principais são significativos e que não haverá efeitos de interacção, que a utilidade global resulta da adição da utilidade dos atributos. O modelo de utilidade a estimar será assim descrito pela equação abaixo:

$$y = \boldsymbol{f}(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) \quad (36)$$

onde y indica a classificação que o público-alvo deu aos produtos analisados (*scores* ou *rankings*), e $f_i(x_i)$ pode assumir qualquer formulação.

Habitualmente $f_i(x_i)$ assume uma forma linear, ou de ponto ideal ou de utilidades

parciais separadas (Green & Srinivasan (1990)). Estes conceitos podem ser caracterizados com as seguintes definições:

a) Linear: Os *scores* ou *rankings* y deverão estar linearmente relacionados com os factores;

b) Ideal: Relação quadrática; onde as preferências apresentam rendimentos crescentes ou decrescentes à escala conforme os níveis utilizados. Haverá assim um ponto ideal a partir do qual oferecer mais (ou menos) aos clientes poderá ser prejudicial.

c) Utilidades parciais discretas: Onde cada nível de um determinado factor apresenta uma relação discreta, ordenada ou não, com os *scores* ou *rankings* obtidos.

No desenho dos produtos há que que responder às seguintes perguntas:

- Qual o número de produtos (combinação de factores e níveis) a questionar?
- Quais os produtos (combinação de factores e níveis) a questionar?

Estas serão respondidas nas secções seguintes.

5.2.1. *Conjoint analysis*: número de produtos a questionar

Seguindo o mesmo exemplo ilustrativo, visto na secção anterior, quando se têm quatro factores, com quatro níveis cada um, tem-se $4 * 4 * 4 * 4 = 256$ produtos a avaliar – se se optar por uma combinação *full profile* e com utilidades parciais discretas.

Porém, o que se pretende é estimar a equação 36. Esta apenas exige 16 observações – 16 comparações/classificações de produtos, já que numa regressão, o número mínimo de produtos classificar (número de observações) deverá ser idêntico à soma dos níveis (k) menos 1.

Esta regra origina um desenho experimental bastante parcimonioso e passível de

ser aplicado. A questão que fica é de saber se este número garante eficiência nos estimadores do modelo tanto mais que não trará nenhum grau de liberdade, pelo que não haverá os habituais indicadores de qualidade do modelo ou possibilidade de questionar a qualidade das estimativas.

Note-se que o modelo a estimar será indivíduo a indivíduo, já que não há comparações individuais de utilidade. O modelo agregado que dará a elasticidade total será a soma das decisões que cada indivíduo elaborou. Este detalhe é muito relevante já que se poderia empilhar os resultados e fazer apenas uma estimativa para o individuo médio, e assim, conseguir os graus de liberdade necessário para fazer as avaliações de qualidade.

A tabela seguinte sistematiza as diferentes opções, bem como os custos a pagar em cada uma das opções e as principais vantagens.

Tabela 5.1 - Vantagens e desvantagens entre um desenho mínimo ou *full profile*

Designação	Descrição do limite e observações
Limite mínimo $n = k - 1$	Vantagens: Modelo parcimonioso relativamente fácil de aplicar através de uma técnica presencial de recolha de informação (PAPI – *Paper And Pencil Interview* ou *CAPI – Computer Assisted Personal Interview*) ou mesmo com a ajuda de uma plataforma informática (CAWI – *Computer Assisted Web Interview*). Desvantagens: será que o mínimo de observações garante uma estimativa eficiente do modelo? Não há possibilidade de avaliar as estimativas do modelo. Nem possibilidade de estimar elementos de interacção entre os diferentes factores.
Limite máximo	Este modelo compara todos os possíveis produtos combinando

Designação	Descrição do limite e observações
$n = (k_i - 1)^{\#\, factores}$	todos os níveis e atributos. Vantagens: Neste modelo é possível aplicar efeitos de interacção. Os estimadores serão, por definição, os mais eficientes, e apenas sujeitos ao erro amostral. Desvantagens: O volume de informação a recolher torna este modelo impossível de recolher.

Fonte: autor

5.2.2. *Conjoint analysis*: quais os produtos a questionar

Apesar de se puder escolher qualquer número de produtos a avaliar entre mínimo e *full profile*, opta-se habitualmente por um desenho ortogonal mínimo.

Por ortogonal, como visto no capítulo anterior, entende-se que não há correlação entre os factores. Do ponto de vista estatístico, a grande vantagem deste desenho experimental é que implica geralmente estimadores centrados e eficientes (variância mínima se o modelo estiver bem desenhado).

5.2.3. Abordagem metodológica: Escolha dos atributos e níveis

A abordagem metodológica de *conjoint analysis* é, na prática, idêntica à abordagem de desenho experimental visto no capítulo anterior. Há apenas uma diferença, é que na *conjoint analysis* a análise é 'esticada' de forma a fazer simulação.

Assim é possível seguir a abordagem associada a um desenho experimental clássico:

- Fase 1: Identificar os factores e níveis que definem o produto seguro;
- Fase 2: Desenho óptimo;
- Fase 3: Recolha de informação;
- Fase 4: Análise.

A identificação dos factores e níveis é habitualmente conseguida com *focus groups* junto dos clientes e demais *stakeholders*, ver novamente Barker (1994).

5.3. *Conjoint analysis*: Estimação

O objectivo no processo de estimação é conseguir apurar o modelo de utilidade descrito pela Equação 36; onde como variável explicada tem-se a ordenação (ou classificação) dos produtos-cartão e, como variáveis explicativas, os diferentes factores e níveis que caracterizam o produto.

O primeiro desafio está em especificar a formulação linear que explica a ordenação (ou classificação) dos produtos, já que, como já afirmado, $f_i(x_i)$ pode assumir uma forma linear, ou de ponto ideal ou de utilidades parciais separadas.

Esta equação pode ser estimada por mínimos quadrados, sendo que, como visto anteriormente, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Onde '$\mathbf{X}$' representa a matriz pelas "n" observações pelas "k" variáveis do modelo; '$\mathbf{y}$' é um vector que representa as 'n' hierarquizações feitas pelos indivíduos e β é um vetor com os "k" coeficientes a estimar do modelo.

É de notar que '$\boldsymbol{y}$' é, na maioria dos casos, uma variável ordinal já que traduz uma hierarquização dos diferentes produtos criados. Ainda assim, o processo de estimação mais utilizado acaba por ser a regressão linear. Nesse caso, sendo '$\boldsymbol{X}$' uma matriz ortogonal, tem-se que:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{I}\,\boldsymbol{X}'\boldsymbol{y} \qquad (37).$$

Ou seja, com recurso à Equação 37, os valores de $\widehat{\boldsymbol{\beta}}$ são muito fáceis de apurar.

Note-se que os valores a estimar podem ser calculados através de outros mecanismos além da regressão linear, mais ajustados à natureza da variável dependente, como o logit ordenado, ou *max diff*. Gonçalves e Reis (2000) indicam ainda regressão monótona MONANOVA e método de Johnson e técnicas de

programação linear – modelo LINMAP.

De facto é difícil sustentar que não sendo a variável dependente contúnua (os dados são contínuos) que o modelo seja estimado por Mínimos Quadrados ordinários. É de notar que do ponto de vista formal, o modelo regressão linear pode ser utilizado para uma variável dependente ordinal. O estimador continuará a ser centrado, se o modelo for bem especificado. O estimador de mínimos quadrados ordinais no caso de *conjoint* continuará a apresentar uma variância mínima, já que se está em presença de variáveis independentes ortogonais. Mas apenas com uma dose extra de certeza se pode concluir que o estimador é centrado e de eficiência mínima, já que a forma funcional e o modelo estimado é sempre incompleto. Adicionalmente é ainda de referir que apesar de tudo, os métodos de regressão linear são os mais usados (Gonçalves e Reis (2000), Aaker, Kumar Day (1998) e Hollensen, Schmidt (2006)) e os disponíveis automaticamente e na maioria dos pacotes de análise estatística de *marketing research*.

Por forma a aferir a qualidade dos resultados estimados o habitual é verificar a correlação de Spearman ou tau-b entre o modelo estimado e as hierarquizações estimadas para cada uma das observações. Eventualmente, colocam-se mais alguns casos em *holdout* que são usados para estimar a sua utilidade e aferir a capacidade de previsão.

5.4. *Conjoint analysis*: Interpretação

Tal como já referido, os resultados deverão ser estimados indivíduo a indivíduo. Há razões teóricas e matemáticas para esta prática. De um ponto de vista teórico, os economistas defendem que não há comparações interpessoais de utilidade. A escala de utilidade não tem de ser a mesma entre todos os indivíduos[24]. De um

[24] Num conexto totalmente diferente, de estudo de desigualdades e liberdade, Amartya Sen (Prémio Nobel) dá a seguinte parábola (Sen, Amartya (1999) Development as Freedom) para explicar as diferentes motivações por detrás de um bem público, e consequentemente caracterizar a utilidade: Dinu, Bishanno e Roginipall querem um emprego de jardineiro disponibilizado por

ponto de vista matemático, o que está em causa é a aplicação de médias no contexto de votações ou, mais formalmente, do paradoxo de Condorcet[25]. Mais formalmente, o produto A, pode ser em média preferido ao Produto B e ao Produto C, mas se os três estiverem em concurso simultaneamente, este fica em último lugar em termos de ordenação.

Habitualmente são produzidos os seguintes resultados por indivíduo, ou por uma questão de simplicidade os mesmos resultados são transformados/aplicados ao indivíduo médio, ou de individuo médio por *cluster*:

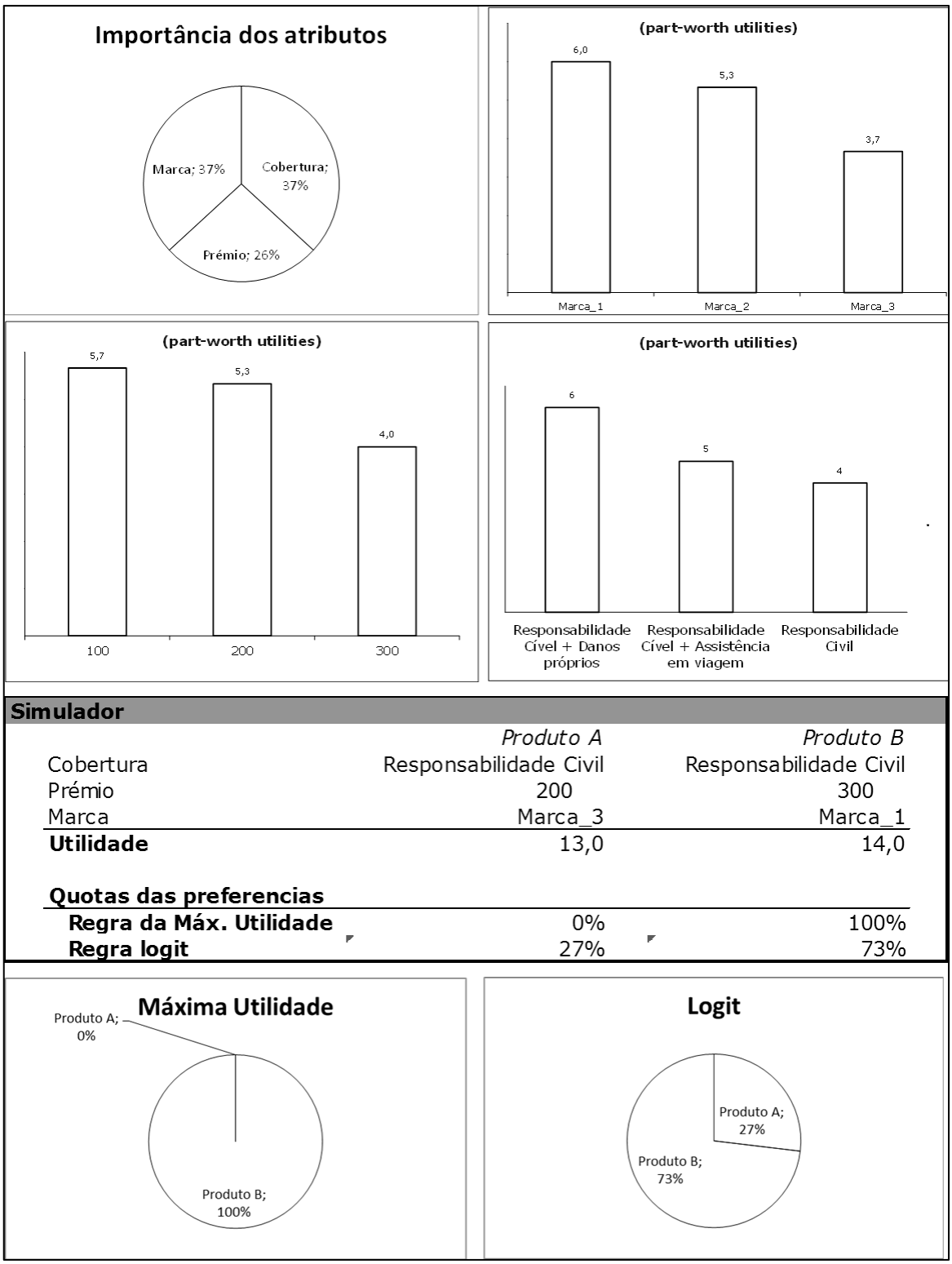
- Importância de cada factor;
- Gráficos de utilidade parciais;
- Simulador.

Alguns exemplos destes resultados são apresentados na figura seguinte. Seguidamente é apresentado o detalhe de cada um destes indicadores.

Annapurna e todos têm a mesma produtividade. A quem deve ser dado o emprego: ao que é doente? Ao mais pobre? Ou ao mais infeliz? "*Annapurna wonders what she really should do.*"

[25] Marie Jean Antoine Nicolas Caritat, Marquês de Condorcet foi um matemático e filósofo francês do século XVIII. O paradoxo de Condorcet explica-se com o facto de muitas vezes as preferências serem ciculares, pelo que a melhor solução não se consegue da agregação (média) das respostas. A sua obra basilar para este contexto é de 1785, e chama-se: *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.*

Figura 5.2 - Resultados de conjoint analysis (exemplo)



Simulador

	Produto A	*Produto B*
Cobertura	Responsabilidade Civil	Responsabilidade Civil
Prémio	200	300
Marca	Marca_3	Marca_1
Utilidade	13,0	14,0
Quotas das preferencias		
Regra da Máx. Utilidade	0%	100%
Regra logit	27%	73%

Fonte: autor

5.4.1. *Conjoint analysis*: Interpretação – Importância

Para o cálculo da importância é utilizada a seguinte fórmula:

$$\frac{\max(\beta_{ik})-\min(\beta_{ik})}{\sum_{k=1}^{K}(\max(\beta_{ik})-\min(\beta_{ik}))} \quad (38).$$

onde i = corresponde ao i-ésimo indivíduo

k = corresponde ao k-ésimo atributo.

A soma das importâncias relativas somará sempre 100%, o que facilita a interpretação da importância de cada atributo em relação à importância conjunta de todos os atributos. Contudo, é de notar que esta importância está muito dependente da amplitude dos atributos. Se um atributo tiver um nível com um muito baixo (ou alto) valor de utilidade associado, devido a um desenho experimental desequilibrado, a importância de um nível fica inflacionada.

Este cálculo é feito por indivíduo. Para a totalidade do mercado, habitualmente faz-se a média dos coeficientes obtidos.

5.4.2. *Conjoint analysis*: Interpretação – Utilidades parciais

As utilidades parciais são facilmente calculadas pelos valores dos coeficientes associados a cada um dos factores-níveis estimados da forma descrita. Este cálculo, nunca é demais salientar, é feito por individuo. São calculados tantas utilidades parciais como o número de inquiridos.

Note-se que estes coeficientes individuais podem ser usados para formar clusters hierárquicos e desta forma segmentar o mercado pela forma como os clientes se associam ao produto. Ou seja, é possível ter os clientes sensíveis ao preço, à marca, etc.

O *cluster* pode ser definido com base nos coeficientes da *conjoint analysis*. Ou seja, tentar-se agregar os indivíduos que valorizam os atributos de forma idêntica. Nesse sentido faz sentido normalizar os coeficientes das variáveis discretas para que todos os números esteiam na mesma escala. A análise de um *cluster*

analysis sei fora do âmbito desta tese.

Guven & McPhail (2013), na sua abordagem de avaliação da elasticidade sugerem que os *clusters* sejam feitos com base em análise empírica ou com recurso a árvores de decisão, sendo no entanto difícil argumentar qual deverá ser a variável objectivo.

5.4.3. *Conjoint analysis*: Interpretação – Simulador

Uma grande vantagem deste modelo é que as utilidades parciais podem ser combinadas com os diferentes atributos podendo ser gerado um simulador de utilidade. Na prática trata-se de aplicar a equação estimada, assumindo que y é o indicador de utilidade, ou seja:

$$\hat{y}_{produto\ específico} = \hat{\beta} X_{produto\ específico}.$$

A aplicação desta fórmula está na Figura 5.2, na secção de Simulador onde a cada produto é calculada a utilidade.

A questão seguinte é a de encontrar um algoritmo que explique como é que um individuo escolhe entre dois (ou mais) produtos concorrenciais. Escolherá o cliente sempre o produto com maior utilidade? Será o cliente assim tão racional?

A forma mais habitual de calcular a procura de um produto é assumir que os clientes optam sempre pelo produto com maior utilidade (*D Max*). Para cada indivíduo haverá assim um concurso entre os diferentes produtos possíveis, onde haverá apenas um produto vencedor. De um ponto de vista matematizado e com dois produtos a concurso, um indivíduo terá as utilidades $\hat{y}_A$ e $\hat{y}_B$:

$$\hat{y}_A = \hat{\beta} X_A.$$

$$\hat{y}_B = \hat{\beta} X_B.$$

sendo que X_A e X_B são as características chave dos produtos A e B. Se, por hipótese, $\hat{y}_A > \hat{y}_B$ então, para este indivíduo 100% das preferências irão para o produto A. A quota de preferência será calculada quando se fizer o saldo para todos os indivíduos analisados numa amostra representativa.

