UNIVERSIDADE ABERTA

ANÁLISE COMPORTAMENTAL – UM ESTUDO POR ANÁLISE DE SOBREVIVÊNCIA, APLICADA A CARTÕES DE CRÉDITO

JOÃO CARLOS DE LIMA CORREIA DE ALBERGARIA CRESPO

MESTRADO EM BIOESTATÍSTICA E BIOMETRIA

ORIENTAÇÃO: Professor Doutor Pedro Miguel Picado de Carvalho Serranho

Fevereiro / 2016

ANÁLISE COMPORTAMENTAL – UM ESTUDO POR ANÁLISE DE SOBREVIVÊNCIA, APLICADA A CARTÕES DE CRÉDITO

João Carlos de Lima Correia de Albergaria Crespo

Mestrado em Bioestatística e Biometria

Orientador: Professor Doutor Pedro Serranho

RESUMO

A capacidade de adaptação e rapidez de decisão, distinguem as empresas que melhor conseguem competir e crescer no mercado global. Para atuar rapidamente, as organizações precisam de sistemas de informação cada vez mais eficazes, surgindo recentemente uma nova função considerada fundamental para as empresas, que é a de Cientista de Dados. É neste contexto e para responder aos desafios atuais e futuros, que surgem sistemas de informação cada vez mais avançados, suportados por modelos de análise e visualização estatística. Este trabalho consiste em criar uma metodologia de desenvolvimento de modelos de previsão de incumprimento e perfil do consumidor, aplicado a cartões de crédito, com base numa exposição de análise comportamental, utilizando técnicas de análise de sobrevivência. São definidas técnicas de tratamento dos dados recolhidos, estimado modelo não-paramétrico de Kaplan-Meier e vários modelos de Cox de riscos proporcionais. Com recurso à curva ROC, dependente do tempo, à AUC e ao índice de Gini, conclui-se que o modelo final apresenta um desempenho positivo para identificar os clientes em situação de incumprimento ou com propensão a incumprir.

Palavras-Chave: Análise de Sobrevivência, Kaplan-Meier, Modelo de Cox, Riscos Proporcionais e Cartões de Crédito.

BEHAVIOURAL ANALYSIS – A STUDIE WITH SURVIVAL ANALYSIS MODELS, APPLIED TO CREDIT CARDS

João Carlos de Lima Correia de Albergaria Crespo

Master Degree in Biostatistics and Biometrics

Advisor: Professor Doutor Pedro Serranho

ABSTRACT

The adaptability and decision time distinguish companies that can better compete and grow in the global market. To act quickly, organizations need information systems increasingly effective, to achieve that goal there is a new job profile and role considered essential for companies, which is the Data Scientist. It is in this context and to respond to current and future challenges, arising from information systems increasingly advanced, supported by models of visualization and statistical analysis. The aim of this work is to create a development methodology of default prediction models and consumer behavior, applied to credit cards, based on empirical exposure of behavioral analysis using survival analysis techniques. Data processing techniques are defined and the Kaplan-Meier non-parametric model and several Cox proportional hazards models are estimated. Using the time dependent ROC curve, the AUC and the Gini coefficient, one is led to conclude that the final model shows a positive performance to identify customers in default or likely to defaulting.

Keywords: Survival Analysis, Kaplan-Meier, Cox Model, Proportional Hazards, Credit Cards.

À Dária e à Inês

AGRADECIMENTOS

O meu especial agradecimento é para o Senhor Professor Doutor Pedro Serranho, foi através da sua constante disponibilidade e incentivo que me foi possível ultrapassar obstáculos e dificuldades.

Agradeço, igualmente, todo o apoio e incentivo da minha Família, ao longo destes meses, para que esta dissertação se tornasse uma realidade.

ÍNDICE

1 Introdução e Revisão de Literatura 3

1.1 Introdução 3

1.1.1 Organização da Dissertação 6

1.2 Revisão de Literatura 7

2 Recolha e Tratamento de Dados 13

2.1 Recolha e preparação de dados 14

2.2 Análise estatística das variáveis 18

2.2.1 *Weight of Evidence* (WoE) 18

2.2.2 Z-score 19

2.2.3 *Information Value* (IV) 20

2.2.4 Coeficiente de Correlação de Spearman 21

2.2.5 Caracterização dos dados de sobrevivência 22

3 Análise de Sobrevivência 27

3.1 As funções de Sobrevivência e Taxa de Falha ou Hazard 27

3.2 Modelo não-paramétrico de Kaplan-Meier 29

3.2.1 O teste *Logrank* 31

3.3 A Especificação do Modelo de Cox de Riscos Proporcionais 31

3.3.1 Interpretação dos Coeficientes 33

3.3.2 Avaliação da Proporcionalidade 33

3.3.3 Critério Akaike (AIC) 35

3.3.4 Teste do Rácio das Verossimilhanças (TRV) 36

3.3.5 Resíduos, Forma Funcional e Valores Atípicos 36

3.4 A Especificação do Modelo de Aalen 39

3.5 Avaliação da Performance do Modelo 40

3.5.1 Área ROC (AUC) 40

3.5.2 Índice de Gini 42
4 Apresentação de Resultados 45
4.1 Resultados do modelo não-paramétrico de Kaplan-Meier 45
4.2 Resultados da estimação do modelo de Cox de riscos proporcionais 47
4.2.1 Capacidade Preditiva das Covariáveis 47
4.2.2 Avaliação da Proporcionalidade dos Riscos de Falha 48
4.2.3 Estimação de Cox de Riscos Proporcionais 54
4.3 Avaliação da performance do modelo de Cox de riscos proporcionais 64
5 Conclusões e Perspetivas Futuras 69
5.1 Revisão da metodologia seguida 69
5.2 Principais conclusões 71
5.3 Trabalho futuro 72
Referências Bibliográficas 75
ANEXOS 79

ÍNDICE DE TABELAS

Tabela 2.1: Transformações logarítmicas e de raiz quadrada 17
Tabela 2.2: Descrição das covariáveis 17
Tabela 2.3: Correlação de Spearman 21
Tabela 3.1: Valores de referência para a AUC 42
Tabela 3.2: Valores de referência para o Índice de Gini 42
Tabela 4.1: Estimativa de Kaplan-Meier 46
Tabela 4.2: Estimação Cox das Covariáveis 47
Tabela 4.3: Teste de Proporcionalidade de Falha no Modelo de Cox 48
Tabela 4.4: Teste *Logrank* de comparação entre grupos 52
Tabela 4.5: Constituição das amostras de Treino e Teste 54
Tabela 4.6: Regressão inicial do Modelo de Cox (modelo irrestrito) 55
Tabela 4.7: Seleção de covariáveis usando o modelo de regressão de Cox 55
Tabela 4.8: Testes de proporcionalidade no modelo ajustado 59
Tabela 4.9: Resultados do ajustamento do modelo de Cox para o Modelo 2 59
Tabela 4.10: Resultados da AUC e Gini para as amostras de treino e teste 65

ÍNDICE DE FIGURAS E GRÁFICOS

Figura 2.1: Janela de Amostragem (exemplo para o período 2005/12) 13
Figura 2.2: Ilustração de dados com censura (A) e falha (B) 23

Gráfico 2.1: Correlação de Spearman 22
Gráfico 4.1: Estimativa de Kaplan-Meier 46
Gráfico 4.2: Resíduos de Schoenfeld de x10 - Cartão de Uso Exclusivo em Combustível 49
Gráfico 4.3: Resíduos de Schoenfeld de x18 - Cliente com Crédito Hipotecário 49
Gráfico 4.4: Resíduos de Schoenfeld de x21 - Rácio entre Débitos e Créditos 50
Gráfico 4.5: Logaritmo da função taxa de falha acumulada estimada versus t da x4 51
Gráfico 4.6: Estimativa de Kaplan-Meier de x4 - Habilitações Literárias 51
Gráfico 4.7: Logaritmo da função taxa de falha acumulada estimada versus t de $x4$ 53
Gráfico 4.8: Resíduos padronizados de Schoenfeld de $x4$ - Habilitações Literárias (revista) .. 53
Gráfico 4.9: Resíduos padronizados de Schoenfeld de x16 - Inibição de Uso de Cheque 56
Gráfico 4.10: Estimativas das funções de regressão acumulada de x16 - Inibição de Uso de Cheque 57
Gráfico 4.11: Logaritmo da função taxa de falha acumulada *versus* tempo para x16 57
Gráfico 4.12: Resíduos padronizados de Schoenfeld para as covariáveis do modelo 2 58
Gráfico 4.13: Curvas de sobrevivência 61
Gráfico 4.14: Resíduos martingala e desviância *versus* preditor linear do modelo de Cox final 62
Gráfico 4.15: Gráficos 3D dos resíduos martingala e da desviância 62
Gráfico 4.16: Resíduos DFBETAS *versus* cada covariável presente no modelo de Cox final 63
Gráfico 4.17: Curvas ROC a 12 e 6 meses, do modelo de Cox final 64
Gráfico 4.18: Evolução temporal da AUC para o período de sobrevivência 65

LISTA DE ACRÓNIMOS

ACP	Análise de Componentes Principais
AIC	*Akaike Information Criterion*
AS	Análise de Sobrevivência
AUC	*Area Under Curve*
IV	*Information Value*
LOWESS	*Locally Weighted Scatterplot Smoothing*
ROC	*Receiver Operating Characteristic*
SI	Sistemas de Informação
WoE	*Weight of Evidence*

CAPÍTULO 1

INTRODUÇÃO E REVISÃO DE LITERATURA

1 Introdução e Revisão de Literatura

1.1 Introdução

O apogeu das arquiteturas de sistemas aplicacionais cliente-servidor, registado na década de noventa, numa tentativa de baixar custos orçamentais das Tecnologias de Informação, teve como consequência a multiplicação e dispersão geográfica dos dados e da informação de suporte à decisão, fundamental para as empresas. Com o mito criado em torno do *bug* do ano 2000 e o rápido progresso tecnológico verificado, foi possível reestruturar os sistemas de informação (SI) em torno da centralização da informação, de arquiteturas *datawarehouse* e mais recentemente *Big Data*. Este passo fez com que as entidades públicas e empresas em geral, refletissem sobre a necessidade de analisar os elevados volumes de dados armazenados nos seus sistemas centrais. É no início do séc. XXI, que se inicia uma grande evolução na criação de sistemas de informação estatístico–econométricos, de suporte à criação de modelos de apoio ao negócio e gestão do risco. Na Banca, impulsionada, também, pelo Acordo de Basileia II, são criadas estruturas próprias de suporte à análise de risco, gerando uma elevada capacidade de criação de modelos estatísticos de risco de crédito e comportamento do consumidor.

Seja pela necessidade de prever a probabilidade de incumprimento, através de técnicas de discriminação entre Bons e Maus pagadores, ou pela necessidade em prever a continuidade de empresas no mercado, separando as Boas das Más e identificando características de apoio à previsão probabilística de se manterem ou não no mercado, que há aproximadamente quarenta anos se desenvolvem trabalhos de investigação em *credit scoring*, sendo o estudo realizado por Beaver (1966) pioneiro neste tipo de previsão.

Fundamental nas instituições financeiras, o *credit scoring* consagra o desenvolvimento de elevados conceitos técnicos utilizados, por exemplo, na avaliação de pessoas que se candidatam ao crédito pela primeira vez (*scoring* aplicacional) ou na avaliação de conduta e

ética de utilização de cartões de crédito e respetivos comportamentos de pagamento (*scoring* comportamental). O *credit scoring* tem, então, como objetivo principal criar estimativas das probabilidades dos clientes (não) incumprirem nos contratos de crédito assumidos, permitindo, a definição de critérios de suporte à decisão de atribuição de crédito que vise a maximização das receitas ou a minimização das perdas. O *credit scoring* não estando no âmbito deste trabalho, é um bom exemplo de como a modelação estatística conquistou um espaço importante nas instituições financeiras, sendo uma referência fundamental nos processos de modelação apresentados nesta dissertação.

O objetivo global deste trabalho é criar uma metodologia de desenvolvimento de modelos de previsão de incumprimento e perfil de risco do consumidor, aplicado a cartões de crédito, com base numa exposição de análise comportamental, utilizando técnicas de análise de sobrevivência (AS). Nesse âmbito, analisam-se um conjunto de covariáveis que influenciam a função de sobrevivência, para estimar quando é que um cliente incumprirá. Algumas das principais vantagens deste tipo de modelação, em relação aos modelos binários, são a incorporação de dados censurados, uma avaliação temporal em vez de uma variável de saída dicotómica e num contexto de análise económico-financeira a utilização de variáveis macroeconómicas (Belloti e Crook, 2007).

Os métodos, atualmente, mais usados nas instituições financeiras, como por exemplo a regressão logística, estão associados a alguma instabilidade com amostras pequenas e à incapacidade de tratar eficientemente características não lineares nos dados. Ao contrário da AS, a modelação binária concentra-se, essencialmente, no risco dos devedores. A qualidade e inovação apresentada sobre modelos que compreendem a dinâmica temporal está relacionada com a introdução de fatores económicos e condições dos mercados financeiros, informação que está omissa nas metodologias associadas a modelos binários, também, designados, contrapondo com os dinâmicos, como estáticos.

Foi a área das ciências biomédicas juntamente com a evolução tecnológica ocorrida, que motivou o desenvolvimento desta área da estatística, designada AS. Como referido por Cristina Rocha (1995), a AS engloba um conjunto de métodos e modelos, destinados à análise de dados

de sobrevivência. Este tipo de dados surge, quando é registado para indivíduos de um determinado grupo o tempo decorrido desde um instante inicial, bem definido, até à ocorrência de um acontecimento de interesse (denominado de falha), nomeadamente:

- Morte de indivíduos ou de animais em estudos clínicos;
- Uma recaída depois de ter sido atingido um estado de remissão, em indivíduos que sofrem de doença;
- Falha de componentes mecânicas ou eletrónicas;
- Fim de um período de greve ou de desemprego.

De destacar o facto dos dois artigos mais citados em toda a literatura estatística no período de 1987 a 1989 foram, segundo Colosimo e Giolo (2006), o do estimador Kaplan-Meier para a função de sobrevivência (Kaplan e Meier 1958) e o do modelo de Cox (Cox, 1972), dois dos modelos mais utilizados em AS.

Apesar do seu enquadramento bioestatístico, a AS tem aplicação em áreas tão diversas como economia, física, engenharia, psicologia e *credit scoring* (Stepanova, M. 2001). Existem, por exemplo, várias técnicas de construção de modelos de risco de crédito, no entanto, considera-se a regressão logística como a que reúne maior consenso, conforme referido por Stepanova e Thomas (2001), para modelos binários. O modelo de sobrevivência estabelece uma perceção diferente dos modelos binários, habitualmente utilizados. Os modelos binários indicam uma estimativa da probabilidade de incumprimento e os modelos de sobrevivência definem a probabilidade de incumprimento ao longo do tempo, ou seja, propõe-se a avaliação de quando é que um cliente incumprirá, em vez de analisar-se se entrará em incumprimento ou não. Nesta dissertação, a ocorrência de falha é o evento de *default* (ou incumprimento). A principal característica dos dados apresentados neste estudo é a presença de censura, ou seja, de sujeitos que por alguma razão saem do estudo antes de ocorrer o evento de falha ou que sobrevivem para além da experiência. Estes dados censurados são ignorados pelos modelos binários, sendo a incorporação desta informação uma das principais vantagens da utilização da AS em modelos estatísticos. A eliminação de dados censurados pode enviesar os resultados, pelo que, é desejável a adoção de métodos que os considerem.

O objetivo desta dissertação é encontrar um modelo de AS para definição da probabilidade de tempo até incumprimento para clientes de uma instituição financeira. Para tal, foi fornecida uma amostra de clientes particulares, de cartões de crédito, com um horizonte temporal de 31/07/2005 a 31/05/2007, sendo omissa qualquer possibilidade de identificação dos mesmos.

1.1.1 Organização da Dissertação

O presente capítulo apresenta a introdução e a revisão de literatura onde é efetuada uma exposição da considerada relevante para o enquadramento metodológico utilizado nesta dissertação.

No segundo capítulo são detalhadas as etapas realizadas no processo de recolha e tratamento de dados. A escolha da janela de amostragem tem aqui um papel muito importante, bem como os processos de tratamento e análise de informação que visam definir, com base nos dados selecionados, uma amostra para a AS.

No terceiro capítulo é apresentada uma introdução teórica à AS e à avaliação de performance de um modelo.

O quarto capítulo é dedicado à apresentação de resultados com base em computação estatística realizada em R e nas bibliotecas *{pastecs}, {aplpack}, {survival}, {xtable}, {knitr}, {ggplot2}, {survMisc}, {MASS}, {caTools}, {risksetROC}, {corrplot}* e *{scatterplot3d}*. Com recurso ao algoritmo *stepwise* (*backward* + *forward*) é criado um primeiro modelo que é posteriormente otimizado com base na significância das covariáveis, nos resultados do teste rácio verosimilhanças, proporcionalidade e forma funcional (análise dos resíduos), obtendo-se o modelo final.

Por fim, no capítulo 5 são apresentadas as conclusões e as perspetivas para estudos futuros.

1.2 Revisão de Literatura

O tema desta dissertação prende-se com a análise do comportamento dos clientes de uma instituição bancária portuguesa, na utilização dos seus cartões de crédito, através de AS. A partir desta análise pode ser criada uma metodologia de suporte à comercialização deste produto financeiro, minimizando por esta via a probabilidade de um cliente entrar em incumprimento com a instituição com quem contratou o seu cartão de crédito.

Em 2001, destacam-se, um conjunto de trabalhos sobre a aplicabilidade do modelo de Cox proporcional, ao sector financeiro, são eles, Stepanova, M. e Thomas, L. C. (2001) *PHAB Scores: Proportional Hazards Analysis Behavioural Scores* no qual se propõe uma metodologia de construção de *scoring* comportamental utilizando o modelo de Cox Proporcional comparando a sua performance com a regressão logística e Stepanova, M. (2001) *Using survival analysis methods to build credit scoring models* apresentando, a partir de AS, várias metodologias de construção de modelos de *credit scoring*. Anteriormente em 1999, os mesmos autores publicam o trabalho *Survival analysis methods for personal loan data* propondo métodos de AS para categorização de covariáveis contínuas e aplicação de algumas extensões ao modelo de Cox de riscos proporcionais. Sendo posteriormente publicados inúmeros trabalhos sobre a aplicabilidade de modelos de sobrevivência nas áreas económico-financeiras.

Com o Acordo de Basileia II e Solvência II a gestão de risco e consequentemente o *credit scoring* conquistam uma importância relevante na gestão da atividade bancária e seguradora, sendo a sua aplicabilidade e metodologia apresentadas por Steven Finlay (2010) em *Credit Scoring, Response Modelling and Insurance Rating. A Practical Guide to Forecasting Consumer Behaviour* e Sarmento Baptista, António Manuel (2012) em *Credit Scoring*, Uma Ferramenta de Gestão Financeira, destacando ambos os autores a AS como metodologia a considerar no *credit scoring*.

Mais recentemente, as arquiteturas de informação *Big Data*, sugerem várias metodologias de modelação estatística de suporte à decisão de gestão e à análise preditiva em áreas da banca, como por exemplo, os departamentos comerciais ou de marketing. Das diferentes soluções de

Big Data existentes, são de destacar a da Oracle e da HPE[1] pela analítica baseada em R, confirmando-se assim o R, como uma das principais soluções de computação estatística da atualidade. Tecnologias, de decisão em tempo real – que combinam regras de negócio com análise estatística, que poderão ser utilizadas na interação com o cliente enquanto esta ocorre – revolucionaram as metodologias de modelação estatística e a sua aplicabilidade. Já em 1993, Davidson, R. e MacKinnon, J. anteviam grandes progressos nos processos de modelação estatística associados aos progressos das ciências da computação:

> Com a redução dos custos de computação, é expectável que cada vez mais profissionais de estatística aplicada se voltarão para variantes do *bootstrap*, para lidar com modelos para os quais a teoria assintótica poderá tornar-se inadequada.
>
> (Davidson, R. e MacKinnon, J., 1993:168)

Para esta dissertação foi utilizado o R como ferramenta de computação estatística, livros como *Applied Econometrics with R*, Kleiber, C., Zeileis, A. (2008) e Análise de Sobrevivência Aplicada, Colosimo e Giolo (2006), são boas introduções para o uso do R neste contexto.

A recolha e tratamento de dados é uma das etapas mais importantes de um processo de modelação estatística, permitindo que a partir de dados recolhidos de um conjunto de bases de dados distintas, possam ser transformados numa amostra de sobrevivência. Em 2001, Thomas, L. C., Ho, J. e Scherer, W. T. em *Time will tell: Behavioural Scoring and the Dynamics of Consumer Credit Assessment* apresentam métodos para a definição do período de observação e período resultado (*outcome period*) propondo intervalos de tempo para os mesmos, num contexto de *scoring* comportamental.

Nesta dissertação foi utilizado o modelo semiparamétrico de Cox proporcional, defendido por Cristina Rocha (1995) devido à componente não-paramétrica e à versatilidade daí decorrente, sendo bastante utilizado para a análise de dados censurados de sobrevivência.

O pressuposto para a utilização deste modelo é o da proporcionalidade no tempo das funções de taxa de falha entre grupos. Para além dos métodos tradicionais de análise deste

[1] HPE- Hewlett Packard Enterprise

pressuposto apresentados por Colosimo e Giolo (2006), foi utilizado o método proposto por Hosmer, D. e Royston, P. (2002) em *Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model*, que consiste na utilização do gráfico da regressão do modelo de Aalen para aferir da proporcionalidade.

Depois de definido o modelo de Cox é pertinente avaliar o seu desempenho. Nesse âmbito, a curva ROC é um método popular para avaliar a performance de modelos binários. No entanto, para analisar o desempenho de um modelo de Cox o mais apropriado é uma análise da ROC que varie em função do tempo, uma opção é a usada por Heagerty e Zheng (2005). A mesma especifica que para cada instante *t*, dividem-se os indivíduos em dois grupos, o primeiro é composto pelos que apresentam o evento de interesse naquele determinado instante de tempo *t* (sensibilidade), e o segundo, composto pelas unidades que apresentam um tempo de falha ou censura superior ao instante de avaliação *t* (especificidade).

Em 2015, Baesens, B. publica o livro *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, no qual dedica um capítulo à importância da AS nos sistemas de informação estatísticos de suporte à decisão.

CAPÍTULO 1

INTRODUÇÃO E REVISÃO DE LITERATURA

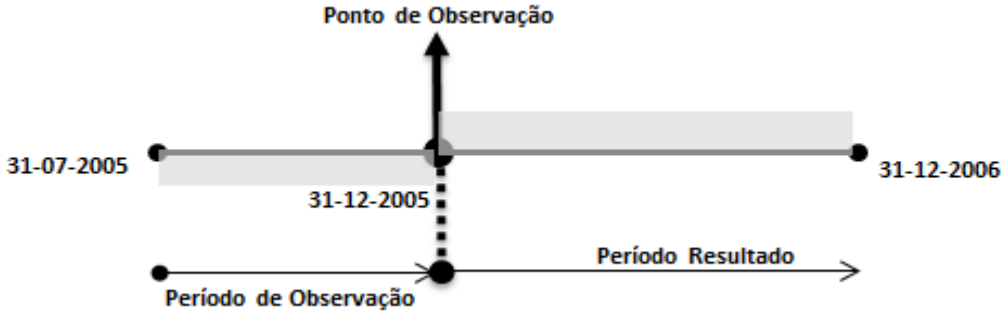
CAPÍTULO 2

RECOLHA E TRATAMENTO DOS DADOS

2 Recolha e Tratamento de Dados

A definição de um horizonte temporal é, provavelmente, mais crítico numa análise comportamental do que em modelos de risco de crédito. Metodologias de análise comportamental têm como objetivo a previsão i.e. avaliar o estado dos clientes no final do período de observação e período resultado. Logo o tempo entre estes dois períodos é crucial para o desenvolvimento de um método robusto de previsão, sugerindo Lyn Thomas (2001) para o efeito, um período de 12 a 18 meses. A partir de uma amostra de 23 meses, 57 variáveis e 149.566 registos, foi definido, para este estudo, como janela de amostragem o período compreendido entre Dez/2005 e Dez/2006, considerado equilibrado de forma a suavizar os efeitos da sazonalidade presente na utilização dos cartões de crédito, eventuais novas funcionalidades introduzidas no sistema de informação da instituição financeira, pedidos de alteração pelas áreas de negócio e marketing, correções de erros e campanhas comerciais. Com base neste período, foi selecionada uma amostra de clientes, observados em diferentes momentos temporais, cuja representação típica é visível na Figura 2.1. A figura ilustra para o ponto de observação de 31/12/2005, um período de observação com duração de 6 meses e um período de resultado de 12 meses.

Figura 2.1: Janela de Amostragem (exemplo para o período 2005/12)



No período anterior ao ponto de observação, designado por período de observação, com a duração 6 meses, podem observar-se as variáveis independentes, ou seja, as variáveis a utilizar na análise comportamental e que caracterizam os clientes durante este período. O período após o ponto de observação tem a duração de 12 meses e é o período onde se avalia se um cliente é classificado, ou não, em incumprimento.

Apesar da regulamentação do Banco de Portugal (BdP), que minimiza inconsistências na informação sobre um cliente, a qualidade dos dados é atualmente um grande desafio das instituições financeiras. A informação introduzida e gerada, por um SI, deverá estar sujeita a metodologias que certifiquem a qualidade da informação, tendo sido, igualmente, um desafio presente neste trabalho.

2.1 Recolha e preparação de dados

Definida a janela de amostragem, inicia-se o processo de preparação dos dados, verificando a existência de inconsistências e omissões. Uma vez que a seleção efetuada foi sujeita a um tratamento informático intensivo, de extração de dados de diferentes aplicações (gestão bancária; crédito; pessoas; etc) e conjunção (consolidação) num só ficheiro, é possível a ocorrência de erros – gerada pelas diferentes linguagens de programação utilizadas e pelos diferentes modelos e repositórios de dados – que deverão ser tratados.

O processo de preparação e "limpeza" de dados tem como objetivo final preparar os mesmos de forma a produzir uma amostra representativa que permita obter um modelo o mais preditivo possível.

Neste ponto sintetizam-se os principais passos que levaram à amostra final. Foi considerado como critério de exclusão clientes:

- que são empresas;
- que são empresários em nome individual;
- sem utilização do cartão de crédito durante o período resultado;
- com estados de cartão diferentes de normal e contencioso.

Adicionalmente, e tratando-se de um modelo cuja unidade amostral é o cliente, torna-se necessário, sempre que existam mais do que um cartão de crédito, definir regras que permita a sua agregação, nomeadamente:

- Agregações de campos financeiros (ex. saldo utilizado para total de saldo utilizado do período);
- Atribuição do estado do cliente (ex. quando no mesmo período em análise forem detetados para o mesmo cliente diferentes estados de crédito – normal, incumprimento, considera-se o cliente como em incumprimento);
- Tratamento de variáveis binárias de agregação e substituição (ex. alertas de cartão de crédito; informação associada a incidentes bancários).

Utilizou-se o mesmo método para a análise de dados corruptos, sendo que, nestes casos e depois de analisados, optou-se por eliminar as observações associadas. A justificação encontrada foi uma deficiente introdução de dados e/ou tratamento informático posterior (ex. processos *batch*[2]) que estiveram na origem da corrupção detetada, mas que não é passível de correção nesta fase.

Posteriormente, foram extraídas as estatísticas descritivas das variáveis contínuas (ver ANEXO I) e contagens de incumprimentos por categoria, das variáveis binárias e categóricas. Com base nesta análise, foi possível identificar registos que apresentavam, para uma percentagem elevada de variáveis, dados omissos ou inválidos, tendo sido os mesmos excluídos da amostra. Da análise efetuada às estatísticas descritivas foram excluídas as covariáveis, por demasiados registos a nulo, cujo enquadramento funcional não o justifica:

- Saldo da operação na moeda de origem;
- Valor do capital vencido;
- Valor dos juros vencidos;
- Especificação do número de dias a descoberto não autorizado no período de operação.

[2] São processos informáticos que são realizados sem a supervisão direta de um utilizador e que estão, geralmente, associados a tarefas recorrentes sobre grandes volumes de informação.

Considera-se normal a existência de uma amplitude elevada entre máximo e mínimo, dada a existência de valores (quantias) nulos e de valores elevados. Na mesma amostra estão clientes, por exemplo, com cartão *gold* e clientes com cartão normal e com limites de crédito significativamente distintos consoante o produto. Esta amplitude justifica igualmente os valores apurados para a mediana e para a média, confirmando a existência de valores extremos. No entanto e uma vez que alguns modelos de sobrevivência são sensíveis à presença de valores atípicos, foi realizada – com recurso ao diagrama de extremos-e-quartis (*boxplot*)[3] e ao *bagplot* - uma análise à existência dos mesmos. Com auxílio do R recorrendo à função boxplot() para a construção destes gráficos e à função bagplot() da biblioteca aplpack, foram calculados os diagramas de extremos e quartis e os *bagplots*, para cada uma das covariáveis, tendo-se obtido os resultados expostos no ANEXO II.

O *bagplot* é uma generalização bivariada do *boxplot*, proposta por Rousseeuw, Ruts e Tukey (1999). No caso bivariado, a caixa-com-bigodes é substituída por um polígono convexo. No polígono estão representadas 50% (*n/2*) das observações com maior profundidade[4] e a linha à volta do mesmo separa as observações que estão dentro e fora deste. As que estão fora são marcadas como valores atípicos e valores atípicos severos. Para minimizar a presença destes, foram efetuadas transformações de algumas covariáveis. Estas transformações não são mais do que expressão dos mesmos dados em unidades de medida diferentes. Neste contexto, foram utilizadas as transformações logaritmo natural e raiz quadrada (ver Tabela 2.1), ou seja, transformação de x em $\sqrt{x}$ e $\ln x$. Por fim e depois de identificados os valores atípicos severos, optou-se pela remoção das observações associadas aos mesmos.

