DISTORÇÃO IDADE-ESCOLARIDADE: UMA ANÁLISE A PARTIR DE MODELOS DE CONTAGEM

Alysson Lorenzon Portella [1]
Tanise Brandão Bussmann [2]
Ana Maria Hermeto Camilo de Oliveira[3]

RESUMO: Grande esforço vem sendo dedicado para melhor compreender como a educação afeta a vida das pessoas, bem como os determinantes de uma boa educação. Este artigo tem como objetivo colaborar com o debate sobre educação a partir do estudo dos determinantes da distorção idade-escolaridade, sendo esta definida como a diferença entre os anos de estudos esperado de um estudante em certa idade e o número efetivo de anos de estudos que ele completou. A partir de uma análise econométrica, com dados da PNAD 2013, é avaliada a maneira como características pessoais e familiares, como sexo, cor da pele, estar trabalhando, residir na mesma residência da mãe, escolaridade materna e renda familiar per capita, estão relacionadas com a ocorrência de defasagens na escolaridade, assim como sua magnitude. Em especial, este trabalho emprega modelos econométricos que tratam especificamente de dados de contagem, como o Poisson e o Binomial Negativo, e também modelos que trabalham com excesso de zeros, como o *Zero-Inflated* Poisson e o *Zero Inflated* Binomial Negativo. Entre os resultados, destaca-se que a medida em que a idade avança, a ausência de distorção reflete-se em menor probabilidade de ela ocorrer no futuro.
PALAVRAS-CHAVE: Distorção Idade-Escolaridade; Modelos de Contagem; Modelos de Contagem com Zeros Inflados.

ABSTRACT: A great deal of efforts has been placed on the study of how education affects people's lives, as well as what are the determinants of a good education. This paper aims to collaborate with the debate on education dealing with the determinants of the age-grade distortion, defined as the difference in the expected grade of a student of certain age and his actual grade. Using an econometric analysis, with 2013 PNAD database, it assess how personal and familiar features, such as gender, race, dwelling with one's mother, maternal schooling or household's per capita income, is related with the emergence and magnitude of age-grade distortion. Specially, this work employs count data models, such as Poisson and Negative Binomial, and also models that work with excessive zeros, such as Zero-inflated Poisson and Zero Inflated Negative Binomial. Among its results, it finds that as age advances the absence of any grade repetition is related to a smaller probability of one ever comes to happen in the future.
KEY WORDS: Age-Grade Distortion; Count Models, Zero Inflated Count Models.

ÁREA TEMÁTICA: 8 - Econometria

Classificação JEL: I25

[1] Mestrando em Economia pelo Centro de Desenvolvimento e Planejamento Regional - CEDEPLAR/ UFMG. Contato: alyssonportella@gmail.com
[2] Doutoranda em Economia pelo Programa de Pós Graduação em Economia – PPGE/PUCRS. Contato: tanisebrandao@gmail.com
[3] Professora Associada do Centro de Desenvolvimento e Planejamento Regional – CEDEPLAR/UFMG. Contato: ahermeto@cedeplar.ufmg.br

INTRODUÇÃO E REVISÃO DA LITERATURA

A atenção dedicada à educação pela literatura e pelos debates públicos é evidência da sua grande importância na atualidade. São diversos os motivos que justificam a relevância desse aspecto da vida das pessoas. Em primeiro lugar, há a literatura sobre Capital Humano, cujo trabalho seminal pode ser atribuído à Becker (1956). Além disso, conforme coloca Hanushek (2003), as análises empíricas também contribuíram para o debate sobre os retornos consideráveis da educação, explicando os motivos para o grande interesse acadêmico e público sobre o tema. Por outro lado, uma boa educação pode ser vista como um constituinte importante da vida das pessoas e, consequentemente, para a avaliação do desenvolvimento de uma sociedade. Esse papel da educação como elemento constitutivo do desenvolvimento fica evidente pela ótica da Abordagem das Capacitações de Amartya Sen (2000), que coloca a educação tanto como uma importante liberdade instrumental, na medida em que propicia a expansão de outras liberdades, como também substantiva, tendo importância por si só. Sendo assim, um acesso a uma melhor educação facilitaria a possibilidade das pessoas levarem uma vida que elas valorizam.

Pode ser visto que o estudo da educação é justificado por diversos motivos. Um enfoque adotado nas principais abordagens sobre o tema é a função de produção educacional, que, de forma similar à teoria microeconômica, o produto diz respeito ao resultado educacional (que pode ser descrito por anos de estudo ou proficiência em um exame) e os insumos são as características dos indivíduos e do ambiente onde este está inserido (HANUSHEK, 2007). Uma questão pouco explorada, mas relevante e que pode ser analisada utilizando a função de produção educacional é a defasagem idade-escolaridade. A defasagem idade-escolaridade é a diferença entre a idade do estudante e aquela adequada para a série onde ele se encontra. O ideal seria que esta diferença fosse zero, ou seja, que o estudante estivesse na série adequada para sua idade.

Um elevado grau de distorção idade-escolaridade pode afetar a acumulação de capital humano por parte da população, trazendo não apenas consequências para os indivíduos, como também para a sociedade como um todo, afetando tanto o crescimento econômico de longo prazo e retardando a queda na desigualdade. Sendo assim, do ponto de vista social, a distorção idade-escolaridade não somente reduz a velocidade com que se acumularia capital humano, como também afeta o nível máximo que este poderia alcançar.

A partir da visão da Abordagem das Capacitações também é possível encontrar importantes motivos para o estudo da distorção idade-escolaridade. Não ser capaz de concluir etapas dos cursos formais pode afetar as habilidades das pessoas de seguir objetivos por elas almejadas, de modo a incapacitar sua agência. Pode também trazer vergonha ou levar a perda de amigos colegas de classe, sendo esses importantes funcionamentos que constituem a vida das pessoas. Sen (2000) também argumenta em torno de outros aspectos importantes da educação, como a relação entre educação feminina e saúde das crianças. Por estes motivos, estudar os fatores relacionados à distorção idade-escolaridade faz-se importante.

O tratamento da variável dependente dentro da categoria de dados em contagem (ou seja, assumindo apenas o valor zero ou positivos inteiros, indicando o número de anos defasados do estudante) é uma inovação. Geralmente, são utilizados modelos de regressão tradicional (sem restrições para os valores que a variável dependente assume) ou modelos probabilísticos. Nestes casos, a variável dependente é o desempenho em uma prova de proficiência ou a ocorrência de avanço/reprovação do estudante. A

utilização da variável dependente como a defasagem idade-escolaridade também não é trivial.

Os principais motivos para a existência de defasagem idade-escolaridade são a reprovação, quando o aluno precisa repetir a série em questão, o abandono escolar, quando o aluno deixa de frequentar a escola por um período ou, por fim, a matrícula tardia do estudante na escola. Estes eventos também são relacionados entre si. Uma elevada distorção idade-escolaridade pode resultar em abandono, conforme mostram Fritsch, Vitelli e Rocha (2014), ao perceber que a defasagem entre a idade recomendada e a série frequentada tem grande impacto sobre a taxa de abandono dos estudantes.

A situação da distorção idade-escolaridade no Brasil apresentou uma melhora nas últimas décadas, estagnando em um patamar onde aproximadamente metade dos estudantes têm pelo menos um ano de defasagem idade-escolaridade. Ao observar a evolução da defasagem idade-escolaridade ao longo do tempo, Riani (2005) mostra que em 1980, 78% dos estudantes apresentava idade superior a adequada, enquanto em 2000 este número foi reduzido para 54%. Fernandes e Natenzon (2003) observam uma redução no percentual de estudantes fora da idade correta. Em 1995, 57% das crianças que deveriam estar na quarta série estavam fora desta série, enquanto em 2009 houve a redução para 44%.

Para Ferrão *et al* (2001) e Franco (2008), há uma relação clara entre a distorção idade-escolaridade e um pior desempenho. Segundo Ferrão *et al* (2001, p.119-120): "Torna-se evidente que os alunos com atraso escolar têm resultados escolares reduzidos comparativamente aos que estão na idade adequada para a série". Ainda, Machado (2005) observa que uma maior distorção idade-escolaridade se relaciona de forma positiva com a probabilidade de abandono escolar.

