DOI: http://dx.doi.org/10.1590/0103-6351/3138

A relação de fatores individuais, familiares e escolares com a distorção idade-série no ensino público brasileiro

The relationship between individual, family and scholar characteristics and the age-grade distorcion in Brazilian public education

Alysson Lorenzon Portella
Insper
Tanise Brandão Bussmann
Fundação Universidade Federal do Pampa
Ana Maria Hermeto de Oliveira
Universidade Federal de Minas Gerais

Abstract

This work discusses factors related to the age-grade gap, indicating the main determinants of it for Brazilian public-school students. Using data from PNAD and INEP from the year of 2013, it is evaluated how personal, family and school characteristics are related to the occurrence of age grade-distortion on public school students. More specifically, this work applies econometric models that deal exclusively with count data, such as Poisson and Negative Binomial, as well as models adjusted for excess zeros, such as Zero-Inflated Poisson and Zero-Inflated Negative Binomial. Among the results, it is highlighted a substitutability between the variables related to the educational and the individual-family context, with a reduction of the individual-family coefficients once educational variables are considered in the estimation. Furthermore, it is found that as students grow older, the absence of any previous age-grade gap is reflected in a smaller probability of it ever happening in the future.

Keywords

age-grade distortion; count models, zero-inflated count models.

JEL Codes A21; C51; I2.

Resumo

Este trabalho discute os fatores relacionados à distorção idade-série, indicando quais são os principais determinantes desta defasagem para estudantes do ensino público do Brasil. A partir de uma análise econométrica, com dados da PNAD e do INEP de 2013, é avaliada a maneira como características pessoais, familiares e escolares estão relacionadas com a ocorrência de distorções. Em especial, este trabalho emprega modelos econométricos que tratam especificamente de dados de contagem, como o Poisson e o Binomial Negativo, bem como outros que trabalham com excesso de zeros, como o Poisson com Zeros Inflados e o Binomial Negativo com Zeros Inflados. Entre os resultados, destaca-se uma substutibilidade entre as variáveis do contexto educacional e as do contexto familiar individual, pela redução dos coeficientes destas com a adição das variáveis educacionais. Além disso, destaca-se que na medida em que a idade avança, a ausência de distorção reflete-se em menor probabilidade de ela ocorrer no futuro.

Palavras-chave

distorção idade-série; modelos de contagem; modelos de contagem com zeros inflados.

Códigos JEL *A21; C51; I21.*

1 Introdução

Desde o *Coleman Report* (Coleman *et al*., 1966), publicação que visava analisar os fatores associados com o desempenho escolar, diversos trabalhos vêm sendo realizados com o objetivo de verificar qual a influência de certas variáveis no resultado escolar dos estudantes. Uma questão importante é observar quais são os fatores relacionados com a defasagem idade-escolaridade. A distorção idade-série é um problema grave, principalmente nos países em desenvolvimento, onde é difícil implementar de forma efetiva a obrigatoriedade para os estudantes do ensino básico.

No caso brasileiro, Fernandes e Natezon (2003) observam uma melhora neste indicador ao longo do tempo. Apesar disso, seu percentual ainda é bastante elevado, ficando em 44% para o ano de 1999. Diversos estudos já foram realizados para verificar a relação entre a defasagem idade-série e o desempenho dos alunos (Fritsch; Vitelli; Rocha, 2014; Machado, 2005, Ferrão *et al*., 2001) ou buscando entender quais são os determinantes da defasagem idade-série (Riani, 2005; Leon, Menezes-Filho, 2002; Machado; Gonzaga, 2007). O presente trabalho se encontra no grupo que procura entender seus determinantes.

O estudo dos principais fatores relacionados com a distorção idade-série pode conduzir a um melhor entendimento do problema e também indicar fatores que, ao se tornarem alvos das políticas governamentais, consigam diminuí-lo. Assim, o presente trabalho pretende contribuir para a literatura sobre o tema a partir do estudo tanto da influência das características individuais e familiares, quanto de algumas características do ambiente escolar sobre o grau de distorções idade-série dos indivíduos. É possível utilizar uma gama de formas funcionais para verificar quais são os fatores relacionados com a defasagem idade-série. Geralmente são aplicados modelos probabilísticos, visando estimar o aumento da probabilidade da ocorrência da defasagem idade-série de acordo com as mudanças de certas características individuais. Aqui, há uma inovação metodológica neste sentido: são utilizados os modelos de contagem, e especificamente modelos com zeros inflados, que até o momento não haviam sido explorados para essa variável dependente, conforme será visto na revisão da literatura.

É possível utilizar tais modelos pelo fato da variável dependente – a defasagem idade-série – apresentar apenas valores inteiros e não-negativos. Sendo assim, é realizado um tratamento distinto do fenômeno, de acordo

com seu número de ocorrências e não com a probabilidade de realização. Embora esses modelos já tenham sido aplicados para o estudo de diversos temas, especialmente na área da saúde[1], até onde é do conhecimento dos autores, eles ainda não foram aplicados para o estudo da distorção idade-série, uma variável que claramente se enquadra no grupo. Assim, este artigo busca fazer uma contribuição metodologicamente original ao tema, embora ainda de maneira exploratória. Como será visto adiante, o uso desses modelos traz à tona resultados interessantes e que corroboram parte dos resultados obtidos pela literatura até o momento.

Com o objetivo de verificar se algumas características do ambiente educacional impactam o número de distorções idade-série que poderiam ser alvos de políticas públicas, a análise se restringe à provisão de educação pelo setor público. Ainda, como a Educação Básica nos anos iniciais é responsabilidade preponderante dos municípios, enquanto a Educação Básica nos anos iniciais residuais e nos anos finais é de responsabilidade dos estados, apenas estas esferas de administração foram levadas em consideração, sem que os alunos de instituições federais fossem considerados na análise. Por fim, foram considerados apenas os estudantes entre 8 e 17 anos, ou seja, aqueles ainda em idade escolar e que poderiam ou não apresentar distorção. A base de dados utilizada é a Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2013, complementada com informações a respeito das características escolares para o mesmo ano, todas obtidas a partir do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Este é outro aspecto pouco explorado na literatura: a utilização de diversas bases de dados. Apenas Rios Neto, Cesar e Riani (2002) e Pontili e Kassouf (2008) utilizaram bases de dados de pesquisa de domicílios (no caso, a PNAD e o Censo Demográfico, respectivamente) com informações do ambiente escolar (no caso, o Censo Escolar).

Além desta seção introdutória, o artigo é dividido em outras cinco seções. A segunda seção trata da revisão de literatura, em que são descritos os principais resultados encontrados. A terceira seção trata dos aspectos metodológicos, na qual os procedimentos realizados na base de dados são explicitados, além dos métodos econométricos empregados. Na quarta seção são apresentadas algumas estatísticas descritivas, seguida da apresentação e análise das estimações, na seção cinco. Por fim, a última seção conclui este trabalho.

1 Para referências de trabalhos na área, consultar Staub e Winkelmann (2013), Cameron e Trivedi (2013) e Winkelmann (2008).

2 **Revisão de literatura**

O estudo dos principais fatores relacionados à qualidade da educação dos indivíduos tem sido um dos temas mais estudados na economia da educação, desde a década de 1960, quando ocorre a divulgação do trabalho pioneiro na área, o *Coleman Report*. Esse relatório tinha como objetivo verificar as distinções no nível de aprendizagem de diferentes grupos da sociedade norte-americana. O relatório mostra que diversos fatores são importantes no desempenho dos estudantes, porém, os mais relevantes deles estariam relacionados com o *background* familiar, enquanto fatores como infraestrutura escolar possuíam pouca influência sobre o desempenho (Coleman *et al*., 1966; Murnane; Wilett, 2011).

Desde o *Coleman Report*, diversos trabalhos realizaram a estimação da função de produção educacional, que, de forma similar à teoria microeconômica, tem como seu produto o resultado educacional (que pode ser descrito por anos de estudo ou proficiência em um exame) e cujos insumos são as características dos indivíduos e do ambiente onde este está inserido (Hanushek, 2007). Uma questão menos explorada, mas igualmente relevante e que pode ser estudada a partir de uma função de produção educacional, é a distorção idade-série. A distorção idade-série (ou defasagem idade-escolaridade) é a diferença entre a idade adequada para a série do estudante e a idade real do estudante. O recomendado é que esta diferença seja zero, isto é, que o estudante esteja na série adequada para sua idade.

Um elevado grau de distorção idade-série 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 o crescimento econômico de longo prazo e retardando a queda na desigualdade social. Sendo assim, do ponto de vista social, a distorção idade-série não somente reduz a velocidade com que se acumula capital humano, como também afeta o nível máximo que este pode alcançar.

Os motivos para a existência de defasagem idade-série 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. Alguns esforços para entender o fenômeno da distorção idade-série foram realizados no caso brasileiro, em parte pela grande quantidade de estudantes com atraso escolar, isto é, com defasagem idade-série positiva. Ao observar a evolução da distorção

idade-série ao longo do tempo no Brasil, Riani (2005) mostra que, em 1980, 78% dos estudantes apresentava idade superior a adequada, enquanto que 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, enquanto em 1999 houve uma redução para 44%.