[3] Diagrama de extremos-e-quartis é um gráfico em que, à escala, se representam o menor valor que não seja valores atípico, o quartil inferior, a mediana, o quartil superior, e o maior valor que não seja valores atípico.

[4] A profundidade de localização ou profundidade de Tukey de um ponto $\theta \in \mathbb{R}^2$ relativamente a um conjunto de dados bivariado $X = \{x_1, \dots, x_n\}$, é definida como o menor número de pontos, pertencentes a *X*, situados em qualquer semi-espaço fechado definido por um hiperplano que passe por θ.

Tabela 2.1: Transformações logarítmicas e de raiz quadrada

Covariável	Descrição	Escala Transformada
x7	Limite de Crédito Contratado	$\sqrt{x7}$
x13	Total Mensal de Créditos em Conta	$\ln x13$
x14	Total Mensal de Débitos em Conta	$\ln x14$
x20	Antiguidade do Cliente em Meses	$\ln x20$
x21	Rácio entre Débitos e Créditos	$\ln x21$

Depois de realizado o tratamento dos dados, foram eleitas para o estudo as covariáveis definidas na Tabela 2.2.

Tabela 2.2: Descrição das covariáveis

Covariável	Tipo de variável	Descrição
x1	Categórica	Estado Civil
x2	Binária	Indicador de Existência de Produtos de Poupança
x3	Categórica	Sexo M=1 ; F =0
x4	Categórica	Habilitações Literárias
x5	Binária	Indicador de Trabalhador por Conta de Outrem
x6	Categórica	Código de Natureza Jurídica
x7	Continua	Limite de Crédito Contratado
x8	Binária	Indicador de Catão de Crédito Normal
x9	Binária	Indicador de Cartão Gold
x10	Binária	Indicador de Cartão de Uso Exclusivo em Combustível
x11	Continua	Numero de Cartões por Cliente
x12	Continua	Idade
x13	Continua	Total Mensal de Créditos em Conta
x14	Continua	Total Mensal de Débitos em Conta
x15	Continua	Diferença entre Débitos e Créditos
x16	Binária	Indicador de Inibição de Uso de Cheque
x17	Binária	Indicador de Cliente com Crédito Particular (consumo)
x18	Binária	Indicador de Cliente com Crédito Hipotecário
x19	Binária	Indicador de Cliente com Ordenado Domiciliado
x20	Continua	Antiguidade do Cliente em Meses
x21	Continua	Rácio entre Débitos e Créditos

2.2 Análise estatística das variáveis

Concluída a construção da amostra, foram analisadas as estatísticas descritivas das variáveis categóricas e a sua capacidade preditiva com recurso às estatísticas: *Weight of Evidence* (WoE), *Information Value* (IV) e Z-*score*, tal como sugerido em Steven Finlay (2010). Os resultados estão disponíveis no ANEXO IV. Os mesmos, não tendo em consideração o tempo de sobrevivência, servem como um indicador explicativo da covariável, no contexto das taxas de incumprimento associadas a cada uma delas. Foram estudadas individualmente, sem o respetivo contexto temporal de uma AS.

As duas primeiras estatísticas, WoE e IV, são úteis na identificação da contribuição de cada atributo para a discriminação entre clientes cumpridores ou incumpridores. As restantes são medidas de associação que permitem avaliar o contributo da variável para o fenómeno em estudo.

2.2.1 *Weight of Evidence* (WoE)

A estatística *Weight of Evidence* mede o risco relativo de um dado atributo na probabilidade de incumprimento. De acordo com Steven Finlay, o cálculo do WoE é dado por:

$$WoE = \ln\left(\frac{g_a/G}{b_a/B}\right) \tag{2.1}$$

onde:

g_a = Número de clientes regulares (BONS)[5] com um determinado atributo

b_a = Número de clientes com incumprimento (MAUS) com um determinado atributo

G = Número total de clientes BONS

[5] Considera-se MAU um cliente que atinja um incumprimento superior ou igual a 90 dias nos 12 meses seguintes à data de referência (*default*) e BOM, caso contrário.

B = Número total de clientes MAUS

O WoE relaciona as chances, regular/incumprimento, de um dado atributo com a média da amostra. Um atributo com uma proporção de regulares superior à média apresentará um WoE positivo e, pelo contrário, um atributo com maior proporção de incumprimento que a média, terá um WoE negativo.

Enquanto o WoE permite ter uma ideia da magnitude da diferença entre a taxa de incumprimento de um determinado atributo e a taxa de incumprimento média, o *z-score* permite avaliar se essa diferença é ou não estatisticamente significativa.

Genericamente, a relação entre o atributo e a covariável deverá ser tão elevada e estatisticamente significativa que justifique a sua entrada no modelo, com consequente aumento da capacidade preditiva do mesmo. As duas estatísticas deverão ser analisadas em conjunto.

2.2.2 Z-score

A estatística *z-score* providencia uma medida de quão semelhantes são as taxas de incumprimento de duas populações, definida pela verificação ou não de um dado atributo. A estatística *z-score* para cada atributo ou intervalo é dada por:

$$z-score = \frac{BR_{\neq a} - BR_a}{\sqrt{\frac{BR_T\left(1-BR_T\right)}{n_{\neq a}} + \frac{BR_T\left(1-BR_T\right)}{n_a}}} \tag{2.2}$$

onde:

$BR_{\neq a}$ = a taxa de MAUS das observações que não verificam o atributo a, ou seja, o complementar de a

BR_a = a taxa de MAUS com o atributo a

BR_T = taxa de MAUS registada no total da amostra,

n_a corresponde ao número de observações com o atributo a

$n_{\neq a}$ corresponde ao número de observações sem o atributo a.

A estatística *z-score* compara a taxa de incumprimento de cada atributo com o resto da amostra. Assim, valores absolutos da estatística entre 1,96 e 3,58 indicam um grau de confiança moderado e valores absolutos acima de 3,58 indicam um forte grau de confiança (> 99,9%) sobre a dissemelhança das duas populações.

2.2.3 *Information Value* (IV)

A estatística *Information Value* é, segundo Steven Finlay (2010), uma das medidas de associação mais populares usadas em problemas de classificação. O IV para uma determinada variável independente, categórica com n atributos, é dado por:

$$IV = \sum_{a=1}^{n} \left(p_{ag} - p_{ab} \right) \ln \left(\frac{p_{ag}}{p_{ab}} \right) \tag{2.3}$$

onde:

n = Número de atributos da variável independente

$p_{ag} = g_a / G$ = Proporção de clientes BONS com o atributo a

$p_{ab} = b_a / B$ = Proporção de clientes MAUS com o atributo a

g_a = Número de clientes BONS com um determinado atributo

b_a = Número de clientes MAUS com um determinado atributo

G = Número total de clientes BONS

B = Número total de clientes MAUS

Genericamente, valores da estatística abaixo de 0,05 indicam uma fraca associação entre a variável categórica em análise e a variável dependente binária; valores entre 0,05 e 0,25 indicam uma relação moderada e valores acima de 0,25 indicam uma relação forte entre as duas variáveis.

2.2.4 Coeficiente de Correlação de Spearman

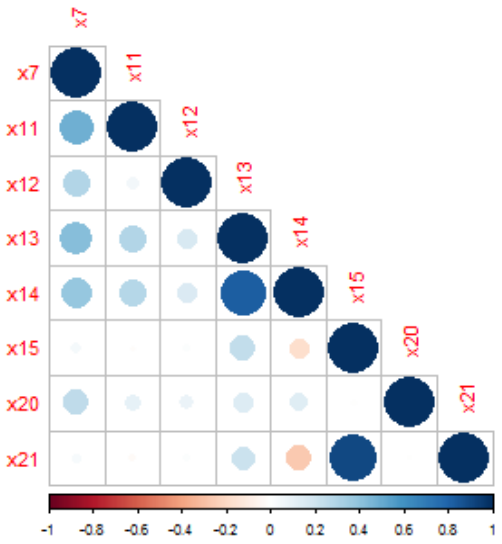
Foi calculado o coeficiente de correlação não paramétrica de Spearman uma vez que não estão garantidos os pressupostos exigidos para o cálculo do coeficiente de correlação de Pearson. O coeficiente de correlação varia entre -1 e 1, sendo que quanto mais próximo estiver de 1 mais perfeita é a relação positiva entre as duas covariáveis e quanto mais se aproximar de -1 melhor a relação negativa.

Tabela 2.3: Correlação de Spearman

	x7	x11	x12	x13	x14	x15	x20	x21
x7	1	0,49	0,30	0,43	0,38	0,04	0,26	0,04
x11	0,49	1	0,06	0,29	0,28	-0,01	0,10	-0,02
x12	0,30	0,06	1	0,17	0,16	0,02	0,09	0,02
x13	0,43	0,29	0,17	1	**0,83**	0,25	0,15	0,21
x14	0,38	0,28	0,16	**0,83**	1	-0,17	0,14	-0,27
x15	0,04	-0,01	0,02	0,25	-0,17	1	-0,01	**0,91**
x20	0,26	0,10	0,09	0,15	0,14	-0,01	1	-0,01
x21	0,04	-0,02	0,02	0,21	-0,27	**0,91**	-0,01	1

Da análise da Tabela 2.3 e do Gráfico 2.1, constata-se uma correlação positiva forte de 83% entre as covariáveis x14 – Total Mensal de Débitos em Conta e x13 – Total Mensal de Créditos em Conta. E uma correlação, igualmente, positiva forte de 91% entre as covariáveis x15 – Diferença entre Débitos e Créditos e x21 – Rácio entre Débitos e Créditos.

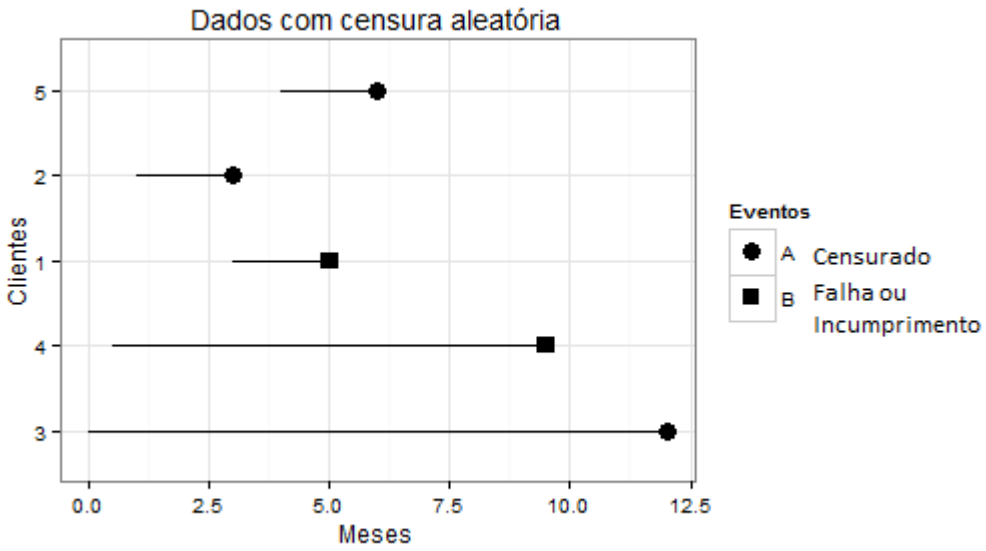
Gráfico 2.1: Correlação de Spearman



2.2.5 Caracterização dos dados de sobrevivência

Concluída a seleção da amostra, foi necessário dotar a mesma de uma representação de dados de sobrevivência. Esta representação pode ser descrita como uma linha horizontal para cada cliente, sendo o seu comprimento caracterizado pelo instante inicial que neste estudo é 31/12/2005 e o final que é 31/12/2006. Desenham-se as linhas da esquerda para a direita e é possível distinguir os clientes que alcançam o limite dos que são censurados. A AS dedica-se ao estudo da variável tempo de sobrevivência, isto é, uma variável que representa o tempo até que determinado acontecimento de interesse suceda (designado como evento de falha). Nos estudos de AS é normal existirem observações censuradas, ou seja, o acontecimento de falha não ocorreu no tempo de observação – a AS tem ferramentas próprias para lidar com os dados censurados. Neste estudo e conforme ilustrado na Figura 2.2, estamos perante censura aleatória, ou seja, a experiência foi planeada para durar 12 meses e teve o seu início com um conjunto de clientes, na evolução mensal da mesma, foram entrando novos clientes e ocorrendo falhas com os existentes.

Figura 2.2: Ilustração de dados com censura (A) e falha (B)



CAPÍTULO 2

RECOLHA E TRATAMENTO DOS DADOS

CAPÍTULO 3

ANÁLISE DE SOBREVIVÊNCIA

3 Análise de Sobrevivência

Neste capítulo descreve-se sumariamente as metodologias estatísticas de análise de sobrevivência e avaliação da performance de um modelo, utilizadas nesta dissertação. No ponto 3.1 define-se função de sobrevivência e taxa de falha. Em 3.2 o estimador não-paramétrico de Kaplan-Meier, no 3.3 é apresentada a especificação teórica do modelo de proporcional de Cox e o 3.4 é dedicado ao modelo de Aalen. Em 3.5 é realizada uma introdução teórica à curva ROC e ao índice de Gini.

3.1 As funções de Sobrevivência e Taxa de Falha ou Hazard

Seja *T* uma variável aleatória não negativa que representa o tempo até acontecimento de interesse ou falha que pode ser especificada pela função de sobrevivência e pela função de taxa de falha.

A função de sobrevivência no instante *t* define a probabilidade de uma observação sobreviver ao tempo *t*. Analogamente, a função de distribuição acumulada da variável T é definida como a probabilidade de uma observação não sobreviver ao tempo *t.*

Temos assim a função de Sobrevivência $S(t)$ e a respetiva função de distribuição acumulada, de T.

$$S(t) = P(T \geq t) \tag{3.1}$$

$$F(t) = 1 - S(t) \tag{3.2}$$

A *S(t)* tem as seguintes propriedades:

- É monótona não crescente.
- É contínua à esquerda.

- $S(0) = 1$, ou seja, todos os elementos estão vivos no instante inicial. Aplicado a esta dissertação significa que todos os clientes estão sem histórico de incumprimento no instante inicial.

A função taxa de falha representa a taxa instantânea de falha no instante *t*, condicional à sobrevivência da observação até esse instante. Esta função descreve a forma em que a taxa instantânea de falha muda com o tempo. A função Taxa de Falha Acumulada fornece a taxa de falha acumulada da observação.
A função Taxa de Falha e função Taxa de Falha Acumulada são dadas então por

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3.3}$$

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{3.4}$$

No contexto dos estudos de sobrevivência as formas mais comuns que a função taxa de falha $\lambda(t)$ apresenta, são as seguintes:

- Monótona Crescente – especifica um período durante o qual, a proporção de eventos de falha aumenta com o tempo.
- Monótona Decrescente – indica que quanto maior o tempo de sobrevivência, menor a probabilidade de morte no instante subsequente.
- Constante – a taxa de falha não se altera com o passar do tempo.
- Em forma de U (*bathtub shaped*) – ocorre em populações em que os indivíduos são observados desde o nascimento até à morte.

3.2 Modelo não-paramétrico de Kaplan-Meier

Para acomodar os dados censurados contidos na amostra, são necessárias técnicas estatísticas próprias, nomeadamente o estimador de Kaplan-Meier. O estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, que é também chamado de estimador limite-produto, é a técnica mais utilizada em estudos clínicos e em crescendo noutras áreas, como por exemplo a de análise de risco.

As estimativas das probabilidades condicionadas[6] em cada momento de tempo são o resultado de multiplicações sucessivas de probabilidades de sobrevivência à medida que os intervalos de tempo diminuem até zero.

$$\hat{S}(t_j +) = \hat{P}(T \geq 0)\hat{P}(T > 1|T \geq 0)$$

Quando não existir qualquer evento numa determinada unidade de tempo, a probabilidade condicionada de sobrevivência, nesse instante, é igual a 1.

Para determinar a estimativa de Kaplan-Meier para a função de sobrevivência, começamos por ordenar os indivíduos por ordem crescente do tempo de observação e definir os instantes de falha t_j, sendo n_j o número de indivíduos ainda em risco no instante t_j. Temos D_j com o número de falhas por cada instante t e as estimativas de Kaplan-Meier $\hat{S}(t_j +)$ para a função de sobrevivência no intervalo entre t_j e o instante de falha seguinte t_{j+1}, considerando que no instante inicial temos S(0)=1.

A expressão geral do estimador de Kaplan-Meier pode ser assim apresentada:

$$\hat{S}(\text{t}) = \prod_{j:\, t_j<t}\left(\frac{n_j-d_j}{n_j}\right) = \prod_{j:\, t_j<t}\left(1-\frac{d_j}{n_j}\right) \qquad (3.5)$$

[6] É a probabilidade de um evento de falha ocorrer num intervalo de tempo, sabendo que não ocorreu até esse instante.

No seu conjunto, estas estimativas chamam-se função de sobrevivência, que são habitualmente apresentadas num gráfico chamado de curva de Kaplan-Meier.

A $\hat{S}$(t) é uma função escada com degraus nos tempos observados de falha de tamanho $^1/_n$, em que n é o tamanho da amostra. A sobrevivência no tempo *t*, traduz a proporção de indivíduos censurados, no final desse tempo.

Se não existirem censuras o estimador de Kaplan-Meier reduz-se à função de sobrevivência:

$$\hat{S}(t) = \frac{\text{Número de observações que não falharam até } t}{\text{Número total de observações no estudo}} \tag{3.6}$$

As principais propriedades do estimador de Kaplan-Meier são:

[1] Converge assintoticamente para um processo Gaussiano;

[2] É estimador de máxima verossimilhança de $S(t)$.

A sua variância assintótica é dada por:

$$\widehat{Var}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)} \tag{3.7}$$

Podemos definir o intervalo de confiança com (100 – α)% de confiança, baseados na formula de Greenwood, como:

$$IC_{100-\propto} = [\hat{S}(t) \pm z_{\propto/2}\sqrt{\widehat{Var}(\hat{S}(t))}] \tag{3.8}$$

De notar que quando os valores obtidos no intervalo forem negativos ou superiores a 1, estes devem ser restringidos ao intervalo [0,1].

3.2.1 O teste *Logrank*

Quando dizemos que dois ou mais grupos de indivíduos ou duas ou mais curvas de Kaplan-Meier são ou não diferentes, não o poderemos fazer apenas pela observação gráfica ou dos resultados do respetivo modelo. Para tal, utiliza-se o teste *logrank* que nos permite testar a hipótese nula de um evento ocorrer com igual frequência em populações diferentes. Por outras palavras, testa a hipótese de igualdade entre várias curvas de Kaplan-Meier,

$$H_0: S_1(t) = S_2(t) \quad vs \quad H_1: S_1(t) \neq S_2(t) \tag{3.9}$$

onde $S_i(t), i = 1, 2$ são as respetivas funções de sobrevivência dos dois grupos de indivíduos de dimensão *m* e *n*. O teste *logrank* compara o número de eventos de falha observados O_i e esperados E_i, para o grupo *i* calculando-se a estatística,

$$T = \frac{\left[\sum_{j=1}^{J}(O_{1j} - E_{1j})\right]^2}{\sum_{j=1}^{J} V_j} \sim \chi_1^2 \tag{3.10}$$

Que sob H_0 segue uma distribuição χ^2 com 1 grau de liberdade, sendo *j = 1,…,J* os eventos relativos aos *m + n* indivíduos e $V_j = \frac{O_j(N_{1j}/N_j)(1-N_{1j}/N_j)(N_j-O_j)}{N_j-1}$.

3.3 A Especificação do Modelo de Cox de Riscos Proporcionais

O modelo de regressão de Cox (Cox, 1972) deu início a uma nova fase na modelação estatística de dados clínicos. Stigler (1994) quantifica, no período de 1987 – 1989, o artigo de Cox como um dos mais referenciados na literatura estatística, sendo ultrapassado apenas por Kaplan-Meier (1958). Uma das principais motivações para esta popularidade é a versatilidade deste modelo, sendo a mesma, devido à presença de uma componente não-paramétrica. O modelo de Cox é considerado um modelo semiparamétrico (Rocha, 1995).

Seja t uma variável aleatória continua que representa o tempo de vida de um individuo, $x = (x_1, \dots, x_p)'$ o vetor de covariáveis (associado ao cliente ou individuo em estudo) e assumindo a proporcionalidade entre as funções de taxa de falha, define-se o modelo de regressão de Cox de riscos proporcionais, por:

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 x_1 + \cdots + \beta_p x_p) \tag{3.11}$$

em que $\beta_1, \dots, \beta_p$ são os coeficientes de regressão que representam o efeito das covariáveis na sobrevivência e $\lambda_0(t)$ [7] é a componente não paramétrica, função não negativa do tempo, que representa a função de taxa de falha para um individuo a que está associado o vetor $x=0$. As funções relacionadas a $\lambda_0(t)$ são, igualmente, importantes no modelo de Cox. A função de taxa de falha acumulada de base é dada por:

$$\Lambda_0(t) = \int_0^1 \lambda_0(u)\, d_u \tag{3.12}$$

A correspondente função de sobrevivência de base:

$$S_0(t) = \exp\{-\Lambda_0(t)\} \tag{3.13}$$

Uma estimativa para $\Lambda_0(t)$ é dada por:

$$\widehat{\Lambda_0}(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\in R_j} exp\{x_l'\widehat{\beta}\}} \tag{3.14}$$

em que d_j é o número de falhas em t_j.

[7] Função de taxa de falha basal

3.3.1 Interpretação dos Coeficientes

Dado tratar-se de um modelo de riscos proporcionais, as funções taxa de falha correspondente a dois indivíduos com covariáveis x_1 e x_2, por exemplo, são proporcionais. De facto, sendo $\beta = (\beta_1, \dots, \beta_2)'$,

$$\frac{\lambda(t;x_1)}{\lambda(t;x_2)} = e^{(\beta'(x_1 - x_2))} \tag{3.15}$$

não depende de *t.*

As covariáveis têm um efeito multiplicativo na função taxa de falha de acordo com o fator $e^{(\beta' x)}$ que é designado por risco relativo.

3.3.2 Avaliação da Proporcionalidade

O modelo de Cox tem como pressuposto que os riscos são proporcionais, entre grupos, em todo o período temporal em estudo. Uma vez identificado o conjunto de covariáveis para modelação, é necessário verificar se cada uma cumpre o pressuposto de riscos proporcionais. Existem vários métodos para avaliar esta suposição, sendo nesta dissertação utilizadas as infra descritas.

3.3.2.1 Método gráfico descritivo

Este método é utilizado apenas em variáveis categóricas e consiste na construção de um gráfico do logaritmo da função taxa de falha acumulada em cada grupo no tempo, também conhecido como gráfico log-log, que representa o logaritmo de $\widehat{\Lambda_0}(t)$ contra o logaritmo do tempo de sobrevivência. Para ser considerado proporcional, o mesmo, deverá apresentar linhas paralelas. A existência de linhas não paralelas significa que o efeito em estudo não é constante ao longo do tempo.

Se a variável for contínua, será necessário categorizá-la para a aplicação deste método.

3.3.2.2 Método com coeficiente dependente do tempo

Este método consiste na análise gráfica dos resíduos padronizados de Schoenfeld e dos coeficientes de correlação de Pearson entre os referidos resíduos e uma função do tempo, *g(t)*, para aferir da referida suposição de taxas de falha proporcionais no modelo de Cox.

Os resíduos de Schoenfeld no modelo de Cox definem-se considerando o vetor de covariáveis $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ e um vetor de resíduos de Schoenfeld $r_i = (r_{i1}, r_{i2}, \dots, r_{ip})$ em que cada componente r_{iq}, para *q = 1, …, p*, é definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} e^{(x_j' \widehat{\beta})}}{\sum_{j \in R(t_i)} e^{(x_j' \widehat{\beta})}} \tag{3.16}$$

Os resíduos são definidos para cada falha e não para censuras.

Nesta dissertação, para avaliar a adequação do modelo em termos da hipótese de proporcionalidade da taxa de falha entre grupos tendo em vista a aplicação do modelo de Cox, considerando individualmente cada uma das covariáveis em estudo, foram utilizados gráficos de uma forma padronizada dos resíduos de Schoenfeld *(scaled Schoenfeld residuals)*, bem como, os respetivos testes de correlação associados a esta análise.

Para se proceder ao estudo individual de cada uma das covariáveis, são analisados os gráficos (vd. ANEXO VI) em que se representam os resíduos, com a curva suavizada com bandas de confiança e a abcissa para facilitar a leitura, por exemplo, das mudanças de sinal. Foi analisado, visualmente, se a variação no efeito da covariável se mantém constante no tempo. Posteriormente, ao resultado da análise gráfica foram associados os resultados do teste de proporcionalidade, ou seja, o coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e uma função do tempo *g(t),* para cada covariável.

Admite-se a proporcionalidade dos riscos quando não rejeitamos a hipótese nula de que a correlação é igual a zero[8].

[8] Consideram-se as hipóteses: $H_0: \rho = 0 \quad vs \quad H_1: \rho \neq 0$.

A interpretação dos gráficos dos resíduos padronizados de Schoenfeld está sujeita a conclusões subjetivas e por vezes distinta dos coeficientes de correlação de Pearson entre os resíduos. Quando perante este cenário, propõe-se que seja analisada a função de regressão acumulada do modelo aditivo de Aalen, para cada uma das covariáveis nas condições supra descritas, para confirmação da decisão de evidência da hipótese de proporcionalidade da taxa de falha. Todavia, recomenda-se utilização parcimoniosa deste método, ou seja, apenas, como um suporte à decisão quando os resultados não são conclusivos ou quando se pretende utilizar, por exemplo, o modelo de Cox com covariáveis dependentes do tempo, conforme proposto por Hosmer e Royston (2002). Neste contexto, estes autores dão particular relevo ao trabalho de Henderson e Milner (1991) que demonstra que, mesmo sob riscos proporcionais, os gráficos de Aalen apresentam uma ligeira curva, não linear, nas covariáveis.

3.3.3 Critério Akaike (AIC)

Akaike (1973, 1974b), apresentou um critério ao qual se denominou *Akaike's Information Criterion (AIC)*. A minimização do mesmo permite avaliar o ajustamento de um modelo estatístico e é definido como,

$$AIC(K) = -2\ln[\mathcal{L}] + 2K \tag{3.17}$$

onde *K* é o número de parâmetros (graus de liberdade) no modelo e $\mathcal{L}$ um estimador de máxima verossimilhança.

O AIC é muito utilizado para comparar modelos. Numa metodologia de modelação, como por exemplo, a *general-to-specific* em que as variáveis não significativas para um determinado nível de significância, vão sendo sucessivamente excluídas e calculado o AIC para cada novo modelo daí resultante, o menor AIC é indicador do melhor ajustamento, no entanto, recomenda-se a utilização deste critério com outros testes, como por exemplo, o Teste do Rácio das Verosimilhanças (TRV).

3.3.4 Teste do Rácio das Verossimilhanças (TRV)

Seja θ um parâmetro genérico de uma certa população de interesse, $\mathcal{L}(\hat{\theta})$ o logaritmo da função de máxima verossimilhança do modelo irrestrito[9] e $\mathcal{L}(\theta_0)$ o logaritmo da função de máxima verossimilhança do modelo restrito[10], a estatística do teste é dada por:

$$TRV = -2log\left[\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\widehat{\theta})}\right] = 2\left[log\mathcal{L}(\widehat{\theta}) - log\mathcal{L}(\theta_0)\right] \quad (3.18)$$

que, sob H_0: $\theta = \theta_0$, segue aproximadamente uma distribuição χ^2 com *p* graus de liberdade. Para amostras grandes, H_0, é rejeitada, a um nível α de significância, se TRV > $\chi^2_{p,1-\alpha}$.
Note-se que para calcular o *TRV* é necessário estimar quer o modelo irrestrito, quer o modelo restrito.

3.3.5 Resíduos, Forma Funcional e Valores Atípicos

Para além da suposição de proporcionalidade, é importante avaliar outros aspetos do ajuste do modelo de Cox, nomeadamente, a forma funcional, verificar a presença de valores atípicos e avaliar a influência que cada individuo exerce nas diferentes covariáveis que compõem o modelo ajustado. Para examinar estes aspetos são, geralmente, utilizados os resíduos Cox-Snell, martingala, da desviância e os DFBETAS.
Podemos definir o resíduo como um valor calculado para cada observação, informando sobre a diferença entre a sobrevivência observada para a observação e o valor estimado pela equação de regressão, quanto maior a diferença maior o valor absoluto do resíduo.
Os resíduos Cox-Snell, para o modelo de Cox, são definidos por:

$$\widehat{e_i} = \widehat{\Lambda_0}(t)exp\left\{\sum_{k=1}^{p} x_{ik}\,\hat{\beta}_k\right\},\ i = 1, \dots, n \quad (3.19)$$

[9] Modelo completo sem restrições.
[10] Modelo após restrição de covariáveis.

Se o ajustamento não for satisfatório a análise gráfica destes resíduos, $\widehat{\Lambda}(\widehat{e_i})$ versus $\widehat{e_i}$, não é informativa, neste sentido, os mesmos não são considerados neste estudo.
Os resíduos martingala são obtidos a partir dos resíduos Cox-Snell e os da desviância são obtidos a partir dos martingala.

