Universidade de Brasília
Departamento de Estatística

Uso da regressão logísitica na estimação da probabilidade de reincidência de jovens infratoras

Felipe Gomes Ribeiro

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2015

Felipe Gomes Ribeiro

Uso da regressão logísitica na estimação da probabilidade de reincidência de jovens infratoras

Orientador:
Prof. Dr. **EDUARDO YOSHIO NAKANO**

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2015

AGRADECIMENTOS

Agradeço primeiramente aos meus pais que sempre me deram auxílio incondicional em toda a minha vida, conselhos e apoio afetivo. Sempre buscaram o melhor pra mim para que eu pudesse ter as melhores condições possíveis.

Agradeço também aos meus amigos que fiz na vida, dentro e fora da unb, que sempre que precisei estavam ao meu lado.

SUMÁRIO

RESUMO

Uso da regressão logísitica na estimação da probabilidade de reincidência de jovens infratoras

O trabalho teve como intuito verificar possíveis fatores que levam jovens infratoras a terem uma possível reincidência. Fazendo uma análise do teste qui-quadrado, o trabalho mostra algumas variáveis que se associam com a reincidência e suas intensidades, assim como a associação entre elas e uma possível existência de multicolinearidade. Utilizando a técnica de Regressão Logística, foi feito um modelo logístico que possa calcular a probabilidade de uma jovem infratora reincidir de acordo com algumas de suas características socias e demográficas.

Palavras-chave: Modelos de regressão logístico, reincidência, p-valor, covariáveis, associação

1 INTRODUÇÃO

Por muitos anos, o crime feminino teve pouco destaque devido ao seu número extremamente reduzido, quando comparado ao masculino (Santos, 2013). Segundo Assis e Constantino (2001), o número de crimes cometidos por mulheres e adolescentes do sexo feminino vem crescendo. Se tratando, especialmente, com adolescentes, alguns fatores podem estar relacionados ao cometimento de um ato infracional, que podem ser devido a desorganizações no circulo familiar (como a ausência da figura paterna, rompimento de vínculos familiares e outros), uso de drogas e afastamento da escola (Santos, 2013).

De acordo com o Panorama Nacional de Medidas Socioeducativas de Internação (BRASIL, 2012), 54% dos adolescentes que cumpriam medida de internação em toda região Centro-Oeste no ano de 2012 eram reincidentes no ato criminal. As medidas sócio-educativas aplicadas nos adolescentes que se encontravam em regime de internação deveriam servir para educar esses adolescentes a fim de retornarem à família, escola e sociedade, repensando e mudando seus atos. Entretanto, observa-se uma ausência de acolhimento e assistência adequada para que esses jovens mudem seus comportamento (BRASIL, 2012). Neste contexto, o objetivo deste trabalho tentar identificar quais são os perfis de jovens com maior risco de reincidência. Essas informações poderão guiar um plano mais adequado de ressocialização dessas jovens na sociedade.

Conhecendo melhor o perfil das jovens infratoras que tem casos de reincidência, torna-se importante o estudo para que se possa ao máximo prever as possibilidades de novos casos criminais serem efetuados por elas. Sendo assim, torna-se importante também por ajudar a gerar insumos para a criação de políticas públicas que visam atender melhor e de uma maneira diferente cada infratora de acordo com suas características.

O objetivo desse estudo é a criação de um modelo estatístico utilizando as técnicas de regressão logística que possa representar as probabilidades de uma nova ocorrência criminal de acordo com suas características sociais e demográficas e, se possível, identificar fatores de riscos que levam à reincidência.

2 MODELO DE REGRESSÃO LOGÍSTICA

Este projeto consiste em modelar a probabilidade de uma jovem infratora cometer novamente um novo delito, de acordo com suas características sócio-demográficas, relacionamento com familiares, uso de drogas e o tipo do ato infracional cometido. A análise será realizada utilizando um modelo de Regressão Logística (Hosmer e Lemeshow, 2000) considerando a reincidência (ou não) como variável resposta. Como algumas covariáveis são qualitativas nominais com muitos níveis, as mesmas serão inicialmente agrupadas segundo suas razões de chances (odds-ratios) individuais. Neste contexto, esse trabalho prevê uma revisão das técnicas de regressão logística, sobretudo métodos de seleção de variáveis, testes de ajuste do modelo e diagnóstico de multicolinearidade. Todas as análises serão realizadas através dos softwares SPSS e R (R Core Team, 2013).