Franco (2008) constrói um painel de escolas com dados do SAEB de 1999 a 2005, e observa uma relação negativa entre a defasagem idade-escolaridade e o desempenho escolar, mensurado pela proficiência no SAEB em um modelo de efeitos fixos. Alves, Ortigão e Franco (2007), observam quais os efeitos de cada um dos diferentes capitais para a repetência escolar. Os estudantes do sexo masculino, negros e que trabalham tem um risco maior, e o nível sócio econômico é especialmente importante para os estudantes brancos.

Machado e Gonzaga (2007) observam o efeito da renda e da educação dos pais sobre a existência de defasagem idade-escolaridade. Os autores utilizam variáveis instrumentais para as variáveis explicativas, buscando evitar a estimação com possível viés. Machado e Gonzaga (2007) encontram uma redução na probabilidade de distorção idade-escolaridade quanto maior a renda e educação dos pais. Outras variáveis, como o gênero masculino e não brancos e amarelos apresentam um nível maior de vulnerabilidade à defasagem idade-escolaridade.

Pontili e Kassouf (2008), com dados do Censo Demográfico e Censo Escolar de 2000, estimam um modelo probit e um probit ordenado para encontrar a probabilidade dos estudantes frequentarem séries adequadas para o seu nível para o primeiro modelo e para o último, quantos anos acima do indicado pela idade-escolaridade. As características individuais e familiares (renda per capita, educação, sexo e idade do chefe de família e sexo, idade e cor do estudante) foram importantes. Houve maior probabilidade de aumento da defasagem idade-escolaridade para estudantes do sexo masculino, negros ou pardos, e com menor renda *per capita*, enquanto que se tomar para o chefe de família essa probabilidade aumenta quando ele tem menor idade e educação e é do sexo masculino. A idade apresentou sinal positivo em Pernambuco e negativo em São Paulo para o aumento da defasagem idade-escolaridade. Para comparação, utilizou-se informações de São Paulo e Pernambuco, sendo que as

variáveis familiares do primeiro grupo obtiveram um maior impacto. Nos dois casos, uma variável importante na escola foi o número de laboratórios de informática.

Soares e Sátyro (2008), com dados do Censo Escolar de 1998 a 2005, observam a relação entre a defasagem idade-escolaridade e variáveis da escola, especificamente de infraestrutura e relativas a qualidade dos professores. Os autores observam uma relação negativa entre essas variáveis, tanto na abordagem paramétrica quanto não-paramétrica.

Riani (2005) utiliza dados do Censo Demográfico de 1980 a 2000 para analisar a defasagem idade-escolaridade. A autora conclui que indivíduos do sexo masculino e também negros são os que tem a maior defasagem. A maior incidência de distorção idade-escolaridade para indivíduos do sexo masculino também foi encontrada para Leon e Menezes-Filho (2002) com dados da PME de 1984 a 1997.

Leon e Menezes-Filho (2002) usando um modelo heckprobit, observam que morar com os pais aumentam as chance de reprovação para o 3º ano do ensino médio, enquanto que para o 4º ano do ensino fundamental a probabilidade é menor, quando comparada a viver com apenas um ou nenhum dos pais. Alunos de maior idade também apresentaram menor probabilidade do avanço escolar, indicando inclusive uma menor chance em concluir os ciclos escolares. Para estudantes reprovados, há uma maior chance de abandono caso eles vivam sem os pais, *vis-a-vis* estudantes que moram com 1 ou os 2 pais. Os resultados não são fixos ao longo do tempo, pelas simulações incluídas por Leon e Menezes-Filho (2002), indicando uma menor dependência na maioria das variáveis independentes, quando comparados os valores de 1984-1985 a 1996-1997.

Não é apenas o desempenho individual dos estudantes que é prejudicado com uma maior dispersão da defasagem idade-escolaridade. Para Machado, Firpo e Gonzaga (2013), com dados do SAEB de 2011, observam que quanto maior a dispersão em termos de idade, menor a proficiência das crianças.

A ocorrência de defasagem idade-escolaridade não é homogênea em todo o território nacional. Para Leon e Menezes-Filho (2002), estudantes da região metropolitana de São Paulo são menos propensos à reprovação em relação àqueles de outras regiões metropolitanas do país. Além disso, para Machado e Gonzaga (2007) os estudantes do nordeste, norte e centro-oeste tem oportunidades menores do que os estudantes do Sul para evitar a defasagem idade-escolaridade. A residência urbana também é um fator importante para evitar a distorção idade-escolaridade, para Machado e Gonzaga (2007).

Apesar de apenas a variável distorção idade-escolaridade ser utilizada de forma direta neste trabalho, a reprovação é uma das maneiras mais efetivas de aumentar a defasagem idade-escolaridade do aluno. Para Ribeiro (1991), a repetência é o principal problema para o aumento da escolaridade no caso brasileiro, aumentando as chances do aluno evadir posteriormente. Barros e Mendonça (1998), com dados do Censo Demográfico, mostram que as reprovações, eventos altamente relacionados com a defasagem idade-escolaridade, apresentam um efeito negativo para a autoestima dos estudantes, que aumentam a probabilidade de reprovações subsequentes e também para o estado, pois há maiores gastos com as reprovações. Este aumento na probabilidade de reprovações posteriores também foi encontrado em um período mais recente por Souza *et al* (2012), com dados da PME.

Tendo em vista o colocado acima, este trabalho busca estudar os fatores relacionados com a distorção idade-escolaridade das crianças e jovens brasileiros a partir da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2013. Na próxima seção é exposta a metodologia utilizada, que engloba algumas definições que

permitiram a construção da base de dados e também os métodos econométricos, que incluem o modelo de Mínimos Quadrados Ordinários, Modelo Poisson e Modelo Binomial Negativo e, os modelos com zeros inflados, Poisson (*Zero-Inflated Poisson* ou ZIP) e Binomial Negativo (*Zero-Inflated Negative Binomial* ou ZINB). Em seguida são apresentadas as estatísticas descritivas, as estimações e as previsões realizadas por estes modelos. Então, é apresentada a síntese dos resultados deste trabalho, na conclusão.

METODOLOGIA

Nesta seção, serão expostas inicialmente as definições adotadas para a obtenção das variáveis de interesse. A base de dados utilizada foi a PNAD, no entanto foram necessárias algumas definições para o cálculo de certas variáveis. Então são expostos os métodos econométricos utilizados.

Com o objetivo de observar quais são os principais fatores individuais e familiares relacionados com a distorção idade-escolaridade, o primeiro passo é construir uma variável indicativa da existência desta distorção. Uma definição de defasagem idade-escolaridade é sugerida por Machado e Gonzaga (2007, p.456): "a criança é considerada atrasada em termos educacionais se não tem o total de anos de estudo completos compatível com a sua idade no início de cada ano letivo". Sendo assim, nos casos onde a diferença entre a idade e o total de anos de estudo ideal é negativo foram considerados como zero, pois, para a defasagem idade-escolaridade apenas interessam os valores positivos, onde o estudante apresenta atraso escolar. Esta definição é distinta da do INEP, que define como defasagem apenas os casos em que os alunos têm dois anos de atraso em relação à idade recomendada. A razão para isso pode ser encontrada em Sampaio e Nespoli (2004), ao afirmarem que se busca evitar o caso em que uma criança inicia seus estudos aos sete anos, mas complete 8 durante o ano letivo.

Como a legislação brasileira coloca como obrigatória a matrícula de estudantes a partir dos 4 anos para a educação infantil, até os 5 anos, e então encaminhamento para o ensino fundamental (BRASIL, 1996, 2013), considerou-se a definição de Machado e Gonzaga (2007) mais coerente. Para fins de comparação com as estatísticas oficiais do Brasil, é necessário lembrar que os dados deste trabalho irão superestimar a defasagem idade-escolaridade, em relação às informações oficiais.