E se os clientes não forem racionais? A atribuição de quotas está muito dependente deste critério e produz resultados muito extremos (Orme & Baker (2000)). A questão intensifica-se, já que será necessário encontrar uma distribuição paramétrica para definir a irracionalidade.

Uma das soluções populares é trabalhar com a distribuição logística, assumindo implicitamente que a irracionalidade segue uma distribuição próxima da normalidade. A probabilidade de escolha será conseguida através do valor da utilidade de um produto, na distribuição logit, sobre a soma do logit de todas as utilidades de todos os outros produtos disponíveis. De um ponto de vista matematizado e usando a formulação de utilidade definida acima calcula-se a quota:

$$Quota\ de\ preferência\ do\ produto\ A = \frac{\exp(\hat{y}_A)}{\exp(\hat{y}_A) + \exp(\hat{y}_B)}$$

$$Quota\ de\ preferência\ do\ produto\ B = \frac{\exp(\hat{y}_B)}{\exp(\hat{y}_A) + \exp(\hat{y}_B)}$$

Se se desejar, estas quotas de preferências podem ser calibradas de forma a que o rácio máximo/mínimo atinja um determinado valor. Tal consegue-se multiplicando as utilidades parciais por uma constante positiva. Uma 'grande' constante tornará as previsões mais extremadas; uma 'pequena' constante tornará as previsões das quotas quase iguais.

Contudo, apesar deste modelo parecer mais interessante, apresenta também forte debilidades. Algo que Orme & Baker (2000) designam como "IIA *problems*",

isto é: *independence from irrelevante alternatives*. Tal pode ser visto com duas situações concretas:

1. Qualquer produto, mesmo que seja completamente deslocado das preferências consumidores terá sempre uma utilidade e nesse sentido uma quota de preferência.
2. Quando dois produtos estão muito próximos, a preferência de um desses produtos fica valorizada. Orme & Baker (2000) apresentam o seguinte exemplo:
    - Imagine-se que se tem dois produtos com as características apresentadas na tabela seguinte:

Tabela 5.2 - Cálculo de preferência para o produto A e B

Produtos	Utilidade	Exp (Utl)	Quota de preferência
A	0	1,00	27%
B	1	2,72	73%

Fonte: adaptado de Orme & Baker (2000).

Se se criar um produto (A') idêntico a A, tem-se os resultados apresentados na tabela seguinte

Tabela 5.3 - Cálculo de preferência para o produto A, A' e B

Produtos	Utilidade	Exp (Utl)	Quota de preferência
A	0	1,00	21%
A'	0	1,00	21%
B	1	2,72	58%

Fonte: adaptado de Orme & Baker (2000).

Ou seja, duplicou-se, sem justificação, a quota associada ao produto A; quando se estava à espera era de ter 50%-50% para A e A', de 27%.

Apesar de tudo, há um efeito de apoio quando dois produtos muito semelhantes são apresentados no mercado – estes ajudam-se a vender a si próprios. Mas é difícil de justificar estes valores quando se tratam de dois produtos iguais.

Um terceiro método pode ser empregue para calcular as quotas de preferência e simultaneamente combater a irracionalidade dos potenciais clientes, e IIA. Com a *Randomized First Choice*, proposta pelos Orme & Baker (2000), pretende-se calibrar as quotas propostas associando às utilidades estimadas um coeficiente de perturbação. Recuperando o caso a formulação matemática anterior para o apuramento das utilidades, tem-se para cada produto i:

$$\hat{y}_i = \hat{\beta} X_i \; + \textit{perturbação aleatória de produto}.$$

Mais, estes especialistas de *conjoint analysis* sugerem igualmente a inclusão de um elemento de perturbação por atributo, definindo um novo vector de *partworth utilities* $\hat{\hat{\beta}}$ como:

$$\hat{\hat{\beta}} = \hat{\beta} + \begin{bmatrix} pertubação\ de\ atributo \\ \dots \\ pertubação\ de\ atributo \end{bmatrix}.$$

As quotas podem ser quantificadas com a média das diferentes simulações. Este último método pode também ser calibrado de forma a conseguir-se resultados entre as quotas máximos e mínimas.

É de referir que o simulador de quotas (seja qual for o método usado) tem de assumir que as restantes variáveis do *marketing-mix* e que compõem a proposta de valor do produto, como a distribuição, a publicidade, o acesso, etc. são idênticos para todos os produtos em análise. Assim, para retractar esta limitação, diz-se que as quotas estimadas retractam antes quotas de preferências, por oposição a quotas de venda ou de mercado como até aqui têm vindo a ser

designadas.

Por fim, para determinar a curva de procura (relação entre preço e quotas de preferência) é necessário determinar vários as características que compõem os produtos das congéneres. Seguidamente, para o nosso produto alvo, deverá fixar-se um preço (dentro do intervalo estudado) e obter as quotas de preferência individuais. Fazendo uma média das quotas preferenciais é possível determinar a quota agregada para um determinado preço. Repete-se o exercício para mais preços. Desta forma, para cada indivíduo será possível calcular uma *switch rate* entre produtos; de forma agregada é possível obter uma curva de procura.

5.5. Caso prático

A análise de *conjoint analysis* não é mais do que um desenho experimental, onde a uma variação controlada dos *inputs* (produtos seguros, neste caso) se mede a reactividade do cliente (neste caso pela ordenação de cartões). Assim, tal como visto no capítulo anterior, a abordagem metodológica para responder a este desafio deverá seguir a abordagem clássica de desenhos experimental.

- Fase 1: Identificar os factores e níveis que definem o produto a avaliar;
- Fase 2: Desenho óptimo;
- Fase 3: Recolha de informação;
- Fase 4: Análise.

A figura seguinte ilustra a abordagem efectuada neste exercício. É de notar que este é um exercício de simulação, assumindo vários cenários para a quota de preferência (D Max: clientes racionais e regra logit); sendo que a curva de procura foi estimada apenas com base na regra D Max (clientes racionais).

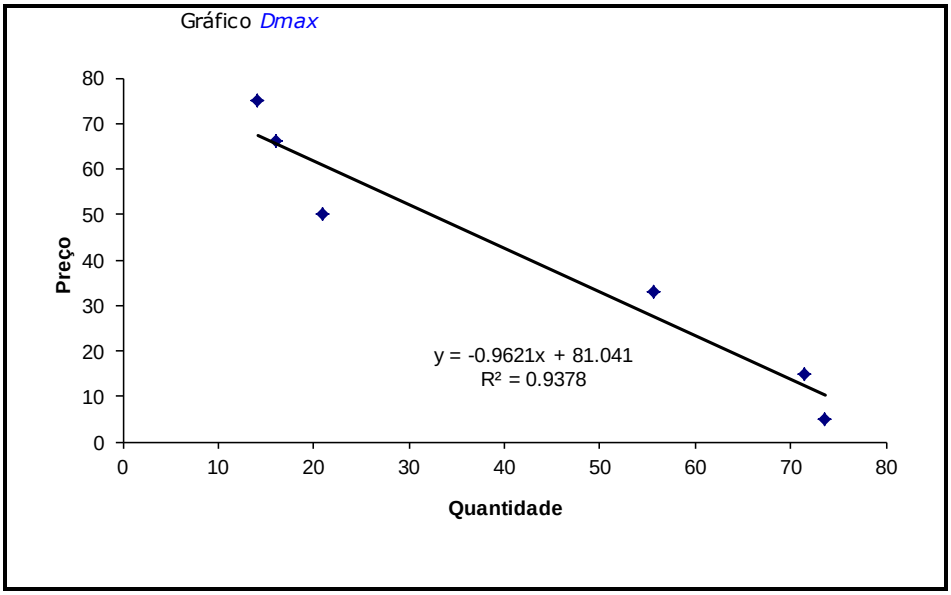
Figura 5.3 - Apresentação de um simulador de curva de procura agregado

	Produto A	Produto B	Produto C	Produto D	Produto E
Atributo_1	*Os atributos dos produsos A a E têm de ser definidos com base nas caracteristicas do mercado. O produto A é o nosso produto alvo cuja a variação de preço pode ser estudada.*				
Atributo_2					
Atributo_3					
Atributo_4					
Atributo_5					

Quotas de mercado	Regra da máxima utilidade #	Regra da máxima utilidade %	Regra Logit %
Produto A	232	55.5	45.8
Produto B	53	12.7	18.5
Produto C	10	2.4	7.9
Produto C	42	10.0	11.8
Produto E	81	19.4	16.1
Total	418	100.0	100.0

Curva da procura do produto A*	€	Dmax	Dlog
Preço máximo	*75*	14	44
	66	16	34
	50	21	40
	33	56	60
	15	71	73
Preço mínimo	*5*	74	75

* mantendo tudo o resto igual



Fonte: autor

6. Optimização

A mensagem chave deste capítulo é demonstrar que custo, valor e preço são algo distintos e que essas diferenças podem ser maximizadas no interesse das companhias.

Assim, tem-se que o 'custo' é o limite mínimo que o cliente tem de pagar, o 'valor' é o limite máximo que o cliente pode pagar, o preço é o valor que o cliente vai pagar. Evidentemente, que o desejável, do ponto de vista da companhia, é estabelecer um preço "próximo" ao valor que o cliente pode pagar, por oposição a um preço "próximo" dos custos. O preço final dependerá da decisão estratégica da companhia:

- Maximização da utilidade;
- Maximização da receita;
- Maximização dos lucros;
- Maximização da quota de preferência;

sendo certo que neste projecto só será estudada a maximização dos lucros e de receita.

É de notar que estes desafios de maximização podem ser complexificados uma companhia pode ter (e habitualmente tem) vários produtos para cada *tariff cell*.

6.1. Maximização dos lucros e de receita com um produto por *tariff cell*

A forma mais simples de compreender o processo de optimização resume-se à compreensão da tabela abaixo. Esta tabela indica por segmento, qual o prémio técnico (que deverá coincidir com os custos que a companhia incorre); os prémios praticados pelo mercado; as quotas de preferência para três cenários de preço e por fim duas soluções de *pricing* vistas no capítulo anterior (preço que maximiza a receita e o preço que maximiza o lucro).

A segmentação de mercado assume que um cliente não pode cruzar *tariff cells* – por definição.

De facto, a *tariff cell* é definida pelas características do bem seguro, do segurado e de contexto. Por absurdo, não é plausível que para reduzir o prémio de seguro, um cliente altere a cilindrada de um carro, mude de sexo ou altere as condições socioeconómicas do local onde mora.

Do lado da seguradora, também não há muita possibilidade em ter subsidiação cruzada entre *tariff cells*. Se uma seguradora agregar linhas de segmentação, outra companhia poderá discriminar o mercado e focar-se apenas nos clientes que convivem abaixo da média. Tal política terá dois efeitos, a primeira companhia terá os clientes com maior risco face ao desejado (selecção adversa), incorrendo porventura em prejuízo, e a segunda companhia terá uma posição mais firme no mercado.

Assim, pelas suas condições de sociodemográficas, contexto, e características do bem seguro um cliente está catalogado e a oferta que tem disponível resume-se à oferta que as outras companhias também oferecem para a mesma *tariff cell*.

Os prémios praticados pelo mercado são conhecidos e foram conhecidos com base na metodologia apresentada no capítulo 4.

Tabela 6.1 - Tabela síntese das condições de mercado

Segmento de risco/ *Tariff cell*	**Prémio de risco (custos)**	**Prémio Cia A**	**…**	**Prémio Cia Z**	**Quota pref. – P_{High}**	**Quota pref. – P_{M}**	**Quota pref. – P_{Low}**	***Profit max. price***	***Revenue max. price***
1									
2									
….									
m-ésimo									

Fonte: autor

Por fim, para explicar completamente esta tabela, vale a pena definir alguns valores ‘preços-óptimos’ da curva de procura e que definem as últimas colunas da

tabela síntese.

Como já visto, a companhia está disposta a vender seguros a diferentes pessoas, mas o preço de cada apólice vendida dentro de cada *tariff cells* é o mesmo. Adicionalmente, considera-se que cada cliente que compra um seguro não pode revendê-lo a outros clientes[26]. Está-se assim nas condições em que os economistas designam por descriminação de terceiro grau.

Nestas condições, considere-se:

$$D^{-1}: p_i(q_i)$$

a função procura inversa associada a *tariff cell* i considerada, onde p indica o preço e q a quota de preferência. Como habitualmente, assume-se que a função procura inversa tem declive negativo: quanto maior o preço menor a quantidade (quota de preferência) consumida. Neste caso o problema de maximização de lucro, para todas as *tariff cells* de uma companhia, pode ser definido da seguinte forma:

$$\max_{q_i} Lucro = \sum_{i=1}^{m} \{p_i(q_i).q_i - c_i(q_i)\} \quad (39).$$

A solução óptima será, evidentemente, onde a receita marginal é igual ao custo marginal em cada *tariff cell*. De facto, se se notar nas Condições de Primeira Ordem (C.P.O.) tem-se:

$$\frac{\partial p_i}{\partial q_i}.q_i + p_i(q_i) = \frac{\partial c_i}{\partial q_i}.$$

Vale a pena analisar o que aconteceria se:

[26] Um cliente não pode revender um seguro, mas pode fazer o seguro em nome de um familiar...

$$c_i = c_j \, , \forall i, j.$$

Se o custo marginal fosse o mesmo em cada *tariff cell*, uma companhia seria indiferente em vender em qualquer uma das *tariff cells*, já que a receita marginal teria de ser a mesma. Mais, se $p_i > p_j$, então a semi-elasticidade do produto j ($\frac{\partial p_j}{\partial q_j}.q_j$) deverá ser maior do que a semi-elasticidade do produto i. Assim, o preço será mais elevado nos produtos onde a procura é mais rígida.

Em qualquer caso, se se definir um custo marginal constante em cada *tariff cell,* mas distinto entre *tariff cells*, e se se assumir que uma determinada formulação para a curva de procura é então possível definir o preço maximizador de lucro. Concretizando, sem perda de generalidade, que a procura linear: $p_i(q_i) = p_o - \beta.q_i$), tem-se duas equações:

$$\begin{cases} D^{-1}: p_i(q_i) = p_o - \beta.q_i \\ C.P.O: -\beta.q_i + p_i(q_i) = \frac{\partial c_i}{\partial q_i} \end{cases};$$

notando evidentemente que $\frac{\partial p_i}{\partial q_i} = \beta$.

Então tem-se que os valores de equilíbrio para cada *tariff cell* serão:

$$\begin{cases} {q_i}^* = \frac{\frac{\partial c_i}{\partial q_i} - p^*}{-2\beta} \\ {p_i}^* = Profit\ max.price = p_o + \frac{p_o - \frac{\partial c_i}{\partial q_i}}{\beta} \end{cases} \quad (40).$$

Estas equações de equilíbrio assumem que uma seguradora não tem problemas de oferta, que tem capacidade de abastecer todo o mercado aquele custo

marginal – pode não ser assim, já que o resseguro ou o apetite pelo risco da seguradora pode bloquear algumas estratégias comerciais mais agressivas.

Usando todo o mesmo racional, podemos dizer que o preço que maximiza a receita é dado quando o custo marginal é zero, ou seja:

$$\begin{cases} {q_i}^{**} = \frac{p_o}{2\beta} \\ {p_i}^{**} = Revenue - maximizing\ price = \frac{\partial c_i}{\partial q_i} - \frac{p_o}{\beta} \end{cases} \quad (41)$$

Note-se que se está a partir do princípio que uma seguradora é tomadora de preço em relação ao mercado, que a sua estratégia de preço não influência a estratégia de preço das restantes companhias. Se a variação de preço da seguradora líder ou do restante mercado for racional é possível calcular as funções reacção e as curvas de iso-lucro que maximizam os resultados da 'nossa' companhia. O modelo de Stakelberg[27] e a sua abordagem é especialmente relevante neste caso. Contudo, no caso em apreço, a função reacção pode ser calculada fazendo variar na tabela síntese o preço da companhia líder (ou das restantes), calculando o preço óptimo em conformidade.

6.2. Maximização dos lucros e de receita com dois (ou mais) produtos por *tariff cell*

As seguradoras apresentam muitas vezes vários produtos para a mesma *tariff cell*. No caso automóvel tem-se um produto de responsabilidade cível e assistência em viagem como o produto básico, e um produto de danos próprios que além das coberturas indicadas, protege igualmente o automóvel próprio e eventualmente os passageiros e condutor. A tabela síntese duplicará assim de

[27] Heinrich Freiherr von Stackelberg (1905 – 1946) foi um economista alemão que estudou os movimentos de preço entre empresas líderes e seguidoras, dando um contributo para a Teoria dos Jogos, economia industrial e a microeconomia em particular. O modelo de Stackelberg que aqui se fala pode ser encontrado num qualquer manual de microeconomia de nível intermédio.

tamanho já que haverá mais colunas analisadas. Mas fora isso, a sua estrutura e racional de análise não serão alterados.