Nos modelos de regressão linear simples ou múltipla, a variável resposta Y é uma variável aleatória contínua com disrtibuição normal. Porém, muitas vezes essa variável resposta tem uma natureza qualitativa, assumindo dois ou mais valores. Quando ocorre essa situação, com a variável resposta sendo uma variável qualitativa, o uso da regressão logística torna possível o uso de um modelo de regressão para que se possa prever a probabilidade de ocorrência de um certo evento. O interesse de qualquer análise de regressão é o valor médio da variável resposta de acordo com o valor da variável explicativa, denotado por $E(Y|x)$. Nesse trabalho, a resposta ficará apenas entre duas categorias, e com isso $E(Y|x)$ é uma proporção, isto é, $0<E(Y|x)<1$. Para variável resposta dicotômica existe duas categorias que poderemos chamar de "sucesso", com probabilidade igual a π, e "fracasso", com probabilidade igual a 1 - π, e que mostraremos mais pra frente.

Primeiramente, antes de se falar da regressão logística, temos que falar um pouco sobre Modelos Lineares Generalizados.

Segundo Nery (2015), os modelos lineares generalizados foram sugeridos quando a variável dependente Y pode ser expressa por alguma distribuição da família exponencial. Esses modelos são especificados por três componentes: uma componente aleatória, a qual identifica a distribuição de probabilidade da variável dependente; uma componente sistemática, que especifica uma função linear entre as variáveis independentes e uma função de ligação que descreve a relação matemática entre a componente sistemática e

o valor esperado da componente aleatória.

Assim como no caso desse estudo, quando a variável resposta Y for uma variável qualitativa e que assume apenas dois valores, seguindo uma distribuição Bernoulli, ela pertence a uma das duas classes importantes de modelos linear generalizados que é constituido pelos modelos *logit* e é aqui que se enquadra a regressão logística.

A distribuição Bernoulli se relaciona com a regressão logística binária pois essa regressão busca representar apenas o "sucesso"e o "fracasso"da população, podendo ser expressas pelos valores 1 e 0, respectivamente.

Sendo Y_i a variável resposta da i-ésima observação, i= 1, 2,..,n, temos que o valor da probabilidade de sucesso (Y=1) e fracasso (Y=0) são expressas respectivamente por:

- $P(Y_i = 1) = \pi$
- $P(Y_i = 0) = 1 - \pi$

Com média $E(Y_i) = \pi$ e $Var(Y_i) = \pi(1 - \pi)$

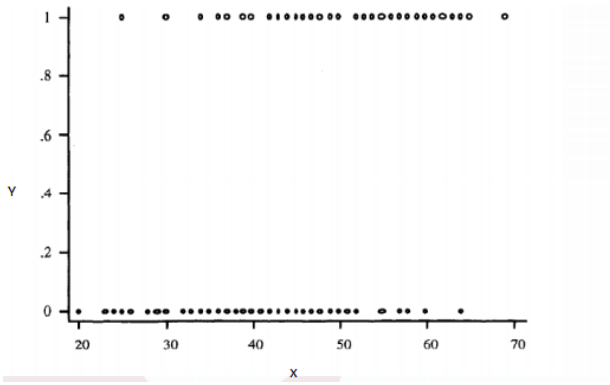

Figura 1 – Exemplo gráfico da variável resposta de uma regressão logística

As principais propriedades da regressão logística são:

1. Pode ser linearizada;

2. É quase linear no intervalo de crescimento e nas extremidades aproxima-se gradualmente de 0 e 1;

3. A função logística é monótona (crescente ou decrescente, que vai de acordo com o sinal

dos parâmetros que estão associados às variáveis explicativas).

2.1 Relação com a distribuição Bernoulli

Para uma variável resposta Y binária e um vetor de covariáveis x, o valor de π(x) é a probabilidade de sucesso. Com isso, os valores de Y seguem uma distribuição de probabilidade Bernoulli com parâmetro π(x).

$$Y \sim Bernoulli(\pi(x)).$$

Os valores das variáveis explicativas formam um vetor X = (x_1, x_2,...,x_k). Esses valores são usados para que possamos chegar a um valor da variável dependente Y. Como π(x)= E(Y|x) assume valores apenas entre 0 e 1, uma representação linear para π sobre todos os valores de x possíveis não é adequada, e para isso considera-se a transformação logística de π(x) sob a forma linear, ao qual chamaremos de logito:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

em que β_0 é o intercepto e β_j é o coeficiente da covariável x_j, j = 1, 2,..., k.

