UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

ANÁLISE DE CLASSES LATENTES: DA TEORIA À PRÁTICA

Autor: Juliana Obino Mastella
Orientadora: Professora Drª. Stela Maris de Jezus Castro
Coorientadora: Professora Drª. Lisiane Priscila Roldão Selau

Porto Alegre, 9 de dezembro de 2015.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

ANÁLISE DE CLASSES LATENTES: DA TEORIA À PRÁTICA

Autor: Juliana Obino Mastella

Trabalho de Conclusão de Curso
apresentado para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Drª. Stela Maris de Jezus Castro
Professora Drª. Lisiane Priscila Roldão Selau
Professor Dr Alvaro Vigo

Porto Alegre, 9 de dezembro de 2015.

AGRADECIMENTOS

Gostaria de deixar registrado meu agradecimento a todos que me apoiaram ao longo deste curso, desde quando decidi que iria fazer uma nova graduação. Inicialmente, um agradecimento especial aos meus pais que sempre me apoiaram nas minhas escolhas.

Agradeço ao meu amor, meu companheiro de todos os momentos, meu marido, Diego, que acompanhou quase toda minha trajetória na Estatística. Agradeço por todo apoio, amor, carinho e principalmente: compreensão, especialmente nos finais dos semestres, e durante os semestres mais atarefados. Teria sido muito mais difícil sem você ao meu lado para me ajudar, ou mesmo me acalmar nos momentos mais estressantes.

Aos meus amigos pelo apoio e paciência de entender que muitas vezes foi preciso ficar ausente para me dedicar aos estudos.

Às minhas professoras e orientadoras Stela e Lisiane por todo aprendizado durante o curso e por terem me acolhido como orientanda nesta etapa final. Obrigada por todo apoio, orientação, disponibilidade, ideias, desde o início deste trabalho.

Aos demais professores do Departamento de Estatística que foram essenciais na minha formação: obrigada por toda a disponibilidade e todo o aprendizado.

À professora Sidia, a qual foi minha primeira professora de uma disciplina de estatística, em uma disciplina extracurricular, quando tive certeza de que iria cursar a graduação em Estatística.

Aos amigos/colegas que fiz na graduação, com quem compartilhei diversos momentos dessa caminhada.

Aos meus amigos/colegas de trabalho que foram sempre disponíveis, com quem tive a oportunidade de colocar em prática e aprimorar conhecimentos adquiridos durante a graduação, além de desenvolver outras habilidades.

Resumo

Em algumas áreas de pesquisa é bastante comum o interesse por variáveis que não são diretamente observáveis, as quais são denominadas variáveis latentes. Nos casos em que tanto a variável latente como as variáveis observadas são categóricas, a análise de classes latentes é uma possível técnica a ser utilizada. Há grande potencial desse método em diferentes áreas de conhecimento, mas, talvez por desconhecimento, ele não seja tão empregado quanto poderia ser. Desta forma, no intuito de facilitar futuras aplicações desse tipo de abordagem, o presente estudo traz um roteiro para a aplicação da análise de classes latentes, utilizando o procedimento PROC LCA do SAS. Na sequência é apresentada uma aplicação a dados reais de clientes proponentes de cartão de crédito em uma rede de farmácias do Rio Grande do Sul, a partir de variáveis sabidamente relacionadas ao comportamento de inadimplência. Para o ajuste do modelo base foram utilizados 8 itens. O modelo se apresentou invariante à variável de grupo testada e todas as covariáveis foram significativas ao nível de 0,0001. Ao final, foi ajustado um modelo de 9 classes latentes, com 3 covariáveis. Por fim, verificou-se que foi possível identificar classes latentes com diferentes comportamentos de pagamento.

Sumário

1. Introdução

Em algumas áreas de pesquisa é bastante comum o interesse por variáveis que não são diretamente observáveis, as quais são denominadas variáveis latentes (GAMA, 2011). Nesses casos, a abordagem comumente utilizada é estimar ou explorar essa variável a partir de variáveis observáveis, sabidamente relacionadas a ela. Esse tipo de interesse é muito comum em ciências que estudam o comportamento, como exemplo: a segmentação de consumidores pelo marketing (BHATNAGAR; GHOSE, 2004; DUBOIS *et. al.,* 2005); a compreensão dos padrões de consumo de álcool pela epidemiologia (SACCO, *et al.* 2009); a identificação de diferentes perfis de depressão pela psiquiatria (SULLIVAN, *et al.*1998); a produção de indicadores socioeconômicos pela economia (SILVA; PREVIDELLI, 2012); a estratificação social conforme o consumo da arte pela sociologia (CHAN; GOLDTHORPE , 2007); entre outros.

Assim como as variáveis observáveis, as variáveis latentes podem ser quantitativas ou categóricas, havendo abordagens adequadas a cada caso. Naqueles em que tanto a variável latente como as variáveis observadas são categóricas (nominais ou ordinais), a análise de classes latentes (LCA, *Latent Class Analysis,* em inglês) é uma técnica interessante a ser utilizada. O objetivo da LCA é identificar uma série de subgrupos mutuamente exclusivos de indivíduos, de acordo com a variável latente de interesse, com base em uma série de variáveis categóricas observadas, notoriamente relacionadas à mesma. Ela pode ser entendida como um modelo de segmentação, similar à análise de cluster (AC). Entretanto, o modelo de classes latentes vai além de uma análise exploratória, possuindo algumas vantagens em relação à AC: é possível estabelecer de forma objetiva o número ideal de classes comparando estatísticas de ajuste para cada um dos modelos; é possível estimar a probabilidade de um indivíduo pertencer a cada uma das classes; é possível testar invariância entre grupos; além de ser possível incluir covariáveis que auxiliem a predizer a classe latente do indivíduo.

Diante do exposto, há grande potencial de aplicação da LCA a diferentes áreas de conhecimento. Contudo, talvez por desconhecimento ou por limitações técnicas

do pesquisador, o método não é tão empregado quanto poderia ser. Desta forma, o objetivo do presente estudo é apresentar um roteiro para a aplicação da análise de classes latentes, utilizando o procedimento PROC LCA do SAS, e na sequência aplicá-lo a dados reais relativos a clientes proponentes de cartão de crédito de uma rede de farmácias do Rio Grande do Sul, aliando teoria e prática, contribuindo para uma maior utilização do método.

2. Modelos de Classes Latentes**:**

O objetivo da análise de classes latentes (LCA) é identificar uma série de subgrupos mutuamente exclusivos de indivíduos com base em um conjunto de variáveis categóricas observadas, sabidamente relacionadas à variável latente de interesse. O modelo tradicional de classes latentes é ajustado a partir de indicadores categóricos, nesse caso o modelo possui a vantagem de não fazer pressupostos a respeito da distribuição desses indicadores além da independência local, ou seja, em uma mesma classe latente os indicadores devem ser independentes (VERMUNT e MAGIDSON, 2004).

Clogg (1992) afirma que os precursores desse tipo de análise vieram das Ciências Sociais e do comportamento com Lazarsfel, em 1950, a partir de um estudo com dados de uma pesquisa conduzida durante a segunda guerra mundial. Segundo Goodman (2002), a utilização desses modelos como ferramenta para auxiliar a compreender a relação entre variáveis categóricas observadas possui história recente, datada do século XX, apesar de afirmar que no início do século XIX já eram utilizados alguns modelos matemáticos que podem ser considerados casos especiais dos modelos de classes latentes. O maior avanço dos modelos LCA se desenvolveu na segunda metade do século XX, sendo que a sua aplicação prática por pesquisadores de diferentes áreas só se tornou mais difundida no último quarto do século, com o desenvolvimento de métodos estatísticos eficientes e modelos de classes latentes mais genéricos (GOODMAN, 2002).

As áreas de aplicação dos modelos de classes latentes são diversas, sendo empregados principalmente em estudos que visam a identificar comportamentos, percepção e outros traços psicológicos. Por exemplo, Sullivan (1998) utilizou a LCA para segregar indivíduos depressivos em diferentes subgrupos a partir de sintomas depressivos. Por sua vez, Sacco (2009) buscou identificar classes latentes que representassem os diferentes padrões de consumo de álcool entre idosos e explorou a associação de fatores de risco com cada uma das classes, incluindo covariáveis no modelo. Já na área de marketing, Bhatnagar e Ghose (2004) aplicaram os modelos de classes latentes para segmentar consumidores virtuais, baseado no seu padrão de compra em diversas categorias de produtos, e na sequência foram avaliadas características demográficas para compreender se havia relação delas

com as classes formadas. Dubois *et. al.* (2005) também utilizaram a LCA para estabelecer uma segmentação de consumidores, mas sob outro aspecto, baseado em suas opiniões a respeito da luxúria a partir de algumas questões com as quais poderiam concordar ou discordar, em uma escala *Likert*. Mais recentemente, em 2014, Swanson *et al.* utilizaram o método para identificar e compreender diferentes classes de pacientes jovens com desordens alimentares.

Nesse cenário, pode-se notar uma pluralidade de ciências cujos problemas de pesquisa podem ser estudados a partir do ajuste e da análise de modelos de classes latentes.

2.1. Definição do modelo:

Inicialmente, é importante definir o modelo base de classes latentes (*basic model*), a partir do qual serão analisadas algumas variações mais complexas. Dessa forma, segundo Collins e Lanza (2010) o modelo base é definido pela seguinte equação:

$$P(Y_i = y) = \sum_{C=1}^{C} \gamma_c \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|c}^{I(y_{im}=k)}$$

$$I(y_{im} = k) = \begin{cases} 1, \text{ se } y_{im} = k \\ 0, \text{ c.c.} \end{cases} \quad (2.1.1)$$

Em que $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iM})$ é o vetor correspondente às respostas do i-ésimo indivíduo aos M itens. Dentre os parâmetros que se deseja estimar em um modelo tradicional de classes latentes tem-se dois grupos: os parâmetros γ_c que representam a probabilidade de pertencer à c-ésima classe latente; os parâmetros $\rho_{mk|c}$ que representam a probabilidade de resposta à k-ésima categoria do m-ésimo item condicionada à classe latente c. O único pressuposto que se faz é a independência local, ou seja, os M itens devem ser independentes dentro de cada classe latente.

Quando é incluída uma variável de grupo a esse modelo, tanto os parâmetros γ_c como os $\rho_{mk|c}$ são condicionados ao grupo. Já caso sejam adicionadas covariáveis, um novo conjunto de parâmetros é estimado: os parâmetros β que

correspondem aos coeficientes da regressão logística multinomial para as covariáveis (x_i).

LANZA *et al.* (2007) definem que, seja um modelo de classes latentes com C classes a serem estimadas a partir de M variáveis categóricas (itens) observadas (*m*= 1, ..., M), e uma variável de grupo com G categorias (*g* = 1, ..., G). Considerando Y_i= ($Y_{i1}, Y_{i2}, \dots, Y_{iM}$) o vetor correspondente às respostas do i-ésimo indivíduo aos M itens, onde cada m item possui r_m categorias de resposta. Seja ainda c_i = 1, 2, ..., C a classe latente a qual o indivíduo *i* pertença e I (*y* = *k*) a função indicadora que será igual a 1 se a resposta de y for igual à k-ésima categoria do item m, e 0 caso contrário. Suponha também que g_i represente o valor do grupo correspondente ao i-ésimo indivíduo, x_i represente o valor da covariável também para o indivíduo i e seu valor possa ser relacionado com as probabilidades (γ_c) de pertencer a cada classe latente. Sendo assim, o modelo de classes latentes incluindo variável de grupo e covariáveis pode ser expresso da seguinte forma:

$$P\,(Y_i = y|x_i, g) = \sum_{C=1}^{C} \gamma_{c|g}(x_i) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|cg}^{I(y_{im}=k)} \quad \textbf{(2.1.2)}$$

$$I(y_{im} = k) = \begin{cases} 1, \text{ se } y_{im} = k \\ 0, \text{ c.c.} \end{cases}$$

Em que $\gamma_{c|g}(x_i)$ representa um modelo logístico multinomial padrão. Por exemplo, para o caso de ter apenas uma covariável x, os parâmetros γ são expressos pela probabilidade condicional:

$$\gamma_{c|g}(x_i) = P(C_i = c|x_i, G_i = g) = \frac{\exp\{\beta_{0c|g} + x_i\beta_{1c|g}\}}{1 + \sum_{j=1}^{C-1} \exp\{\beta_{0j|g} + x_i\beta_{1j|g}\}} \quad \textbf{(2.1.3)}$$

Dessa forma, uma vez que a probabilidade de um indivíduo pertencer a determinada classe é estimada como função das covariáveis, as quais variam conforme o indivíduo, no caso de modelos LCA com covariáveis é calculado um vetor de estimativas da probabilidade de pertencer a cada classe latente, para cada indivíduo. Assim, a prevalência para cada classe latente é calculada como uma

média entre as probabilidades específicas por indivíduo de pertencer a cada classe latente (LANZA *et al.,* 2007).