A população de interesse é composta pelos estudantes em idade escolar no Brasil. Como ainda existem indivíduos no sistema antigo de ensino básico, com 11 ao invés de 12 anos de estudo e ingresso aos 7 anos completos na 1ª série – ao invés de 6 anos completos no primeiro ano, consideram-se defasados apenas os estudantes que, com 8 anos de idade, não tinham 1 ano completo de escolaridade, e assim sucessivamente. Sendo assim a amostra utilizada foram todas as crianças e adolescentes de 8 a 18 anos de idade, estando elas matriculadas ou não em qualquer instituição de ensino, de todos os estados brasileiros. Espera-se que uma criança de 8 anos tenha completado ao menos um ano de escolarização, enquanto um jovem de 18 anos deve ter completado todos os anos do Ensino Fundamental e Médio, totalizando 11 anos de estudos. Diferença nesses números resultam em distorção idade-escolaridade, que pode ir de zero (para quem não apresenta distorção) até 11 anos - quando o aluno não completou nenhum ano de estudo.

Os modelos propostos empregaram o mesmo conjunto de variáveis. A variável dependente será o número de anos de distorção idade-escolaridade da criança ou adolescente, calculado conforme especificado acima. As variáveis independentes

englobam a idade, sexo, cor da pele (brancos e não-brancos, sendo amarelos enquadrados como brancos), se está matriculado ou não em escola e em qual rede de ensino, a pessoa trabalhar ou não, uma variável indicativa da moradia da mãe no mesmo domicílio que a pessoa, a educação da mãe, o logaritmo da renda *per capita* familiar, se o domicílio se encontra em zona rural ou urbana, *dummy* de região e um intercepto.

Para observar a relação entre a distorção idade-escolaridade e suas covariadas, são propostos quatro métodos, além da estimação via Mínimos Quadrados Ordinários (MQO): i) modelo Poisson, ii) modelo Poisson com zeros inflados (*Zero-Inflated Poisson* ou ZIP), iii) modelo Binomial Negativo (*Negative Binomial* ou NB) e iv) modelo Binomial Negativo com zeros inflatos (*Zero-Inflated Negative Binomial* ou ZINB).

Todos os modelos estimados baseiam-se em dados de contagem, ou seja, dados que tomam apenas a forma de inteiros não-negativos, sem um limite claro (WINKELMANN, 2008). Entre os exemplos propostos por Cameron e Trivedi (2013) para ilustrar esse tipo de dados estão o número ligações em um *call* center, número de faltas no trabalho e até mesmo avaliações de crédito de agências de *rating*. Logicamente, a variável distorção idade-escolaridade, conforme aqui definida, se enquadra nessa categoria de dados.

O Modelo Poisson é o ponto de partida quando se lida com dados de contagem. Como o próprio nome diz, o modelo se baseia na especificação de uma distribuição Poisson para os dados, conforme descrito em (1).

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2 \ldots \quad (1)$$

Onde, μ denota tanto a média como a variância de y – a chamada propriedade de equidispersão. O modelo de regressão então assume que o parâmetro μ está relacionado com o vetor de regressores x de acordo com a equação (2).

$$\mu_i = \exp(x_i'\beta), i = 1, \ldots N \quad (2)$$

A forma exponencial garante que μ seja positivo para qualquer combinação de parâmetros ou variáveis explicatórias (WINKELMANN, 2008). A partir dessas especificações e supondo que as observações são independentes, pode-se estimar o vetor de coeficientes β por máxima verossimilhança. A consistência do Modelo Poisson requer a correta especificação da média condicional, não sendo necessária a correta especificação da distribuição da variável dependente. Porém, para inferência, também é necessário que a variância seja corretamente especificada e igual a média, ou seja, requer-se a propriedade equidispersão[4] (CAMERON; TRIVEDI, 2013). Esta última hipótese é geralmente violada, havendo diversas causas para sua ocorrência e uma série de possibilidades de tratamento[5]. Quando da interpretação do modelo, é importante ter em mente que o efeito marginal de x_j na esperança condicional de y é

[4] Equidispersão estaria para o Modelo Poisson como homocedasticidade está para o MQO, embora a magnitude da distorção nas estimativas do erro padrão e estatística t provocada pela ausência do primeiro pode vir a ser bem maior, de acordo com Cameron e Trivedi (2013, pg. 89).

[5] Ver Cameron e Trivedi (2013) para maiores detalhes.

dada por (3). Ou seja, o efeito da variação de uma variável explicativa irá variar conforme o indivíduo[6].

$$\frac{\partial E(y|x)}{\partial x_j} = \beta_j \exp(x'\beta) \quad (3)$$

O modelo de Poisson é uma das formas funcionais de modelos de contagem mais utilizadas. Porém, existem algumas limitações no uso deste modelo. Entre as possíveis fontes de problemas para a estimação do modelo Poisson, pode-se citar duas principais: heterogeneidade não observada e excesso de zeros (o modelo prevê menos zeros do que a quantia presente nos dados). Essa primeira decorre do fato de o modelo Poisson padrão não permitir a modelagem das heterogeneidades não-observadas. Isso porque a média (ou a taxa na qual os eventos ocorrem) é uma função determinística dos regressores, como pode ser visto em (2)[7]. Porém, caso haja algum fator não observado que afete a taxa na qual os eventos ocorrem, haverá problemas para a estimação, especialmente pelo fato de que heterogeneidade não observada resultar em sobredispersão (variância maior que a média).

Seguindo Winkelmann (2008) é possível lidar com algumas das limitações do modelo Poisson. O procedimento padrão para tratar essa heterogeneidade é adicionar um termo aleatório multiplicativo na equação da média condicional, u, onde u é definido como $\exp(v)$. Desta forma, é possível reescrever (2) adicionando este termo e chegando em (4).

$$\mu_i = E(y|x,u) = \exp(x_i'\beta)\, u, \;\; i = 1, \dots N \quad (4)$$

Assume-se que $E[\exp(v)\,|x] = E[\exp(v)]$. É preciso então modelar essa heterogeneidade. A metodologia mais aplicada em trabalhos empíricos consiste nos chamados modelos mistos paramétricos (CAMERON; TRIVEDI, 2013), que irão combinar uma distribuição Poisson com uma distribuição paramétrica para o termo u. Entre as possibilidades, a mais utilizada é a distribuição Gama, com u seguindo uma distribuição de acordo com os parâmetros α e β, ou seja, $u \sim \Gamma(\alpha, \beta)$. Assumindo que os valores de α e β são iguais, de modo a adicionar apenas um parâmetro, a integração das distribuições Gama e Poisson leva a uma distribuição Binomial Negativa, que pode ser descrita em (5).

$$f(y|\alpha,\mu) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\mu+\alpha}\right)^{\alpha} \left(\frac{\mu}{\mu+\alpha}\right)^{y} \quad (5)$$

Especificando o valor de $\mu = \exp(x'\beta)$ e também $\alpha = \sigma^{-2}$, é possível reescrever a esperança de y e também sua variância condicional a x de acordo com (6) e (7):

$$E(y|x) = \mu = \exp(x'\beta) \quad (6)$$

$$Var(y|x) = exp(x'\beta) + \sigma^2[\exp(x'\beta)]^2 \quad (7)$$

[6] Para evitar esse efeito individual, pode-se trabalhar com o Efeito Marginal Médio, o Efeito Marginal na Média ou até mesmo o Efeito Marginal em um valor representativo. Esses não serão aqui empregados, por isso são omitidos. Ver Cameron e Trivedi (2013, p. 92-93).

7 O fato de a taxa na qual o evento ocorre não significa que a variável dependente não seja aleatória. Ela mantém essa característica pois μ determina apenas a taxa na qual o evento ocorre, não se ele irá ocorrer ou em qual quantidade de vezes, que é sujeito a processo Poisson, que é aleatório.

Por (6) e (7) observa-se que a variância será sempre maior que a média, o que significa que este modelo é capaz de lidar com a sobredispersão. Essa parametrização leva o nome de Binomial Negativa II (NB2). Assim como o modelo Poisson, o NB2 é estimado a partir de sua função de máxima verossimilhança. Nesse caso, o parâmetro α toma a forma de uma constante. Caso este parâmetro seja estimado de acordo com as observações da amostra, tem-se um modelo Binomial Negativo Generalizado .

