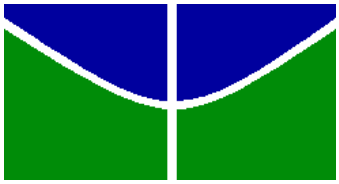


Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

VANESSA QUEIROZ BASTOS

JUVENTUDE E EDUCAÇÃO NA ÁREA METROPOLITANA DE BRASÍLIA

Brasília

2013

VANESSA QUEIROZ BASTOS

JUVENTUDE E EDUCAÇÃO NA ÁREA METROPOLITANA DE BRASÍLIA

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientadora: Profª Dra. Ana Maria Nogales

Brasília

2013

Resumo

O trabalho tem como objetivo analisar o perfil dos jovens da Área Metropolitana de Brasília , tendo como base os diferentes níveis educacionais, relacionando esses níveis com diferentes características sócio-econômicas. A faixa etária utilizada é a de 18 a 24 anos, idade em se espera que os jovens estejam consolidando sua formação educacional. O banco de dados utilizado é a amostra de pessoas do Censo Demográfico de 2010 e o método estatístico adotado é a Análise de Regressão Logística.

Sumário

1 Introdução

Não há como negar que as taxas de escolarização no Brasil têm melhorado nos últimos anos. A porcentagem da população analfabeta diminuiu, ainda que timidamente (segundo dados das PNAD de 2009 e 2011, a taxa de analfabetismo das pessoas de 15 anos ou mais caiu de 9,7 para 8,6), a taxa de frequência à escola aumentou. Em relação aos jovens de 18 a 24 anos, as políticas públicas relacionadas à educação provocaram um considerável aumento de jovens no ensino superior – a taxa de frequência na educação superior mais do que dobrou no período 1996-2007.

Um problema que continua, porém, é a desigualdade de acesso à educação entre os jovens de família de baixa renda em relação aos jovens de família de renda mais alta. Uma pesquisa realizada pelo IPEA mostra que 50% dos universitários ultrapassam o nível de escolaridade dos pais apenas ao ingressar no ensino superior. Porém, essa mesma pesquisa mostra que o nível de escolaridade dos pais está altamente relacionado com o acesso dos filhos ao nível superior. Os dados do Censo 2010 mostram que a maior faixa de concentração de pessoas com mais de 25 anos situa-se entre aquelas sem instrução ou com ensino fundamental incompleto, e a pesquisa mostra que apenas 10% dos pais dos alunos na universidade situam-se nessa faixa de instrução. Isso nos leva a crer que os jovens provenientes de famílias com maior escolaridade têm maior acesso ao ensino superior.

O INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) realiza anualmente o Censo da Educação Superior, que coleta informações sobre Instituições de Educação Superior, cursos de graduação e os alunos e professores desses cursos e tem por finalidade retratar a educação superior brasileira. Pelo censo da educação superior 2010 pôde-

se constatar significativo aumento do atendimento na educação superior: A **taxa de escolarização bruta**[1]**,** por exemplo, passou de 15,1% para 26,7%.

Tabela 1 - Número médio de anos de estudo[2] para a faixa etária de 18 a 24 anos

	Ano		
Subgrupos	**2001**	**2005**	**2009**
1º quartil da renda	7,4	8,4	9,2
Campo	5,1	6,4	7,5
Região Nordeste	6,3	7,4	8,5
Negros (pretos e pardos)	6,8	8,0	8,7
Brancos	8,8	9,6	10,2
Média Nacional	7,9	8,8	9,4

Fonte: Pnad/IBGE; elaborado por MEC/Inep

Por meio da tabela anterior, extraída do relatório do Censo do Ensino Superior 2010, realizado pelo INEP, podemos notar que por mais que tenha diminuído, ainda há considerável diferença em relação a quantidade de anos de estudos, se comparando a média nacional com pessoas do campo, negros, e a população mais pobre. Apenas os anos de estudos dos brancos estão acima da média nacional.

Outro fator bastante relacionando à renda é o ingresso precoce no mercado de trabalho. As vezes, quando a renda familiar é muito baixa, o jovem começa a procurar emprego mais , e acaba ficando mais complicado conciliar os estudos com o trabalho. Além disso, sem o devido preparo técnico, fica difícil conseguir um emprego com boa remuneração.

[1] **Taxa de escolarização bruta** consiste em um indicador que permite comparar o total de matrículas de determinado nível de ensino com a população na faixa etária teoricamente adequada a esse nível.

[2] **Anos de estudo** - Segundo o INEP, " a classificação segundo os anos de estudo foi obtida em função da série e do grau que a pessoa estava freqüentando ou haviam freqüentado, considerando a última série concluída com aprovação. A correspondência foi feita de forma que cada série concluída com aprovação correspondeu a 1 ano de estudo. A contagem dos anos de estudo teve início em 1 ano, a partir da lª série concluída com aprovação de curso de 1º grau ou elementar; em 5 anos de estudo, a partir da lª série concluída com aprovação de curso de médio 1º ciclo; em 9 anos de estudo, a partir da 1º série concluída com aprovação de curso de 2º grau ou de médio 2º ciclo; em 12 anos de estudo, a partir da 1º série concluída com aprovação de curso superior. As pessoas que não declararam a série e *o* grau ou com informações incompletas ou que não permitissem a sua classificação foram reunidas no grupo de anos de estudo 'não determinados ou sem declaração'".

E quando se larga os estudos para focar no emprego, essa dificuldade tende a se perpetuar ao longo da vida, o que acaba impedindo uma ascensão social.

Ao estudarmos o Brasil, podemos constatar uma diferença considerável das regiões mais pobres em relação as outras. Nas mais pobres geralmente existem as mais altas taxas de analfabetismo e maior índice de evasão escolar. É de se esperar que, dentro da área metropolitana de Brasília, ao compararmos as cidades satélites mais pobres com as RA's que apresentam maior renda, como Brasília, Lago Sul, Lago Norte, Sudoeste/Octogonal, haja também considerável discrepância nos níveis educacionais.

2 Metodologia

Foram utilizados no trabalho, dados da amostra do Censo Demográfico de 2010. O Censo Demográfico é uma operação realizada pelo IBGE, onde são investigadas características de toda a população no território nacional. Em todas as residências do país é aplicado um questionário básico, e naquelas selecionadas na amostra, além das perguntas contidas no questionário básico, são investigadas outras informações sociais, econômicas e demográficas consideradas importantes. Todos os setores censitários estão contidos na amostra, mas a seleção dos domícilios é feita por uma amostragem sistemática.

Consideramos população jovem as pessoas de 18 a 24 anos (VASCONCELOS; CESAR; COSTA. 2013).

O local considerado no estudo é a Área Metropolitana de Brasília, AMB, (PAVIANI et al., 2010) e é formada por todo o DF e os municípios goianos de Águas Lindas de Goiás,

Cidade Ocidental, Cristalina, Formosa, Luziânia, Novo Gama, Padre Bernardo, Planaltina, Santo Antônio do Descoberto, e Valparaíso de Goiás.

Em relação a característica 'trabalho', foram consideradas as quatro variáveis presentes no Censo:

V0641 - Trabalhou ganhando em dinheiro, produtos, mercadorias ou benefícios

V0642 - Tinha trabalho remunerado do qual estava temporariamente afastado(a)

V0643 – Ajudou, sem qualquer pagamento, no trabalho remunerado de algum morador do domicílio

V0644 - Trabalhou na plantação, criação de animais ou pesca, somente para alimentação dos moradores do domicílio (Inclusive caça e extração vegetal)

Um jovem que teve qualquer uma dessas respostas positivas foi considerado um jovem que trabalha.

Os jovens foram separados em 4 regiões. Três dentro do DF, sendo a região 1 a mais rica (Plano Piloto, Sudoeste/Octogonal, Lago Norte, Lago Sul) a 2, intermediária (Cruzeiro, Candangolândia, Núcleo Bandeirante, Guará, Gama, Taguatinga, Águas Claras, Vicente Pires, Riacho Fundo I e São Sebastião) e a 3, a mais pobre (Brazlândia, Ceilândia, Itapoã, Planaltina, Santa Maria, Recanto das Emas, Riacho Fundo II, Samambaia e áreas rurais). A quarta região agrupa os municípios goianos que fazem parte da AMB que, em geral, tem renda ainda menor do que a 3ª região do DF (VASCONCELOS; CESAR; COSTA, 2013).[3]

Para visualizar as diferentes situações educacionais dos jovens, foram montados 3 grupos com diferentes níveis de ensino. O grupo 1 contempla os jovens que não terminaram a

[3] Varjão está inserido no Lago Norte, Vila Estrutural e Setor de Industria no Guará, Jardim Botânico distribuído entre São Sebastião e Paranoá, Sobradinho II em Sobradinho e Park Way no Núcleo Bandeirante.

educação básica[4] (Lei nº 9394/1996) , agrupando dos sem instrução até os com ensino médio incompleto. Quem não tinha largado os estudos mas apresentava grande defasagem também foi incluído nesse grupo, na suposição de que a probabilidade de eles concluírem o ensino médio é baixa. O grupo 2 abrange quem completou o ensino médio e parou de estudar e os que chegaram no ensino superior mas largaram os estudos . Os jovens que estavam frequentando o ensino médio regular também foram incluídos nesse grupo, com exceção dos de 18 anos cursando o 3º ano do ensino médio. O grupo 3 contém os jovens que estavam cursando o ensino superior e os que já o haviam concluído, além dos jovens com 18 anos que frequentavam o 3º ano do ensino médio, que não apresentavam defasagem.

A renda considerada no estudo foi a variável **V6531** do Censo, que é o rendimento *per capita* domiciliar em reais.

[4] **Educação básica** : Segundo a Lei de Diretrizes e Bases da Educação brasileira (Lei nº 9394 de 20 de dezembro de 1996), "O dever do Estado com educação escolar pública será efetivado mediante a garantia de: a educação básica obrigatória e gratuita dos 4 (quatro) aos 17 (dezessete) anos de idade organizada na forma: a) pré-escola; b) ensino fundamental; c) ensino médio".

3 Regressão Logística

Métodos de regressão buscam estabelecer relações entre uma variável resposta e variáveis explicativas. A principal diferença entre a regressão logística e a regressão linear, é que na logística a variável é categórica, podendo assumir *k* valores.

3.1 Modelo Logístico Binário

O problema chave é encontrar o valor de *Y* quando para *X's* diferentes. Essa quantidade é chamada de esperança condicional e é expresa por *E(Y|x)*, onde *Y* é a variável resposta e *x* é o valor da variável independente. Na regressão linear, essa esperança é expressa como :

$$E(Y|x) = \beta_0 + \beta_1 x$$

Por meio dessa expressão, a esperança condicional de *Y* dado *x* pode variar de $-\infty$ a $+\infty$. Quando a variável *Y* é binária, a esperança condicional deve estar contida no intervalo [0,1]

Na regressão logística binária, a variável Y pode assumir os valores 0 ou 1, e sua distribuição é uma bernoulli especificada pelas probabilidades

$$P(Y = 1 \,|X = x) = \pi(x)$$

E

$$P(Y = 0 \,|X = x) = 1 - \pi(x)$$

Essa distribuição tem média $\pi(x)$ e variância $\pi(x)[1 - \pi(x)]$

Consideramos

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

E usamos a transformação linear, que chamaremos de transformação logito :

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x$$

Essa transformação é importante porque g(x) tem propriedades desejadas do modelo de regressão linear.

O erro é dado pelo $\boldsymbol{\varepsilon}$ da expressão $\boldsymbol{\varepsilon} = y - \pi(x)$, que é a diferença entre o valor real e o estimado. No caso de um modelo binário, se $y = 1$, $\varepsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$, $\varepsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$. O erro terá média 0 e variância $\pi(x)[1 - \pi(x)]$, ou seja, terá com distribuição binomial com probabilidade $\pi(x)$, enquanto que na regressão linear ele segue uma distribuição normal.