É possível dividir os estudos sobre a distorção idade-série no Brasil em dois grandes grupos. O primeiro deles está focado em estudar os efeitos da distorção sobre outras variáveis escolares, como o desempenho dos estudantes, de acordo com as provas de proficiência, a repetência e o abandono escolar. Outro grupo de estudos busca conhecer quais são os determinantes da distorção idade-série. Nesse caso são empregadas funções de produção escolar, levando em consideração variáveis relativas aos alunos, sua família, seu *background* socioeconômico e as condições de infraestrutura da escola.

Dentro do primeiro grupo de estudos, Ribeiro (1991) observa que a repetência é o principal problema no que tange o aumento da escolaridade no caso brasileiro, pois há um aumento das chances de o aluno evadir posteriormente. Utilizando dados da PNAD de 1982 e o modelo matemático PROFLUXO, ele encontra taxas de repetência muito superiores às apresentadas pelo MEC, de modo que as estatísticas oficiais estariam distorcendo a realidade do sistema educacional brasileiro da década de 80, ao superestimar o problema da evasão e subestimar o da repetência. Barros e Mendonça (1998), com dados da PNAD de 1985 e modelos probabilísticos, mostram que as reprovações, eventos relacionados positivamente com a defasagem idade-escolaridade, apresentam um efeito negativo tanto para a autoestima dos estudantes, aumentando a probabilidade de reprovações subsequentes, quanto para o estado, pois há maiores gastos com as reprovações. A relação entre distorção idade-série e abandono também foi encontrada por Machado (2005), utilizando dados do Sistema Nacional de Avaliação da Educação Básica (SAEB) de 2003 com estimativas de regressão com efeitos fixos da escola.

Para Ferrão *et al*. (2001) e Franco (2008), há uma relação clara entre a distorção idade-série 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". Ferrão *et al*. (2001) utiliza dados do SAEB de 1999 e um modelo multinível. 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. Um aumento na probabilidade de reprovações posteriores também foi encontrado em um período mais recente por Souza *et al*. (2012), com dados da Pesquisa Mensal do Emprego (PME) de 2002 a 2009 e uso de um modelo de probabilidade linear.

Não é apenas o desempenho individual dos estudantes que é prejudicado com uma maior dispersão da defasagem idade-escolaridade. Machado, Firpo e Gonzaga (2013), com dados do SAEB de 2011 e o método de regressão com efeitos fixos, observam que quanto maior a dispersão em termos de idade em sala de aula, menor a proficiência média das crianças. Finalmente, Fritsch, Vitelli e Rocha (2014) observam que a defasagem entre a série recomendada e a série frequentada tem grande impacto sobre a taxa de abandono dos estudantes, indicando uma relação positiva entre essas duas variáveis. A base de dados é composta por escolas públicas do Município de São Leopoldo em 2001/2002 e são realizadas análises descritivas.

No segundo grupo de trabalhos, Riani (2005) utiliza dados do Censo Demográfico de 1980 a 2000 para analisar a defasagem idade-série. A autora conclui que indivíduos do sexo masculino e negros são os que têm a maior defasagem. A autora também desenvolve um modelo hierárquico-espacial com dados do Censo Demográfico e Censo Escolar de 2000. A maior incidência de distorção idade-escolaridade para indivíduos do sexo masculino também foi encontrada para Leon e Menezes-Filho (2002), que utilizam um modelo *heckprobit* a partir de dados da PME de 1984 a 1997. Eles também observam que morar sem pai e mãe aumenta as chances de reprovação para o 3º ano do ensino médio, enquanto que para o 4º ano do ensino fundamental morar com ambos os pais reduz as chances de reprovação em relação a viver com apenas um deles. 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-à-vis estudantes que moram com apenas pai e mãe ou ambos. Esses resultados não são fixos ao longo do tempo, de acordo com 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. Alves, Ortigão e Franco (2007), com dados do SAEB de 2001 e uso de modelos de risco, observam

os efeitos de cada um dos diferentes capitais (econômico, social e cultural) para a repetência escolar. Os estudantes do sexo masculino, negros e que trabalham têm um risco maior de repetência, e o nível socioeconômico 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, com dados da PNAD de 1996. Os autores utilizam modelos *probit* e de regressão linear combinados a variáveis instrumentais, buscando evitar possíveis vieses. Os autores encontram uma redução na probabilidade de distorção idade-série 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* tradicional para encontrar a probabilidade de os estudantes frequentarem séries adequadas para a sua idade e também quantos anos acima do indicado pela idade-escolaridade eles se encontram, utilizando um modelo ordenado. As características individuais e familiares (renda *per capita*, educação, sexo e idade do chefe de família, e idade e cor da pele 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 em relação ao chefe de família essa probabilidade aumenta quando ele tem menor idade e educação e é do sexo masculino. Os autores incluíram também variáveis do ambiente escolar. Em termos de variáveis da escola, uma variável importante foi o número de laboratórios de informática.

Rios-Neto, César e Riani (2002) relacionam características individuais, de acordo com a PNAD e características das escolas locais das décadas de 1980 e 1990, fazendo o uso um modelo *logit* hierárquico. Com a inclusão de características da escola, há uma redução da importância das características da família. Especificamente, a educação da mãe, que afeta positivamente a probabilidade de progressão educacional, tem seu coeficiente reduzido com a inclusão de variáveis sobre a escolaridade média local dos professores do ensino fundamental, para a 1ª série do Ensino Fundamental. Este resultado indica um efeito substituição entre estes dois fatores. Soares e Sátyro (2008), com dados do Censo Escolar de 1998 a 2005 e comparando métodos de regressão linear com não-paramétricos, observam a relação entre a defasagem idade-escolaridade e variáveis da escola,

especificamente de infraestrutura e relativas à qualidade dos professores. Os autores observam uma relação negativa entre essas variáveis, tanto na abordagem paramétrica quanto não-paramétrica.

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 têm 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).

Por conta dos efeitos adversos da defasagem idade-série, diversos autores buscaram analisar se a adoção de um sistema de avaliação distinto (como o de ciclos ou progressão automática) traria diferenciais positivos no desempenho dos estudantes. Menezes-Filho, Vasconcellos e Werlang (2005), com dados do Censo Escolar (2002) e SAEB (2003), verificam que a diferença de desempenho não se modifica de acordo com o regime educacional, utilizando a regressão simples e após emparelhamento com *propensity score*. Resultado semelhante foi encontrado por Menezes-Filho *et al*. (2008), com dados da Prova Brasil de 2005, utilizando estimativas por mínimos quadrados tradicional e com pareamento. Nos dois trabalhos, pode ser verificado que o desempenho não é estatisticamente distinto, apesar da melhora nos índices de reprovação e evasão. Menezes-Filho, Vasconcellos e Werlang (2005), com dados do SAEB e Censo Escolar de 1999, concluem também que a promoção automática pode ser uma boa alternativa para reduzir a defasagem idade-série dos estudantes.

Esta seção revisou brevemente a literatura sobre o tema. Verifica-se que a defasagem idade-série se relaciona de maneira positiva com o abandono, aumentando as chances de os alunos abandonarem a escola, além de indicarem um desempenho pior e mais chance de reprovações. Além desses resultados individuais, uma maior variância na idade dos estudantes relacionou-se com uma proficiência média menor. Ao analisar as características dos indivíduos, vemos que, em sua maioria, as características relacionadas com maiores probabilidades de defasagem são relacionadas com piores níveis socioeconômicos, de educação familiar/capital cultural e também com a cor negra e com o sexo masculino. Finalmente, conforme

exposto pelos trabalhos de Rios-Neto, César e Riani (2002) e Soares e Sátyro (2008), características relativas à escola também podem influenciar o total de distorções idade-escolaridade do indivíduo. Pode-se concluir que há consequências perversas na ocorrência de defasagem idade-escolaridade e também que os diversos grupos da população são expostos de maneira distinta a este evento.

3 **Metodologia**

Nesta seção, são expostas inicialmente as definições adotadas para a obtenção das variáveis de interesse e, num segundo momento, os métodos econométricos utilizados. A base de dados utilizada foi a PNAD 2013 que consiste de uma amostra representativa da população brasileira, cobrindo regiões rurais e urbanas de todas as Unidades da Federação, com data de referência o dia 28 de setembro de 2013. A população de interesse é composta pelos estudantes em idade escolar matriculados no ensino básico regular público no Brasil.[2] 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 um ano completo de escolaridade, e assim sucessivamente. A amostra utilizada é composta de todas as crianças e adolescentes de 8 a 17 anos de idade, matriculadas em escolas da rede estadual ou municipal, sendo descartadas as observações referentes a alunos matriculados em escolas federais[3] de todos os estados brasileiros. Após exclusão dos alunos com dados faltantes para as variáveis de interesse, a amostra disponível contou com 47.047 observações. Por utilizar este recorte, é possível a existência de um viés negativo, subestimando os resultados encontrados.