O outro problema que pode ocorrer em alguns modelos de contagem é a dificuldade na previsão do número adequado de eventos zero, ou seja, o modelo prevê poucos resultados zero em relação ao que seria esperado, havendo um excesso de zeros. O excesso de zeros também ser uma das causas da sobredispersão (WINKELMANN, 2008). Uma solução é dividir o processo em duas partes: primeiro, introduz-se um componente distinto com o objetivo de aumentar a probabilidade de ocorrência de zeros em π. O outro componente seria um processo de contagem padrão, com função densidade f_2, que ocorre com probabilidade $(1-\pi)$. A forma funcional de f_2 pode ser alterada de acordo com as hipóteses do modelo, assim como a probabilidade π pode ser uma constante ou função de um dado conjunto de regressores por meio de um processo binário qualquer. Então, por (8) pode-se descrever o modelo com zeros inflados (CAMERON; TRIVEDI, 2013).

$$\Pr[y=j] = \begin{cases} \pi + (1-\pi)f_2(y), \ se\ y = 0 \\ (1-\pi)f_2(y), se\ y \geq 1 \end{cases} \tag{8}$$

No caso do modelo de Poisson, esta modificação é denominada *Zero Inflated Poisson* ou ZIP. Para o Binomial Negativo, o modelo é chamado de *Zero Inflated Negative Binomial* ou ZINB. Estes modelos suplementam a estimação inicial com uma nova distribuição binária que irá especificar se a variável dependente toma um valor zero ou será descrita por um processo estocástico. Sendo assim, há um processo binário inicial, f_1, que é uma função densidade de probabilidade degenerada, e, caso o resultado seja a ocorrência do evento (o processo gerar o resultado 1), o valor previsto para y será zero. Caso contrário, parte-se para o processo de contagem, descrito por f_2 em (8). Ainda, observa-se que é possível que ocorra a previsão do valor zero neste caso também. Dessa forma, os modelos de zeros inflados podem resultar em zeros de duas formas: a partir do processo binário f_1, ou pela função de densidade padrão f_2. O processo binário geralmente é descrito por uma função *logit* enquanto a função densidade é um processo de Poisson ou Binomial Negativo[8].

Para escolha do modelo mais adequado, uma possibilidade é observar os critérios de informação AIC e BIC. Estes critérios são capazes de comparar a qualidade de cada modelo. A avaliação de cada modelo é realizada de acordo com a função de verossimilhança, onde os critérios AIC e BIC são modificações desta (GREENE, 2012). O teste Vuong (1989), utilizado para comparar modelos não-aninhados permite a comparação entre os modelos com zeros inflados e o Poisson. Com relação a sobredispersão, há algumas possibilidade de testes, conforme apresentados em Cameron e Trivedi (2013, seção 3.4) embora aqui destaque-se apenas a significância

8 A interpretação e o cálculo dos efeitos marginais desses modelos se faz mais complexa do que os modelos tradicionais de contagem, uma vez que são compostos por dois processos. Para maiores detalhes, ver Winkelmann (2008).

do parâmetro α que modela a variância. Se este for significativamente diferente de zero, os modelos com sobredispersão são considerados mais adequados que o Poisson.

Sendo assim, passa-se a exposição da análise descritiva e estimação dos modelos de mínimos quadrados, Poisson, Binomial Negativo, e suas versões para zeros inflados.

ANÁLISE DESCRITIVA E ECONOMÉTRICA

Para conseguir estimar de forma adequada os principais fatores relacionados com a defasagem idade-escolaridade, optou-se pela utilização da PNAD, com a base de dados de indivíduos, para o ano de 2013. Ainda, foi realizado o recorte de idade, de 8 a 18 anos, com o objetivo de conseguir contemplar todos os indivíduos em idade escolar que poderiam ter pelo menos um ano de defasagem idade-escolaridade.

Todos os modelos estimados consideram o mesmo conjunto de variáveis: idade, mulher (1 para sexo feminino e 2 para masculino), branco (1 para brancos e amarelos e 0 para negros e pardos), variável categórica ensino, que se divide em duas variáveis: na primeira, ensino_publico, o valor 1 é para os indivíduos de escolas públicas e zero para indivíduos que não estão matriculados em nenhuma escola ou em uma escola privada, e na variável ensino_privado, onde o valor 1 é para os indivíduos que estão matriculados em escolas privadas e zero caso contrário. Também foi incluída a variável trabalha (1 para quem trabalhou na semana de referência e 0 para quem não trabalhou), mae_mora (1 se a mãe mora com a pessoa e 0 se não mora), educ_mae (anos de estudo da mãe), log_rend_pc (logaritmo da renda familiar *per capita*), urbano (1 se a pessoa mora na zona urbana e 0 se mora na zona rural) e a variável categórica região, onde, todas as regiões menos a região norte possuem uma variável indicativa. (1 região norte, 2 nordeste, 3 sudeste, 4 sul e 5 centro-oeste).

Em relação às estatísticas descritivas destas variáveis, pela tabela 1, observa-se que 48% da amostra é feminina, com uma média de 13 anos de idade, com 41,17% de cor branca ou amarela. A grande maioria frequenta o ensino público e mora com a mãe. Uma parcela pequena da amostra trabalhou na semana de referencia (12,9%). A grande maioria das famílias mora em residências urbanas e na região sudeste e nordeste do país. A renda média *per capita* foi de R$1.053,e a educação da mãe foi em média de 7 anos de estudo. As duas ultimas variáveis, que indicam se o indivíduo está defasado, observa-se que 52% apresentam uma defasagem positiva. Em média, esta defasagem é de pouco mais de um ano, conforme mostra a tabela 1.

Os resultados dos modelos estimados podem ser observados nas tabelas 2 e 4, com os resultados dos modelos MQO, Poisson, NB2, ZIP e ZINB, usando como variável dependente o número de defasagens escolares. Os modelos ZIP e ZINB contam com a coluna referente ao processo *logit*, responsável pelo aumento no número de zeros. Os modelos NB2 e ZINB ainda apresentam uma variável referente à modelagem do termo de heterogeneidade não-observada, responsável pelo aumento da variância em relação a média, que aqui é tomada apenas como uma constante, referida como lnalpha nas tabelas 2 e 4 e α na equação (6).

Os resultados esperados são que todas as variáveis independentes apresentem sinal negativo, com exceção da idade, que deve ter uma relação positiva com a variável dependente. Também se espera um sinal positivo de trabalha, uma vez que crianças e jovens que trabalham teriam menos tempo para os estudos. Desse modo, espera-se que os modelos apontem um número menor de defasagens para mulheres ou para aqueles que moram com suas mães. De modo equivalente, espera-se que quanto maior a educação da mãe ou a renda familiar *per capita*, menor o número de defasagens.

Tabela 1 – Estatísticas Descritivas da Amostra

Variável	Média	Variável	Média
Mulher	0,489	log_rend_pc	6,960
Idade	13,170	Nordeste	0,311
Branco	0,412	Sudeste	0,388
mae_mora	0,867	Centro-Oeste	0,133
Ensino_Público	0,770	Sul	0,062
Ensino_Privado	0,129	educ_mae	7,677
trabalha	0,120	defasado	0,528
urbano	0,821	ndefasado	1,075

Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Pode-se observar pelas tabelas 2 e 4 que os coeficientes apresentaram o sinal esperado, assim como todos foram altamente significativos. Estes resultados são coerentes com a literatura sobre o tema. O modelo Poisson não apresentou coeficientes muito diferentes do modelo NB2, sendo no geral levemente superiores. Entre os resultados, o mais impressionante é o fato de mulheres terem mais de 30% menos defasagens que homens em média, enquanto os brancos tendem a ter cerca de 15% menos. A educação da mãe também apresenta um importante papel, sendo que cada ano a mais de educação leva a uma redução de 6% no número de defasagens. Coeficiente similar foi encontrado para a presença da mãe no domicílio. Já uma variação de 1% na renda familiar *per capita* - que por estar na forma logarítmica pode ser interpretado como elasticidade - tem um efeito relativo na defasagem de 0,5-0,6% no número de defasagens. Logicamente esse número parece pequeno, mas o acúmulo de grandes variações nesse parâmetro podem se revelar importantes. Estar matriculado em escola privada mostra ter um efeito maior sobre o número de defasagens, sendo este praticamente o dobro com relação ao ensino público, quando compara-se ambos aos resultados alcançados por aqueles que não estão matriculados.