É de suma importância que estimemos valores para os parâmetros β_0 e β_1. Enquanto na regressão linear normalmente utilizados o método de mínimos quadrados para estimação desses parâmetros, geralmente na regressão logística usa-se o método da máxima verossimilhança.

Método de máxima verossimilhança para estimação dos parâmetros

Antes de mais nada, tem-se a *função de verossimilhança* e os estimadores para os parâmetros são aqueles que maximizam essa função. No caso binário, $\pi(x)$ nos dá $P(Y = 1|x)$ e $1 - \pi(x)$, $P(Y = 0|x)$. Então os pares (x_i, y_i) onde $y_i = 1$, a contribuem com $\pi(x)$ para a função de verossimilhança, e quando $y_i = 0$, com $1 - \pi(x)$. Então, a função de verossimilhança é dada por :

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

Comumente se utiliza o log da função de verossimilhança,

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{ y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]\}$$

Para encontrarmos os valores de $\boldsymbol{\beta}$ que maximiza essa função, derivamos em relação a β_0 e a β_1, e igualamos os resultados a zero, resultando nas seguintes equações, respectivamente

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

e

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0$$

As equações descritas acima são não lineares e, por isso, necessitam de métodos especiais iterativos de estimação. McCullagh e Nelder (1989) mostraram que a solução pode ser obtida usando um processo iterativo de mínimos quadrados ponderados.

Os valores de **β** das equações acima são os estimadores de máxima verossimilhança, $\widehat{\boldsymbol{\beta}}$.

Teste de Significância para os parâmetros

Uma questão importante é : o modelo que inclui a variável em questão nos diz mais sobre a variável resposta do que o modelo sem essa variável?

Se o modelo contendo a variável ajusta melhor os dados do que sem a variável, chamamos essa variável de significante.

Na regressão logística, a comparação entre os valores preditos e os valores observados é baseada na função de log verissomilhança. A comparação é baseada na estatística *D* dada por

$$D = -2ln\left[\frac{verossimilhança\ do\ modelo\ ajustado}{verossimilhança\ do\ modelo\ saturado}\right]$$

A estatística D é comumente chamada de *deviance* (desvio) e é muito importante na avaliação de ajuste do modelo. Ela pode ser reescrita da forma

$$D = -2ln[verossimilhança\ do\ modelo\ ajustado]$$

Para avaliarmos a significância de uma variável, temos:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável)$$

Que pode ser expressa por :

$$G = -2ln\left[\frac{verossimilhança\ sem\ a\ variável}{verossimilhança\ com\ a\ variável}\right]$$

Para o simples caso de apenas uma variável independente, a estatística G segue uma χ^2 com 1 grau de liberdade sob a hipótese nula de β,=0. Se $P[\chi^2(1) > G] < \alpha$, onde α é o nível de significância determinado, então a variável é significativa. Essa teste é chamado de razão de máxima verossimilhança.

Outros testes equivalentes são o *Teste de Wald* e o *Score Test*

O teste de Wald é o mais comum nos softwares estatísticos. Sob a hipótese nula de que o parâmetro é zero, a estatística a seguir segue uma distribuição normal

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$$

Onde $\hat{\beta}_1$ é a estimativa para β_1 por máxima verossimilhança, e $\widehat{SE}(\hat{\beta}_1)$ é a estimativa do erro padrão desse coeficiente.

E o p-valor do teste bilateral normalmente é apresentado nas saídas dos softwares.

Score Test

Tanto o teste da razão de verossimilhança quanto o teste de Wald requerem o cálculo do estimador de máxima verossimilhança de β_1. O Score Test não necessita precisa desse cálculo, sendo esse o fato de maior importância do estimador. Ele é dado por:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Para amostras grandes, os três têm resultados parecidos. Porém, Hauck e Domer (1977) e Jennings (1986) estudaram a performance desses testes e indicam que o teste de razão de verossimilhança é mais adequado do que o Score Test e do teste de Wald.

Intervalos de Confiança

O intervalo de confiança para os parâmetros é baseado na estatística do teste de Wald para sua significância. Os limites de um intervalo de 100(1-α)% de confiança são

$$\hat{\beta}_i \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_i)$$

Onde $z_{1-\alpha/2}$ é o ponto da distribuição normal , tal que $\mathrm{P}(\mathrm{Z} > \mathrm{z}) = \mathrm{P}(\mathrm{Z} < -\mathrm{z}) = \alpha/2$

O intervalo de confiança para o logito é dado por

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\widehat{Var}[\hat{g}(x)] = \widehat{Var}(\hat{\beta}_0) + x^2\widehat{Var}(\hat{\beta}_1) + 2x\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

Os limites do intervalo de confinça para o logito são dados por

$$\hat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]$$

Onde $\widehat{SE}[\hat{g}(x)]$ é a raiz do estimador da variância $\widehat{Var}[\hat{g}(x)]$.

Para achar a estimativa para $\hat{\pi}(x)$, substituímos o valor encontrando de $\hat{g}$(x) na equação

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

E, analogamente, podemos encontrar os limites do intervalo de confiança para $\hat{\pi}(x)$ fazendo

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]}}{1+e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]}}$$

O valor para $\hat{\pi}(x)$ deve ser interpretado como o valor médio da proporção de Y=1 numa população com a característica x (lembrando que $\pi(x) = P(Y = 1|x)$) . x pode ser, por exemplo, idade, sexo, cor, etc.

3.2 Odds Ratio (Razão de Chances)

O primeiro passo seria ver qual função da variável dependente é uma função linear das variáveis independentes. McCullagh e Belder (1983) e Dobson (1990) chamaram essa função de '*link function*', e no caso da regressão logística seria o logito g(x).

No caso de apenas uma variável independente, onde vimos que o logito é dado por

$$g(x) = \beta_0 + \beta_1 x\,, \qquad \beta_1 = g(x+1) - g(x)$$

Ou seja, esse coeficiente representa a mudança no logito para a mudança de 1 unidade na variável x.

Para o caso em que a variável independente é dicotômica, temos que

$$\beta_1 = g(1) - g(0)$$

Para interpretarmos esses resultados mais claramente, é importante definir um termo muito importante na regressão logística, chamado *odds ratio.*

O *odds* (chance) de y=1 estar presente em indivíduos com a caracterítisca x=1 é definido como

$$\frac{\pi(1)}{1-\pi(1)}$$

Ou seja, é a razão entra a probabilidade de ocorrer y=1 dado que x=1, sobre a probabilidade de ocorrer y=0, dado que x=1. Da mesma forma, o *odds* para quando x=0 é dado por

$$\frac{\pi(0)}{1-\pi(0)}$$

O *odds ratio* (que chamaremos de OR) é definido como a razão entre o *odds* de x=1 e o de x=0, ou seja, é dado pela equação

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Lembrando que $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$, temos que

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

O *odds ratio* é uma medida que dá quanto mais chance existe de y=1 ocorrer entre os indivíduos com a característica x=1 do que com a característica x=0. É muito utilizado em epidemiologia, para comparar, por exemplo, a incidência de uma doença em um grupo controle com um grupo que toma determinado medicamento. Quando OR= 1, as duas variáveis são independentes. Quanto mais essa razão se afasta de um, maior a dependência.

O Risco Relativo, dado pela razão $\frac{\pi(1)}{\pi(0)}$, se aproxima do *odds ratio* quando $\pi(x)$ é pequeno tanto para x=1 quanto para x=0, ou seja, quando y=1 é um evento raro.

Para grandes amostras, a distribuição do *odds ratio* estimado ($\widehat{OR}$) é normal, mas nem sempre esse requisito é satisfeito nos estudos. Por isso geralmente as inferências são baseadas na distribuição de $\ln(\widehat{OR}) = \hat{\beta}_1$, que segue uma distribuição normal para amostras menores.

Um intervalo de confiança a 100(1-α)% para o *odds ratio* é obtido calculando-se os limites do intervalo de confiança para $\hat{\beta}_1$, e depois exponencializando os valores pra esses limites.

$$\exp[\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)]$$

Caso a variável com dois níveis que não sejam 0 e 1, há algumas diferenças. Consideremos que os nívels são **a** e **b;**

$$\widehat{OR}(a,b) = e^{[\hat{\beta}_1(a-b)]}$$

E

$$\widehat{OR}(a,b) = \frac{\hat{\pi}(x=a)/[1-\hat{\pi}(x=a)]}{\hat{\pi}(x=b)/[1-\hat{\pi}(x=b)]}$$

E os limites do intervalo de confiança

$$\exp[\hat{\beta}_1(a-b) \pm z_{1-\alpha/2}|a-b|\widehat{SE}(\hat{\beta}_1)$$

Porém, na maioria das situações, é recomendado que adote-se 0 ou 1, para facilitar as análises.

3.3 Modelo Logístico Múltiplo

A regressão logística múltipla usa, no geral, as mesmas idéias da regressão logística binária, só que neste caso consideramos $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$ e os vetores $\boldsymbol{x}' = (1, x_1, x_2, \dots, x_p)$ e $\boldsymbol{\beta}' = (\beta_0, \beta_{1,} \dots, \beta_p)$ **.**

O logito, no caso da regressão múltipla, é dado por:

$$g(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Da mesma forma,

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

Quando uma variável independente é qualitativa ordinal (categórica), ou seja, existem diferentes níveis da variável, criamos variáveis indicadoras (*dummys)* para incluí-las no modelo. Se uma variável x_i tem *k* níveis diferentes, são necessárias *k-1 dummys.* Por exemplo, se existem 3 níveis para uma variável, criamos D_1 *e* D_2 e fazemos $D_1 = 0$ *e* $D_2 = 0$ quando a variável está no nível 1, $D_1 = 1$ *e* $D_2 = 0$ quando está no nível 2 e $D_1 = 0$ *e* $D_2 = 1$ quando está no nível 3. O modelo onde a j-ésima variável independente é uma variável categórica seria

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k-1} \beta_{jl} D_l + \beta_p x_p$$

O método de estimação de variâncias e covariancias dos coeficientes estimados é obtido através ca teoria de estimação da máxima verossimilhança (Rao, 1973), que diz que esses estimadores são obtidos da matriz de derivadas parciais da função de log verossimilhança. Essas derivadas parciais tem a seguinte forma geral:

$$\frac{d^2 L(\boldsymbol{\beta})}{d{\beta^2}_j} = -\sum_{i=1}^{n} {x^2}_{ij} \pi(x_i)[1 - \pi(x_i)]$$

E

$$\frac{d^2 L(\boldsymbol{\beta})}{d\beta_j d\beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi(x_i)[1 - \pi(x_i)]$$

Para j=*0,1,2,....,p* e l=*0,1,2,.....p.* Seja a matriz (p+1)×(p+1) que contém o negativo dos termos dados pelas equações acima denonimada $\boldsymbol{I}(\boldsymbol{\beta})$. Essa matriz é chamada de *matriz de informação observada.* As variâncias e as covariâncias dos coeficientes estimados são obtidos pelo inverso dessa matriz, $Var(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$**.** Usaremos a notação $Var(\beta_j)$ para denotar o j-ésimo elemento da diagonal da matriz, que é a variância de $\hat{\beta}_j$ e $Cov(\beta_j, \beta_l)$ que é a covariância entre $\hat{\beta}_j$ e $\hat{\beta}_l$.

Uma formulação importante da matriz de informação é $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$

E **X** é a matriz n×(p+1) que contem os dados, que é escrita como

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

E a matriz **V** é uma matriz n× n é dada por

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

Teste de significância para o modelo

Da mesma forma que no modelo binário, o teste de significância para o modelo é baseado na estatística *G* citada anteriormente

$$G = -2ln\left[\frac{verossimilhança\ sem\ a\ variável}{verossimilhança\ com\ a\ variável}\right]$$

Só que neste caso, o numerador da da razão é a verossimilhança do modelo contendo todas as variáveis. Para testarmos a hipótese de que os *p* coeficientes para as covariáveis no modelo são iguais a zero, a estatística *G* segue uma χ^2 com *p* graus de liberdade. Rejeitamos a hipótese nula quando $P[\chi^2(p) > G] < \alpha$ -onde α é nosso nível de significância- e concluímos que pelo menos um dos *p* coeficientes é diferente de zero.