No concreto, a determinação dos preços óptimos no entanto será alterada, já que a definição de preço de um produto irá canibalizar a oferta do outro produto. Assim, as funções procura inversa deverão ser definida atendendo também ao preço do outro(s) produtos. Se se assumir que a procura é linear e apenas para uma *tariff cell* tem-se:

$$D_{low}{}^{-1}: p_l(q_l, q_h) = p_{o,l} - \beta_l q_l - \alpha_l q_h$$

$$D_{high}{}^{-1}: p_h(q_l, q_h) = p_{o,h} - \beta_h q_h - \alpha_h q_l$$

Assim, o desafio de definição de preço, para 2 produtos, pode ser dado:

$$\underset{q_{lz}, q_{hz}}{max\ Lucro} = \sum_{z=l,h} \sum_{i=1}^{m} \{p_{iz}(q_{iz}).\, q_{iz} - \ c_{iz}(q_{iz})\} \qquad (42).$$

O resultado deste problema não é tão elegante, mas a interpretação é a mesma: a receita marginal, agora ponderada pelas elasticidades cruzadas, tem de ser igual aos custos marginais. Assim para o caso das procuras lineares e apenas para uma *tariff cell*:

$$\begin{cases} q_l = p_{o,l} - 2\beta_l - (\alpha_l + \alpha_h) q_h - \frac{\partial c_l}{\partial q_l} \\ q_h = p_{o,h} - 2\beta_h - (\alpha_l + \alpha_h) q_l - \frac{\partial c_h}{\partial q_h} \end{cases}.$$

O que simplificado, se tem:

$$\begin{cases} {q_h}^* = \dfrac{(\alpha_l + \alpha_h)p_{o,l} + 2\beta_l\left(\dfrac{\partial c_l}{\partial q_l} + \dfrac{\partial c_h}{\partial q_h}\right) - (\alpha_l + \alpha_h)\dfrac{\partial c_l}{\partial q_l}}{{\alpha_l}^2 + 4(\alpha_l\alpha_h) - 4\beta_h\beta_l + {\alpha_h}^2} \\ {q_l}^* = \dfrac{(\alpha_l + \alpha_h)p_{o,h} + 2\beta_h\left(\dfrac{\partial c_l}{\partial q_l} + \dfrac{\partial c_h}{\partial q_h}\right) - (\alpha_l + \alpha_h)\dfrac{\partial c_h}{\partial q_h}}{{\alpha_l}^2 + 4(\alpha_l\alpha_h) - 4\beta_h\beta_l + {\alpha_h}^2} \\ p_l(q_l, q_h) = p_{o,l} - \beta_l q_l - \alpha_l q_h \\ p_h(q_l, q_h) = p_{o,h} - \beta_h q_h - \alpha_h q_l \end{cases}$$

6.3. Caso prático

Para melhor compreender a aplicação deste conceito é usada a mesma Base de Dados que serviu de suporte ao capítulo 4. Note-se que com o trabalho efectuado nesse capítulo é possível determinar o preço de todas as congéneres em todas as *tariff cells*, quer o preço tenha sido inquirido ou não. Os valores apresentados para a companhia 1, descontados de 30%, serão usados como os associados aos custos totais da "nossa" companhia.

Há no entanto uma decisão a fazer, deverá ser usado o preço estimado ou, se existir, o preço observado, para o preenchimento dos prémios das congéneres? Tendo em conta que o modelo estimado elimina o efeito de *outliers*, e 'espalha' o erro de observação ao longo de todas as *tariff cells* optou-se por usar o modelo estimado.

A elasticidade de cada uma das *tariff cells* foi obtida com recurso às técnicas do capítulo 5, mas aplicadas a um contexto de retalho alimentar. Mais uma vez o detalhe de recolha não será dado. Como ambos os exercícios exigiam marca, houve um emparelhamento das mesmas, sem grande critério. Houve necessidade de abdicar de uma das companhias do capítulo 4, já que no contexto de retalho alimentar foram consideradas apenas 5 insígnias. As restantes características de

retalho alimentar foram consideradas idênticas para todos os produtos a concurso.

Pretende-se com este caso prático construir da tabela síntese e determinar o preço óptimo e da quota de preferência associada apenas a uma das *tariff cells* (#1). Alguns indicadores de rendibilidade serão construídos, bem como a aplicação da regra de que os custos terão de ser sempre menores do que os preços a cobrar.

Assim, os dados que se tem são dados pela tabela abaixo.

Tabela 6.2 - Condições de mercado para a Tariff cell 1

Segmento de risco/ *Tariff cell*	**Custos técnicos**	**Companhia 2**	**Companhia 3**	**Companhia 4**	**Companhia 5**	**Companhia 6**
ID=1	48 515	56 260	55 331	75 220	61 882	114 736

Fonte: autor

O simulador para a *Tariff cell* 1 pode ser visto na tabela abaixo. Note-se que aqui, havendo apenas um produto, assumiu-se a regra Logit para a definição de quotas. A curva de procura inversa foi assumida com base nos 5 pontos calculados. Neste caso, a intercepção (p_0) assume o valor de 64354 e o declive é de -253, fazendo com que a procura inversa possa ser escrita como:

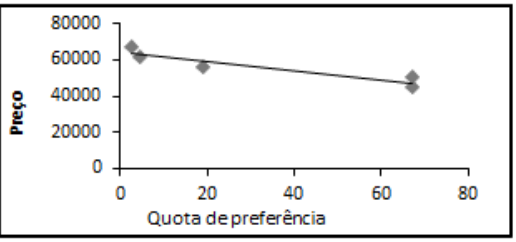
$$D^{-1}: p_i(q_i) = 64354 \ - 253.q_i.$$

O cálculo do *Profit-maximizing price* e do *Revenue-maximizing price* foi elaborado com base nas equações 40 e 41. Porém, o *Revenue-maximzing price* implica uma quota de 127%... para um preço de 48769. Trata-se de um valor evidentemente não credível. Assume-se então que *Revenue-maximzing price* deverá ser entre os valores testados o que maximiza a receita. Ou seja, 50634.

Os resultados para uma única *tariff cell* pode ser visto na tabela abaixo. O resultado para a seguradora deveria ser replicado para todas as *tariff cells*.

Tabela 6.3 - Condições de mercado para a Tariff cell 1

Cenários de preço	Preço #	Regra Logit %	Elasticidade Logit
P. elevado	67 512	2.4	-9.4
P. medio alto	61 886	4.3	-37.1
P. médio	56 260	19.0	-25.3
P. medio baixo	50 634	67.0	0.0
P. baixo	45 008	67.0	ND
Custo mínimo	48 515	Profit-max price	64 417
		Rev-max price	50 634



Fonte: autor

7. Conclusões

Pretendia-se, com este projecto, fornecer uma base de trabalho geral para desenhar os preços de seguros não-vida que levasse em linha de conta:

- o risco;
- o comportamento da concorrência;
- a elasticidade do clientes.

Os objectivos gerais do projecto foram assim conseguidos. Através de uma abordagem sistemática foi apresentada uma abordagem de análise de risco com modelos GLM, de apuramento das tarifas da concorrência com modelo SUR com correcção de ponto de massa, de avaliação da elasticidade dos clientes através de *conjoint analysis* e de optimização cruzando todas estas análises.

Os objectivos específicos deste projecto para um ramo não vida foram também atingidos. Vale no entanto a pena analisar este cumprimento com detalhe.

7.1. Compreender e explicitar, de forma breve, os principais mecanismos de formulação dos custos de seguro (com especial destaque para os modelos GLM)

No capítulo 3 houve um estudo aprofundado das principais técnicas de apuramento tarifário. Outras técnicas poderiam ser estudadas. Por exemplo, no cálculo da frequência a regressão logística é muito usada – sobretudo no sector bancário. A regressão exponencial negativa também tem igualmente os seus adeptos.

Para de alguma forma incluir o *contigency loading* um modelo tarifário com uma componente de *bootstrap* poderia ser igualmente usado.

No estudo dos diferentes tipos de regressão pretendeu-se trabalhar de forma muito cuidada a intuição só introduzindo no fim o estudo dos modelos de dispersão. A regressão Tweedie que de alguma forma agrega todas as equações mencionadas continua a ser alvo de estudo por parte da academia. A determinação do p em contexto de regressão linear (e posteriormente de desenho

experimental) precisa ainda de ser calibrada. A brevíssima programação proposta nesta tese é uma forma expedita de resolver a questão e um contributo para o estudo desta regressão.

7.2. Observar e encontrar formas incluir os preços das companhias congéneres no modelo tarifário de uma companhia

O objectivo principal do capítulo 4 era o de apresentar um método de recolha e captura do modelo tarifário de uma seguradora congénere, pelo que o objectivo foi cumprido. Assim foi:

- apresentada uma abordagem para o desenho amostral, baseada nos princípios de ortogonalidade - esse modelo é potencialmente mais eficiente em termos de custos do que uma recolha *full profile* (aliás, esta é na prática impossível de realizar);
- apresentado um modelo linear para fazer uma primeira análise (modelo de regressão linear com transformação Box-Cox com correcção de ponto de massa). Modelo este que deverá acompanhar razoavelmente o modelo de distribuição Tweedie de atribuição dos prémios;
- apresentado como integrar a informação de mais do que uma congénere, aumentando a eficiência dos estimadores através de um modelo SUR;
- ainda deixada a possibilidade de desenhar um modelo mais complexo por GLM – Tweedie, potencialmente com ainda maior aderência aos dados, indiciando como é que se poderá obter uma estimativa grosseira para o factor de dispersão p melhorando o processo de estimação.

Do ponto de vista académico será interessante avaliar oportunamente a generalização da abordagem SUR no caso de GLM; bem como é que o modelo Box-Cox pode contribuir para uma estimação eficiente da Tweedie – determinação de p.

Na sua forma actual, e do ponto de vista académico este documento sistematiza

algumas técnicas de *pricing* de seguros não vida.

7.3. Observar e encontrar formas incluir a reacção dos clientes (elasticidade e função reacção/optimização de resultados) face às melhores ofertas.

Os pontos seguintes, sistematizam as vantagens, desvantagens, hipóteses e limitações do método de *conjoint analysis*:

a) Vantagens

- Obriga as pessoas a fazerem escolhas em tudo semelhantes ao que acontece na vida real;
- É possível construir um simulador de forma a testar produtos inovadores ou da concorrência.

b) Resultados:

- Cálculo das quotas de preferências;
- Análise de sensibilidade ao preço, marca, etc.;
- Prever a elasticidade num ponto (switch rates) do "nosso" produto actual para novos produtos (canibalismo), ou para produtos da concorrência;
- Possibilidade de criar novos produtos (combinação de atributos) e obtenção da função reacção dos consumidores e da concorrência.

c) Hipóteses e limitações

- Parte do princípio que o que não é questionado não tem relevância (nomeadamente a distribuição e a promoção);
- É necessário decompor o produto em poucos atributos e níveis;
- Muito sensível ao desenho dos analistas;

- A resposta do público-alvo é por questionário e por isso não comprometida, isto é pode ser tendencialmente enviesada.

Oportunamente poderá ser estudado outros mecanismos de avaliação da elasticidade. Ainda no contexto de *conjoint analysis* o processo de estimação por *choice-based* (Logit) parece ser o mais eficiente (Orme & Baker (2000)). Contudo, o acesso a dados de compradores de seguros *online* poderá alterar o processo de recolha de informação e criar outros mecanismos de análise de elasticidade.

O processo de simulação de quotas de preferência por perturbação merecia um estudo mais aprofundado de forma a compreender, sobretudo, a magnitude desses choques e a sua real capacidade de previsão.

7.4. Encontrar formas possíveis de optimizar (de forma muito linear/simples) os resultados técnicos da companhia no ramo em estudo.

O objectivo principal do capítulo 6 era o de apresentar um método de optimização do produto seguro tendo em conta o risco, o comportamento da restante oferta a elasticidade dos clientes. O facto de o cliente não puder cruzar *tariff cells* facilitou em muito a análise já que tornou cada *tariff cell* num mercado único e com nenhuma influência dos restantes.

No entanto, a determinação dos preços finais deve ser feita com cautela. Apesar dos clientes não puderem cruzar *tariff cells*, a verdade é que mais tarde podem trocar de objecto seguro, ficarão mais velhos, etc.. Assim, uma harmonia entre as *tariff cells* poderá garantir uma maior sustentabilidade a prazo.

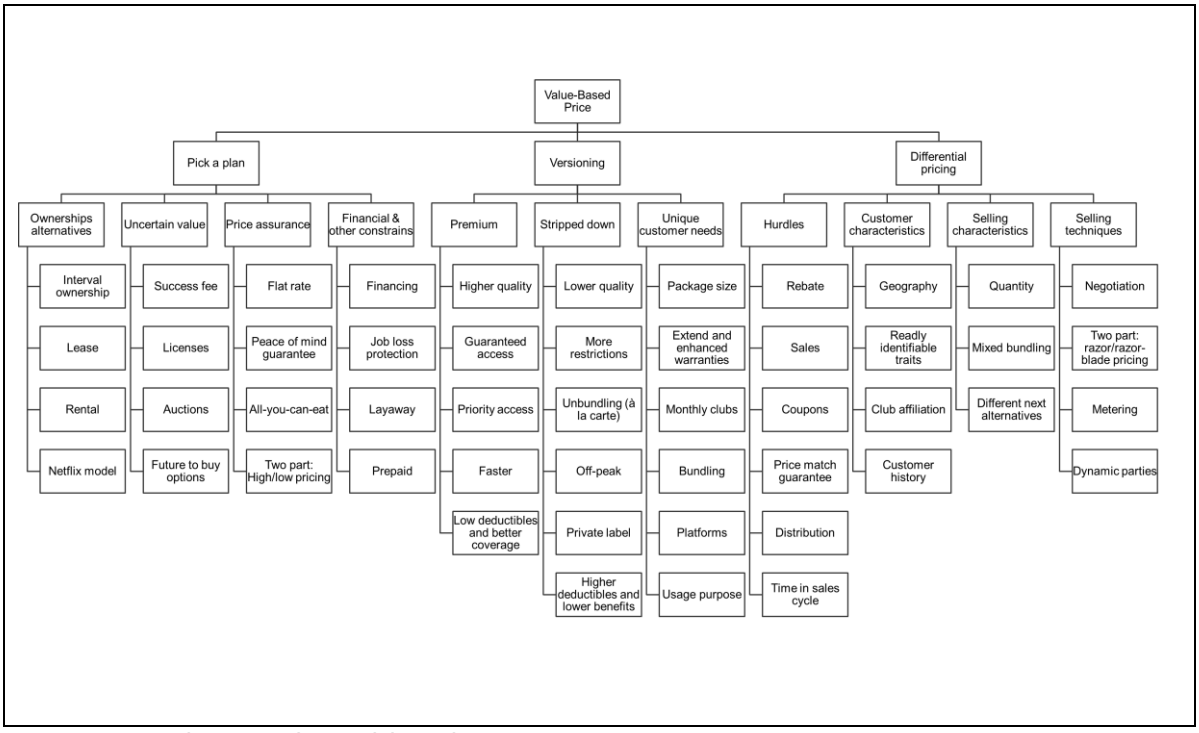
Também é de salientar que as equipas comerciais muitas vezes não conseguem vender às *tariff cells* mais adequadas. Para fins ilustrativos considere-se o caso em que há uma enorme margem de crescimento junto do mercado masculino já que público tem um risco menor que a média homem/mulher mas a equipa comercial apenas consegue vender junto do público feminino.

Além disso, por vezes não há dimensão nas *tariff cells*. Considere-se a seguinte ilustração: onde há para um preço baixo em Portalegre e um preço alto em Lisboa

– contudo não há clientes suficientes que no interior do país para compensar agravamento da tarifa na malha urbana. Estes dois casos, poderão obrigar a uma união de duas *tariff cells*, provocando uma subsidiação cruzada de risco.

Por fim, há ainda que salientar que o processo de *value based pricing* não se resume ao cálculo de elasticidades. Outras estratégias podem ser seguidas, conforme retratado na figura abaixo. Este projecto apenas retractou as de *versionaning- premium* e sobretudo de diferenciação de preço pelas características do cliente (*customer characteristics*). Outras estratégias de *value based pricing* poderão ser seguidas e eventualmente com mais sucesso. Ainda num contexto matematizado e de seguros, o *bundling*, e a distribuição poderão trazer resultados muito positivos e que deverão ser estudados, sobretudo na área seguradora.

Figura 7.1 - Estratégias de *value based pricing*



Fonte: Baseado em Mohamed (2010)

8. Bibliografia

AAKER, D., KUMAR, V.; DAY, G. (1998). Marketing Research. John Wiley & Sons.

ADDELMAN, S. (1962). Symmetrical and asymmetrical fractional factorial plans. Technometrics , 4: 47–58

ANDERSON, Duncan; FELDBLUM, Sholom; MODLIN, Claudine; SCHIRMACHER, Doris; SCHIRMACHER, Ernesto; THANDI, Neeza (2007). A Practitioner's Guide to Generalized Linear Models – a foundation for theory, interpretation and applicattion. 3.th ed. CAS Discussion Paper Program.

BAILEY, R.A.; SIMON, LeRoy J. (1960). Two studies in automobile insurance. ASTIN Bulletin, 192-217. Site: http://www.actuaries.org/LIBRARY/ASTIN/vol1no4/vol1no4.pdf e http://www.statsci.org/data/general/carinsca.html consultado em Fevereiro de 2015

BARKER, Thomas B. (1994). Quality By Experimental Design, 2 th. Chapman and Hall/CRC

BERRY, Michael J.; LINOFF, Gordon S. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons

BERRY, Michael J.; LINOFF, Gordon S. (2009). Oversampling in general. Blog:dataminers, site: http://blog.data-miners.com/2009/11/oversampling-in-general.html

BOX, G. E. P.; COX, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Statist. Soc.* B, **26**, 211-252. Site: http://www.ime.usp.br/~abe/lista/pdfQWaCMboK68.pdf consultado em Junho de 2013.

BOX-COX transformation. Encyclopedia of Mathematics. Site: http://www.encyclopediaofmath.org/index.php?title=Box-Cox_transformation&oldid=22177 , consultado em Junho de 2013

BROCKMAN, M J; WRIGHT, T S (1992). Statistical motor rating: making effective use of your data. Journal of the Institute of Actuaries [JIA] (1992) 119: 457-543. Site: http://www.actuaries.org.uk/research-and-resources/documents/statistical-motor-rating-making-effective-use-your-data consultado em Maio de 2013

CARDOSO, Margarida (2006). Modelos discriminantes lógicos na caracterização de uma estrutura de segmentos in Temas em Métodos Quantitativos 3 (Reis, Elizabeth; Magalhães Hill, Manuela (ed)). Edições Sílabo

CHARPENTIER, Arthur (2010). On Box-Cox transform in regression models. Blog: http://www.r-bloggers.com/on-box-cox-transform-in-regression-models/ consultado em Maio de 2015

COUTTS, S M (1984). Motor insurance rating, an actuarial approach, Journal of the Institute of Actuaries [JIA] 111: 87-148. Site: http://www.actuaries.org.uk/research-and-resources/documents/motor-insurance-rating-actuarial-approach consultado em Maio de 2015

DIAS, José Gonçalves; REIS, Elizabeth (2000). Análise Conjunta – Uma apresentação dos princípios in Temas em Métodos Quantitativos 1 (Reis, Elizabeth; Ferreira, Manuel (ed)). Edições Sílabo

DIXON, J. W.; MOOD, A. M. (1948). "A Method for Obtaining and Analyzing Sensitivity Data," Journal of the American Statistical Association, 43, 109–126

DONLAN, Brian M.; TURNACIOGLU, Theresa A. (2006). Introduction to Ratemaking Relativities - CAS Seminar on Ratemaking.