Com o objetivo de observar quais são os principais fatores relacionados com a distorção idade-série, o primeiro passo é construir uma variável in-

2 Deste modo, descarta-se da análise aqueles alunos que evadiram ou abandonaram o sistema escolar. Entretanto, no caso de crianças e adolescentes que estão fora do sistema escolar, não faz sentido falar em distorção idade-escolaridade, uma vez que seu problema é que eles deixaram de estudar permanentemente, não simplesmente que eles estão atrasados.Como visto na revisão da literatura, a evasão, o abandono e o atraso são problemas intimamente ligados, mas não iguais. Cada um deles tem sua especificidade e deve ser analisado individualmente.

3 Os estudantes de escolas públicas federais representavam menos de 1% da amostra.

dicativa da existência dessa 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". Espera-se, então, que uma criança de 8 anos tenha completado ao menos um ano de escolarização, com 9 anos, dois anos de estudos e assim sucessivamente, até que um jovem de 17 anos deve ter completado 10 anos de estudos. Diferenças nesses números resultam em distorção idade-série, que pode ir de zero, para quem não apresenta distorção, até 10 anos, quando o aluno não completou nenhum ano de estudo aos 17 anos. A variável dependente, portanto, é o número de anos de distorção idade-série da criança ou adolescente, calculado descontando-se de sua idade o número de anos de estudos completos e mais 7 anos. Números negativos – representando alunos com escolaridade mais avançada para sua idade – foram convertidos em distorção zero. É importante sublinhar que aqui a variável dependente está mais relacionada às repetências do que ao abandono, uma vez que são consideradas apenas as crianças e jovens que estão matriculadas em estabelecimentos de ensino público.[4]

Além das informações relativas aos indivíduos, o ambiente escolar também é levado em conta. Dados a respeito do percentual de docentes com ensino superior, da média de horas-aula e também da média de alunos por turma são utilizados. Tais informações foram retiradas dos Indicadores Educacionais do INEP, também para o ano de 2013. O recorte utilizado será a Unidade da Federação, a região ser Urbana ou Rural e a esfera do setor público (Estadual ou Municipal) relacionado.

Dois conjuntos de variáveis independentes são utilizados nas estimações. O primeiro conjunto engloba variáveis relativas à: idade; mulher (sexo feminino = 1 e masculino = 0); cor de pele; branco (cor de pele - brancos = 1 e não-brancos = 0, sendo amarelos enquadrados como brancos); se a mãe mora ou não no mesmo domicílio (se mora = 1, se não mora = 0, o que inclui mães que já morreram ou que não se tem conhecimento);

4 É possível justificar a escolha pelo recorte etário de 8 a 17 anos pela maneira como a variável dependente é construída, pois é apenas neste intervalo que é possível a existência de defasagem idade-série positiva ou zero, sendo que para estudantes com menos de 8 anos ou mais de 17, a defasagem idade-série será sempre nula ou positiva, respectivamente. Mais especificamente para o caso de alunos com 18 anos, como se considera apenas alunos matriculados, esses podem estar no máximo no terceiro ano do ensino médio, tendo completado então o segundo ano, representando um total de 10 anos de estudos e, necessariamente, resultado em ao menos um ano de distorção idade-escolaridade.

a educação do indivíduo de referência ou seu cônjuge, a que for maior[5]; o logaritmo da renda familiar *per capita*; o número de membros da família; urbano - se reside na zona urbana ou rural (urbano = 1 e rural = 0); estadual, se a escola é da rede municipal ou estadual (municipal = 0 e estadual = 1, com escolas federais excluídas da amostra); se trabalha (representado por duas *dummies*: uma na qual trabalhar até vinte horas semanais equivale a 1 e outra no qual trabalhar mais de vinte horas semanais equivale a 1, enquanto não trabalhar se refere a zero em ambas); se faz algum tipo de trabalho doméstico (representado de forma semelhante ao trabalho, embora o recorte seja em dez horas de trabalho doméstico semanal, com uma *dummy* para até 10 horas e outra para mais de 10 horas); e controles por região (com cinco *dummies* referentes às regiões Nordeste, Sudeste, Sul, Centro-Oeste e o Distrito Federal, tendo como resultado base o da região Norte). O segundo conjunto de variáveis é composto pelo primeiro conjunto acrescido de variáveis relativas às características das escolas. São elas: média de alunos por turma, número de horas-aula diárias e o percentual de professores com ensino superior. Essas variáveis são disponibilizadas pelo INEP, de acordo com Unidade da Federação, a esfera de ensino (estadual ou municipal) e a localização (urbana ou rural). Além disso, como a totalização ocorre em dois períodos para o ensino fundamental (anos iniciais e finais) e também para o ensino médio, cada indivíduo recebeu os valores relativos à sua unidade da federação, com a localização informada, esfera de ensino e período (inicial do ensino fundamental, final do ensino fundamental ou ensino médio).

Neste trabalho, são estimados cinco modelos de contagem, sendo três deles com zeros inflados: Poisson, Binomial Negativo (NB – *Negative Binomial*), Poisson com Zeros Inflados (ZIP – *Zero-Inflated Poisson*), Binomial Negativo com Zeros Inflados (ZINB – *Zero-Inflated Negative Binomial*) e Poisson com Zeros Inflados estimado por máxima quasi-verossimilhança

5 Aqui optou-se pela educação do indíviduo de referência ou seu cônjuge ao invés da educação da mãe, obtida a partir de seu indicador disponível na PNAD. Essa escolha permite incorporar mais de 8 mil observações na amostra, incluindo alunos que moram apenas com o pai ou aqueles que moram com parentes. Dessa amostra, mais de 85% das observações são classificadas como filhos, 13% como parentes e o restante como agregados, pessoa de referência ou cônjuge. A correlação da variável utilizada com a educação da mãe é superior a 0,80, sendo de 0,95 se se considerar apenas as pessoas de referência do sexo feminino. Finalmente, as regressões conduzidas neste trabalho foram estimadas para ambas as definições, não havendo diferenças significativas nos resultados dos coeficientes estimados, sendo a única alteração marcante na variável relativa a de se o aluno mora ou não com a mãe, que se torna maior em magnitude, embora não haja mudança qualitativa na interpretação dessa variável.

(PQL – *Poisson quasi-likelihood*). Além disso, são apresentados resultados estimados a partir de MQO, de modo a comparar o desempenho dos modelos de contagem com lineares.

O modelo Poisson é o mais simples dos modelos de contagem, cuja distribuição das probabilidades dos resultados y depende da média μ, de acordo com a equação (1).

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

A média μ é modelada a partir de um vetor de covariadas x e seus respectivos coeficientes, β, pela equação (2).

$$\mu_i = \exp\left(x_i'\beta\right) \quad , \quad i = 1,\ldots N \tag{2}$$

A partir desses parâmetros, a estimação procede por meio de máxima verossimilhança e a consistência dos coeficientes estimados requer a correta especificação da média condicional, não sendo dependente da especificação correta da distribuição da variável dependente. É importante lembrar que a média de uma distribuição Poisson é igual a sua variância, a chamada propriedade de equidispersão (Winkerlmann, 2008). Essa propriedade é necessária para inferência, sendo muitas vezes infringida, uma vez que muitas distribuições apresentam sobredispersão – variância maior que a média. Nesses casos, um possível tratamento é por meio do acréscimo de um componente multiplicativo u, responsável por representar a heterogeneidade não observada, conforme (3).

$$\mu_i = E\left(y|x,u\right) = \exp\left(x_i'\beta\right)u, \; i = 1,\ldots N \tag{3}$$

Entre as possibilidades, a mais utilizada para a modelagem deste termo é a distribuição Gama, $u \sim \Gamma(\alpha,\beta)$. Assumindo valores iguais para α e β, a combinação de ambas, Poisson e Gama, leva o nome de Binomial Negativa[6] (NB). Assumindo que o valor de α seja constante e igual a σ^{-2}, a variância do modelo é capaz de lidar com a sobredispersão.

6 Diferentes especificações do termo multiplicativo u levam a diferentes modelos NB. Ver Cameron e Trivedi (2013) ou Winkelmann (2008) para um tratamento pormenorizado dos modelos NB.