Uma variável que apresentou sinal diferente do esperado foi trabalha, com um sinal negativo, significando que aqueles que trabalhavam tinham um número menor de defasagens que aqueles que não trabalhavam, contradizendo o que se deveria esperar. Esse fato pode ser explicado pela tabela 3 abaixo. Como pode ser visto, a diferença nas proporções de presença ou não de distorções é relacionada com o fato da criança ou jovem estar matriculada ou não em alguma rede de ensino (na tabela sintetizado apenas como estuda ou não estuda). Desta forma, ao controlar por essa variável, o efeito de trabalhar ou não parece favorecer a ausência de distorção. Uma possível explicação seria que alunos que trabalham e estudam podem ser muito mais motivados que aqueles que não trabalham nem estudam, o que poderia se refletir em uma presença menor de distorção idade-escolaridade, embora o uso de uma variável de iteração não tenha sido empregada para captar tal efeito.

Tabela 2- Defasagem idade escolaridade: Estimações via MQO e modelos de contagem

	MQO	Poisson	NB2
Variáveis	ndefasado	ndefasado	ndefasado
idade	0,116***	0,117***	0,119***
	(0,002)	(0,002)	(0,0018)
mulher	-0,366***	-0,336***	-0,319***
	(0,012)	(0,012)	(0,001)
branco	-0,168***	-0,169***	-0,152***
	(0,013)	(0,013)	(0,0113)
Ensino_público	-0,985***	-0,447***	-0,434***
	(0,022)	(0,019)	(0,0152)
Ensino_privado	-1,253***	-0,970***	-0,977***
	(0,029)	(0,035)	(0,0261)
trabalha	-0,191***	-0,162***	-0,157***
	(0,02)	(0,018)	(0,0151)
mae_mora	-0,087***	-0,0744***	-0,0729***
	(0,019)	(0,017)	(0,0146)
educ_mae	-0,07***	-0,062***	-0,061***
	(0,002)	(0,002)	(0,0013)
log_rend_pc	-0,005***	-0,005***	-0,006***
	(0,002)	(0,001)	(0,0011)
urbano	-0,082***	-0,055***	-0,063***
	(0,016)	(0,013)	(0,013)
Nordeste	-0,145***	-0,117***	-0,08***
	(0,021)	(0,013)	(0,013)
Sudeste	-0,443***	-0,382***	-0,278***
	(0,021)	(0,016)	(0,015)
Sul	-0,387***	-0,314***	-0,227***
	(0,025)	(0,021)	(0,018)
Centro-Oeste	-0,355***	-0,284***	-0,225***
	(0,03)	(0,022)	(0,021)
Constante	1,797***	-0,057	-0,139***
	(0,047)	(0,039)	(0,037)
lnalpha	-	-	-0,859***
	-	-	(0,02)
Número de Observações	57.880	57.880	57.880

Desvios Padrões Entre Parênteses. ***p<0,01,**p<0,05,*p<0,1
Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Na análise para os casos ZIP e ZINB, na tabela 4, vê-se duas mudanças importantes. A primeira delas é que alguns coeficientes se alteram, no geral reduzindo seu impacto (ou seja, coeficientes menores em termos de modulo, uma vez que nenhum sinal inverte) quando toma-se para comparação apenas a porção Poisson (ou binomial negativo) do modelo, desconsiderando a parte *logit*. Desta forma, considerando intervalos de confiança de 95%, pode-se dizer que a mudança no coeficiente das variáveis mulher e educ_mae irão reduzir significativamente de magnitude, enquanto

que as variáveis referentes à idade e ensino em escola pública irão aumentar sua magnitude, conforme o APÊNDICE A. Outras mudanças nos coeficientes não se mostraram significativas pela comparação dos intervalos também no APÊNDICE A, embora em duas delas (branco e ensino privado) a mudança foi significativa para o modelo ZIP, mas não ZINB.

Tabela 3 – Proporção de estudantes que trabalham e estudam de acordo com a presença de distorção idade-escolaridade

Distorção	Trabalha			Não Trabalha		
	Estuda	Não Estuda	Total	Estuda	Não Estuda	Total
Sem distorção	38,05	39,22	38,4	45,48	33,77	44,76
Com distorção	61,95	60,78	61,6	54,52	66,23	55,24

Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Porém, é análise da porção *inflated* do modelo que traz aspectos mais interessantes dessas regressões. Antes de mais nada, é importante que o leitor tenha claro o funcionamento dos modelos com zeros inflados, que dividem o processo de geração dos eventos em duas partes (conforme explicado na seção metodológica). Deve-se também deixar claro que os resultados apresentados estão na forma de coeficientes, não *odds ratio*. Dadas essas considerações, aponta-se primeiro para a não significância estatística das variáveis trabalha, log_rend_pc e urbano. Ou seja, elas parecem não afetar a probabilidade das crianças e jovens de apresentarem ou não distorção, embora elas estejam ligadas à magnitude dessa distorção, conforme mostra a parte Poisson ou NB2 da modelagem. Sendo assim, espera-se que, *ceteris paribus*, um jovem que trabalha, cuja renda familiar *per capita* seja alta e que mora na zona urbana, apresente um número menor de defasagens, *caso ele esteja sujeito a ocorrência de distorção*. Agora, quando se refere a determinar se essa criança ou jovem está sujeito ou não à ocorrência da defasagem idade-escolaridade, tratado na porção *logit* do modelo, essas variáveis não se mostram significativas. Logo, as chances das crianças e jovens pertencerem ao grupo de alunos que nunca irão ter qualquer distorção idade-série não são afetadas pelas variáveis mencionadas.

Outro resultado interessante é que estudar em escolas públicas aumenta a chance de ocorrência de distorções em relação àqueles que não estão matriculados, enquanto que aqueles que estudam em escolar particulares tem suas chances reduzidas em relação àqueles. Esse é um resultado de certa forma curioso, ao imaginar que aqueles matriculados em uma escola pública tem mais chances de apresentarem defasagens que aqueles não matriculados. Esse paradoxo pode ser respondido ao se considerar que uma parte daqueles que não estão matriculados realmente terminaram seus estudos no período adequado. De fato, a proporção de jovens sem distorção é de 35% entre aqueles que não estão matriculados em escolas, sendo 41% para escola públicas e 67% para privadas. Porém, ao se selecionar apenas aqueles com 18 anos de idade, que já podem ter concluído seus estudos, vê-se que quase 50% deles não apresentam qualquer distorção entre não matriculados, enquanto esse número cai para menos de 10% nas escolas públicas, mas sobe para 77% no caso das escolas privadas (TABELA 5). Desta forma, ao controlar por idade, é possível entender o motivo do ensino público estar mais ligado à ocorrência de repetências do que não estar matriculado em qualquer rede de ensino.