DROR, Hovav A.; STEINBERG, David M. (2005). Approximate Local D-optimal Experimental Design for Binary Response, Technical Report RP-SOR-0501, Tel Aviv University.

DROR, Hovav A.; STEINBERG, David M. (2006). Robust Experimental Design for Multivariate Generalized Linear Models, Technometrics Vol. 48, No. 4, 520-529 .

DROR, Hovav A.; STEINBERG, David M.(2008). Sequential Experimental Designs for Generalized Linear Models, Journal of the American Statistical Association, 103, 288-298.

DUNN, Peter K.; SMYTH Gordon K. (2005). Series evaluation of Tweedie exponential dispersion model densities. Statistics and Computing , Volume 15, pages 267–280 site: http://www.statsci.org/smyth/pubs/tweediepdf-series-preprint.pdf consultado em Junho de 2013

ELKAN, Charles (2001). Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000. Site: http://cseweb.ucsd.edu/users/elkan/kddcoil.pdf consultado em Junho de 2014

FELDBLUM, Sholom; BROSIUS, J.Eric (2002). The Minimum Bias Procedure A Practitioner's Guide. CAS. Forum. VolII page(s).591-684. Site: www.casact.org/pubs/forum/02fforum/02ff591.pdf consultado em Maio de 2013

FINLAY, Steven (2012). Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior. Palgrave Macmillan; 2nd edition edition

FONSECA, Jaime (1992). Introdução à estatística matemática aplicações – volume 1. Litografia Amorim.

GILLAM, William R.(1993). Injured Worker Mortality. Proceedings: 1993 Volume LXXX, Numbers 152 and 153. Site: http://www.casualtyactuarialsociety.com/pubs/proceed/proceed93/93034.pdf consultado em Maio de 2013

GREEN P.E., SRINIVASAN V. (1978), Conjoint Analysis in Consumer Research: Issues and Outlook "Journal of Consumer Research", September, 5, 103-123

GREEN, Paul E.; KRIEGER, A.M.; WIND, Yoram (Jerry). (2001) Thirty Years of Conjoint Analysis: Reflections and Prospects INTERFACES 31: 3, Part 2 of 2; pp. S56–S73

GREEN, Paul E.; SRINIVASAN, V. (1990) Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice, Journal of Marketing, Vol. 54, No. 4 (Oct., 1990), pp. 3-19, Published by: American Marketing Association Article Stable URL: http://www.jstor.org/stable/1251756

GREEN, Paul E.; WIND, Jerry (1975), “New Way to Measure Consumers’ Judgments,” Harvard Business Review, (July-August),107-117

GRIFFITHS, William; HILL, R. Carter; JUDGE, George (1993). Learning and Practicing Econometrics, John Wiley and Sons

GUJARATI, Damodar N. (1995). Basic econometrics; McGraw Hill, 3 ed

GUVEN, Serhat; MCPHAIL, Michael (2013). Beyond the Cost Model: Understanding Price Elasticity and its Applications. CAS E-Forum, Spring 2013. Site: http://www.casact.org/pubs/forum/13spforum consultado em Maio de 2013.

HAINES, L. M.; PEREVOZSKAYA, I.; ROSENBERGER, W. F. (2003). “Bayesian-Optimal Designs for Phase I Clinical Trials,” Biometrics, 59, 591–600

HALLIN, M.; INGENBLEEK, J.-F. (1983). The Swedish automobile portfolio in 1977. A statistical study. Scandinavian Actuarial Journal, 49-64.

HARRISON, M. J.; MCCABE., B. P. M. (1979) A Test for Heteroscedasticity Based on Ordinary Least Squares Residuals. Journal of the American Statistical AssociationVol. 74, No. 366, pp. 494-499. Site: http://www.jstor.org/stable/2286361 consultado em Fevereiro de 2015

HASSETT, M. J., STEWART, D.(2006). Probability for Risk Management, ACTEX Publications

HAUSER, John R.; RAO,Vithala R. (2003) "Conjoint analysis, related modeling, and applications." Marketing Research and Modeling: Progress and Prospects. Springer US, 2004. 141-168.

HAYASHI, Fumio (2001) Econometrics; Princeton University Press

HOLLENSEN, S.; SCHMIDT, M. (2006). Marketing Research: An International Approach. FT Prentice Hall.

ISP (2013), Guia de Seguros. Site: http://www.isp.pt/NR/rdonlyres/95A10CFE-C814-4B73-B0E9-CBBFD3D46B34/0/Guiawebuv.pdf consultado em Janeiro de 2013

IVANOVA, A.; WANG, K. (2004).“A Non-Parametric Approach to the Design and Analysis of Two-Dimensional Dose-Finding Trials,” Statistics in Medicine, 23, 1861–1870

JØRGENSEN, B. (1989). The Theory of Exponential Dispersion Models and Analysis of Deviance, Lecture notes for short course, School of Linear Models, University of São Paulo, 129 pages. Second Edition 1992: Monografias de Matemática #51, IMPA, Rio de Janeiro. Site: http://www.impa.br/opencms/pt/biblioteca/mono/Mon_51.pdf consultado em Fevereiro de 2015

JØRGENSEN, Bent (1987). Exponential dispersion models. Journal of the Royal Statistical Society, Series B 49 (2): 127–162.

JØRGENSEN, Bent; de SOUZA, M.P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. Scand. Actuarial J. 1994, 69-93

KARVANEN, J.; VARTIAINEN, J. J.; TIMOFEEV, A.; PEKOLA, J. (2007). "Experimental Designs for Binary Data in Switching Measurements on Superconducting Josephson Junctions," Applied Statistics, 56, 167–181

KHURI, André I; MUKHERJEE, BHRAMAR; SINHA, Bikas K.; GHOSH, Malay (2006). Design Issues for Generalized Linear Models: A Review, Statistical Science Statistical Science, Vol. 21, No. 3, 376–399, Site: http://www.stat.ufl.edu/~ufakhuri/pdf/STS159.pdf consultado em Maio/Junho de 2013

LONG, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Sage Publications.

MACCULLAGH, Peter; NELDER, ASHWORTH, John (1989). Generalized Linear Models. Chapman and Hall: Volume 37 de Monographs on statistics and applied probability

MADDALA, G.S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge: Cambridge University press.

MEYERS, Glenn (2009). Pure Premium Regression with the Tweedie Model Brainstorms – The Actuarial Review, site: http://www.casact.org/newsletter/index.cfm?fa=viewart&id=5756 consultado em Maio 2013

MINKA, T. P. (2002). Estimating a Gamma distribution. Microsoft Research, Cambridge, UK, Tech. Rep.

MOHAMMED, Rafi (2010). The 1% Windfall: How Successful Companies Use Price to Profit and Grow. HarperBusiness

NELDER, J. A.; WEDDERBURN, R. W. M. (1972) Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3, pp.370-384 site: http://www.jstor.org/stable/2344614 consultado em Janeiro de 2015.

OESTERREICHISCHE NATIONALBANK (OeNB) & FINANCIAL MARKET AUTHORITY (FMA) (2004). Guidelines on Credit Risk Management. Oesterreichische Nationalbank

OHLSSON, Esbjörn; JOHANSSON, Björn (2010). Non-Life Insurance Pricing with Generalized Linear Models. Springer

OOI, HONG (2013) Where does the *offset* go in Poisson/negative binomial regression? Site: http://stats.stackexchange.com/questions/66791/where-does-the-*offset*-go-in-poisson-negative-binomial-regression consultado em Maio de 2015.

ORME, Bryan K.;BAKER, Gary C. (2000) Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations; Sawtooth Software research paper series. Site: http://www.sawtoothsoftware.com/download/techpap/rfcdrw.pdf consultado em Junho de 2015.

OSBORNE, Jason (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research & Evaluation, 15(12). site: http://pareonline.net/getvn.asp?v=15&n=12 consultado em Maio de 2015.

PETERS, Gareth W.; SHEVCHENKO, Pavel V.; WÜTHRICH, Mario V. (2009). Model uncertainty in claims reserving within Tweedie's compound Poisson models. ASTIN Bulletin 39(1), pp.1-33 Site: http://www.actuaries.org/LIBRARY/ASTIN/vol39no1/1.pdf consultado em Junho de 2013

PLACKETT, R. L.; BURMAN, J. P. (1946). The design of optimum multifactorial experiments. Biometrika , 33: 305–325.

PRENTICE, R. L. (1974) A Log Gamma Model and Its Maximum Likelihood Estimation, Biometrika, Vol. 61, No. 3

RIPLEY, Brian; VENABLES, Bill; BATES, Douglas M.; HORNIK, Kurt; GEBHARDT, ALBRECHT; FIRTH, David (2015). Package 'MASS'. Site: http://www.stats.ox.ac.uk/pub/MASS4 consultado em Maio de 2015

SIDDIQI, Naeem (2005) Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, SAS Institute Inc.

SMYTH, Gordon K. (2014) Third Party Motor Insurance in Sweden in http://www.statsci.org/data/general/motorins.html e consultado em Janeiro de 2015

SMYTH, Gordon K.; JORGENSEN, Bent (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modeling. ASTIN bulletin, Vol. 32, No. 1, pp. 143-157 Site: http://www.casualtyactuarialsociety.com/library/astin/vol32no1/143.pdf e consultado em Maio de 2013

STACY, E. W. (1962) A Generalization of the Gamma Distribution. The Annals of Mathematical Statistics. Vol. 33, No. 3, pp. 1187-1192

VARIAN, H. R. (2006). Revealed Preference. Samuelsonian Economics and the Twenty-First Century. MSzenberg, L. Ramrattan e A.A.Gottesman (eds), Oxford Univerty press, pp-99-115 Site: http://people.ischool.berkeley.edu/~hal/Papers/2005/revpref.pdf consultado em Maio de 2015

WITTMAN (2011). Fisher Matrix for Beginners. Site: http://www.physics.ucdavis.edu/~dWittman/ Fisher-matrix-guide.pdf consultado em Abril de 2013

WOODS, D. C.; LEWIS, S. M.; ECCLESTON, J. A.; RUSSELL, K. G. (2006). Designs for Generalized Linear Models With Several Variables and Model Uncertainty, Technometrics, 48, 284–292.

WOOLDRIDGE, Jeffrey (2003) Introductory econometrics: A modern approach. Thomson South-Western, 2 ed

WÜTHRICH, Mario V. (2003). Claims reserving using Twedies's compound Poisson Model, Astin Bulletin, Vol 33, N2, pp 331-346. Site: http://www.actuaries.org/LIBRARY/ASTIN/vol33no2/331.pdf consultado em Junho de 2013

ZEILEIS, KLEIBER E JACKMAN (2008). Regression Models for Count Data in R, Journal of Statistical Software, Volume 27, Issue 8. Site: http://www.jstatsoft.org/v27/i08/paper

ZELLNER (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for. Aggregation Bias. Vol. 57, No. 298 (Jun., 1962), pp. 348-368

ZUANETTI, Daiane Aparecida; DINIZ, Carlos A. R.; LEITE, José Galvão (2006). A Lognormal Model For Insurance Claims Data, REVSTAT – Statistical Journal. Site: http://www.ine.pt/revstat/pdf/rs060203.pdf consultado em Junho de 2013

Anexo A – Casos práticos associados ao estudo dos modelos GLM

Poisson: aplicação para o tema em estudo

Para a construção desta tese serão usadas várias Bases de Dados reais, mas de origem e contexto distinto. De um ponto de vista largo e tendo em conta que o objectivo último deste projecto é o de construir uma abordagem de análise de *Value Based Pricing*, não há problema.

Note-se ainda que para a análise de risco, o objectivo final é compreender e explicitar, de forma breve, os principais mecanismos de formulação dos custos de seguro (com especial destaque aos modelos GLM).

Assim, é usada uma Base de Dados de tarifação automóvel de 1977, aplicada ao contexto sueco – Smyth (2014) baseado em Hallin & Ingenbleek (1983). À data, todas as companhias de seguro suecas usavam os mesmos factores de tarifação, pelo que as suas carteiras podiam ser comparadas e agregadas. Nessa altura, o Comité para a Análise de Risco no Seguro Automóvel sueco foi convidado a analisar se a estrutura presente se ajustava à realidade. É neste contexto que a esta Base de Dados foi construída.

O resultado, pelo autor, do processo de estimação de Poisson pode ser encontrado abaixo[28]. Assim, pode-se entender que:

1. "*Kilometres: Kilometres travelled per year*": tem um impacto (tendencialmente) decrescente no número de sinistros. Ou seja, quanto mais forem os quilómetros percorridos, menor o número de participações de sinistros – eventualmente um resultado contra intuitivo.
2. "*Bonus: No claims bonus. Equal to the number of years, plus one, since last claim*": tem um impacto parabólico, dando ideia de que os sinistros são uma (quase) inevitabilidade ao fim de quatro anos.
3. "*Car*": Cada tipo de carros tem um impacto distinto, indiciando que mais do que os modelos isolados seja talvez mais relevante trabalhar com outras

[28] Note-se, os resultados obtidos com a regressão de Poisson estão logaritmizados. Para obter a contagem dos sinistros é necessário somar os coeficientes e exponencializalos. Assim, a interpretação dos coeficientes não é directa: depende da posição de cada uma das posições de "x".

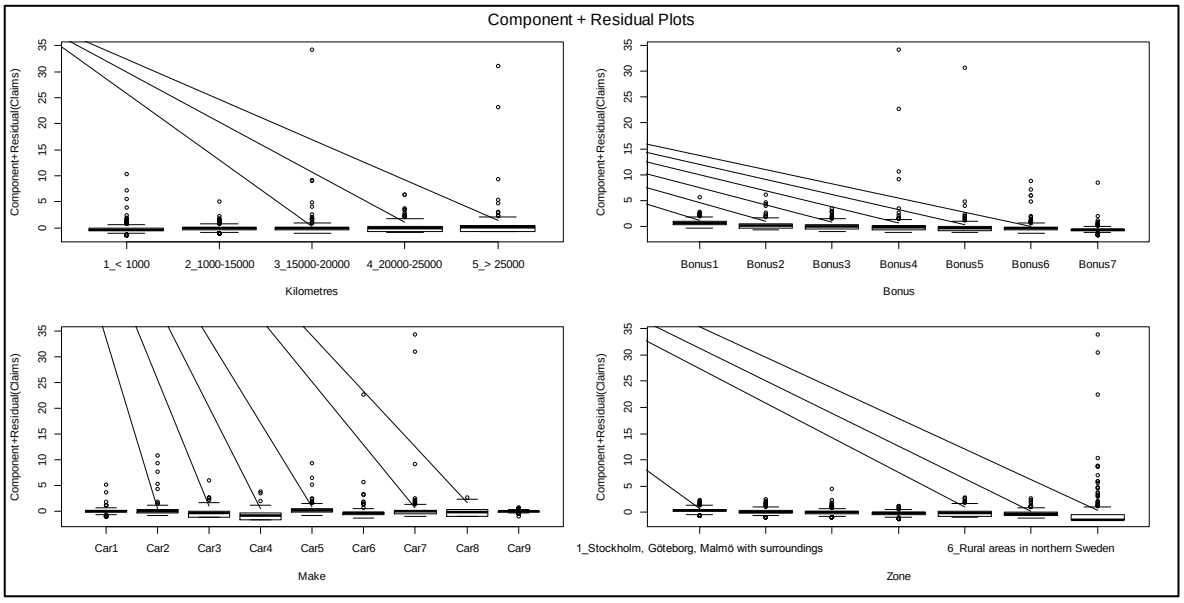
variáveis caracterizadoras do perfil automóvel, como a relação peso potência.

4. "*Zone*": tem um impacto difícil de interpretar, sendo talvez mais relevante associar variáveis à zona indicadores de urbanidade com uma característica ordenada.

Em relação à distribuição associada a cada coeficiente, no caso sem ponderação e sem *offset*, é notório a existência de alguns *outlier*s. A solução de estimação escolhida é assim por *offset*.