3. Introdução ao procedimento PROC LCA

Alguns programas estatísticos apresentam pacotes específicos para realizar a análise de classes latentes, entre eles o pacote *Latent GOLD*® da *Statistical Innovations*; já dentre os programas livres é possível citar os pacotes *poLCA* e *MCLUST* do R. Uma comparação entre esses três procedimentos, assim como uma definição mais detalhada de cada um deles, é apresentada por Haughton *et al.* (2009). Ainda dentre os programas pagos, outro bastante utilizado para essa finalidade é o M*plus*®*.*

Além desses, temos um procedimento desenvolvido pelo Centro de Metodologia da Universidade do Estado da Pensilvânia (The Metodology Center, The Pennsylvania State University) destinado à análise de classes latentes no *software* SAS, o PROC LCA (PROC LCA & PROC LTA, 2015), objeto de estudo do presente trabalho. Outro procedimento disponibilizado em conjunto com o PROC LCA é o PROC LTA, o qual tem a finalidade de ajustar um modelo de análise transacional latente (em inglês *Latent Transition Analysis),* uma versão especial da LCA para dados longitudinais. Importante salientar que, uma vez que o procedimento não foi desenvolvido pela empresa SAS, esta não presta suporte, nem garante seu funcionamento. Entretanto, em casos de necessidade de apoio técnico, é possível encaminhar as eventuais dúvidas aos desenvolvedores do PROC LCA[1].

3.1. Passo a Passo do PROC LCA:

3.1.1 Preparação dos dados

Inicialmente, é importante que o banco de dados esteja na estrutura que o procedimento exige. Deve haver uma linha por observação, ou ainda é possível trabalhar com dados agrupados por perfil de resposta, contendo uma variável extra com a frequência de cada perfil de resposta, a qual, neste caso, deverá ser identificada no código. Apenas nos modelos com covariáveis não é possível

[1] É indicado que dúvidas a respeito do PROC LCA sejam encaminhadas para o endereço eletrônico: mchelpdesk@psu.edu.

trabalhar com dados agrupados. Além disso, é importante que as categorias dos itens estejam codificadas com números inteiros de forma sequencial, a partir de 1. Sendo M o número de itens que compõem o modelo, caso algumas observações tenham dimensão<M, ou seja, há *missings* em um ou mais itens, eles são tratados como *missings* ao acaso, considerando as informações válidas do indivíduo para estimar os parâmetros dos itens. Em contrapartida, no caso de qualquer *missing* na variável de grupo ou em alguma covariável, a observação como um todo é excluída da análise.

Quanto à variável de grupo, apenas uma variável pode ser incluída. Entretanto, nos casos em que há mais do que uma variável de grupo, é possível cruzá-las, criando os diferentes perfis, construindo uma variável única. As categorias dessa variável também devem ser codificadas como números inteiros positivos, de forma sequencial.

Já em se tratando das covariáveis, no caso das numéricas elas podem ser mantidas em sua forma original, enquanto covariáveis categóricas também podem ser incluídas, mas para tanto é necessário que sejam transformadas em variáveis *dummies*.

3.1.2 Ajuste do modelo base

A Tabela 10, disponível no Anexo I, apresenta um resumo das opções e declarações disponíveis no procedimento PROC LCA. Os desenvolvedores do procedimento (LANZA *et al.,* 2007) sugerem que antes de serem testados modelos mais complexos, contendo variáveis de grupo e covariáveis, sejam ajustados diferentes modelos base, definindo-se o número de classes latentes que será utilizado, para após serem testados modelos mais complexos a partir do modelo base escolhido.

Os elementos mínimos a serem identificados para o ajuste do modelo LCA são: DATA, contendo o banco de dados SAS a ser analisado; NCLASS definindo o número de classes latentes do modelo; ITENS declarando as variáveis que serão usadas para ajustar as classes latentes; CATEGORIES declarando o número de categorias para cada item incluído no modelo.

3.1.3 Avaliação do Modelo

Alguns autores (HAUGHTON *et al.*, 2009; LANZA *et al.*, 2007) sugerem que sejam ajustados sequencialmente modelos com duas classes, três classes e assim por diante, para após ser escolhido o modelo LCA base ótimo. O procedimento PROC LCA oferece diferentes ferramentas para essa escolha, dentre elas: estatística G^2 da razão de verossimilhança; Critério de Informação de Akaike (AIC); Critério de Informação de Akaike Corrigido (AICc); Critério de Informação Bayesiano (BIC); entropia, entre outros.

Apesar dos modelos com diferente número de classes latentes serem tecnicamente aninhados, a distribuição da estatística da razão de verossimilhança comparando dois modelos não deveria ser comparada a uma χ^2, dessa forma a diferença entre as estatísticas G^2 só poderia ser utilizada de forma “grosseira” para comparar o ajuste dos modelos (VERMUNT e MAGIDSON, 2004; COLLINS e LANZA, 2010). Assim, seria mais recomendada a utilização dos critérios de informação, os quais são medidas que penalizam a verossimilhança pelo acréscimo de parâmetros (Emiliano *et al.*, 2014), sendo que quanto menor os seus valores, melhor será o ajuste do modelo, já balanceando entre um modelo mais parcimonioso e melhor ajustado.

O critério AIC é dado pela expressão:

$$AIC = 2k - 2\ln(L) \qquad \textbf{(3.1.3.1)}$$

Em que k representa o número de parâmetros e L o valor maximizado da função verossimilhança do modelo ajustado (MIOLA, 2013). A utilização do AIC permite grande flexibilidade na comparação de modelos, os quais podem ser lineares, não lineares, aninhados ou não aninhados. A sua base teórica combina a teoria da máxima verossimilhança, a teoria da informação e o conceito de entropia da informação (FLORIANO *et al.*, 2006).

Proposto por Schwarz (1978), o critério BIC é semelhante ao critério AIC, porém aplica uma penalidade maior ao acréscimo de parâmetros a serem estimados. É dado pela seguinte expressão:

$$BIC = k\ln(n) - 2\ln(L) \qquad \textbf{(3.1.3.2)}$$

Em que *k* é o número de parâmetros estimados no modelo, *n* o tamanho da amostra e *L* o valor maximizado da função de verossimilhança para o modelo estimado (MIOLA, 2013).

Outra possibilidade para seleção do número de classes é utilizar o critério da entropia. A medida de entropia avalia a incerteza na classificação dos indivíduos, no PROC LCA ela é apresentada como RAW ENTROPY. No intuito de avaliar a qualidade da classificação, é comum ser utilizada a sua versão relativizada, a qual varia entre 0 e 1, sendo que quanto mais próximo de 1 ela for, melhor será o ajuste do modelo (RAMASWAMY *et al.*, 1993). Essa medida é apresentada no PROC LCA como ENTROPY, por meio da opção OUTPUTPARAM e também na saída do procedimento.

A Entropia bruta (RAW ENTROPY) é baseada nas probabilidades à posteriori e é dada pela seguinte fórmula:

$$S = \sum_{i=1}^{n} \sum_{c=1}^{C} -p_{ic} ln\, p_{ic} \quad \textbf{(3.1.3.3)}$$

Em que p_{ic} é a probabilidade à posteriori do i-ésimo indivíduo pertencer à c-ésima classe latente (COLLINS, LANZA, 2010). Para avaliar o grau de separação entre as classes, costuma-se utilizar a Entropia relativizada (ENTROPY) que é dada pela seguinte fórmula:

$$ENTROPY = 1 - \frac{S}{(n \ln(C))} \quad \textbf{(3.1.3.4)}$$

Em que S é a entropia bruta (RAW ENTROPY), n é o tamanho da amostra e C é o número de classes latentes (RAMASWAMY *et al.*, 1993).

Haughton *et al.* (2009) apontam que, conforme a semente utilizada, o ajuste do modelo pode ficar levemente diferente, portanto sugerem que sejam gerados diferentes modelos, com sementes aleatórias para cada quantidade de classes, a fim de se obter um ajuste ótimo. Vermunt e Magidson (2004) afirmam que esse problema de estimação está relacionado com a presença de máximos locais na função de verossimilhança, dessa forma o algoritmo pode convergir para diferentes máximos, conforme o valor inicial. Os autores do PROC LCA afirmam que é esperado que ocorram essas pequenas diferenças no ajuste partindo de sementes variadas, entretanto o resultado ótimo deve repetir-se na maior parte das vezes,

caso contrário poderia indicar problemas de identificação do modelo (LANZA *et al*., 2007).

Emiliano *et al.* (2014) demonstraram, via simulação a partir de diferentes modelos e com diferentes tamanhos de amostra, que o desempenho dos critérios de informação AIC, AICc e BIC variam conforme o tipo de modelo e o tamanho de amostra, no entanto AIC e AICc tendem a possuir melhor desempenho para amostras muito pequenas, enquanto o BIC tende a apresentar um desempenho superior para a seleção de modelos a partir de amostras grandes. No mesmo sentido, Nylund (2007) testou via simulação de Monte Carlo o desempenho de diferentes critérios para a seleção do número ótimo de classes em modelos LCA, para diferentes tamanhos de amostra, verificando que o BIC apresentou o melhor desempenho dentre os critérios de informação testados, especialmente conforme o tamanho de amostra aumenta.

Tein *et al.* (2013) realizaram um estudo via simulação de diferentes modelos, comparando alguns métodos para a seleção do número de classes latentes, dentre eles: AIC, BIC e entropia relativizada. Nesse estudo os critérios AIC e entropia apresentaram desempenho inferior para a seleção do número de classes, enquanto BIC esteve entre os critérios que apresentaram melhor desempenho para diferentes quantidades de classes, tamanhos de amostra e distâncias entre as classes.

Dessa forma, após o teste de diferentes modelos base, seleciona-se aquele que oferecer melhor ajuste, conforme o critério adotado pelo pesquisador e considerando a facilidade de interpretação do modelo (HAUGHTON *et al.*, 2009).

3.1.4 Modelos Expandidos:

3.1.4.1 LCA com múltiplos grupos

Similar à situação quando há funcionamento diferencial do item (DIF) nos modelos TRI (Teoria da Resposta ao Item), existem situações em que há motivos para se acreditar que a probabilidade de pertencer à determinada classe latente, relacionada às possíveis respostas de cada item, varie conforme determinados grupos. Diante disso, é possível testar a invariância do modelo de classes latentes quanto a essa variável, reestimando todos os parâmetros condicionando-os ao grupo. Neste caso, essa variável deve ser incluída na opção GROUPS do PROC

LCA. Contudo, é importante verificar a presença de *missings* para a variável grupo, pois nesses casos toda a observação é excluída da análise de classes latentes (LANZA *et al.*, 2007).