A fórmula também pode ser escrita isolando-se π(x), e fica da seguinte maneira:

$$P(Y = 1|x) = E(Y|x) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}$$

2.2 Estimação dos Parâmetros

Na regressão linear, os parâmetros do modelo são estimados através do método dos mínimos quadrados. Na regressão logística, isso não é possível e, com isso, os parâmetros do modelo são estimados através do método de máxima verossimilhança. Este método maximiza o logaritmo da função de verossimilhança. Seu uso somente é possibilitado pelo fato de

conhecermos a distribuição que está associada a variável resposta binária, que é a distribuição Bernoulli, que tem a seguinte forma:

$$P(Y = y) = \pi(x)^y(1 - \pi(x))^{1-y}, y = 0, 1.$$

Seja Y_i a resposta do i-ésimo indivíduo da amostra com vetor de covariáveis $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$. A função de verossimilhança é dada por:

$$L(\beta|Y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

em que

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki}}}{1 + e^{\beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki}}}$$

e β_0, β_1,..., β_k são os coeficientes do modelo a serem estimados.

Os estimadores de máxima verossimilhança é um vetor de valores dos parâmetros que maximiza a função de verossimilhança $L(\beta|Y)$, e para a sua obtenção derivamos a função em relação a cada parâmetro e igualamos ao valor zero. Porém, como essas equações são não-lineares nos parâmetros, resolvemos isso aplicando métodos numéricos, como o Newton-Raphson.

2.3 Interpretação dos Parâmetros

A interpretação dos parâmetros do modelo de regressão logística é obtida através da comparação entre a probabilidade de sucesso e a probabilidade de fracasso. Para isso, é usada a função "*odds ratio*"- OR (Razão de chances). Essa comparação consiste na razão da probabilidade de sucesso sobre a probabilidade de fracasso *(ODDS)*.
No caso de uma regressão linear simples, temos:

$$Odds1 = \frac{P[Y=1|X=x]}{P[Y=0|X=x]} = \frac{\frac{e^{\beta_0+\beta_1*x_1}}{1+e^{\beta_0+\beta_1*x_1}}}{1-\frac{e^{\beta_0+\beta_1*x_1}}{1+e^{\beta_0+\beta_1*x_1}}} = e^{\beta_0+\beta_1 x}$$

Ao tomarmos uma variação unitária da variável explicativa X, o Odds2 é:

$$Odds2 = \frac{P[Y=1|X=x+1}{P[Y=2|X=x+1} = e^{\beta_0+\beta_1(x+1)}$$

Para calcularmos a razão de chance, dividimos o Odds2 em relação ao Odds1:

$$OR = Odds2/Odds1 = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1}} = e^{\beta_1}$$

Ao substituir o valor de β_1 na exponencial, iremos analisar o valor obtido, e se ele for maior que 1, a chance de sucesso tende a aumentar com a aumento da variável explicativa.

2.3.1 Intervalo de confiança

Estimado os valores dos parâmetros do modelo através do método de máxima verossimilhança, o intervalo de (1-α) x 100% de confiança para esses parâmetros é dado por:

$$\widehat{\beta_j} \pm z_{1-\frac{\alpha}{2}} DP(\widehat{\beta_j}) = \widehat{\beta_j} \pm z_{1-\frac{\alpha}{2}} \sqrt{var(\widehat{\beta_j})}$$

Onde $\widehat{\beta_j}$ é o valor do estimador de máxima verossimilhança do parâmetro β_j; o valor $z_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ de uma Normal padrão; e a $var(\widehat{\beta_j})$ é a estimativa da variância de $\widehat{\beta_j}$, obtida pela matriz de informação de Fisher observada.

2.4 Seleção de variáveis do modelo Logístico

Para a seleção de variáveis serão usados três métodos que são baseados na checagem da importância de cada variável, incluindo ou excluindo-as do modelo.

1. *Forward*: Neste método as variáveis são adicionais sucessivamente no modelo, ou seja, partimos da suposição que não há variável no modelo, contém apenas o intercepto. Adicionamos as variáveis no modelo até que a inclusão de uma variável não seja mais significativa para o modelo, e assim, interrompe-se o processo. O critério de adição de variáveis no modelo é a análise do p-valor do teste de significância de cada uma delas. Em cada etapa de adição de variáveis ao modelo, o componente que tiver o menor p-valor obtido quando se testam as significâncias das variáveis é incluído no modelo. A comparação é feita com um nível de significância α pré determinado.

2. *Backward*: Nesse processo, a variável de pior desempenho pode ser eliminada, desde que não atenta a outros critérios mínimos exigidos. O método backward se baseia na retirada do componente com maior p-valor obtido quando se testam as suas significâncias. Partindo do modelo completo, com todas as variáveis dentro dele, o processo de retirada se encerra quando não houver mais nenhuma variável com p-valor maior que um nível de significância α pré determinado.