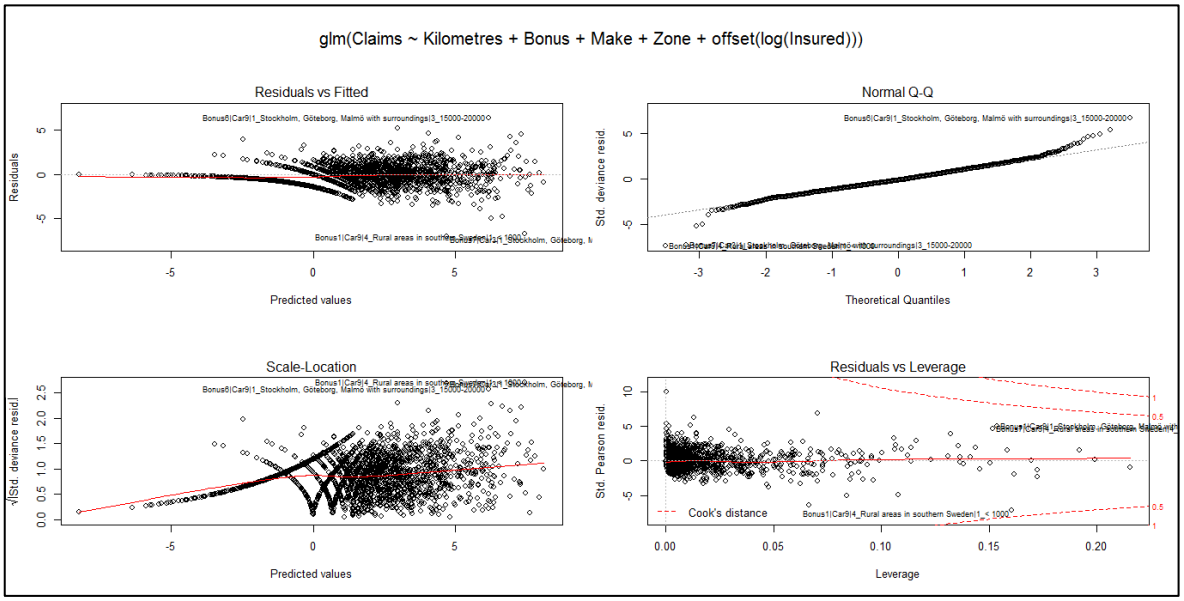
A qualidade do ajustamento, R2, é de 0.9956136. Outras medidas de qualidade de ajustamento podem ser encontradas na abaixo **Error! Reference source not ound.**, onde é visível que o modelo não sofre de heterocedasticidade, e tem uma boa capacidade de previsão nos casos mais regulares.

Figura A.1. - Avaliação gráfica dos resultados para a evolução de frequência para o swedish automobile portfólio in 1977



Fonte: autor

Figura A.2. - Avaliação gráfica da qualidade de ajustamento do modelo de frequência para o swedish automobile portfolio in 1977



Fonte: autor

Tabela A.1. - Resultados para o modelo de contagem baseado no swedish automobile portfolio in 1977

```
> Call:
glm(formula = Claims ~ Kilometres + Bonus + Make + Zone + offset(log(Insured)),
    family = poisson(link = log), data = DB_ClaimsR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9854  -0.8626  -0.1718   0.5997   6.4005

Coefficients:
                                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                                    -1.812840   0.013757 -131.775  < 2e-16 ***
Kilometres[T.2_1000-15000]                      0.212586   0.007524   28.255  < 2e-16 ***
Kilometres[T.3_15000-20000]                     0.320226   0.008661   36.974  < 2e-16 ***
Kilometres[T.4_20000-25000]                     0.404657   0.012054   33.571  < 2e-16 ***
Kilometres[T.5_> 25000]                         0.575954   0.012830   44.892  < 2e-16 ***
Bonus[T.Bonus2]                                -0.478993   0.012094  -39.607  < 2e-16 ***
Bonus[T.Bonus3]                                -0.693172   0.013508  -51.316  < 2e-16 ***
Bonus[T.Bonus4]                                -0.827397   0.014584  -56.735  < 2e-16 ***
Bonus[T.Bonus5]                                -0.925632   0.013968  -66.269  < 2e-16 ***
Bonus[T.Bonus6]                                -0.993457   0.011629  -85.429  < 2e-16 ***
Bonus[T.Bonus7]                                -1.327406   0.008685 -152.845  < 2e-16 ***
Make[T.Car2]                                    0.076245   0.021239    3.590 0.000331 ***
Make[T.Car3]                                   -0.247413   0.025094   -9.859  < 2e-16 ***
Make[T.Car4]                                   -0.653524   0.024185  -27.022  < 2e-16 ***
Make[T.Car5]                                    0.154924   0.020235    7.656 1.91e-14 ***
Make[T.Car6]                                   -0.335581   0.017375  -19.314  < 2e-16 ***
Make[T.Car7]                                   -0.055940   0.023343   -2.396 0.016554 *
Make[T.Car8]                                   -0.043933   0.031604   -1.390 0.164493
```

```
Make[T.Car9]                                                -0.068054   0.009956   -6.836 8.17e-12 ***
Zone[T.2_Other large cities with surroundings]              -0.238168   0.009496  -25.082  < 2e-16 ***
Zone[T.3_Smaller cities with surroundings in southern Sweden] -0.386395   0.009670  -39.959  < 2e-16 ***
Zone[T.4_Rural areas in southern Sweden]                    -0.581902   0.008654  -67.243  < 2e-16 ***
Zone[T.5_Smaller cities with surroundings in northern Sweden] -0.326128   0.014530  -22.446  < 2e-16 ***
Zone[T.6_Rural areas in northern Sweden]                    -0.526234   0.011877  -44.309  < 2e-16 ***
Zone[T.7_Gotland]                                           -0.730999   0.040698  -17.962  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 34070.6  on 2181  degrees of freedom
Residual deviance:  2966.1  on 2157  degrees of freedom
AIC: 10654

Number of Fisher Scoring iterations: 4
```

Gama: aplicação para o tema em estudo

Para melhor compreender a distribuição de custos, e como indicado na regressão de Poisson, é usada uma Base de Dados de tarifação automóvel de 1977, aplicada ao contexto sueco. À data todas as companhias de seguro suecas usavam os mesmos factores de tarifação, pelo que as suas carteiras podiam ser comparadas e agregadas. Nessa altura, o Comité para a Análise de Risco no Seguro Automóvel sueco foi convidado a analisar se a estrutura presente se ajustava à realidade. É neste contexto que a esta Base de Dados foi construída.

O resultado, pelo autor, do processo de estimação de gama pode ser encontrado abaixo. No entanto para melhor compreender é mais fácil analisar a Figura A.4. - Avaliação gráfica dos resultados para a evolução dos custos médios para o swedish automobile portfolio in 1977

que retrata os “partial regression coefficients”. Assim, pode-se entender que:

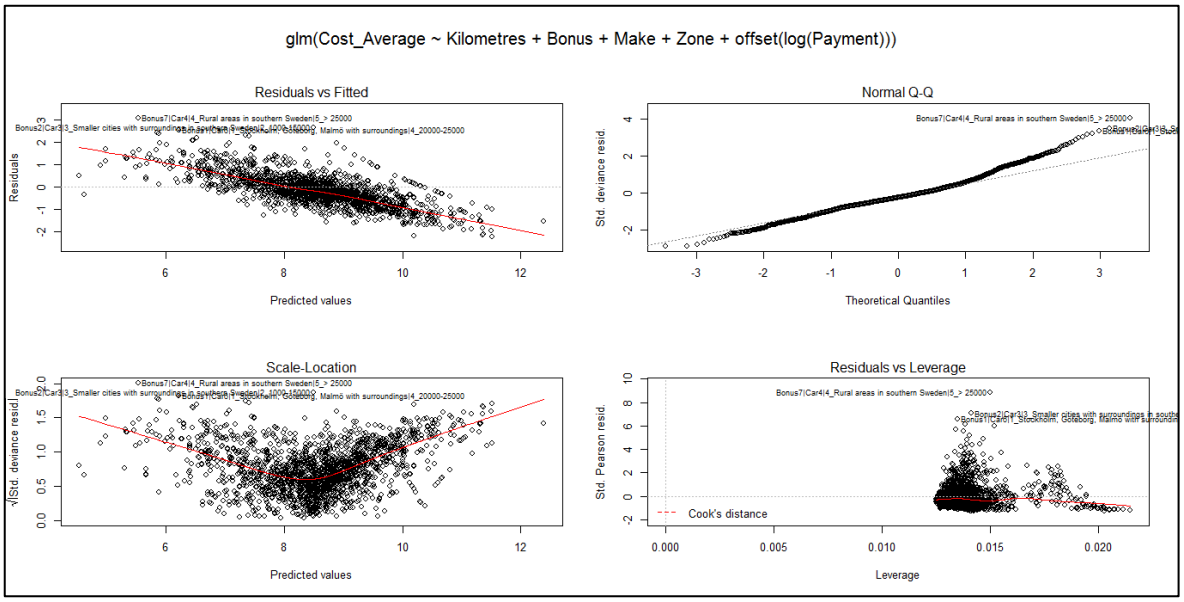
1. “Kilometres: Kilometres travelled per year”: tem um impacto (tendencialmente) decrescente no custo de sinistros. Ou seja, quanto mais forem os quilómetros percorridos, maior o custo com sinistro – eventualmente um resultado contra intuitivo, indiciando a possibilidade de efeitos cruzados.
2. “Bonus: No claims bonus. Equal to the number of years, plus one, since last claim”: tem um impacto oscilante, dando ideia de que os custos de sinistros são uma (quase) inevitabilidade ao fim de quatro anos.
3. “Car”: Cada tipo de carros tem um impacto distinto, indiciando que mais do que os modelos isolados seja talvez mais relevante trabalhar com outras variáveis caracterizadoras do perfil automóvel, como a relação peso potência.
4. “Zone”: tem um impacto difícil de interpretar, sendo talvez mais relevante associar variáveis à zona indicadores de urbanidade com uma característica ordenada.

Em relação à distribuição associada a cada coeficiente, é notório a existência de alguns *outlier*s e de heterocedasticidade mesmo com o padrão de coeficiente de

variação constante. Apesar da correcção por ponderação ou *offset* o modelo simples parece ser o mais adequado. Valeria a pena estudar uma estimação robusta com um padrão de heterocedasticidade desconhecida, aqui desconsiderada.

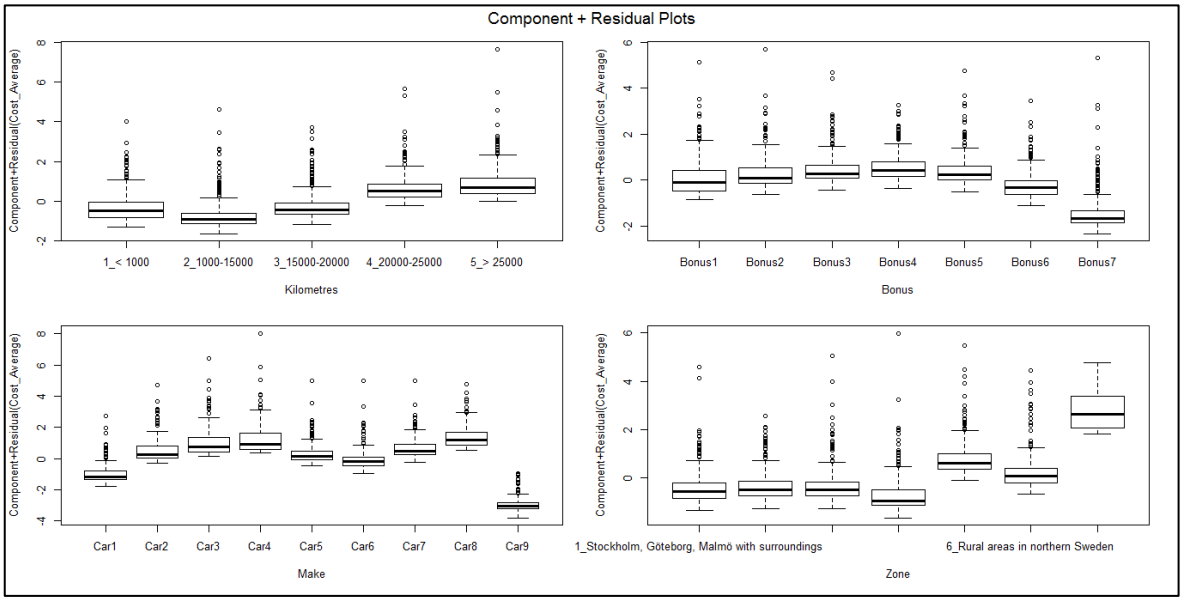
Se acreditar que o modelo de custos pode ser apresentado com uma Tweedie o p óptimo é de 0,4. O modelo de custos está entre a normal e a Poisson; e não uma gama. Ainda assim, os dados parecem mostrar um limite nas coberturas de capital. Se assim for, esta conclusão deve ser repensada com mais detalhe.

Figura A.3. - Avaliação gráfica da qualidade de ajustamento do modelo de custos médios para o swedish automobile portfolio in 1977



Fonte: autor

Figura A.4. - Avaliação gráfica dos resultados para a evolução dos custos médios para o swedish automobile portfolio in 1977



Fonte: autor

Tabela A.2. - Resultados para o modelo de contagem baseado no swedish automobile portfolio in 1977

```
Call:
glm(formula = Cost_Average ~ Kilometres + Bonus + Make + Zone +
    offset(log(Payment)), family = Gamma(link = log), data = DB_Claims_custosR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2014  -0.5361  -0.2008   0.1886   3.0635

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      -3.67947    0.08703 -42.280  < 2e-16 ***
Kilometres[T.2_1000-15000]                       -0.42528    0.05501  -7.731 1.78e-14 ***
Kilometres[T.3_15000-20000]                       0.08502    0.05578   1.524   0.1276
Kilometres[T.4_20000-25000]                       0.98815    0.05813  16.999  < 2e-16 ***
Kilometres[T.5_> 25000]                           1.25784    0.05912  21.274  < 2e-16 ***
Bonus[T.Bonus2]                                   0.19894    0.06883   2.890   0.0039 **
Bonus[T.Bonus3]                                   0.33958    0.06884   4.933 8.86e-07 ***
Bonus[T.Bonus4]                                   0.45129    0.06926   6.516 9.40e-11 ***
Bonus[T.Bonus5]                                   0.31460    0.06876   4.575 5.09e-06 ***
Bonus[T.Bonus6]                                  -0.33270    0.06731  -4.943 8.42e-07 ***
Bonus[T.Bonus7]                                  -1.53958    0.06597 -23.338  < 2e-16 ***
Make[T.Car2]                                      1.47532    0.07404  19.926  < 2e-16 ***
Make[T.Car3]                                      1.98480    0.07716  25.722  < 2e-16 ***
Make[T.Car4]                                      2.24841    0.08055  27.913  < 2e-16 ***
Make[T.Car5]                                      1.27976    0.07397  17.300  < 2e-16 ***
Make[T.Car6]                                      0.89798    0.07335  12.242  < 2e-16 ***
Make[T.Car7]                                      1.62493    0.07488  21.701  < 2e-16 ***
```

```
Make[T.Car8]                                                2.36868    0.07782  30.438  < 2e-16 ***
Make[T.Car9]                                               -2.03784    0.07090 -28.742  < 2e-16 ***
Zone[T.2_Other large cities with surroundings]              0.05534    0.06320   0.876   0.3813
Zone[T.3_Smaller cities with surroundings in southern Sweden]  0.06694    0.06332   1.057   0.2906
Zone[T.4_Rural areas in southern Sweden]                   -0.33342    0.06264  -5.323 1.15e-07 ***
Zone[T.5_Smaller cities with surroundings in northern Sweden]  1.21592    0.06721  18.092  < 2e-16 ***
Zone[T.6_Rural areas in northern Sweden]                    0.64046    0.06512   9.835  < 2e-16 ***
Zone[T.7_Gotland]                                           3.18334    0.08773  36.285  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.5889822)

    Null deviance: 3556.08  on 1796  degrees of freedom
Residual deviance:  848.95  on 1772  degrees of freedom
AIC: 33801

Number of Fisher Scoring iterations: 11
```

Tweedie: aplicação para o tema em estudo

Para melhor compreender a distribuição de custos, e como indicado anteriormente, é usada novamente a Base de Dados de tarifação automóvel de 1977, aplicada ao contexto sueco. Mais uma vez é de salientar que o objectivo desta secção mais prática é o de demonstrar a aplicabilidade destas técnicas em contextos reais.

À data, todas as companhias de seguro suecas usavam os mesmos factores de tarifação, pelo que as suas carteiras podiam ser comparadas e agregadas. Nessa altura, o Comité para a Análise de Risco no Seguro Automóvel sueco foi convidado a analisar se a estrutura presente se ajustava à realidade. É neste contexto que a esta Base de Dados foi construída.

O resultado, pelo autor, do processo de estimação de gama pode ser encontrado na abaixo. No entanto para melhor compreender é mais fácil analisar a Figura A.4. - Avaliação gráfica dos resultados para a evolução dos custos médios para o swedish automobile portfolio in 1977

que retrata os “partial regression coefficients”. Assim, pode-se entender que:

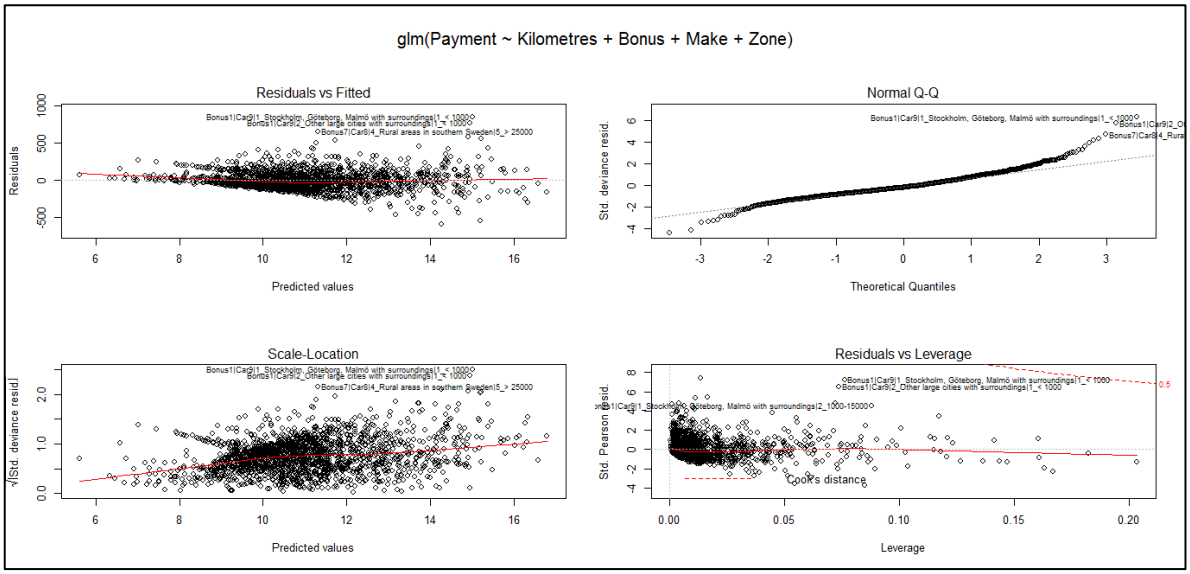
1. “*Kilometres: Kilometres travelled per year*”: tem um impacto (tendencialmente) decrescente no custo total de sinistros. Ou seja, quanto mais forem os quilómetros percorridos, maior o custo com sinistro – eventualmente um resultado contra intuitivo, indiciando a possibilidade de efeitos cruzados.
2. “*Bonus: No claims bonus. Equal to the number of years, plus one, since last claim*”: tem um impacto oscilante, dando ideia de que os custos de sinistros são uma (quase) inevitabilidade ao fim de quatro anos.
3. “*Car*”: Cada tipo de carros tem um impacto distinto, indiciando que mais do que os modelos isolados seja talvez mais relevante trabalhar com outras variáveis caracterizadoras do perfil automóvel, como a relação peso potência.
4. “*Zone*”: tem um impacto difícil de interpretar, sendo talvez mais relevante associar variáveis à zona indicadores de urbanidade com uma

característica ordenada.