No intuito de testar se o modelo é invariante quanto a uma determinada variável, é recomendado ajustar o modelo sem a variável de grupo e reestimar os parâmetros condicionando-os ao grupo. Uma vez que esses modelos são naturalmente aninhados e a distribuição da estatística de teste da diferença na razão de verossimilhança tem distribuição χ^2, é possível avaliar a mudança na qualidade do ajuste. Para tanto, considera-se:

$$W = \prod_{m=1}^{M} R_m \quad \textbf{(3.1.4.1.1)}$$

$$G^2 = 2\sum_{w=1}^{W} f_w ln\left(\frac{f_w}{\hat{f}_w}\right), \quad \textbf{(3.1.4.1.2)}$$

$$gl_{G^2} = W - P - 1, \quad \textbf{(3.1.4.1.3)}$$

$$G_\Delta^2 = G_B^2 - G_A^2, \quad \textbf{(3.1.4.1.4)}$$

$$gl_{G_\Delta^2} = gl_{G_B^2} - gl_{G_A^2}, \quad \textbf{(3.1.4.1.5)}$$

Em que, P representa o número de parâmetros do modelo; M representa o número de itens; R_m o número de categorias no m-ésimo item; f_wa frequência observada da w-ésima célula da tabela de contigência e $\hat{f}_w$ a frequência esperada da w-ésima célula. Maiores detalhes podem ser encontrados em Collins e Lanza (2010).

Para testar a hipótese nula de que existe invariância entre os grupos, utiliza-se a estatística G_Δ^2 que segue uma distribuição χ^2 com $gl_{G_\Delta^2}$ graus de liberdade, quando há relativamente poucos graus de liberdade. Nos casos em que há um grande número de parâmetros ρ no modelo, a estatística G_Δ^2 pode não ser bem aproximada por uma distribuição χ^2 quando há muitos graus de liberdade. Nesses casos, é comum utilizar-se as estatísticas AIC e BIC para decidir qual modelo possui melhor ajuste (COLLINS e LANZA, 2010).

3.1.4.2 LCA com covariáveis

Outra opção disponível no procedimento estudado é a inclusão de covariáveis a serem declaradas na opção COVARIATES do PROC LCA. Neste caso, a probabilidade de pertencer a cada classe latente será ajustada para as covariáveis, por meio de um modelo de regressão logística multinomial em que a variável resposta será a classe latente. Por padrão, a classe de referência é a primeira, todavia ela pode ser alterada através da opção REFERENCE. Da mesma forma que para a inclusão de uma variável de grupo, a presença de *missing* em alguma covariável leva à exclusão de todos os dados da observação para a análise.

3.1.5 Opções de estimação

Conforme apresentado anteriormente, no modelo base temos os parâmetros γ e ρ para serem estimados. Para tanto, é utilizado o método de estimação pela maximização da função de verossimilhança, por meio de um procedimento iterativo que busca, dentre os possíveis parâmetros para o modelo, aqueles que são mais verossímeis de serem os reais, dado a informação da amostra (COLLINS e LANZA, 2010).

O procedimento PROC LCA utiliza um algoritmo de otimização (EM – maximização da esperança) para gerar as estimativas de máxima verossimilhança. Assim, os parâmetros são estimados com base no banco de dados, valores iniciais e especificações do modelo, de forma iterativa até atingir o critério de convergência ou o número máximo de iterações estabelecido.

O critério de convergência adotado é o desvio máximo absoluto (DMA) entre as estimativas para os parâmetros a cada iteração, que por padrão é 0,000001, mas pode ser alterado pelo comando CRITERION. Já o número máximo de iterações padrão é 5.000, mas também pode ser modificado pelo comando MAXITER.

Nos casos em que há pouca quantidade de informação (tamanho de amostra, número de itens) pode ser difícil identificar uma solução ótima para o modelo. De maneira geral, modelos complexos, ou seja, com um maior número de classes latentes, grupos, covariáveis, requerem mais informação do que os mais simples.

O guia do usuário PROC LCA (LANZA *et al*., 2015) ainda traz algumas sugestões para contornar possíveis problemas na estimação dos parâmetros, que normalmente estão relacionados a alta variabilidade nos dados ou quando o valor

estimado para algum parâmetro fica próximo de zero. No caso de modelos com covariáveis, em caso de dificuldade para estimar os coeficientes de regressão, é possível acrescentar uma distribuição a priori para esses parâmetros, a partir dos próprios dados. Para tanto, utiliza-se o comando BETA PRIOR definindo um valor real positivo que indicará o peso dessa priori, sendo que 0 não utilizaria a priori e o sugerido pelos autores é uma priori igual a 1. A mesma abordagem pode ser usada, se necessário, para os parâmetros γ (GAMMA PRIOR) e ρ (RHO PRIOR).

Grande parte dos problemas de estimação, devido à alta variabilidade dos dados, são solucionados com a inclusão da priori, os quais são mais comuns em amostras pequenas. Entretanto, em casos extremos essa abordagem não é suficiente, sendo recomendável reduzir a complexidade do modelo: o número de classes ou covariáveis.

3.1.6 Resultados opcionais

As estimativas para os parâmetros são apresentadas no *output* padrão do PROC LCA, mas caso haja interesse em armazená-las em uma tabela SAS isso é possível através da opção OUTEST (versão em uma linha) ou OUTPARAM (versão “amigável”).

Outra opção útil é armazenar as probabilidades a posteriori, as quais representam a probabilidade de o indivíduo pertencer a cada classe latente, utilizando o comando OUTPOST. Além dos itens, das variáveis de grupo e das covariáveis utilizadas para ajuste do modelo que por padrão são mantidas na tabela de saída, é possível manter outras variáveis de interesse, identificando-as na opção ID separadas por espaços. Além disso, nessa tabela é acrescentada a variável BEST que representa a classe latente a que o sujeito é atribuído, considerando a classe modal, ou seja, que apresenta maior probabilidade à posterori (VERMUNT e MAGIDSON, 2004). A partir da separação dos indivíduos nas classes latentes, é possível estudá-las, compará-las e desenvolver outras análises.

Uma forma de avaliar a acurácia do modelo seria verificar a média da probabilidade à posteriori. Caso esse valor seja muito próximo de um, indica que na maior parte das vezes os indivíduos estão sendo atribuídos à classe que realmente pertencem (COLLINS, LANZA, 2010). Ou seja, o modelo consegue separar bem as

classes, sendo a soma do vetor das probabilidades à posteriori igual a 1, se a classe à qual foi atribuído apresentada valor próximo de um, as demais apresentam valor próximo de zero.

Ainda em relação às opções de saída, quando há interesse em um *output* mais completo, pode-se utilizar o comando VERBOSE_OUTPUT, assim serão apresentadas também as restrições utilizadas para os parâmetros, valores iniciais, além do histórico completo das iterações até o método de otimização atingir o critério de parada.

4. Aplicação a dados reais

Após ser definido um roteiro para utilizar o procedimento PROC LCA do SAS, ele foi aplicado a dados reais no intuito de identificar se seria possível estabelecer classes latentes relacionadas ao perfil de inadimplência, num grupo de proponentes de cartão de crédito, a partir de variáveis cadastrais sabidamente relacionadas à inadimplência.

Neste exemplo, o traço latente de interesse seria o comportamento de inadimplência de clientes da farmácia, no produto cartão de crédito. No intuito de verificar posteriormente se foi possível identificar e formar classes com diferentes padrões de inadimplência, as classes foram comparadas com o padrão ouro, máximo dias de atraso ao longo de um ano após a concessão.

4.1. Apresentação do banco de dados

A amostra do presente estudo foi disponibilizada por uma rede de farmácias do Rio Grande do Sul, sendo composta por 12.257 clientes que solicitaram cartão de crédito e não apresentavam relacionamento anterior com a empresa.

4.2. Definição das variáveis:

As variáveis disponíveis foram separadas em três tipos: (1) itens usados para ajuste do modelo base de classes latentes; (2) grupo a ser testado para invariância; e (3) covariáveis utilizadas como preditoras na regressão logística multinomial.

4.2.1 Itens

No intuito de definir as variáveis que seriam utilizadas como itens para compor o modelo base de classes latentes, foram identificadas aquelas que sabidamente estão relacionadas ao perfil de inadimplência segundo a literatura da área. Os modelos de crédito prestam um importante suporte à tomada de decisão nas instituições que os utilizam, dessa forma normalmente seu conteúdo é de caráter sigiloso. Devido ao sigilo de informação associado aos modelos de crédito, grande parte dos estudos com dados reais não divulga as variáveis utilizadas para o

desenvolvimento dos modelos, ou quando o fazem, não especificam detalhes a respeito delas, apenas trazendo categorias codificadas. Abaixo são apresentadas as variáveis utilizadas como itens na presente análise de classes latentes.

Idade

Variável originalmente contínua referente ao tempo, em anos, decorrido desde o nascimento da pessoa até a data de realização do cadastro. Parece haver consenso na literatura de que esta variável esteja inversamente relacionada à probabilidade de inadimplência, independente do segmento do crédito, ou seja, espera-se que clientes mais jovens sejam mais arriscados (DINIZ; LOUZADA, 2013; LIMA, 2004; FERREIRA *et. al.,* 2012).

Profissão

Variável categórica que corresponde à profissão declarada pelo indivíduo no momento do cadastro, que geralmente é agrupada em grandes grupos considerando a probabilidade de inadimplência.

Tipo de residência

Variável categórica que corresponde ao tipo de residência (própria/ não própria) declarado pelo indivíduo no momento do cadastro. Há estudos que apontam que indivíduos com residência própria são menos arriscados, uma vez que em uma eventual adversidade o patrimônio poderia tornar-se recurso, mesmo que de baixa liquidez, além da inexistência de despesa com aluguel aumentar a disponibilidade de sua renda (CAOUETTE *et al.*, 2008; LIMA, 2004; VASCONCELLOS, 2002).

Estado Civil

Variável categórica que corresponde ao estado civil declarado pelo indivíduo no momento do cadastro. Parece haver consenso de que indivíduos casados apresentem menor probabilidade de inadimplência em relação a solteiros (CAOUETTE *et al.*, 2008; LIMA, 2004; VASCONCELLOS, 2002).

CEP Comercial

Variável categórica que corresponde ao CEP Comercial declarado pelo indivíduo no momento do cadastro, que geralmente é agrupada em grandes grupos considerando a probabilidade de inadimplência.

CEP Residencial

Variável categórica que corresponde ao CEP Residencial comprovado pelo indivíduo no momento do cadastro, que geralmente é agrupada em grandes grupos considerando a probabilidade de inadimplência.

Cidade de Nascimento

Variável categórica que corresponde à cidade de nascimento declarada pelo indivíduo no momento do cadastro, que geralmente é agrupada em grandes grupos considerando a probabilidade de inadimplência.

Tipo de Ocupação

Variável categórica que corresponde ao tipo de ocupação declarado pelo indivíduo no momento do cadastro. Não foi encontrado um consenso a respeito das ocupações, mas Vanconcellos (2002) obteve resultados que indicam que clientes desempregados apresentam maior probabilidade de se tornarem inadimplentes, enquanto profissionais liberais, funcionários de empresas privadas, aposentados e funcionários públicos apresentaram menor risco em relação às demais ocupações avaliadas.

4.2.2 Grupo

Sexo

Parece não haver consenso na literatura quanto a essa variável, alguns autores mostram que as mulheres apresentam menor inadimplência (DINIZ; LOUZADA, 2013), outros não encontraram essa relação em seus estudos, ou afirmam que não faria sentido haver diferença (LIMA, 2004; FERREIRA *et al.,* 2012). Já Miola (2013), que avaliou a inadimplência em uma carteira de crédito direto ao consumidor, identificou uma interação dessa variável com renda, apontando que na população estudada, para aquele produto, mulheres com baixa renda são menos

arriscadas, enquanto homens com alta renda apresentam maior probabilidade de inadimplência.

Portanto, após definido o modelo base, ele foi testado para invariância entre sexo (masculino/feminino).