A estatística para testar a significância de cada coeficiente separamente ainda é a estatística de Wald, citada anteriormente. Caso alguma das variáveis não seja significativa pelo teste de Wald, nós testamos a diferença entre o modelo contendo essas variáveis não

significativas e o modelo inicial. A G terá distribuição χ^2, sob H_0, com grau de liberdade igual ao número de variáveis não significativas que estamos excluindo (m).

Se $P[\chi^2(m) > G] > \alpha$, concluimos que o modelo reduzido é tão bom quanto o modelo anterior. Caso contrário, pelo menos uma das variáveis excluídas seria considerada importante para o modelo.

Intervalos de Confiança

Um estimador da variância do logito é dado por:

$$\widehat{Var}[g(\boldsymbol{x})] = \sum_{i=0}^{p} x_i^2 \widehat{Var}(\hat{\beta}_i) + \sum_{i=0}^{p} \sum_{k=i+1}^{p} 2x_i\, x_k \widehat{Cov}(\hat{\beta}_i, \hat{\beta}_k)$$

E a variância para o vetor $\widehat{\boldsymbol{\beta}}$ é

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

Considerando essa variância, então o estimador da variância do logito por ser reescrito como

$$\widehat{Var}[\hat{g}(\boldsymbol{x})] = \boldsymbol{x}' Var(\widehat{\boldsymbol{\beta}})\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}$$

Analogamente ao caso binário, o intervalo de confiança para os coeficientes é dado por

$$\hat{\beta}_i \pm z_{1-\alpha/2} \widehat{SE}(\hat{\beta}_i)$$

E para o logito

$$\hat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(\boldsymbol{x})]$$

Variável independente politômica

Quando a variável independente em questão tem mais de 2 níveis, a interpretação do *odds ratio* fica um pouco mais complexa. Calculamos os *odds ratio* relacionando cada grupo com um grupo que definimos de referência. O grupo de referência é indicado por um valor de 1 para o *odds ratio*. Por exemplo : *odds ratio* para raça, comparando negros com brancos e hispanicos também com brancos.

Os intervalos de confiança para o *odds ratio* são dados do jeito que foi visto anteriormente:

$$\exp[\hat{\beta}_j \pm z_{1-\alpha/_2}\widehat{SE}(\hat{\beta}_j)]$$

Variável Independente Contínua

O log *odds ratio* para uma mudança de c unidades em x é obtido pela diferença

$$g(x+c) - g(x) = c\beta_1$$

E o *odds ratio* é dado por

$$\widehat{\mathrm{OR}}(\mathrm{c}) = \exp(c\hat{\beta}_1)$$

O erro padrão de $c\hat{\beta}_1$ é obtido multiplicando-se o erro padrão de $\hat{\beta}_1$ pela constante c, logo os limites pro intervalo de confiança de $\widehat{\mathrm{OR}}(\mathrm{c})$ são dados por:

$$\exp[c\hat{\beta}_1 \pm z_{1-\alpha/2} c\widehat{SE}(\hat{\beta}_1)]$$

O valor de c deve ser escolhido conforme nossa conveniência. As vezes quando c é igual a 1 é muito pequena para ser considerada importante, ou pode ser até muito grande, dependendo da escala de x.

Quando um intervalo de confiança para um *odds ratio* contém o 1, significa que a diferença de x não influencia Y. Da mesma forma que um intervalo de confiança para os coeficientes de regressão (betas) contendo zero, quer dizer que eles não são significativos.

Por exemplo, um *odds ratio* com c igual a 10 anos, que tem valor igual a 3, deve ser interpretado da seguinte forma : a cada aumento de 10 anos, a chance de se apresentar a característica Y=1 aumenta 3 vezes.

3.4 Confundimento e Interação

O termo variável de confundimento é usado para descrever uma variável que é associada com alguma variável independente e também com a variável resposta. Quando existe uma variável assim, ela deve ser considerada no modelo, pois se é ignorada, ela afeta as interpretações. Por exemplo, se quero comparar o peso de pessoas em dois grupos em relação a um determinado tipo de alimentação, mas esses dois grupos não tem a mesma distribuição de idade, que influencia muito no peso, é indicado introduzir a variável idade no modelo.

Se a associação entre uma covariável e a variável resposta é a mesma para todos os níveis, então não há interação entre elas. Os gráfico de regressão nos k níveis das covariáveis seriam retas paralelas. Uma maneira de ver se a interação é significativa é fazendo o teste da razão de máxima verossimilhança, e vendo se o modelo que inclui a interação é significativamente melhor do que o que não inclui.

3.5 Seleção de variáveis

Tipicamente, existem muitas variáveis que podem ser incluídas no modelo. O objetivo de qualquer método é selecionar aquelas que resultam no melhor modelo, sendo o mais simples possível. Um modelo que minimiza o número de variáveis é mais provável de ser numericamente estável, e mais facilmente generalizado.

Stepwise selection

Deve-se verificar a importância de cada variável a ser incluída no modelo. Essa verificação pode ser feita:

Fazendo teste de Wald para o coeficiente de cada variável (univariado), e comparando o coeficiente estimado da variável no modelo completo, com o coeficiente no modelo contendo apenas essa variável. Variáveis que não contribuam para o modelo segundo esses critérios devem ser retiradas, e o novo modelo deve ser comparado com o inicial através do teste de *likelihood ratio*. Devemos comparar também os coeficientes das variáveis em comum

nos dois modelos. Se um coeficiente sofre uma alteração muito grande com a eliminação de alguma variável, pode ser que a variável excluída seja importante para o ajuste.

Em qualquer passo no procedimento, a variável mais importante é a que gera a maior diferença na log verossimilhança .

Passo 1: Começa com um modelo contendo apenas o intercepto, e estimando sua log verossimilhança, que chamaremos de L_0. Então, para as p variáveis possíveis de entrar no modelo, são criadas regressões que contenham apenas elas (p regressão univariadas) e comparando suas respectivas log verossimilhanças. A log verossimilhança para um modelo contendo a variável x_j será tido como L_j.

A variável considerada mais importante, será a com menor p-valor no teste da razão de verossimilhança, comparando cada modelo univariado com o modelo com o intercepto. Um p-valor limite é estipulado (p_ϵ), e o processo só continua se o menor p-valor encontrado, ou seja, o p-valor da variável que será adicionada, for menor do que o valor limite.

Passo 2: O modelo agora contém o intercepto e a variável selecionada no passo anterior, que chamaremos de $x_{\varepsilon 1}$. Ajustamos agora as p-1 regressões contendo $x_{\varepsilon 1}$ e x_j e encontramos suas log verossimilhanças. Como no passo 1, efetuamos o teste da razão de máxima verossimilhança para todos os modelos ajustados, e escolhemos o que tem o menor p-valor , se esse p-valor for menor do que p_ϵ.

Passo 3: Depois que mais uma variável foi adicionada ao modelo, talvez a primeira variável passe a não ter tanta importância. Portanto fixamos um p-valor (p_r) para a exclusão da variável. São feitos testes comparando os modelos univariados com o modelo as duas variáveis. Se o p-valor de algum deles for maior do que p_r(que deve ser maior do que p_ϵ), a variável é excluída.

Passo 4: Análogo ao passo 3, o processo continua dessa forma até o último passo.

Último passo: Esse passo ocorre quando todas as p variáveis estão no modelo, ou todas as variáveis dentro do modelo tem p-valor de saída menores do que p_r e as fora do modelo tem p-valores de entrada maiores do que p_ϵ.

3.6 Modelo de regressão logística multinomial

Uma regressão logística politômica é um modelo onde a variável resposta pode assumir k diferentes categorias. Falaremos da regressão com 3 categorias, que pode ser facilmente generalizado para k.

Quando a variável resposta é binária, o modelo é parametrizado em termos do logito de Y=1 e Y=0. No modelo com 3 níveis, precisamos de dois logitos, logo, temos que escolher quais categorias comparar. Normalmente usa-se Y=0 como referência, comparando a Y=1 e Y=2. (a comparação entre Y=1 e Y=2 será dada pela diferença entre os dois logitos).

Assumindo o vetor $\boldsymbol{x}' = (1, x_1, x_2, \dots, x_p)$ citado anteriormente, as duas funções logito são dadas por

$$g_1(\boldsymbol{x}) = \ln\left[\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p = \boldsymbol{x}'\boldsymbol{\beta_1}$$

$$g_2(\boldsymbol{x}) = \ln\left[\frac{P(Y=2|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p = \boldsymbol{x}'\boldsymbol{\beta_2}$$

E

$$\ln\left[\frac{P(Y=2|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right] = \ln\left[\frac{P(Y=2|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right] - \ln\left[\frac{P(Y=1|\boldsymbol{x})}{P(Y=0|\boldsymbol{x})}\right]$$

A probabilidade condicional é dada por

$$P(Y=j|\boldsymbol{x}) = \frac{e^{\boldsymbol{g_j(x)}}}{\sum_{k=0}^{2} e^{g_k(\boldsymbol{x})}}$$

onde $g_0(\boldsymbol{x}) = 0$

Para construir a função de máxima verossimilhança, criamos três variáveis binárias, que assume os valores 0 e 1 para indicar a qual grupo a observação pertence. Se Y=0, então $Y_0 = 1$, $Y_1 = 0$ e $Y_2 = 0$. Se Y=1, então $Y_0 = 0$, $Y_1 = 1$ e $Y_2 = 0$. E se Y=2, $Y_0 = 0$, $Y_1 = 0$ e $Y_2 = 1$.

Logo, a função de verossimilhança para uma amostra de *n* observações independentes é

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi_0(\boldsymbol{x_i})^{y_0} \pi_1(\boldsymbol{x_i})^{y_1} \pi_2(\boldsymbol{x_i})^{y_2}]$$

Tirando o logaritmo é usando o fato de que $Y_0 + Y_1 + Y_2 = 1$ para todo *i*, a função de logverossimilhança é

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_{1i} g_1(\boldsymbol{x_i}) + y_{2i} g_2(\boldsymbol{x_i}) - \ln\left(1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})}\right)$$

As equações de verossimilhança são obtidas por meio das derivadas parcial de $L(\boldsymbol{\beta})$, em relação a cada um dos 2(p+1) parâmetros.

$$\frac{dL(\boldsymbol{\beta})}{d\beta_{jk}} = \sum_{i=1}^{n} x_{ki}[y_{ji} - \pi_j(x_i)]$$

O estimador de máxima verossimilhança, $\widehat{\boldsymbol{\beta}}$**,** é obtido igualando essas equações a zero.

A matriz de segundas derivadas parciais é necessária para a obtenção da matriz de informção e a matriz de covariâncias estimadas. Os elementos dessa matriz tem a forma

$$\frac{d^2L(\boldsymbol{\beta})}{d\beta_{jk}\beta_{jk'}} = -\sum_{i=1}^{n} x_{k'i}x_{ki}\pi_j(x_i)\,[1 - \pi_j(x_i)]$$

E

$$\frac{d^2L(\boldsymbol{\beta})}{d\beta_{jk}\beta_{j'k'}} = -\sum_{i=1}^{n} x_{k'i}x_{ki}\pi_j(x_i)\,\pi_{j'}(x_i)$$

A matriz de informação observada , $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$**,** é 2(p+1)×2(p+1) é a matriz cujos elementos são o negativo dos valores das equações acima. A matriz de covariâncias dos estimadores é dada pela inversa da matriz de informação observada.

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1}$$

Sendo a matriz **X** a matriz n×(p+1) contendo os valores das covariáveis, $\boldsymbol{V_j}$ a matriz diagonal n×n com elemento geral $\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$ para j=1,2 e i=1,2,...,n, e seja $\boldsymbol{V_3}$ a matriz diagonal n×n $\hat{\pi}_{1i}\hat{\pi}_{2i}$.