3.3.5.1 Resíduos Martingala

Os resíduos martingala são calculados tendo por base os resíduos Cox-Snell e poderão ser assim definidos:

$$\widehat{m_i} = \delta_i - \widehat{e_i}\,, \quad i = 1, \dots, n \tag{3.20}$$

sendo $\widehat{e_i}$ os resíduos de Cox-Snell e $\delta_i = 0$ se a observação for censurada e $\delta_i = 1$ se se verificar o evento de falha. Os resíduos martingala são utilizados para avaliar a forma funcional do modelo obtido ou de cada uma das covariáveis contínuas individualmente. Para amostras grandes o valor esperado da soma dos resíduos é zero, no entanto e apesar do modelo estar correto, os mesmos não estão distribuídos simetricamente em torno de zero o que dificulta a análise gráfica. Todavia, nesta dissertação, foi acrescentada uma linha representando a regressão linear local LOWESS (*Locally Weighted Scatterplot Smoothing*) ao gráfico dos resíduos martingala, sendo que se a mesma estiver em torno de zero sem demonstrar qualquer tendência, significa que a forma funcional do modelo é valida.

3.3.5.2 Resíduos da Desviância

Como referido anteriormente, o facto dos resíduos martingala não estarem distribuídos em torno de zero dificulta a sua análise. A partir destes, foram criados os resíduos da desviância que resultam de uma tentativa de tornar os resíduos martingala simétricos em torno de zero,

facilitando assim a deteção de valores atípicos, os resíduos da desviância são assim definidos no modelo de Cox:

$$\widehat{d_i} = sinal(\widehat{m_i})[-2(\widehat{m_i} + \delta_i \log(\delta_i - \widehat{m_i}))]^{1/2} \quad (3.21)$$

O gráfico de d_i *versus* o preditor linear $\sum_{l=1}^{p} x_{il} \times \beta_l$, com $i = 1, \dots, n$ e $\beta = (\beta_1, \dots \beta_p)$ o vetor de parâmetros, é utilizado para avaliar a presença de valores atípicos.

Os valores de $\widehat{d_i}$ variam entre $-\infty$ e $+\infty$. Quando positivos os resíduos da desviância significam que as respetivas observações tiveram um tempo de sobrevivência inferior ao esperado. Inversamente, resíduos com sinal negativo indicam observações com um tempo de sobrevivência superior ao esperado. Valores absolutos superiores a um determinado valor (neste estudo, 3) deverão ser analisados individualmente dado que são potenciais valores atípicos.

3.3.5.3 Estatística DFBETA

A estatística DFBETA indica a influência que cada individuo exerce nas diferentes covariáveis do modelo ajustado. Valores de DFBETAS negativos, significam que o coeficiente da variável aumenta quando a observação é removida do estudo (Paul Allison, 1995).

Analisando a influência ou impacto de cada observação, através dos gráficos dos resíduos DFBETAS [11], considera-se que valores absolutos superiores a 1 evidenciam pontos influentes no ajustamento e logo possivelmente valores atípicos (Schutte e Violette, 1994).

[11] https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect040.htm, Belsley, Kuh, e Welsch (1980) recomendam o 2 para valor genérico de *cutoff* e $2/\sqrt{n}$ como um *cutoff* ajustado.

3.4 A Especificação do Modelo de Aalen

Aalen (1980) propôs um modelo linear não-paramétrico, denominado modelo de Aalen ou modelo aditivo de Aalen para análise de regressão de dados censurados, com uma característica muito importante, de que é permitido aos seus coeficientes de regressão variarem ao longo do tempo. O modelo de Aalen permite assim que os parâmetros das covariáveis variem com o tempo, sendo o modelo capaz de fornecer informações detalhadas relativamente à influência temporal de cada covariável. Desta forma, e contrariamente ao modelo de Cox, o modelo de Aalen é completamente não-paramétrico, de facto, são muito comuns os estudos em que se observam os indivíduos ao longo do tempo para se analisar a ocorrência de um determinado evento.

Tendo em consideração o conjunto de valores das covariáveis no instante zero, $x_i(t) = (1, x_{i1}(t), x_{i2}(t), \dots, x_{ip}(t))'$, dependentes do tempo, *n* o número de indivíduos, *p* o número de covariáveis e $\lambda(t|X_i(t))$ a função de taxa de falha para o tempo de sobrevivência t do individuo *i*, o modelo de Aalen, que assume que $\lambda(t|X_i(t))$ é uma combinação linear dos $x_{ij}(t)$, é dado por:

$$\lambda\big(t\big|X_i(t)\big) = \beta_0(t) + \sum_{J=1}^{p} \beta_j(t) x_{ij}(t) \tag{3.22}$$

O primeiro elemento $\beta_0(t)$ pode ser interpretado como uma função taxa de falha de base, enquanto que $\beta_j(t)$, *j=1,…,p*, denominadas funções de regressão ou coeficientes de risco, medem a influência das respetivas covariáveis e são permitidas variar com o tempo.

Considerando a sua forma matricial, temos:

$$\lambda\big(t\big|x(t)\big) = X(t)\mathrm{B}(t), \tag{3.23}$$

com $\mathrm{B}(t) = (\beta_0(t), \beta_1(t), \dots, \beta_p(t))'$ um vetor de funções do tempo, desconhecido.

O estimador de risco acumulado para um individuo com vetor $x = (1, x_1, x_2, \dots, x_p)'$ é:

$$\widehat{\Lambda}(t|x) = x'\widehat{\mathrm{B}}(t) = \widehat{\mathrm{B}}_0(t) + \sum_{j=1}^{p} \widehat{\mathrm{B}}_j(t)x_j, \;\; t \leq \tau \tag{3.24}$$

com $\widehat{\mathrm{B}}_j(t)$ estimadores de mínimos quadrados e τ o valor maximal de t para o qual a matriz $X(t)$ é não singular.

A função de sobrevivência é estimada por:

$$\tilde{S}(t|x) = \exp\{-\widetilde{\Lambda}(t|x)\}. \tag{3.25}$$

Foram apresentadas algumas aplicações deste modelo para explorar a linearidade analogamente à teoria clássica dos modelos lineares (McKeague e Utikal, 1991). Outra das vantagens deste modelo é que para além de ser não-paramétrico, não exige tamanhos de amostra excessivamente grandes (Colosimo e Giolo, 2006).

3.5 Avaliação da Performance do Modelo

A avaliação da performance do modelo é realizada recorrendo aos resultados da AUC, ou seja, área abaixo da curva ROC (*Receiver Operating Characteristics*) e do Índice de Gini.

3.5.1 Área ROC (AUC)

Curvas ROC são um método bastante popular para avaliar a performance, por exemplo, de um modelo de resposta binária, todavia perante a existência de dados censurados na amostra, muitos dos métodos estatísticos tradicionais para avaliar a performance de um modelo preditivo não poderão ser utilizados em AS. No nosso caso o evento em estudo contém dados censurados e é dependente do tempo, neste sentido será apropriado apresentar curvas ROC como função do tempo de sobrevivência.

Heagerty and Zheng (2005) apresentam um método de variação temporal da sensibilidade e especificidade baseado na definição de incidente/dinâmico:

$$sensibilidade^{\mathbb{I}}(c,t): Pr\{M_i > c \mid \delta_i(t) = 1\} = Pr\{M_i > c | T_i = t\} \quad (3.26)$$

$$especificidade^{\mathbb{D}}(c,t): Pr\{M_i \leq c \mid \delta_i(t) = 0\} = Pr\{M_i \leq c | T_i > t\} \quad (3.27)$$

M = marcador continuo

c = truncatura

t = instante de tempo

$\delta(t) = \mathbb{I}(T \leq C)$ o indicador de evento de falha ou censura, considerando *C* o tempo de censura e *T* o tempo de sobrevivência.

Nesta abordagem, para cada instante *t*, divide-se as unidades em risco em dois grupos. A sensibilidade mede o grupo de indivíduos com marca superior à truncatura, *c*, que entraram em incumprimento no instante *t*, o segundo grupo de indivíduos, a especificidade, é composta pelos indivíduos que apresentam um tempo de falha ou censura superior ao instante de tempo, t. Com base na abordagem incidente/dinâmico e nas definições de sensibilidade e especificidade, é possível estimar a curva ROC. Trata-se da representação gráfica dos pares ordenados "sensibilidade (*c, t*)" *versus* "1 – especificidade (*c, t*)", para todos os valores de *c*. O modelo será tanto melhor quanto mais côncava for a curva.

Uma medida numérica comummente utilizada pode ser obtida pela área abaixo da curva ROC (AUC), em que o valor 1 representa o modelo perfeito, enquanto uma área próxima de 0,50 representa o modelo aleatório, indiciando uma fraca aderência aos dados. A AUC mede a capacidade preditiva do modelo classificar corretamente os indivíduos em regulares e em incumprimento. Referindo-se a Hosmer e Lemeshow (2000, p. 162), Chorão (2005) apresenta os seguintes valores de referência para a AUC.

Tabela 3.1: Valores de referência para a AUC

Valores	Comentário
AUROC = 0,50	Nenhuma discriminação
0,70 ≤AUROC≤0,80	Modelo Aceitável
0,80 ≤AUROC≤0,90	Boa discriminação
AUROC > 0,90	Excelente modelo

3.5.2 Índice de Gini

O Índice de Gini pode ser obtido através da sua relação linear com a área abaixo da curva ROC (AUC), ou seja, Gini =2 ROC -1, existe uma equivalência direta entre a área abaixo da ROC e o Índice de Gini.

Em risco de crédito, por exemplo, existem valores de referência por tipo de modelo sendo habitual tomar como referência num modelo comportamental o valor de 65% para o Índice de Gini. Em *Plug & Score*[12] sugerem-se os seguintes valores de referência por tipologia de modelo de *credit scoring* (Tabela 3.2), sendo os mesmos utilizados nesta dissertação.

Tabela 3.2: Valores de referência para o Índice de Gini

Scoring Aplicacional		Scoring Comportamental	
Valores	Qualidade do modelo	Valores	Qualidade do modelo
Gini < 0,25	Baixa	Gini < 0,45	Baixa
0,25 ≤Gini≤0,45	Média	0,45 ≤Gini≤0,65	Média
0,45 ≤Gini≤0,60	Boa	0,65 ≤Gini≤0,80	Boa
Gini > 0,60	Muito boa	Gini > 0,80	Muito boa

[12] http://www.plug-n-score.com/learning/gini-and-roc-curve.htm

CAPÍTULO 4

APRESENTAÇÃO DE RESULTADOS

4 Apresentação de Resultados

Neste capítulo apresentam-se os principais resultados dos modelos considerados no estudo empírico, recorrendo à AS com enfoque à modelação de Cox de Riscos Proporcionais e ao R versão 3.0.1 como ferramenta de computação estatística de suporte aos cálculos realizados a partir da versão 0.98.1091 do R Studio.

O ponto 4.1 é dedicado à aplicação do método não-paramétrico de Kaplan-Meier, o 4.2 à avaliação da adequação do modelo de Cox e o 4.3 apresenta os principais resultados da avaliação da performance dos modelos, recorrendo aos resultados da área abaixo da curva ROC e do Índice de Gini.

A partir de uma amostra de 4.358 clientes de uma instituição financeira, observados no período de Dez/2005 a Dez/2006 e titulares de cartão de crédito, foi aplicado o método não-paramétrico de Kaplan-Meier (4.1) e em 4.2.1 avaliada a capacidade preditiva de cada uma das covariáveis individualmente. Antes de se iniciar a modelação (4.2.2), foram geradas duas amostras aleatórias simples, de treino e de teste, sendo a primeira constituída por 70% dos clientes retirados aleatoriamente do total e a segunda constituída pelos restantes 30%. Por fim, é realizada com base na amostra de treino a avaliação da performance do modelo selecionado. Sendo posteriormente considerados estes resultados com a avaliação do modelo selecionado, utilizando a amostra de teste.

4.1 Resultados do modelo não-paramétrico de Kaplan-Meier

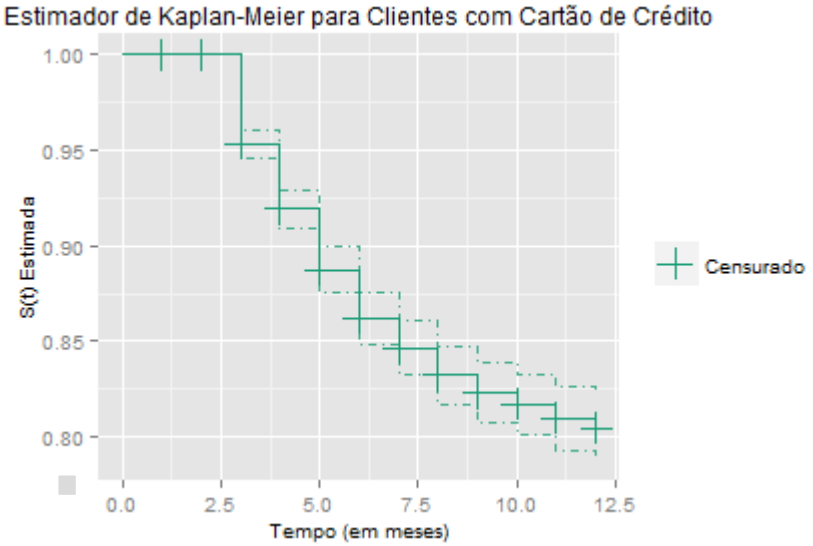
Apresenta-se a estimativa de Kaplan-Meier, na Tabela 4.1, para a amostra em análise neste estudo. Nos três primeiros meses não se verificam falhas uma vez que a mesma é considerada após 90 dias de incumprimento com a instituição financeira.

Tabela 4.1: Estimativa de Kaplan-Meier

t_j	n_j	D_j	$\hat{S}(t_j+)$	$I.C.\left(S(t_j+)\right)_{95\%}$	
3	3021	141	0,953	0,946	0,961
4	2498	89	0,919	0,909	0,929
5	2110	73	0,888	0,876	0,9
6	1820	52	0,862	0,849	0,876
7	1618	29	0,847	0,832	0,861
8	1468	25	0,832	0,817	0,848
9	1340	15	0,823	0,807	0,839
10	1222	9	0,817	0,801	0,833
11	1137	10	0,81	0,793	0,826
12	1062	7	0,804	0,787	0,821

A tabela apresenta os tempos ordenados até ao limite superior de 12 meses, que é o tempo de acompanhamento do estudo.

Gráfico 4.1: Estimativa de Kaplan-Meier



Para o cálculo da estimativa de Kaplan-Meier foi utilizada a função *survfit()* da biblioteca *{survival}* do R. Para os restantes argumentos foram utilizados os valores por omissão, como por exemplo, o *conf.int=0,95* que define intervalo de 95% de confiança do tipo *plain,* ou seja, linear[13]. Para a construção do Gráfico 4.1 foi utilizada a função *autoplot()* da biblioteca *{survMisc}* do R.

4.2 Resultados da estimação do modelo de Cox de riscos proporcionais

4.2.1 Capacidade Preditiva das Covariáveis

Posteriormente e para suporte à avaliação da capacidade preditiva, foram estimadas individualmente, pelo modelo de Cox, cada uma das variáveis identificadas na Tabela 2.2. Para tal, foi utilizada a função *coxph()* da biblioteca *{survival}* do R, tendo-se obtido os resultados (Tabela 4.2):

Tabela 4.2: Estimação Cox das Covariáveis

Covariável	Coeficiente	Erro-Padrão	Valor *p*	Covariável	Coeficiente	Erro-Padrão	Valor *p*
x1	-0,33220	0,09470	0,00045	x8	0,86400	0,16200	$0{,}99 \times 10^{-6}$
x2	-2,12600	0,29300	$0{,}38 \times 10^{-15}$	x9	-1,23500	0,17900	$0{,}47 \times 10^{-9}$
x3	-0,14360	0,09920	0,15000	x10	-0,34200	0,09500	0,00031
x4=1	-0,35900	0,15900	0,02410	x11	-0,08870	0,07200	0,22000
x4=2	-0,36500	0,12000	0,00230	x12	-0,03111	0,00470	$0{,}63 \times 10^{-10}$
x4=3	-0,65200	0,22900	0,00440	x13	-0,25240	0,02620	$0{,}2 \times 10^{-15}$
x4=4	-0,96200	0,15400	$0{,}43 \times 10^{-10}$	x14	-0,20270	0,02580	$0{,}39 \times 10^{-14}$
x5	0,35600	0,20600	0,08400	x15	$0{,}15 \times 10^{-5}$	$0{,}4 \times 10^{-5}$	0,68000
x6=2	0,11450	0,70950	0,84000	x16	1,46400	0,11100	$0{,}2 \times 10^{-15}$
x6=3	0,06820	0,30720	0,82000	x17	0,78880	0,09710	$0{,}44 \times 10^{-16}$
x6=4	0,41880	1,00170	0,68000	x18	-0,49500	0,11400	0,00002
x6=13	-12,9002	1062,68	0,99000	x19	0,4940	0,10	$0{,}46 \times 10^{-6}$
x6=99	0,81872	0,10203	$0{,}1 \times 10^{-14}$	x20	-5,28306	0,26909	$0{,}2 \times 10^{-15}$
x7	-0,03187	0,00368	$0{,}2 \times 10^{-15}$	x21	-0,12980	0,04870	0,00770

[13] Baseado na fórmula de Greenwood

Excluem-se as covariáveis não significativas ao nível de 5%. Face aos resultados obtidos na Tabela 4.2, são excluídas do modelo as covariáveis, x3 – Sexo, x5 – Indicador de trabalhador por conta de outrem, x6 – Código de natureza jurídica, x11 – Número de cartões por cliente e x15 – Diferença entre débitos e créditos.

4.2.2 Avaliação da Proporcionalidade dos Riscos de Falha

Da análise dos gráficos dos resíduos padronizados de Schoenfeld para cada uma das covariáveis em estudo, pode-se observar a ausência de proporcionalidade em x10 e x18. Depois de analisados os gráficos logaritmo da função taxa de falha acumulada estimada em *t* (vd. ANEXO V) verifica-se em x4 ausência de proporcionalidade e que em x1, x19 são observadas, embora não muito acentuadas, tendências ao longo do tempo. Tais tendências sugerem uma possível violação da suposição de taxas de falha proporcionais, mas que a análise gráfica não consegue confirmar. Para as restantes, para além da análise gráfica é realizado o teste de significância de correlação para os resíduos de Schoenfeld, cujos resultados são apresentados na Tabela 4.3.

Tabela 4.3: Teste de Proporcionalidade de Falha no Modelo de Cox

Covariáveis	ρ	χ^2	Valor-p	Covariáveis	ρ	χ^2	Valor-p
x1	-0,0292	0,384	0,535	x10	0,12	6,41	0,0114
x2	-0,00636	0,0182	0,893	x12	0,0119	0,0814	0,775
x4 = 1	-0,04247	0,812	0,3674	x13	-0,0017	0,0011	0,974
x4 = 2	-0,08229	3,06	0,0802	x14	0,0298	0,288	0,591
x4 = 3	0,00977	0,043	0,8358	x16	0,087	3,44	0,064
x4 = 4	-0,09595	4,175	0,041	x17	-0,0159	0,114	0,735
	Global x4	5,907	0,2062	x18	0,12	6,42	0,012
x7	0,0682	3,07	0,0799	x19	0,0768	2,65	0,103
x8	0,0003	0,00004	0,995	x20	-0,0463	6,2	0,013
x9	0,0683	2,09	0,149	x21	-0,0619	1,51	0,219

Verifica-se, pelo teste de correlação apresentado na tabela supra, que os valores dos coeficientes de correlação de Pearson para cada uma das variáveis são próximos de zero. Deste modo, os resultados sugerem não existir evidências de violação da suposição de taxas de falha

proporcionais de cada uma das variáveis em análise, para uma significância de 5%, exceto para a x10 e x18, que confirmam as conclusões da análise gráfica (Gráfico 4.2 e Gráfico 4.3).

Gráfico 4.2: Resíduos de Schoenfeld de x10 - Cartão de Uso Exclusivo em Combustível

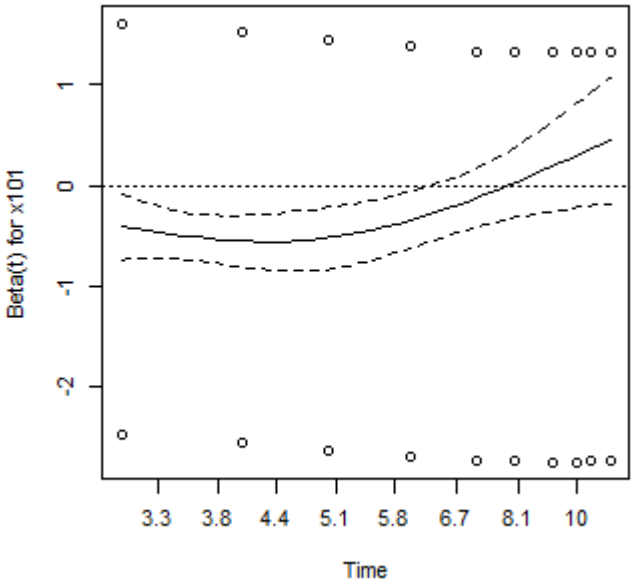
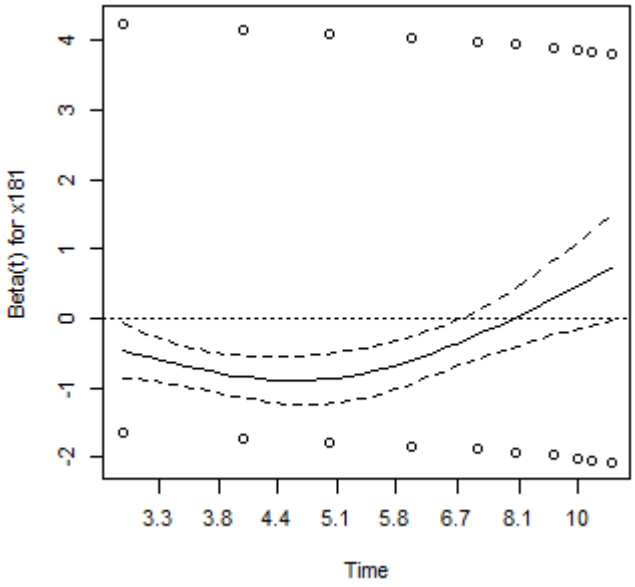


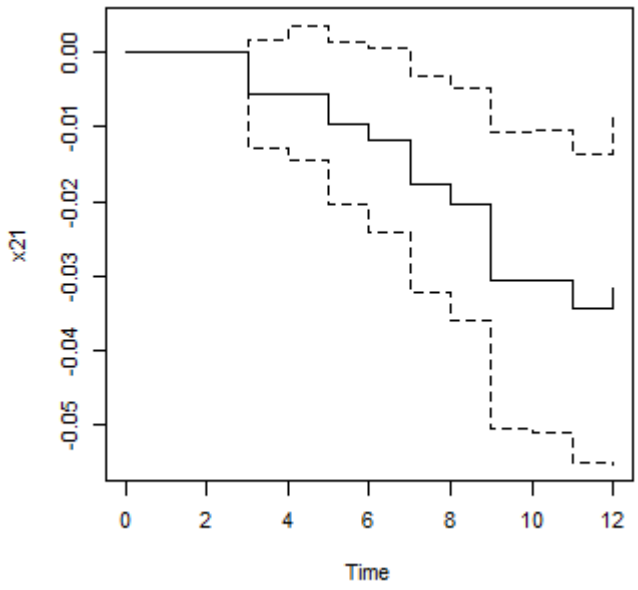
Gráfico 4.3: Resíduos de Schoenfeld de x18 - Cliente com Crédito Hipotecário



Para x20, cuja análise gráfica dos resíduos de Schoenfeld sugere a proporcionalidade (disponível no ANEXO VI), apesar do valor-p de 0,013 como o ρ é muito próximo de zero (-0,04) opta-se pela sua inclusão para o modelo inicial.

Para a covariável x21, cuja análise dos resíduos de Schoenfeld sugere alteração de sinal, foi analisada a função de regressão acumulada do modelo aditivo de Aalen. Da análise gráfica (Gráfico 4.4) constata-se para a covariável x21, que no décimo segundo mês existe uma inversão da tendência o que sugere violação da hipótese de proporcionalidade. Neste sentido, a referida covariável será retirada deste estudo.

Gráfico 4.4: Resíduos de Schoenfeld de x21 - Rácio entre Débitos e Créditos



A covariável x4, que agrupa os clientes por habilitações literárias, apesar do gráfico do logaritmo da função taxa de falha acumulada estimada em *t*, sugerir que não é proporcional, os resíduos de Schoenfeld e os valores da correlação de Pearson sugerem o inverso. Dado que se tratam de cinco categorias e duas delas são, visivelmente sobrepostas, pela análise das curvas de Kaplan-Meier (Gráfico 4.6) e do logaritmo da função taxa de falha acumulada estimada *versus t* (Gráfico 4.5).

Gráfico 4.5: Logaritmo da função taxa de falha acumulada estimada versus t da x4

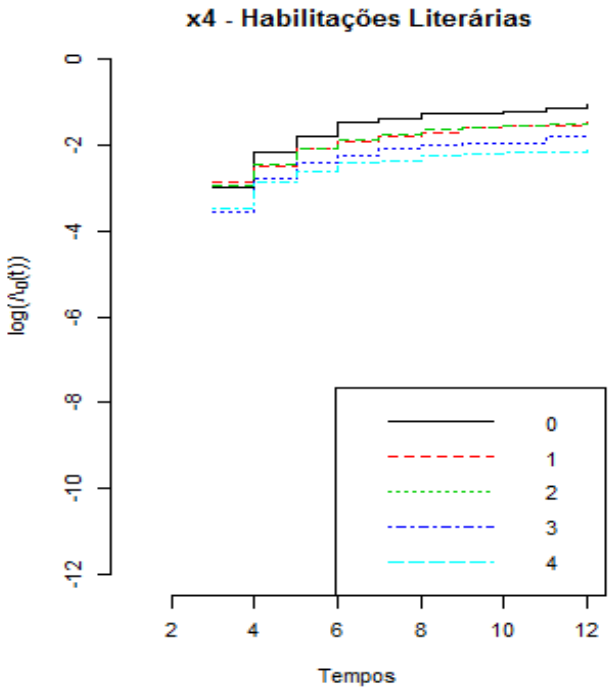
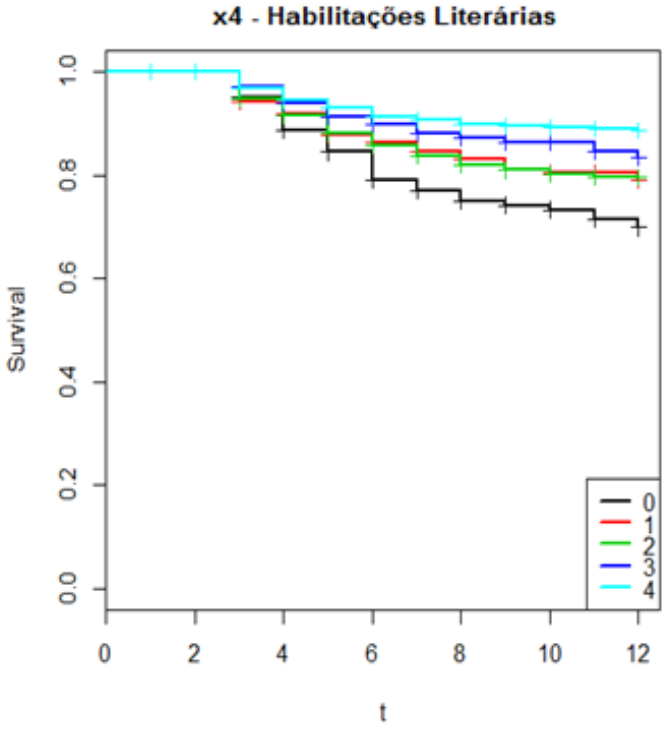


Gráfico 4.6: Estimativa de Kaplan-Meier de x4 - Habilitações Literárias



Para aferir da diferença entre grupos, primeiro foi verificado através do teste *logrank* de comparação das cinco estimativas de Kaplan-Meier das funções de sobrevivência por grupo, tendo sido obtido um valor-p de $6{,}33 \times 10^{-9}$, inferior a 0,05, ou seja, para uma significância de 5% existe uma diferença significativa entre os cinco grupos. Foi efetuada a análise dos grupos, dois a dois, cujos resultados se apresentam na tabela seguinte:

Tabela 4.4: Teste *Logrank* de comparação entre grupos

Sub-conjunto de Categorias	Valor-*p*
0 x 1	0,01910
0 x 2	0,00169
0 x 3	0,00213
0 x 4	$4{,}15 \times 10^{-11}$
1 x 2	0,97600
1 x 3	0,21500
1 x 4	0,00053
2 x 3	0,18800
2 x 4	0,00002
3 x 4	0,18900

Para um nível de significância de 5%, constata-se que não podemos garantir que exista uma diferença significativa entre os grupos 1 x 2, 1 x 3, 2 x 3 e 3 x 4. Face ao exposto, a covariável x4 foi transformada em três categorias (x4t ou $\tilde{x}4$):

[1] Habilitações Literárias Desconhecidas

[2] Ensino Obrigatório ou Curso Médio

[3] Curso Superior

Após a transformação na covariável $\tilde{x}4$, obtemos um valor-p para o teste *logrank* de $p = 7{,}09 \times 10^{-10}$, para uma significância de 5% existe uma diferença significativa entre os três grupos.

Para os cálculos de teste *logrank* foi utilizada a função *surdiff()* da biblioteca *{survival}* do R.

Através da análise dos gráficos dos resíduos padronizados de Schoenfeld e do logaritmo da função taxa de falha acumulada estimada em *t*, constata-se a proporcionalidade de $\tilde{x}4$.