Tabela 4 - Distorção idade escolaridade: Estimações via Modelos *zero-inflated*

	ZIP		ZINB	
Variáveis	ndefasado	inflado	Ndefasado	Inflado
idade	0,148***	0,234***	0,142***	0,343***
	(0,0019)	(0,0094)	(0,002)	(0,0149)
mulher	-0,193***	0,707***	-0,223***	0,944***
	(0,012)	(0,0488)	(0,0117)	(0,0695)
branco	-0,111***	0,314***	-0,132***	0,357***
	(0,0137)	(0,0499)	(0,0132)	(0,0686)
Ensino_público	-0,623***	-0,681***	0,696***	-1,372***
	(0,0168)	(0,0535)	(0,018)	(0,101)
Ensino_privado	-0,789***	0,580***	0,914***	0,399***
	(0,0452)	(0,0919)	(0,0427)	(0,119)
trabalha	-0,141***	0,0615	0,142***	0,01
	(0,0171)	(0,0557)	(0,0175)	(0,0732)
mae_mora	-0,015	0,486***	-0,036**	0,696***
	(0,0163)	(0,0787)	(0,0162)	(0,109)
educ_mae	-0,042***	0,106***	-0,048***	0,135***
	(0,0017)	(0,0067)	(0,0016)	(0,0096)
log_rend_pc	-0,004***	0,003	-0,003***	0,006
	(0,0013)	(0,0041)	(0,0012)	(0,0055)
urbano	-0,059***	-0,002	-0,06***	-0,017
	(0,0133)	(0,0651)	(0,0128)	(0,092)
Nordeste	-0,097***	0,153**	-0,109***	0,127
	(0,0136)	(0,0643)	(0,0132)	(0,0916)
Sudeste	-0,322***	0,286***	-0,341***	0,301***
	(0,0166)	(0,0688)	(0,0157)	(0,0954)
Sul	-0,237***	0,483***	-0,276***	0,441***
	(0,0222)	(0,0823)	(0,0213)	(0,118)
Centro-Oeste	-0,262***	0,108)	-0,258***	0,164
	(0,0236)	(0,0926)	(0,0222)	(0,126)
Constante	-0,455***	-6,413***	-0,298***	-8,968***
	(0,0378)	(0,198)	(0,0414)	(0,314)
lnalpha	-	-	-1,524***	-
	-	-	(0,041)	-
Número de Observações	57.880	57.880	57.880	57.880

Desvios Padrões Entre Parênteses. ***p<0,01,**p<0,05,*p<0,1 Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Ainda destaca-se o coeficiente relativo à variável idade. Levando-se em conta os resultados presentes nas outras regressões, seria esperado que a probabilidade de ocorrência de distorção fosse diretamente proporcional a idade. Entretanto, no caso da porção *logit* o coeficiente foi positivo, indicando que as chances de ocorrência de nenhuma distorção aumenta conforme as crianças e jovens vão ficando mais velhos, o que parece um contrassenso. A explicação para esse suposto sinal trocado se encontra

na forma que o modelo ZIP e ZINB assumem. Como eles dividem em dois o processos de geração dos eventos, eles também acabam por dividir a amostra em dois grupos de indivíduos: aqueles que estão sujeitos a apresentar alguma distorção, e aqueles que nunca irão apresentar qualquer distorção. Assim, conforme mais velho fica um aluno, menores as chances de ele apresentar uma distorção, *caso isso não tenha ocorrido até então*. Ou seja, quando se toma o processo Poisson (ou NB2), se está pensando em um aluno que está sujeito à ocorrência da distorção idade-escolaridade, restando apenas a determinação de quantas serão (podendo inclusive ser zero, conforme explicado na seção metodológica). Já quando se toma o processo *logit*, o que será determinado é se o aluno estará ou não sujeito à ocorrência da distorção. Caso ele venha a ter sucesso nesse processo (o processo binário gerar um 1), então o aluno nunca irá apresentar um valor positivo na defasagem idade-escolaridade, nem mesmo de um ano. Logo, quanto maior a sua idade, maiores as chances de um aluno nunca vir a ter defasagem idade-escolaridade, caso isso nunca tenha ocorrido. Um aluno que não apresentou defasagem até o primeiro ano do Ensino Médio, terá grandes chances de não apresentar a defasagem idade-escolaridade nos dois anos que lhe restam de escolarização. Já um aluno que já possuía um valor positivo na defasagem idade-escolaridade na quinta série do Ensino Fundamental, terá mais chances de aumentar o valor desta nas séries seguintes.

Tabela 5- *Status* de Ensino de acordo com a idade e presença de distorção idade-escolaridade

Distorção	**Ensino Público**		**Ensino Privado**		**Não Matriculado**	
	Qualquer idade	18 anos	Qualquer idade	18 anos	Qualquer idade	18 anos
Sem distorção	41,93	8,62	67,69	77,25	35,88	48,9
Com distorção	58,07	91,38	32,31	22,75	64,12	51,1

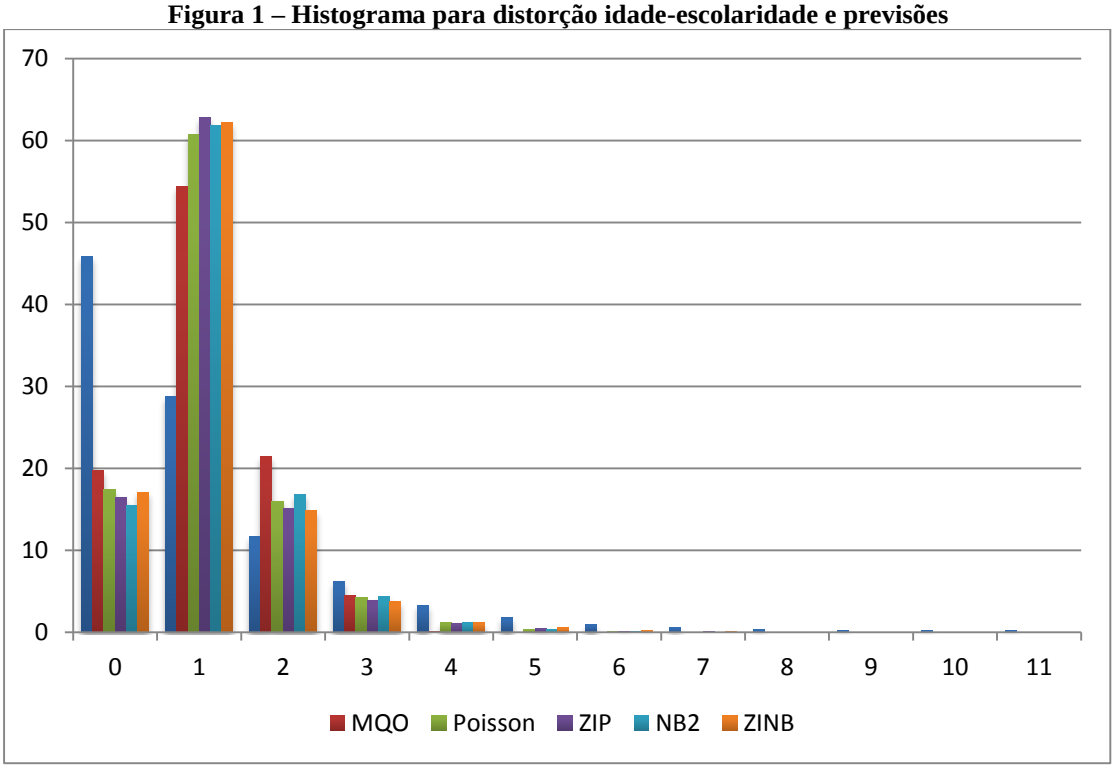
Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Esse resultado é de fato interessante, e está ligado com a literatura sobre o tema. Em especial, Leon e Menezes-Filho (2002) encontram que a taxa de evasão escolar é muito maior entre os alunos que reprovam do que entre aqueles que são aprovados, principalmente em séries iniciais. Isso levaria a uma "seleção" dos alunos, ficando nas séries mais altas aqueles com características pessoais que reduzem suas chances de reprovar, inclusive em séries futuras. Existem outros trabalhos com resultados semelhantes como Ribeiro (1991), Barros e Mendonça (1998) e Souza *et al* (2012), relacionando a reprovação – a principal razão para a distorção idade-escolaridade nos alunos que continuam na escola – com maior probabilidade de evasão ou probabilidade de reprovações posteriores. Ou seja, quanto mais velho um aluno, mais chances ele tem de terminar os estudos sem reprovação, caso isso nunca tenha ocorrido. Já quanto mais cedo a reprovação, maiores a chance de futuras reprovações e eventual evasão escolar. Estes resultados são confirmados por este trabalho.