4.2.3 Covariáveis

Renda

Variável contínua que corresponde à renda (expressa em R$100,00 reais para facilitar a interpretação da razão de chances) declarada pelo indivíduo no momento do cadastro. Parece não haver consenso a respeito da relação dessa variável com o padrão de inadimplência, alguns autores citam que indivíduos com alta renda estariam associados com menor risco de inadimplência (VASCONCELLOS, 2002), entretanto outros mostram o contrário, afirmando que o fato de possuírem maior renda permite que esses indivíduos consigam concessão de maiores valores, dessa forma se arriscando mais (FERREIRA, *et al*. 2012). Diante disso, recomenda-se que a relação dessa variável com a inadimplência seja avaliada especificamente para a população e tipo de operação de interesse.

Tempo de Emprego

Variável originalmente contínua que corresponde ao tempo (em anos) decorrido entre o início do emprego atual e o momento do cadastro. Há indícios de que quanto maior esse tempo, menor será a probabilidade de inadimplência do indivíduo (CAOUETTE *et al.*, 2008; LIMA, 2004).

Escolaridade

Variável categórica, originalmente com quatro categorias (ensino fundamental, ensino médio, ensino superior incompleto, ensino superior completo). De forma geral, espera-se que clientes com maior grau de instrução sejam menos arriscados, pois tendem a possuir maior conhecimento sobre o mercado de crédito, taxa de juros, custo efetivo, além de em média possuírem maior capacidade de pagamento. Dessa forma, seria esperada de pessoas mais instruídas uma maior consciência sobre a operação contratada (FERREIRA *et al.*, 2012; LIMA, 2004).

4.3. Categorização das variáveis:

Uma vez que os itens do modelo LCA devem ser variáveis categóricas, foi necessário incluir esta etapa na preparação das variáveis. A variável “idade” teve que ser categorizada e aquelas variáveis com um número excessivo de categorias tiveram que ser reagrupadas no intuito de facilitar a interpretação e reduzir o número de parâmetros a serem estimados.

Idade

A fim de garantir-se um volume adequado de casos por categoria, os pontos de corte foram feitos a partir dos quantis, utilizando o procedimento PROC RANK do SAS, originando as seguintes categorias: 16 a 24 anos; 25 a 33 anos; 34 a 43 anos; 44 a 55 anos; 56 a 93 anos.

CEP Residencial, CEP Comercial, Profissão, Cidade de Nascimento

Devido ao grande número de categorias originais, sendo algumas com volume muito baixo de clientes para representar aquele grupo, optou-se por utilizar o método de agrupamento de categorias proposto por Selau (2008, p.70-71), baseado no Risco Relativo (RR) para ser utilizado em modelos de crédito. Dessa forma, chegou-se a cinco categorias com probabilidade de inadimplência homogênea em cada categoria e decrescente entre elas. A categorização final obtida é apresentada no Apêndice I.

Tempo de Emprego

Apesar de a variável ser originalmente quantitativa contínua, havia muitos casos de *missings* (73%) que poderiam indicar falha no preenchimento do cadastro, mas também real ausência de emprego ou ainda uma possível “fuga à resposta”, caso a pessoa estivesse há pouco tempo no atual emprego, por exemplo. Dessa forma, para considerar os casos de *missings* como uma possível informação no modelo, a variável foi categorizada como: *missings* (1); até 2 anos (2); mais do que 2 anos (3).

Escolaridade

Uma vez que as categorias “superior completo” e “superior incompleto” representavam um baixo percentual da base trabalhada, aproximadamente 4% cada uma, avaliou-se o comportamento de cada categoria em relação ao padrão ouro, identificando-se que “superior incompleto” era similar ao “ensino médio”, indicando que possuem comportamento similar em relação à inadimplência. Assim, essas categorias foram agrupadas resultando em: ensino fundamental (1); ensino médio/superior incompleto (2); ensino superior completo (3).

4.4. Objetivo da aplicação:

A partir das variáveis apresentadas acima, o objetivo foi identificar um número ótimo de classes latentes existentes dentre os proponentes de cartão de crédito das farmácias em questão, ajustar o modelo base para esse número de classes, verificar se há invariância entre sexo e, na sequência, testar as covariáveis propostas para então ser ajustado o modelo final, acrescentando ao modelo base as covariáveis e a variável de grupo que fossem significativas.

4.5. Escolha do modelo base

Conforme apresentado no capítulo 3.1.3, sugere-se que sejam gerados diferentes modelos, com sementes diferentes para cada quantidade de classes latentes, a fim de se obter um ajuste ótimo e avaliar se o modelo está bem identificado. Sendo assim, foram gerados modelos de 2 a 20 classes latentes, sendo 1024 repetições a partir de sementes aleatórias, armazenando-se as respectivas estatísticas de ajuste, assim como a respectiva semente que gerou cada modelo. Em modelos mais complexos o procedimento não foi capaz de estimar os parâmetros ρ, diante disso, o problema foi contornado adicionando-se uma priori de peso 1 para esses parâmetros. A sintaxe utilizada é apresentada no Apêndice II.

Conforme apresentado anteriormente, há indícios de que a estatística BIC forneça melhor desempenho em relação ao AIC e AICc para amostras grandes (EMILIANO *et al.* 2014). Além disso, o BIC aplica uma penalidade maior ao acréscimo de parâmetros, o que seria adequado considerando a complexidade já existente advinda da quantidade elevada de variáveis no modelo proposto. Dessa

forma, optou-se por trabalhar com o critério BIC para seleção do modelo de classes latentes. A partir disso, obteve-se o resultado apresentado na Figura 1:

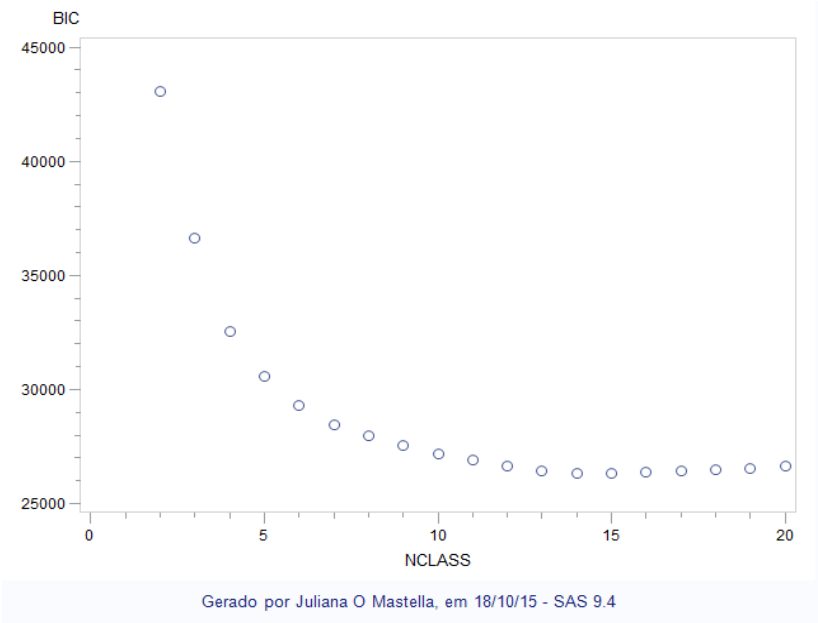


Figura 1 Gráfico de dispersão para escolha do modelo base de Classes Latentes. BIC x Número de Classes (NCLASS)

Conforme o esperado, a Figura 1 demonstra que à medida que se aumenta o número de classes latentes do modelo, há uma melhora no seu ajuste evidenciado pela redução na estatística BIC. Entretanto, após 15 classes latentes começa a ocorrer um aumento no valor deste critério de informação, que por sua vez indica uma piora no ajuste do modelo, apontando que o número ótimo neste caso seria 15 classes latentes.

Na Tabela 1 é apresentada a saída correspondente ao ajuste do modelo base com 15 classes latentes:

Tabela 1 Estatísticas de ajuste para o modelo base com 15 Classes Latentes

Estatística de Ajuste	Valor
Log-verossimilhança	-110.793,90
G^2	22.108,87
AIC	23.006,87
BIC	26.335,69
Entropia	0,84
Graus de liberdade	155.800

Além do menor BIC dentre os modelos gerados, o valor de entropia 0,84, próximo de 1, indica um bom ajuste do modelo.

4.6. Teste LCA com dois grupos

Após a definição do modelo de base, foi ajustado um modelo condicionado à variável de grupo (sexo), conforme Tabela 2.

Tabela 2 Estatísticas de ajuste do modelo com 15 Classes Latentes, incluindo variável de grupo: sexo.

Estatística de Ajuste	Valor
Log-verossimilhança	-109.973,10
G^2	28.237,09
AIC	30.033,09
BIC	36.690,73
Entropia	0,85
Graus de liberdade	311.601

Uma vez que o modelo possui muitos graus de liberdade, conforme apresentado no item 3.1.4, seria mais indicado basear-se nas estatísticas BIC e AIC para a tomada de decisão. Neste caso, as Tabelas 1 e 2 mostram que o modelo base possui AIC e BIC menores do que o modelo condicionado ao grupo, sugerindo melhor ajuste do modelo mais simples. Assim, considerou-se que o modelo apresentou invariância entre os grupos de sexo, e as próximas etapas foram desenvolvidas sem a variável de grupo.

4.7. Teste LCA com covariáveis

Na sequência, tentou-se ajustar o modelo de 15 classes latentes incluindo as covariáveis apresentadas anteriormente: renda, escolaridade e tempo de emprego. Entretanto, dado a complexidade do modelo, o procedimento não foi capaz de estimar os parâmetros. A fim de contornar esta limitação, inicialmente, tentou-se incluir uma priori de peso 1 para os parâmetros β, o que não foi suficiente. Na sequência, tentou-se diminuir a complexidade do modelo reduzindo sequencialmente o número de classes, sendo apenas possível estimar os parâmetros do modelo com covariáveis, quando se utilizou 9 classes latentes.

Uma vez que a variável renda foi incluída na sua forma numérica, seria recomendado conferir se o pressuposto de linearidade da variável foi atendido. Nesse sentido, foram realizadas tentativas para verificar a linearidade, mas o modelo não apresentou convergência em nenhuma delas. Sendo assim, para as seguintes análises, pressupõe-se que a linearidade foi atendida.

Para futuros estudos, uma ideia seria utilizar a informação de localização espacial (CEP residencial e comercial) como sugerido por Fernandes e Artes (2016), criando uma variável numérica, a partir das latitudes e longitudes, que represente a dependência espacial em relação à variável de interesse.

Além disso, pressupõe-se que a independência local esteja atendida, ou seja, os itens idade, profissão, tipo de residência, estado civil, CEP residencial, CEP comercial, cidade de nascimento e tipo de ocupação sejam independentes em uma mesma classe latente. Os resultados deste modelo são apresentados nas Tabelas 3, 4, 5 e 6.

Tabela 3 Testes de Significância para os parâmetros Beta no modelo LCA com 9 Classes Latentes.

Covariável	log-L Exclusão	Variação 2Log-L	Graus de Liberdade	p-valor
Renda	- 109.092,70	709,38	8	<0,0001
Tempo de Emprego (0-2 anos)	- 110.098,04	2.720,05	8	<0,0001
Tempo de Emprego (>2 anos)	- 109.948,96	2.421,89	8	<0,0001
Escolaridade (Ens. Fundamental)	- 109.235,11	994,21	8	<0,0001
Escolaridade (Ens. Superior)	- 108.795,83	115,63	8	<0,0001

Tabela 4 Estimativas dos parâmetros Beta (EP) da Regressão Logística Multimonial, sendo a Classe 4 (Péssimo) a classe de referência. Ausência de informação (*missings*) categoria de referência para variável tempo de emprego; Ensino Médio categoria de referência para variável Escolaridade.