Em relação à distribuição associada a cada coeficiente, é notório a existência de alguns *outliers*. Mais relevante, é de salientar a existência de alguma heterocedasticidade empurrando o modelo para uma estimação robusta, aqui desconsiderada.

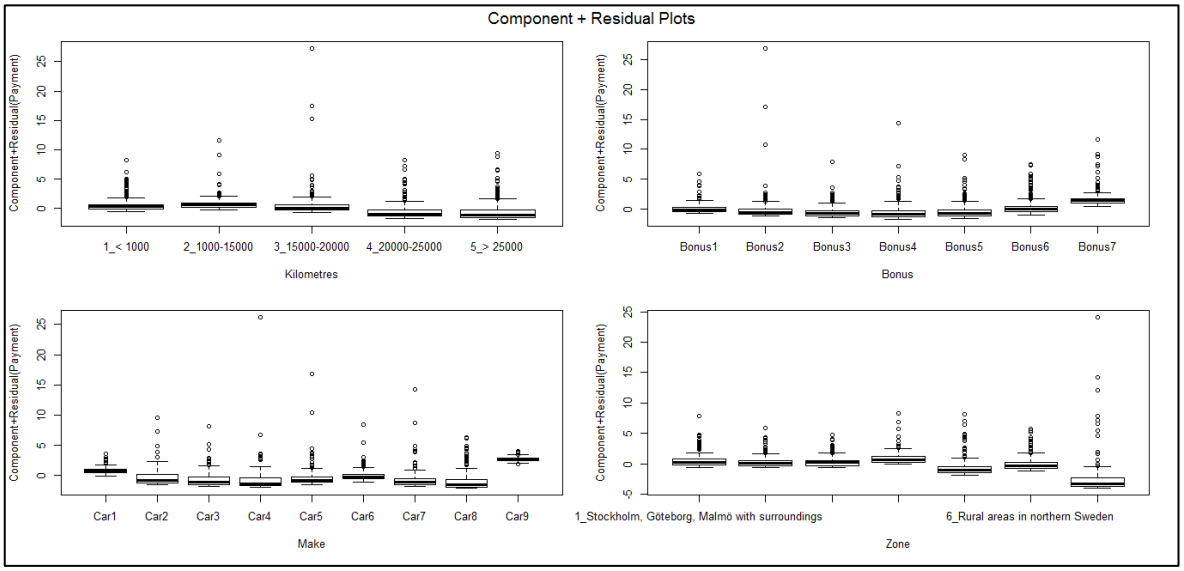
Os resultados estão em linha com os resultados anteriormente apresentados.

Figura A.5. - Avaliação gráfica da qualidade de ajustamento do modelo de custos médios para o swedish automobile portfolio in 1977



Fonte: autor

Figura A.6. - Avaliação gráfica dos resultados para a evolução dos custos médios para o swedish automobile portfolio in 1977



Fonte: autor

Tabela A.3. - Resultados para o modelo de contagem baseado no swedish automobile portfolio in 1977

```
Call:
glm(formula = Payment ~ Kilometres + Bonus + Make + Zone, family = tweedie(var.power = p,
    link.power = 0), data = DB_Claims_custos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.5433  -3.7851  -1.1537   0.6189  17.7317

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       12.76942    0.13229  96.529  < 2e-16 ***
Kilometres[T.2_1000-15000]                         0.24214    0.08356   2.898  0.00380 **
Kilometres[T.3_15000-20000]                       -0.21278    0.08548  -2.489  0.01288 *
Kilometres[T.4_20000-25000]                       -1.23640    0.09030 -13.692  < 2e-16 ***
Kilometres[T.5_> 25000]                           -1.31131    0.09068 -14.460  < 2e-16 ***
Bonus[T.Bonus2]                                   -0.28548    0.10607  -2.691  0.00717 **
Bonus[T.Bonus3]                                   -0.66930    0.10822  -6.184 7.42e-10 ***
Bonus[T.Bonus4]                                   -0.73599    0.10861  -6.776 1.58e-11 ***
Bonus[T.Bonus5]                                   -0.57618    0.10769  -5.350 9.70e-08 ***
Bonus[T.Bonus6]                                    0.15219    0.10376   1.467  0.14259
Bonus[T.Bonus7]                                    1.54442    0.09735  15.864  < 2e-16 ***
Make[T.Car2]                                      -1.35259    0.11777 -11.485  < 2e-16 ***
Make[T.Car3]                                      -1.71888    0.12028 -14.290  < 2e-16 ***
Make[T.Car4]                                      -1.79209    0.12080 -14.835  < 2e-16 ***
Make[T.Car5]                                      -1.30422    0.11745 -11.104  < 2e-16 ***
Make[T.Car6]                                      -0.86090    0.11460  -7.512 8.44e-14 ***
Make[T.Car7]                                      -1.72909    0.12036 -14.366  < 2e-16 ***
Make[T.Car8]                                      -1.98153    0.12217 -16.220  < 2e-16 ***
Make[T.Car9]                                       1.98943    0.10018  19.859  < 2e-16 ***
Zone[T.2_Other large cities with surroundings]    -0.16496    0.09890  -1.668  0.09546 .
Zone[T.3_Smaller cities with surroundings in southern Sweden] -0.16650    0.09891  -1.683  0.09243 .
```

```
Zone[T.4_Rural areas in southern Sweden]                         0.41900    0.09609   4.361 1.36e-05 ***
Zone[T.5_Smaller cities with surroundings in northern Sweden] -1.31945    0.10522 -12.540  < 2e-16 ***
Zone[T.6_Rural areas in northern Sweden]                        -0.67773    0.10158  -6.672 3.19e-11 ***
Zone[T.7_Gotland]                                               -3.79935    0.12285 -30.926  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Tweedie family taken to be 15.02572)

    Null deviance: 126353  on 2204  degrees of freedom
Residual deviance:  33584  on 2180  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 11
```

Anexo B – Programação utlizada para o estudo dos modelos GLM

```
MotorSweden1977 <- read.table("http://www.statsci.org/data/general/motorins.txt", header=TRUE, sep="\t",
na.strings="NA", dec=".", strip.white=TRUE)

############################################################
#Dicionário
############################################################
#Kilometres   Kilometres travelled per year
##1: < 1000
##2: 1000-15000
##3: 15000-20000
##4: 20000-25000
##5: > 25000

#Zone   Geographical zone
##1: Stockholm, Göteborg, Malmö with surroundings
##2: Other large cities with surroundings
##3: Smaller cities with surroundings in southern Sweden
##4: Rural areas in southern Sweden
##5: Smaller cities with surroundings in northern Sweden
##6: Rural areas in northern Sweden
##7: Gotland

#Bonus  No claims bonus. Equal to the number of years, plus one, since last claim

# Make  1-8 represent eight different common car models. All other models are combined in class 9
##The other makes could not be identified because of the potential for the data to impact on sales of
those cars
##1: Car 1
##2: Car 2
##3: Car 3
##4: Car 4 #4 is the Volkswagen 1200, which was discontinued shortly after 1977
##5: Car 5
##6: Car 6
##7: Car 7
##8: Car 8
##9: Car 9

# Insured  Number of insured in policy-years
# Claims  Number of claims
# Payment  Total value of payments in Skr

############################################################
#Factores
############################################################

MotorSweden1977$k<- factor(MotorSweden1977$Kilometres,
                           levels = c(1:5),
                           labels = c("1_< 1000",
                                      "2_1000-15000",
                                      "3_15000-20000",
                                      "4_20000-25000",
                                      "5_> 25000"),ordered = is.ordered(MotorSweden1977$Kilometres))

MotorSweden1977$z<- factor(MotorSweden1977$Zone,
                           levels = c(1:7),
                           labels = c("1_Stockholm, Göteborg, Malmö with surroundings",
```

```
                                    "2_Other large cities with surroundings",
                                    "3_Smaller cities with surroundings in southern Sweden",
                                    "4_Rural areas in southern Sweden",
                                    "5_Smaller cities with surroundings in northern Sweden",
                                    "6_Rural areas in northern Sweden",
                                    "7_Gotland") )

MotorSweden1977$m<- factor(MotorSweden1977$Make,
                           levels = c(1:9),
                           labels = c("Car1",
                                      "Car2",
                                      "Car3",
                                      "Car4",
                                      "Car5",
                                      "Car6",
                                      "Car7",
                                      "Car8",
                                      "Car9"))

MotorSweden1977$b<- factor(MotorSweden1977$Bonus,
                           levels = c(1:7),
                           labels = c("Bonus1",
                                      "Bonus2",
                                      "Bonus3",
                                      "Bonus4",
                                      "Bonus5",
                                      "Bonus6",
                                      "Bonus7"),ordered = is.ordered(MotorSweden1977$Bonus))




#Gerar as tariff cells


############################################################
#Gerar as tariff cells
##através da geração de todas as combinações
############################################################
tariff_cells = expand.grid (k=c("1_< 1000",
                                "2_1000-15000",
                                "3_15000-20000",
                                "4_20000-25000",
                                "5_> 25000"),
                            z=c("1_Stockholm, Göteborg, Malmö with surroundings",
                                "2_Other large cities with surroundings",
                                "3_Smaller cities with surroundings in southern Sweden",
                                "4_Rural areas in southern Sweden",
                                "5_Smaller cities with surroundings in northern Sweden",
                                "6_Rural areas in northern Sweden",
                                "7_Gotland") ,
                            m=c("Car1",
                                "Car2",
                                "Car3",
                                "Car4",
                                "Car5",
```

```
                                "Car6",
                                "Car7",
                                "Car8",
                                "Car9"),
                            b=c("Bonus1",
                                "Bonus2",
                                "Bonus3",
                                "Bonus4",
                                "Bonus5",
                                "Bonus6",
                                "Bonus7"))


#################################################################
##todas as combinações darão 2205 casos: dim(tariff_cells)
#################################################################
Insured=rep(0, dim(tariff_cells)[1])
Claims=rep(0, dim(tariff_cells)[1])
Payment=rep(0, dim(tariff_cells)[1])

tariff_cells<- data.frame(Insured,Claims,Payment,tariff_cells)


#################################################################
##empilhar os ficheiros de forma a garantir que tenho a combinação de todas as tariff cells
#################################################################
library("Rcmdr", lib.loc="C:/Program Files/R/R-3.1.0/library") #para empilhar os dados, este pacote deve
estar instalado.

MotorSweden1977_aux<- mergeRows(MotorSweden1977, tariff_cells, common.only=FALSE)

#################################################################
###Criar um index
#################################################################
MotorSweden1977_aux$index<-
paste(MotorSweden1977_aux$b,MotorSweden1977_aux$m,MotorSweden1977_aux$z,MotorSweden1977_aux$k,sep="|" )

#################################################################
###Criar a BD contagens e Custos por tariff cells
#################################################################
DB_Claims<-data.frame(
  Insured=tapply(MotorSweden1977_aux$Insured,  MotorSweden1977_aux$index ,sum ) ,
  Claims=tapply(MotorSweden1977_aux$Claims,  MotorSweden1977_aux$index ,sum ) ,
  Payment=tapply(MotorSweden1977_aux$Payment,  MotorSweden1977_aux$index ,sum ) ,
  Kilometres=tariff_cells$k,
  Zone=tariff_cells$z,
  Make=tariff_cells$m,
  Bonus=tariff_cells$b)

DB_Claims$Cost_Average=DB_Claims$Payment/DB_Claims$Claims


#################################################################
###Modelo de freq
#################################################################
```

```
#Modelo sem ponderador
DB_Claims_freq<-DB_Claims [! DB_Claims$Insured %in% 0,] #tirar os casos que não podem ser ponderados por falta de variáveis

GLM.2 <- glm(Claims ~ Kilometres +Bonus + Make + Zone, family=poisson(link=log), data=DB_Claims_freq)
summary(GLM.2)

#Modelo com ponderador
## Criar ponderador
DB_ClaimsR<-DB_Claims [! DB_Claims$Insured %in% 0,] #tirar os casos que não podem ser ponderados por falta de variáveis

DB_ClaimsR$Ponderador<-(sum(DB_ClaimsR$Insured)/nrow(DB_ClaimsR))/DB_ClaimsR$Insured

DB_ClaimsR$teste<-DB_ClaimsR$Ponderador*DB_ClaimsR$Insured # garantir que o Ponderador foi bem calculado

## Fim do ponderador

GLM.3 <- glm(Claims ~ Kilometres +Bonus + Make + Zone, family=poisson(link=log), weights=Ponderador, data=DB_ClaimsR)
summary(GLM.3)


#Modelo com offset

GLM.4 <- glm(Claims ~ Kilometres +Bonus + Make + Zone +offset(log(Insured)), family=poisson(link=log), data=DB_ClaimsR)
summary(GLM.4)


Poisson<-
data.frame(DB_ClaimsR$Claims,round(predict(GLM.2,type="response")),round(predict(GLM.3,type="response"))
,round(predict(GLM.4,type="response")))
R2_poisson<-array(cor(Poisson)^2)
cat("Apuramento do R^2 GLM2=")
R2_poisson[2]
cat("Apuramento do R^2 GLM3 (ponderador=")
R2_poisson[3]
cat("Apuramento do R^2 GLM4 (ponderador=")
R2_poisson[4]

oldpar <- par(oma=c(0,0,3,0), mfrow=c(1,3))
plot(DB_ClaimsR$Claims,predict(GLM.2,type="response"), main="poisson")
plot(DB_ClaimsR$Claims,predict(GLM.3,type="response"),  main="poisson com ponderador (o modelo mais fraquito, reparar escala de yy")
plot(DB_ClaimsR$Claims,predict(GLM.4,type="response"), main="poisson com offset")
par(oldpar)


##Para compreender a distribuição de y_i  , para uma dada contagem específica (m) considerando que o vector x_i é conhecido
crPlots(GLM.4)

##Qualidade do modelo de freq
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(GLM.4)
```

```
par(oldpar)


############################################################
###Modelo de custo médio
############################################################

DB_Claims_custos<-DB_Claims[! DB_Claims$Cost_Average == 0, ] #o ! faz a condição negativa
hist(DB_Claims_custos$Cost_Average,breaks = "Scot") #para confirmar se ficou tudo bem

GLM.5 <- glm(Cost_Average ~ Kilometres +Bonus + Make + Zone, family=Gamma(link=log),
data=DB_Claims_custos)


#Modelo com ponderador
## Criar ponderador
DB_Claims_custosR<-DB_Claims_custos [! DB_Claims_custos$Payment %in% c(NA),] #tirar os casos que não
podem ser ponderados por falta de variáveis

DB_Claims_custosR$Ponderador<-
(sum(DB_Claims_custosR$Payment)/nrow(DB_Claims_custosR))/DB_Claims_custosR$Cost_Average

DB_Claims_custosR$teste<-DB_Claims_custosR$Ponderador*DB_Claims_custosR$Payment # garantir que o
Ponderador foi bem calculado

## Fim do ponderador

GLM.5 <- glm(Cost_Average ~ Kilometres +Bonus + Make + Zone, family=Gamma(link=log),
data=DB_Claims_custosR ) #novo modelo GLM5 para granatir a comparabilidade

GLM.6 <- glm(Cost_Average ~ Kilometres +Bonus + Make + Zone, family=Gamma(link=log), weight=Ponderador,
data=DB_Claims_custosR )

GLM.7 <- glm(Cost_Average ~ Kilometres +Bonus + Make + Zone + offset(log(Payment)),
family=Gamma(link=log), data=DB_Claims_custosR )


Gama<-
data.frame(DB_Claims_custosR$Cost_Average,predict(GLM.5,type="response"),predict(GLM.6,type="response"),
predict(GLM.7,type="response"))
R2_Gama<-array(cor(Gama)^2)
cat("Apuramento do R^2 GLM5=")
R2_Gama[2]
cat("Apuramento do R^2 GLM6 (ponderador=")
R2_Gama[3]
cat("Apuramento do R^2 GLM7 (ponderador=")
R2_Gama[4]


oldpar <- par(oma=c(0,0,3,0), mfrow=c(1,3))
plot(DB_Claims_custosR$Cost_Average,predict(GLM.5,type="response"), main="Gama")
plot(DB_Claims_custosR$Cost_Average,predict(GLM.6,type="response"), main="Gama com ponderador (o modelo
mais fraquito, reparar escala de yy")
plot(DB_Claims_custosR$Cost_Average,predict(GLM.7,type="response"), main="Gama com offset")
par(oldpar)
```

```
##Para compreender a distribuição de y_i   , para um dado impacto específico (m) considerando que o
vector x_i é conhecido
library("Rcmdr", lib.loc="C:/Program Files/R/R-3.1.0/library")
crPlots(GLM.7)


##Qualidade do modelo de impacto
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(GLM.7)
par(oldpar)


############################################################
###Modelo de custo totais
############################################################
#Instalar o pacore com Tweedie
library("statmod", lib.loc="C:/Program Files/R/R-3.1.0/library")
DB_Claims_custos<-DB_Claims[! DB_Claims$Payment == 0, ] #o ! faz a condição negativa
hist(DB_Claims_custos$Payment,breaks = "Scot") #para confirmar se ficou tudo bem

#vamos definir p por tentativa e erro.

n= 9999  #podemos por n até m-1; mas por vezes o modelo não converge quando se está tão próximo do limite
         #pelo que por vezes é dificil obter o mínimo
m=10000
yy=rep(NA, n)


for (i in 1:n) {
  xx=1+i/m
  # var.power   index of power variance function NO NOSSO CASO TEM DE SER UM NUMERO ENTRE (1,2)
  #  link.power   index of power link function. link.power=0 produces a log-link. Defaults to the
canonical link, which is 1-var.power.
  GLM.aux <- glm(Payment ~ Kilometres +Bonus + Make + Zone, family=tweedie (var.power=xx,link.power=0),
data=DB_Claims_custos )
  aux<- (exp(predict(GLM.aux)-DB_Claims_custos$Payment)^2)
  yy[i]<-sum(aux,na.rm = any(!is.na(x)))
#http://www.r-bloggers.com/perculiar-behaviour-of-the-sum-function/
}


cat("Descoberta da iteração do p.")
which (yy ==min(yy))

cat("Definição do p =")
p=1+which (yy ==min(yy))/m
p
plot(yy)


GLM.tweedie <- glm(Payment ~ Kilometres +Bonus + Make + Zone, family=tweedie (var.power=p,link.power=0),
data=DB_Claims_custos )
summary(GLM.tweedie)

tweedie<-data.frame(DB_Claims_custos$Payment,predict(GLM.tweedie))
R2_Gama<-array(cor(tweedie)^2)
```

```
cat("Apuramento do R^2 GLM.tweedie=")
R2_Gama[2]


##Para compreender a distribuição de y_i   , para um dado impacto específico (m) considerando que o vector x_i é conhecido
library("Rcmdr", lib.loc="C:/Program Files/R/R-3.1.0/library")
crPlots(tweedie)


##Qualidade do modelo de impacto
oldpar <- par(oma=c(0,0,3,0), mfrow=c(2,2))
plot(tweedie)
par(oldpar)
```