3. *Stepwise*: Este procedimento é uma forma similar do *Forward*, pois se inicia com o modelo mais simples, apenas com o intercepto. Após cada etapa de incorporação de uma variável, temos uma etapa em que uma das variáveis já selecionadas pode ser descartada, e por isso é considerado o método mais iterativo entre esses três. O primeiro processo é o cálculo do p-valor de cada variável fora do modelo e após essa etapa é incluido no modelo o componente de menor p-valor. Com a variável no modelo, a próxima etapa é a analise simultânea de inclusão de um novo componente e a exclusão do que já esta dentro do modelo, semelhante ao método *Backward*. O processo é repetido até chegar a um modelo quando não existir mais nenhuma variável para ser adicionada o retirada do modelo.

O método de seleção de variáveis *Stepwise* requer algumas considerações:

• O modelo final depende do valor do nível de significância pré determinado;

• Algumas variáveis são extremamente importantes para o modelo, porém pode ocorrer que sejam excluidas dele, e isso ocorre devido à multicolinearidade entre as variáveis.

2.5 Qualidade de Ajuste do Modelo

Para verificar a qualidade do ajuste do modelo de regressão logística de resposta binária, o teste de Hosmer e Lemeshow é a forma mais usual para isso. O teste consiste basicamente em avaliar o modelo comparando com as frequências observadas e esperadas, propondo dois tipos de agrupamento que se baseiam nas probabilidades estimadas. Inicialmente, classificamos as observações e os eventos de probabilidade são estimados. As observações são divididas em cerca de 10 grupos. Seja N o número total de indivíduos e M o número alvo de indivíduos que podemos calcular da seguinte forma:

$$M = [0,1 * N + 0,5]$$

Para que a estatística de Hosmer - Lemeshow possa ser determinada, é necessário que existam pelo menos 3 grupos. A estatística proposta segue uma distribuição Qui-quadrado quando não há replicação em qualquer uma das subpopulações. A estatística do teste:

$$H = \sum_{g=1}^{G} \frac{(O_g - N_g)^2}{N_g(1 - \pi_g)}$$

Onde,

• N_g é a frequência total de indivíduos no g-ésimo grupo;

• O_g é a frequência total de resultados de evento no g-ésimo grupo;

• π_g é a probabilidade média estimada prevista de um resultado de eventos para o g-ésimo grupo, com g=1,2,...,G

Achado o valor da estatística do teste, compara-se com uma distribuição Qui-quadrada com (G-2) graus de liberdade. Caso o valor da estatística do teste seja menor que o p-valor, isso indica que o modelo proposto pode explicar bem o que se observa.

2.6 Validação do modelo

Para que possamos saber se o modelo reflete a realidade, ou seja, se podemos usa-lo para outras observações é necessário que passemos por um certo processo. Assim como na regressão linear, aqui na regressão logística o procedimento de validação é semelhante. Existe algumas maneiras para a validação do modelo como:

1. Dividir a observação em dois grupos, não necessariamente de tamanhos iguais, para trabalharmos com uma dessas subamostra para estimar os parâmetros do modelo e a outra subamostra para validar os parâmetros e verificar se podemos realmente usar o modelo para outras observações.

2. O processo conhecido como Leave-one-out é feito quando retira-se uma única observação da amostra e ajusta-se o modelo com as n-1 observações que restaram. O valor retirado é predito pelo modelo ajustado. A seguir, a observação retirada é devolvida na amostra e passa-se a retirar uma outra observac ao da amostra, ajustando o modelo e fazendo a previsao com os n-1 valores restantes. O processo e repetido ate que todas as observações da amostra sejam retiradas (e preditas).

2.7 Multicolinearidade

O uso dos modelos de regressão logística dependem direta ou indiretamente das estimativas dos seus coeficientes. Porém, a presença de multicolinearidade pode ocasionar problemas no ajuste do modelo por causar impactos nas estimativas dos parâmetros do modelo. O problema da multiolinearidade existe quando há uma dependência linear exata ou aproximada entre as covariáveis do modelo. Dependendo do nível de associação entre as varáveis independentes, a estimação dos parâmetros pode ficar imprecisa.

Para a testar a existência de multicolinearidade no modelo, pode-se usar o VIF (Fator de Inflação da Variância). O VIF mede o quanto a variância do coeficiente

$$VIF = \frac{1}{1 - R^2}$$

onde R^2 é o coeficiente de determinação ao se fazer uma regressão usando a covariável j como resposta e as demais covariáveis com variáveis explicativas.

