4
Metodologia

Serão apresentadas duas formas de se estimar a persistência. A primeira é de forma mais agregada e se utiliza de dados em forma de triângulos de *run-off* e é conhecida como *Chain Ladder*, uma técnica comum para se estimar provisão de sinistros ocorridos e não avisados. Como este é um método muito subjetivo, em seguida é aplicado a técnica de Mínimos Quadrados para se otimizar o método de *Chain Ladder*.

A segunda forma busca inspiração na tarifação de seguros de automóveis, esta é uma forma menos agregada de se estimar a persistência.

Os modelos matemáticos podem ser classificados em:

- Modelos dinâmicos e estáticos: Dependentes do tempo e independentes do tempo
- Modelos determinísticos e estocásticos: Modelos determinísticos são aqueles que têm uma única resposta, enquanto que nos modelos estocásticos (ou probabilísticos) possíveis respostas dependem de uma distribuição de probabilidade. A distribuição normal é a distribuição de probabilidade mais empregada nos modelos estocásticos. Para construir um modelo estocástico basta adicionar um fator de variação aleatória (e_i). Os modelos estocásticos permitem a estimação dos parâmetros de dispersão, os quais medem a variabilidade dos indivíduos que compõem a população.
- Modelos lineares e não lineares: Um modelo é considerado linear quando todos os parâmetros estão linearmente dispostos, mesmo quando existirem termos elevados ao quadrado, ao cubo etc. Assim, um exemplo de modelo linear é o modelo de regressão polinomial:

$$Y_{ij} = b_0 + b_1X_i + b_2X_{i2} + e_{ij} \qquad (12)$$

Um exemplo de modelo não-linear é a função de Gompertz (Gous, 1986, 1998; Hruby et al., 1994; Emmans, 1995), cuja expressão é dada por:

$$Wt = W_0\exp\{(L/K)(1\text{-}e\text{-}kt)\} \qquad (8)$$

Supondo Ft o percentual de saídas do plano, os modelos estudados podem ser classificados como:

Chain ladder: estático, não-linear e determinístico

$$Ft = \alpha / (\beta + t) \qquad (9)$$

Mínimos quadrados: estático, não-linear e estocástico

$$Ft = \alpha / (\beta + t) + e_t \qquad (10)$$

Modelos lineares generalizados: estático, não-linear e estocástico

$$Ft = \text{Poisson} \qquad (11)$$

4.1
Chain Ladder

Um breve histórico deste método de acordo com o artigo do Verral (2002): Murphy (1994) considerou *Chain Ladder* num contexto com Regressão Linear; Renshaw e Verral (1998) não foram os primeiros a verificar a relação da técnica *Chain Ladder* com a distribuição de Poisson, mas foram os primeiros a implementar. Mack (1991) mostrou que as estimativas do *Chain Ladder* podem ser obtidas maximizando a verossimilhança de Poisson pelo método dos marginais totais. Uma discussão sobre a base estocástica deste método pode ser encontrada em Mack (1994), Verral (2002), Mack e Venter (2000) e Verral e England (2000).

Os dados em forma de triângulo permitem o estudo dos seguintes questões:

- Na coluna tem-se a evolução das saídas, o *run-off* , cada coluna representa a distância entre ocorrência e aviso;
- Na linha tem-se a evolução da implantação, pode-se acompanha através das linhas do triângulo a tendência das saídas em função da época de implantação, assim como, se a carteira está aumentando ou se está em extinção, ou seja, não é mais comercializada;
- E por último, tem-se a evolução de ativos na diagonal do triângulo, que são a fotografia atual da carteira.

Partindo do triângulo de sinistros acumulados abaixo temos a seguinte situação: na coluna parte-se da ocorrência (coluna zero) e em seguida para a célula (i,j) que contém a quantidade de avisos até i períodos depois.

Tabela 5: Sinistros acumulados

Mês	0	1	2	3	4
1	300	390	398	400	400
2	400	430	445	500	
3	350	376	399		
4	330	340			
5	380				

Para se estimar o valor final dos sinistros, no geral, esta técnica estima uma taxa de evolução dos sinistros como o exemplo simplificado abaixo, onde vamos acumulando os avisos de sinistros com ocorrência num mesmo período, origem.