O estimador da matriz de informação pode ser expresso como

$$\hat{I}(\hat{\beta}) = \begin{bmatrix} \hat{I}(\hat{\beta})_{11} & \hat{I}(\hat{\beta})_{12} \\ \hat{I}(\hat{\beta})_{21} & \hat{I}(\hat{\beta})_{22} \end{bmatrix}$$

Onde

$$\hat{I}(\hat{\beta})_{11} = (X'V_1X),$$

$$\hat{I}(\hat{\beta})_{22} = (X'V_2X)$$

E

$$\hat{I}(\hat{\beta})_{12} = -(X'V_3X)$$

$$\hat{I}(\hat{\beta})_{22} = -(X'V_3X)$$

Interpretando a siginificância dos coeficientes estimados

O *odds ratio* da resposta Y=j com Y=0, para os valores *a* de covariaveis, versus os valores *b* e dado por:

$$\widehat{OR}(a,b) = \frac{P(Y=j|x=a)/P(Y=0|x=a)}{P(Y=j|x=b)/P(Y=0|x=b)}$$

Para compararmos Y=1 com Y=2, a melhor forma é é obtendo a diferença entre $\hat{\beta}_{21} e\, \hat{\beta}_{11}$ e o intervalo de confiança pra essa diferença.

O estimador da variância pra essa diferença é

$$\widehat{Var}(\hat{\beta}_{21} - \hat{\beta}_{11}) = \widehat{Var}(\hat{\beta}_{21}) + \widehat{Var}(\hat{\beta}_{11}) - 2\widehat{Cov}(\hat{\beta}_{21}, \hat{\beta}_{11})$$

E o intervalo de confiança para a diferença é dado da mesmo jeito que foi visto anteriormente,

$$\left(\hat{\beta}_{21} - \hat{\beta}_{11}\right) \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_{21} - \hat{\beta}_{11})$$

Se esse intervalo conter zero, significa que o intervalo do *odds ratio*, que é dado pela exponencialização desses limites, conterá o 1. Ou seja, não poderíamos concluir que há diferente entre os *odds* pra Y=1 e para Y=2.

3.7 Modelo de regressão logística ordinal

Em casos em que a variável resposta é ordinal, como por exemplo, dor (pouca, média, muita), o modelo de regressão multinomial não levaria em conta a ordem das respostas, e as estimativas podem não sair como esperado. Nesse caso, os modelos mais utilizado são os *proportional odds models.*

Logitos cumulativos

Seja $P(Y \leq j|\boldsymbol{x}) = \pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})$, j =1,....,J

Onde J são os níveis que a variável resposta pode adotar. (mas $P(Y \leq j|\boldsymbol{x})$ quando j=J é 1, então daqui para frente consideraremos j=1,....,J-1)

O logito cumulativo é dado por

$$logit[P(Y \leq j|\boldsymbol{x})] = \log\frac{P(Y \leq j|\boldsymbol{x})}{1 - P(Y \leq j|\boldsymbol{x})} = log\frac{\pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}$$

, j = 1,...., J-1

Modelo *Odds* Proporcional

$$logit[P(Y \leq j|\boldsymbol{x})] = \boldsymbol{\alpha_j} + \boldsymbol{\beta'x}, \qquad \text{j = 1, ... , J-1}$$

Cada logito cumulativo *j* tem seu próprio intercepto.

O logito cumulativo satisfaz

$$logit[P(Y \leq j|\boldsymbol{x_1})] - logit[P(Y \leq j|\boldsymbol{x_2})] \;=\; log\left[\frac{P(Y \leq j|\boldsymbol{x_1})/P(Y \leq j|\boldsymbol{x_1})}{P(Y \leq j|\boldsymbol{x_2})/P(Y \leq j|\boldsymbol{x_2})}\right]$$

$$= \boldsymbol{\beta'}(\boldsymbol{x_1} - \boldsymbol{x_2})$$

Um *odds ratio* de probabilidade acumulada é chamado *odds ratio acumulado.* A chance de se fazer a resposta $\leq j$ em $\boldsymbol{x_1}$ é $\exp[\boldsymbol{\beta'}(\boldsymbol{x_1} - \boldsymbol{x_2})]$ a chance em $\boldsymbol{x_2}$.

O *odds ratio acumulado* é igual a e^{β} quando $\boldsymbol{x_1} - \boldsymbol{x_2} = 1$.

Sejam $(y_{i1}, \dots, y_{iJ})$ indicadores binários para a resposta em i, a função de verossimilhança é dada por

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(x_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}(P(Y \le j|\boldsymbol{x_1}) - P(Y \le j-1|\boldsymbol{x_1}))\right]$$

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}(\frac{\exp(\alpha_j + \boldsymbol{\beta'x_i})}{1+\exp(\alpha_j + \boldsymbol{\beta'x_i})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta'x_i})}{1-\exp(\alpha_{j-1} + \boldsymbol{\beta'x_i})})^{y_{ij}}\right]$$

Para a utilização de uma regressão logística, deve-se aplicar um o teste de *proportional odds*, visto que o modelo contém apenas uma estimativa para os coeficientes $\boldsymbol{\beta}$, assumindo que a mudança do odds de um nível para Y=1 comparado Y=0 e para Y=2 comparado Y=1. O software SAS utiliza o Score Test nesse caso.

4 Análise Exploratória dos Dados

Toda a análise descritiva das variáveis levou em conta o peso atribuído a cada uma das observações da amostra.

Tabela 2 – Renda per capita em reais da população jovem (18 a 24 anos) na AMB. 2010

Amostra Expandida	Média	Desvio Padrão	Mínimo	Máximo
456.994	1.144	18.271	0	591.043

Fonte: IBGE - Censo Demográfico 2010

Tabela 3 – Quantis da renda *per capita* em reais da da população jovem (18 a 24 anos) na AMB. 2010

Quantis	**Renda**
100%	591.043
99%	8.233
95%	4.000
90%	2.553
75%	1.140
50% (Mediana)	553
25%	301
10%	168
5%	85
1%	0
0%	0

Fonte: IBGE - Censo Demográfico 2010

Ao analisar as tabelas anteriores observa-se a grande desigualdade na distribuição de renda na área estudada. Na tabela 3, fica claro que apenas 25% da população tem renda acima da média, ou seja, essa média é afetada por valores extremamente altos, o que demonstra a forte concentração de renda em uma pequena parcela da população.

Tabela 4- População de 18 a 24 anos na AMB que frequenta a escola. 2010

Frequenta Escola	n	%
Sim	161.456	35,33%
Não, já frequentou	290.074	63,47%
Não, nunca frequentou	5.464	1,20%
Total	456.994	

Fonte: IBGE - Censo Demográfico 2010

Apenas 35,33% dos jovens de 18 a 24 anos na AMB frequentam a escola, sendo que entre esses jovens que ainda estudam, 69.048 (que equivale a 42,8%) não estão cursando o ensino superior.

Tabela 5 – Nível educacional dos maiores de 24 anos e dos jovens da AMB. 2010

	Maiores de 24 anos		**Jovens de 18 a 24 anos**	
Nível Educacional	n	%	n	%
1	1.000.472	50,61	166.887	36,52
2	514.673	26,04	169.931	37,18
3	461.495	23,35	120.176	26,30
Total	1.976.640		456.994	

Fonte: IBGE - Censo Demográfico 2010

Na comparação os níveis de ensino dos jovens entre 18 a 24 anos com a população com mais de 24 anos, observa-se um aumento no nível de escolaridade dos jovens em relação à população mais velha. Enquanto 50,61% dessa população mais velha não tem a educação básica, ou está em enorme defasagem de estudo, entre os jovens, essa porcentagem cai para 36,52%.

Tabela 6– Características sócio-demográficas dos jovens da AMB de 18 a 24 anos segundo nível de ensino. 2010

Características	Total		Nível Educacional					
			1		2		3	
	n	%	n	%	n	%	n	%
Total	**456.994**		166.887		169.931		120.176	
Idade								
18 anos	61.761	13,5	23.415	14,0	20.929	12,3	17.417	14,5
19 anos	61.247	13,4	22.386	13,4	24.244	14,3	14.617	12,2
20 anos	64.803	14,2	23.607	14,1	25.054	14,7	16.141	13,4
21anos	64.200	14,0	23.299	14,0	24.093	14,2	16.808	14,0
22 anos	67.797	14,8	25.397	15,2	24.321	14,3	18.079	15,0
23 anos	67.563	14,8	24.271	14,5	25.788	15,2	17.503	14,6
24 anos	69.624	15,2	24.512	14,7	25.501	15,0	19.611	16,3
Sexo								
Masculino	224.768	49,2	93.861	56,2	77.890	45,8	53.018	44,1
Feminino	232.226	50,8	73.027	43,8	92.040	54,2	67.158	55,9
Cor								
Branca	165.333	36,2	44.918	26,9	58.130	34,2	62.284	51,8
Preta	41.242	9,0	18.218	10,9	15.635	9,2	7.389	6,1
Amarela	9.791	2,1	3.386	2,0	4.107	2,4	2.299	1,9
Parda	239.461	52,4	99.829	59,8	91.643	53,9	47.989	39,9
Indígena	1.122	0,3	492	0,3	416	0,2	215	0,2
Ignorado	45	0,0	45	0,0	-	-	-	-
Situação Domicílio								
Urbano	437.814	95,8	155.163	93,0	163.804	96,4	118.847	98,9
Rural	19.180	4,2	11.724	7,0	6.126	3,6	1.330	1,1
Relação com Responsável								
Responsável ou cônjuge	124.009	27,1	59.906	35,9	47.286	27,8	16.818	14,0
Filho ou Enteado	247.276	54,1	69.811	41,8	90.632	53,3	86.834	72,3
Outros	85.708	18,8	37.171	22,3	32.013	18,8	16.524	13,7
Trabalho								
Só estuda	80.778	17,7	18.678	11,2	15.868	9,3	46.233	38,5
Estuda e Trabalha	80.678	17,7	13.506	8,1	13.662	8,0	53.511	44,5
Só trabalha	184.694	40,4	79.797	47,8	90.531	53,3	14.367	12,0
Não estuda e não trabalha	110.843	24,3	54.906	32,9	49.871	29,3	6.066	5,0
Região de Moradia								
Região 1 - DF	39.491	8,6	4.315	2,6	7.799	4,6	27.377	22,8
Região 2 - DF	124.921	27,3	34.058	20,4	43.160	25,4	47.703	39,7
Região 3 - DF	170.415	37,3	70.108	42,0	69.663	41,0	30.646	25,5
Municípios de Goiás	122.166	26,7	58.406	35,0	49.309	29,0	14.450	12,0
Quartil de Renda*								
1	116.799	25,5	69.177	41,5	37.872	22,3	9.750	8,1
2	113.363	24,8	50.514	30,3	49.802	29,3	13.047	10,9
3	113.198	24,8	34.654	20,8	50.626	29,8	27.918	23,2
4	113.634	24,9	12.542	7,5	31.630	18,6	69.461	57,8

Fonte: IBGE - Censo Demográfico 2010

* Os quartis apresentam erro de arredondamento

Nota-se que a maior porcentagem de pretos e a maior porcentagem de pardos (que juntos formam os negros) está no nível 1 de ensino. Enquanto nesse nível eles somam 70,7% das pessoas, no nível 3 essa porcentagem é de 46%. O grupo 3 também apresenta a maior proporção de brancos, 51,8%, muito maior do que nos outros níveis, que apresentam 26,9% e 34,2%.

Embora a porcentagem de moradores da zona rural seja muito baixa no total estudado, ainda assim nota-se diferença entre os níveis educacionais. Enquanto no nível mais baixo essa população chega a 7%, no nível 3 ela é de apenas 1,1%.