2.7.1 Diagnóstico da Multicolinearidade

Segundo Neter et al.(2005), as seguintes técnicas informais de diagnóstico podem ser utilizados para avaliar indícios de multicolinearidade:

• Ocorrem grandes mudanças nas estimativas dos coeficientes quando variáveis são adicionadas ou excluídas do modelo;

• Resultados dos testes individuais para os coeficientes do modelo de importantes variáveis preditoras são não significativas;

• Duas ou mais covariáveis independentes são correlacionadas.

3 DESCRIÇÃO DA APLICAÇÃO

3.1 Banco de Dados

O banco de dados do estudo realizado tem como origem o Trabalho de Conclusão de Curso feito por Samantha Lima dos Santos (Santos, 2013), aluna do curso de Terapia Ocupacional da Universidade de Brasília. No seu trabalho foram feitas várias análises para se saber mais sobre o perfil de adolescentes do sexo feminino que estão em conflito com as leis no Distrito Federal, levando-se em conta o uso de drogas, a violência e a reincidência. O resultado do trabalho de conclusao de curso da Samantha gerou algumas variáveis: o uso de drogas, ato infracional, idade, reincidência, estuda, série, familia usa drogas, tem filho ou grávida, mora com quem, e cidade onde mora.

3.2 Ajuste das Variáveis

Para essa análise, foram utilizadas apenas 7 das variáveis mencionadas, que são: uso de drogas, idade, reincidência, estuda, família usa drogas, tem filho ou grávida e mora com quem. Dessas sete, seis são chamadas de variáveis explicativas, e uma é chamada de variável dependente, que é o casa da reincidência.

Tabela 4.1 - Variáveis a serem analisadas

V.A	Descrição
Y	Reincidência
X1	Usa drogas
X2	Idade
X3	Estuda
X4	Mora com quem
X5	Família usa drogas
X6	Filho ou grávida

As variáveis excluídas foram consideradas variáveis sem grandes significâncias em relação à variável resposta, a reincidência.

Algumas medidas foram tomadas com o objetivo de deixá-las cada uma agrupada em 2 categorias. Duas dessas variáveis eram quantitativas e, com isso, foi necessário

categorizá-las. A variável idade tinha um intervalo de 12 anos até os 18 anos, e após a categorização, ficou agrupada em até 16 anos e maiores de 16 anos. A outra variável quantitativa que sofreu uma categorização foi a variável resposta. Inicialmente era contabilizada o número de reincidências que a jovem tinha, e após a categorização foram agrupadas em 2 grupos: com reincidência, e sem reincidência.

Houve também um caso em que se existia 3 categorias que foi reduzida a duas, que é o caso da variável mora com quem. Na origem do estudo, os valores pertenciam a 3 grupos: sozinho, parentes, e outros familiares. Sendo assim, reduziu-se a 2 o número de categorias desta variável, que passaram a ficar entre sozinho ou genitores/outros familiares.

A amostra contou com 284 jovens infratoras que estiveram internadas entre os anos de 2004 e 2011 na unidade de internação do Plano Piloto (Brasília/DF).

4 RESULTADOS

A análise para a obtenção dos resultados do banco de dados foi feita com o auxilio do software estatístico SPSS e R.

4.1 Teste Qui-Quadrado

Muitas vezes queremos saber o grau de associação entre duas variáveis. Quando essas variáveis são classificadas com variáveis qualitativas, utilizamos o teste qui-quadrado para sabermos se existe ou não uma associação entre duas variáveis. Ele é um teste não paramétrico, o que significa que não depende de parâmetros populacionais, como a variância e a média.

O teste consiste basicamente em verificar as possíveis divergências entre as freqüências esperadas e observadas de um certo evento.

A análise é feita a partir dessas divergências. Quanto menor for essa divergência, podemos dizer que os dois grupos se comportam de maneira bem semelhante. Para testar essa divergência, trabalhamos com duas hipóteses:

• Hipótese nula, onde as freqüências esperadas e observadas não são diferentes, ou seja, não há associação entre as variáveis.

• Hipótese alternativa, onde as freqüências esperadas e observadas são diferentes, e portanto há associação entre as variáveis.

4.1.1 Associação das variáveis explicativas com a resposta

Para uma melhor análise do estudo, decidiu-se inicialmente avaliar a associação existente entre a variável resposta e cada variável explicativa separadamente.