Gráfico 4.7: Logaritmo da função taxa de falha acumulada estimada versus t de $\tilde{x}4$

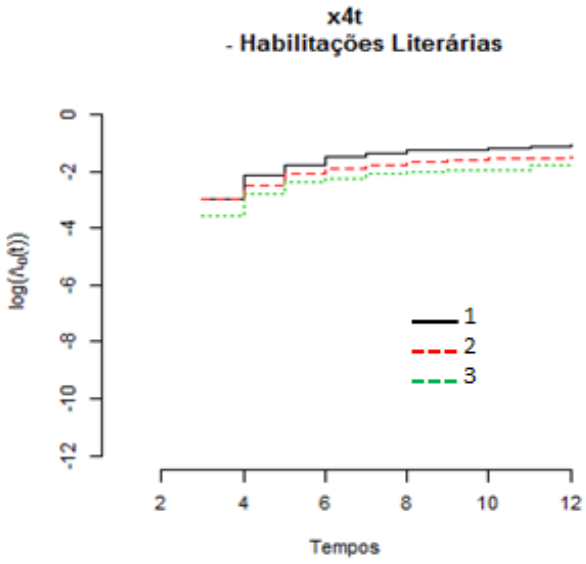
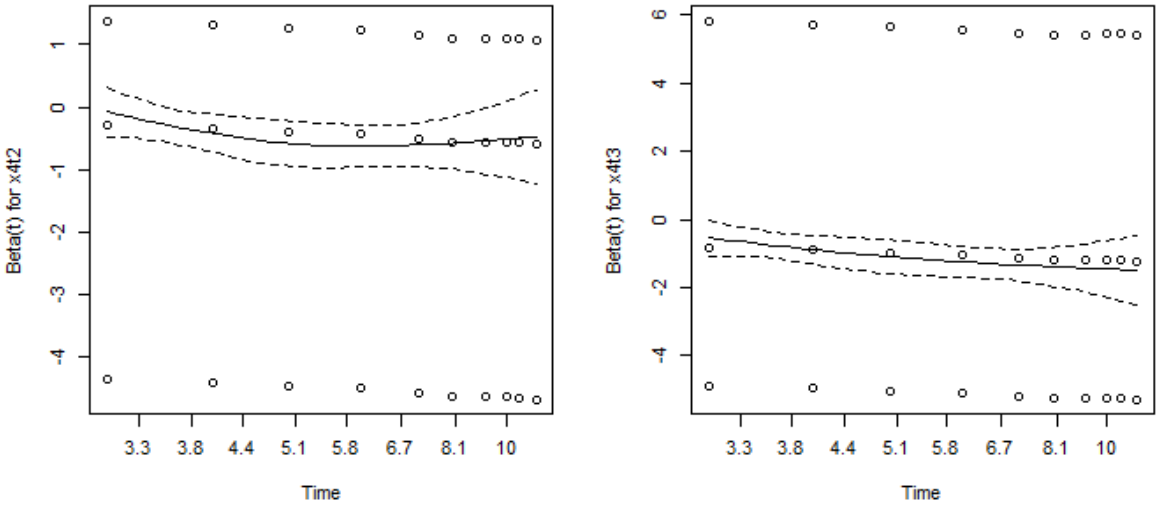


Gráfico 4.8: Resíduos padronizados de Schoenfeld de $\tilde{x}4$ - Habilitações Literárias (revista)



Terminada a análise exploratória, foram eleitas para modelação as covariáveis, x1, x2, $\tilde{x}4$, x7, x8, x9, x12, x13, x14, x16, x17, x19 e x20.

Antes de se iniciar a modelação, foram geradas duas amostras aleatórias simples, de treino e de teste, sendo a primeira constituída por 70% dos clientes retirados aleatoriamente do total e a

segunda constituída pelos restantes 30%. A amostra de treino[14] será utilizada na construção dos vários modelos e a amostra de teste[15] será utilizada para confirmar a performance do modelo (Tabela 4.5).

Tabela 4.5: Constituição das amostras de Treino e Teste

Amostra	Periodo	Nr Contratos
Treino	Dez/2005 a Dez/2006	3.050
Teste	Dez/2005 a Dez/2006	1.308

4.2.3 Estimação de Cox de Riscos Proporcionais

Para a estimação Cox de Riscos Proporcionais, foi utilizado o método *stepwise* (*backward* + *forward*), recorrendo à função *step()* da biblioteca *{stats}* do R, com a opção *direction* igual *"both"*. O uso deste procedimento é adequado num estudo empírico de análise preditiva em que o principal objetivo é obter um modelo e um conjunto de preditores para explicar o fenómeno em estudo, conforme Scott Menard (2001).

Apresentam-se os resultados do modelo de Cox ajustado aos dados, com as covariáveis eleitas para modelação, i.e., x1, x2, $\tilde{x}4$, x7, x8, x9, x12, x13, x14, x16, x17, x19 e x20:

[14] Amostra sobre a qual são estimadas as regras de classificação do modelo.

[15] Esta amostra não participa na estimação do modelo, é utilizada para conferir a medida do seu desempenho.

Tabela 4.6: Regressão inicial do Modelo de Cox (modelo irrestrito)

Covariável	Coeficiente	Erro-Padrão	Valor-*p*	Covariável	Coeficiente	Erro-Padrão	Valor-*p*
x1	0,0422	0,1240	0,7337	x12	-0,0154	0,0060	0,0102
x2	-0,8538	0,2996	0,0043	x13	-0,1502	0,0608	0,0132
$\tilde{x}4=2$	-0,1680	0,1389	0,2262	x14	0,0521	0,0619	0,3901
$\tilde{x}4=3$	-0,4940	0,1912	0,0097	x16	1,0455	0,1348	$6{,}8 \times 10^{-15}$
x7	0,0046	0,0052	0,3766	x17	0,4891	0,1220	$4{,}7 \times 10^{-5}$
x8	-0,0640	0,2511	0,7991	x19	0,3253	0,1206	0,0068
x9	-0,5269	0,3212	0,1018	x20	-4,6067	0,2997	$0{,}2 \times 10^{-16}$

A partir deste conjunto de covariáveis, foi calculado o modelo base a partir do método *stepwise* (*backward* + *forward*), tendo-se obtido x2, $\tilde{x}4$, x9, x12, x13, x16, x17, x19 e x20. O objetivo é encontrar a partir deste, um modelo o mais parcimonioso possível, desta forma a metodologia utilizada para a modelação foi *general-to-specific*, ou seja, para um nível de significância de 5% foram, sucessivamente, excluídas do modelo as variáveis estatisticamente não significativas através do teste de Wald e analisada a proporcionalidade de cada uma.

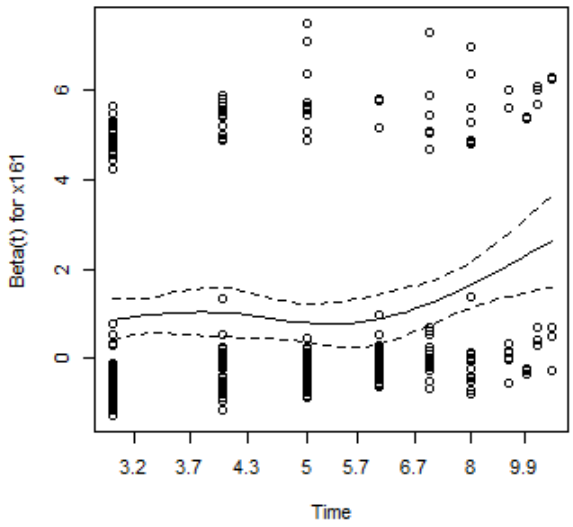
Tabela 4.7: Seleção de covariáveis usando o modelo de regressão de Cox

	Modelo	$\mathcal{L}$	TRV	Valor *p*	AIC
0	x1 + x2 + $\tilde{x}$4 + x7 + x8 + x9 + x12 + x13 + x14 + x16 + x17 + x19 + x20	-1953,64	n.a	n.a	3935,29
1	x2 + $\tilde{x}$4 + x9 + x12 + x13 + x16 + x17 + x19 + x20	-1954,64	1,9916	0,7372	3929,28
2	x2 + $\tilde{x}$4 + x12 + x13 + x16 + x17 + x19 + x20	-1955,73	4,1847	0,5231	3929,47
3	x2 + x12 + x13 + X16 + x17 + x19 + x20	-1959,47	11,6394	0,0705	3932,93

O TRV e o AIC foram utilizados para comparar os diferentes modelos obtidos. Da Tabela 4.7 conclui-se, através dos resultados do TRV, que todos os modelos são adequados, significando que o efeito da exclusão das covariáveis x1, $\tilde{x}4$, x7, x8, x9 e x14, evidenciam a validade destas restrições adicionais. Face ao exposto e tendo como referência os resultados do TRV (TRV=4,1847, *valor-p*=0,5231) não se rejeitando a hipótese nula de que o modelo de interesse é adequado e a minimização do AIC (3929,47), considera-se que o modelo 2 é o que melhor se ajusta aos dados.

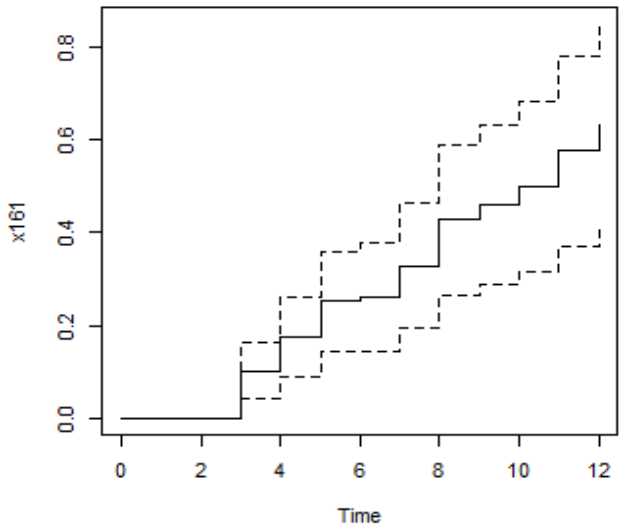
Não se considera o modelo 1 dado que a covariável x9 é não significativa a 5% (com um *valor-p* de 0,1539), optando-se por não se considerar a mesma na modelação (resultados no ANEXO VII). Para verificar a suposição de taxas de falha proporcionais no modelo de Cox ajustado para os dados em análise, foram utilizados os métodos gráficos utilizados em 3.3.2. Desta forma, da análise dos gráficos dos resíduos padronizados de Schoenfeld e dos respetivos resultados do teste de Pearson (Tabela 4.8), para cada uma das covariáveis do modelo 2, pode-se observar uma possível ausência de proporcionalidade em x16 (Gráfico 4.9).

Gráfico 4.9: Resíduos padronizados de Schoenfeld de x16 - Inibição de Uso de Cheque



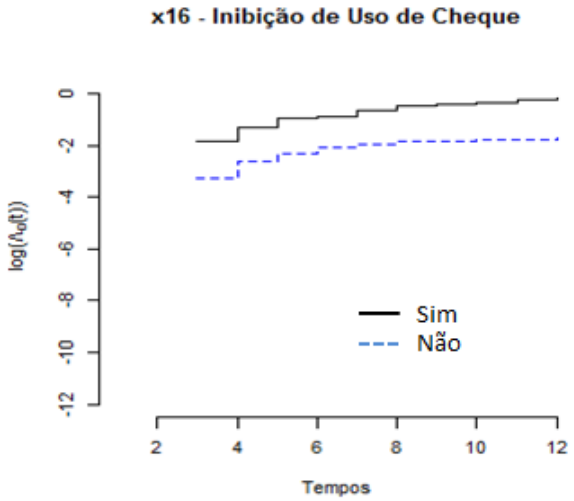
Analisando a função de regressão acumulada do modelo aditivo de Aalen, para o modelo 2 (Gráfico 4.10), verifica-se uma inclinação constante nos 12 meses em estudo, notando-se apenas um ligeiro abrandamento no 6º mês.

Gráfico 4.10: Estimativas das funções de regressão acumulada de x16 - Inibição de Uso de Cheque



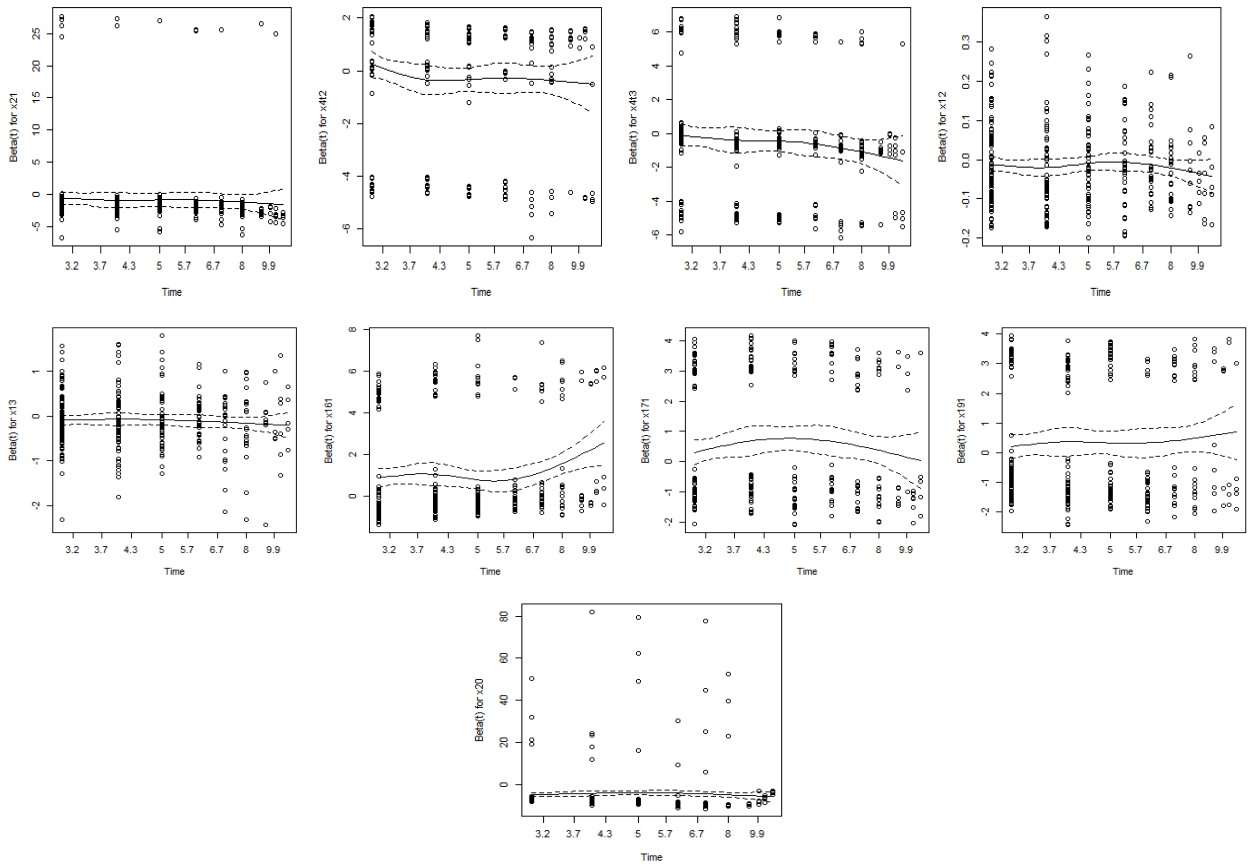
Dado tratar-se de uma variável categórica (binária), é analisado o gráfico do logaritmo da função taxa de falha acumulada estimada ($\widehat{\Lambda_{0j}}(t)$) em *t* (Gráfico 4.11).

Gráfico 4.11: Logaritmo da função taxa de falha acumulada *versus* tempo para x16



Uma vez que as curvas do logaritmo apresentam diferenças aproximadamente constantes ao longo do tempo e apesar do resultado do teste de proporcionalidade (Tabela 4.8) apurado para x16 – valor-p = 0,0115 – considera-se válida a suposição de proporcionalidade para os clientes com marca de inibição de uso de cheque.

Gráfico 4.12: Resíduos padronizados de Schoenfeld para as covariáveis do modelo 2



Na Tabela 4.8, são apresentados os resultados do teste de proporcionalidade das taxas de falha, no modelo 2.

Tabela 4.8: Testes de proporcionalidade no modelo ajustado

Covariáveis	ρ	χ^2	Valor p
x2	-0,0466	0,724	0,3949
x̃4=2	-0,0871	2,428	0,1192
x̃4=3	-0,1192	4,655	0,0310
x12	-0,0494	0,886	0,3466
x13	-0,0409	0,587	0,4436
x16	0,1359	6,383	0,0115
x17	-0,0219	0,154	0,6947
x19	0,0517	0,849	0,3569
x20	-0,0075	0,106	0,7446
GLOBAL		16,079	0,0652

Apresentam-se, na Tabela 4.9, os resultados do modelo de Cox ajustado aos dados (modelo 2), a partir das covariáveis eleitas para modelação.

Tabela 4.9: Resultados do ajustamento do modelo de Cox para o Modelo 2

Covariável	Coeficiente	$e^{\widehat{\beta}_i}$	Erro-Padrão	Valor p
x2	-0,8502	0,4273	0,2985	0,0044
x̃4=2	-0,1540	0,8573	0,1382	0,2651
x̃4=3	-0,4877	0,6140	0,1852	0,0084
x12	-0,0151	0,9850	0,0055	0,0057
x13	-0,1073	0,8983	0,0345	0,0018
x16	1,0567	2,8769	0,1340	$3{,}11 \times 10^{-15}$
x17	0,5050	1,6570	0,1189	$2{,}14 \times 10^{-5}$
x19	0,3354	1,3985	0,1194	0,0049
x20	-4,6167	0,0099	0,2983	2×10^{-16}

A partir da expressão (3.11), podemos assim definir o modelo de Cox ajustado para o modelo 2:

$$\lambda(t|x) = \lambda_0(t)\exp(-0{,}8502x_2 - 0{,}1540\tilde{x}_{4,2} - 0{,}4877\tilde{x}_{4,3} - 0{,}0151x_{12} - 0{,}1073x_{13} + 1{,}0567x_{16} + 0{,}5050x_{17} + 0{,}3354x_{19} - 4{,}6167x_{20})$$

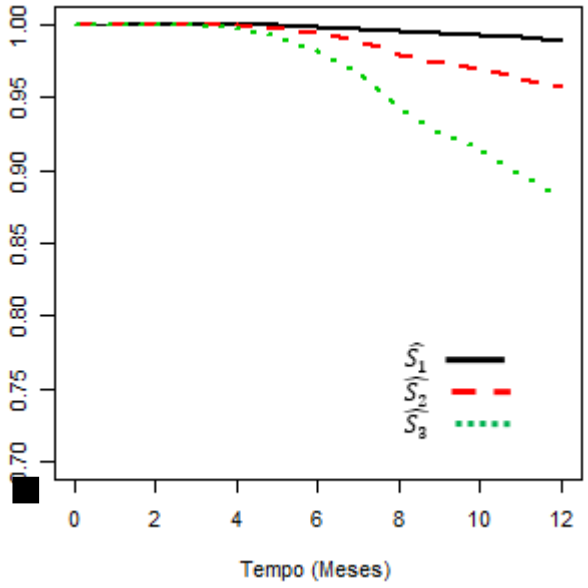
Considerando as razões de taxa de falha $(e^{\widehat{\beta_i}})$ podem ser obtidas as seguintes interpretações a partir da Tabela 4.9:

a) O risco de incumprimento nos clientes desta instituição financeira (x2), que tiverem um produto de poupança contratado, relativamente aos que não têm é 0,43.
b) O risco de incumprimento nos clientes com ensino superior ($\tilde{x}42$= 3), relativamente aos que não têm ($\tilde{x}42$= 2), é 0,32 vezes.
c) O aumento da idade de um cliente (x12), por exemplo de 30 anos comparativamente a um cliente de 22 anos, fixadas as restantes covariáveis, está associado um risco de incumprimento de 0,88 vezes.
d) O aumento do logaritmo do total de crédito em conta à ordem de um cliente (x13), de 7,31 comparativamente a um cliente com 6,91, fixadas as restantes covariáveis, tem uma taxa de incumprimento estimada em 0,96 vezes.
e) O risco de incumprimento nos clientes com inibição de uso de cheque (x16) é estimada em 2,88 vezes a de um cliente que pode utilizar cheques.
f) A taxa de incumprimento nos clientes com crédito particular (x17) é estimada em 1,66 vezes a de um cliente que não têm este tipo de produto.
g) A taxa de incumprimento nos clientes com ordenado domiciliado ou conta ordenado (x19) é estimada em 1,40 vezes a de um cliente que não têm este tipo de produto.
h) Um cliente com 5 meses de antiguidade (x20) comparativamente a um cliente com 6 meses, fixadas as restantes covariáveis, tem associado um aumento de risco de incumprimento, estimado em 101,16 vezes. O inverso, tem um risco associado de 0,01.

No Gráfico 4.13 apresentam-se, a título de exemplo, curvas de sobrevivência estimadas para diferentes perfis de cliente. A curva $\widehat{S_1}$ilustra clientes com 1 mês de antiguidade, 30 anos de idade, titulares de um curso superior, com um total mensal de crédito em conta positivo, sem incidentes na utilização de cheques, não são titulares de um crédito particular e não têm conta ordenado. A curva $\widehat{S_2}$, define clientes com 1 mês de antiguidade, 30 anos de idade, sem curso superior, com

um total mensal de crédito em conta nulo, sem incidentes na utilização de cheques, titulares de um crédito particular e com conta ordenado. $\widehat{S_3}$ é semelhante a $\widehat{S_2}$, excetuando na variável x16 que identifica clientes que tiveram incidentes bancários na utilização de cheques. Do Gráfico 4.13 observa-se que as curvas de sobrevivência estimadas para $\widehat{S_1}$ e $\widehat{S_2}$ não apresentam diferenças acentuadas. Na $\widehat{S_3}$ observa-se um decréscimo desta curva para os clientes com incidente bancário na utilização de cheques.

Gráfico 4.13: Curvas de sobrevivência



Conclui-se que a função de sobrevivência é afetada pelas covariáveis x16 – Indicador de Inibição de Uso de Cheque, x17 – Indicador de Cliente com Crédito Particular e x19 – Indicador de Cliente com Conta Ordenado, ou seja, a presença destas implica uma função de sobrevivência menor que a sua ausência conjunta.

Analisando os resíduos martingala e da desviância, do modelo 2, constata-se que os resíduos martingala confirmam a forma funcional do modelo, assimétrica à direita e a linha LOWESS sem tendência em torno de zero, os resíduos da desviância apresentam valores positivos superiores a 3, o que significa que as referidas observações são potenciais valores atípicos.

Gráfico 4.14: Resíduos martingala e desviância *versus* preditor linear do modelo de Cox final

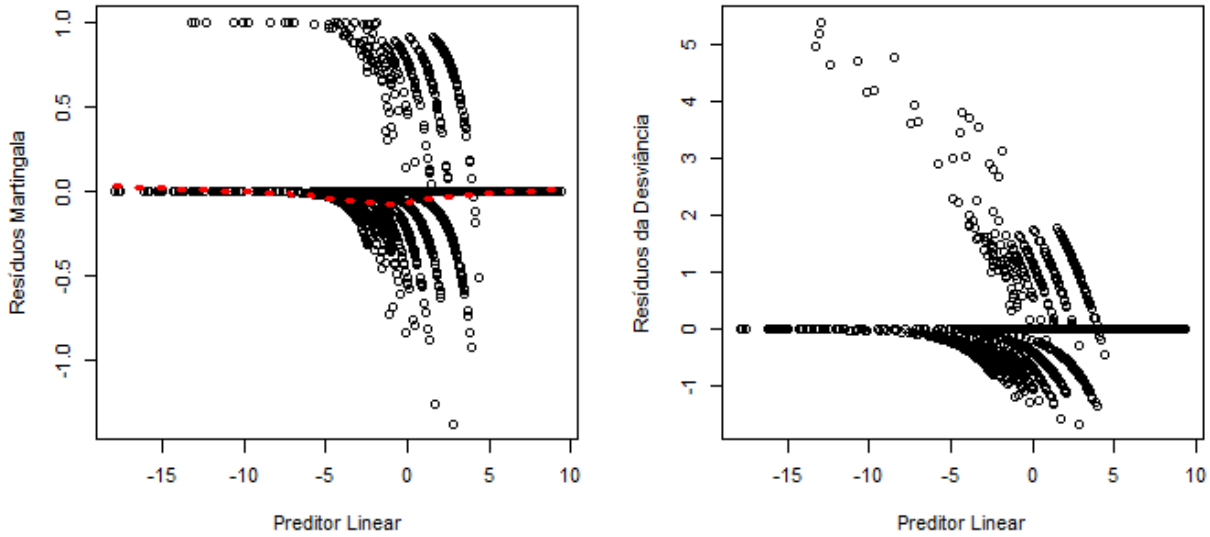
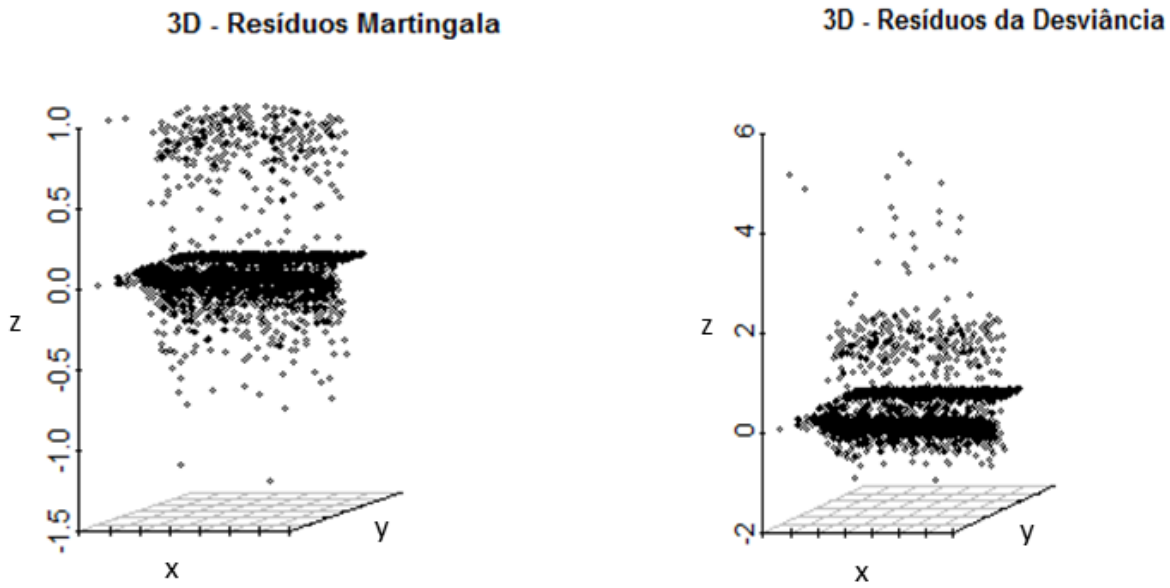


Gráfico 4.15: Gráficos 3D dos resíduos martingala e da desviância



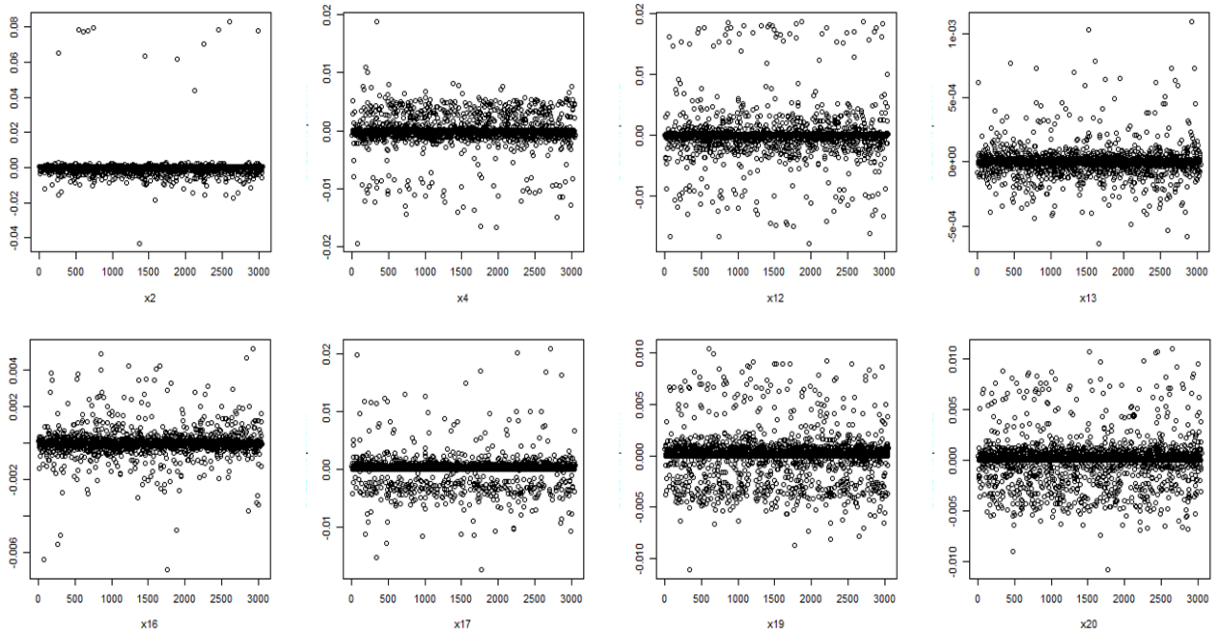
Para a obtenção do Gráfico 4.15, foi utilizada a função *scatterplot3d()*, da biblioteca com o mesmo nome, cujo objetivo é a visualização de dados multivariados num espaço tridimensional. Apesar

de só existirem estruturas de dados bidimensionais para os resíduos apresentados, a função *scatterplot3d()* gera a estrutura tridimensional.

O sinal positivo dos resíduos da desviância das observações em análise significa que as mesmas tiveram um tempo de sobrevivência inferior ao esperado, inversamente, resíduos com sinal negativo significa observações com um tempo de sobrevivência superior ao esperado (Paul Allison, 1995). Observando os dados para valores superiores a 3, constata-se que se trata de valores atípicos que representam clientes com uma antiguidade elevada que entraram em incumprimento.