Já os modelos com zeros inflados são empregados quando o número de zeros da amostra é muito elevado, ou seja, há excesso de zeros, que também pode vir a ser uma das causas da sobredispersão (Winkelmann, 2008). Nessas situações, os modelos de contagem tradicionais não fazem previsões adequadas. Em tais casos, a solução passa por dividir o processo gerador dos dados em duas partes: no primeiro, são modelados os zeros estruturais, a partir de uma distribuição binária qualquer que ocorre com probabilidade π. O outro componente seria um processo de contagem, com função densidade f_2, que ocorre com probabilidade $(1 - \pi)$. Assim, a distribuição de probabilidades toma a forma descrita em (4).

$$\Pr[y = j] = \begin{cases} \pi + (1-\pi) f_2(y), & se \ \ y = 0 \\ (1-\pi) f_2(y), & se \ \ y \geq 1 \end{cases} \tag{4}$$

No caso do modelo de Poisson, esta modificação é denominada Poisson com Zeros Inflados ou ZIP (*Zero-Inflated Poisson*). Para o Binomial Negativo, o modelo é chamado de Binomial Negativo com Zeros Inflados ou ZINB (*Zero-Inflated Negative Binomial*). Nesses modelos há um processo binário inicial, f_1. 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 (4). Ainda, observa-se que também é possível que ocorra a previsão do valor zero neste segundo processo. Dessa forma, os modelos de zeros inflados podem resultar em zeros de duas formas: a partir do processo binário, f_1 – zeros estruturais – ou pela função de densidade padrão f_2 – zeros acidentais. O processo binário é geralmente descrito por uma função *logit*, que dependerá de $z'\gamma$, onde z é um conjunto de covariadas que independe de x, i.e., o conjunto de covariadas utilizados na porção *logit* pode ser igual ou diferente do conjunto empregado na porção de contagem. Esta, por sua vez, pode seguir uma distribuição qualquer, embora neste trabalho sejam tratados os casos Poisson e Binomial Negativo.

A estimação dos modelos com zeros inflados é baseada na máxima verossimilhança. Este processo, porém, gera coeficientes viesados caso o modelo não seja corretamente especificado. Em vista disso, Staub e Winkelmann (2013) propõem um processo de estimação por máxima quasi-verossimilhança (PQL), que requer menos pressupostos e em função disso

é robusto à má especificação do modelo[7]. O custo, porém, é uma perda na precisão. Experimentos de Monte Carlo realizados pelos autores mostraram um bom desempenho do PQL para amostras grandes (50 mil), apresentando viés muito baixo.

Dada a forma exponencial da parametrização da média, nos modelos Poisson, a derivada parcial de $E(y|x)$ dependerá dos valores de $x'\beta$. Assim, os efeitos marginais serão diferentes para cada indivíduo de acordo com seus atributos. Porém, o coeficiente β_j também pode ser interpretado diretamente como a semielasticidade (Winkelmann, 2008). Isto é, enquanto as mudanças marginais variam de acordo com os indivíduos, as mudanças relativas são constantes. Caso a variável independente esteja na forma logarítmica, então a interpretação é dada diretamente através de elasticidades. Interpretação similar aos modelos Poisson pode ser dada ao modelo Binomial Negativo no caso dos efeitos parciais relativos, mas não para efeitos marginais, que acarretam maiores dificuldades.[8] Quanto aos modelos com zeros inflados, estes possuem vários efeitos marginais de interesse. Importante é notar, porém, que o valor esperado de y condicional a uma covariada que está presente tanto na porção de contagem como na porção inflada do modelo será ambíguo, dependendo tanto da razão entre os coeficientes β_j e γ_j como dos próprios valores das variáveis independentes.[9] Neste trabalho a interpretação será focada em cada um dos coeficientes separadamente, em que os coeficientes γ_j têm a interpretação padrão dos modelos *logit*.

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, em que 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 à sobredispersão, há algumas possibilidades de testes, conforme apresentados em Cameron e Trivedi (2013, seção 3.4), embora aqui destaque-se apenas a significância do parâmetro α que modela a variância. Se este for significativamente

7 A implementação dessa estimação está disponível no anexo de Staub e Winkelmann (2013).

8 Consultar Winkelmann (2008, p. 128) para a explicação, cujas dificuldades derivam justamente do termo multiplicativo, u.

9 Ver Winkelmann (2008, p. 192).

diferente de zero, os modelos com sobredispersão são considerados mais adequados que o Poisson.

Duas especificações do modelo são propostas. Uma leva em conta apenas variáveis relativas ao aluno, obtidas a partir dos microdados da PNAD. Nesta especificação (Modelo 1), assume-se que a distorção idade-série está relacionada com as características dos alunos de acordo com (5).

$$d_i = f\left(C_{p,i}, C_{f,i}, C_{g,i}\right) \tag{5}$$

onde d_i representa a distorção do aluno *i*, modelado como uma função de suas características pessoais, $C_{p,i}$; suas características familiares, $C_{f,i}$; e suas características geográficas, $C_{g,i}$. Essas características incluem, respectivamente, as seguintes variáveis: idade, mulher, branco, *dummies* representando se trabalha até 20 horas ou mais de 20 horas semanais, *dummies* representando se dedica até 10 horas ou mais de 10 horas semanais às atividades domésticas e se é de escola estadual ou municipal; se mora com a mãe, a educação máxima da pessoa de referência, o logaritmo da renda *per capita* familiar e o tamanho da família; e se vive em zona urbana, juntamente com *dummies* para a região Nordeste, Sudeste, Sul, Centro-Oeste e Distrito Federal.

Com vistas de avaliar até que ponto variáveis escolares podem afetar a distorção idade-série, é proposta uma segunda especificação (Modelo 2) que mantém as mesmas variáveis anteriores e inclui algumas variáveis relativas ao ambiente escolar. Essa segunda especificação assume a seguinte forma:

$$d_i = f\left(C_{p,i}, C_{f,i}, C_{g,i}, C_{e,i}\right) \tag{6}$$

onde $C_{e,i}$ designa as variáveis do ambiente escolar obtidas a partir dos dados do INEP: média de alunos por turma, número de horas-aula e porcentagem de professores com ensino superior. Tendo em vista os objetivos deste artigo, esse segundo modelo seria o mais interessante e ao qual será dada maior atenção, justamente por conter variáveis relativas à qualidade das escolas.

4 **Análise descritiva**

Quando se observa o total de indivíduos entre 8 e 17 anos matriculados em escolas públicas, 46% deles apresentam alguma distorção idade-série.

A Tabela 1 apresenta a distribuição do número de alunos e a frequência referente a cada valor da distorção idade-série. Como pode ser visto, conforme o número de distorções aumenta, sua frequência diminui continuamente, até alcançar número muito baixo de ocorrências para os maiores valores.

Tabela 1 **Distribuição da distorção idade-série**

Distorção	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Número	21.732	14.617	5.670	2.691	1.368	594	271	132	69	42	14
Porcentagem	45,95	30,90	12,20	5,69	2,89	1,26	0,57	0,28	0,15	0,09	0,03

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

Em relação às estatísticas descritivas das variáveis independentes, pela Tabela 2, observa-se que 49% da amostra são do sexo feminino, com uma média de 12 anos de idade, com 38,01% de cor branca ou amarela. A grande maioria reside com a mãe (86,57%) em área urbana (78,99%) e a educação máxima dos pais é de 7 anos. Em torno de 8% dos indivíduos trabalharam fora de casa na semana de referência sendo que 4,76% trabalharam 20 horas ou mais, e cerca de metade trabalhou em casa. A renda média *per capita* foi de R$ 460,70 e a média do logaritmo da renda *per capita* familiar foi 5,74, valor bastante inferior ao da amostra total, o que já era esperado pelo recorte escolar. Em relação às variáveis do ambiente escolar, observamos que a média de horas-aula diária foi de 4,54, enquanto foram 26 alunos em média por turma. O percentual de docentes com ensino superior foi de 83%.

A Tabela 2 também apresenta as médias e desvios padrão das variáveis independentes relativas ao grupo de alunos sem qualquer distorção e para aqueles com algum número de distorção. Embora não esteja baseada na variável de interesse, a tabela ajuda a entender quais atributos estão mais relacionados a uma maior distorção idade-série. Como pode ser visto, o grupo de alunos sem distorção teve uma maior participação de pessoas do sexo feminino e brancas que ainda vivem com suas mães em zonas urbanas, cujos pais tem uma maior renda e nível educacional. Além disso, a maioria está matriculada em escolas estaduais, trabalha menos e está proporcionalmente mais representada na região Sul e Sudeste. No geral, estudam em turmas menores, têm mais horas de aula e mais professores com ensino superior. Sendo assim, é possível ter uma ideia dos coeficientes estimados a seguir.