Alguns testes foram conduzidos para a comparação entre os modelos. Seus resultados podem ser visto no APÊNDICE B. De fato, foi possível captar sobredispersão, indicando o uso dos modelos NB2, que seriam adequados à modelagem desse fenômeno. Teste de razão de verossimilhança foram realizados para comparar os modelos Poisson com os modelos NB2 (incluindo em suas formas com zeros inflados), mostrando que a constante α - responsável por modelar a variância - é significativa. Os resultados obtidos no teste Vuong é que os modelos com zeros inflados são mais adequados. Por este motivo, juntamente com a característica da sobredispersão, o modelo que parece mais adequado é justamente o ZINB. Finalmente, utilizando os

critérios de informação de AIC e BIC, há a confirmação de que este é o modelo mais adequado.

Figura 1 – Histograma para distorção idade-escolaridade e previsões



Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

Cabe ainda uma análise das previsões obtidas com estes modelos, assim como sua comparação entre eles e com o dados utilizados. Eles estão resumidos na figura 1. Como é possível visualizar, o adequação dos dados previstos com a amostra não ficou ideal, principalmente ao se levar em consideração as defasagens zero e um. O problema de excesso de zeros era esperado e ainda deverá ser melhor trabalhado em aplicações futuras dos modelos de contagem para a distorção escolar. Sua correção muito possivelmente também levará a uma previsão melhor no caso de apenas uma defasagem. A partir da segunda defasagem os modelos de contagem passam a se sair melhor, principalmente se comparados contra o MQO, uma vez que eles preveem maior número de distorções, embora não consigam reproduzir distorções maiores que 8 anos.

Esses resultados reforçam que o correto tratamento dessa variável dependente deva ser como dados de contagem. Futuras aplicações poderiam ainda se beneficiar do uso de dados mais completos, que levassem em consideração características pessoais das crianças e jovens, como motivação, QI ou habilidades socioemocionais. Ainda, fatores que seriam mais facilmente observados, mas que não foram incluídos nos modelos propostos, como a infraestrutura escolar ou *background* familiar, podem ter efeito sobre o número de defasagens. De fato, essas variáveis já foram consideradas em outros estudos sobre educação e podem se mostrar importantes no contexto da análise econométrica da distorção idade-escolaridade a partir de modelos de dados de contagem.

CONCLUSÃO

O objetivo deste artigo foi fazer uma análise econométrica do número de distorções idade-escolaridade a partir de modelos de contagem. Esse tema já havia sido discutido em outros artigos aqui citados, embora nenhum tenha utilizado a variável dependente como número de anos de distorção idade-escolaridade e portanto empregado modelos de dados de contagem para tal análise. Desta forma, este trabalho busca contribuir de modo original para a literatura sobre o tema, embora ainda de maneira exploratória.

Quatro modelos de contagem foram estimados, além de um modelo de Mínimos Quadrados Ordinário: o modelo Poisson, o modelo Poisson com Zeros Inflados, o Modelo Binomial Negativo e o modelo Binomial Negativo com Zeros Inflados, cujos resultados estão resumidos nas tabelas 2 e 4. Destaca-se que em todos os modelos a grande maioria das variáveis utilizadas apresentaram o coeficiente significativo e com o sinal esperado. Sendo assim, um número menor de distorções idade-escolaridade está associada à pessoas do sexo feminino, brancas ou amarelas, que frequentam ensino privado, cuja mãe mora no mesmo domicílio e que tem maior educação, cuja renda familiar *per capita* é maior e que vivem na zona urbana. Um resultado distinto do esperado foi o coeficiente da variável trabalha, que se mostrou negativo, de modo que aqueles que trabalham tiveram menor número de defasagens. Esse resultado poderia ser explicado pela ocorrência de distorção estar mais associado ao fato das crianças e jovens não estudarem do que com o fato delas trabalharem. Um estudo que se aprofunde mais nesse sentido é demandado para melhor compreender como trabalhar ou não trabalhar (inclusive levando em consideração a carga horária) realmente afeta a ocorrência e magnitude da distorção escolar.

Quando tomamos em consideração os modelos com zeros inflados, algumas variáveis não se mostram significativas no processo *logit*, como trabalha, renda familiar *per capita* e zona urbana. A variável categórica de ensino apresentou coeficiente negativo para ensino em escola pública, indicando que no processo *logit* os que estão matriculados em escolas pública apresentam maiores chances de ocorrência de alguma distorção em relação aqueles não matriculados em qualquer sistema de ensino. Esse paradoxo pode ser explicado por haver muitos concluintes do ensino médio no tempo correto entre os que não estão matriculados em qualquer escola. Em uma análise futura é importante realizar a comparação entre aqueles matriculados com não matriculados, pois, embora distorção idade-escolaridade e evasão sejam fenômenos intimamente relacionados, não são equivalentes, estando a não-matricula associada mais à evasão que distorção. Além disso, associar a não-matricula a outras variáveis, como trabalhar ou estar matriculado em um curso de ensino superior ou técnico também poderia contribuir para melhor compreender o resultado aqui obtido.

Por fim, deve-se destacar as vantagens metodológicas obtidas a partir do uso de modelos de contagem com zeros inflados. Ao dividir o processo gerador de contagens em dois, um binário (neste caso, *logit*) e outro de contagem (Poisson ou Negativo Binomial) é possível dividir a amostra em dois grandes grupos de crianças e jovens: aqueles que nunca apresentarão qualquer distorção, terminando os estudos no tempo adequado, e aqueles que podem vir a apresentar alguma distorção, embora não necessariamente. Essa inovação em termos de modelagem tem um importante desdobramento, que encontra respaldo na literatura sobre o tema: que a reprovação está positivamente associada a maiores chances de ocorrência de outra reprovação ou até mesmo evasão escolar. No modelo aqui apresentado, este resultado é obtido a partir do coeficiente positivo encontrado para a variável idade na porção *logit* de ambos os

modelos com zeros inflados. Assim, uma maior idade estaria associada a maiores chances de crianças e jovens não estarem sujeitos à ocorrência de qualquer distorção idade-escolaridade. Este resultado parace ser contra intuitivo, dado que o coeficiente desta covariada na parcela de contagem dos modelos mostrou que idade está positivamente relacionada com o número de distorções. A interpretação, porém, é que a medida que os anos passam e nenhuma defasagem ocorre, maiores se tornam as chances de que nenhuma distorção venha a ocorrer no futuro, reforçando a tendência de não ocorrência de distorção ao longo do tempo. Já, caso haja a possibilidade de ocorrência de distorção, uma maior idade relaciona-se com maior número delas. Deste modo, a ocorrência de uma defasagem estaria associada a uma maior probabilidade de ocorrência de outra defasagem no futuro, seja por meio de uma reprovação ou evasão. Este resultado é bastante coerente com os trabalhos já realizados para o Brasil.

Embora o emprego desta nova metodologia tenha trazido contribuições ao estudo da distorção idade-escolaridade, melhorias ainda podem ser feitas dentro deste mesmo paradigma. Um deles seria a estimação de outros modelos com zeros inflados, como o modelo Hurdle. Também seria interessante incluir outras variáveis relacionadas ao *background* familiar, como número de filhos da família, e até mesmo variáveis de infraestrutura escolar, como acesso à laboratório de informática e proporção aluno-professor, além é claro de buscar novas combinações entre os regressores já presentes. Finalmente, uma base de dados mais abrangente, incluindo observações de diversos anos para os mesmos indivíduos ou com acesso a informações relativas à atributos individuais das pessoas, como motivação, QI ou habilidades socioemocionais, pode ser uma grande contribuição, na medida em que haveria um melhor tratamento das heterogeneidades não observadas.

REFERÊNCIAS

ALVES, F.; ORTIGÃO, I.; FRANCO, C.. Origem Social e Risco de Repetência: Interação Raça-Capital Econômico. **Cadernos de Pesquisa,** v. 37, n.130, p. 161-180,2007.

BARROS, R. P. de; MENDONÇA, R.. **Consequências da repetência sobre o desempenho educacional**. Brasília: Ministério da Educação. Projeto de Educação Básica para o Nordeste, 1998.

BECKER, G.S. Investment in Human Capital: A Theoretical Analysis. **The Journal of Political Economy**, Vol. 70, No. 5, Part 2: Investment in Human Beings, 1962

BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996.** Estabelece as Diretrizes e Bases da Educação Nacional. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 10 jun. 2013.