Na tabela anterior pode-se ver a relação do grau de parentesco com o responsável do domicílio com o nível educacional. Embora em todos os níveis a relação principal seja de filho ou enteado, no maior nível educacional essa proporção é de 72,3% enquanto que no menor é 41,8%. Inversamente a isso, nos níveis educacionais mais baixos, 1 e 2, a proporção de responsáveis ou conjuges é muito maior do que no 3.

A porcentagem de jovens que não trabalham e também não estudam nos níveis educacionais mais baixos é consideravelmente maior do que no grupo 3. Enquanto nos níveis 1 e 2 essa proporção é de 32,9% e de 29,3%, respectivamente, no nível 3 é de apenas 5%.

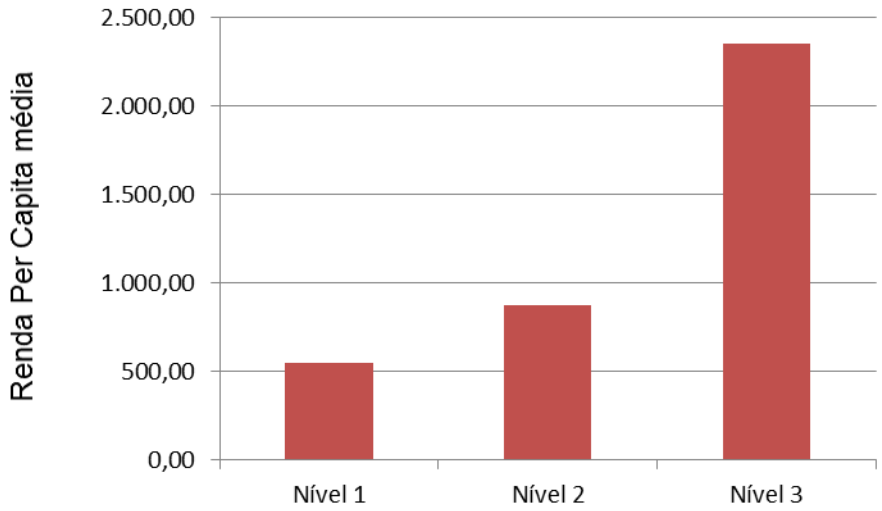
As características ‘região de moradia’ e ‘quartil de renda’ têm interpretações similares, visto que essas regiões foram classificadas em termos de renda. Enquanto entre os jovens que estão no nível 1, 7,5% se encontram no quartil de renda mais alta e 23% residem nas regiões 1 e 2 (sendo que na regiao 1 essa porcentagem é de apenas 2,6%), 62,5% dos jovens no grupo 3 moram nas regiões 1 e 2, e 57,8% estão no quartil de maior renda. Fica nítida a grande relação entre situação financeira e educação.

Tabela 7 - Média da renda do jovem de 18 a 24 anos da AMB × Nível de Ensino. 2010

Nível Educacional	Renda Per Capita Média
Nível 1	544,99
Nível 2	870,88
Nível 3	2.353,58

Fonte: IBGE - Censo Demográfico 2010

Gráfico 1 – Média da renda do jovem de 18 a 24 anos da AMB × Nível de Ensino.



Observa-se que apenas o grupo 3 tem renda acima da média geral, que é R$ 1.140,00 per capita.

5 Aplicações e Resultados

Quando uma variável tem uma ordem, deve-se utilizar uma regressão logística ordinal. Como a variável ensino foi criada com o objetivo de ser ordinal, sendo o 1 mais baixo, o 2 médio e o 3 o mais alto, devemos testar se os odds mudam na mesma forma (ou seja, o odds ratio do 3 pro 2 é igual ao do 2 pro 1). Através do score test , testamos 'proportional odds test' no SAS. Como a amostra é muito grande, para testar essa hipótese foram selecionadas várias amostras diferentes com o objetivo de avaliar com que freqüência o teste é significante. Foram selecionadas 50 amostras para cada tamanho de n, onde n=100,200,....,1000.

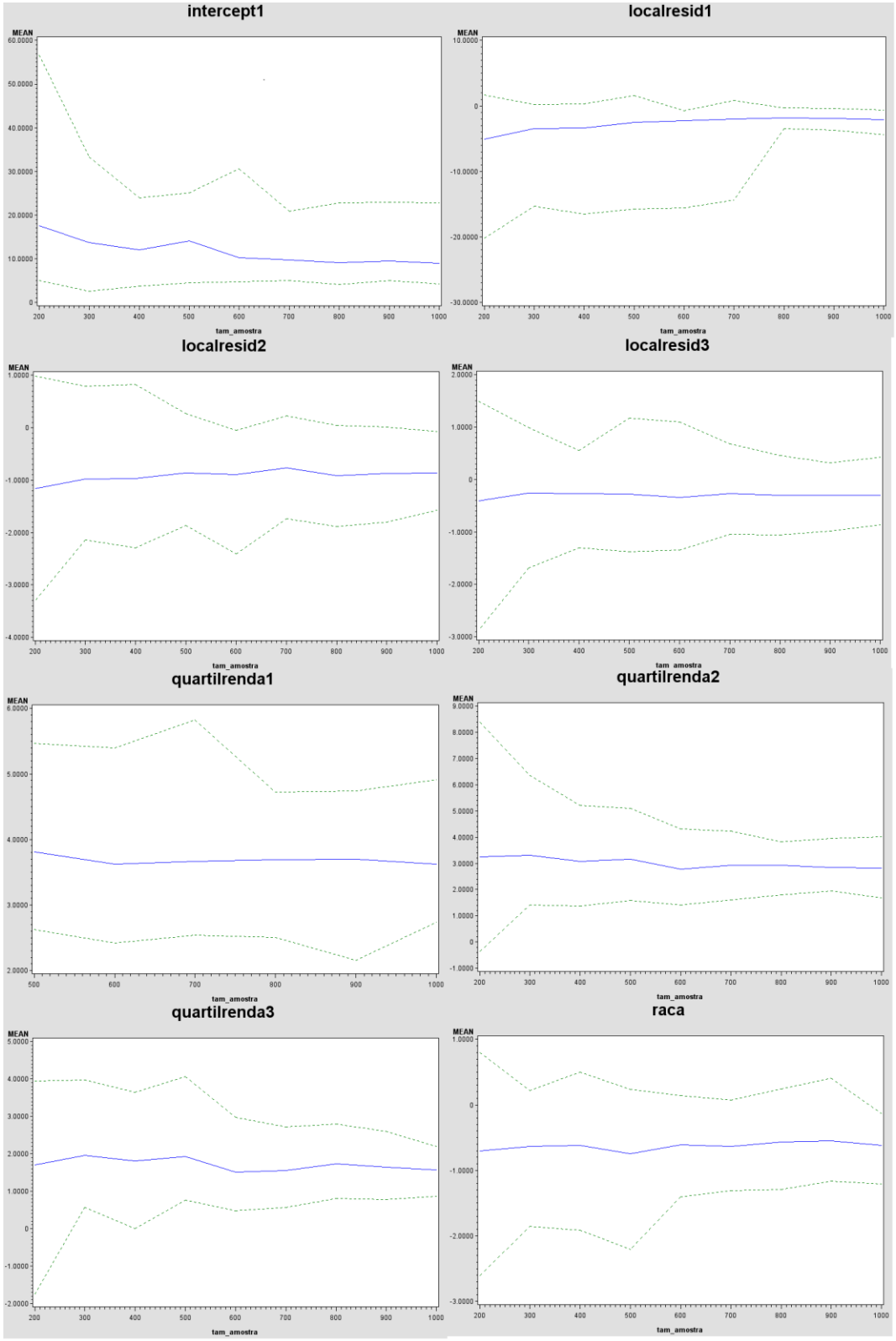
As amostras selecionadas nessa parte do trabalho não consideram os pesos para expansão da amostra. Foram selecionadas, portanto, da população de aproximadamente 26.000 jovens.

Tabela 8 - Proportional Odds Score Test

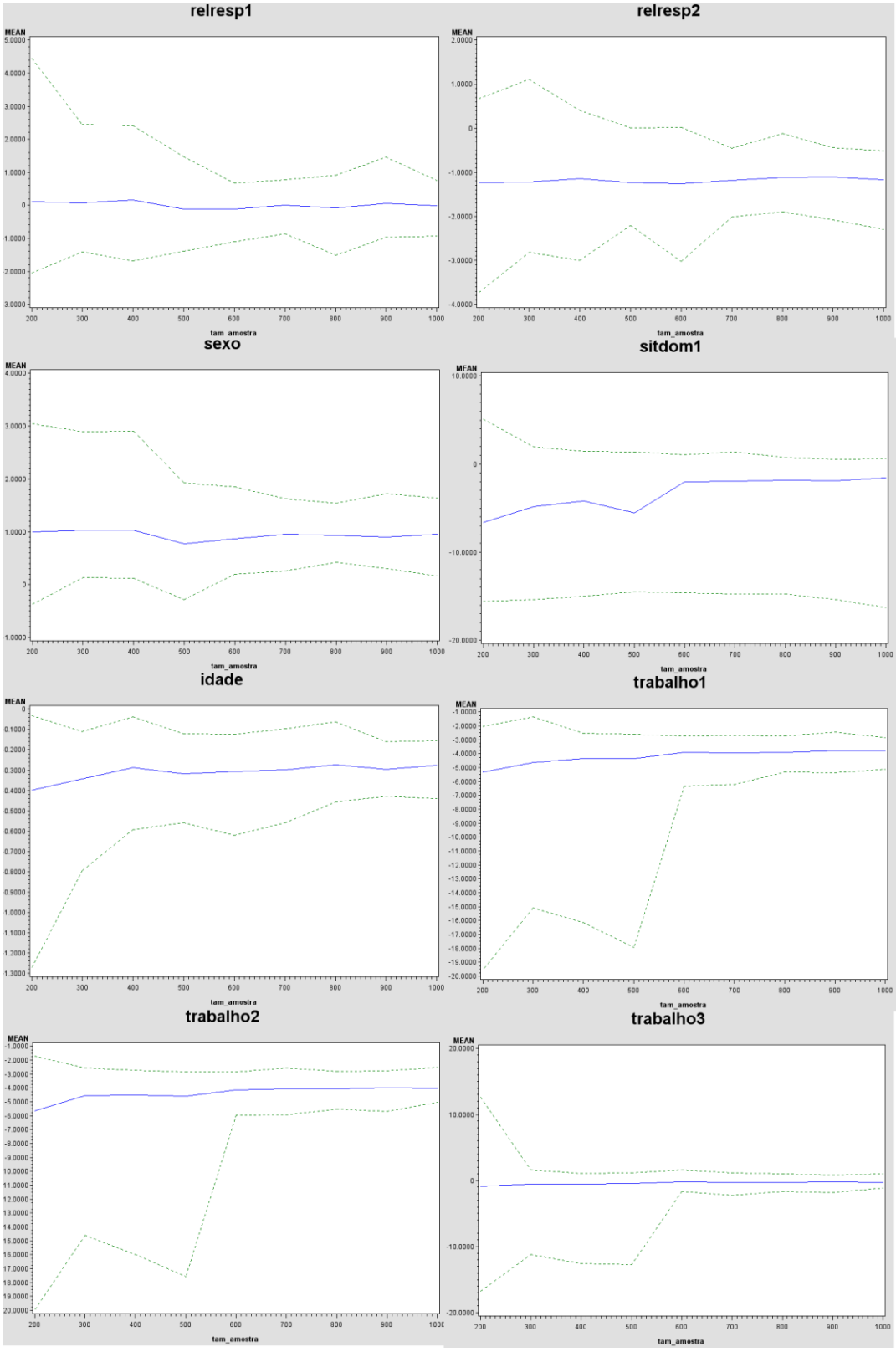
P-valor	Frequência
< 0,05	479
≥ 0,05	21
Total	500

Do total de 500 amostras, apenas 21 deram significativas a um nível de 5%, o que faz supor que um modelo considerando as variáveis qualitativas nominais seja mais adequado (link=glogit).