Tabela 2 **Estatísticas descritivas da amostra**

Variável	Base completa		Sem distorção		Com distorção	
	Média	Desvio padrão	Média	Desvio padrão	Média	Desvio padrão
Idade	12,4651	2,7829	12,1577	2,8552	11,9509	2,7450
Mulher	0,4875	0,4998	0,5387	0,4985	0,4848	0,4998
Branco	0,3802	0,4854	0,4376	0,4961	0,3690	0,4825
Mora com a mãe	0,8657	0,3410	0,8906	0,3122	0,8669	0,3397
Educação da pessoa de referência	7,7640	4,0532	8,6521	3,8972	7,6504	3,9180
Logaritmo da renda *per capita* familiar	5,7147	1,1198	5,8759	1,1059	5,6712	1,1079
Tamanho da família	4,4916	1,5473	4,3197	1,3954	4,5043	1,5517
Urbano	0,7899	0,4074	0,8245	0,3804	0,7874	0,4092
Estadual	0,4755	0,4994	0,5498	0,4975	0,4377	0,4961
Trabalha até 20 horas semanais	0,0379	0,1910	0,0350	0,1839	0,0315	0,1746
Trabalha mais de 20 horas semanais	0,0415	0,1995	0,0407	0,1976	0,0303	0,1714
Trabalhos domésticos até 10 horas semanais	0,3212	0,4669	0,3320	0,4709	0,3158	0,4648
Trabalhos domésticos mais de 10 semanais	0,1801	0,3843	0,1821	0,3859	0,1617	0,3682
Nordeste	0,3065	0,4611	0,2617	0,4395	0,2984	0,4576
Sudeste	0,3748	0,4841	0,4200	0,4936	0,3862	0,4869
Sul	0,1336	0,3402	0,1544	0,3613	0,1257	0,3316
Centro-Oeste	0,0583	0,2343	0,0583	0,2343	0,0645	0,2456
Distrito Federal	0,0121	0,1094	0,0132	0,1143	0,0116	0,1070
Média de alunos por turma	26,8223	5,2806	27,6006	5,1847	26,2565	5,2779
Horas-aula	4,5444	0,4326	4,5793	0,4310	4,5386	0,4367
Porcentual de docentes com ensino superior	83,0615	17,8010	86,1367	15,2679	82,2294	18,1065

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

Finalmente, a Tabela 3, reproduz a porcentagem de alunos com e sem distorção idade-escolaridade de acordo com a sua idade. Como pode ser visto, a proporção de alunos defasados cresce continuamente até alcançar

um pico para 13 anos de idade, quando sofre uma abrupta queda, possivelmente devido ao grande número de abandono escolar nessa idade.

Tabela 3 **Porcentagem de alunos com e sem distorção idade-série de acordo com a idade**

Idade	8	9	10	11	12	13	14	15	16	17	Total
Sem distorção	58,42	55,56	51,11	46,04	40,83	37,43	42,76	45,87	43,46	39,03	45,93
Com distorção	41,58	44,44	48,89	53,96	59,17	62,57	57,24	54,13	56,54	60,97	54,07

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

5 **Resultados**

A apresentação dos resultados está dividida em duas partes. Na primeira são apresentados os resultados relativos aos modelos sem zeros inflados: MQO, Poisson e NB. Num primeiro momento, são analisados os modelos sem a adição das variáveis relativas à escola. Em seguida, seguindo o mesmo padrão, são apresentados os resultados relativos aos modelos com zeros inflados: ZIP, ZINB e PQL.

Como pode ser visto na Tabela 4, todos os coeficientes foram significativos a um grau de significância de 1%. A grande maioria das variáveis também teve seu coeficiente no sinal esperado, exceto em relação à variável se mora em zona urbana, em que se esperava que os estudantes das zonas urbanas tivessem um número menor de repetências. Em parte, pode-se justificar tal fato pelo recorte amostral utilizado, que considera apenas indivíduos matriculados, se se assumir que na zona rural a evasão escolar ocorre mais cedo. Além disso, as variáveis indicativas de trabalho estiveram relacionadas ao número menor de defasagens idade-escolaridade. Uma possível explicação desse resultado são fatores motivacionais, como a manutenção do emprego na ocasião da conclusão da escola. Importante ainda sublinhar o melhor desempenho das escolas da rede estadual em relação àquelas da rede municipal.

Ao adicionar variáveis relativas à escola, observamos que as mudanças foram no sentido de arrefecer a importância de algumas variáveis, conforme a Tabela 4. Os resultados de um menor número de defasagens idade-escolaridade relacionados a docentes mais escolarizados e com um número maior de horas-aula já era esperado. Porém, um menor tamanho de turma não está relacionado a uma defasagem idade-escolaridade menor.

Para turmas muito pequenas, é possível que este efeito não ocorra, havendo um tamanho ótimo de turma.

Tabela 4 **Fatores relacionados com a distorção idade-série dos estudantes de escolas públicas de acordo com modelos de contagem**

	Modelo 1			Modelo 2		
	MQO	**Poisson**	**NB**	**MQO**	**Poisson**	**NB**
Idade	***0,175	***0.180	***0.177	***0.267	***0.247	***0.249
	(0,00282)	(0.00238)	(0.00238)	(0.00350)	(0.00248)	(0.00252)
Mulher	***-0,224	***-0.235	***-0.233	***-0.207	***-0.211	***-0.212
	(0,0118)	(0.0127)	(0.0127)	(0.0111)	(0.0119)	(0.0119)
Branco	***-0,118	***-0.138	***-0.139	***-0.0890	***-0.105	***-0.106
	(0,0125)	(0.0138)	(0.0138)	(0.0119)	(0.0130)	(0.0130)
Mora com a mãe	***-0,176	***-0.160	***-0.167	***-0.154	***-0.142	***-0.147
	(0,0189)	(0.0155)	(0.0157)	(0.0178)	(0.0145)	(0.0146)
Educação da pessoa de referência	***-0,0480	***-0.0469	***-0.0479	***-0.0433	***-0.0431	***-0.0437
	(0,00165)	(0.00153)	(0.00153)	(0.00155)	(0.00143)	(0.00143)
Logaritmo da renda *per capita* familiar	***-0,0670	***-0.0705	***-0.0735	***-0.0592	***-0.0628	***-0.0643
	(0,00582)	(0.00479)	(0.00501)	(0.00541)	(0.00448)	(0.00456)
Tamanho da família	***0,0572	***0.0488	***0.0500	***0.0577	***0.0474	***0.0491
	(0,00417)	(0.00331)	(0.00340)	(0.00392)	(0.00314)	(0.00319)
Urbano	0,00844	**0.0339	***0.0394	***1.185	***1.102	***1.103
	(0,0158)	(0.0135)	(0.0136)	(0.0266)	(0.0245)	(0.0242)
Estadual	***-0,582	***-0.590	***-0.570	***-0.315	***-0.309	***-0.308
	(0,0142)	(0.0141)	(0.0140)	(0.0142)	(0.0144)	(0.0143)
Trabalha até 20 horas semanais	***-0,0959	***-0.117	***-0.131	***-0.163	***-0.168	***-0.176
	(0,0362)	(0.0270)	(0.0276)	(0.0342)	(0.0253)	(0.0257)
Trabalha mais de 20 horas semanais	***-0,119	***-0.113	***-0.124	***-0.175	***-0.147	***-0.156
	(0,0393)	(0.0294)	(0.0298)	(0.0370)	(0.0272)	(0.0277)
Trabalhos domésticos até 10 horas semanais	***-0,115	***-0.0878	***-0.0937	***-0.104	***-0.0861	***-0.0896
	(0,0134)	(0.0133)	(0.0134)	(0.0126)	(0.0124)	(0.0124)
Trabalhos domésticos mais de 10 semanais	***-0,202	***-0.141	***-0.149	***-0.175	***-0.121	***-0.124
	(0,0184)	(0.0175)	(0.0176)	(0.0174)	(0.0166)	(0.0166)
Nordeste	***-0,227	***-0.199	***-0.198	***-0.280	***-0.154	***-0.177
	(0,018)	(0.0139)	(0.0139)	(0.0176)	(0.0136)	(0.0138)

(continua)

Tabela 4 **(continuação)**

			Modelo 1			Modelo 2
	MQO	**Poisson**	**NB**	**MQO**	**Poisson**	**NB**
Sudeste	***-0,335	***-0.324	***-0.317	***-0.305	***-0.199	***-0.222
	(0,0174)	(0.0165)	(0.0164)	(0.0196)	(0.0183)	(0.0183)
Sul	***-0,235	***-0.205	***-0.205	***-0.624	***-0.513	***-0.541
	(0,021)	(0.0215)	(0.0213)	(0.0234)	(0.0242)	(0.0242)
Centro-Oeste	***-0,177	***-0.134	***-0.121	***-0.333	***-0.218	***-0.232
	(0,0232)	(0.0232)	(0.0233)	(0.0243)	(0.0245)	(0.0244)
Distrito Federal	***-0,232	***-0.207	***-0.209	***-0.133	*-0.0735	**-0.0954
	(0,0381)	(0.0423)	(0.0426)	(0.0381)	(0.0426)	(0.0427)
Média de alunos por turma	***0,194	***-1.046	***-1.007	***-0.114	***-0.102	***-0.104
	(0,0513)	(0.0456)	(0.0463)	(0.00235)	(0.00211)	(0.00211)
Horas-aula				***0.185	***0.0967	***0.105
				(0.0188)	(0.0149)	(0.0154)
Porcentual de docentes com ensino superior				***-0.0127	***-0.00937	***-0.00917
				(0.000582)	(0.000497)	(0.000495)
Constante				***1.206	0.0560	0.0527
				(0.0880)	(0.0733)	(0.0751)
Alpha			***-1.601			***-2.411
			(0.0476)			(0.0837)
Observações	47.047	47,047	47,047	47,047	47,047	47,047

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

*Erros-Padrões Robustos entre Parênteses *** p<0,01, ** p<0,05, * p<0,1.*

Além disso, com a inclusão das variáveis do ambiente escolar, o coeficiente indicativo de gênero, de cor e da moradia da mãe e da renda familiar reduziu em magnitude, indicando uma maior igualdade entre os diferentes indivíduos. Já o coeficiente da variável relativa a morar em zona urbana aumentou, possivelmente indicando desigualdade entre a estrutura das escolas rurais e urbanas em que as urbanas estão em pior situação. As escolas estaduais também apresentaram redução em seu coeficiente, e alguns coeficientes regionais modificaram-se bastante, indicando que parte da desigualdade regional pode ser explicada pela estrutura das escolas em termos de alunos por turma, percentual de docentes com ensino superior e número de horas-aula.