Tabela 6: Taxa de evolução dos sinistros

Mês	0-1	1-2	2-3	3-4
1	1,30	1,02	1,01	1,00
2	1,08	1,03	1,12	
3	1,07	1,06		
4	1,03			
Média	1,12	1,04	1,06	1,00

Adaptando a idéia acima, ao invés de acumular avisos de sinistros, se retirarmos as saídas em função de cancelamentos / resgate, temos um novo triângulo, porém de decréscimos. No triângulo abaixo temos a seguinte situação: na coluna, parte-se da implantação (coluna zero) e em seguida para a célula (i,j) que contém a quantidade de ativos que foram implantados no período i e que continuam ativos j períodos depois.

Tabela 7: Quantidade de inscrições ao longo do tempo

Mês	0	1	2	3	4
1	300	290	275	273	273
2	400	350	325	300	
3	350	325	300		
4	330	300			
5	380				

A Persistência é o percentual de ativos na célula (i,j) com relação a implantação. Também se pode estimar as saídas ou a persistência com relação ao período anterior. Com este tipo de informação, se consegue estimar quanto tempo mais um participante deverá permanecer no plano em função da sua implantação e do tempo em que está no plano. O estudo apresentará as saídas como função do tempo de permanência no plano. A idéia de se utilizar o triângulo de *run-off* para estimar a persistência de forma agrupada foi apontada por André di Montiny[1].

Tabela 8: Persistência com relação a implantação

Mês	Implantação	0-1	1-2	2-3	3-4
1	100%	97%	92%	91%	91%
2	100%	88%	81%	75%	
3	100%	93%	86%		
4	100%	91%			
5	100%				

Pode-se buscar estimar a média das taxas de saídas através da fórmula a ser vista no item 2, ou com outros métodos. O método de *Chain Ladder* consiste em uma análise quase que visual dos dados, verifica-se o que se pretende com o resultado, se desejo usar as informações mais recentes ou com maior histórico, e analisa-se a existência de *outliers,* ou seja, informações fora do padrão que podem prejudicar as estimativas.

[1]Conforme trabalho realizado por André di Montiny em 2004. Tanto o agrupamento de forma triangular como a fórmula por partes foram utilizadas em seu trabalho.

4.2
Cain Ladder com Mínimos Quadrados

Para se estimar o conteúdo de cada célula modela-se uma curva aos dados obtidos para se verificar os valores para prazos longos. Algumas possíveis curvas de saídas, no caso por morte, podem ser encontradas no material de tábuas de mortalidade do Kaizô (2004), assim como as curvas de taxa de resgate apresentadas no item 3.2.

Um modelo proposto para persistência nos planos de sobrevivência / risco segue com a curva apresentada abaixo[1], ou seja, a curva será por partes, de acordo com o intervalo de tempo, pois cada intervalo possui um comportamento particular :

$$F_t = (\alpha / (\beta + t)) + \gamma \qquad (12)$$

Onde,

$$\begin{cases} \gamma = \gamma_1,\ 1 \leq t \leq 5; \\ \gamma = 0,\ 6 \leq t \leq 23 \text{ e } 36 < t < n; \\ \gamma = \gamma_2,\ 24 \leq t \leq 36 \end{cases}$$

F - probabilidade de saída da carteira;

t - tempo decorrido em meses;

$\gamma 1$ - Parâmetro utilizado para confirmação da primeira venda (intervenção no modelo), onde a desistência é mais acentuada;

$\gamma 2$ - Parâmetro utilizado para período após carência para resgate (intervenção no modelo), onde a desistência é mais acentuada;

α - Parâmetro utilizado para a concavidade da função;

β - Parâmetro utilizado para a posição da função no eixo y;

n – número de períodos disponíveis na amostra.