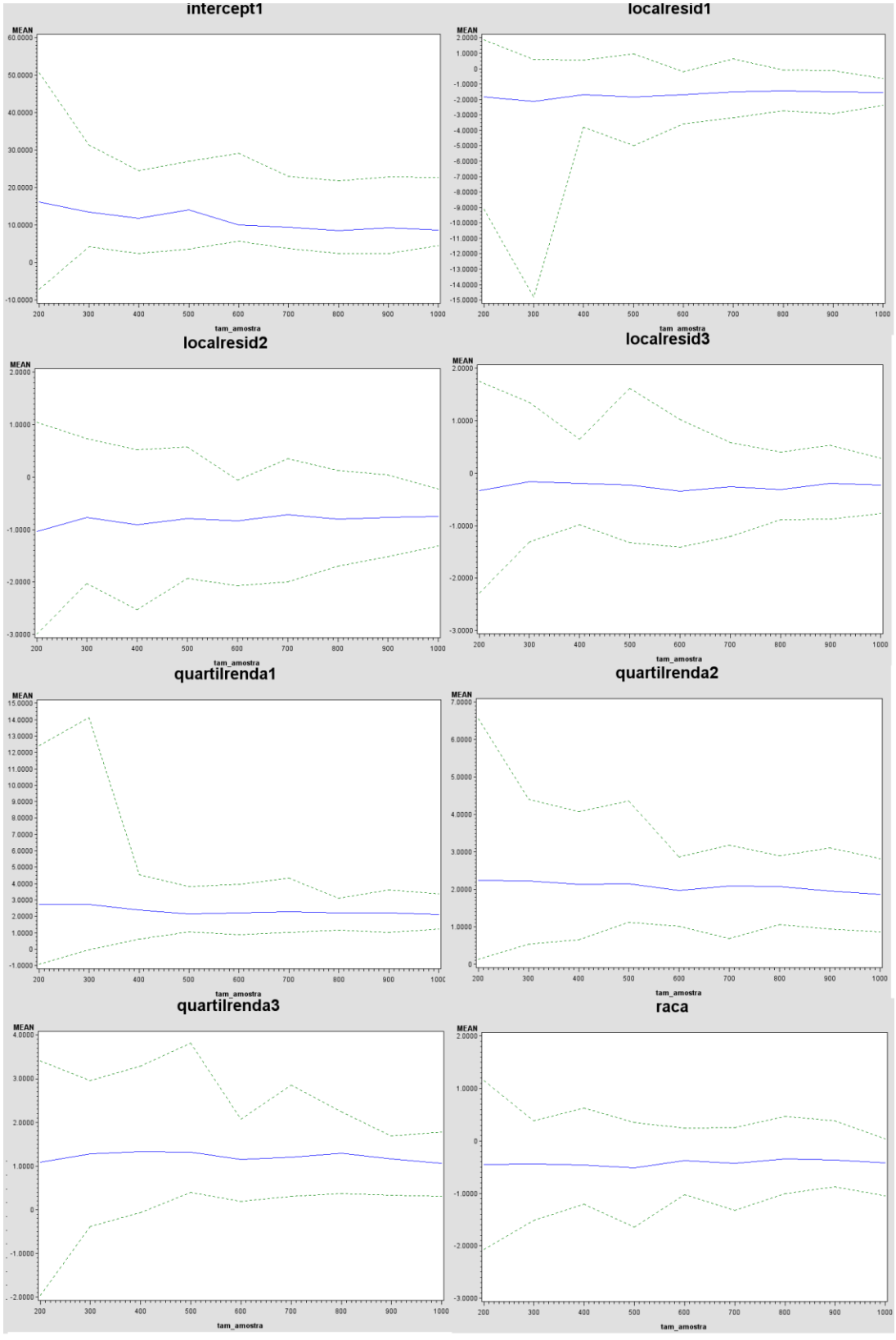
Logito 1


intercept1
MEAN
tam_amostra
localresid1
MEAN
tam_amostra
localresid2
MEAN
tam_amostra
localresid3
MEAN
tam_amostra
quartilrenda1
MEAN
tam_amostra
quartilrenda2
MEAN
tam_amostra
quartilrenda3
MEAN
tam_amostra
raca
MEAN
tam_amostra

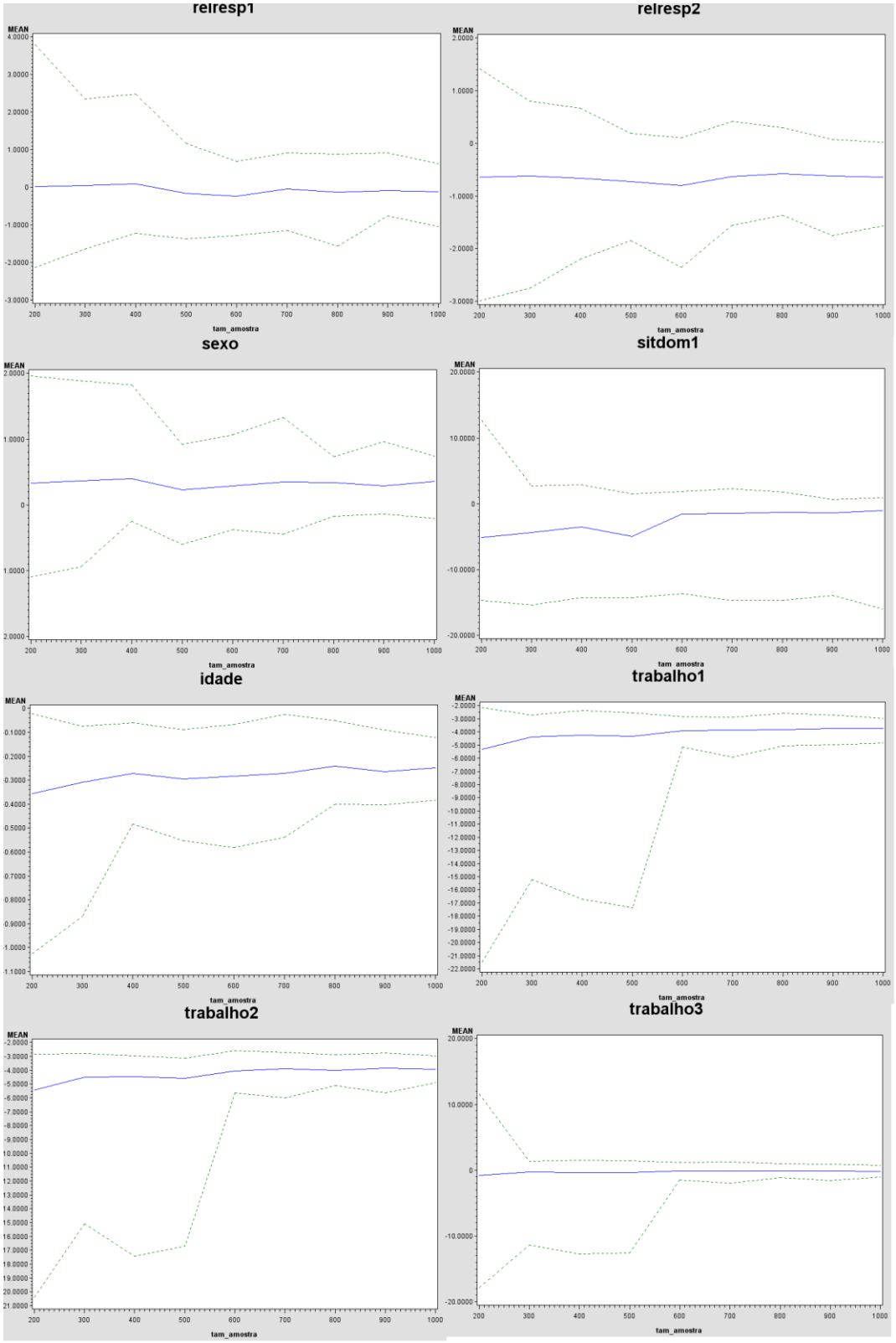
Logito 1


relresp1
relresp2
sexo
sitdom1
idade
trabalho1
trabalho2
trabalho3
MEAN
tam_amostra

Logito 2


intercept1
localresid1
localresid2
localresid3
quartilrenda1
quartilrenda2
quartilrenda3
raca
MEAN
tam_amostra

Logito 2


relresp1
relresp2
sexo
sitdom1
idade
trabalho1
trabalho2
trabalho3
MEAN
tam_amostra

Como a amostra é muito grande (aproximadamente 26.000 pessoas, e 450.000 na amostra expandida), deve-se selecionar uma amostra menor para aplicação do método de análise.

Depois de constatado que o modelo ordinal não deveria ser usado, foi feita uma simulação do modelo multinomial contendo todas as variáveis de interesse. Foram selecionadas 50 amostras para cada tamanho de amostra, que variou de 200 a 1.000 (para amostras de tamanho 100 nenhum modelo se mostrou bem ajustado) e calculada uma média das estimativas dos coeficientes regressão para cada tamanho de amostra. Os gráficos anteriores foram utilizados para auxiliar na escolha do tamanho da amostra. A linha azul é a média das estimativas para cada tamanho de amostra, e as verdes são o máximo e mínimo. Para amostras pequenas essa variação é muito grande, mas depois de um certo tamanho, os valores começam a convergir, e variar menos do mínimo para o máximo. Com base nessa observação, o valor de estabilidade escolhido foi o de 700 pessoas.

Para a análise das variáveis mais selecionadas para o modelo, foi feita uma *stepwise selection* em 1000 amostras diferentes com 700 pessoas com o propósito de verificar quais variáveis entraram mais vezes no modelo. As freqüências de entrada estão apresentadasna tabela abaixo, em ordem decrescente.

Tabela 9 – Frequência das variáveis selecionadas em 1.000 amostras de 700 pessoas

Variável	Frequência
Trabalho	1000
Quartil de Renda	1000
Idade	951
Sexo	986
Relação com Responsável	985
Local de Residência	905
Raça	704
Situação de Domicílio	551

As variáveis ‘quartil de renda’ e ‘trabalho’ foram selecionadas em todas as amostras, enquanto ‘situação de domícilio’ e ‘raça’ apresentaram freqüências mais baixas. Embora raça e local de residência sejam correlacionadas com a variável resposta, ensino, elas também têm correlação com renda.

Para aplicação a regressão logística, foi selecionada apenas uma amostra de tamanho 700. Para essa amostra, as variáveis selecionadas pelo *stepwise selection* foram sexo, quartil de renda, trabalho, relação com o responsável e idade.

O modelo multinomial, no caso em que a variável resposta tem 3 níveis, cria 2 logitos diferentes. O nível de referência utilizado foi ensino = 3, então o logito 1 é o nível 1 em relação ao 3, e o logito 2, o 2 também em relação a o 3.

Tabela 10 – Estatísticas de Ajuste do modelo

Logito	Hosmer & Lemeshow	g.l.	P-valor	χ^2 de Pearson	g.l.	P-valor
1	4,63	8	0,80	184,04	256	0,99
2	6,31	8	0,61	207,76	242	0,95

No caso da regressão logística politômica, para interpretação do ajuste do modelo foram separados os dois logitos e analisadas as estatísticas individaulmente. Para os dois logitos, ambos os testes de Hosmer & Lemeshow e do χ^2 de Pearson têm p-valor alto, o que indica que o modelo está bem ajustado.