Nos modelos com zeros inflados, apresentados nas Tabelas 5 e 6 abaixo, pode-se observar algumas diferenças em relação aos modelos anteriores.

Em relação à significância estatística, os modelos apresentam coeficientes, no geral, menos significativos, embora os coeficientes não significativos estejam relacionados majoritariamente à porção *logit*. Com relação aos sinais das porções relativas ao processo de contagem, elas não mostram grandes surpresas para o modelo restrito, com exceção dos coeficientes relativos ao trabalho e trabalhos domésticos, embora em alguns casos não tenham sido significativos. Já no que diz respeito aos coeficientes da porção *logit*, destaque especial é dado para as variáveis relativas a trabalho e trabalho doméstico, que não foram significativas ou apenas fracamente. Com relação aos sinais, esses mostraram padrão de comportamento similar ao do processo de contagem. Isso significa que as variáveis que estão relacionadas ao maior (menor) número de distorções também estão relacionados à maior (menor) possibilidade de ocorrência de distorção. Exceção é a variável relativa à idade, que merece uma análise mais atenta.

A variável idade tem sinal positivo no processo de contagem, indicando que o número de distorções cresce na medida em que a idade avança, um resultado esperado. Porém, na porção inflada dos modelos, o sinal também foi positivo, indicando uma menor probabilidade de ocorrência de distorção, ou maior probabilidade de não ocorrer qualquer distorção. À primeira vista, este parece ser um resultado inusitado, uma vez que se espera que uma maior idade esteja associada a maiores distorções. Entretanto, esses coeficientes passam a fazer sentido quando se leva em consideração que a divisão do processo de geração de dados acaba por dividir a amostra em dois grupos de estudantes: aqueles que nunca vão ter qualquer distorção e aqueles que podem vir a apresentar algum número de distorção, restando determinar o número delas. No primeiro caso, fala-se dos zeros estruturais, enquanto que no segundo fala-se em zeros acidentais. Sendo assim, conforme a idade avança e não ocorre qualquer distorção, as chances de se pertencer ao grupo dos alunos que nunca apresentam qualquer repetência crescem, o que é uma manifestação do chamado processo de "seleção" dos alunos, no qual apenas aqueles com bom desempenho se mantêm na escola.[10]

Os coeficientes estimados para as variáveis ligadas às escolas para a porção da contagem apresentam sinal negativo, estando associados à menor distorção. Para a porção *logit*, porém, as evidências são mais divergen-

10 Tal resultado foi encontrado para outras especificações de amostra, incluindo amostras que incorporaram todas as observações de crianças e jovens entre 8-18 anos de idade, estando estes matriculados ou não em instituições de ensino públicas ou privadas.

tes: para os modelos estimados por máxima verossimilhança, um maior número de alunos parece estar associado a uma menor probabilidade de ocorrência de distorções, uma maior quantidade de aulas está associada a maior probabilidade de ocorrência e a porcentagem de docentes com ensino superior não foi significativa. Uma hipótese é que a dedicação do docente (medido em um maior número de horas-aula) e também sua formação, com a conclusão do ensino superior, possivelmente traz benefícios aos estudantes, mas também pode aumentar seu grau de exigência. Sendo assim, a escolaridade e o número de horas-aula parecem ser uma faca de dois gumes: o ambiente de aprendizado melhora ao mesmo tempo em que aumenta o nível de conhecimento exigido dos alunos.

Ao adicionar as variáveis de contexto escolar, observa-se que para os modelos ZIP e ZINB houve, de modo geral, uma redução no coeficiente do processo de contagem, com exceção dos coeficientes relativos ao trabalho, zona urbana, idade e região Sul. Já na porção *logit* desses modelos ocorreu um comportamento menos homogêneo, com alguns coeficientes aumentando e outros caindo, a depender do modelo, embora os coeficientes tenham em sua maioria aumentado de magnitude. Como as variáveis associadas estão diretamente relacionadas às características escolares, as variáveis de zona urbana e escola estadual merecem destaque. A variável indicativa da localização em área urbana tem significância estatística apenas com as variáveis contextuais, estando relacionada ao maior número de distorção e à menor probabilidade de nenhuma distorção ocorrer. A variável indicativa de escola estadual perde magnitude com o maior número de variáveis relativas às escolas, estando mais associada a um número menor de distorções e a uma probabilidade maior de não ocorrer qualquer distorção.

Tabela 5 **Fatores relacionados com a distorção idade-série dos estudantes de escolas públicas de acordo com modelos de contagem zero inflados – Modelo 1**

	ZIP – Contagem	ZIP - Logit	ZINB – Contagem	ZINB - Logit	PQL - Contagem	PQL – Logit
Idade	***0,204	***0,327	***0,200	***0,357	***0,231	***0,284
	(0,00245)	(0,0164)	(0,00247)	(0,0176)	(0,00982)	(0,0211)
Mulher	***-0,159	***0,595	***-0,166	***0,643	***-0,148	***0,384
	(0,0136)	(0,0722)	(0,0134)	(0,0822)	(0,0166)	(0,066)

(continua)

Tabela 5 **(continuação)**

	ZIP – Contagem	ZIP - Logit	ZINB – Contagem	ZINB - Logit	PQL - Contagem	PQL – Logit
Branco	***-0,0941	***0,304	***-0,0989	***0,322	***-0,0922	0,0856
	(0,0147)	(0,0718)	(0,0145)	(0,0796)	(0,0163)	(0,0525)
Mora com a mãe	***-0,119	***0,489	***-0,126	***0,542	***-0,0873	***0,286
	(0,0162)	(0,105)	(0,0163)	(0,123)	(0,0217)	(0,0844)
Educação da pessoa de referência	***-0,0410	***0,0484	***-0,0422	***0,0469	***-0,0389	***0,0260
	(0,00174)	(0,0123)	(0,00169)	(0,0133)	(0,00189)	(0,00819)
Logaritmo da renda *per capita* familiar	***-0,0435	**0,339	***-0,0448	***0,437	**-0,0328	**0,192
	(0,00833)	(0,135)	(0,00668)	(0,131)	(0,0138)	(0,0768)
Tamanho da família	***0,0426	*-0,0536	***0,0451	-0,0377	***0,0482	-0,00244
	(0,00351)	(0,028)	(0,00345)	(0,0294)	(0,00384)	(0,0167)
Urbano	0,0195	*-0,171	0,0205	**-0,215	*-0,0326	***-0,332
	(0,0142)	(0,0924)	(0,0142)	(0,107)	(0,0179)	(0,0683)
Estadual	***-0,402	***1,550	***-0,407	***1,952	0,0238	***2,638
	(0,0162)	(0,142)	(0,0164)	(0,23)	(0,0771)	(0,68)
Trabalha até 20 horas semanais	***-0,0737	*0,210	**-0,0727	*0,259	-0,0377	*0,188
	(0,0284)	(0,119)	(0,0288)	(0,134)	(0,0358)	(0,103)
Trabalha mais de 20 horas semanais	-0,0452	0,12	-0,0497	0,12	0,00597	0,127
	(0,0313)	(0,108)	(0,0322)	(0,121)	(0,0388)	(0,0913)
Trabalhos domésticos até 10 horas semanais	***-0,110	-0,124	***-0,114	*-0,165	***-0,105	-0,0795
	(0,014)	(0,0774)	(0,0139)	(0,0871)	(0,0155)	(0,0564)
Trabalhos domésticos mais de 10 semanais	***-0,164	-0,123	***-0,167	-0,152	***-0,161	*-0,130
	(0,0189)	(0,0923)	(0,0188)	(0,104)	(0,0218)	(0,0696)
Nordeste	***-0,170	***0,271	***-0,174	***0,279	***-0,176	*-0,116
	(0,0145)	(0,0919)	(0,0144)	(0,105)	(0,0173)	(0,0657)
Sudeste	***-0,302	0,166	***-0,311	0,0888	***-0,260	0,0204
	(0,0184)	(0,109)	(0,0181)	(0,12)	(0,0244)	(0,0842)
Sul	***-0,129	***0,537	***-0,140	***0,518	***-0,114	0,116
	(0,0227)	(0,113)	(0,0223)	(0,122)	(0,0274)	(0,0857)
Centro-Oeste	***-0,152	-0,142	***-0,150	-0,215	***-0,171	**-0,250
	(0,0258)	(0,139)	(0,0255)	(0,155)	(0,0335)	(0,127)
Distrito Federal	***-0,242	-0,266	***-0,244	-0,371	***-0,273	-0,31
	(0,0454)	(0,311)	(0,0451)	(0,367)	(0,0564)	(0,211)