Analisando a influência ou impacto de cada observação (Gráfico 4.16), através dos gráficos dos resíduos DFBETAS, verifica-se a existência de valores reduzidos – com valor absoluto menor que 1 – não evidenciando pontos influentes no ajustamento.

Gráfico 4.16: Resíduos DFBETAS *versus* cada covariável presente no modelo de Cox final

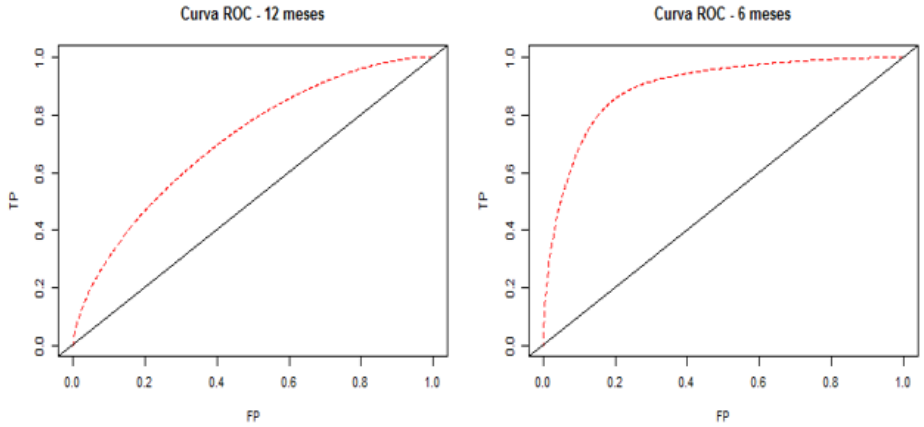


4.3 Avaliação da performance do modelo de Cox de riscos proporcionais

Para avaliar a performance do modelo de Cox final, e ter uma informação mais vasta sobre a qualidade de previsão na amostra, utiliza-se a curva ROC. A mesma foi calculada recorrendo à biblioteca do R *{risksetROC}* e respetiva função com o mesmo nome.

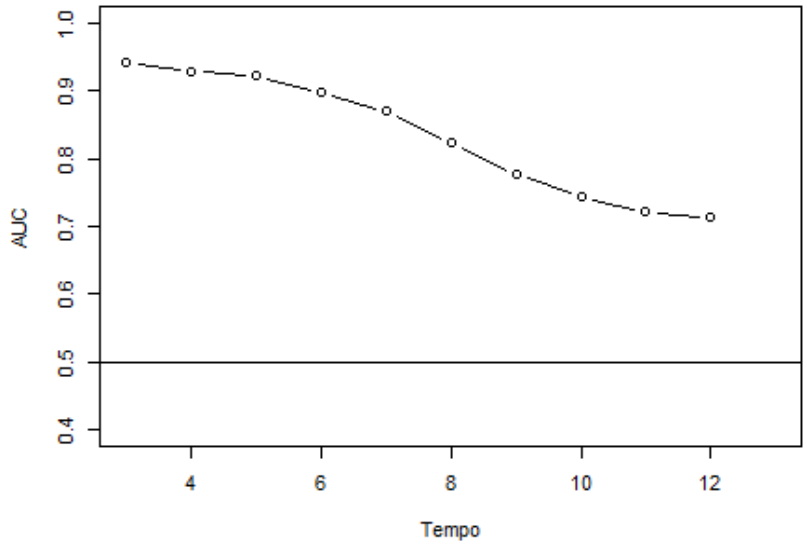
Uma vez que se trata de uma ROC dependente do tempo e considerando 12 meses como o tempo máximo da experiência, a partir da opção *predict.time* da função *risksetROC()* foram avaliadas duas curvas ROC para 6 e 12 meses respetivamente.

Gráfico 4.17: Curvas ROC a 12 e 6 meses, do modelo de Cox final



Da análise aos resultados obtidos a 6 meses, para a amostra de treino, constata-se que o modelo apresenta uma qualidade muito boa evidenciada por uma curva ROC muito côncava. A 12 meses, para a amostra de treino, o modelo apresenta uma qualidade média-baixa para responder aos objetivos da análise comportamental.

Gráfico 4.18: Evolução temporal da AUC para o período de sobrevivência



Os resultados apurados para a AUC e Gini da amostra de teste, confirmam os resultados supra obtidos para a amostra de treino (Tabela 4.10).

Tabela 4.10: Resultados da AUC e Gini para as amostras de treino e teste

	AUC		Gini	
Meses	6	12	6	12
Treino	89,69%	71,28%	79,39%	42,56%
Teste	94,95%	76,54%	89,90%	53,09%

Com base nos resultados obtidos, podemos afirmar que o modelo, para as amostras de treino e teste, a 6 meses tem uma boa capacidade de identificar a percentagem de clientes incumpridores em cada instante. Sendo que a 12 meses os modelos apresentam uma faculdade média-baixa para a identificação dos clientes em análise (resultados na Tabela 4.10).

CAPÍTULO 4

APRESENTAÇÃO DE RESULTADOS

CAPÍTULO 5

CONCLUSÕES

5 Conclusões e Perspetivas Futuras

Numa altura em que se preveem grandes alterações nos sistemas de informação informáticos para as organizações, tendo por base a denominada 3ª plataforma, i.e., *Sistemas Cloud, Big Data, Mobile* e *Social Business*, as instituições financeiras desenvolvem sofisticadas técnicas de modelação estatística. O acesso direto a dados, a utilização diária de ferramentas de análise e avaliação dos mesmos no seio da direção de empresas, são ingredientes fundamentais para o êxito nas vendas e o aumento das receitas. De acordo com a unidade britânica de consultoria *Economist Intelligence Unit* (EIU), 97% das empresas que alcançaram ou excederam as suas metas de vendas anteriores revelaram que haviam disponibilizado aos seus funcionários e gerentes acesso direto e em tempo real a dados e/ou contas de clientes[16], a análise estatística do comportamento do consumidor torna-se um fator de sucesso para o negócio.

Esta dissertação tinha como objetivo criar uma metodologia de modelação estatística para o tempo de sobrevivência até ao incumprimento na utilização do cartão de crédito, tendo por base uma amostra de clientes particulares, de cartões de crédito, fornecidos por uma instituição financeira, com um horizonte temporal de Dez/2005 a Dez/2006, tendo sido omissa qualquer possibilidade de identificação dos mesmos.

5.1 Revisão da metodologia seguida

Foram analisadas individualmente cada uma das covariáveis, de uma amostra inicial de dados de uma instituição financeira, para decisão da sua inclusão neste estudo. Para tal foram analisadas estatísticas descritivas, existência de omissos ou evidência de erros na recolha da informação. Depois de realizado o tratamento dos dados, foram avaliadas as estatísticas WoE, z-score e IV

[16] Oracle ONE Magazine http://www.oracle.com/us/dm/one-magazine/home.html?qInfo=Winter2016&qCountry=Portugal&elq_mid=34177&sh=171226122492281112152654 25&cmid=EMEAFM15034858MPP381C003

para aferir do potencial preditivo de cada uma das covariáveis categóricas candidatas a esta dissertação. Posteriormente, com base nas covariáveis selecionadas, foi criada uma amostra de sobrevivência. Foram excluídas as observações em incumprimento nos três primeiros meses da análise de sobrevivência e considerado apenas a censura à direita.

Para aferir da probabilidade de sobrevivência de um cliente não incumprir em qualquer momento de tempo posterior ao início da experiência, foi aplicado inicialmente o método não-paramétrico de Kaplan-Meier. Posteriormente as covariáveis foram estimadas individualmente, pelo modelo de Cox e para um nível de significância de 5%, foram excluídas do estudo aquelas identificadas como estatisticamente não significativas através do teste de Wald.

Para as restantes foi avaliada a proporcionalidade da taxa de falha, recorrendo à análise gráfica do logaritmo da função taxa de falha acumulada em cada grupo no tempo, aos gráficos dos resíduos de Schoenfeld e coeficientes de correlação de Pearson entre os resíduos padronizados e uma função do tempo. Para a tomada de decisão sobre a evidência ou não de proporcionalidade, são avaliadas estas três abordagens em conjunto sendo a análise gráfica determinante para a decisão. Caso a informação disponível não seja suficiente para decisão, foi utilizada a metodologia proposta por Hosmer, D. e Royston, P. (2002) na qual são avaliadas as variações no tempo das covariáveis em análise, através do gráfico da regressão do modelo aditivo de Aalen. Foram excluídas do estudo as covariáveis que falharam a condição de proporcionalidade de taxa de falha.

A partir do conjunto de covariáveis eleitas para modelação, foi calculado o modelo base a partir do método *stepwise* (*backward* + *forward*). A partir do modelo base a metodologia utilizada foi a *general-to-specific*, ou seja, para um nível de significância de 5% foram, sucessivamente, excluídas do modelo as variáveis estatisticamente não significativas através do teste de Wald. Dos modelos obtidos, foi avaliada a capacidade de ajustamento dos mesmos aos dados com o TRV e o AIC. São analisados os resíduos martingala e da desviância do modelo eleito, constatando-se a partir dos resíduos da desviância a existência de valores atípicos. A análise das observações com resíduos da desviância superiores a 3 permitiu identificar clientes com uma antiguidade elevada e que

entraram em incumprimento. Tendo em consideração o tamanho da amostra e o número reduzido de valores atípicos optou-se por mantê-los no estudo.

5.2 Principais conclusões

Conclui-se que as covariáveis Indicador de Cliente com Crédito Particular e Indicador de Cliente com Conta Ordenado, afetam a função de sobrevivência e que a introdução da covariável Indicador de Inibição de Uso de Cheque acentua o decréscimo da curva de sobrevivência. A presença destas covariáveis implica uma função de sobrevivência menor que a sua ausência conjunta. Este é um resultado expectável e confirma que um cliente com histórico de incidentes bancários, neste estudo, na utilização do cheque, tem uma propensão superior a voltar a incumprir.

De destacar o risco de incumprimento de um cliente com conta ordenado ser 1,40 vezes a de um cliente sem este produto. Este facto parece indiciar que clientes detentores de conta ordenado tiveram, no passado, garantida a contratação de outros produtos, como o cartão de crédito, sem uma análise de risco prévia. Esta situação deveu-se possivelmente a políticas de captação de clientes. Sugere-se em futuras campanhas de marketing uma análise de risco do cliente que inclua algumas das covariáveis aqui apresentadas, em particular o indicador de domiciliação de ordenado e respetiva conta ordenado. A antiguidade do cliente é, igualmente, relevante para a comercialização de cartões de crédito nas instituições financeiras, neste sentido, quanto mais antiga é a relação do cliente com o banco menor é o risco de incumprimento.

Conclui-se para a informação em análise, que as campanhas de comercialização de cartões de crédito desta instituição financeira, deveriam ser direcionadas a clientes com algum histórico de relacionamento com a mesma, com mais 30 anos de idade, titulares de um curso superior, com um total mensal de crédito em conta positivo, sem incidentes na utilização de cheques, sem crédito particular e sem conta ordenado.

Por fim é efetuada uma avaliação da performance do modelo considerado no estudo empírico para as amostras de treino e teste, recorrendo aos resultados da área abaixo da curva ROC

dependente do tempo e do índice de Gini. Conclui-se que o modelo a 6 meses apresenta uma qualidade muito boa evidenciada por uma curva ROC muito côncava, a 12 meses apresenta uma qualidade média-baixa, tendo como referência os *benchmarks* utilizados nesta dissertação.

5.3 Trabalho futuro

As instituições financeiras têm como principal negócio, o da compra e venda de riscos. Com as atuais condições macroeconómicas, o momento em que vive a banca em Portugal e a expectável entrada de novos *players* no mercado da banca de retalho[17], torna-se cada mais importante conhecer o cliente e decidir rapidamente a venda de um produto sem colocar em perigo o negócio. Neste sentido, um dos pontos que merecerá pesquisa futura é a regressão de Cox com covariáveis dependentes do tempo e o modelo aditivo de Aalen, para possibilitar a inclusão de outras covariáveis que poderão ser pertinentes para a análise, como por exemplo o rácio entre débitos e créditos excluída por se realizar a condição de proporcionalidade da taxa de falha.

A amostra utilizada nesta dissertação é censurada à direita, ou seja, os indivíduos são observados por um período de tempo até falha ou até final da experiência sem que antes se observe qualquer falha, no entanto, é usual a recolha inicial dos dados conter múltiplas observações no tempo de um mesmo individuo. Utilizar modelos de AS em dados de painel, poderão minimizar esta questão observando o individuo por um período tempo, avaliando se o mesmo regista eventos de falha repetidos.

Com a crescente volumetria de dados e covariáveis suscetíveis de contextualização em análise comportamental, que está a ocorrer nas organizações[18], surge um ponto que merece pesquisa, é a análise de componentes principais (ACP) em dados de sobrevivência. O método apresentado por Shen e Huang (2006), o PC-CR (*Principal Components for Cox Regression*) será um aspeto a

[17] Jornal Económico - http://economico.sapo.pt/noticias/apple-google-e-facebook-preparam-revolucao-na-banca_196053.html

[18] Jornal Público - https://www.publico.pt/economia/noticia/big-data-nova-forma-de-fazer-sentido-1715348?page=2#/follow

desenvolver para otimizar o processo de seleção de covariáveis, tal como a utilização de ACP para detalhar conclusões a partir do modelo final obtido.

CAPÍTULO 5

CONCLUSÕES

Referências Bibliográficas

Allison, Paul (1995) *Survival Analysis Using SAS a Practical Guide*. Cary, NC: SAS Institute Inc., 165-181

Baesens, B. (2015), *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, Wiley and SAS Business Series

Bellotti, T., Crook, J. (2007) Credit Scoring With Macroeconomic Variables Using Survival Analysis, Credit Research Centre, Management School and Economics, University of Edinburgh

Beaver, W.H. (1966), *Financial Ratios as Predictors of Failure,* Journal of Accounting Research, Vol. 4, Issue Empirical Research in Accounting: Selected Studies, Pages 71-111.

Chorão, L.R. (2005*). Credit Scoring: Logit vs Redes Neuronais Artificiais. Um exemplo aplicado a cartões de Crédito.* Dissertação de Mestrado em Estatística e Gestão de Informação, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Lisboa., 7-17, 43-59, 111-123

Finlay, Steven (2010) *Credit Scoring, Response Modelling and Insurance Rating. A Practical Guide to Forecasting Consumer Behaviour*. London. Palgrave Macmillan, 66-159

Giolo, S., Colosimo, E. (2006) *Análise de Sobrevivência Aplicada.* Edgard Blucher, 1-34, 115-127, 155-199, 225-227

Heagerty, J., P., Zheng, Y. (2005), *Survival Model Predictive Accuracy and ROC Curves*, Biometrics No. 61, pp. 92-105

Hosmer, D., Royston, P., (2002) *Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model*. The Stata Journal, Number 4, pp. 331–350, em http://www.stata-journal.com/sjpdf.html?articlenum=st0024

Li, H. e Gui, J., (2004) *Partial Cox regression analysis for high-dimensional microarray gene expression data.* Bioinformatics 20: 1208-215

Ligges, U. e Machler, M. (2003) *Scatterplot3d – an R Package for Visualizing Multivariate Data.* Journal of Statistical Software 8(11), 1–20.

Kleiber, C., Zeileis, A. (2008). *Applied Econometrics with R*. Springer

McKeague, I. W., Utikal, K. J. (1991). Goodness-of-Fit Tests for Additive Hazards and Proportional Hazards Models. Scandinavian Journal of Statistics, 18(3), 177–195. Em http://www.jstor.org/stable/4616202

Mello, F. M. e Guimarães, R. C. (2015), *Métodos Estatísticos para o Ensino e Investigação nas Ciências da Saúde.* Edições Sílabo, 357-403

Menard, S. (2002). *Applied Logistic Regression Analysis (Sage University Paper series on Quantitative Applications in the Social Sciences, 2nd ed. Vol. 07-106)*. Thousand Oaks, CA: Sage., 63-66, 80-90

Paulino, C. D., Pestana, D., Branco, J., Singer, J., Barroso, L. e Bussab, W. (2011). *Glossário Inglês – Português de Estatística.* 2ª Ed. Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística.

Pestana, D. e Velosa, S. (2010), Introdução à Probabilidade e Estatística (Volume I, 4ª ed), Fundação Calouste Gulbenkian, 52-88, 105-123

Plug & Score, http://www.plug-n-score.com/learning/gini-and-roc-curve.htm [10/05/2015]

Rocha, C. (1995) *Modelos de Sobrevivência.* Working Paper Nº 40, ISSN: 0872 - 895X, Depósito Legal Nº 90631/95, ISEGI – UNL

Rousseeuw, P., Ruts I. e Tukey, John W. (1999): *The Bagplot: A Bivariate Boxplot, The American Statistician*, 53:4, pp. 382-387

Sarmento Baptista, António Manuel (2012) *Credit Scoring, Uma ferramenta de gestão financeira.* Vida Económica

Schutte, J. M., Violette, D. M. (1994), *The Treatment of Outliers and Influential Observations in Regression-Based Impact Evaluation,* American Council for an Energy Efficient Economy, ACEEE SUMMER STUDY ON ENERGY EFFICIENCY IN BUILDINGS; 8; 8.171-8.176; Vol 8

Shen, YJ., e Huang, SG., (2006) *Improve Survival Prediction Using Principal Components of Gene Expression Data*, Volume 4, Issue 2, pp. 110-119, em http://www.sciencedirect.com/science/article/pii/S1672022906600223

Snipes M., Taylor D. C.,(2014) *Model selection and Akaike Information Criteria: An example from wine ratings and prices*, em www.sciencedirect.com, Wine Economics and Policy 3 (2014) 3–9

Stepanova, M. and Thomas, L. C. (2001) *PHAB Scores: Proportional Hazards Analysis Behavioural Scores.* The Journal of the Operational Research Society, Vol. 52, No. 9, Special Issue: Credit Scoring and Data Mining (Sep., 2001), pp. 1007-1016

Stepanova, M. (2001) *Using survival analysis methods to build credit scoring models.* University of Southampton, British Thesis Service, Suplied by The British Library

https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect040.htm, Suporte SAS [18/10/2015]

Thomas, L. C., Ho, J., Scherer, W. T. (2001) *Time will tell: Behavioural scoring and the dynamics of consumer credit assessment*. IMA Journal of Management Mathematics, Volume 12, Issue 1, Pp 89-103., 6-10

Thomas, L., Stepanova, M. (1999) *Survival analysis methods for personal loan data.* Working Paper Series Nº 99/4.Credit Research Centre, The University of Edinburgh, 102-135

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 685-691, 714

ANEXOS

ANEXO I ESTATÍSTICAS DESCRITIVAS

Estatísticas descritivas das covariáveis contínuas.

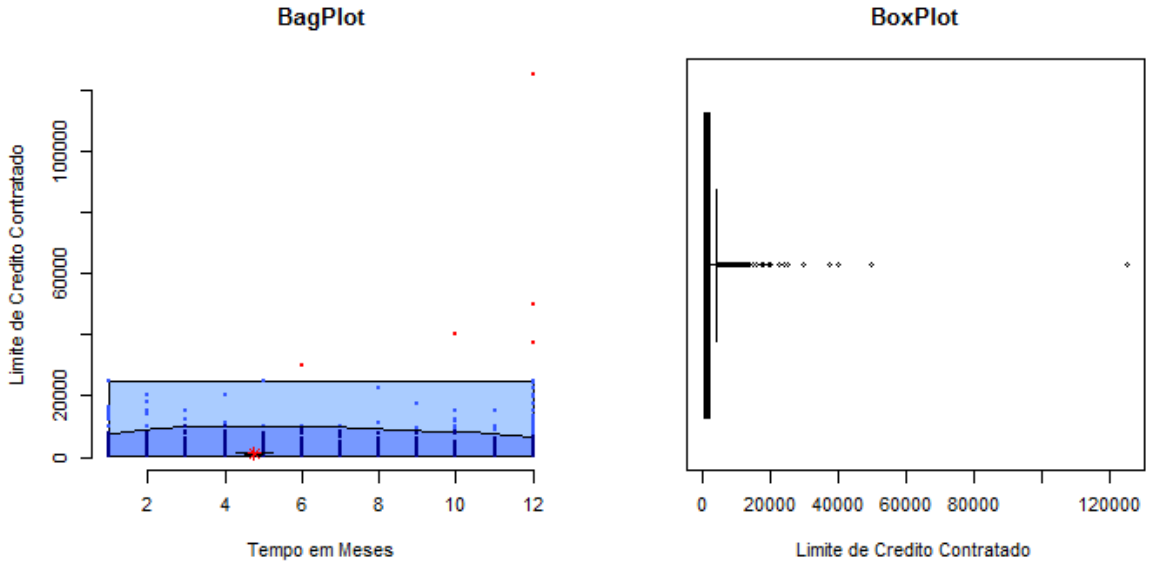
Estatística	x7	x11	x12	x13
N	4358	4358	4358	4358
Missing	0	0	0	0
Mínimo	500	1	4	1,07
Máximo	37500	7	87	827931,84
Amplitude Min-Max	37000	6	83	827930,77
Soma	7819579	5809	178062	32798729,51
Mediana	1000	1	38	938,69
Média	1794,304	1,333	40,859	7526,097
SE (média)	35,982	0,010	0,182	485,201
I.C. 95% (média)	70,543	0,019	0,357	951,240
Variância	5642366,104	0,418	144,160	1025958507,377
Desvio-padrão	2375,367	0,646	12,007	32030,587
Coeficiente de variação	1,324	0,485	0,294	4,256

Estatística	x14	x15	x20	x21
N	4358	4358	4358	4358
Missing	0	0	0	0
Mínimo	0,850	-258033,400	1,000	0,0051
Máximo	850526,87	550729,79	168,00	3529,41
Amplitude Min-Max	850526,02	808763,19	167,00	3529,41
Soma	31908535,40	890194,11	33830,00	21556,62
Mediana	933,165	1,425	5	1,001357705
Média	7321,830	204,267	7,763	4,946
SE (média)	456,074	187,309	0,189	1,063
I.C. 95% (média)	894,136	367,221	0,370	2,085
Variância	906478002,549	152899318,372	155,201	4928,822
Desvio-padrão	30107,773	12365,246	12,458	70,206
Coeficiente de variação	4,112	60,535	1,605	14,193

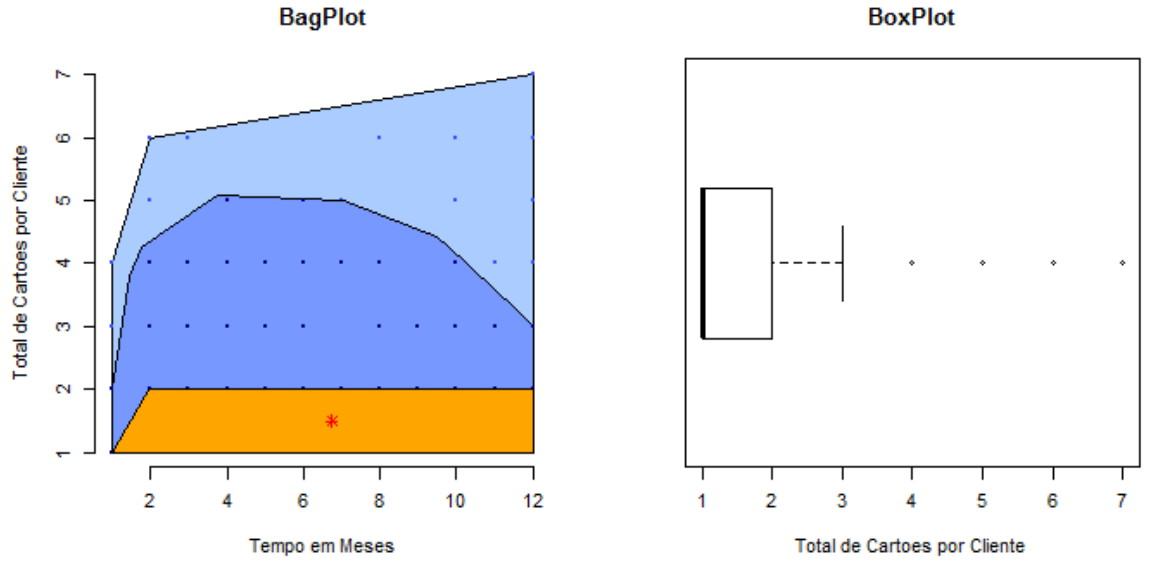
ANEXO II DIAGRAMAS DE EXTREMOS E QUARTIS E *BAGPLOT*

Diagramas de extremos e quartis e *bagplot*, para cada uma das covariáveis, antes de remoção dos valores atípicos severos e respetivas transformações. Aplicável a covariáveis contínuas.