Estimativas	Classes Latentes							
	8- Muito Ruim	7- Ruim	3- Neutro-Ruim	2- Neutro	1- Neutro-Bom	5- Bom	9- Muito Bom	6- Excelente
Intercepto	-0,12 (0,1054)	-1,56 (0,5032)	-1,23 (0,1149)	-1,65 (0,1091)	-0,7 (0,0989)	-1,26 (0,1116)	-0,29 (0,1141)	-0,64 (0,0881)
Renda	-0,06 (0,0109)	-0,61 (0,0547)	0,06 (0,0075)	0,06 (0,0077)	-0,01 (0,0091)	0,07 (0,0075)	-0,1 (0,0126)	0,06 (0,0074)
Tempo de Emprego (0-2 anos)	4,79 (0,3484)	-0,24 (1,0891)	0,34 (0,5233)	4,34 (0,3533)	4,19 (0,3494)	-0,89 (0,5922)	3,84 (0,3538)	-0,74 (0,4823)
Tempo de Emprego (>2 anos)	4,46 (0,3837)	2,17 (0,5988)	1,72 (0,4422)	5,35 (0,3812)	4,42 (0,3813)	0,22 (0,5044)	4,67 (0,3834)	1,71 (0,3956)
Escolaridade (Ens. Fundamental)	-0,44 (0,1017)	3,49 (0,4607)	0,39 (0,129)	0,73 (0,0997)	0,46 (0,0948)	1,63 (0,1021)	-0,21 (0,1064)	1,51 (0,085)
Escolaridade (Ens. Superior)	0,86 (0,2584)	1,76 (1,2852)	1,4 (0,2563)	1,31 (0,2454)	1,46 (0,2424)	-0,31 (0,3451)	2,04 (0,2432)	1,12 (0,2375)

Tabela 5 Estimativas das Razões de Chance, e respectivos intervalos de confiança 95%, da Regressão Logística Multimonial, sendo a Classe 4 (Péssimo) a classe de referência. Ausência de informação (*missings*) categoria de referência para variável tempo de emprego; Ensino Médio categoria de referência para variável Escolaridade.

Razão de Chances	Classes Latentes							
	8- Muito Ruim	7- Ruim	3- Neutro-Ruim	2- Neutro	1- Neutro-Bom	5- Bom	9- Muito Bom	6- Excelente
Renda (R$100,00)	**0,94**	**0,54**	**1,07**	**1,06**	**0,99**	**1,07**	**0,9**	**1,06**
Intervalo de Confiança	[0,92-0,96]	[0,49-0,6]	[1,05-1,08]	[1,05-1,08]	[0,97-1,01]	[1,05-1,08]	[0,88-0,93]	[1,04-1,07]
Tempo de Emprego (0-2 anos)	**119,7**	**0,79**	**1,41**	**76,89**	**66,31**	**0,41**	**46,57**	**0,48**
Intervalo de Confiança	[60,47-236,98]	[0,09-6,64]	[0,51-3,94]	[38,47-153,66]	[33,44-131,53]	[0,13-1,31]	[23,28-93,17]	[0,19-1,23]
Tempo de Emprego (>2 anos)	**86,58**	**8,78**	**5,57**	**210,48**	**83,17**	**1,25**	**107,09**	**5,5**
Intervalo de Confiança	[40,82-183,65]	[2,72-28,41]	[2,34-13,25]	[99,71-444,31]	[39,39-175,61]	[0,47-3,36]	[50,52-227,03]	[2,53-11,95]
Escolaridade (Ens. Fundamental)	**0,65**	**32,74**	**1,48**	**2,08**	**1,58**	**5,1**	**0,81**	**4,54**
Intervalo de Confiança	[0,53-0,79]	[13,27-80,76]	[1,15-1,9]	[1,71-2,52]	[1,31-1,9]	[4,18-6,24]	[0,66-1]	[3,85-5,37]
Escolaridade (Ens.Superior)	**2,36**	**5,83**	**4,06**	**3,7**	**4,31**	**0,74**	**7,68**	**3,07**
Intervalo de Confiança	[1,42-3,91]	[0,47-72,42]	[2,46-6,71]	[2,29-5,98]	[2,68-6,94]	[0,37-1,45]	[4,77-12,37]	[1,93-4,89]

Tabela 6 Probabilidade à Posteriori de pertencer à classe latente atribuída, expressa em média (EP).

Classe Latente	Probabilidade à Posteriori
1	0,93 (0,003)
2	0,87 (0,004)
3	0,80 (0,007)
4	0,83 (0,005)
5	0,83 (0,004)
6	0,96 (0,002)
7	0,93 (0,009)
8	0,88 (0,004)
9	0,94 (0,004)

Nas Tabelas 4 e 5 são apresentadas, respectivamente, as estimativas dos parâmetros da Regressão Logística Multimonial e as estimativas da razão de chances com os respectivos intervalos de confiança (95%).

Quanto à variável renda, na Tabela 5, temos que um aumento de R$100,00 na renda do cliente está associado a um acréscimo de 0,06 na chance de ele pertencer à classe 6 (Excelente) em relação à classe de referência, classe 4 (Péssimo).

Em relação às variáveis *dummies*, pode-se observar que a chance de clientes que possuem mais do que 2 anos no atual emprego (categoria 3 tempo de emprego) pertencerem à classe 6 (Excelente) é 5,5 vezes a chance dentre aqueles clientes que não apresentam informação a respeito do tempo de emprego (categoria 1 tempo de emprego), ajustado para a presença das demais variáveis.

Já em relação à escolaridade, percebe-se que a chance de pertencer à classe 6 (Excelente) entre clientes que possuem ensino superior completo é 3,07 vezes a chance entre clientes que possuem apenas o ensino médio. Enquanto isso, a chance de pertencer à classe 6 (Excelente) entre clientes que possuem apenas ensino fundamental é 4,54 vezes a chance entre clientes que possuem apenas o ensino médio, o que se justifica por ser uma classe composta principalmente por aposentados. O mesmo tipo de interpretação pode ser feito para as demais variáveis.

Por sua vez, a Tabela 6 indica que o modelo ajustado possui grande acurácia na classificação dos indivíduos, uma vez que as probabilidades à posteriori ficaram próximas de 1 para todas as classes.

4.8. Modelo final: estimativas e interpretação da prevalência de cada classe

Após verificar-se que todas as covariáveis testadas se mostraram estatisticamente significativas ($p<0,0001$), foi realizada uma análise descritiva das classes formadas. Neste sentido, a partir das variáveis utilizadas para seu ajuste,

assim como a informação de atraso dos clientes, foi avaliado se foi possível agrupar os clientes em classes que possuam diferentes comportamentos de atraso no pagamento do cartão, conforme objetivo inicial. Abaixo são apresentados os resultados (Figuras 2 e 3).

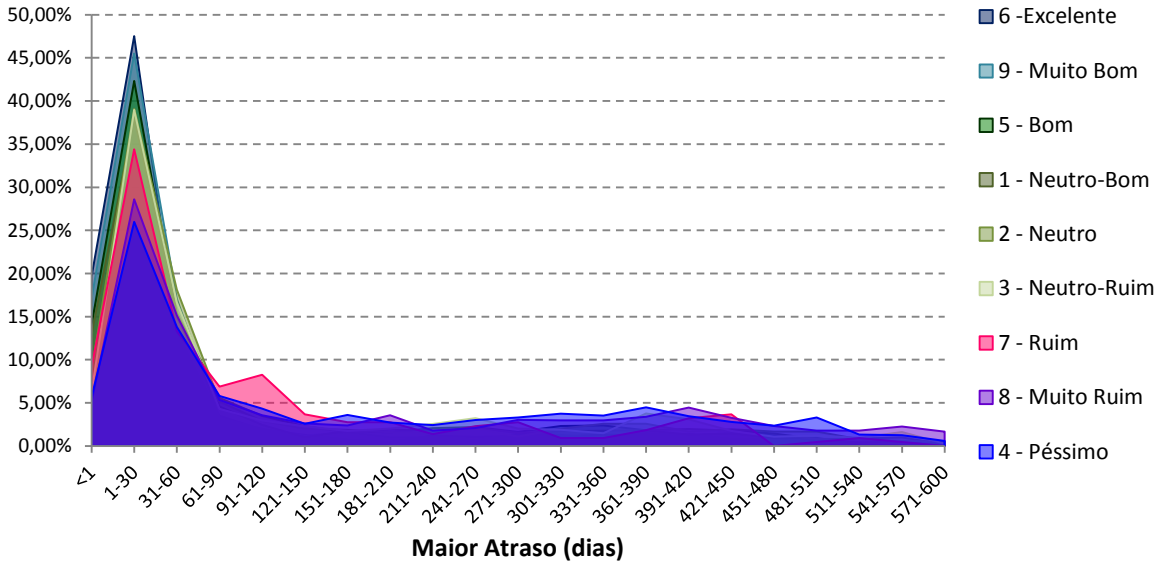


Figura 2 Distribuição do maior dia de atraso por classes com nomenclatura sugerida.

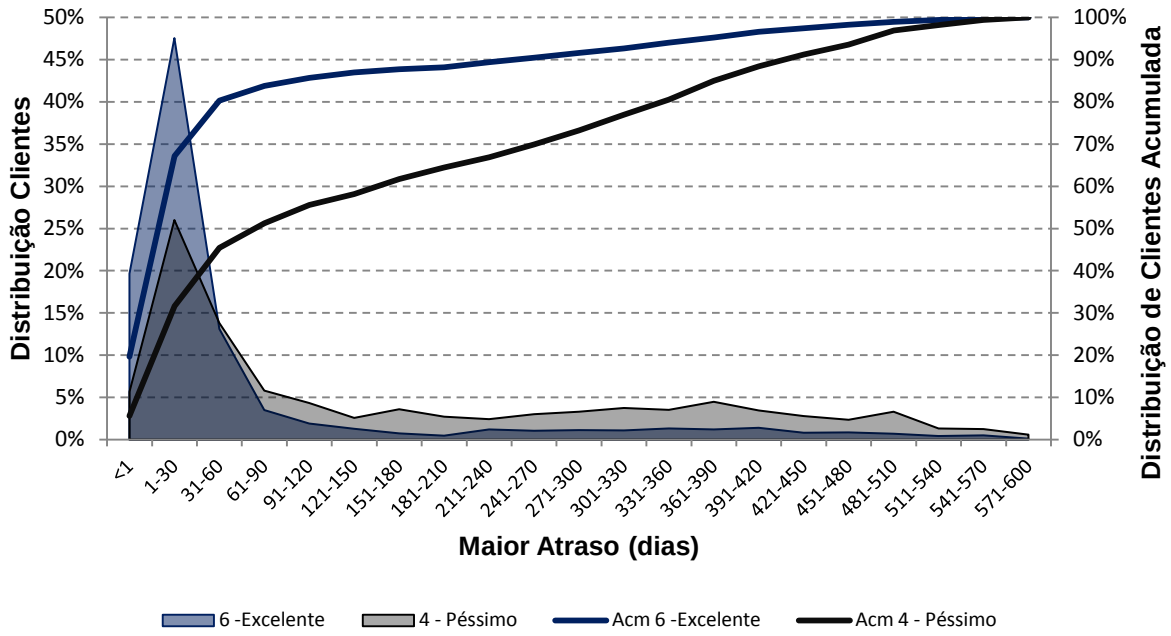


Figura 3 Distribuição do maior dia de atraso por classes com nomenclatura sugerida.

Avaliando a distribuição do maior dia de atraso por classes (Figura 2), observa-se que há diferença no comportamento. Quando se observa apenas as classes extremas (Figura 3), a diferença se torna mais evidente. No gráfico de linhas da Figura 3 verifica-se que na classe 6 - “Excelente” cerca de 20% dos clientes não apresentaram atraso (Maior atraso<1dia), e quase 70% dos casos apresentaram no máximo 30 dias de atraso. Por outro lado, na classe 4 - “Péssimo” menos de 5% dos clientes não apresentaram atraso, sendo que apenas aproximadamente 30% desses clientes apresentaram no máximo 30 dias de atraso. Assim, percebe-se que algumas classes estão mais concentradas em menores dias de atraso (menos arriscadas) do que outras, conforme nomenclatura atribuída.