Anexo C – Programação utlizada para o apuramento do modelo tarifário das companhias congéneres: Box-Cox, SUR e correcção do ponto de massa

```
rm(list=ls(all=TRUE))
#dados de 2010 da companhia F
#dados confidenciais e ligeiramente ajustados
BDExp<- read.csv("file:///C:/BD_companhias.csv", sep=";", na.strings=c(".", "NA", "", "?"),
strip.white=TRUE, encoding="UTF-8")


library(MASS)
library(Rcmdr) # senão o comando predict não funciona...
library("dummies")
library(systemfit)
library("Matrix")

#apagar as variáveis colunas que não me interessam
keep<-c("Categoria","Idveiculo",
"IdCondutor","IdCarta","Sexo","Sinistros","Dist_Concelho","y1","y2","y3","y4","y5","y6","y7")
BDExp_aux<-BDExp[keep]


#variáveis a dummificar. Vou esquecer a variável pagamento para colocar maior heterogeneidade nos dados
#"Idveiculo",  "IdCondutor",  "IdCarta" não vamos dummificar, já que são variáveis contínuas
var=c("Categoria",  "Sexo",  "Sinistros",  "Dist_Concelho")

#BD de trabalho
BDExp_r<-dummy.data.frame(BDExp_aux, names = var, omit.constants=TRUE)


step<-0.001
min_lambda<- -5 # só pode ir até -2
max_lambda<-  1 # só pode ir até +2

companhia<-
c("Companhia_1","Companhia_2","Companhia_3","Companhia_4","Companhia_5","Companhia_6","Companhia_7")

oldpar <- par(oma=c(0,0,4,0), mfrow=c(2,4))
n<-length(companhia)
lambda<-rep(NA,n)

lambda_mv1 <- boxcox(y1 ~ . -y2 -y3 -y4 -y5 -y6 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda1<-min_lambda+step*which (lambda_mv1$y == max (lambda_mv1$y))-step

lambda_mv2 <- boxcox(y2 ~ . -y1 -y3 -y4 -y5 -y6 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda2<-min_lambda+step*which (lambda_mv2$y == max (lambda_mv2$y))-step

lambda_mv3 <- boxcox(y3 ~ . -y1 -y2 -y4 -y5 -y6 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda3<-min_lambda+step*which (lambda_mv3$y == max (lambda_mv3$y))-step

lambda_mv4 <- boxcox(y4 ~ . -y1 -y2 -y3 -y5 -y6 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda4<-min_lambda+step*which (lambda_mv4$y == max (lambda_mv4$y))-step

lambda_mv5 <- boxcox(y5 ~ . -y1 -y2 -y3 -y4 -y6 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
```

```
lambda5<-min_lambda+step*which (lambda_mv5$y == max (lambda_mv5$y))-step

lambda_mv6 <- boxcox(y6 ~ . -y1 -y2 -y3 -y4 -y5 -y7 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda6<-min_lambda+step*which (lambda_mv6$y == max (lambda_mv6$y))-step

lambda_mv7 <- boxcox(y7 ~ . -y1 -y2 -y3 -y4 -y5 -y6 ,
data=BDExp_r,lambda=seq(min_lambda,max_lambda,by=step),plotit = TRUE)
lambda7<-min_lambda+step*which (lambda_mv7$y == max (lambda_mv7$y))-step
par(oldpar)

lambda<-c(lambda1,lambda2,lambda3,lambda4,lambda5,lambda6,lambda7)
lambda

#[1] -3.504 -1.723 -4.055 -3.469 -1.380 -3.276 -3.024

BDExp_r$yboxcox1<-log(BDExp_r$y1)
BDExp_r$yboxcox2<-log(BDExp_r$y2)
BDExp_r$yboxcox3<-log(BDExp_r$y3)
BDExp_r$yboxcox4<-log(BDExp_r$y4)
BDExp_r$yboxcox5<-log(BDExp_r$y5)
BDExp_r$yboxcox6<-log(BDExp_r$y6)
BDExp_r$yboxcox7<-log(BDExp_r$y7)


#BDExp_r$yboxcox1<-((BDExp_r$y1^lambda[1])-1)/lambda[1]
#BDExp_r$yboxcox2<-((BDExp_r$y2^lambda[2])-1)/lambda[2]
#BDExp_r$yboxcox3<-((BDExp_r$y3^lambda[3])-1)/lambda[3]
#BDExp_r$yboxcox4<-((BDExp_r$y4^lambda[4])-1)/lambda[4]
#BDExp_r$yboxcox5<-((BDExp_r$y5^lambda[5])-1)/lambda[5]
#BDExp_r$yboxcox6<-((BDExp_r$y6^lambda[6])-1)/lambda[6]
#BDExp_r$yboxcox7<-((BDExp_r$y7^lambda[7])-1)/lambda[7]



lm1<-BDExp_r$yboxcox1~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
```

```
BDExp_r$Sinistros1sin_0anos_1anos+
BDExp_r$Sinistros1sin_0anos_2anos+
BDExp_r$Sinistros1sin_0anos_3anos+
BDExp_r$Sinistros1sin_0anos_4anos+
BDExp_r$Sinistros1sin_0anos_5anos+
BDExp_r$Sinistros1sin_1anos_5anos+
BDExp_r$Sinistros1sin_2anos_9anos+
BDExp_r$Sinistros1sin_4anos_10anos+
BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_168+
BDExp_r$Dist_Concelhodc_17+
BDExp_r$Dist_Concelhodc_194+
BDExp_r$Dist_Concelhodc_241+
BDExp_r$Dist_Concelhodc_25+
BDExp_r$Dist_Concelhodc_259+
BDExp_r$Dist_Concelhodc_262+
BDExp_r$Dist_Concelhodc_305+
BDExp_r$Dist_Concelhodc_319+
BDExp_r$Dist_Concelhodc_322+
BDExp_r$Dist_Concelhodc_344+
BDExp_r$Dist_Concelhodc_345+
BDExp_r$Dist_Concelhodc_357+
BDExp_r$Dist_Concelhodc_359+
BDExp_r$Dist_Concelhodc_376+
BDExp_r$Dist_Concelhodc_40+
BDExp_r$Dist_Concelhodc_421+
BDExp_r$Dist_Concelhodc_425+
BDExp_r$Dist_Concelhodc_427+
BDExp_r$Dist_Concelhodc_437+
BDExp_r$Dist_Concelhodc_470+
BDExp_r$Dist_Concelhodc_483+
BDExp_r$Dist_Concelhodc_491+
BDExp_r$Dist_Concelhodc_493+
BDExp_r$Dist_Concelhodc_530+
BDExp_r$Dist_Concelhodc_538+
BDExp_r$Dist_Concelhodc_557+
BDExp_r$Dist_Concelhodc_565+
BDExp_r$Dist_Concelhodc_585+
BDExp_r$Dist_Concelhodc_631+
BDExp_r$Dist_Concelhodc_641+
BDExp_r$Dist_Concelhodc_650+
BDExp_r$Dist_Concelhodc_651+
BDExp_r$Dist_Concelhodc_654+
BDExp_r$Dist_Concelhodc_656+
BDExp_r$Dist_Concelhodc_672+
BDExp_r$Dist_Concelhodc_710+
BDExp_r$Dist_Concelhodc_737+
BDExp_r$Dist_Concelhodc_771+
BDExp_r$Dist_Concelhodc_789+
BDExp_r$Dist_Concelhodc_796+
BDExp_r$Dist_Concelhodc_807+
BDExp_r$Dist_Concelhodc_858+
BDExp_r$Dist_Concelhodc_861+
```

```
  BDExp_r$Dist_Concelhodc_869+
  BDExp_r$Dist_Concelhodc_886+
  BDExp_r$Dist_Concelhodc_896+
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984


lm2<-BDExp_r$yboxcox2~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
  BDExp_r$Sinistros1sin_0anos_1anos+
  BDExp_r$Sinistros1sin_0anos_2anos+
  BDExp_r$Sinistros1sin_0anos_3anos+
  BDExp_r$Sinistros1sin_0anos_4anos+
  BDExp_r$Sinistros1sin_0anos_5anos+
  BDExp_r$Sinistros1sin_1anos_5anos+
  BDExp_r$Sinistros1sin_2anos_9anos+
  BDExp_r$Sinistros1sin_4anos_10anos+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_168+
  BDExp_r$Dist_Concelhodc_17+
  BDExp_r$Dist_Concelhodc_194+
  BDExp_r$Dist_Concelhodc_241+
  BDExp_r$Dist_Concelhodc_25+
  BDExp_r$Dist_Concelhodc_259+
  BDExp_r$Dist_Concelhodc_262+
  BDExp_r$Dist_Concelhodc_305+
  BDExp_r$Dist_Concelhodc_319+
  BDExp_r$Dist_Concelhodc_322+
  BDExp_r$Dist_Concelhodc_344+
```

```
  BDExp_r$Dist_Concelhodc_345+
  BDExp_r$Dist_Concelhodc_357+
  BDExp_r$Dist_Concelhodc_359+
  BDExp_r$Dist_Concelhodc_376+
  BDExp_r$Dist_Concelhodc_40+
  BDExp_r$Dist_Concelhodc_421+
  BDExp_r$Dist_Concelhodc_425+
  BDExp_r$Dist_Concelhodc_427+
  BDExp_r$Dist_Concelhodc_437+
  BDExp_r$Dist_Concelhodc_470+
  BDExp_r$Dist_Concelhodc_483+
  BDExp_r$Dist_Concelhodc_491+
  BDExp_r$Dist_Concelhodc_493+
  BDExp_r$Dist_Concelhodc_530+
  BDExp_r$Dist_Concelhodc_538+
  BDExp_r$Dist_Concelhodc_557+
  BDExp_r$Dist_Concelhodc_565+
  BDExp_r$Dist_Concelhodc_585+
  BDExp_r$Dist_Concelhodc_631+
  BDExp_r$Dist_Concelhodc_641+
  BDExp_r$Dist_Concelhodc_650+
  BDExp_r$Dist_Concelhodc_651+
  BDExp_r$Dist_Concelhodc_654+
  BDExp_r$Dist_Concelhodc_656+
  BDExp_r$Dist_Concelhodc_672+
  BDExp_r$Dist_Concelhodc_710+
  BDExp_r$Dist_Concelhodc_737+
  BDExp_r$Dist_Concelhodc_771+
  BDExp_r$Dist_Concelhodc_789+
  BDExp_r$Dist_Concelhodc_796+
  BDExp_r$Dist_Concelhodc_807+
  BDExp_r$Dist_Concelhodc_858+
  BDExp_r$Dist_Concelhodc_861+
  BDExp_r$Dist_Concelhodc_869+
  BDExp_r$Dist_Concelhodc_886+
  BDExp_r$Dist_Concelhodc_896+
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984


lm3<-BDExp_r$yboxcox3~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
```

```
BDExp_r$Sinistros0sin_15anos_15anos+
BDExp_r$Sinistros0sin_2anos_2anos+
BDExp_r$Sinistros0sin_4anos_4anos+
BDExp_r$Sinistros0sin_5anos_10anos+
BDExp_r$Sinistros0sin_5anos_12anos+
BDExp_r$Sinistros0sin_5anos_5anos+
BDExp_r$Sinistros0sin_5anos_6anos+
BDExp_r$Sinistros0sin_5anos_7anos+
BDExp_r$Sinistros0sin_5anos_8anos+
BDExp_r$Sinistros0sin_7anos_9anos+
BDExp_r$Sinistros1sin_0anos_1anos+
BDExp_r$Sinistros1sin_0anos_2anos+
BDExp_r$Sinistros1sin_0anos_3anos+
BDExp_r$Sinistros1sin_0anos_4anos+
BDExp_r$Sinistros1sin_0anos_5anos+
BDExp_r$Sinistros1sin_1anos_5anos+
BDExp_r$Sinistros1sin_2anos_9anos+
BDExp_r$Sinistros1sin_4anos_10anos+
BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_168+
BDExp_r$Dist_Concelhodc_17+
BDExp_r$Dist_Concelhodc_194+
BDExp_r$Dist_Concelhodc_241+
BDExp_r$Dist_Concelhodc_25+
BDExp_r$Dist_Concelhodc_259+
BDExp_r$Dist_Concelhodc_262+
BDExp_r$Dist_Concelhodc_305+
BDExp_r$Dist_Concelhodc_319+
BDExp_r$Dist_Concelhodc_322+
BDExp_r$Dist_Concelhodc_344+
BDExp_r$Dist_Concelhodc_345+
BDExp_r$Dist_Concelhodc_357+
BDExp_r$Dist_Concelhodc_359+
BDExp_r$Dist_Concelhodc_376+
BDExp_r$Dist_Concelhodc_40+
BDExp_r$Dist_Concelhodc_421+
BDExp_r$Dist_Concelhodc_425+
BDExp_r$Dist_Concelhodc_427+
BDExp_r$Dist_Concelhodc_437+
BDExp_r$Dist_Concelhodc_470+
BDExp_r$Dist_Concelhodc_483+
BDExp_r$Dist_Concelhodc_491+
BDExp_r$Dist_Concelhodc_493+
BDExp_r$Dist_Concelhodc_530+
BDExp_r$Dist_Concelhodc_538+
BDExp_r$Dist_Concelhodc_557+
BDExp_r$Dist_Concelhodc_565+
BDExp_r$Dist_Concelhodc_585+
BDExp_r$Dist_Concelhodc_631+
BDExp_r$Dist_Concelhodc_641+
BDExp_r$Dist_Concelhodc_650+
BDExp_r$Dist_Concelhodc_651+
BDExp_r$Dist_Concelhodc_654+
```

```
  BDExp_r$Dist_Concelhodc_656+
  BDExp_r$Dist_Concelhodc_672+
  BDExp_r$Dist_Concelhodc_710+
  BDExp_r$Dist_Concelhodc_737+
  BDExp_r$Dist_Concelhodc_771+
  BDExp_r$Dist_Concelhodc_789+
  BDExp_r$Dist_Concelhodc_796+
  BDExp_r$Dist_Concelhodc_807+
  BDExp_r$Dist_Concelhodc_858+
  BDExp_r$Dist_Concelhodc_861+
  BDExp_r$Dist_Concelhodc_869+
  BDExp_r$Dist_Concelhodc_886+
  BDExp_r$Dist_Concelhodc_896+
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984


lm4<-BDExp_r$yboxcox4~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
  BDExp_r$Sinistros1sin_0anos_1anos+
  BDExp_r$Sinistros1sin_0anos_2anos+
  BDExp_r$Sinistros1sin_0anos_3anos+
  BDExp_r$Sinistros1sin_0anos_4anos+
  BDExp_r$Sinistros1sin_0anos_5anos+
  BDExp_r$Sinistros1sin_1anos_5anos+
  BDExp_r$Sinistros1sin_2anos_9anos+
  BDExp_r$Sinistros1sin_4anos_10anos+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_168+
```

```
BDExp_r$Dist_Concelhodc_17+
BDExp_r$Dist_Concelhodc_194+
BDExp_r$Dist_Concelhodc_241+
BDExp_r$Dist_Concelhodc_25+
BDExp_r$Dist_Concelhodc_259+
BDExp_r$Dist_Concelhodc_262+
BDExp_r$Dist_Concelhodc_305+
BDExp_r$Dist_Concelhodc_319+
BDExp_r$Dist_Concelhodc_322+
BDExp_r$Dist_Concelhodc_344+
BDExp_r$Dist_Concelhodc_345+
BDExp_r$Dist_Concelhodc_357+
BDExp_r$Dist_Concelhodc_359+
BDExp_r$Dist_Concelhodc_376+
BDExp_r$Dist_Concelhodc_40+
BDExp_r$Dist_Concelhodc_421+
BDExp_r$Dist_Concelhodc_425+
BDExp_r$Dist_Concelhodc_427+
BDExp_r$Dist_Concelhodc_437+
BDExp_r$Dist_Concelhodc_470+
BDExp_r$Dist_Concelhodc_483+
BDExp_r$Dist_Concelhodc_491+
BDExp_r$Dist_Concelhodc_493+
BDExp_r$Dist_Concelhodc_530+
BDExp_r$Dist_Concelhodc_538+
BDExp_r$Dist_Concelhodc_557+
BDExp_r$Dist_Concelhodc_565+
BDExp_r$Dist_Concelhodc_585+
BDExp_r$Dist_Concelhodc_631+
BDExp_r$Dist_Concelhodc_641+