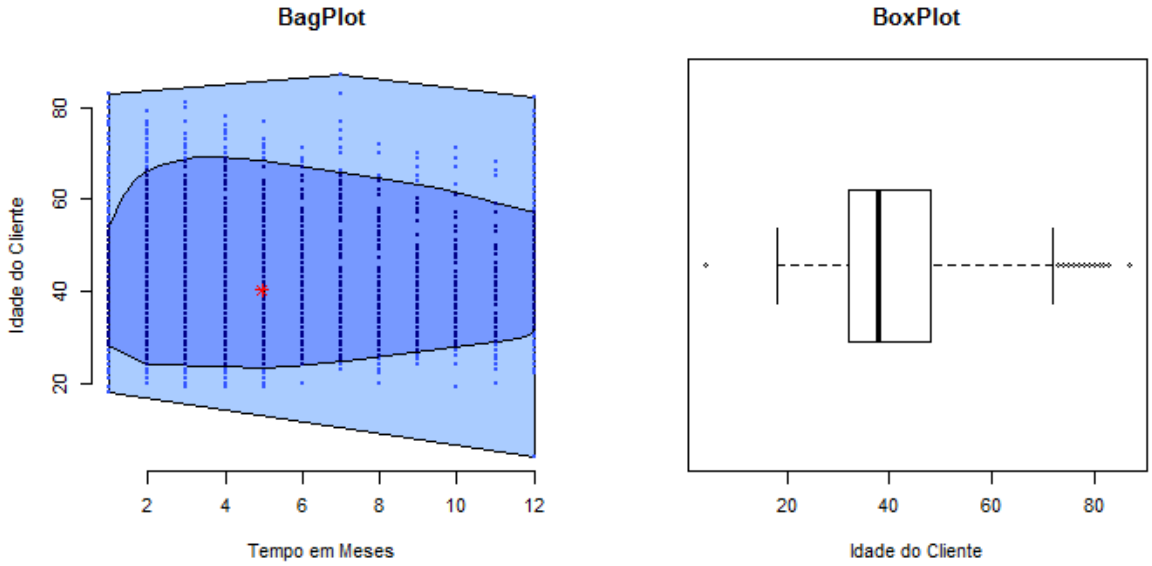
X7 – Limite de Crédito Contratado



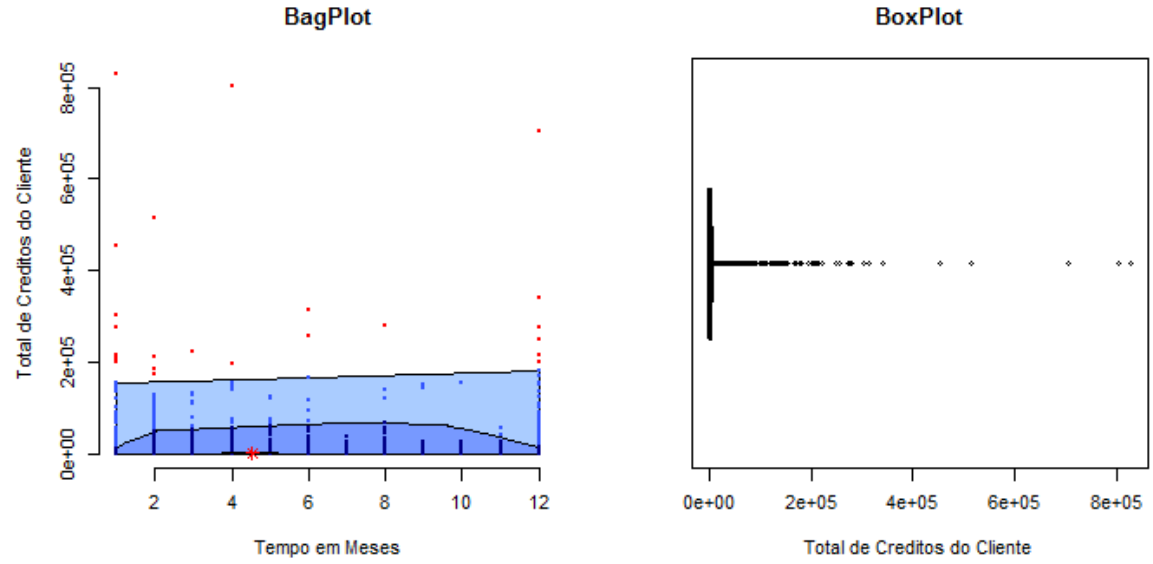
X11 – Total de Cartões por Cliente



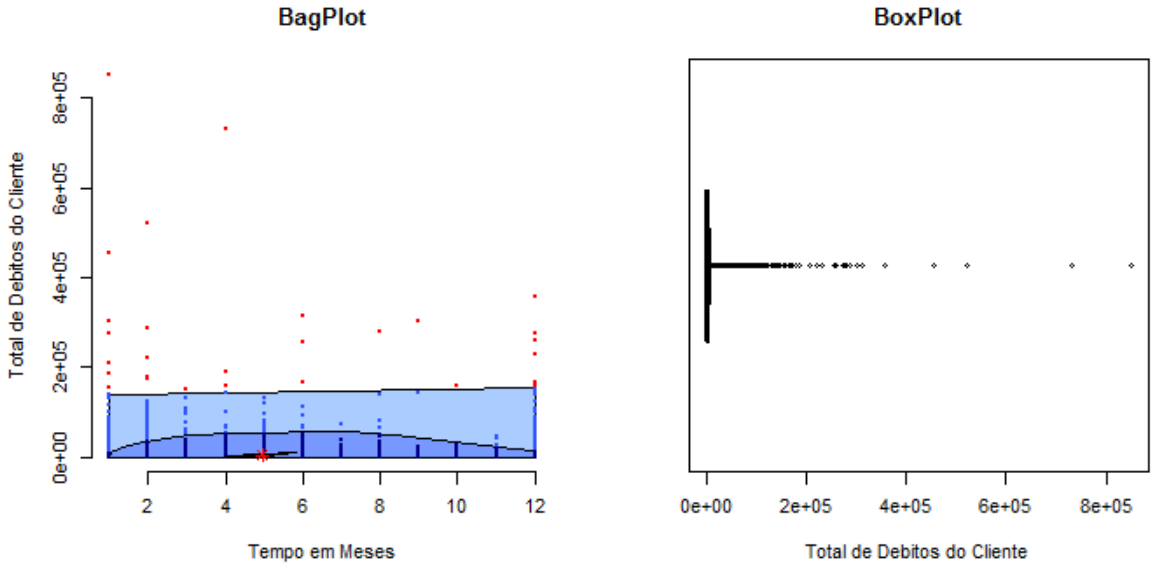
X12 – Idade do Cliente



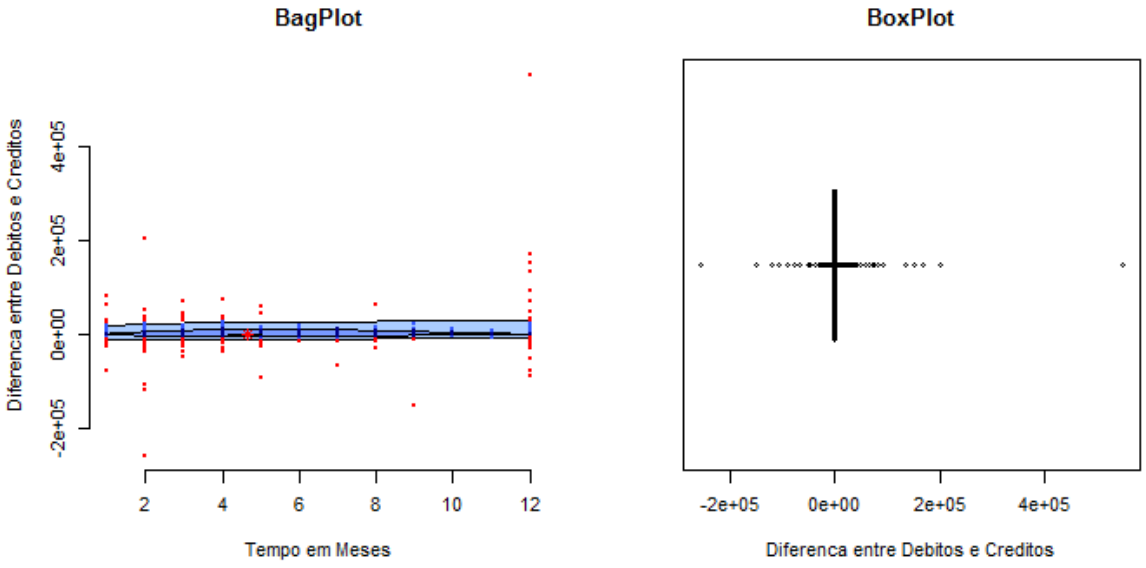
X13 –Total de Créditos do Cliente



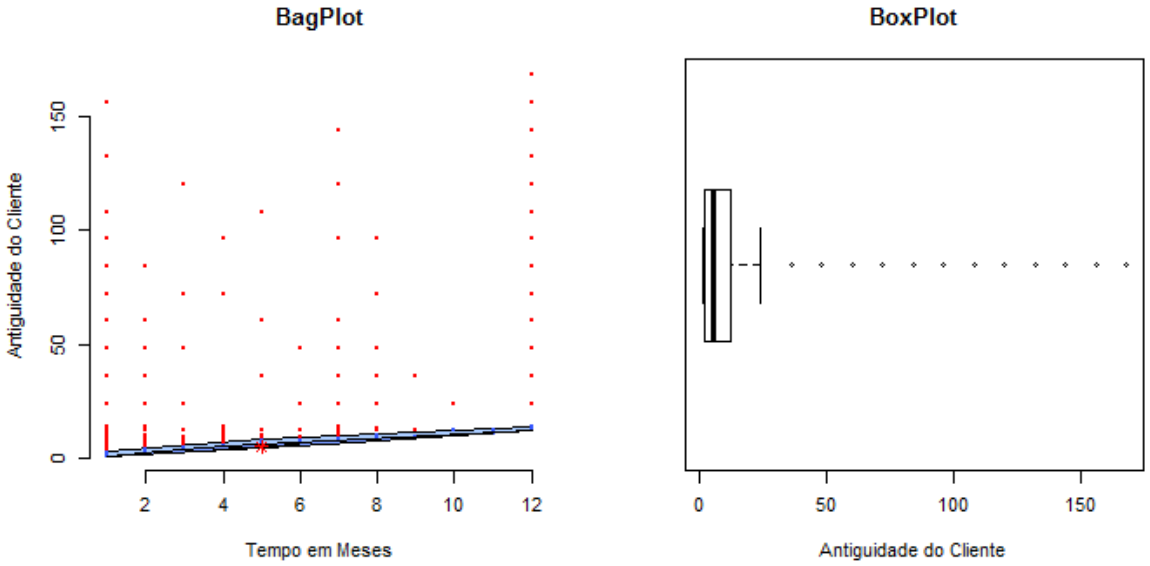
X14 – Total de Débitos do Cliente



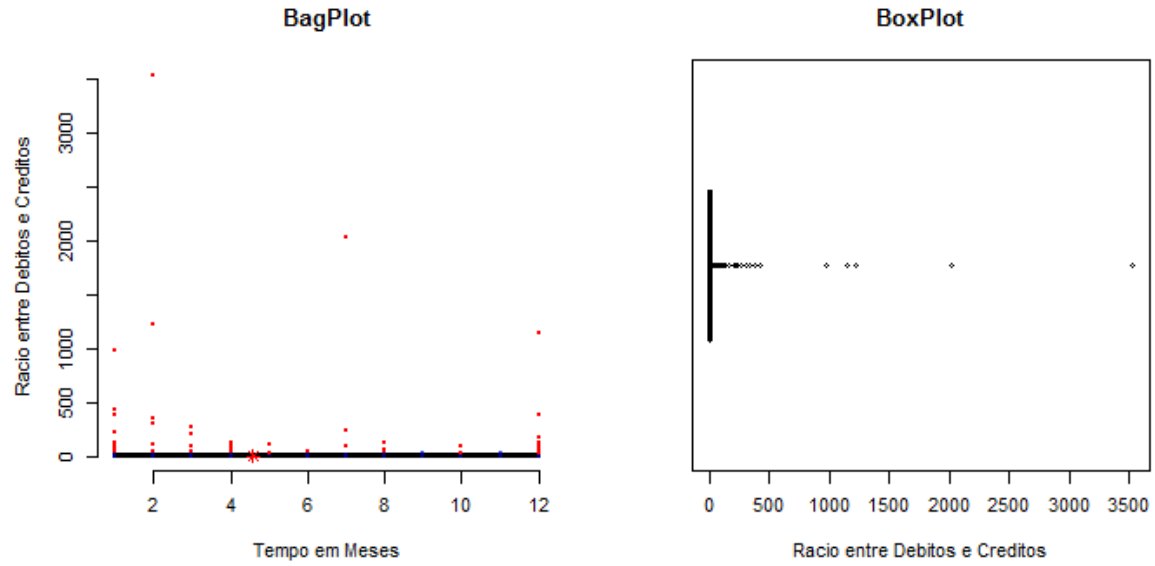
X15 – Diferença entre Débitos e Créditos



X20 – Antiguidade do Cliente



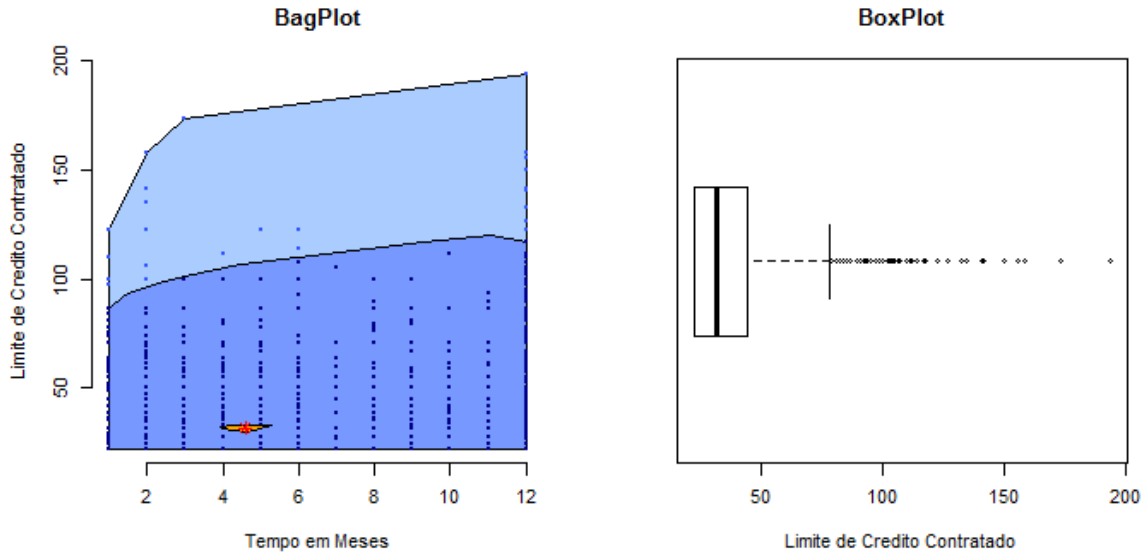
X21 – Rácio entre Débitos e Créditos



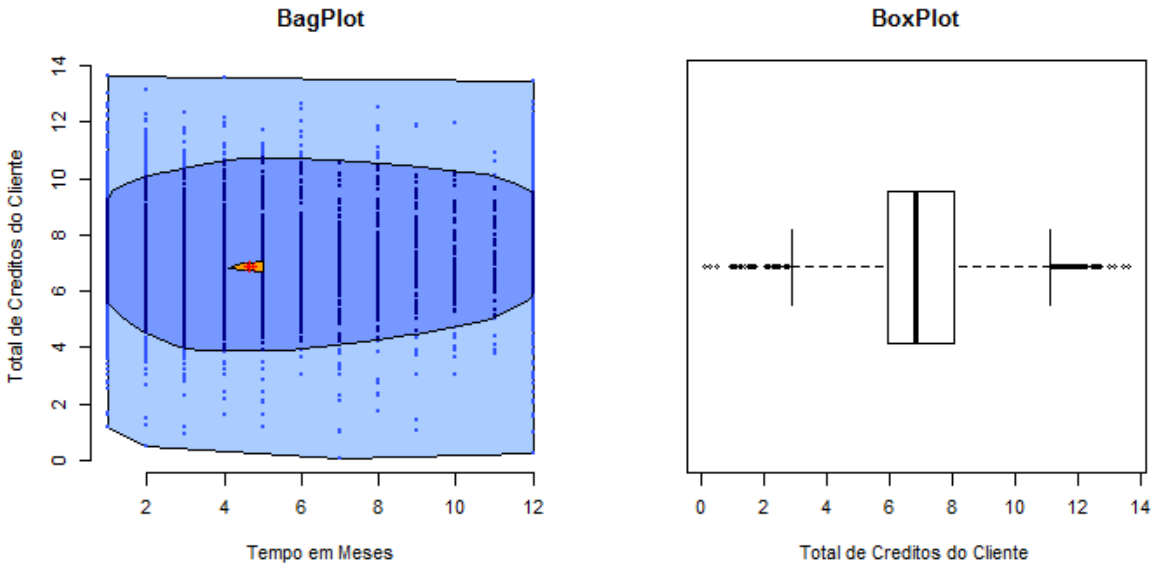
ANEXO III DIAGRAMAS DE EXTREMOS E QUARTIS E BAGPLOT DAS COVARIÁVEIS CONTÍNUAS APÓS REMOÇÃO DE VALORES ATÍPICOS SEVEROS E TRANSFORMAÇÃO

Diagramas de extremos e quartis e *bagplot* para cada uma das covariáveis, após remoção dos valores atípicos severos e respetivas transformações.

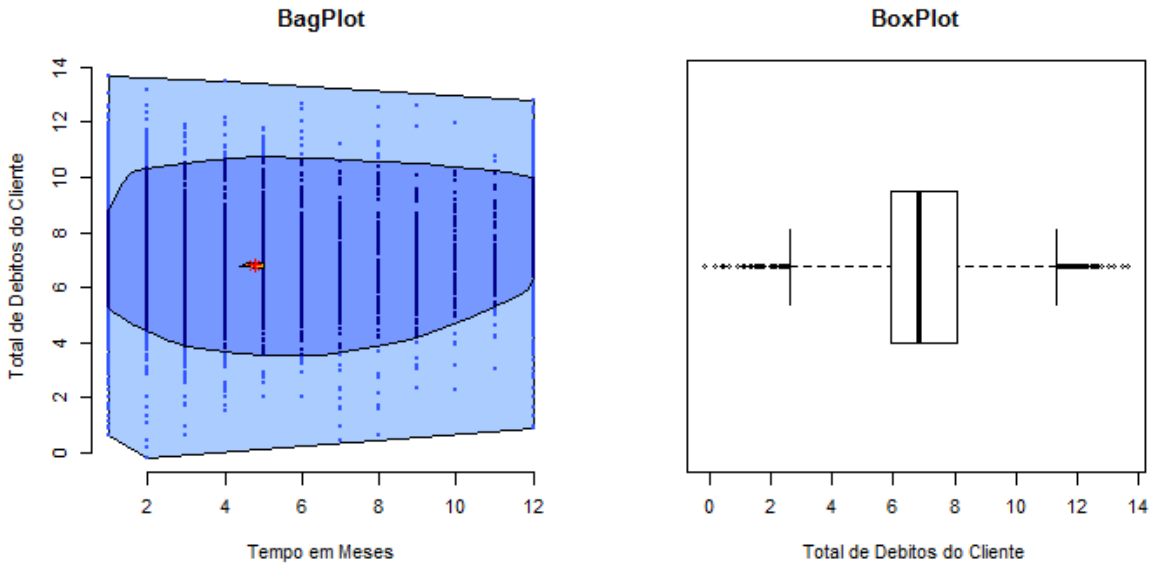
X7 – Limite de Crédito Contratado após transformação



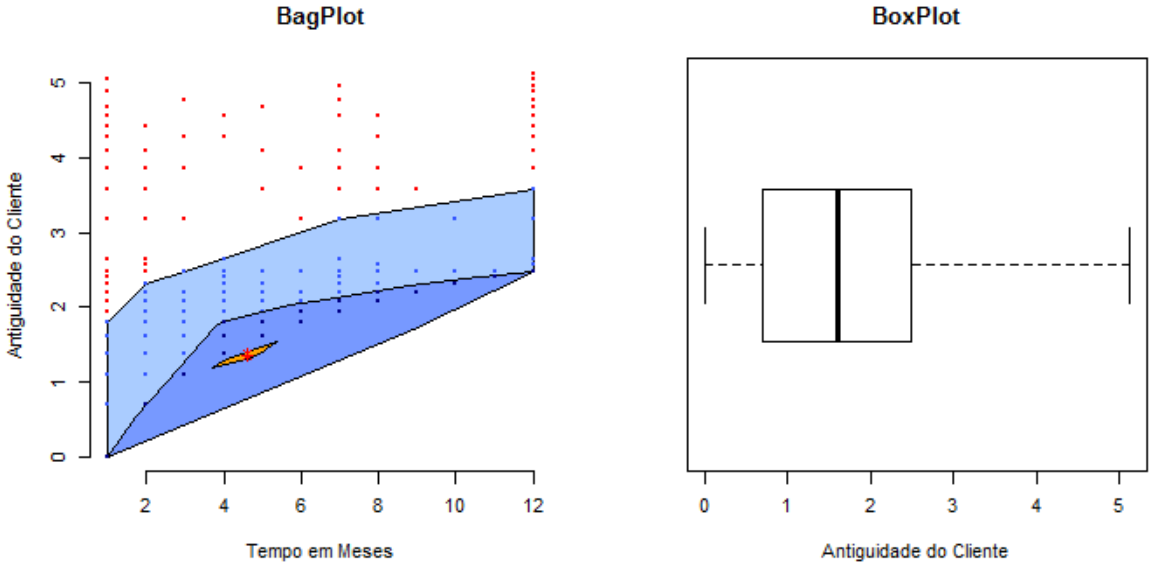
X13 – Total de Créditos do Cliente após transformação



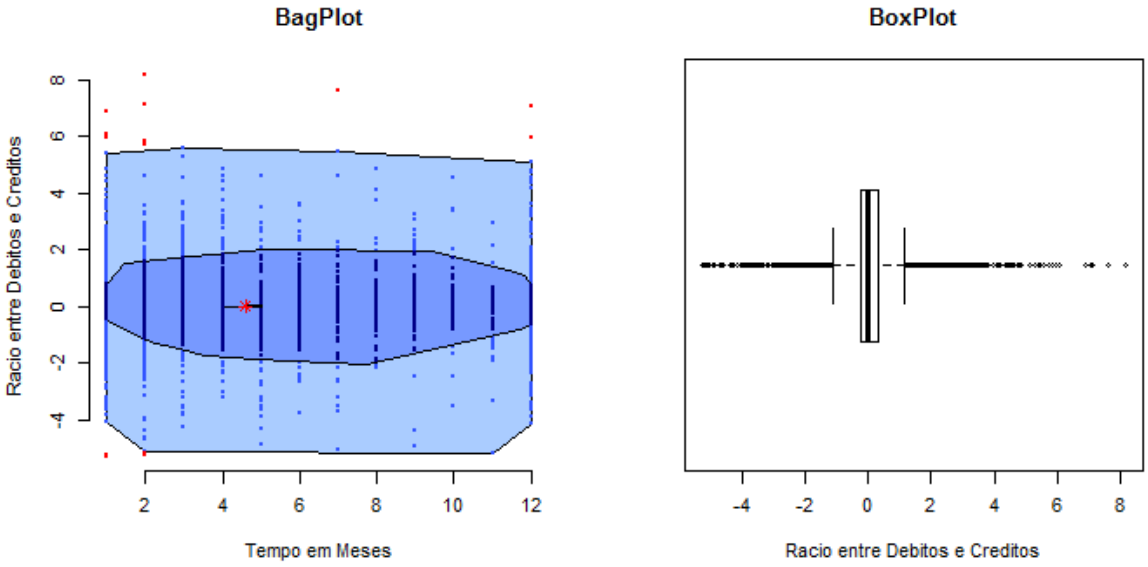
X14 – Total de Débitos do Cliente após transformação



X20 – Antiguidade do Cliente após transformação



X21 – Rácio entre Débitos e Créditos após transformação



ANEXO IV ESTATÍSTICAS WEIGHT OF EVIDENCE (WoE), INFORMATION VALUE (IV) E Z-SCORE DAS VARIÁVEIS CATEGÓRICAS

Análise desenvolvida em covariáveis categóricas.

X1 – Estado Civil

ID	x1	Contratos	WoE	IV	Z-score
0	SOLTEIRO(A)	2 104	-0,135	0,0093	-2,7639
1	CASADO(A)	2 254	0,1406	0,0097	2,7639
	Total	4 358	0	0,0189	

X2 – Indicador de Existência de Produtos de Poupança

ID	x2	Contratos	WoE	IV	Z-score
0	Não	3 710	-0,1506	0,0205	-7,6831
1	Sim	648	1,8088	0,2461	7,6831
	Total	4 358	0	0,2666	

X3 – Sexo

ID	x3	Contratos	WoE	IV	Z-score
F	Feminino	1 444	-0,0433	0,0006	-0,6234
M	Masculino	2 914	0,022	0,0003	0,6234
	Total	4 358	0	0,001	

X5 – Indicador de Trabalhador por Conta de Outrém

ID	x5	Contratos	WoE	IV	Z-score
1	Sim	4 031	-0,0233	0,0005	-1,6563
0	Não	327	0,33	0,0072	1,6563
	Total	4 358	0	0,0077	

X6 – Código de Natureza Jurídica

ID	x6	Contratos	WoE	IV	Z-score
1	SINGULAR - NACIONAL	3 116	0,1242	0,0105	3,8325
2	SINGULAR-EMIGRANTE	19	0,7264	0,0017	0,7268
3	SINGULAR - ESTRANGEIRO	113	0,0631	0,0001	0,2093
4	EMPRES. EM NOME INDIVIDUAL	10	0,0332	0	0,0339
13	SOCIEDADE POR QUOTAS	2	n.a.	n.a	0,48
99	NÃO APLICAVEL	1 098	-0,3126	0,0278	-4,0847
	Total	4 358	0	0,0402	

X8 – Indicador de Cartão de Crédito Normal

ID	x8	Contratos	WoE	IV	Z-score
0	Não	741	0,6504	0,0556	4,5737
1	Sim	3 617	-0,0991	0,0085	-4,5737
	Total	4 358	0	0,0641	

X9 – Indicador de Cartão Gold

ID	x9	Contratos	WoE	IV	Z-score
0	Não	3 648	-0,1114	0,0108	-5,2995
1	Sim	710	0,8283	0,0807	5,2995
	Total	4 358	0	0,0915	

X10 – Indicador de Cartão de Uso Exclusivo em Combustível

ID	x10	Contratos	WoE	IV	Z-score
0	Não	1 864	-0,0485	0,001	-0,8579
1	Sim	2 494	0,0375	0,0008	0,8579
	Total	4 358	0	0,0018	

X17 – Indicador de Cliente com Crédito Particular (consumo)

ID	x17	Contratos	WoE	IV	Z-score
0	Não	3 390	0,25	0,044	8,5084
1	Sim	968	-0,6223	0,1096	-8,5084
	Total	4 358	0	0,1536	

X18 – Indicador de Cliente com Crédito Hipotecário

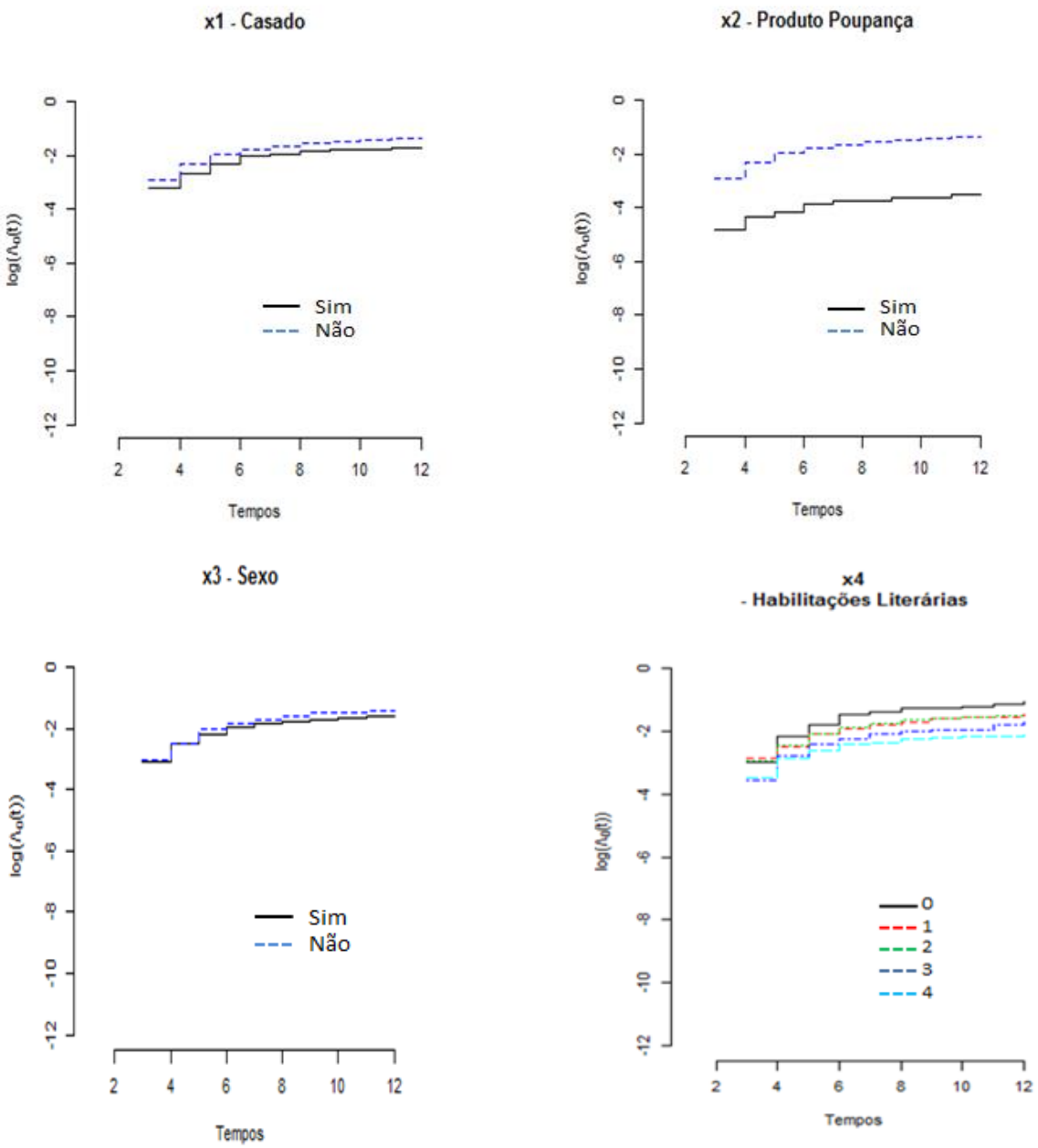
ID	x18	Contratos	WoE	IV	Z-score
0	Não	3 216	-0,0652	0,0032	-2,2551
1	Sim	1 142	0,2043	0,0101	2,2551
	Total	4 358	0	0,0133	

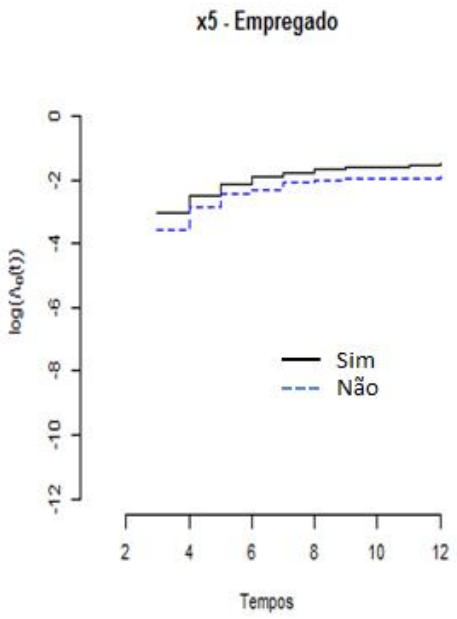
X19 – Indicador de Cliente com Ordenado Domiciliado

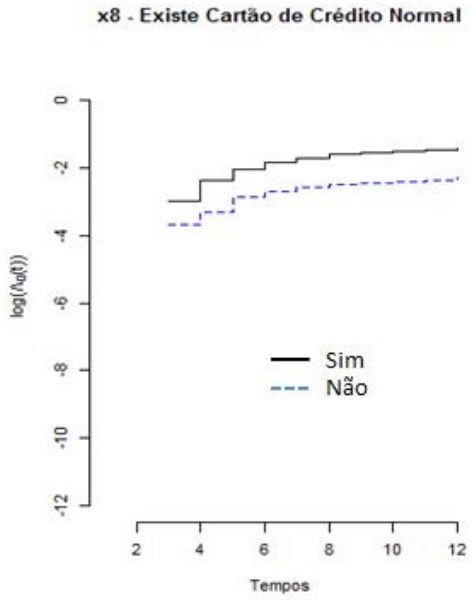
ID	x19	Contratos	WoE	IV	Z-score
0	Não	3 182	0,1536	0,0162	4,774
1	Sim	1 176	-0,3417	0,036	-4,774
	Total	4 358	0	0,0522	

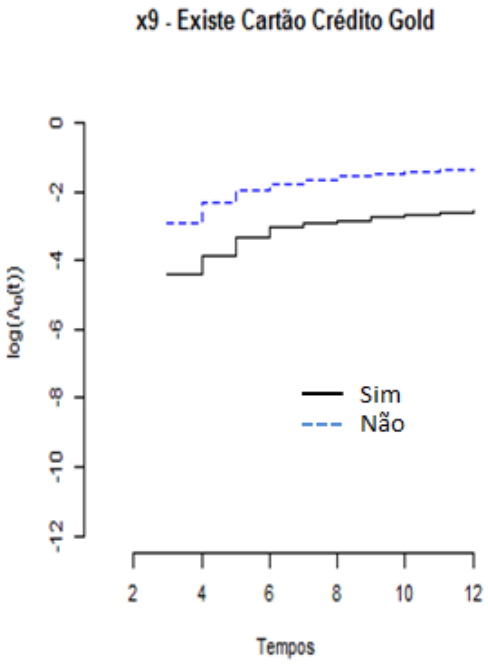
ANEXO V LOGARITMO DA FUNÇÃO TAXA DE FALHA ACUMULADA EM CADA GRUPO NO TEMPO

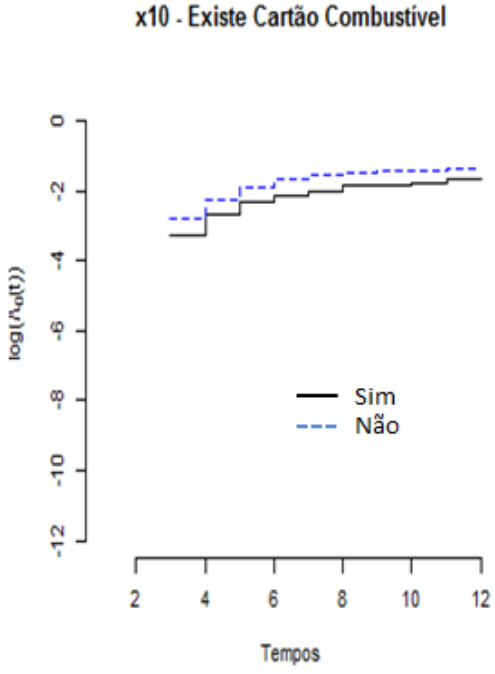
Para ser considerado proporcional, o mesmo, deverá apresentar linhas paralelas. A existência de linhas não paralelas significa que o efeito em estudo não é constante ao longo do tempo. Análise desenvolvida para variáveis categóricas.