Pode-se observar através da Tabela 5.1 que as variáveis usa droga, idade e estuda estão relacionadas à variável reincidência tanto com um nível de significância (α) igual a 0,05 quanto igual a 0,10. Já as variáveis família usa droga e filho ou grávida são as variáveis que não apresentam associação significativa com a variável resposta em nenhum dos casos. Já a variável mora com quem relaciona-se com a reincidência apenas quando o valor do nível de significância é igual a 0,10. Todas as análises de associação das variáveis explicativas com a a variável resposta foram feitas levando em consideração um nível de significância igual a

0,05 e 0,10. A Tabela 5.1 também contém as frequências cruzadas de cada variável com a variável resposta (reincidência). A análise dessas frequências permite verificar a existência de uma associação entre as variáveis explicativas com a variável resposta, o que é confirmada através da análise do P-valor.

Tabela 5.1 - Frequências cruzadas e associações em relação à resposta

		Reincidencia		P-valor (X^2)	alfa = 0,10	alfa = 0,05
		Sim	Não			
Usa drogas	Sim	58 (38,7%)	92 (61,3%)	0	Há associação	Há associação
	Não	99 (73,9%)	35 (26,1%)			
Idade	Até 16 anos	111 (62%)	68 (38%)	0,003	Há associação	Há associação
	Maior que 16 anos	46 (43,8%)	59 (56,2%)			
Estuda	Sim	63 (63,6%)	36 (36,4%)	0,038	Há associação	Há associação
	Não	94 (50,8%)	91 (49,2%)			
Familia usa drogas	Sim	42 (60,9%)	27 (39,1%)	0,283	Não há	Não há
	Não	115 (53,5%)	100 (46,5%)			
Filho ou grávida	Sim	36 (47,4%)	40 (52,6%)	0,103	Não há	Não há
	Não	120 (58,3%)	86 (41,7%)			
Mora com quem	Sozinho	41 (47,7%)	45 (52,3%)	0,089	Há associação	Não há
	Genitores/outros familiares	116 (58,6%)	82 (41,4%)			

Essa associação pode ser melhor visualizada através do diagrama apresentado pela Figura 2 que mostra, de acordo com a espessura da linha e a sua respectiva ausência, a existencia ou não de uma associação entre uma determinada variável explicativa com a reincidência.

A linha mais espessa e sólida evidencia uma associação com um nível de significância igual a 0,05. Já a linha mais fraca e pontilhada, no caso da variável mora com quem, mostra a existência de uma associação apenas com um nível de significância igual a 0,10. A falta de uma linha indica ausência de associação em ambos os casos, como visto nas variáveis família usa drogas e filho ou grávida.

Figura 2 – Diagrama de associação entre as variáveis explicativas e a resposta. A linha sólida representa uma associação ao nível de 5% e uma linha pontilhada ao nível de 10%

4.1.2 Associação entre as variáveis explicativas

É de interesse da regressão a análise da associação entre as variáveis explicativas. O diagrama abaixo mostra as relações existentes entre cada uma das 6 variáveis explicativas analisadas. A análise dessas associações foram feitas considerando um nível de significância de 0,05.

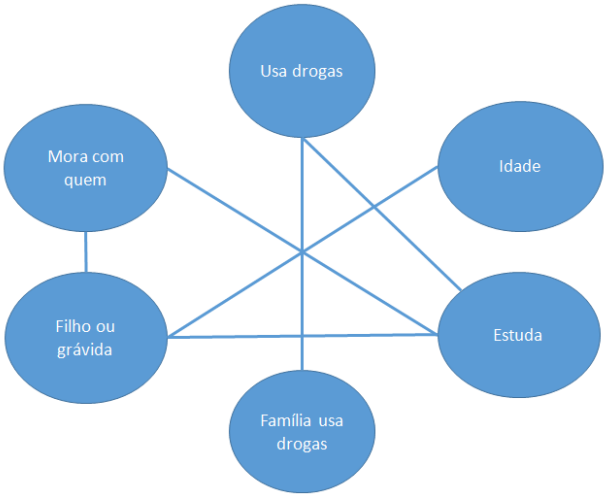

Figura 3 – Diagrama de associação entre as variáveis explicativas

Pelo diagrama, notamos que algumas variáveis apresentam mais associações do que outras, variando de 3 a 1 associação. Usa drogas é uma variável que se associa com estuda e família usa drogas. Já idade se associa apenas com filho ou grávida. A variável estuda é

uma das que mais tem associação entre elas, associando-se com usa drogas, mora com quem e filho ou grávida. A única variável que se associa à familia usa drogas é a variável usa drogas. Assim como estuda, filho ou grávida tem 3 associações, com idade, mora com quem e estuda. Mora com quem se associa com filho ou grávida e estuda.