Tabela 11 – Logitos Estimados

Logito	Variável	Coeficiente	P-valor de Wald
1	Intercepto	2,99	0,0689
	Sexo	0,81	0,0079
	Idade	-0,34	<0,0001
	RelResp1	1,47	0,0004
	RelResp3	1,09	0,0109
	Quartilrenda1	4,56	<0,0001
	Quartilrenda2	2,68	<0,0001
	Quartilrenda3	4,52	<0,0001
	Trabalho2	-0,72	0,0919
	Trabalho3	3,13	<0,0001
	Trabalho4	2,83	<0,0001
2	Intercepto	2,69	0,0954
	Sexo	0,35	0,2407
	Idade	-0,26	0,0015
	RelResp1	0,94	0,0232
	RelResp3	0,60	0,1587
	Quartilrenda1	2,68	<0,0001
	Quartilrenda2	3,14	<0,0001
	Quartilrenda3	1,92	<0,0001
	Trabalho2	-1,37	0,0036
	Trabalho3	3,43	<0,0001
	Trabalho4	3,24	<0,0001

Os modelos de regressão estão contidos na tabela anterior. Embora o software SAS mostre na saída as razões de chances para a comparação entre dois níveis de uma variáveis, sabe-se que esses valores são calculados fazendo-se e^{β} e são dados na tabela a seguir.

Tabela 12 – Odds Ratios Estimados

Logito	Variável		Odds Ratio	Intervalo de Confiança de Wald (95%)	
1	sexo		2,249	1,237	4,089
	idade		0,714	0,608	0,838
	relresp	1 vs 2	4,356	1,917	9,897
	relresp	3 vs 2	2,977	1,286	6,893
	quartilrenda	1 vs 4	95,857	34,946	262,938
	quartilrenda	2 vs 4	92,101	33,135	256,004
	quartilrenda	3 vs 4	13,86	5,832	32,942
	trabalho	2 vs 1	0,485	0,209	1,125
	trabalho	3 vs 1	22,921	8,757	59,995
	trabalho	4 vs 1	16,952	6,302	45,602
2	sexo		1,417	0,791	2,538
	idade		0,774	0,661	0,906
	relresp	1 vs 2	2,555	1,137	5,743
	relresp	3 vs 2	1,824	0,791	4,21
	quartilrenda	1 vs 4	14,635	5,599	38,25
	quartilrenda	2 vs 4	23,12	8,838	60,48
	quartilrenda	3 vs 4	6,823	3,127	14,887
	trabalho	2 vs 1	0,255	0,102	0,639
	trabalho	3 vs 1	30,723	12,209	77,313
	trabalho	4 vs 1	25,426	9,868	65,514

Tabela 13 – Média dos Odds Ratios

Logito	Variável		Média Odds Ratio	Intervalo de Incerteza 5%	95%
1	sexo		2,57	1,45	4,12
	idade		0,74	0,65	0,84
	relresp	1 vs 2	3,44	1,57	6,25
	relresp	3 vs 2	3,13	1,44	6,00
	quartilrend	1 vs 4	83,64	31,26	169,13
	quartilrend	2 vs 4	33,32	14,52	71,72
	quartilrend	3 vs 4	8,75	3,89	16,54
	trabalho	2 vs 1	1,11	0,44	2,16
	trabalho	3 vs 1	47,46	18,47	95,89
	trabalho	4 vs 1	53,68	19,22	121,32
2	sexo		1,43	0,84	2,23
	idade		0,76	0,67	0,86
	relresp	1 vs 2	2,22	0,98	3,40
	relresp	3 vs 2	1,90	0,92	3,43
	quartilrend	1 vs 4	16,19	6,87	31,35
	quartilrend	2 vs 4	11,83	6,06	22,17
	quartilrend	3 vs 4	4,94	2,63	8,24
	trabalho	2 vs 1	1,09	0,53	2,08
	trabalho	3 vs 1	47,46	20,41	90,83
	trabalho	4 vs 1	47,17	18,38	97,34

O modelo foi rodado com apenas uma amostra de 700 pessoas. Com fim de analisar a que ponto essa amostra estaria sendo representativa, foram geradas 500 amostras também de tamanho 700, e para cada amostra foi gerado um modelo de regressão com as variáveis selecionadas anteriormente. Foram calculadas médias para os *odds ratios* e o intervalo de incerteza foi feito ordenando a amostra e selecionando o quantil que corta 5% das pessoas, e o que corta 95%.

Quando o intervalo de confiança para um odds ratio contém o valor 1, significa que não há diferença entre as chances . No logito 1, que compara o nível de ensino um com o dois, quando compara-se os jovens que só estudam e os jovens que estudam e trabalham, não parece haver diferença entre as chances desse jovem estar no nível 1 de ensino. No logito dois, que compara o nível 2 de ensino com o nível 3, os odds não signifcativos são o que compara sexo masculino e feminino e o que compara a relação com o responsável filho e outros. O intervalo de incerteza das médias dos odds para os três casos citados acima também contém o valor 1, e se consideramos a interpretação semelhante ao intervalo de confinça, a diferença entre o jovem que estuda com o jovem que estuda e trabalha também não seria significante no segundo logito, junto com a diferença entre os jovems que são filho ou enteado do responsável do domícilio para o jovem que é responsável ou cônjuge.

Embora na prática um valor de odds ratio de 90 seja absurdamente grande, pode-se observar que as variáveis com maior odds ratio são as mais importantes para explicar a variável ensino. As variáveis quartil de renda e trabalho foram as únicas selecionadas em todas as amostras da simulação, e são as com maior correlação com a nossa variável resposta. Mesmo que os odds médios para essas variáveis sejam menores do que os estimados para uma única amostra de 700 pessoas, ainda sim eles são consideravelmente altos.

Em ambos os logitos, a variável idades tem um odds em torno de 0,7. Isso significa que para cada ano de idade, um jovem tem aproximadamente 0,7 a chance de estar em um nível mais baixo do que um jovem mais novo. Para calcular a diferença entre um jovem de 18 com um de 24 anos, no logito 1, deve-se fazer a conta $e^{\beta(24-18)} = e^{-0,34(24-18)} = 0,13$. Ou seja, um jovem de 24 anos tem 0,13 vezes a chance de estar no nível mais baixo de ensino, comparado com o nível 3, do que um jovem de 18 anos.

Embora no logito 2 a variável sexo não seja significativa, no logito 1 considera-se que um homem tem 2,25 vezes a chance de estar no nível mais baixo de ensino do que a mulher.

Um jovem que só trabalha e um jovem que não trabalha nem estuda tem, respectivamente, 22,91 e 16,90 a chance de estar no nível 1 de ensino do que um jovem que só estuda. Na média do odds ratio, essa ordem é invertida, e consideraria-se que um jovem que não trabalha nem estuda tem maior diferença em relação a quem só estuda do que o jovem que só trabalha.

No logito 2, enquanto a chance de um jovem no primeiro quartil de renda estar no nível 2 de ensino é 14,635 vezes a chance de um jovem no quartil de renda mais alto estar no nível 2, um jovem no segundo quartil tem 23 vezes essa chance em relação a um jovem no quartil de renda mais alto. Era de se esperar que essa diferença fosse maior para o jovem do quartil 1 do que para o jovem do quartil 2. No odds ratio médio, essa ordem é invertida. Os valores passam a ser 16,19 e 11,83.

Não é simples analisar esses valores para os odds ratios estimados, até porque uma amostra de apenas 700 pessoas é muito pequena para uma população tão grande. Um fato que fica claro, porém, é a forte relação entre quartil de renda e nível de ensino, e trabalho e nível de ensino. Essas variáveis foram as únicas selecionadas em todas as amostras e são as com maior

correlação com nossa variável resposta. Também são as que apresentam maiores odds ratios, médios e estimados, para os dois logitos, porém a comparação entre quem apenas estuda e quem estuda e também trabalha não parece ser tão relevante quanto a comparação entre os jovens que só trabalham e os jovems que não trabalham nem estudam com os jovens que apenas estudam.

6 Conclusões

As variáveis renda e trabalho foram as consideradas mais importantes na explicação do modelo. Como era esperado, há uma forte relação entre a renda do jovem e seu nível de escolaridade. A variável renda, além de ser a mais correlacionada com a variável resposta, também é a mais correlacionada com raça, situação de domícilio e local de residência. Acredita-se que a inclusão dessa variável no modelo já explica suficientemente essas outras, o que torna desnecessário adicioná-las. Isso volta a velha discussão das cotas raciais, que geraram muita polêmica quando foram implementadas. Até hoje a Universidade de Brasília, por exemplo, ainda reserva parte das vagas para negros e indígenas, e há indícios de que a cor da pele não é tão importante no acesso à educação. Sabe-se que uma pessoa negra normalmente tem nível educacional mais baixo, mas ela também tem, num geral, renda consideravelmente mais baixa. Acredito, portanto, que não é a cor da pele ou o local de moradia que influencia diretamente no acesso ao ensino, e sim a renda. A renda do jovem parece ser o fator mais determinante em seu nível educacional, e nota-se a importância de outros estudos em relação a esse assunto sempre tão polêmico.

Uma dificuldade do trabalho foi utilizar uma amostra muito grande. Com uma base de dados grande, como a utilizada neste trabalho, não se pode fazer as inferências da estatística clássica. Os testes de hipóteses são muito afetados e as significâncias dos parâmetros não são representativas. Qualquer variável ou interação seria considerada importante, mas não há como conhecer se é em decorrência do tamanho da amostra ou pelo fato dela ser efetivamente importante. No presente caso, a alternativa encontrada foi de selecionar várias amostras e analisar o comportamento das estimativas. Pode não ser o ideal, pois implica na necessidade de desconsiderar os pesos dados a cada indivíduo, mas foi o método que se apresentou mais adequado.

Tal dificuldade suscita uma discussão importante no ramo da estatística, uma vez que , por mais complicado que seja trabalhar com uma base de dados muito grande, é cada vez mais importante pensar em alternativas com esse objetivo. Grandes empresas como bancos, operadoras de telefonia, e outras, dispõe de base de dados de grande dimensão e enfrentam o desafio de encontrar métodos adequados para a manipulação e inferência para tantos dados.

7 Trabalhos Futuros

Uma das restrições do trabalho foi a utilização do Censo 2010. O trabalho não é longitudinal, é um corte na data em que foi efetuado o Censo. Um estudo interessante seria considerar os recortes de outros anos para observar como evolui o perfil desses jovens (como por exemplo, antes e depois das cotas na UnB). Como o foco do trabalho era utilizar as informações

do Censo, e apenas elas, podem ter ficado faltando informações que não são contidas no questionário da amostra. Para um trabalho mais abrangente, pode-se também utilizar informações, por exemplo, do INEP, como o Censo da Educação Superior, que contém informações extremamente relevantes.

A Lei nº 12711/2012, em vigor desde o segundo semestre de 2012, estabelece que as instituições federais de educação superior terão que reservar no mínimo 50% das vagas a alunos provenientes de escolas públicas. No último edital, a UnB ofereceu 12,5% das vagas a esses alunos e, em quatro anos, essa porcentagem deve chegar a 50%. A Universidade de Brasília adota um sistema de cotas raciais desde 2003, e reserva atualmente um total de 32,5% das vagas para os dois tipos de cotas.

Além do mais, do total de vagas reservado a quem estudou integralmente em colégios públicos, metade das vagas devem ser destinadas a jovens provenientes de famílias com renda per capita igual ou inferior a 1,5 salários mínimos.

Assim, pode-se perceber como seria interessante efetuar uma análise ao longo de um período maior de tempo, mostrando como o acesso à educação superior tem evoluído, desde a implementação do sistema de cotas raciais, e vai evoluir até os próximos anos, com o aumento de vagas destinadas a pessoas de baixa renda e provenientes de escolas públicas.

8 Referências Bibliográficas

AGRESTI, Alan. Categorical Data Analysis, Second Edition. Wiley, New York, 2002

IBGE. Censo Demográfico, 2010.

IBGE. Pesquisa Nacional por Amostra de Domicílios PNAD, 2011.

IPEA. Estudo Comparado sobre a Juventude Brasileira e Chinesa, Brasília, 2012.

IPEA. Juventude e Políticas Sociais no Brasil, Brasília, 2009.

INEP. Censo do Ensino Superior, 2010.

HOSMER, David W.; LEMEWHOW, Stanley. *Applied Logistic Regression, Second Edition. Wiley, New York*, 2000.