A persistência da carteira é obtida através da seguinte fórmula:

$$P_t = P_{t-1} * (1 - F_t), \text{ com } P_0 = 100\%. \qquad (13)$$

Note que o modelo toma por base uma função complexa, sem solução analítica. P_0 é 100% pois os dados são do status no final do mês, ou seja, somente serão avaliadas as propostas que foram implantadas, as negadas não serão

analisadas. Os dados não mostram as propostas que entraram e não se confirmaram dentro do próprio mês, pois o proponente paga a primeira contribuição no ato de assinatura da proposta e só é cancelado após 3 meses de atraso.

Para escolhermos os melhores parâmetros: α, β, γ_1 e γ_2, podemos utilizar o método de Mínimos Quadrados que propõe parâmetros que minimizem a soma dos erros ao quadrado.

Para os nossos dados temos:

$$\text{Min} \sum_{1}^{n-1} (P_t - \hat{P}_t)^2 \qquad (14)$$

Restrição: Para t= 0, $P_0 = 100\%$

O modelo acima foi utilizado sobre a média das persistências obtida com os triângulos de *run-off*.

Para minimizar a função acima foi utilizada a ferramenta *solver* do MS Excel.

Esta forma de modelagem é simples e, conforme veremos ver a seguir, eficaz. Também é de fácil entendimento, o que facilita a aceitação desta forma de modelagem por parte da área estratégica da empresa.

4.3
Modelos Lineares Generalizados

A forma de modelagem que vai ser apresentada nesta seção é muito utilizada para tarifação de seguros de automóveis, onde para calcular o prêmio estima-se primeiro o número de sinistros e depois a severidade (valor médio do sinistro).

A classe de Modelos Lineares Generalizados é uma extensão dos modelos lineares tradicionais que permite a média da população depender de um preditor linear através de uma função de ligação não linear e permite a probabilidade da variável ser estimada por qualquer membro da família exponencial. Muitos dos modelos estatísticos utilizados são modelos lineares generalizados.

Um modelo linear tradicional possui a seguinte forma:

$$y_i = x_i'\beta + e_i \qquad (15)$$

Onde y_i é a variável resposta para a i-ésima informação. A quantidade x_i é o vetor coluna de variáveis explicativas para a observação i que é conhecida. O vetor de coeficientes β é estimado por mínimos quadrados para modelar y. Assume-se que os erros e_i sejam variáveis aleatórias independentes e normalmente distribuídas com média zero e variância constante. O valor esperado de y_i, denotado por m_i, é dado por:

$$m_i = x_i'\beta \qquad (16)$$

Enquanto modelos lineares tradicionais são extensivamente utilizados em análises estatísticas, existem alguns problemas nos quais eles não são apropriados:

- Pode não ser razoável assumir que os dados são normalmente distribuídos. Por exemplo: a distribuição normal, que é contínua, talvez não seja adequada para modelar contagens ou medidas de proporções, que são distribuições discretas;
- Se a média dos dados é naturalmente restrita a um intervalo de valores, o modelo linear tradicional pode não ser apropriado, pois o preditor linear $x_i'\beta$ pode resultar qualquer valor. Por exemplo: a medida de uma medida de proporção está entre 0 e 1, mas o preditor linear da média num modelo linear tradicional não está restrito a este intervalo;
- Pode não ser realista que a variância dos dados seja constante para todas as observações. Por exemplo: é comum observar dados onde a variância aumenta de acordo com a média.

Um modelo linear generalizado estende o modelo linear tradicional e é aplicável a problemas de análise de dados mais genéricos. Um modelo linear generalizado consiste nos seguintes componentes:

- Componente linear definido na mesma forma do modelo linear tradicional:

$$n_i = x_i'\beta \qquad (17)$$

- Função de ligação monotônica diferenciável g que descreve como o valor esperado de y_i está relacionado ao preditor linear n_i:

$$g(m_i) = x_i'\beta \qquad (18)$$

- Variável resposta y_i independente para i = 1, 2,...com distribuição de probabilidade da família exponencial, ver anexo sobre "distribuições

pertencentes a família exponencial". Isto implica que a variância da variável resposta depende da média através de uma função de variância V:

$$var(yi) = \phi V(m_i)/w_i \qquad (19)$$

Onde ϕ é uma constante e w_i é um peso conhecido para cada observação. O parâmetro de dispersão ϕ também é conhecido, por exemplo, para a distribuição binomial, ou pode ser estimado.