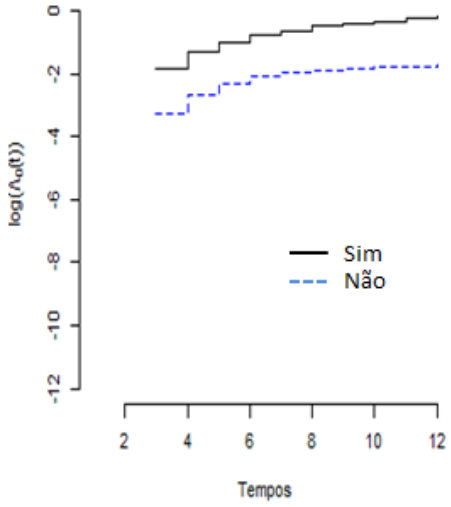

x5 - Empregado
log(Λ0(t))
0
-2
-4
-6
-8
-10
-12
2
4
6
8
10
12
Tempos
Sim
Não


x8 - Existe Cartão de Crédito Normal
log(Λ0(t))
0
-2
-4
-6
-8
-10
-12
2
4
6
8
10
12
Tempos
Sim
Não

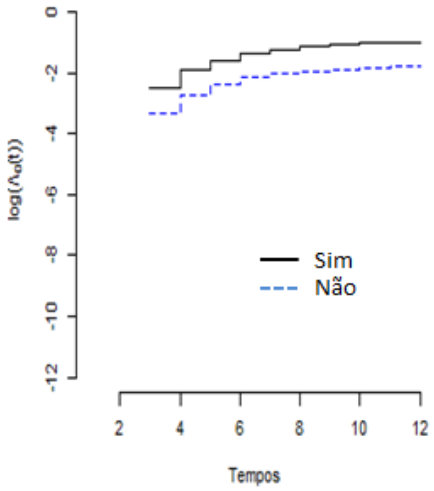

x9 - Existe Cartão Crédito Gold
log(Λ0(t))
0
-2
-4
-6
-8
-10
-12
2
4
6
8
10
12
Tempos
Sim
Não


x10 - Existe Cartão Combustível
log(Λ0(t))
0
-2
-4
-6
-8
-10
-12
2
4
6
8
10
12
Tempos
Sim
Não

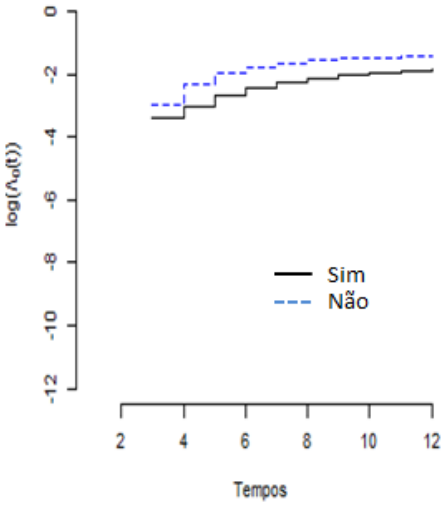
x16 - Inibição de Uso de Cheque



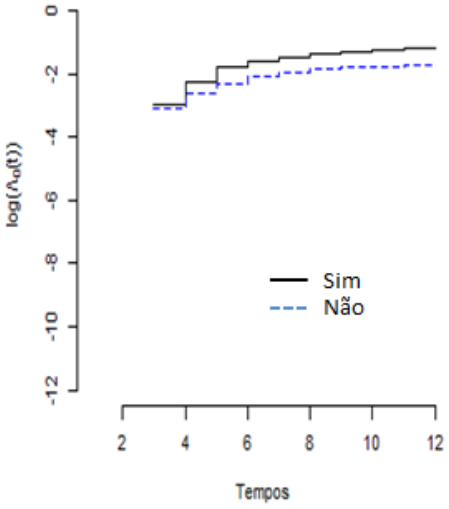
x17 - Existe Crédito Particular



x18 - Existe Crédito Hipotecário



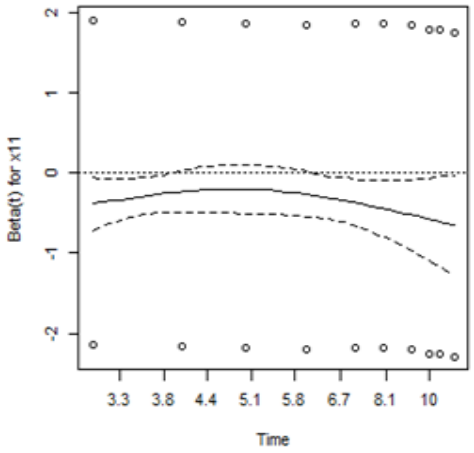
x19 - Ordenado Domiciliado

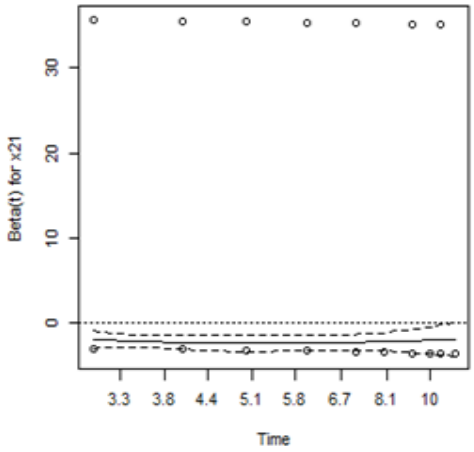


ANEXO VI RESÍDUOS PADRONIZADOS DE SCHOENFELD

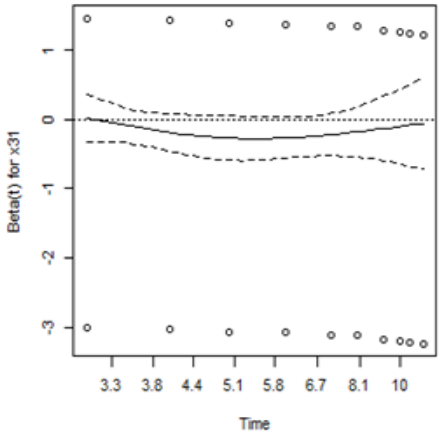
Gráficos dos resíduos padronizados de Schoenfeld, em que se representam os resíduos, com a curva suavizada com bandas de confiança.

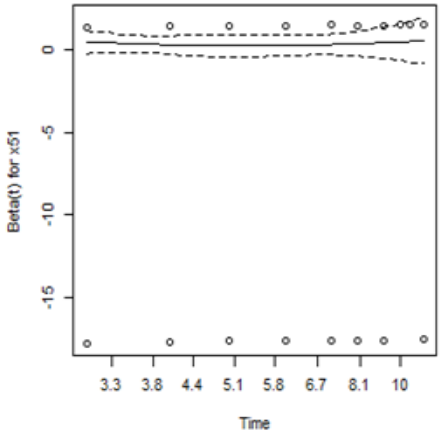
X1 - Estado Civil e X2 - Indicador de Existência de Produtos de Poupança



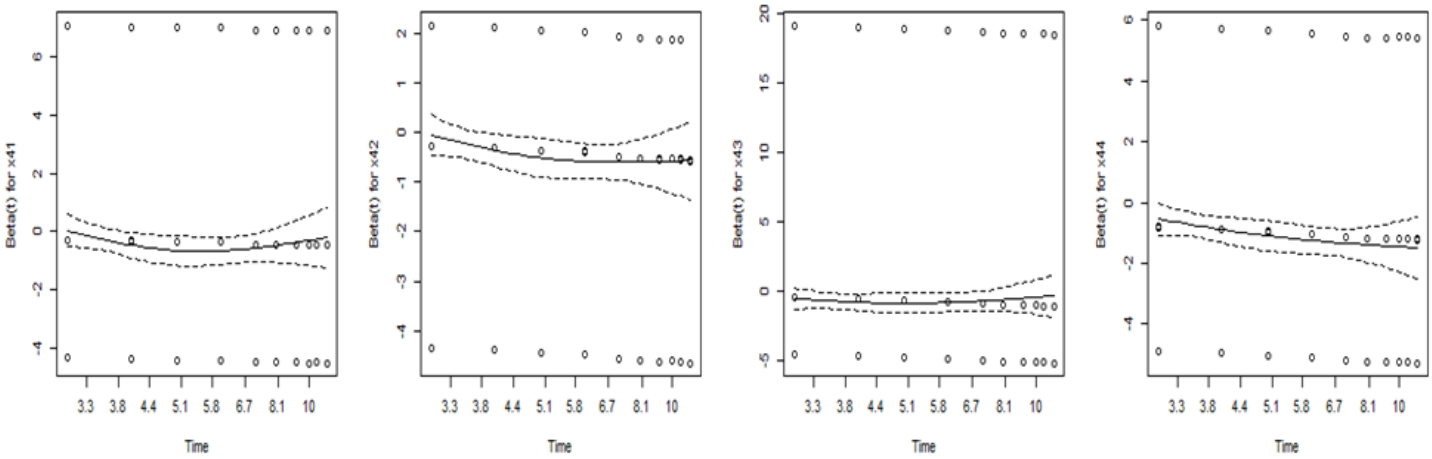


X3 – Sexo e X5 - Indicador de Trabalhador por Conta de Outrem

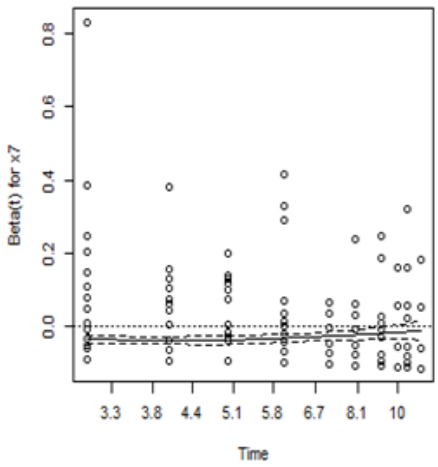


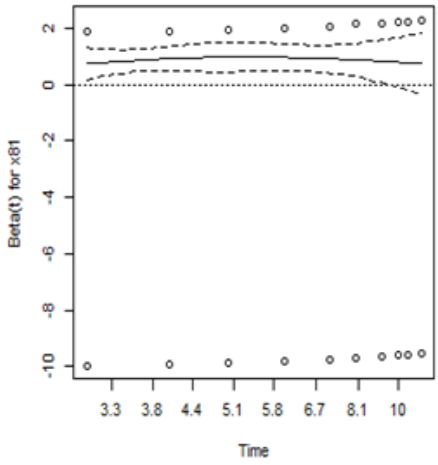


X4 - Habilitações Literárias

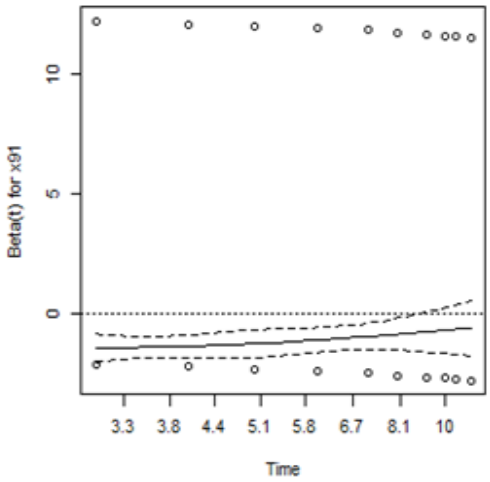


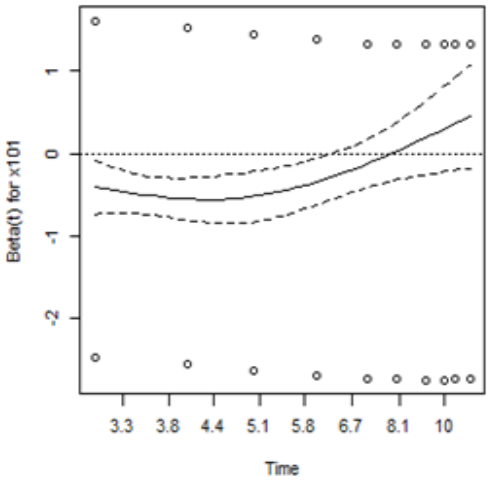
X7 - Limite de Crédito Contratado e X8 - Indicador de Cartão de Crédito Normal



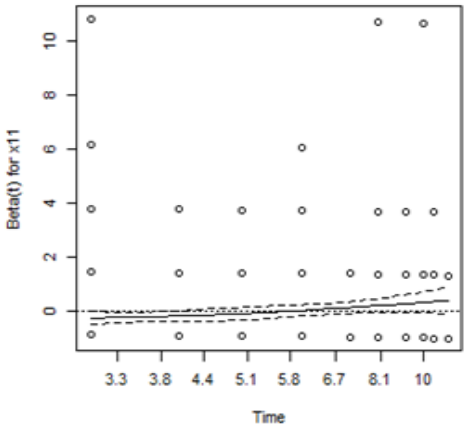


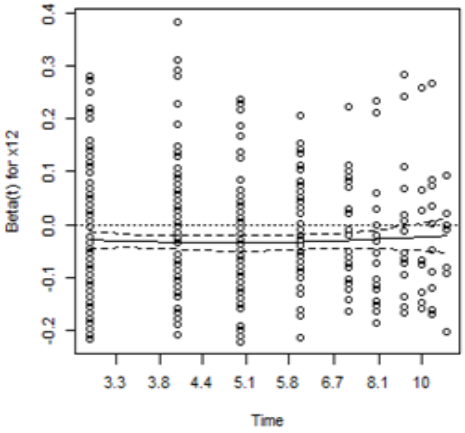
X9 - Indicador de Cartão Gold e X10 - Indicador de Cartão de Uso Exclusivo em Combustível



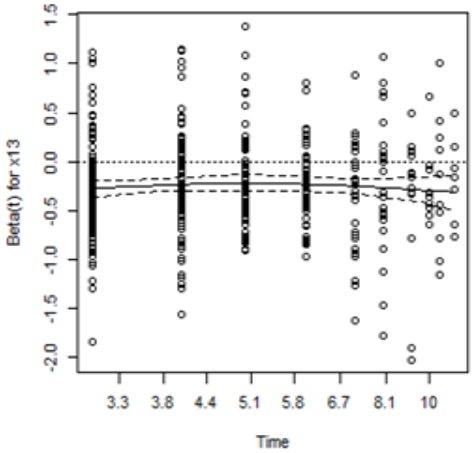


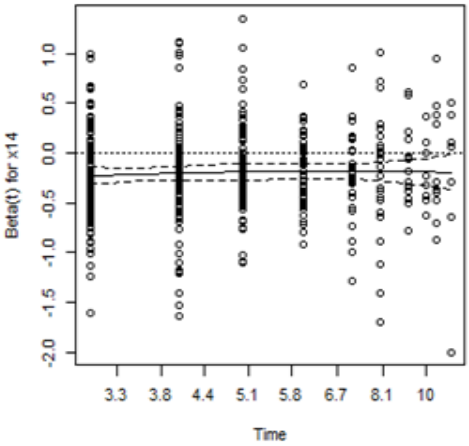
X11 - Numero de Cartões por Cliente e X12 - Idade



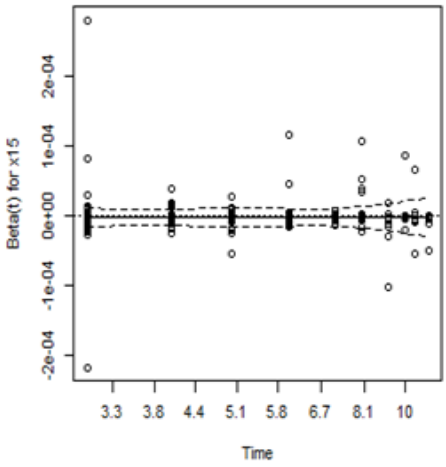


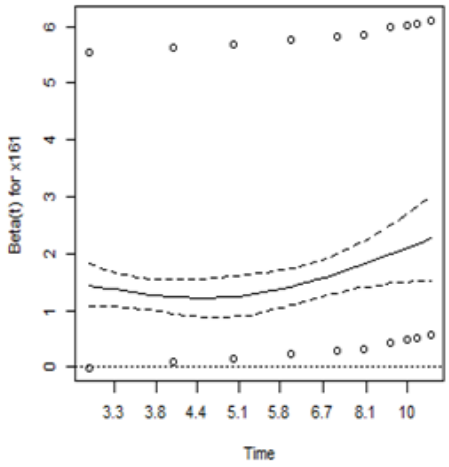
X13 - Total Mensal de Créditos em Conta e X14 - Total Mensal de Débitos em Conta



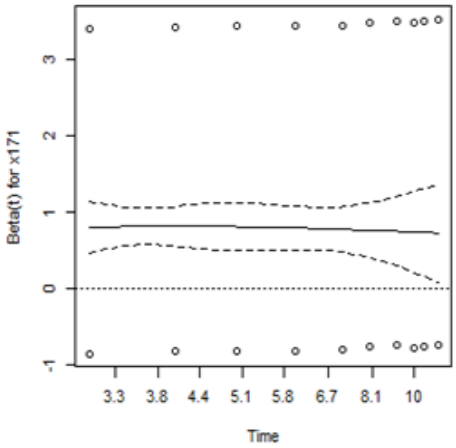


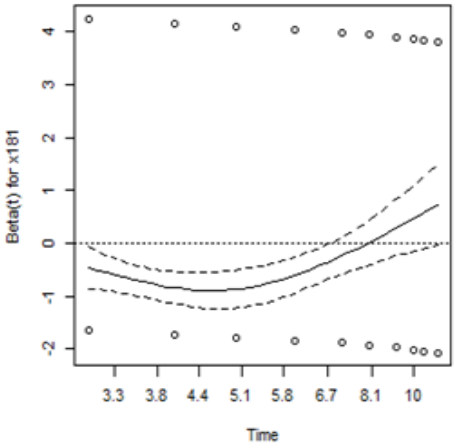
X15 - Diferença entre Débitos e Créditos e X16 - Indicador de Inibição de Uso de Cheque



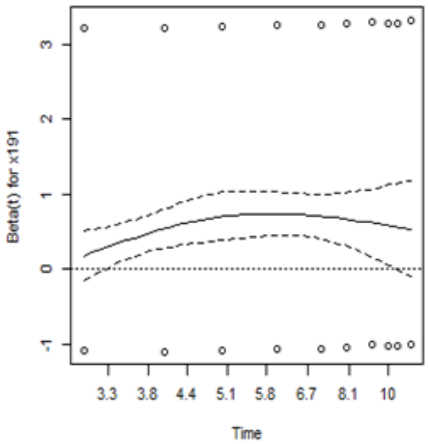


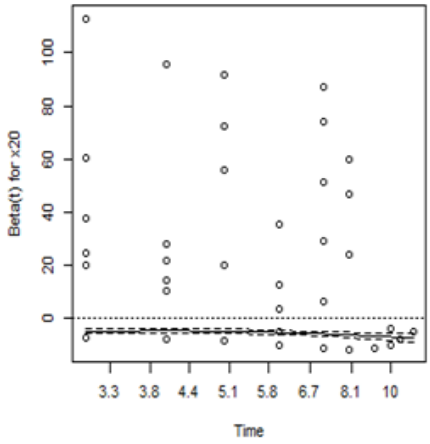
X17 - Cliente com Crédito Particular (consumo) e X18 - Cliente com Crédito Hipotecário



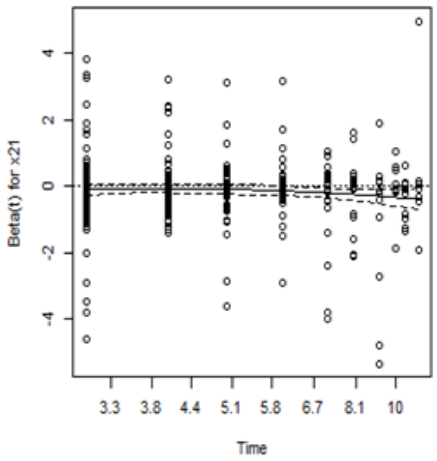


X19 - Cliente com Ordenado Domiciliado e X20 - Antiguidade do Cliente em Meses





X21 - Rácio entre Débitos e Créditos



ANEXO VII RESULTADOS DO AJUSTAMENTO DA REGRESSÃO DE COX PARA OS MODELOS 1 E 3

Resultados da regressão do Modelo de Cox – Modelo 1

Covariável	Coeficiente	Erro-Padrão	Valor-*p*
x2	-0,8314	0,2987	0,0054
x̃4=2	-0,1547	0,1383	0,2633
x̃4=3	-0,4584	0,1862	0,0138
x9	-0,3328	0,2334	0,1539
x12	-0,0138	0,0056	0,0126
x13	-0,0975	0,0351	0,0056
x16	1,0522	0,1341	$4{,}2 \times 10^{-15}$
x17	0,4874	0,1194	$4{,}5 \times 10^{-5}$
x19	0,3271	0,1195	0,0062
x20	-4,5932	0,2984	$< 2 \times 10^{-16}$

Resultados da regressão do Modelo de Cox – Modelo 3

Covariável	Coeficiente	Erro-Padrão	Valor-*p*
x2	-0,8683	0,2984	0,00361
x12	-0,0153	0,0055	0,0055
x13	-0,1101	0,0342	0,0013
x16	1,0937	0,1334	$2{,}2 \times 10^{-16}$
x17	0,5158	0,1187	$1{,}4 \times 10^{-5}$
x19	0,3402	0,1193	0,0044
x20	-4,6353	0,2981	$< 2 \times 10^{-16}$