LANGRAND, Claude; PINZÓN, Luz Mary. *Análises de datos: Métodos y ejemplos. Escuela Colombiana de Ingeniería*, 2009.

PAVIANI, Aldo [et al]. Brasília 50 anos, da capital a metrópole. Brasília : Editora UNB, 2010.

VASCONCELOS, Ana Maria; CESAR, Layla; COSTA, Maria Teresa. A universidade de Brasília e o acesso ao ensino superior na Área Metropolitana de Brasília, 2013. A publicar

_____. **Lei nº 9.394, de 20 de dezembro de 2001. Estabelece as diretrizes e bases da educação nacional**. Disponível em: <http:// *www.planalto.gov.br/ccivil_03/**leis**/l**9394**.htm*>. Acesso em 11/06/2011.

_____. **Lei nº 12.711** , **de 29 de agosto de 2012. Dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino técnico de nível médio e dá outras providências.** Disponível em: <*http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12711.htm*>. Acesso em 11/06/2011.

9 Apêndice – Programação em SAS

```
/* níveis de ensino */
data x.basejovem;
set x.basejovem;
if V0628 =4 then ensino=1;
if v0628 =3 and v6400=1 then ensino=1;
if v0628 =3 and v6400=2 then ensino=1;
if v0628 =3 and v6400=3 then ensino=2;
if v0628 =3 and v6400=4 then ensino=3;
if V0628=04 then ensino=1;
if V0629=07 and V6036=18 and V0631=03 then ensino=3;
if V0629=07 and V6036=18 and V0631=04 then ENSINO=2;
if V0629=07 and V6036=18 and V0631=05 then ENSINO=2;
if V0629=07 and V6036=18 and V0631=01 or V0631=02 then ensino=2;
if V0629=07 and V6036>18 then ensino=2;
if V0629=9 or V0629=10 or V0629=11 or V0629=12 then ensino=3;
if V0629=04 or v0629=05 or V0629=06 or V0629=08 then ensino=1;
RUN;


/* quartis de renda */
data x.basejovem;
set x.basejovem;
if rendimentoreaispc=<300.8 then quartilrenda=1;
if 553.33>rendimentoreaispc>=300.8 then quartilrenda=2;
if 1140>rendimentoreaispc>=553.33 then quartilrenda=3;
if rendimentoreaispc>=1140 then quartilrenda=4;
run;


/* local de residência */

DATA x.basejovem;
      SET x.basejovem;
    length LocalResid  $ 2;
        /* **** DF - Alta Renda  ****** */
      if v0011='00108005001' then LocalResid1='1';
    if v0011='00108005002' then LocalResid1='1';
      if v0011='00108005003' then LocalResid1='1';
      if v0011='00108005035' then LocalResid1='1';
      if v0011='00108005045' then LocalResid1='1';
      if v0011='00108005048' then LocalResid1='1';

    /* **** DF - Média Renda ****** */
    if v0011='00108005004' then LocalResid1='2';
      if v0011='00108005005' then LocalResid1='2';
      if v0011='00108005006' then LocalResid1='2';
      if v0011='00108005007' then LocalResid1='2';
      if v0011='00108005008' then LocalResid1='2';
      if v0011='00108005009' then LocalResid1='2';
      if v0011='00108005010' then LocalResid1='2';
    if v0011='00108005011' then LocalResid1='2';
      if v0011='00108005012' then LocalResid1='2';
```

```
      if v0011='00108005015' then LocalResid1='2';
      if v0011='00108005016' then LocalResid1='2';
      if v0011='00108005021' then LocalResid1='2';
      if v0011='00108005031' then LocalResid1='2';
      if v0011='00108005032' then LocalResid1='2';
      if v0011='00108005033' then LocalResid1='2';
      if v0011='00108005034' then LocalResid1='2';
      if v0011='00108005042' then LocalResid1='2';
    if v0011='00108005046' then LocalResid1='2';
    if v0011='00108005049' then LocalResid1='2';

          /* **** DF - Baixa Renda  ****** */
      if v0011='00108005013' then LocalResid1='3';
      if v0011='00108005014' then LocalResid1='3';
      if v0011='00108005017' then LocalResid1='3';
      if v0011='00108005018' then LocalResid1='3';
      if v0011='00108005019' then LocalResid1='3';
      if v0011='00108005020' then LocalResid1='3';
    if v0011='00108005022' then LocalResid1='3';
      if v0011='00108005023' then LocalResid1='3';
      if v0011='00108005024' then LocalResid1='3';
      if v0011='00108005025' then LocalResid1='3';
      if v0011='00108005026' then LocalResid1='3';
      if v0011='00108005027' then LocalResid1='3';
      if v0011='00108005028' then LocalResid1='3';
      if v0011='00108005029' then LocalResid1='3';
      if v0011='00108005030' then LocalResid1='3';
      if v0011='00108005036' then LocalResid1='3';
      if v0011='00108005037' then LocalResid1='3';
      if v0011='00108005038' then LocalResid1='3';
      if v0011='00108005039' then LocalResid1='3';
      if v0011='00108005040' then LocalResid1='3';
      if v0011='00108005041' then LocalResid1='3';
    if v0011='00108005043' then LocalResid1='3';
      if v0011='00108005044' then LocalResid1='3';
      if v0011='00108005047' then LocalResid1='3';

      if v0011='00108005050' then LocalResid1='3';
    if v0011='00108005051' then LocalResid1='3';

 if v0001=52 then LocalResid1=4;
run;




/* trabalho */

data x.basejovemfiltro3;
set x.basejovemfiltro2;
/* só estuda */
if V0628=1 and v0641=2 or V0628=1 and V0642=2 or V0628=1 and V0643=2  or
V0628=1 and V0644=2
then trabalho=1;
```

```
if V0628=2 and v0641=2 or V0628=2 and V0642=2 or V0628=2 and V0643=2  or
V0628=2 and V0644=2
then trabalho=1;
/* trabalha e estuda */
if V0628=1 and v0641=1 or V0628=1 and V0642=1 or V0628=1 and V0643=1  or
V0628=1 and V0644=1
then trabalho=2;
if V0628=2 and v0641=1 or V0628=2 and V0642=1 or V0628=2 and V0643=1  or
V0628=2 and V0644=1
then trabalho=2;
/* só trabalha*/
if V0628=3 and V0641=1 or V0628=3 and V0642=1 or V0628=3 and V0643=1 or
V0628=3 and V0644=1
then trabalho=3;
if V0628=4 and V0641=1 or V0628=4 and V0642=1 or V0628=4 and V0643=1 or
V0628=4 and V0644=1
then trabalho=3;
/* não estuda e nem trablha */
if V0628=3 and V0641=1 or V0628=3 and V0642=1 or V0628=3 and V0643=1 or
V0628=3 and V0644=1
then trabalho=4;
if V0628=4 and V0641=1 or V0628=4 and V0642=1 or V0628=4 and V0643=1 or
V0628=4 and V0644=1
then trabalho=4;
run;


/* proportional odds test */

%macro simula;
proc sql;
drop table model;
quit;
%do a=100 %to 1000 %by 100;
%do i=1 %to 50;
proc surveyselect data=x.mono1 sampsize=&a seed=&i out=amostra noprint
stats;id _all_;run;
title "Amostra &i";
ods select cumulativemodeltest;
ods output cumulativemodeltest=model&i;
   proc logistic data=amostra;
class sexo raca sitdom relresp quartilrenda trabalho localresid1 /param=ref;
   model ensino= raca idade sitdom relresp localresid1 quartilrenda trabalho;
run;
data model&i;set model&i;amostra=&i;tam_amostra=&a;run;
proc append base=model data=model&i;run;
proc sql;drop table model&i;quit;
%end;
%end;
%mend;
%simula;

/*seleção do tamanho da amostra */

%macro simula;
proc sql;drop table parameters;quit;
```

```
%do a=200 %to 1000 %by 100;
%do i=1 %to 50;
proc surveyselect data=x.mono1 sampsize=&a seed=&i out=amostra noprint
stats;id _all_;run;
title "Amostra &i";
ods select parameterestimates;
ods output parameterestimates=parameters&i;
   proc logistic data=amostra;
class sexo raca sitdom relresp quartilrenda trabalho localresid1 /param=ref;
   model ensino= sexo raca idade sitdom relresp localresid1 quartilrenda
trabalho /link=glogit;
run;
data parameters&i;set parameters&i;amostra=&i;tam_amostra=&a;run;
proc append base=parameters data=parameters&i;run;
proc sql;drop table parameters&i;quit;
%end;
%end;
%mend;
%simula;

/* gráficos média max e min */

title "intercept1";
ods html file='medias.html';
proc gplot data=medias4;
plot mean*tam_amostra min*tam_amostra max*tam_amostra /overlay;
where Variable='Intercept' and ClassVal0='1' and Response='1' ;
symbol1 i=join c=blue;
symbol2 i=join c=green l=2;
symbol3 i=join c=green l=2;
run;
ods html close;

/* seleção das variáveis */

%macro simula;
proc sql;drop table model;quit;
%do i=1 %to 1000;
proc surveyselect data=x.mono1 sampsize=700 seed=&i out=amostra noprint
stats;id _all_;run;
title "Amostra &i";
ods select modelbuildingsummary;
ods output modelbuildingsummary=model&i;
   proc logistic data=amostra;
class sexo raca sitdom relresp quartilrenda trabalho localresid1 /param=ref;
   model ensino= sexo raca idade sitdom relresp localresid1 quartilrenda
trabalho
  / link=glogit selection=stepwise
                    slentry=0.20
                    slstay=0.25;

run;
data model&i;set model&i;amostra=&i;tam_amostra=700;run;
proc append base=model data=model&i;run;
proc sql;drop table model&i;quit;
%end;
```

```
%mend;
%simula;


proc freq data=model;
tables EffectEntered;
run;

/* modelo */

proc logistic data=amostranova;
class sexo raca sitdom relresp quartilrenda trabalho localresid1;
   model ensino= sexo raca idade sitdom relresp localresid1 quartilrenda
trabalho / link=glogit selection=stepwise
                    slentry=0.20
                    slstay=0.25
                    details
                    lackfit;
run;

proc logistic data=amostranova;
class ensino(ref='3') sexo raca sitdom relresp(ref='1') quartilrenda
trabalho(ref=first) ;
   model ensino= sexo idade relresp quartilrenda trabalho / link=glogit;
run



proc logistic data=X.amostranova;
where ensino ~=1;
class ensino(ref='3') sexo raca sitdom relresp(ref='2') quartilrenda
trabalho(ref=first) localresid1/param=ref;
   model ensino=  sexo idade relresp quartilrenda trabalho /link=glogit
aggregate scale=none lackfit;
     run;

proc logistic data=X.amostranova;
where ensino ~=2;
class ensino(ref='3') sexo raca sitdom relresp(ref='2') quartilrenda
trabalho(ref=first) localresid1/param=ref;
   model ensino=  sexo idade relresp quartilrenda trabalho /link=glogit scale=
none aggregate lackfit;
     run;

/* odds ratio - média e intervalo */

%macro simul;
proc sql;drop table odds;quit;
%do i=1 %to 500;
proc surveyselect data=x.mono1 sampsize=700 seed=&i out=amostra noprint
stats;id _all_;run;
title "Amostra &i";
ods select oddsratios;
ods output oddsratios=odds&i;
   proc logistic data=amostra;
```

```
class ensino(ref='3') sexo raca sitdom relresp(ref='2') quartilrenda
trabalho(ref=first) localresid1/param=ref;
   model ensino= sexo idade relresp quartilrenda trabalho /link=glogit ;
run;
data odds&i;set odds&i;amostra=&i;tam_amostra=700;run;
proc append base=odds data=odds&i;run;
proc sql;drop table odds&i;quit;
%end;
%end;
%mend;
%simul;



proc univariate data=odds;
var oddsratioest;
class response effect;
run;
```