Tabela 7 Estatísticas descritivas para caracterização da classes. Variáveis categóricas estão expressas em frequência (%) e variáveis contínuas como média (DP).

		Padrão Ouro		Proporção Clientes	Profissão*	CEP COM*	CEP RES*	Cidade Nasc.*	Possui Residência Própria	Estado Civil (n=12257)				
	CLASSE	Ever60	Ever90	(n=12257)	(n=12257)	(n=12257)	(n=12257)	(n=12257)	(n=12257)	Casado	Divorc./ Sep.	Solteiro	Viúvo	Outro
6	Excelente	19,68%	16,21%	20,98%	1,05%	8,40%	38,97%	28,12%	86,50%	32,48%	13,42%	18,24%	32,40%	3,46%
9	Muito Bom	20,26%	17,09%	8,17%	30,54%	1,00%	0,00%	9,28%	53,39%	41,22%	7,49%	43,21%	0,80%	7,29%
5	Bom	28,88%	24,12%	13,03%	84,10%	1,82%	20,66%	22,54%	91,17%	41,58%	11,08%	33,44%	7,70%	6,20%
1	Neutro-Bom	33,51%	27,88%	12,43%	39,00%	0,79%	0,00%	22,06%	66,91%	34,54%	6,83%	47,08%	1,84%	9,72%
2	Neutro	34,11%	28,96%	12,69%	36,14%	89,84%	55,31%	45,47%	86,37%	50,29%	17,04%	24,12%	2,70%	5,85%
3	Neutro-Ruim	38,11%	33,73%	6,07%	38,98%	44,09%	94,89%	72,04%	85,08%	43,28%	15,73%	30,78%	7,26%	2,96%
7	Ruim	43,12%	36,24%	1,78%	49,08%	97,71%	97,71%	82,11%	99,08%	3,21%	0,00%	87,16%	0,00%	9,63%
8	Muito Ruim	51,19%	45,89%	13,75%	44,09%	87,36%	62,61%	72,52%	60,42%	10,09%	0,53%	82,91%	0,12%	6,35%
4	Péssimo	54,94%	49,29%	11,11%	77,02%	4,48%	48,31%	68,87%	56,17%	8,59%	0,00%	86,34%	0,00%	5,07%

* % relativo às 2 categorias mais arriscadas, conforme agrupamento no Apêndice I.

Tabela 8 Estatísticas descritivas para caracterização da classes. Variáveis categóricas estão expressas em frequência (%) e variáveis contínuas como média (DP).

		Ocupação (n=12257)					Grau de Instrução (n=12257)			IDADE	Renda (R$100,00)	Tempo Emprego
	CLASSE	Aposentado	Assalariado	Autônomo	Func. Púb.	Prof. Lib.	Ens. Fundamental	Ens. Médio	Ens. Superior	(n=12257)	(n=12257)	(n=3319)
6	Excelente	95,76%	0,89%	1,21%	0,62%	1,52%	69,66%	27,23%	3,11%	59,48(11,67)	8,25(9,7)	13,95(11,4)
9	Muito Bom	1,10%	73,75%	12,48%	11,28%	1,40%	29,94%	61,78%	8,28%	32,37(10,11)	6,25(3,82)	5,07(6,62)
5	Bom	0,13%	0,69%	98,87%	0,00%	0,31%	76,33%	23,17%	0,50%	45,1(12,64)	9,22(22,28)	17,97(21,64)
1	Neutro-Bom	1,18%	72,88%	19,57%	5,65%	0,72%	43,73%	51,28%	4,99%	33,17(11,06)	6,98(4,83)	3,59(5,3)
2	Neutro	0,77%	80,90%	2,70%	14,28%	1,35%	45,72%	46,82%	7,46%	43,21(9,11)	10,32(11,28)	6,92(7,83)
3	Neutro-Ruim	5,24%	1,34%	88,31%	2,69%	2,42%	38,04%	52,42%	9,54%	43,98(12,35)	10,75(8,59)	8,57(10,07)
7	Ruim	10,55%	0,00%	89,45%	0,00%	0,00%	97,25%	2,75%	0,00%	38,92(16,26)	3,6(1,26)	7,77(7,12)
8	Muito Ruim	0,36%	92,76%	4,63%	1,13%	1,13%	24,93%	72,46%	2,61%	26,29(6,07)	6,19(3,08)	1,99(3,31)
4	Péssimo	3,01%	0,00%	91,92%	0,00%	5,07%	36,71%	61,89%	1,40%	25,93(7,59)	7,18(3,4)	1,55(1,16)

Tabela 9 Estimativas de prevalência para as classes, ajustadas para a presença das covariáveis.

Classe Latente	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Gamma	0,1234	0,1275	0,069	0,1089	0,1277	0,2091	0,0175	0,1377	0,0793
EP	(0,0034)	(0,0041)	(0,0038)	(0,0044)	(0,0045)	(0,0039)	(0,0012)	(0,0042)	(0,0025)

A partir das tabelas 7 e 8 é possível avaliar descritivamente as classes formadas, observando a Tabela 7, é possível verificar que as classes formadas apresentam comportamento de pagamento diferente, verificado pelos indicadores de inadimplência[2]. Na classe 6, de excelente perfil de pagamento, 19,68% dos clientes apresentaram pelo menos um atraso superior a 60 dias durante o ano após a concessão, enquanto isso, dentre os clientes que pertencem à classe 4, péssimo perfil de pagamento, mais da metade (54,94%) dos clientes tiveram pelo menos um atraso acima de 60 dias. A mesma análise pode ser feita para os perfis intermediários, onde a diferença em termos de inadimplência não é tão evidente.

A classe 4 teve sua prevalência estimada em 10,89% (Tabela 9), de forma descritiva foi observado que ela possui péssimos pagadores, é composta por indivíduos em maioria jovens, solteiros (86,34%), com média de aproximadamente 26 (DP=7,59) anos, com pouco tempo no atual emprego, em média 1,55 ano (DP=1,16), sendo que uma grande parcela não possui moradia própria (43,83%). São indivíduos em sua maioria com ensino médio completo (61,89%), trabalham como autônomos (91,92%) e sua profissão é relacionada a mau perfil de pagamento, sendo que 77,02% dos casos possui profissão arriscada em termos de inadimplência, conforme agrupamento prévio (categorias 1 e 2 apresentadas no Apêndice I). 48,31% dos casos possui residência em locais relacionados a mau perfil de pagamento, entretanto o local de trabalho não é relacionado a mau perfil de pagamento (4,48%), enquanto em sua maioria possuem cidade de nascimento relacionada a mau perfil de pagamento (68,87%).

Já quanto à classe 6, cuja prevalência foi estimada em 20,91%, de forma descritiva foi observado que ela possui excelentes pagadores, é composta por indivíduos em maioria com idade mais avançada sendo em média 59,48 (DP=11,67) anos, casados (32,48%) ou viúvos (32,40%). Em sua maioria possuem moradia própria (86,50%). Chama a atenção de que são indivíduos em sua maioria apenas com ensino fundamental completo (69,66%), aposentados (95,76%), apenas 38,97% dos casos possui residência em locais relacionados a mau perfil de pagamento, além de em sua maioria não terem nascido em cidades relacionadas a mau perfil de

[2] Ever60 representa a proporção de casos que apresentaram pelo menos um atraso superior a 60 dias, durante um ano após a concessão do cartão de crédito. Ever90 representa a proporção de casos que apresentaram pelo menos um atraso superior a 90 dias, durante um ano após a concessão do cartão de crédito.

pagamento (28,12%). O mesmo tipo de análise pode ser estendido às demais classes latentes identificadas.

5 Considerações finais

Apesar das limitações existentes, foi possível identificar classes latentes com diferentes comportamentos de pagamento entre os proponentes de cartão de crédito da rede de farmácias do Rio Grande do Sul avaliada neste estudo, a partir de variáveis sabidamente relacionadas à inadimplência segundo a literatura da área. Uma vez que o perfil de inadimplência é um comportamento que pode variar conforme a natureza do produto, tipo de pessoa (física/jurídica) e população alvo, é recomendado que para a utilização da informação das classes de clientes, a empresa desenvolva um modelo interno partindo dessas premissas. A população amostrada neste estudo possui perfil específico com respeito à renda, moradia, escolaridade e demais variáveis avaliadas.

Ao identificar as classes latentes, conhecer suas características e relação com o perfil de inadimplência seria possível a empresa criar estratégias ou políticas diferenciadas para cada grupo. Como exemplo: prospecção de novos clientes de forma direcionada àquele perfil menos arriscado; estratégias de cobrança preventiva direcionadas àqueles indivíduos pertencentes às classes historicamente mais arriscadas; condições diferentes para conjuntos de classes com perfil mais ou menos arriscado; entre outros.

Cabe ressaltar que a base de dados utilizada no presente estudo não foi estruturada especificamente para o tipo de análise utilizada. Além disso, a variável idade, de natureza quantitativa contínua, foi categorizada para ser inserida como item no modelo, levando a uma possível perda de informação contida na amostra. Adicionalmente, parece haver problemas de preenchimento no cadastro da variável tempo de emprego a qual apresentava elevada quantidade de dados faltantes.

Futuramente, seria interessante avaliar o desempenho do modelo para identificar classes de perfil de inadimplência a partir de variáveis previamente pensadas para este fim, além da variável idade ser incluída em sua forma quantitativa como covariável. Além disso, no intuito de tornar o modelo mais simples, reduzir o número de parâmetros, seria interessante substituir a informação de CEP por uma variável de dependência espacial. No presente

estudo tinha-se a particularidade de se tratarem de clientes sem histórico de relacionamento com a empresa, assim, outra sugestão para estudos futuros seria identificar classes de clientes que já tenham esse relacionamento e, portanto, possuam informações mais ricas a respeito do seu comportamento com a instituição.

Portanto, apesar das limitações específicas do banco de dados aplicado, foi possível obter-se um resultado satisfatório ajustando um modelo de classes latentes capaz de segmentar os clientes em classes com diferentes comportamentos de inadimplência. Somado a isso, verificou-se que mesmo o modelo de Classes Latentes sendo complexo, há grande potencial de aplicação a diversas áreas do conhecimento no intuito de identificar subgrupos de indivíduos em relação a uma variável latente de interesse. Por último, em termos de programas disponíveis, o procedimento PROC LCA se mostrou uma opção muito útil, flexível e completa para esses tipos de análise.

Referências Bibliográficas

BHATNAGAR, A.; GHOSE, S. A latent class segmentation analysis of e-shoppers. **Journal Of Business Research,** USA, v. 57, n. 7, p.758-767, jul. 2004. Elsevier BV. DOI: 10.1016/s0148-2963(02)00357-0.

CHAN, T. W.; GOLDTHORPE, J. H. Social stratification and cultural consumption: The visual arts in England. **Poetics,** [s.l.], v. 35, n. 2-3, p.168-190, abr. 2007. Elsevier BV. DOI: 10.1016/j.poetic.2007.05.002.

CLOGG, C. C. The Impact of Sociological Methodology on Statistical Methodology. **Statistical Science**, v. 7, n. 2, p. 183-196, 1992.

CAOUETTE, J. B.; ALTMAN, E. I.; NARAYANAN, P.; NIMMO, R. **Managing credit risk: the great challenge for the global financial markets**. John Wiley & Sons, Inc. 2nd ed., c.12, p.201-221, 2008.

COLLINS, L. M.; LANZA, S. T. **Latent Class and Latent Transition Analysis**: with applications in the Social, Behavioral, and Health Sciences. John Wiley & Sons, Wiley Series in Probability and Statistics, c.3, p.74-75, 2010.

DINIZ, C.; LOUZADA, F. Métodos estatísticos para análise de dados de crédito. **6th Brazilian Conference on Statistical Modelling in Insurance and Finance.** Maresias, SP, 2013.

DUBOIS, B.; CZELLAR, S.; LAURENT, G. Consumer Segments Based on Attitudes Toward Luxury: Empirical Evidence from Twenty Countries. **Marketing Letters**, Netherlands, v. 16, n. 2, p. 115-128, 2005.