4.2 Modelo estatístico

Sem usar algum método específico de seleção de variáveis, as quatro variáveis explicativas que apresentavam associação com a resposta (Figura 2, Tabela 5.1) a um nível de significância de 10% foram incluídas no modelo estatistico e assim foram obtidas suas respectivas estimativas dos parâmetros. Portanto, as variáveis família usa drogas e filho ou grávida não participaram do modelo final. Ademais, vimos que essas duas vairáveis não auxiliam muito o percentual de acerto do modelo final.

Os quadros abaixo mostram respectivamente o percentual de acerto do modelo com as variáveis família usa drogas e filho ou grávida, e sem as duas variáveis.

Tabela 5.3 - Quadro de acerto com família usa drogas e filho ou grávida

Observado	Esperado Y=0	Esperado Y=1	Acertos(%)
Y=0	109	48	69,4
Y=1	45	82	64,6
Acerto Geral (%)			67,3

Tabela 5.3 - Quadro de acerto sem família usa drogas e filho ou grávida

Observado	Esperado Y=0	Esperado Y=1	Acertos(%)
Y=0	109	48	69,4
Y=1	45	82	64,6
Acerto Geral (%)			67,3

Além disso, no estudo vimos (Figura 3) que há existência de multicolinearidade nessas duas variáveis, além de que, as mesmas não auxiliam significamente na explicação do modelo de regressão logística e também não estão associadas à resposta, como pode ser visto na Tabela 5.5.

Sendo assim, foi decidido retirá-las da modelagem, visto que, é mais vantajoso para o modelo suas respectivas ausências (resultando um modelo mais simples).

Assim sendo, com as quatro variáveis explicativas restantes, foram feitos alguns testes de seleção de variáveis e em todos os casos o modelo era formado apenas por duas variáveis explicativas: uso de drogas e idade.

Porém, de acordo com o estudo feito pela Samantha, a reincidência apresenta uma significância estatística quanto a evasão escolar e ao fato de morar sozinha. A variável mora com quem se torna importante para o modelo a partir da conclusão que a família é um fator essencial de proteção à adolescente já que a família tem forte influência nas decisões da jovem. A variável estuda tem como importância no modelo pelo fato da evasão escolar ser um fator de risco para a reincidencia, já que é na escola que se há informação e instrução para a vida. Sendo assim, não foi utilizado nenhum método de seleção de variáveis, apenas a inclusão das quatro analisadas.

O teste de Hosmer e Lemeshow teve significância de 0,881, o que pressupõe que o modelo logístico considerando as 4 variáveis está bem ajustado.

4.2.1 Estimativas dos parâmetros do modelo

Excluídas as variáveis família usa drogas e filho ou grávida, foram selecionadas para a modelagem apenas as quatro variáveis restantes: uso de drogas, idade, estuda e mora com quem.

Tabela 5.4 - Variáveis do modelo logístico

V.A	Descrição
Y	Reincidência
X1	Usa drogas
X2	Idade
X3	Estuda
X4	Mora com quem

Temos que a probabilidade de reincidencia criminal das adolescentes de acordo com cada variável é estimada por

$$\widehat{P}(Y=1|x) = \widehat{\pi(x)} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\ldots+\widehat{\beta_4}x_4}}{1+e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\ldots+\widehat{\beta_4}x_4}}$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$ e $\widehat{\beta_4}$ são apresentados pela Tabela 5.5.

Tabela 5.5 - Estimativas dos parâmetros e p-valores

		Univariada				Multivariada			
		β()	Exp β	(IC 95%) Exp β	p	β()	Exp β	(IC 95%) Exp β	p
Usa drogas	Sim	1,501	4,487	(2,704 , 7,446)	0,000	1,439	4,216	(2,508 , 7,087)	0,000
	Não	0	0	0	0	0	0	0	0
Idade	Até 16 anos	0	0	0	0	0	0	0	0
	Maior que 16 anos	0,739	2,094	(1,283 , 3,416)	0,003	0,68	1,975	(1,168 , 3,340)	0,011
Estuda	Sim	0	0	0	0	0	0	0	0
	Não	0,527	1,694	(1,027 , 2,795)	0,039	0,224	1,251	(0,720 , 2,172)	0,427
Mora com quem	Sozinho	0,44	1,553	(0,933 , 2,583)	0,09	0,267	1,307	(0,747 , 2,287)	0,349
	Genitores/outros familiares	0	0	0	0	0	0	0	0
Familia usa drogas	Sim	-0,302	0,739	(0,425 , 1,285)	0,284				
	Não	0	0	0	0				
Filho ou grávida	Sim	0	0	0	0				
	Não	-0,439	0,645	(0,380 , 1,094)	0,104				
Intercepto						-1,49	0,225		0,000

Assim, os valores positivos das estimativas dos parâmetros fazem com que a probabilidade de reincidência seja maior. Ao mesmo tempo, os valores negativos das estimativas dos parâmetros fazem com que a probabilidade de reincidência diminua.