(continua)

Tabela 5 **(continuação)**

	ZIP – Contagem	ZIP - Logit	ZINB – Contagem	ZINB - Logit	PQL - Contagem	PQL – Logit
Constante	***-1,552	***-10,53	***-1,507	***-12,20	***-1,972	***-8,088
	(0,0573)	(0,777)	(0,0523)	(0,772)	(0,145)	(1,16)
Alpha			***-2,534			
			(0,0998)			
Observações	47.047	47.047	47.047	47.047	47.047	47.047

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

*Erros-Padrões Robustos entre Parênteses *** p<0,01, ** p<0,05, * p<0,1.*

Tabela 6 **Fatores relacionados com a distorção idade-série dos estudantes de escolas públicas de acordo com modelos de contagem zero inflados – Modelo 2**

	ZIP – Contagem	ZIP - Logit	ZINB – Contagem	ZINB - Logit	PQL - Contagem	PQL – Logit
Idade	***0,263	***0,288	***0,263	***0,287	***0,467	***0,416
	(0,00278)	(0,0302)	(0,00285)	(0,0322)	(0,0509)	(0,031)
Mulher	***-0,153	***0,615	***-0,154	***0,616	*-0,0516	***0,274
	(0,013)	(0,0845)	(0,0132)	(0,0852)	(0,0295)	(0,0438)
Branco	***-0,0584	***0,353	***-0,0585	***0,355	0,0137	***0,160
	(0,0135)	(0,0813)	(0,0135)	(0,0827)	(0,0283)	(0,0458)
Mora com a mãe	***-0,107	***0,509	***-0,108	***0,509	0,0236	***0,264
	(0,0152)	(0,129)	(0,0154)	(0,13)	(0,034)	(0,0572)
Educação da pessoa de referência	***-0,0386	***0,0483	***-0,0386	***0,0481	***-0,0344	**0,0131
	(0,00165)	(0,014)	(0,00168)	(0,0141)	(0,00341)	(0,00645)
Logaritmo da renda *per capita* familiar	***-0,0442	**0,260	***-0,0444	*0,259	0,00379	***0,124
	(0,0073)	(0,131)	(0,00747)	(0,134)	(0,0142)	(0,0298)
Tamanho da família	***0,0402	***-0,103	***0,0404	***-0,103	***0,0392	-0,00782
	(0,00321)	(0,0306)	(0,00322)	(0,0309)	(0,00707)	(0,0128)
Urbano	***1,042	***-1,178	***1,042	***-1,200	***1,188	**0,218
	(0,0261)	(0,309)	(0,0264)	(0,353)	(0,0748)	(0,0991)
Estadual	***-0,167	***1,389	***-0,168	***1,396	***0,461	***1,198
	(0,0173)	(0,249)	(0,0178)	(0,254)	(0,0727)	(0,0846)
Trabalha até 20 horas semanais	***-0,123	0,239	***-0,123	0,241	-0,0283	0,135
	(0,0262)	(0,15)	(0,0262)	(0,152)	(0,0771)	(0,114)

(continua)

Tabela 6 **(continuação)**

	ZIP – Contagem	ZIP - Logit	ZINB – Contagem	ZINB - Logit	PQL - Contagem	PQL – Logit
Trabalha mais de 20 horas semanais	***-0,0998	-0,00732	***-0,1000	-0,00765	0,175	0,282
	(0,0292)	(0,127)	(0,0292)	(0,128)	(0,18)	(0,214)
Trabalhos domésticos até 10 horas semanais	***-0,0999	-0,119	***-0,100	-0,121	***-0,150	-0,081
	(0,0127)	(0,0883)	(0,0127)	(0,0891)	(0,0309)	(0,0493)
Trabalhos domésticos mais de 10 semanais	***-0,140	-0,149	***-0,140	-0,15	***-0,195	-0,102
	(0,0175)	(0,106)	(0,0175)	(0,107)	(0,0433)	(0,0689)
Nordeste	***-0,136	***0,488	***-0,137	**0,494	***-0,410	***-0,550
	(0,0141)	(0,197)	(0,0143)	(0,207)	(0,0825)	(0,0789)
Sudeste	***-0,0924	***1,160	***-0,0922	***1,186	-0,0317	**0,206
	(0,029)	(0,386)	(0,0301)	(0,441)	(0,0477)	(0,08)
Sul	***-0,300	***2,674	***-0,300	***2,723	***-0,316	0,213
	(0,0375)	(0,588)	(0,039)	(0,685)	(0,0813)	(0,136)
Centro-Oeste	***-0,117	***1,420	***-0,117	***1,451	***0,250	***0,575
	(0,0311)	(0,413)	(0,0316)	(0,474)	(0,0826)	(0,101)
Distrito Federal	0,0188	***1,432	0,0193	***1,467	0,0748	0,247
	(0,0491)	(0,497)	(0,0498)	(0,56)	(0,102)	(0,174)
Médio de alunos por turma	***-0,0844	***0,240	***-0,0844	***0,244	***-0,114	-0,0357
	(0,00295)	(0,0471)	(0,00304)	(0,055)	(0,0167)	(0,0218)
Número de horas-aula	0,0267	***-0,539	0,0268	***-0,545	***-0,211	***-0,357
	(0,0185)	(0,128)	(0,0186)	(0,134)	(0,054)	(0,0757)
Porcentagem de professor com ensino superior	***-0,0122	-0,00643	***-0,0122	-0,0066	***-0,0183	***-0,0173
	(0,00061)	(0,0081)	(0,00063)	(0,00824)	(0,0012)	(0,00265)
Constante	**-0,229	***-13,62	**-0,225	***-13,72	-0,371	***-2,889
	(0,0926)	(1,175)	(0,0939)	(1,251)	(0,234)	(0,666)
Alpha			***-5,709			
			(2,047)			
Observações	47.047	47.047	47.047	47.047	47.047	47.047

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

*Erros-Padrões Robustos entre Parênteses *** p<0,01, ** p<0,05, * p<0,1.*

A Tabela 7 sumariza os resultados de alguns testes comparando os modelos Poisson, Binomial Negativo, ZIP e ZINB. Como pode ser visto, os critérios de informação AIC e BIC foram menores para o caso do modelo ZINB quando a estimação levou em consideração variáveis relativas às escolas, enquan-

to que quando essas variáveis não foram incluídas, o critério AIC foi menor para o ZINB e o BIC para o ZIP. Os testes Vuong, utilizados para comparar os modelos inflados com suas contrapartes sem aumento do número de zeros favoreceu os modelos inflados, enquanto que o teste de razão de verossimilhança, usado para comparar os modelos Binomial Negativo a suas contrapartes Poisson favoreceram os modelos Binomial Negativo. Além disso, todas as estimativas do termo α relativo aos modelos Binomial Negativo foram significativas, como pode ser visto nas Tabelas 4, 5 e 6, o que leva a crer que a modelagem mais adequada seja aquela com inflação de zeros e heterogeneidade não observada estimados a partir dos modelos ZINB.

Tabela 7 **Testes de comparações de modelos**

Modelos		Vuong	LR	AIC	BIC
Sem Escolas	Poisson	---	---	6.20e+07	6.20e+07
	ZIP	19.60 Pr>z = 0.0000	---	115789.2	116174.6
	NB2	---	352.33 Prob>=chibar2 = 0.000	117305.9	117507.4
	ZINB	19.17 Pr>z = 0.0000	8.03 Pr>=chibar2 = 0.0023	115783.1	116177.3
Com Escolas	Poisson	---	---	6.53e+07	6.53e+07
	ZIP	20.53 Pr>z = 0.0000	---	120807	121140
	NB2	---	1066.49 Prob>=chibar2 = 0.000	121895.1	122070.4
	ZINB	17.77 Pr>z = 0.0000	174.64 Pr>=chibar2 = 0.0000	120634.3	120976.1

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

Há ainda que se considerar a estimação por PQL proposta por Staub e Winkelmann (2013), capaz de fornecer estimativas consistentes mesmo para o caso de especificação incorreta da distribuição. Conforme esses autores apontam, a estimação por meio de máxima quasi-verossimilhança faz com que haja uma perda de precisão em detrimento da maior robustez. Esse *trade-off* está refletido na considerável queda nos níveis de significância dos coeficientes estimados das variáveis relativas ao trabalho, dependência administrativa e local de moradia, nos processos de contagem, e das variáveis de cor da pele, tamanho da família, trabalho e trabalho doméstico, na porção *logit*.