Assim como para os modelos lineares tradicionais, modelos lineares generalizados podem ser resumidos através de estatísticas como estimativas dos parâmetros, seus erros padrões e estatísticas de ajuste do modelo.

Também se pode fazer inferência estatística sobre os parâmetros utilizando intervalos de confiança e testes de hipóteses. Porém, procedimentos específicos de inferência são usualmente baseados em considerações assintóticas, pois não há teoria de distribuição exata ou não é prática para qualquer modelo linear generalizado.

4.3.1
Exemplos de Modelos Lineares Generalizados

Um modelo linear generalizado é construído se escolhendo a variável resposta e as explicativas de acordo com os dados e escolhendo uma função de ligação e distribuição de probabilidade da variável resposta. Alguns exemplos de modelos lineares generalizados seguem abaixo.

Variáveis explicativas podem ser combinações de variáveis contínuas, de classificação ou interações.

- Modelo linear tradicional

  Variável resposta: variável contínua

  Distribuição: normal

  Função de ligação: identidade n = m

- Regressão logística

  Variável resposta: proporção

  Distribuição: binomial

  Função de ligação: logit n = log(m/(1-m))

- Regressão de Poisson no modelo log-linear

  Variável resposta: dados de contagem

  Distribuição: poisson

  Função de ligação: log n = log(m)

- Modelo Gamma com ligação log

  Variável resposta: variável contínua positiva

  Distribuição: gamma

  Função de ligação: log n = log(m)

4.3.2
Procedimento GENMOD

A procedure GENMOD do software SAS modela GLM, como definido por Nelder e Wedderburn (1972).

Este procedimento ajusta modelos lineares generalizados aos dados por estimativa de máxima verossimilhança dos parâmetros β, de forma numérica através de um processo iterativo. O parâmetro de dispersão ϕ também é estimado por máxima verossimilhança. Covariâncias, erros padrões e p-valores são calculados para os parâmetros estimados baseando-se na normalidade assintótica dos estimadores de máxima verossimilhança.

Mudanças nas estatísticas de adequação do modelo são utilizadas para avaliar a contribuição da variável explicativa para o modelo.

O procedimento GENMOD permite ajustar uma seqüência de modelos, começando com um modelo simples até um modelo com o máximo de variáveis explicativas. Esta é a análise do tipo I, que resulta numa tabela com os valores de desvios nas verossimilhanças dos modelos. O resultado deste processo depende da ordem das modelagens.

A análise do tipo III não depende da ordem em que as variáveis são especificadas. Esta análise consiste em calcular a estatística raio de verossimilhança e estatística de Wald para os contrastes entre o termos para um determinado modelo.

Exemplo com Regressão de Poisson:

A distribuição de Poisson pode ser usada para modelar uma distribuição de contagens numa tabela de contingência. Vamos utilizar este método para modelar quantidade de sinistros para seguros de automóveis. Suponha uma experiência de sinistros classificada por dois fatores: idade (duas faixas etárias, ou seja, dois níveis) e tipo de carro (divido em três níveis: pequeno, médio e grande), conforme tabela abaixo.

Tabela 9: Dados de sinistro de automóveis

Exposição	**Número de sinistros**	**Tipo de carro**	**Faixa etária**
500	42	Pequeno	1
1200	37	Médio	1
100	1	Grande	1
400	101	Pequeno	2
500	73	Médio	2
300	14	Grande	2

Assuma que o número de sinistros possui distribuição Poisson e que a média da distribuição m_i está relacionada aos fatores tipo de carro (C) e faixa etária do segurado (I) para a observação i através da função de ligação log:

$$\log(m_i) = \log(\text{exposição}) + \beta_0 + C_i(1)\,\beta_1 + C_i(2)\,\beta_2 + C_i(3)\,\beta_3 + I_i(1)\,\beta_4 + Ii(2)\,\beta_5 \qquad (20)$$