BRASIL. **Lei nº 12.796, de 4 de abril de 2013.** Altera a Lei nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional, para dispor sobre a formação dos profissionais da educação e dar outras providências.

CAMERON, A. C.; TRIVEDI, P. K**. Microeconometrics: methods and applications**. Cambridge: Cambridge University Press, 2005.

_____.**Regression Analysis of Count Data**. Second Edition.Cambridge: Cambridge University Press. 2013.

FERNANDES, R.. NATENZON, P. E.. A evolução recente do rendimento das escolas brasileiras: uma reavaliação dos dados do Saeb. **Estudos em Avaliação Educacional**, n. 28, p.3-22, 2003.

FERRÃO, M. E.. et al. O SAEB - Sistema Nacional de Avaliação da Educação Básica: objetivos, características e contribuições na escola eficaz. **Revista Brasileira de Estudos de População**, v. 18, n. 1/2, p.111-130, 2001.

FRANCO, A. M. de P. . **Os determinantes na qualidade da educação no Brasil**. 2008. 146f. Tese (Doutorado em Economia) – Departamento da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (USP), São Paulo, 2008.

FRITSCH, R.; VITELLI, R.; ROCHA, C. S. Defasagem Idade-Série em Escolas Estaduais de Ensino Médio do Rio Grande do Sul. **Revista Brasileira de Estudos Pedagógicos**. v. 95,n.239, 2014.

GREENE, W. H. **Econometric Analysis**. Seventh Edition. Boston: Pearson. 2012.

HANUSHEK, E. A. The failure of input-based schooling policies. **The Economic Journal**, vol. 113, pg. 64-98, 2003.

_____. Education Production Function. **Palgrave Dictionary**. 2007.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Microdados da Pesquisa Nacional por Amostra de Domicílios (PNAD).** 2013.

LEON, F. L. L. de; MENEZES-FILHO, N. A. Reprovação, avanço e evasão escolar no Brasil. **Pesquisa e Planejamento Econômico**, Rio de Janeiro, v. 32, n. 3, p.417-451, dez. 2002.

MACHADO, Danielle Carusi. Escolaridade das crianças no Brasil: três ensaios sobre a defasagem idade série. 142f. Tese de Doutorado. Programa de Pós-Graduação em Economia, Departamento de Economia. Pontifícia Universidade Católica do Rio de Janeiro, PUC-RIO. Rio de Janeiro, RJ, 2005.

MACHADO, D. C.; GONZAGA, G.. O impacto dos fatores familiares sobre a defasagem idade-série de crianças no Brasil. **Rev. Bras. Econ.**, Rio de Janeiro , v. 61, n. 4,. 2007 .

MACHADO, ,D. C.; FIRPO, S.; GONZAGA, G.. A Relação entre Proficiência e Dispersão de Idade na Sala de Aula: A influência do nível de qualificação do Professor. **Pesquisa e Planejamento Econômico.** v.43,n.3,2013.

PONTILI, R. M.; KASSOUF, A. L. Is Age-Grade Distortion in Brazil´s primary education system more closely associated to school infrastructure or to family characteristics? **Well-Being and Social Policy.** v.4, n.1,p.29-54.

RIANI, J. de L. R.. **Determinantes do resultado educacional no Brasil**: Família, Perfil Escolar dos Municípios e Dividendo Demográfico numa abordagem Hierárquica e Espacial. 2005. 218f. Tese (Doutorado em Demografia) - Centro de Desenvolvimento e Planejamento Regional da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, 2005.

RIBEIRO, S. C.. A pedagogia da Repetência. **Estudos Avançados**, São Paulo, n. 5, v. 12, p.7-21, mai./jun. 1991.

SOARES, S.; SÁTYRO, N.. O impacto da infraestrutura escolar na taxa de distorção idade-série das escolas brasileiras de ensino fundamental - 1998 a 2005. **Texto para Discussão do IPEA. n.1338**, 25p. 2008.

SOUZA, A. P. et al. Fatores Associados ao Fluxo Escolar no Ingresso e ao Longo do Ensino Médio no Brasil. **Texto para Discussão da Escola de Economia de São Paulo da FGV 1/2012**, mar. 2012.

SEN, A. **Desenvolvimento como liberdade**. São Paulo: Editora Companhia das Letras, 2010 [2000]

VUONG. Q. H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. **Econometrica.** n 57, v.2,307-333.1989

WILKELMANN, R. **Econometric Analysis of Count Data**. Fifth Edition. Berlin: Springer. 2008.

APÊNDICE A- INTERVALOS DE CONFIANÇA DOS COEFICIENTES ESTIMADOS PARA MODELOS POISSON, NB2, ZIP E ZINB

Intervalos de Confiança dos Coeficientes Estimados para Modelos Poisson, NB2, ZIP e ZINB

Variáveis	Poisson		NB2[1]	
idade	0,113625	0,121001	0,1154838	0,1227512
mulher	-0,3583657	-0,3134727	-0,3383763	-0,2992257
branco	-0,1948377	-0,1426686	-0,174115	-0,1296607
Ensino_público	-0,4834349	-0,4099276	-0,4641761	-0,4044623
Ensino_privado	-1,038092	-0,9013647	-1,027851	-0,9253851
trabalha	-0,1963089	-0,1267335	-0,1863548	-0,1272352
mae_mora	-0,1066776	-0,0420806	-0,1014996	-0,0442495
educ_mae	-0,0650773	-0,0590716	-0,0631238	-0,0580365
log_rend_pc	-0,007702	-0,0024831	-0,008224	-0,0039114
Urbano	-0,0801365	-0,0290027	-0,0880139	-0,0386148
Nordeste	-0,1435169	-0,0909984	-0,1057705	-0,0534787
Sudeste	-0,4137698	-0,3509673	-0,3072284	-0,2484995
Sul	-0,3541416	-0,2728774	-0,263448	-0,191425
Centro-Oeste	-0,327672	-0,2404494	-0,2662863	-0,1834695
Constante	-0,1329541	0,0180602	-0,21176	-0,0659442

(cont.)

Variáveis	ZIP[2]		ZINB[1 2]	
idade	0,1442569	0,1515277	0,1378254	0,1456397
mulher	-0,216454	-0,1692694	-0,2462161	-0,2002499
branco	-0,1375706	-0,0838335	-0,1579595	-0,1061914
Ensino_público	-0,6563237	-0,5906164	-0,7311531	-0,6607459
Ensino_privado	-0,8773356	-0,7001711	-0,9975664	-0,8302551
trabalha	-0,1747514	-0,1076952	-0,1761176	-0,1076274
mae_mora	-0,0471742	0,0165377	-0,0675801	-0,0041684
educ_mae	-0,0453234	-0,0387923	-0,0507623	-0,0444678
log_rend_pc	-0,0060023	-0,0010916	-0,005825	-0,0009991
urbano	-0,0851575	-0,0331257	-0,0853129	-0,035005
Nordeste	-0,123872	-0,0706062	-0,1350619	-0,0833708
Sudeste	-0,3550254	-0,2898702	-0,3721294	-0,3105383
Sul	-0,2806572	-0,1936953	-0,3175859	-0,2341326
Centro-Oeste	-0,3080183	-0,2156943	-0,3013676	-0,2143591
Constante	-0,5292327	-0,3810897	-0,3795205	-0,2173118

1. Excluidos o coeficiente de lnalpha 2. Excluidos os coeficientes do *logit*

Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)

APÊNDICE B- TESTES PARA COMPARAÇÕES DE MODELOS

Testes de Comparações de Modelos

	Vuong	Likelihood Ratio	AIC	BIC
Poisson	-	-	165307,8	165442,3
ZIP	33,87 (Pr>z = 0,0000)	-	158216,5	158485,5
NB2	-	Chi^2 = 5823,51 Prob>=chibar2 = 0,000	159486,3	159629,8
ZINB	26,95 (Pr>z = 0,0000)	Chi^2 = 1388,59 Prob>=chibar2 = 0,000	156829,9	157107,9

Fonte: Elaborado pelos autores a partir de dados da PNAD(2013)