Tomando como exemplo uma jovem de 17 anos, usuária de drogas, que estuda e mora sozinha, a probabilidade dela ter uma reincidência é de:

$$\widehat{P}(Y=1|x) = \frac{e^{-1,49+0,68+1,439+0+0,267}}{1+e^{-1,49+0,68+1,439+0+0,267}} = 0,71.$$

Através da análise dos valores das estimativas dos parâmetros, vemos a im-

portância de cada variável explicativa no modelo para que a reincidencia ocorra, e com isso podemos concluir que:

- O uso de drogas é o fator que mais aumenta a probabilidade de reincidencia
- A variável estuda é a que menos impacta na reincidencia

A Tabela 5.5 mostra também qual seria o valor de cada parâmetro se o modelo fosse formado por apenas uma variável, e comparando com o modelo multivariado formado, vimos uma certa diferença em seus parâmetros. Isso reforça mais ainda a existencia de multicolinearidade entre as variáveis. Um exemplo seria x_5 que não tem muito influência em relação a Y, porém pode ser uma variável que tem grande influência em relação a x_1, que é a variável mais influente em relação à reincidencia.

5 CONSIDERAÇÕES FINAIS

O uso da regressão logística foi eficiente na construção de um modelo que consiga estimar a probabilidade de uma jovem cometer um novo ato criminal. O modelo logístico apresentou um bom ajuste nos dados, alcançando uma taxa de 70% de classificação correta. O modelo final contou com quatro variáveis explicativas (idade, usa drogas, estuda e mora com quem), dentre as seis inicialmente consideradas. Devido à existência de multicolinearidade entre essas quatro variáveis, as mesmas não poderem ser identificadas como fatores de risco para reincidência. No entanto, o modelo é útil para estimar a probabilidade de uma jovem infratora reincidir. Os resultados obtidos mostraram que o fato da família da jovem usar ou não drogas não influenciou na reincidência. Assim como o fato da jovem ter filho ou estar gravida também não teve efeito significativo na reincidência. O uso de drogas e a idade foram os fatores que tiveram muita influência na reincidência. Ao contrário do estudo de Santos (2003), estudo e morar sozinho ou com genitores não apresentaram efeito significativo neste estudo. Essa diferença possivelmente ocorreu devido ao tipo de análise realizada. O trabalho de Santos (2013) considerou como resposta o número de reincidências e o modelou através de um modelo linear generalizado Binomial Negativo. Neste trabalho o número de reincidência foi categorizado permitindo o uso de uma regressão logística, que apresenta uma melhor (e mais fácil) interpretação dos resultados.

REFERÊNCIAS

Agresti, Alan. An Introduction to Categorical Data Analysis, Jonh Wiley Sons, New York, 2edition, 2007.

ASSIS, S.G.; CONSTANTINO, P. Filhas do mundo:infração juvenil feminina no Rio de Janeiro. Rio de Janeiro: Fiocruz, 2001.

BARRETO, ALEXANDRE SERRA., Modelos de regressão: teoria e com o programa estatstico R, Ed. do autor, Brasilia, 2011.

BRASIL. Panorama Nacional de Execução de Medidas Socioeducativas. Conselho Nacional de Justiça. Brasília, DF, 2012.

Hosmer, D. W., and Lemeshow, S. (2000). Applied Logistic Regression, Second Edition. Wiley, New York.

Nery (2015). Risco de concessão de crédito bancário para empresas: Uma aplicação dos modelos de regressão logística. Bacharelado em Estatística. Universidade de Brasília.

Neter et al.(2005) John Neter, Michel Kutner, William Wasserman e Christopher Nachtsheim. Applied Linear Statistical Models. McGraw-Hill Higher Education. Citado na pag. 22.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject.org/.

Santos (2013). Perfil de adolescents do sexo feminine em conflito com a lei no Distrito Federal: violência, uso de drogas e reincidência. Monografia de graduação. Bacharelado em Terapia Ocupacional. Universidade de Brasília. 42f.