$C_i(j)$ e $I_i(j)$ são variáveis indicadoras associadas ao j-ésimo nível dos fatores tipo de carro (C) e faixa etária do segurado (I):

$$Ci(j) = \begin{cases} 1 \text{ se } C = j \\ \\ 0 \text{ c.c.} \end{cases} \qquad (21)$$

Para a observação i. Os βs são parâmetros desconhecidos que serão estimados. O logaritmo da exposição é utilizado como variável *offset*, ou seja, variável da regressão com coeficiente 1. A relação log linear entre a média e os

fatores é especificada pela função de ligação. A função de ligação log garante que a média do número de sinistros para cada classe de carro e idade estimado seja positivo.

O código abaixo fornece quatro tipos de resultados que serão avaliados a seguir:

```
proc genmod data=insure;
class carro idade;
model c = carro idade /          dist = poisson
                                 link = log
                                 offset = ln
                                 type1
                                 type3;
run;
```

Tabela 10: Ajuste do modelo (*Criteria for assessing goodness of fit*)

Criterion	DF	Value	Value/DF
Deviance	2	2,8207	1,4103
Scaled deviance	2	2,8207	1,4103
Pearson Chi-square	2	2,8416	1,4208
Scaled Pearson X2	2	2,8416	1,4208
Log likelihood	.	837,4533	.

A tabela acima contém as estatísticas de um modelo específico. Estas estatísticas são em avaliar o ajuste do modelo e para comparação com outros modelos. Ao se comparar o desvio 2,8207 com seu qui-quadrado assintótico com 2 graus de liberdade, se encontra o p-valor de 0,24. Isto indica que o modelo específico ajusta razoavelmente bem os dados.

Tabela 11: Análise dos parâmetros estimados (*Analysis of parameter estimates*)

Parameter	DF	Estimate	Std error	X2	Pr>Chi
Intercept	1	-1,3168	0,09	212,73	0,0000
Carro - pequeno	1	-1,7643	0,27	41,96	0,0000
Carro – médio	1	-0,6928	0,13	29,18	0,0000
Carro - pequeno	0	0	0	.	.
Idade -1	1	-1,3199	0,14	94,34	0,0000
Idade -2	0	0	0	.	.

Esta tabela apresenta para cada parâmetro do modelo os graus de liberdade, o valor estimado, o erro padrão, a estatística qui-quadrado de Wald e o p-valor associado para testar a significância do parâmetro para o modelo.

Tabela 12: Análise tipo I (*LR statistics for type1 analysis*)

Source	Deviance	DF	X2	Pr>Chi
Intercept	175,15	0	.	.
Carro	107,46	2	67,69	0,0000
Idade	2,82	1	104,64	0,0000

Na análise do tipo I cada entrada na coluna desvio representa o desvio do modelo contendo o efeito daquela linha e todos os efeitos precedentes na tabela, ou seja, o desvio correspondente a carro é o desvio do modelo contendo o intercepto e carro. Quanto mais termos são incluídos no modelo o desvio diminui.

Entradas na coluna qui-quadrado são os raios de verossimilhança (*LR statistics*) para testar a significância do efeito adicionado ao modelo contendo todos os efeitos anteriores. O valor 67,69 para carro representa a diferença entre os desvios do modelo apenas com o intercepto e outro com o intercepto e a variável carro. O p-valor (Pr>Chi) ser zero significa que esta variável é altamente significante para o modelo.

Tabela 13: Análise tipo III (*LR statistics for type3 analysis*)

Source	DF	X2	Pr>Chi
Carro	2	72,82	0,0000
Idade	1	104,64	0,0000

A análise do tipo III leva as mesmas conclusões que a análise do tipo I. O qui-quadrado do tipo III para carro, por exemplo, é a diferença entre o desvio do modelo com intercepto, carro e idade e o modelo sem carro. A hipótese testada neste caso é a significância de carro no modelo que já possua idade. Os valores do raio de verossimilhança para carro e idade indicam que ambos os fatores são altamente significantes em determinar o comportamento dos sinistros de seguros de automóveis.

A análise feita par o exemplo acima será a mesma a ser feita nos modelos de persistência.