ANEXO VIII COMPUTAÇÃO ESTATÍSTICA EM R

Apresenta-se o módulo de computação estatística de suporte aos resultados apurados neste estudo.

```
#
# Bibliotecas
#
library(pastecs)
library(aplpack)
library(survival)
library(xtable)
library(knitr)
library(ggplot2)
library(survMisc)
library(MASS)
library(risksetROC)
library(survivalROC)
library(corrplot)
library(scatterplot3d)
#
# ######################################################
#  Gerador do relatorio em batch
#  Console > stitch("C:/RWKS/Rfinal/rScriptCovars.R")
# ######################################################
# Exemplo Retirado de "http://stats.stackexchange.com/questions/30496/plotting-interval-censored-follow-up-time-as-a-line-chart"
# Adaptado com referência a "http://docs.ggplot2.org/current/"
#
dat <- structure(list(ID = 1:5, eventA = c(0L, 1L, 1L, 0L, 1L),
                      eventB = c(1L, 0L, 0L, 1L, 0L),
                      t1 = c(4, 2, 6, 5, 5),
                      t2 = c(1, 1, 6, 4.5, 1),
                      censored = c(0, 0, 0, 0, 1)),
                      .Names = c("ID", "eventA","eventB", "t1", "t2", "censored"),
                      class = "data.frame",
                      row.names = c(NA, -5L))

# Criar a variável Evento
dat$event <- with(dat, ifelse(eventA, "A", "B"))
dat$id.ordered <- factor(x = dat$ID, levels = order(dat$t2, decreasing = T))
#
ggplot(dat, aes(x = id.ordered)) +
  geom_linerange(aes(ymin = t1-t2, ymax = t1+t2)) +
  geom_point(aes(y = ifelse(censored, t1 + t2, t1 + t2), shape = event),
          size = 4) +
  coord_flip() +
  scale_shape_manual(name = "Eventos", values = c(19, 15)) +
```

```
 labs(list(title = "Dados com censura aleat?ria", x = "Clientes", y = "Meses")) +
 theme_bw()
#
#
#
dados <- read.csv2("C:/sample.csv", header=T)
vcont <- read.csv2("C:/sample_cont.csv", header=T)
#
# AMOSTRA
#
cens<-dados[,3]
tempos<-dados[,4]
#
x1<-as.factor(dados[,5])
x2<-as.factor(dados[,6])
x3<-as.factor(dados[,7])
x4<-as.factor(dados[,8])
x5<-as.factor(dados[,9])
x6<-as.factor(dados[,10])
x7<-sqrt(as.numeric(dados[,11]))
x8<-as.factor(dados[,12])
x9<-as.factor(dados[,13])
x10<-as.factor(dados[,14])
x11<-as.numeric(dados[,15])
x12<-as.numeric(dados[,16])
x13<-log(as.numeric(dados[,17]))
x14<-log(as.numeric(dados[,18]))
x15<-as.numeric(dados[,19])
x16<-as.factor(dados[,20])
x17<-as.factor(dados[,21])
x18<-as.factor(dados[,22])
x19<-as.factor(dados[,23])
x20<-log(as.numeric(dados[,24]))
x21<-log(as.numeric(dados[,25]))
#
# Transformar x4 em 3 categorias
#
x4t<-0
#
for (i in 1:length(x4))
{
 if (x4[i]==3)
 {
   x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==2)
```

```
 {
  x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==1)
 {
  x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==4)
 {
  x4t[i]<-3
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==0)
 {
  x4t[i]<-1
 }
}
#
x4t<-as.factor(x4t)
#
#
#
c1<-sqrt(as.numeric(vcont[,1]))
c2<-as.numeric(vcont[,2])
c3<-as.numeric(vcont[,3])
c4<-log(as.numeric(vcont[,4]))
c5<-log(as.numeric(vcont[,5]))
c6<-as.numeric(vcont[,6])
c7<-log(as.numeric(vcont[,7]))
c8<-log(as.numeric(vcont[,8]))
#
# K-M
#
ekm0<-survfit(Surv(tempos,cens)~1, conf.type="plain")
ekm0.dados<-summary(ekm0)
print(ekm0.dados)
#
autoplot(ekm0, xLab = "Tempo (em meses)", yLab = "S(t) Estimada",
```

```
      title = "Estimador de Kaplan-Meier para Clientes com Cartão de Crédito", titleSize = 10, axisTitleSize = 10,
axisLabSize = 10, survLineSize = 0.5,
      type = "CI", palette = "Dark2", jitter = "none", censShape = 3, censSize = 5, legend = TRUE,
      legLabs = c("Censurado"), legTitle = NULL, legTitleSize = 9,legLabSize = 10, alpha = 0.8,
      CIline = 10, fillLineSize = 0.05,pVal = FALSE, sigP = 1, pX = 0.1, pY = 0.1, timeTicks = "major",
      tabTitle = "Estimador de Kaplan-Meier", tabTitleSize = 15, tabLabSize = 5, nRiskSize = 5)
#
#
#
options(scipen=100)
options(digits=2)
#
DescStat<-stat.desc(dados)
write.table(DescStat, "C:/Users/iNet/Desktop/TESE AS UAb/01 R/RWKS/Rout/descStats.csv", sep=";")
#
# Análise de outliers
#
#
# Limite Credito Contratado
#
par(mfrow=c(2,4))
bagplot(tempos, x7, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Limite de Credito Contratado",
main="BagPlot")
boxplot(x7, horizontal=TRUE, main="BoxPlot", xlab="Limite de Credito Contratado")
#
# Numero de cartões por cliente
#
par(mfrow=c(2,4))
bagplot(tempos, x11, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Total de Cartoes por Cliente",
main="BagPlot")
boxplot(x11, horizontal=TRUE, main="BoxPlot", xlab="Total de Cartoes por Cliente")
#
# Idade
#
par(mfrow=c(2,4))
bagplot(tempos, x12, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Idade do Cliente",
main="BagPlot")
boxplot(x12, horizontal=TRUE, main="BoxPlot", xlab="Idade do Cliente")
#
# Total de Créditos
#
par(mfrow=c(2,4))
bagplot(tempos, x13, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Total de Creditos do Cliente",
main="BagPlot")
boxplot(x13, horizontal=TRUE, main="BoxPlot", xlab="Total de Creditos do Cliente")
#
# Total de Débitos
#
par(mfrow=c(2,4))
```

```
bagplot(tempos, x14, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Total de Debitos do Cliente",
main="BagPlot")
boxplot(x14, horizontal=TRUE, main="BoxPlot", xlab="Total de Debitos do Cliente")
#
# Diferenca entre debitos e creditos
#
par(mfrow=c(2,4))
bagplot(temposo, x15, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Diferenca entre Debitos e
Creditos", main="BagPlot")
boxplot(x15, horizontal=TRUE, main="BoxPlot", xlab="Diferenca entre Debitos e Creditos")
#
# Antiguidade do cliente em meses
#
par(mfrow=c(2,4))
bagplot(tempos, x20, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Antiguidade do Cliente",
main="BagPlot")
boxplot(x20, horizontal=TRUE, main="BoxPlot", xlab="Antiguidade do Cliente")
#
# Rácios entre débitos e créditos
#
par(mfrow=c(2,4))
bagplot(tempos, x21, show.whiskers=FALSE, xlab="Tempo em Meses", ylab="Racio entre Debitos e Creditos",
main="BagPlot")
boxplot(x21, horizontal=TRUE, main="BoxPlot", xlab="Racio entre Debitos e Creditos")
#
# Avaliar a correlação entre as COVARS
#
par(mfrow=c(2,4))
corrData<-cor(vcont, use="complete.obs", method="spearman")
corrplot.mixed(corrData)
corrplot(corrData, type = "lower")
write.table(corrData, "C:/Users/iNet/Desktop/TESE AS UAb/01 R/RWKS/Rout/corrDados15112015.csv", sep=";")
#
# Estimação Cox para cada uma das covariáveis individualmente
#
auxCPH<-0
destino<-0
z<-0
nomevarcoxph=c("fit0x1", "fit0x2", "fit0x3", "fit0x4", "fit0x5", "fit0x6", "fit0x7", "fit0x8", "fit0x9", "fit0x10",
"fit0x11", "fit0x12", "fit0x13", "fit0x14", "fit0x15", "fit0x16", "fit0x17", "fit0x18", "fit0x19", "fit0x20")
for (z in 1:20)
{
  auxCPH[z]<-paste("resFitx", z, sep="")
  eval(substitute(variable<-coxph(Surv(tempos,cens)~(dados[,z+4]),x=T,method="breslow"),
list(variable=as.name(nomevarcoxph[z]))))
  s1<-"summary (fit0x"
  s2<-")"
  print(eval(parse(text=paste(s1, z, s2,sep=""))))
}
#
```

```
# Análise de Proporcionalidade das Covariáveis em estudo
#
# x1
#
fit<-coxph(Surv(tempos[x1==1],cens[x1==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x1 -
Casado",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x1==0],cens[x1==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "N?o"),
    col=c("black","blue"),lty=1:4, cex=0.4,box.lty=0)
#
# x2
#
fit<-coxph(Surv(tempos[x2==1],cens[x2==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x2 -
Produto Poupan?a",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x2==0],cens[x2==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")

legend("bottomright", legend=c("Sim", "N?o"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x3
#
fit<-coxph(Surv(tempos[x3==1],cens[x3==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x3 -
Sexo",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x3==0],cens[x3==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
```

```
lines(ssx$time,log(h0),type="s",lty=2, col="blue")

legend("bottomright", legend=c("M", "F"),
      col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
#  x4
#
fit<-coxph(Surv(tempos[x4==1],cens[x4==1])~1, x=T, method="breslow")
ssx41<-survfit(fit)
s0<-round(ssx41$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx41$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x4 -
Habilita??es Liter?rias", bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x4==2],cens[x4==2])~1, x=T, method="breslow")
ssx42<-survfit(fit)
s0<-round(ssx42$surv, digits=5)
h0<- -log(s0)
lines(ssx42$time,log(h0),type="s",lty=2, col="blue")
fit<-coxph(Surv(tempos[x4==3],cens[x4==3])~1, x=T, method="breslow")
ssx43<-survfit(fit)
s0<-round(ssx43$surv, digits=5)
h0<- -log(s0)
lines(ssx43$time,log(h0),type="s",lty=3, col="red")
fit<-coxph(Surv(tempos[x4==4],cens[x4==4])~1, x=T, method="breslow")
ssx44<-survfit(fit)
s0<-round(ssx44$surv, digits=5)
h0<- -log(s0)
lines(ssx44$time,log(h0),type="s",lty=4, col="green")

legend("bottomright", legend=c("Line 1", "Line 2", "Line 3", "Line 4"),
      col=c("black","blue", "red","green"), lty=1:4, cex=0.3,box.lty=0)
#
#
#
par(mfrow=c(2,4))
fit<-coxph(Surv(tempos[x4t==1],cens[x4t==1])~1, x=T, method="breslow")
ssx41t<-survfit(fit)
s0<-round(ssx41t$surv, digits=5)
h0<- -log(s0)
plot(ssx41t$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x4t
    - Habilitações Literárias", bty="n",type="s",lty=1)
#
fit<-coxph(Surv(tempos[x4t==2],cens[x4t==2])~1, x=T, method="breslow")
ssx42t<-survfit(fit)
s0<-round(ssx42t$surv, digits=5)
h0<- -log(s0)
lines(ssx42t$time,log(h0),type="s",lty=2, col=2)
#
fit<-coxph(Surv(tempos[x4==3],cens[x4==3])~1, x=T, method="breslow")
```

```
ssx43t<-survfit(fit)
s0<-round(ssx43t$surv, digits=5)
h0<- -log(s0)
lines(ssx43t$time,log(h0),type="s",lty=3, col=3)
#
# x5
#
fit<-coxph(Surv(tempos[x5==1],cens[x5==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x5 -
Empregado",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x5==0],cens[x5==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "N?o"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x8
#
fit<-coxph(Surv(tempos[x8==1],cens[x8==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x8 -
Existe Cartão de Crédito Normal",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x8==0],cens[x8==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "Não"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x9
#
fit<-coxph(Surv(tempos[x9==1],cens[x9==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x9 -
Existe Cartão de Crédito Gold",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x9==0],cens[x9==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
```

```
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "Não"),
     col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x10
#
fit<-coxph(Surv(tempos[x10==1],cens[x10==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x10 -
Existe Cart?o Combust?vel",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x10==0],cens[x10==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")

legend("bottomright", legend=c("Sim", "N?o"),
     col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x16
#
fit<-coxph(Surv(tempos[x16==1],cens[x16==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x16 -
Inibição de Uso de Cheque",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x16==0],cens[x16==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "Não"),
     col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x17
#
fit<-coxph(Surv(tempos[x17==1],cens[x17==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x17 -
Existe Crédito Particular",bty="n",type="s",lty=1)
```

```
fit<-coxph(Surv(tempos[x17==0],cens[x17==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")

legend("bottomright", legend=c("Sim", "Não"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x18
#
fit<-coxph(Surv(tempos[x18==1],cens[x18==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x18 -
Existe Crédito Hipotecário",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x18==0],cens[x18==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "Não"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# x19
#
fit<-coxph(Surv(tempos[x19==1],cens[x19==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x19 -
Ordenado Domiciliado",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x19==0],cens[x19==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
legend("bottomright", legend=c("Sim", "Não"),
    col=c("black","blue"),lty=1:4, cex=0.3,box.lty=0)
#
# Analise aos Resíduos de Schoenfeld
#
fit1x1<-coxph(Surv(tempos,cens)~(x1),x=T,method="breslow")
aux1<-resid(fit1x1,type="scaledsch")
cox.zph(fit1x1,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x1), df=3)
```

```
abline(h=0, lty=3)
#
fit1x2<-coxph(Surv(tempos,cens)~(x2),x=T,method="breslow")
aux2<-resid(fit1x2,type="scaledsch")
cox.zph(fit1x2,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x2), df=3)
abline(h=0, lty=3)
#
fit1x3<-coxph(Surv(tempos,cens)~(x3),x=T,method="breslow")
aux3<-resid(fit1x3,type="scaledsch")
cox.zph(fit1x3,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x3), df=3)
abline(h=0, lty=3)
#
fit1x4<-coxph(Surv(tempos,cens)~(x4),x=T,method="breslow")
aux4<-resid(fit1x4,type="scaledsch")
cox.zph(fit1x4,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x4), df=3)
#abline(h=0, lty=3)
#
# x4t
#
fit1x4t<-coxph(Surv(tempos,cens)~(x4t),x=T,method="breslow")
aux4t<-resid(fit1x4t,type="scaledsch")
cox.zph(fit1x4t,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x4t), df=3)
#
fit1x5<-coxph(Surv(tempos,cens)~(x5),x=T,method="breslow")
aux5<-resid(fit1x5,type="scaledsch")
cox.zph(fit1x5,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x5), df=3)
abline(h=0, lty=3)
#
fit1x6<-coxph(Surv(tempos,cens)~(x6),x=T,method="breslow")
aux6<-resid(fit1x6,type="scaledsch")
cox.zph(fit1x6,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x6), df=3)
#
fit1x7<-coxph(Surv(tempos,cens)~(x7),x=T,method="breslow")
aux7<-resid(fit1x7,type="scaledsch")
cox.zph(fit1x7,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x7), df=3)
abline(h=0, lty=3)
```

```
#
fit1x8<-coxph(Surv(tempos,cens)~(x8),x=T,method="breslow")
aux8<-resid(fit1x8,type="scaledsch")
cox.zph(fit1x8,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x8), df=3)
abline(h=0, lty=3)
#
fit1x9<-coxph(Surv(tempos,cens)~(x9),x=T,method="breslow")
aux9<-resid(fit1x9,type="scaledsch")
cox.zph(fit1x9,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x9), df=3)
abline(h=0, lty=3)
#
fit1x10<-coxph(Surv(tempos,cens)~(x10),x=T,method="breslow")
aux10<-resid(fit1x10,type="scaledsch")
cox.zph(fit1x10,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x10), df=3)
abline(h=0, lty=3)
#
fit1x11<-coxph(Surv(tempos,cens)~(x11),x=T,method="breslow")
aux11<-resid(fit1x11,type="scaledsch")
cox.zph(fit1x11,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x11), df=3)
abline(h=0, lty=3)
#
fit1x12<-coxph(Surv(tempos,cens)~(x12),x=T,method="breslow")
aux12<-resid(fit1x12,type="scaledsch")
cox.zph(fit1x12,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x12), df=3)
abline(h=0, lty=3)
#
fit1x13<-coxph(Surv(tempos,cens)~(x13),x=T,method="breslow")
aux13<-resid(fit1x13,type="scaledsch")
cox.zph(fit1x13,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x13), df=3)
abline(h=0, lty=3)
#
fit1x14<-coxph(Surv(tempos,cens)~(x14),x=T,method="breslow")
aux14<-resid(fit1x14,type="scaledsch")
cox.zph(fit1x14,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x14), df=3)
abline(h=0, lty=3)
#
```

```
fit1x15<-coxph(Surv(tempos,cens)~(x15),x=T,method="breslow")
aux15<-resid(fit1x15,type="scaledsch")
cox.zph(fit1x15,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x15), df=3)
abline(h=0, lty=3)
#
fit1x16<-coxph(Surv(tempos,cens)~(x16),x=T,method="breslow")
aux16<-resid(fit1x16,type="scaledsch")
cox.zph(fit1x16,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x16), df=3)
abline(h=0, lty=3)
#
fit1x17<-coxph(Surv(tempos,cens)~(x17),x=T,method="breslow")
aux17<-resid(fit1x17,type="scaledsch")
cox.zph(fit1x17,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x17), df=3)
abline(h=0, lty=3)
#
fit1x18<-coxph(Surv(tempos,cens)~(x18),x=T,method="breslow")
aux18<-resid(fit1x18,type="scaledsch")
cox.zph(fit1x18,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x18), df=3)
abline(h=0, lty=3)
#
fit1x19<-coxph(Surv(tempos,cens)~(x19),x=T,method="breslow")
aux19<-resid(fit1x19,type="scaledsch")
cox.zph(fit1x19,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x19), df=3)
abline(h=0, lty=3)
#
fit1x20<-coxph(Surv(tempos,cens)~(x20),x=T,method="breslow")
aux20<-resid(fit1x20,type="scaledsch")
cox.zph(fit1x20,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x20),  df=3)
abline(h=0, lty=3)
#
fit1x21<-coxph(Surv(tempos,cens)~(x21),x=T,method="breslow")
aux21<-resid(fit1x21,type="scaledsch")
cox.zph(fit1x21,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x21), df=3)
abline(h=0, lty=3)
#
# Aalens Linear Hazards Model - analise grafica
```

```
#
aafitx21 <- aareg(Surv(tempos, cens) ~ x21)
par(mfrow=c(2,4))
plot(aafitx21)
#
# Transformação x4
#
fit0x4<-coxph(Surv(tempos,cens)~(x4),x=T,method="breslow")
resFit0x4<-summary(fit0x4)
print(resFit0x4)
#
fit<-coxph(Surv(tempos[x4==1],cens[x4==1])~1, x=T, method="breslow")
ssx41<-survfit(fit)
s0<-round(ssx41$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx41$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x4 -
Habilitações Literárias", bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x4==2],cens[x4==2])~1, x=T, method="breslow")
ssx42<-survfit(fit)
s0<-round(ssx42$surv, digits=5)
h0<- -log(s0)
lines(ssx42$time,log(h0),type="s",lty=2, col="blue")
fit<-coxph(Surv(tempos[x4==3],cens[x4==3])~1, x=T, method="breslow")
ssx43<-survfit(fit)
s0<-round(ssx43$surv, digits=5)
h0<- -log(s0)
lines(ssx43$time,log(h0),type="s",lty=3, col="red")
fit<-coxph(Surv(tempos[x4==4],cens[x4==4])~1, x=T, method="breslow")
ssx44<-survfit(fit)
s0<-round(ssx44$surv, digits=5)
h0<- -log(s0)
lines(ssx44$time,log(h0),type="s",lty=4, col="green")
#
# Teste log-rank
#
survdiff(Surv(tempos, cens) ~ x4, rho=0)
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==0 | x4==1))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==0 | x4==2))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==0 | x4==3))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==0 | x4==4))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==1 | x4==2))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==1 | x4==3))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==1 | x4==4))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==2 | x4==3))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==2 | x4==4))
survdiff(Surv(tempos, cens) ~ x4, rho=0, subset=(x4==3 | x4==4))
x4t<-0
#
for (i in 1:length(x4))
```

```
{
 if (x4[i]==3)
 {
  x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==2)
 {
  x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==1)
 {
  x4t[i]<-2
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==4)
 {
  x4t[i]<-3
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==0)
 {
  x4t[i]<-1
 }
}
#
#
#
survdiff(Surv(tempos, cens) ~ x4t, rho=0)
#
survdiff(Surv(tempos, cens) ~ x4t, rho=0, subset=(x4t==1 | x4t==2))
#
survdiff(Surv(tempos, cens) ~ x4t, rho=0, subset=(x4t==1 | x4t==3))
#
survdiff(Surv(tempos, cens) ~ x4t, rho=0, subset=(x4t==2 | x4t==3))
#
#
```

```
ekmx4t<-survfit(Surv(tempos,cens)~x4t, conf.type="plain")
ekmx4t.dados<-summary(ekmx4t)
print(ekmx4t.dados)
#
autoplot(ekmx4t, xLab = "Tempo (em meses)", yLab = "S(t) Estimada",
      title = "Estimador de Kaplan-Meier para x4t", titleSize = 10, axisTitleSize = 10, axisLabSize = 10, survLineSize
= 0.5,
      type = "single", palette = "Dark2", jitter = "all", censShape = 3, censSize = 5, legend = TRUE,
      legLabs = NULL, legTitle = "Categoria", legTitleSize = 9,legLabSize = 10, alpha = 0.8,
      CIline = 10, fillLineSize = 0.05,pVal = FALSE, sigP = 1, pX = 0.1, pY = 0.1, timeTicks = "major",
      tabTitle = "Estimador de Kaplan-Meier", tabTitleSize = 15, tabLabSize = 5, nRiskSize = 5)
#
fit<-coxph(Surv(tempos[x4t==1],cens[x4t==1])~1, x=T, method="breslow")
ssx41t<-survfit(fit)
s0<-round(ssx41t$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx41$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x4t -
Habilitações Literárias", bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x4t==2],cens[x4t==2])~1, x=T, method="breslow")
ssx42t<-survfit(fit)
s0<-round(ssx42t$surv, digits=5)
h0<- -log(s0)
lines(ssx42t$time,log(h0),type="s",lty=2, col="blue")
fit<-coxph(Surv(tempos[x4t==3],cens[x4t==3])~1, x=T, method="breslow")
ssx43t<-survfit(fit)
s0<-round(ssx43t$surv, digits=5)
h0<- -log(s0)
lines(ssx43t$time,log(h0),type="s",lty=3, col="red")
#
#
#
fit1x4t<-coxph(Surv(tempos,cens)~(x4t),x=T,method="breslow")
aux4c<-resid(fit1x4t,type="scaledsch")
cox.zph(fit1x4t,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1x4t), df=3)
#
#
#
fit0x4t<-coxph(Surv(tempos,cens)~(x4t),x=T,method="breslow")
resFit0x4t<-summary(fit0x4t)
print(resFit0x4t)
# ###############################################################################
#  Modelação Cox Riscos Proporcionais
###############################################################################
#
# Bibliotecas
#
library(pastecs)
```

```
library(survival)
library(xtable)
library(knitr)
library(ggplot2)
library(survMisc)
library(MASS)
library(risksetROC)
library(survivalROC)
library(scatterplot3d)
#
# ####################################################
#  Gerador do relatorio em batch
#  Console > stitch("C:/RWKS/RTREINO/rtreinocc.R")
# ####################################################
# Funções desenvolvidas em R
# ####################################################
# Exemplo Retirado de "http://www.gettinggeneticsdone.com/2011/02/split-data-frame-into-testing-and.html"
# Adaptado para gerar amostra de treino e teste
#
splitdf <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(length(index)*0.7))
  trainset <- dataframe[trainindex, ]
  testset <- dataframe[-trainindex, ]
  list(trainset=trainset,testset=testset)
}
#
#
#
dados <- read.csv2("C:/sample.csv", header=T)
#
# Gerar amostra de treino + teste
#
gerAmostra<-splitdf(dados, seed=500)
#
lapply(gerAmostra, nrow)
lapply(gerAmostra, head)
#
treinocc<-gerAmostra$trainset
testecc<-gerAmostra$testset
#
# TREINO
#
cens<-treinocc[,3]
tempos<-treinocc[,4]
#
x1<-as.factor(treinocc[,5])
x2<-as.factor(treinocc[,6])
x3<-as.factor(treinocc[,7])
```

```
x4<-as.factor(treinocc[,8])
x5<-as.factor(treinocc[,9])
x6<-as.factor(treinocc[,10])
x7<-sqrt(as.numeric(treinocc[,11]))
x8<-as.factor(treinocc[,12])
x9<-as.factor(treinocc[,13])
x10<-as.factor(treinocc[,14])
x11<-as.numeric(treinocc[,15])
x12<-as.numeric(treinocc[,16])
x13<-log(as.numeric(treinocc[,17]))
x14<-log(as.numeric(treinocc[,18]))
x15<-as.numeric(treinocc[,19])
x16<-as.factor(treinocc[,20])
x17<-as.factor(treinocc[,21])
x18<-as.factor(treinocc[,22])
x19<-as.factor(treinocc[,23])
x20<-log(as.numeric(treinocc[,24]))
x21<-log(as.numeric(treinocc[,25]))
#
#
#
x4t<-0
#
for (i in 1:length(x4))
{
  if (x4[i]==3)
  {
    x4t[i]<-2
  }
}
#
for (i in 1:length(x4))
{
  if (x4[i]==2)
  {
    x4t[i]<-2
  }
}
#
for (i in 1:length(x4))
{
  if (x4[i]==1)
  {
    x4t[i]<-2
  }
}
#
for (i in 1:length(x4))
{
  if (x4[i]==4)
```

```
 {
  x4t[i]<-3
 }
}
#
for (i in 1:length(x4))
{
 if (x4[i]==0)
 {
  x4t[i]<-1
 }
}
#
x4t<-as.factor(x4t)
#
# TESTE
#
censz<-testecc[,3]
temposz<-testecc[,4]
#
z1<-as.factor(testecc[,5])
z2<-as.factor(testecc[,6])
z3<-as.factor(testecc[,7])
z4<-as.factor(testecc[,8])
z5<-as.factor(testecc[,9])
z6<-as.factor(testecc[,10])
z7<-sqrt(as.numeric(testecc[,11]))
z8<-as.factor(testecc[,12])
z9<-as.factor(testecc[,13])
z10<-as.factor(testecc[,14])
z11<-as.numeric(testecc[,15])
z12<-as.numeric(testecc[,16])
z13<-log(as.numeric(testecc[,17]))
z14<-log(as.numeric(testecc[,18]))
z15<-as.numeric(testecc[,19])
z16<-as.factor(testecc[,20])
z17<-as.factor(testecc[,21])
z18<-as.factor(testecc[,22])
z19<-as.factor(testecc[,23])
z20<-log(as.numeric(testecc[,24]))
z21<-log(as.numeric(testecc[,25]))
#
#
#
z4t<-0
#
for (i in 1:length(z4))
{
 if (z4[i]==3)
 {
```

```
   z4t[i]<-2
  }
}
#
for (i in 1:length(z4))
{
  if (z4[i]==2)
  {
    z4t[i]<-2
  }
}
#
for (i in 1:length(z4))
{
  if (z4[i]==1)
  {
    z4t[i]<-2
  }
}
#
for (i in 1:length(z4))
{
  if (z4[i]==4)
  {
    z4t[i]<-3
  }
}
#
for (i in 1:length(z4))
{
  if (z4[i]==0)
  {
    z4t[i]<-1
  }
}
#
z4t<-as.factor(z4t)
#
# Modelo 0 - Modelo de Regressao de Cox
#
fitxmodelBase<-coxph(Surv(tempos,cens)~(x1 + x2 + x4t + x7 + x8 + x9 + x12 + x13 + x14 + x16 + x17 + x19 +
x20), x=T, method = "breslow")
resFitxmodelBase<-summary(fitxmodelBase)
print(resFitxmodelBase)
loglikBase<-fitxmodelBase$loglik[2]
print(loglikBase)
#
AIC_Base<-(-2*loglikBase)+(2*2)
AIC_Base
AIC(fitxmodelBase)
```

```
#
# Stepwise regression
#
step(fitxmodelBase, direction = "both", trace=FALSE)
#
# Modelo 1 - Modelo final identificado pelo STepwise
#
fitxmodelStep<-coxph(Surv(tempos,cens)~(x2 + x9 + x4t + x12 + x13 + x16 + x17 + x19 + x20), x=T, method =
"breslow")
resFitxmodelStep<-summary(fitxmodelStep)
print(resFitxmodelStep)
loglikStep<-fitxmodelStep$loglik[2]
print(loglikStep)
#
TRV_Step<-2*(loglikBase-loglikStep)
pvalue_Step<-1-pchisq(TRV_Step,4)
TRV_Step
pvalue_Step
#
AIC_Step<-(-2*loglikStep)+(2*2)
AIC_Step
AIC(fitxmodelStep)
#
par(mfrow=c(2,4))
rd<-resid(fitxmodelStep, type="deviance")
rm<-resid(fitxmodelStep, type="martingale")
pl<-fitxmodelStep$linear.predictors
plot(pl,rm,xlab="Linear Predictor", ylab="Resíduos martingala", pch=16)
plot(pl,rd,xlab="Linear Predictor", ylab="Resíduos da desviância", pch=16)
#
# Modelo 2
#
fitxmodel2<-coxph(Surv(tempos,cens)~(x2 + x4t + x12 + x13 + x16 + x17 + x19 + x20), x=T, method = "breslow")
resFitxmodel2<-summary(fitxmodel2)
print(resFitxmodel2)
loglikm2<-fitxmodel2$loglik[2]
print(loglikm2)
#
TRV_m2<-2*(loglikBase-loglikm2)
pvalue_m2<-1-pchisq(TRV_m2,5)
TRV_m2
pvalue_m2
#
AIC_m2<-(-2*loglikm2)+(2*2)
AIC_m2
AIC(fitxmodel2)
#
auxresid<-resid(fitxmodel2,type="scaledsch")
cox.zph(fitxmodel2,transform="identity")
par(mfrow=c(2,4))
```

```
plot(cox.zph(fitxmodel2))
#
fit<-coxph(Surv(tempos[x16==1],cens[x16==1])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
par(mfrow=c(2,4))
plot(ssx$time,log(h0),xlab="Tempos",ylim=range(c(-12,1)),ylab=expression(log(Lambda[0]*(t))),main="x16 -
Inibição de Uso de Cheque",bty="n",type="s",lty=1)
fit<-coxph(Surv(tempos[x16==0],cens[x16==0])~1, x=T, method="breslow")
ssx<-survfit(fit)
s0<-round(ssx$surv, digits=5)
h0<- -log(s0)
lines(ssx$time,log(h0),type="s",lty=2, col="blue")
#
par(mfrow=c(2,4))
rd<-resid(fitxmodel2, type="deviance")
rm<-resid(fitxmodel2, type="martingale")
pl<-fitxmodel2$linear.predictors
plot(pl,rm,xlab="Linear Predictor", ylab="Resíduos martingal", pch=16)
plot(pl,rd,xlab="Preditor Linear", ylab="Resíduos da Desviância", pch=1)
#
plot(pl,rm, xlab="Preditor Linear", ylab= "Resíduos Martingala")
lines(lowess(pl,rm), lwd=3, col=2, lty = "dotted")
#
par(mfrow=c(2,4))
scatterplot3d(pl,rm, angle=10,x.ticklabs=" ", y.ticklabs=" ", xlab="x", ylab="y", zlab="z", scale.y=.7, pch = 1,
main="3D - Resíduos Martingala", grid=TRUE, box=FALSE)
scatterplot3d(pl,rd,angle=10, x.ticklabs=" ", y.ticklabs=" ", xlab="x", ylab="y", zlab="z", scale.y=.7, pch = 1,
main="3D - Resíduos da Desviância", grid=TRUE, box=FALSE)
#
par(mfrow=c(2,4))
dfbetas<-resid(fitxmodel2,type="dfbeta")
plot(dfbetas[,1],xlab="x2",ylab="DFBETAS para x2")
plot(dfbetas[,2],xlab="x4",ylab="DFBETAS para x4t")
plot(dfbetas[,3],xlab="x12",ylab="DFBETAS para x12")
plot(dfbetas[,4],xlab="x13",ylab="DFBETAS para x13")
plot(dfbetas[,5],xlab="x16",ylab="DFBETAS para x16")
plot(dfbetas[,6],xlab="x17",ylab="DFBETAS para x17")
plot(dfbetas[,7],xlab="x19",ylab="DFBETAS para x19")
plot(dfbetas[,8],xlab="x20",ylab="DFBETAS para x20")
#
#
#
par(mfrow=c(2,3))
eta <- fitxmodel2$linear.predictor
ROCfitxmodelFinal_12=risksetROC(Stime=tempos, status=cens, marker=eta, predict.time=12, method="Cox",
main="Curva ROC - 12 meses", lty=2, col="red")
ROCfitxmodelFinal_6=risksetROC(Stime=tempos, status=cens, marker=eta, predict.time=6, method="Cox",
main="Curva ROC - 6 meses", lty=2, col="red")
```

```
#
AUCTreino=risksetAUC(Stime=tempos,status=cens, marker=eta, method="Cox", tmax=12, type="b",
xlab="Tempo");
#
auc12<-ROCfitxmodelFinal_12$AUC
auc12
auc6<-ROCfitxmodelFinal_6$AUC
auc6
#
gini12<-(2*auc12)-1
gini12
gini6<-(2*auc6)-1
gini6
#
beta1<-fitxmodel2$coefficients[1]
beta2<-fitxmodel2$coefficients[2]
beta3<-fitxmodel2$coefficients[3]
beta4<-fitxmodel2$coefficients[4]
beta5<-fitxmodel2$coefficients[5]
beta6<-fitxmodel2$coefficients[6]
beta7<-fitxmodel2$coefficients[7]
beta8<-fitxmodel2$coefficients[8]
beta9<-fitxmodel2$coefficients[9]
#
Ht<-basehaz(fitxmodel2,centered=F)
temp<-Ht$time
H0<-Ht$hazard
s0<-exp(-H0)
s1<-s0^exp(beta1*0+beta2*0+beta3*1+beta4*30+beta5*2.13+beta6*0+beta7*0+beta8*0+beta9*1)
plot(temp, s1, lty=1, lwd=2, type="l",
   xlim=range(c(0,12)),ylim=range(c(0.7,1)), xlab="Tempo (Meses)",
   ylab="S(t|x)")
lines(c(0,temp),c(1,s1),lty=1, col=1)
#
s2<-s0^exp(beta1*0+beta2*1+beta3*0+beta4*30+beta5*0+beta6*0+beta7*1+beta8*1+beta9*1)
lines(c(0,temp), c(1,s2), lty=2, lwd=2, col=2)
#
s3<-s0^exp(beta1*0+beta2*1+beta3*0+beta4*30+beta5*0+beta6*1+beta7*1+beta8*1+beta9*1)
lines(c(0,temp), c(1,s3), lty=3, lwd=2, col=3)
#
# Modelo 3
#
fitxmodel3<-coxph(Surv(tempos,cens)~(x2 + x12 + x13 + x16 + x17 + x19 + x20), x=T, method = "breslow")
resFitxmodel3<-summary(fitxmodel3)
print(resFitxmodel3)
loglikm3<-fitxmodel3$loglik[2]
print(loglikm3)
#
TRV_m3<-2*(loglikBase-loglikm3)
pvalue_m3<-1-pchisq(TRV_m3,6)
```

```
TRV_m3
pvalue_m3
#
AIC_m3<-(-2*loglikm3)+(2*2)
AIC_m3
AIC(fitxmodel3)
#
# TESTE
#
fitxmodelTeste<-coxph(Surv(temposz,censz)~(z2 + z4t + z12 + z13 + z16 + z17 + z19 + z20), x=T, method =
"breslow")
resFitxmodelTeste<-summary(fitxmodelTeste)
print(resFitxmodelTeste)
#
# Curva ROC + AUC + Gini
#
par(mfrow=c(2,3))
eta1 <- fitxmodelTeste$linear.predictor
ROCfitxmodelTeste_12=risksetROC(Stime=temposz, status=censz, marker=eta1, predict.time=12,
method="Cox", main="ROC Curve", lty=2, col="red")
ROCfitxmodelTeste_6=risksetROC(Stime=temposz, status=censz, marker=eta1, predict.time=6, method="Cox",
main="ROC Curve", lty=2, col="red")
#
AUCTeste=risksetAUC(Stime=temposz,status=censz, marker=eta1, method="Cox", tmax=12, type="b",
xlab="Tempo");
#
auc12t<-ROCfitxmodelTeste_12$AUC
auc12t
auc6t<-ROCfitxmodelTeste_6$AUC
auc6t
#
gini12t<-(2*auc12t)-1
gini12t
gini6t<-(2*auc6t)-1
gini6t
```