BDExp_r$Dist_Concelhodc_650+
BDExp_r$Dist_Concelhodc_651+
BDExp_r$Dist_Concelhodc_654+
BDExp_r$Dist_Concelhodc_656+
BDExp_r$Dist_Concelhodc_672+
BDExp_r$Dist_Concelhodc_710+
BDExp_r$Dist_Concelhodc_737+
BDExp_r$Dist_Concelhodc_771+
BDExp_r$Dist_Concelhodc_789+
BDExp_r$Dist_Concelhodc_796+
BDExp_r$Dist_Concelhodc_807+
BDExp_r$Dist_Concelhodc_858+
BDExp_r$Dist_Concelhodc_861+
BDExp_r$Dist_Concelhodc_869+
BDExp_r$Dist_Concelhodc_886+
BDExp_r$Dist_Concelhodc_896+
BDExp_r$Dist_Concelhodc_910+
BDExp_r$Dist_Concelhodc_933+
BDExp_r$Dist_Concelhodc_943+
BDExp_r$Dist_Concelhodc_950+
BDExp_r$Dist_Concelhodc_971+
BDExp_r$Dist_Concelhodc_982+
BDExp_r$Dist_Concelhodc_984
```

```
lm5<-BDExp_r$yboxcox5~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
  BDExp_r$Sinistros1sin_0anos_1anos+
  BDExp_r$Sinistros1sin_0anos_2anos+
  BDExp_r$Sinistros1sin_0anos_3anos+
  BDExp_r$Sinistros1sin_0anos_4anos+
  BDExp_r$Sinistros1sin_0anos_5anos+
  BDExp_r$Sinistros1sin_1anos_5anos+
  BDExp_r$Sinistros1sin_2anos_9anos+
  BDExp_r$Sinistros1sin_4anos_10anos+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_168+
  BDExp_r$Dist_Concelhodc_17+
  BDExp_r$Dist_Concelhodc_194+
  BDExp_r$Dist_Concelhodc_241+
  BDExp_r$Dist_Concelhodc_25+
  BDExp_r$Dist_Concelhodc_259+
  BDExp_r$Dist_Concelhodc_262+
  BDExp_r$Dist_Concelhodc_305+
  BDExp_r$Dist_Concelhodc_319+
  BDExp_r$Dist_Concelhodc_322+
  BDExp_r$Dist_Concelhodc_344+
  BDExp_r$Dist_Concelhodc_345+
  BDExp_r$Dist_Concelhodc_357+
  BDExp_r$Dist_Concelhodc_359+
  BDExp_r$Dist_Concelhodc_376+
  BDExp_r$Dist_Concelhodc_40+
  BDExp_r$Dist_Concelhodc_421+
  BDExp_r$Dist_Concelhodc_425+
  BDExp_r$Dist_Concelhodc_427+
  BDExp_r$Dist_Concelhodc_437+
  BDExp_r$Dist_Concelhodc_470+
  BDExp_r$Dist_Concelhodc_483+
  BDExp_r$Dist_Concelhodc_491+
```

```
  BDExp_r$Dist_Concelhodc_493+
  BDExp_r$Dist_Concelhodc_530+
  BDExp_r$Dist_Concelhodc_538+
  BDExp_r$Dist_Concelhodc_557+
  BDExp_r$Dist_Concelhodc_565+
  BDExp_r$Dist_Concelhodc_585+
  BDExp_r$Dist_Concelhodc_631+
  BDExp_r$Dist_Concelhodc_641+
  BDExp_r$Dist_Concelhodc_650+
  BDExp_r$Dist_Concelhodc_651+
  BDExp_r$Dist_Concelhodc_654+
  BDExp_r$Dist_Concelhodc_656+
  BDExp_r$Dist_Concelhodc_672+
  BDExp_r$Dist_Concelhodc_710+
  BDExp_r$Dist_Concelhodc_737+
  BDExp_r$Dist_Concelhodc_771+
  BDExp_r$Dist_Concelhodc_789+
  BDExp_r$Dist_Concelhodc_796+
  BDExp_r$Dist_Concelhodc_807+
  BDExp_r$Dist_Concelhodc_858+
  BDExp_r$Dist_Concelhodc_861+
  BDExp_r$Dist_Concelhodc_869+
  BDExp_r$Dist_Concelhodc_886+
  BDExp_r$Dist_Concelhodc_896+
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984


lm6<-BDExp_r$yboxcox6~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
  BDExp_r$Sinistros1sin_0anos_1anos+
  BDExp_r$Sinistros1sin_0anos_2anos+
```

```
BDExp_r$Sinistros1sin_0anos_3anos+
BDExp_r$Sinistros1sin_0anos_4anos+
BDExp_r$Sinistros1sin_0anos_5anos+
BDExp_r$Sinistros1sin_1anos_5anos+
BDExp_r$Sinistros1sin_2anos_9anos+
BDExp_r$Sinistros1sin_4anos_10anos+BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_137+
BDExp_r$Dist_Concelhodc_141+
BDExp_r$Dist_Concelhodc_168+
BDExp_r$Dist_Concelhodc_17+
BDExp_r$Dist_Concelhodc_194+
BDExp_r$Dist_Concelhodc_241+
BDExp_r$Dist_Concelhodc_25+
BDExp_r$Dist_Concelhodc_259+
BDExp_r$Dist_Concelhodc_262+
BDExp_r$Dist_Concelhodc_305+
BDExp_r$Dist_Concelhodc_319+
BDExp_r$Dist_Concelhodc_322+
BDExp_r$Dist_Concelhodc_344+
BDExp_r$Dist_Concelhodc_345+
BDExp_r$Dist_Concelhodc_357+
BDExp_r$Dist_Concelhodc_359+
BDExp_r$Dist_Concelhodc_376+
BDExp_r$Dist_Concelhodc_40+
BDExp_r$Dist_Concelhodc_421+
BDExp_r$Dist_Concelhodc_425+
BDExp_r$Dist_Concelhodc_427+
BDExp_r$Dist_Concelhodc_437+
BDExp_r$Dist_Concelhodc_470+
BDExp_r$Dist_Concelhodc_483+
BDExp_r$Dist_Concelhodc_491+
BDExp_r$Dist_Concelhodc_493+
BDExp_r$Dist_Concelhodc_530+
BDExp_r$Dist_Concelhodc_538+
BDExp_r$Dist_Concelhodc_557+
BDExp_r$Dist_Concelhodc_565+
BDExp_r$Dist_Concelhodc_585+
BDExp_r$Dist_Concelhodc_631+
BDExp_r$Dist_Concelhodc_641+
BDExp_r$Dist_Concelhodc_650+
BDExp_r$Dist_Concelhodc_651+
BDExp_r$Dist_Concelhodc_654+
BDExp_r$Dist_Concelhodc_656+
BDExp_r$Dist_Concelhodc_672+
BDExp_r$Dist_Concelhodc_710+
BDExp_r$Dist_Concelhodc_737+
BDExp_r$Dist_Concelhodc_771+
BDExp_r$Dist_Concelhodc_789+
BDExp_r$Dist_Concelhodc_796+
BDExp_r$Dist_Concelhodc_807+
BDExp_r$Dist_Concelhodc_858+
BDExp_r$Dist_Concelhodc_861+
BDExp_r$Dist_Concelhodc_869+
BDExp_r$Dist_Concelhodc_886+
BDExp_r$Dist_Concelhodc_896+
```

```
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984


lm7<-BDExp_r$yboxcox7~
  BDExp_r$CategoriaCaminheta+
  BDExp_r$CategoriaLigeiroParticular+
  BDExp_r$CategoriaMisto+
  BDExp_r$CategoriaMonovolume+
  BDExp_r$CategoriaPickup+
  BDExp_r$Idveiculo+
  BDExp_r$IdCondutor+
  BDExp_r$IdCarta+
  BDExp_r$SexoFeminino+
  BDExp_r$Sinistros0sin_10anos_15anos+
  BDExp_r$Sinistros0sin_15anos_15anos+
  BDExp_r$Sinistros0sin_2anos_2anos+
  BDExp_r$Sinistros0sin_4anos_4anos+
  BDExp_r$Sinistros0sin_5anos_10anos+
  BDExp_r$Sinistros0sin_5anos_12anos+
  BDExp_r$Sinistros0sin_5anos_5anos+
  BDExp_r$Sinistros0sin_5anos_6anos+
  BDExp_r$Sinistros0sin_5anos_7anos+
  BDExp_r$Sinistros0sin_5anos_8anos+
  BDExp_r$Sinistros0sin_7anos_9anos+
  BDExp_r$Sinistros1sin_0anos_1anos+
  BDExp_r$Sinistros1sin_0anos_2anos+
  BDExp_r$Sinistros1sin_0anos_3anos+
  BDExp_r$Sinistros1sin_0anos_4anos+
  BDExp_r$Sinistros1sin_0anos_5anos+
  BDExp_r$Sinistros1sin_1anos_5anos+
  BDExp_r$Sinistros1sin_2anos_9anos+
  BDExp_r$Sinistros1sin_4anos_10anos+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_137+
  BDExp_r$Dist_Concelhodc_141+
  BDExp_r$Dist_Concelhodc_168+
  BDExp_r$Dist_Concelhodc_17+
  BDExp_r$Dist_Concelhodc_194+
  BDExp_r$Dist_Concelhodc_241+
  BDExp_r$Dist_Concelhodc_25+
  BDExp_r$Dist_Concelhodc_259+
  BDExp_r$Dist_Concelhodc_262+
  BDExp_r$Dist_Concelhodc_305+
  BDExp_r$Dist_Concelhodc_319+
  BDExp_r$Dist_Concelhodc_322+
  BDExp_r$Dist_Concelhodc_344+
  BDExp_r$Dist_Concelhodc_345+
  BDExp_r$Dist_Concelhodc_357+
  BDExp_r$Dist_Concelhodc_359+
```

```
  BDExp_r$Dist_Concelhodc_376+
  BDExp_r$Dist_Concelhodc_40+
  BDExp_r$Dist_Concelhodc_421+
  BDExp_r$Dist_Concelhodc_425+
  BDExp_r$Dist_Concelhodc_427+
  BDExp_r$Dist_Concelhodc_437+
  BDExp_r$Dist_Concelhodc_470+
  BDExp_r$Dist_Concelhodc_483+
  BDExp_r$Dist_Concelhodc_491+
  BDExp_r$Dist_Concelhodc_493+
  BDExp_r$Dist_Concelhodc_530+
  BDExp_r$Dist_Concelhodc_538+
  BDExp_r$Dist_Concelhodc_557+
  BDExp_r$Dist_Concelhodc_565+
  BDExp_r$Dist_Concelhodc_585+
  BDExp_r$Dist_Concelhodc_631+
  BDExp_r$Dist_Concelhodc_641+
  BDExp_r$Dist_Concelhodc_650+
  BDExp_r$Dist_Concelhodc_651+
  BDExp_r$Dist_Concelhodc_654+
  BDExp_r$Dist_Concelhodc_656+
  BDExp_r$Dist_Concelhodc_672+
  BDExp_r$Dist_Concelhodc_710+
  BDExp_r$Dist_Concelhodc_737+
  BDExp_r$Dist_Concelhodc_771+
  BDExp_r$Dist_Concelhodc_789+
  BDExp_r$Dist_Concelhodc_796+
  BDExp_r$Dist_Concelhodc_807+
  BDExp_r$Dist_Concelhodc_858+
  BDExp_r$Dist_Concelhodc_861+
  BDExp_r$Dist_Concelhodc_869+
  BDExp_r$Dist_Concelhodc_886+
  BDExp_r$Dist_Concelhodc_896+
  BDExp_r$Dist_Concelhodc_910+
  BDExp_r$Dist_Concelhodc_933+
  BDExp_r$Dist_Concelhodc_943+
  BDExp_r$Dist_Concelhodc_950+
  BDExp_r$Dist_Concelhodc_971+
  BDExp_r$Dist_Concelhodc_982+
  BDExp_r$Dist_Concelhodc_984

#equações com OLS
reg1<-lm(lm1)
reg2<-lm(lm2)
reg3<-lm(lm3)
reg4<-lm(lm4)
reg5<-lm(lm5)
reg6<-lm(lm6)
reg7<-lm(lm7)

summary(reg1)
summary(reg2)
summary(reg3)
summary(reg4)
summary(reg5)
summary(reg6)
```

```
summary(reg7)


#equações com SUR
fitsur <- systemfit(list(lm1,lm2,lm3,lm4,lm5,lm6,lm7),"SUR" )
summary(fitsur)
surboxcox<-predict(fitsur)

#previsão SUR (sem ponto de massa)
BDExp_r$ychapeu1<-exp(surboxcox$eq1.pred)
BDExp_r$ychapeu2<-exp(surboxcox$eq2.pred)
BDExp_r$ychapeu3<-exp(surboxcox$eq3.pred)
BDExp_r$ychapeu4<-exp(surboxcox$eq4.pred)
BDExp_r$ychapeu5<-exp(surboxcox$eq5.pred)
BDExp_r$ychapeu6<-exp(surboxcox$eq6.pred)
BDExp_r$ychapeu7<-exp(surboxcox$eq7.pred)


#previsão SUR (com ponto de massa)
##apuramento do ponto de massa

options(digits=5)
massa1<-lm(y1 ~ ychapeu1-1, data=BDExp_r)
massa2<-lm(y2 ~ ychapeu2-1, data=BDExp_r)
massa3<-lm(y3 ~ ychapeu3-1, data=BDExp_r)
massa4<-lm(y4 ~ ychapeu4-1, data=BDExp_r)
massa5<-lm(y5 ~ ychapeu5-1, data=BDExp_r)
massa6<-lm(y6 ~ ychapeu6-1, data=BDExp_r)
massa7<-lm(y7 ~ ychapeu7-1, data=BDExp_r)

massa<-
c(massa1$coefficient[1],massa2$coefficient[1],massa3$coefficient[1],massa4$coefficient[1],massa5$coeffic
ient[1],massa6$coefficient[1],massa7$coefficient[1])
massa

##medir o impacto de nao fazer com SUR
lixo1<-exp(predict(reg1))
lixo2<-exp(predict(reg2))
lixo3<-exp(predict(reg3))
lixo4<-exp(predict(reg4))
lixo5<-exp(predict(reg5))
lixo6<-exp(predict(reg6))
lixo7<-exp(predict(reg7))

lixo_S1<-lm(BDExp_r$y1 ~ lixo1-1)
lixo_S2<-lm(BDExp_r$y2 ~ lixo2-1)
lixo_S3<-lm(BDExp_r$y3 ~ lixo3-1)
lixo_S4<-lm(BDExp_r$y4 ~ lixo4-1)
lixo_S5<-lm(BDExp_r$y5 ~ lixo5-1)
lixo_S6<-lm(BDExp_r$y6 ~ lixo6-1)
lixo_S7<-lm(BDExp_r$y7 ~ lixo7-1)
lixo_S<-
c(lixo_S1$coefficient[1],lixo_S2$coefficient[1],lixo_S3$coefficient[1],lixo_S4$coefficient[1],lixo_S5$co
efficient[1],lixo_S6$coefficient[1],lixo_S7$coefficient[1])

##previsão SUR (com ponto de massa)
```

```
BDExp_r$ychapeuchapeu1<-BDExp_r$ychapeu1*massa[1]
BDExp_r$ychapeuchapeu2<-BDExp_r$ychapeu2*massa[2]
BDExp_r$ychapeuchapeu3<-BDExp_r$ychapeu3*massa[3]
BDExp_r$ychapeuchapeu4<-BDExp_r$ychapeu4*massa[4]
BDExp_r$ychapeuchapeu5<-BDExp_r$ychapeu5*massa[5]
BDExp_r$ychapeuchapeu6<-BDExp_r$ychapeu6*massa[6]
BDExp_r$ychapeuchapeu7<-BDExp_r$ychapeu7*massa[7]

##avaliação da qualidade do modelo com previsão SUR & com ponto de massa
cor(BDExp_r$ychapeuchapeu1,BDExp_r$y1)^2
cor(BDExp_r$ychapeuchapeu2,BDExp_r$y2)^2
cor(BDExp_r$ychapeuchapeu3,BDExp_r$y3)^2
cor(BDExp_r$ychapeuchapeu4,BDExp_r$y4)^2
cor(BDExp_r$ychapeuchapeu5,BDExp_r$y5)^2
cor(BDExp_r$ychapeuchapeu6,BDExp_r$y6)^2
cor(BDExp_r$ychapeuchapeu7,BDExp_r$y7)^2

##avaliação da qualidade do modelo com previsão SUR & com ponto de massa
cor(lixo1,BDExp_r$y1)^2
cor(lixo2,BDExp_r$y2)^2
cor(lixo3,BDExp_r$y3)^2
cor(lixo4,BDExp_r$y4)^2
cor(lixo5,BDExp_r$y5)^2
cor(lixo6,BDExp_r$y6)^2
cor(lixo7,BDExp_r$y7)^2


oldpar <- par(oma=c(0,0,4,0), mfrow=c(2,4))
  plot(BDExp_r$ychapeuchapeu1,BDExp_r$y1)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu2,BDExp_r$y2)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu3,BDExp_r$y3)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu4,BDExp_r$y4)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu5,BDExp_r$y5)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu6,BDExp_r$y6)
    abline(0, 1)

  plot(BDExp_r$ychapeuchapeu7,BDExp_r$y7)
    abline(0, 1)

par(oldpar)
```