EMILIANO, P. C.; VIVANCO, M. J. F.; MENEZES, F. S. Information criteria: How do they behave in different models? **Computational Statistics and Data Analysis**, v.69, p.141-253, 2014.

FERNANDES, G. B.; ARTES, R. Spatial dependence in credit risk and its improvement in credit scoring. **European Journal of Operational Research**, v.249, p.517-524, 2016.

FERREIRA, M. A. M.; CELSO, A. S. S.; NETO, J. E. B. Aplicação do modelo logit binominal na análise do risco de crédito em uma instituição bancária. **Revista de Negócios**, Blumenau, v.17, n.1, p.41-59, 2012.

FLORIANO, E. P.; MÜLLER, I.; FINGER, C. A. G.; SCHNEIDER, P. R. Ajuste e seleção de modelos tradicionais para série temporal de dados de altura de árvores. **Ciência Florestal**, Santa Maria, v.16, n.2, p.177-199, 2006.

GAMA, H. C. **Mensuração e regressão de variáveis latentes contínuas**. 2011. Dissertação (Mestrado em Estatística) – Universidade de Brasília, Brasília, 2011.

GOODMAN, L. A. Latent class analysis: the empirical study of latent types, latent variables and latent structures. Separata de: HAGENAARS, C. A.; MCCUTCHEON, A. L. (Ed.) **Applied latent class analysis**, Cambridge University Press, Cambridge, United Kingdom, p. 3-55, 2002.

HAUGHTON, D.; LEGRAND, P.; WOOLFORD, S. Review of Three Latent Class Cluster Analysis Packages: Latent GOLD, poLCA, and MCLUST. **The American Statistician**, v. 63, n. 1, p. 81-91, 2009.

LANZA, S. T.; COLLINS, L. M.; LEMMON, D. R.; SCHAFER, J. L. PROC LCA: A SAS Procedure for Latent Class Analysis. **Structural Equation Modeling**, v.14, n. 4, p. 671-694, 2007.

LANZA, S. T., DZIAK, J. J., HUANG, L., WAGNER, A., COLLINS, L. M. (2015). *PROC LCA & PROC LTA users' guide* (Version 1.3.2). **University Park: The Methodology Cente**r, Penn State. Disponível em: http://methodology.psu.edu

LAZARSFELD, P. F. **The logical and mathematical foundations of latent structure analysis**. In: STOUFFLER, S. A.; GUTTMAN, L.; SUCHMAN, E. A.; LAZARSFELD, P. F.; STAR, S. A.; CLAUSEN, J. A. (Hrsg.): Studies in social psychology in World War II. Band IV: Measurement and Prediction. Princeton, S. 362-412. 1950.

LIMA, E. M. B. C. **Análise de determinantes da inadimplência (pessoa física) tomadores de crédito: uma abordagem econométrica**. 2004. Dissertação (Mestrado em Economia) – Universidade Federal do Ceará, Fortaleza, 2004.

MIOLA, R. F. **Uso de modelos estatísticos para dados de escore de crédito de uma instituição financeira.** 2013. Dissertação (Mestrado em Engenharia de Produção) – Universidade Estadual Paulista “Julio de Mesquita Filho”, Bauru, SP, 2013.

NYLUND, K. L. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. **Structural Equation Modeling**, v.14, n.4, p.535-569, 2007.

PROC LCA & PROC LTA (Versão 1.3.2) [*Software*]. University Park: The Methodology Center, Penn State. Disponível em: http://methodology.psu.edu, 2015.

RAMASWAMY, V.; DESARBO, W. S.; REIBSTEIN, D. J.; ROBINSON, W. T. Na empirical pooling approach for estimating marketing mix elasticities with pims data. **Marketing Science,** v. 12, n.1, 1993.

SACCO, P.; BUCHOLZ, K. N. K.; SPITZNAGEL, E. L. Alcohol use among older adults in the national epidemiologic survey on alcohol and related conditions: a latent class analysis**. Journal of Studies on Alcohol and Drugs**, Washington University in St. Louis, St. Louis, Missouri, 2009.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v.6, n.2, p.461-464.

SELAU, L. P. R. **Construção de modelos de previsão de risco de crédito**. 2008. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

SILVA, V. R.; PREVIDELLI, I. T. S. Item response theory in the production of indicators of socioeconomic metropolitan region of Maringá, Paraná State, Brazil. **Acta Scientiarum. Technology,** Maringá, v. 34, n. 4, p. 427-431, 2012.

SULLIVAN, P. F.; KESSLER. R. C.; KENDLER, K. S. Latent Class Analysis of Lifetime Depressive Symptoms in the National Comorbidity Survey. **American Journal of Psychiatry**, v. 155, p. 1398-1406, 1998.

SWANSON, S. A.; HORTON, N. J.; CROSBY, R. D.; MICALI, N.; SONNEVILLE, K. R.; EDDY, K.; FIELD, A. E. A Latent Class Analysis to Empirically Describe Eating Disorders Through Developmental Stages. **International Journal of Eating Disorders,** v. 47, n. 7, p. 762-772, 2014.

TEIN, J. Y.; COXE, S.; CHAM, H. Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis. **Structural Equation Modeling**, v.20, n.4, p.640-657, 2013.

VASCONCELLOS, M. S. **Proposta de método para análise de concessões de crédito a pessoas físicas**. 2002. Dissertação (Mestrado em Economia) – Universidade de São Paulo, Faculdade de Economia, Administração e Contabilidade, São Paulo, 2002.

VERMUNT, J. K.; MAGIDSON, J. Latent Class Analysis. Separata de: LEWIS-BECK, M. S.; BRYMAN, A.; LIAO, T. F.(Ed.) **The Sage Encyclopedia of Social Sciences Research Methods,** Thousand Oaks, CA, p. 549-553, 2004.

APÊNDICE I

Agrupamento de variáveis realizado pelo critério de inadimplência Ever 60 12 meses, conforme metodologia proposta por Selau (2008, p.70-71):

AGRUPAMENTOS DE CEP COMERCIAL			
		2 PRIMEIRAS POSIÇÕES	**3 PRIMEIRAS POSIÇÕES**
1	**Péssimo Pagador**	91 PORTO ALEGRE - RS 92 CANOAS - RS	932 SAPUCAIA DO SUL
2	**Mau Pagador**	90 PORTO ALEGRE - RS 94 GRAVATAI - RS	961 PELOTAS 967 SAO JERONIMO
3	**Pagador Neutro**		933 NOVO HAMBURGO 962 RIO GRANDE
4	**Bom Pagador**	95 CAXIAS DO SUL	934 NOVO HAMBURGO 935 NOVO HAMBURGO 938 SAPIRANGA
5	**Excelente Pagador**	97 SANTA MARIA - RS 99 PASSO FUNDO - RS 98 CRUZ ALTA - RS	965 CACHOEIRA DO SUL 937 CAMPO BOM 939 IVOTI 966 RIO PARDO

AGRUPAMENTOS DE CEP RESIDENCIAL			
		2 PRIMEIRAS POSIÇÕES	3 PRIMEIRAS POSIÇÕES
1	Péssimo Pagador	90 PORTO ALEGRE - RS	919 PORTO ALEGRE
2	Mau Pagador		913 PORTO ALEGRE 915 PORTO ALEGRE 917 PORTO ALEGRE 923 PORTO ALEGRE 932 SAPUCAIA DO SUL 940 GRAVATAI 941 GRAVATAI 948 ALVORADA 961 PELOTAS
3	Pagador Neutro		911 PORTO ALEGRE 912 PORTO ALEGRE 924 PORTO ALEGRE 925 GUAIBA 934 NOVO HAMBURGO 944 VIAMAO 949 CACHOEIRINHA 967 SAO JERONIMO
4	Bom Pagador	95 CAXIAS DO SUL	914 PORTO ALEGRE 922 PORTO ALEGRE 931 PORTAO 933 NOVO HAMBURGO 935 NOVO HAMBURGO 936 ESTANCIA VELHA 938 SAPIRANGA 960 PELOTAS 962 RIO GRANDE
5	Excelente Pagador	97 SANTA MARIA - RS 99 PASSO FUNDO - RS 98 CRUZ ALTA - RS	965 CACHOEIRA DO SUL 937 CAMPO BOM 939 IVOTI 966 RIO PARDO 968 STA CRUZ DO SUL

AGRUPAMENTOS DE PROFISSÕES			
1	Péssimo Pagador	ALMOXARIFE BABA GERENTE PINTOR PROMOTOR VENDAS SUPERVISOR AUX COZINHA CABELEIREIRO COZINHEIRO	ELETRICISTA MANICURE MECANICO PORTEIRO RECEPCIONISTA SERVENTE TEC ENFERMAGEM VENDEDOR
2	Mau Pagador	ATENDENTE AUTONOMO AUX ADMINISTRATIVO AUX PRODUCAO	AUX SERVICOS GERAIS PEDREIRO VIGILANTE
3	Pagador Neutro	AUXILIAR CAIXA COMERCIANTE DIARISTA	DOMESTICA INDUSTRIARIO MOTORISTA
4	Bom Pagador	AGRICULTOR BALCONISTA COMERCIARIO COSTUREIRO	OPERADOR PENSIONISTA SECRETARIA
5	Excelente Pagador	AGENTE DO LAR	APOSENTADO PROFESSOR

AGRUPAMENTOS DE CIDADES DE NASCIMENTO			
1	Péssimo Pagador	ESTEIO ALVORADA CRUZ ALTA GRAVATAI	IJUI PORTO ALEGRE RIO GRANDE TRAMANDAI
2	Mau Pagador	CAMAQUA CANOAS NOVO HAMBURGO PASSO FUNDO	SANTO ANGELO SAPUCAIA DO SUL URUGUAIANA
3	Pagador Neutro	CAMPO BOM GUAIBA PELOTAS SANTA CRUZ DO SUL SANTANA DO LIVRAMENTO SANTIAGO SAO FRANCISCO DE PAULA	SAO GABRIEL SAO JERONIMO SAO LEOPOLDO SAO LUIZ GONZAGA SAPIRANGA VIAMAO
4	Bom Pagador	ALEGRETE BAGE CACAPAVA DO SUL CAXIAS DO SUL HORIZONTINA MONTENEGRO	OSORIO PALMEIRA DAS MISSOES RIO PARDO SANTA MARIA SAO BORJA
5	Excelente Pagador	CACHOEIRA DO SUL ENCRUZILHADA DO SUL SANTA ROSA SANTA VITORIA DO PALMAR SAO LOURENCO DO SUL TAQUARA TRIUNFO BUTIA	CANGUCU GIRUA ROLANTE SANTO ANTONIO DA PATRUL SAO SEPE TAPES TORRES TRES DE MAIO

APÊNDICE II

CÓDIGO SAS – PROC LCA

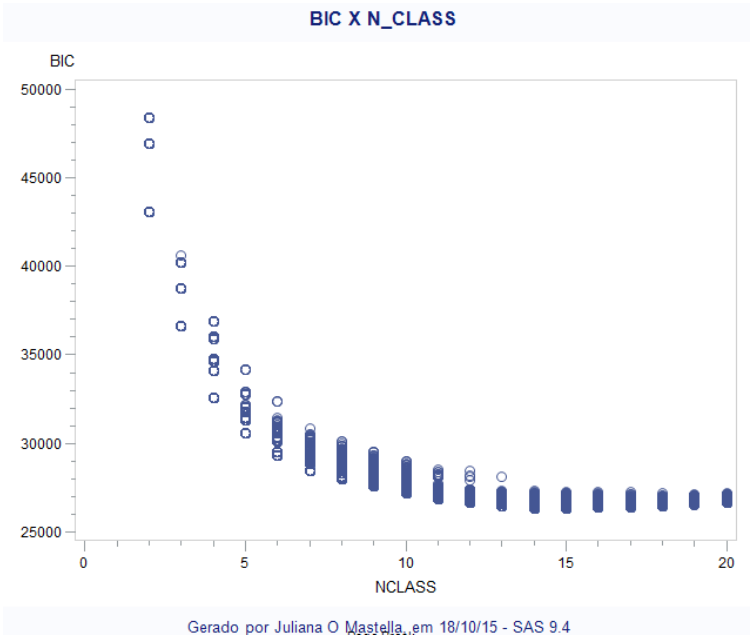
```
/*-------------------------------------------------------------+
|                  Modelo de Classes Latentes Base
|
|   Itens: PROFISSAO CEP_COM CID_NASC CEP_RES ESTADO_CIVIL
|          TP_RESIDENCIA TP_OCUPACAO IDADE
|   Variável de Grupo a ser testada: SEXO
|   Covariáveis a serem testadas: RENDA D1_ESCOLARIDADE
|                                 D3_ESCOLARIDADE
|                                 D2_TEMPO_EMPREGO
|                                 D3_TEMPO_EMPREGO
|
+-------------------------------------------------------------*/

libname LIB "F:\classes latentes";

%let base  = LIB.BD_LCA;
%let itens = PROFISSAO CEP_COM CID_NASC CEP_RES ESTADO_CIVIL
TP_RESIDENCIA TP_OCUPACAO IDADE;

/*-------------------------------------------------------------+
|                   BASELINE MODEL
|
|  Antes de testar grupos e covariáveis, verificar o número
|  ótimo de classes latentes.
|  Macro: baseline_LCA(itens,nc_max,n_rep,base,tab_saida).
|     Onde:
|     itens: Identificar os nomes das variáveis que serão os
|            itens do modelo, separados por espaços.
|     nc_max: Identificar o número máximo de classes latentes
|             a ser testado
|     n_rep: Número de repetições a serem realizadas para
|            cada modelo LCA, de 2 a nc_max classes latentes,
|            a partir de uma semente aleatória.
|     base: Identificar a libname.nome_base_dados que contém
|           os dados a serem analisados
|     tab_saida: identificar a tabela de saída onde serão
|     armazenadas as informações dos modelos ajustados
|     (OUTEST).
|
+-------------------------------------------------------------*/

%macro baseline_LCA(itens,nc_max,n_rep,base,tab_saida);

/* Identificar o número de categorias distintas em cada um
dos itens e armazenar na macrovariável 'cats' */
%global cats;
```

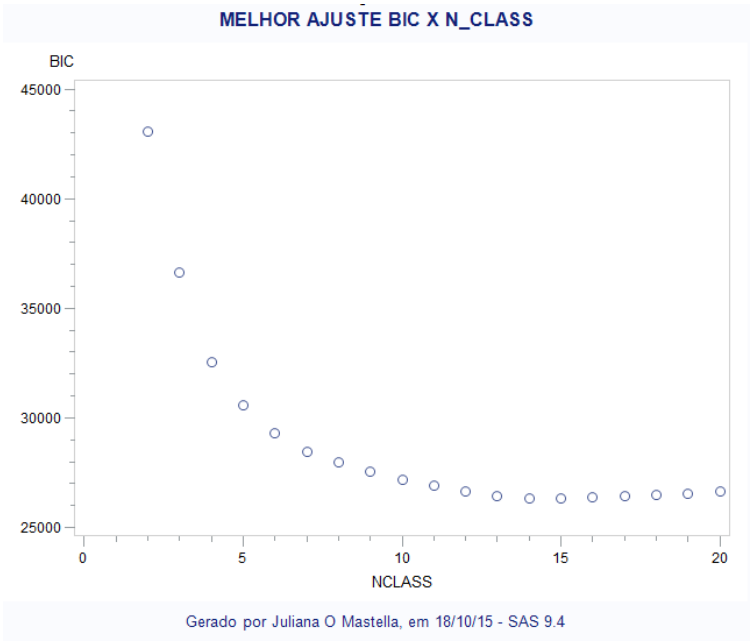
```
/*-------- Inicializar macrovariáveis --------*/
%let i=1;
%let item = %scan(&itens,&i.);
/*----------------------------------------------*/
%do %while (&item ne %str( ));
    proc sql NOPRINT;
        select distinct
        count(&item.) format=1. as tot_cats
        into: cat_
        from (select distinct &item. from &base.);
    quit;
    %if &i = 1 %then
        %do;
            %let cats =&cat_.;
        %end;
    %else
        %do;
            %let cats = %sysfunc(catx(%str(
),&cats.,&cat_.));
        %end;
    %let i=%eval(&i+1);
    %let item = %scan(&itens,&i.);
%end;

%put ------------------------------------------------------;
%put Itens:;
%put &itens;
%put Número de categorias em cada item:;
%put &cats.;
%put Base de Dados utilizada:;
%put &base.;
%put Número Máximo de Classes Latentes a ser testada:;
%put &nc_max;
%put Número de Repetições:;
%put &n_rep.;
%put ------------------------------------------------------;

/* Ajustar 'n_rep' modelos de classes latentes de 2 a
'NC_MAX' classes*/

%DO I=2 %TO &NC_MAX.;
        %DO j=1 %TO &n_rep.;
            %let
seed=%sysfunc(RAND(BINOMIAL,%sysfunc(RAND(UNIFORM)),10000));
            PROC LCA DATA=&base. NOPRINT OUTEST =
STAT_AJU_CLASS(drop=RHO: GAMMA:);
                NCLASS &i.;
                ITEMS &itens;
                RHO PRIOR=1;
                CATEGORIES &cats.;
                ID cod_cli;
```

```
                SEED &seed.;
            RUN;

            DATA STAT_AJU_CLASS;
            SET STAT_AJU_CLASS;
                NCLASS= &i.;
                NREP= &J.;
                SEED = &seed.;
            RUN;

                %IF &I=2 AND &J=1 %THEN %DO;
                    DATA &tab_saida.;SET STAT_AJU_CLASS;RUN;
                    %END;
                %ELSE %DO;
                    DATA &tab_saida.; SET &tab_saida.
STAT_AJU_CLASS;RUN;
                    %END;
            /*Limpa para próxima iteração*/
            proc datasets library=work nolist; delete
STAT_AJU_CLASS;run;
        %END;
%END;

/*CRIA MACROVARIAVEL COM A DATA DO DIA*/
DATA _NULL_;
    CALL SYMPUT('HOJE',PUT(today(),DDMMYY8.));
RUN;

/*  RESULTADOS */

/*retém resultado com menor BIC por classe*/
proc sql;
     create table sort as select
     *,
     min(bic) as minim_bic
     from pd.param_cla_ju
     group by nclass
     having bic = min(bic);
quit;

/*Plotar Gráficos*/
FOOTNOTE1 "Gerado por Juliana O Mastella, em &HOJE. - SAS
9.4";
TITLE1 " BIC X N_CLASS ";

PROC GPLOT DATA=pd.param_cla_ju;
    PLOT BIC * NCLASS;
RUN;

TITLE1 "MELHOR AJUSTE BIC X N_CLASS ";
PROC GPLOT DATA=SORT;
    PLOT BIC * NCLASS;
```

```
RUN;

%mend;

/*baseline_LCA(itens,nc_max,n_rep,base,tab_saida)*/
%baseline_LCA(&itens.,20,1024,&base.,param)
```

SAÍDAS:





```
/*-----------------------------------------------------------+
|                    Modelo de Classes Latentes
|                Condicionado à variável de Grupo
|   Variável de Grupo a ser testada: SEXO
|
|   NCLASS: Número de classes latentes identificado como
|           ótimo na etapa anterior.
|   RHO PRIOR= : opção não obrigatória, identificar uma
|   priori para os parâmetros RHO caso haja interesse. Quando
|   o método apresenta problemas de convergência, uma solução
|   de contorno é utilizar uma priori pouco informativa 1.
|   GROUPS: Identificar a variável de grupo.
|   GROUPNAMES: Identificar a legenda para cada categoria da
|   variável grupo, em ordem crescente do respectivo código
|   da categoria, separados por espaços.
|   SEED: Identificar uma das sementes que gerou o resultado
|   ótimo na etapa anterior.
|
+-----------------------------------------------------------*/

PROC LCA DATA=&base.;
NCLASS 15;
ITEMS &itens;
CATEGORIES &cats.;
GROUPS SEXO;
GROUPNAMES FEMININO MASCULINO ;
SEED 7965;
RUN;
```

```
/*-----------------------------------------------------------+
|                    Modelo de Classes Latentes
|                Ajustado para a presença de Covariáveis
|
|   OUTPOST: Identificar a tabela onde serão salvas as
|            probabilidades a posteriori e classes latentes
|            atribuídas ao indivíduo.
|   OUTPARAM: Identificar a tabela onde serão salvas os
|            parâmetros do modelo ajustado para a presença de
|            covariáveis.
|   ID: Identificar as variáveis que estão na tabela de
|   origem e deverão permanecer na tabela com as
|   probabilidades a posteriori.
|   de contorno é utilizar uma priori pouco informativa 1.
|   COVARIATES: Identificar as covariáveis.
|   REFERENCE: Opção não obrigatória, é possível identificar
|   a classe latente de referência. Caso a opção não seja
|   utilizada, por padrão a classe de referência será a 1.
|
+-----------------------------------------------------------*/


PROC LCA DATA=&base. OUTPOST = pd.post OUTPARAM=pd.parametros_fim;
NCLASS 9;
ITEMS &itens;
CATEGORIES &cats.;
ID COD_CLI BOM_MAU MAIO_ATR ESCOLARIDADE TEMPO_EMPREGO RENDA
IDADE;
COVARIATES &covar.;
REFERENCE 4;
SEED 7965;
RUN;
```

ANEXO I

Tabela 10 Resumo das opções e declarações disponíveis no procedimento PROC LCA:

Sintaxe	Necessário	Descrição
PROC LCA	✓	**Invoca o procedimento.**
Opções		
DATA	✓	**Especifica o arquivo de dados SAS a ser analisado.**
VERBOSE_OUTPUT		Exibe na saída os valores iniciais, restrições aos parâmetros, desvio máximo absoluto e log-verossimilhança a cada iteração.
OUTEST		Salva as estimativas dos parâmetros em um arquivo de dados SAS de um registro/linha.
OUTPARAM		Salva as estimativas dos parâmetros em um arquivo de dados SAS.
OUTPOST		Salva as probabilidades a posteriori a um arquivo de dados SAS.
NOPRINT		Suprime a exibição da saída padrão do procedimento.
START		Permite ao usuário determinar os valores iniciais para os parâmetros.
RESTRICT		Permite ao usuário especificar restrições às probabilidades de resposta aos itens.
NOBETATEST		Suprime os testes de significância para as covariáveis.
Declarações		
NCLASS	✓	**Especifica o número de classes latentes do modelo.**
ITEMS	✓	**Declara as variáveis/itens que serão usadas para ajuste das classes latentes.**
CATEGORIES	✓	**Especifica o número de categorias de cada item.**
ID		Declara a variável identificadora das observações e demais variáveis que se tenha interesse em reter no arquivo de dados SAS que contém as probabilidades à posteriori.
GROUPS		Declara a variável categórica de grupo.
GROUPNAMES		Especifica a legenda para cada grupo.
COVARIATES		Declara as variáveis a serem incluídas como covariáveis.
REFERENCE		Especifica a classe latente a ser usada como referência na presença de covariáveis para ajuste da regressão logística multinomial.
FREQ		Declara variável que contém a frequência de casos com determinado perfil, no caso de utilizar dados agrupados.
ESTIMATION		Especifica o procedimento de estimação.
SEED	✓[a]	**Especifica a semente para o gerador de números aleatórios.**
MEASUREMENT		Invoca a mensuração de invariância entre grupos
MAXITER		Especifica o número máximo de iterações
CRITERION		Especifica o critério de convergência para o desvio máximo absoluto.

[a] A declaração SEED é necessária apenas se a opção START for incluída.
Tradução feita pela autora referente à tabela apresentada por LANZA *et al.* (2007).