Para a estimação por máxima quasi-verossimilhança, na porção de contagem, todas as variáveis relativas ao contexto escolar foram significativas e associadas a um menor número de distorções. Já na porção *logit*, tanto um maior número de aulas como uma maior porcentagem de professo-

res com ensino superior estão associados a uma maior possibilidade de ocorrência de distorção, enquanto o número de alunos por turma não foi estatisticamente significativo.

Para os modelos PQL é mais difícil encontrar um padrão na modificação dos coeficientes após a introdução das variáveis relativas ao contexto escolar, havendo inclusive coeficientes que mudaram de sinal ao adicionar tais variáveis. Para a variável indicativa de zona urbana, esse coeficiente ganha magnitude na porção de contagem, estando mais associado a maiores distorções, enquanto ele inverte sinal para a porção *logit*, estando associado a uma maior probabilidade de não ocorrer distorção. O coeficiente da variável indicativa de escola estadual também aumenta em magnitude, no caso do modelo de contagem, associado a maior número de distorções, enquanto que para a porção *logit* ele perde magnitude e está ligado a uma probabilidade maior de não ocorrência de distorções. No caso desta variável, os resultados, que são similares aos encontrados para os modelos estimados por máxima verossimilhança, podem ser explicados ao notar-se que cabe, de modo geral, ao governo estadual fornecer o ensino médio enquanto os governos municipais fornecem majoritariamente o ensino fundamental. Assim como o processo de "seleção" dos alunos pode explicar o coeficiente da idade, essa divisão de atribuições também pode explicar o coeficiente da variável indicativa de escola estadual.

Importante notar que, embora os coeficientes tenham mudado de magnitude com a estimação a partir da máxima quasi-verossimilhança, essas diferenças foram diferentes de acordo com a variável considerada. Os coeficientes relativos à idade mudaram bastante, enquanto os relativos à educação dos pais quase não foram modificados. Assim, há evidências conflitantes com relação ao processo pelo qual os valores da variável dependente são gerados, uma vez que há coeficientes similares estimados, indicando que o processo gerador de dados pode seguir uma distribuição de Poisson ou Binomial Negativo, enquanto os coeficientes diferentes apontam para uma distribuição distinta das enumeradas. Um teste de Wald para a equivalência dos coeficientes dos modelos ZIP, ZINB e PQL resultou na rejeição da hipótese nula, como pode ser visto na Tabela A1, no anexo[11]. Assim, há

11 Nos casos em que ocorrem comparações entre modelos iguais (ZIP com ZIP, por exemplo), se está comparando os coeficientes das estimativas que levam em conta as variáveis relativas às escolas (porcentagem de professores com ensino superior, tamanho das turmas e horas-aula) com aquelas estimações que deixam de lado essas variáveis. Todos os outros testes são referentes a modelos com as mesmas variáveis.

indícios de que a estimação por máxima quasi-verossimilhança é a mais adequada, devido a sua robustez.

Sendo assim, destaca-se as vantagens dos modelos com zeros inflados por dois motivos. Em primeiro lugar, eles se adequaram melhor aos dados, ao se mostrar superiores aos seus pares que não inflam o número de zeros com os resultados dos testes supracitados. De fato, dado que a proporção de zeros chega quase à metade da amostra, o uso desses modelos se faz necessário. Em segundo lugar, a introdução dos zeros inflados tem uma consequência teórica interessante, que é dividir a amostra em dois grupos de indivíduos: aqueles que nunca terão qualquer distorção e aqueles que podem eventualmente apresentar alguma distorção, restando determinar quantas serão. O coeficiente da porção *logit* revela que, conforme a idade avança, as probabilidades de não ocorrer qualquer distorção crescem, atuando o efeito "seleção": os melhores alunos continuam avançando sem qualquer repetência, enquanto os que repetem passam a apresentar maiores chances de voltar a repetir conforme a idade avança.

6 **Conclusão**

O objetivo deste artigo é fazer uma análise econométrica do número de distorções idade-série de alunos matriculados em escolas da rede pública a partir de modelos de contagem. Nos modelos estimados, destaca-se que a grande maioria das variáveis utilizadas apresentou o coeficiente significativo e com o sinal esperado. Sendo assim, um número menor de distorções idade-série está associado às pessoas do sexo feminino, brancas ou amarelas, cuja mãe mora no mesmo domicílio, com maior nível de educação familiar e cuja renda familiar *per capita* é maior. Um resultado distinto do esperado foi o coeficiente da variável referente ao trabalho, que se mostrou negativo, de modo que aqueles que trabalham tiveram menor número de defasagens. Além disso, as variáveis do contexto educacional, além de serem significativas e associadas a um menor número de distorções, foram capazes de tornar mais fracos alguns coeficientes importantes na literatura, como a moradia com a mãe e a educação máxima dentro da família. Este resultado é importante, pois ressalta que, ao agir de forma proativa nestas variáveis, é possível que os resultados educacionais em termos de distorção idade-série melhorem.

Também deve-se destacar as vantagens advindas do uso de modelos de contagem com zeros inflados. Em primeiro lugar, os testes realizados reforçam a importância de se introduzir o processo binomial na estimação. Além disso, ao dividir o processo gerador de contagens em dois, é possível dividir a amostra em dois grandes grupos de estudantes: aqueles que nunca apresentarão qualquer distorção 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, tal resultado é obtido a partir do coeficiente positivo encontrado para a variável idade na porção *logit* dos 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-série. Este resultado parece ser contra-intuitivo, dado que o coeficiente desta covariada na parcela de contagem dos modelos mostrou que a idade está positivamente relacionada com o número de distorções. A interpretação, porém, é que à medida que os anos passam e nenhuma distorção 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. Desse modo, a ocorrência de uma distorção estaria associada a uma maior probabilidade de ocorrência de outra no futuro, seja por meio de reprovação ou até mesmo de uma eventual evasão.

O emprego desta metodologia trouxe contribuições ao estudo da distorção idade-série, e desdobramentos podem ser feitos dentro deste mesmo paradigma. Destaca-se a necessidade de um estudo mais aprofundado das variáveis relativas ao contexto escolar, uma vez que as aqui empregadas foram obtidas a partir de agregações que podem estar escondendo importantes desigualdades na distribuição dos valores relativos às unidades escolares individuais. Finalmente, uma base de dados mais abrangente, incluindo observações de diversos anos para os mesmos indivíduos ou com acesso a informações relativas aos atributos individuais dos estudantes, 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.

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

CAMERON, A. C.; TRIVEDI, P. K. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. 2013.

COLEMAN, J. S. *et al*. *Equality of Educational Opportunity*. US Department of Health, Education and Welfare. Washington: U.S. Government Printing Office, 1966. Disponível em: <http://sociology.sunimc.net/htmledit/uploadfile/system/20110618/20110618140836102.pdf>. Acesso em: 12 set. 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. ½, p. 111-130, 2001.

FRANCO, A. M. De P. *Os determinantes na qualidade da educação no Brasil*. 2008. 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. 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, D. C. *Escolaridade das crianças no Brasil*: três ensaios sobre a defasagem idade série. 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. *Revista Brasileira de Economia*, 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.

MENEZES-FILHO, N.; VASCONCELLOS, L.; WERLANG, S. R. Avaliando o Impacto da Progressão continuada no Brasil. *Relatório de avaliação econômica 2*. Política de Progressão Continuada. São Paulo: Itaú, 2005.

MENEZES-FILHO, N. *et al*. *Avaliando o Impacto da Progressão Continuada nas Taxas de Rendimento e o Desempenho Escolar no Brasil*. 2008, p. 4-29.

MURNANE, R. J.; WILLETT, J. B. *Methods Matter*: Improving Causal Inference in Educational and Social Science Research. New York: Oxford University Press, 2011.

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 Polic*,.v. 4, n.1, p. 29-54, 2008.

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. 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.

RIOS-NETO, E. L. G.; CÉSAR, C. C; RIANI, J. de L. R. Estratificação educacional e progressão escolar por série no Brasil. *Política e Planejamento Econômico*, Brasília, v. 32, n. 3, 2002.

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, 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.

STAUB, K, E; WINKELMANN, R. Consist estimation of zero-inflated count models. *Health Economics*. n. 22, p. 673-686, 2013.

VUONG. Q. H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, n. 57, v. 2, p. 307-333, 1989

WILKELMANN, R. *Econometric analysis of count data*. Berlin: Springer, 2008.

Sobre os autores

Alysson Lorenzon Portella - alyssonlp@al.insper.edu.br
Mestre em Economia (CEDEPLAR/UFMG). Doutorando em Economia dos Negócios (INSPER).
Tanise Brandão Bussmann - tanisebussmann@unipampa.edu.br
Doutora em Economia do Desenvolvimento (PPGE/PUCRS). Professora Adjunta na UNIPAMPA.
Ana Maria Hermeto de Oliveira - ahermeto@cedeplar.ufmg.br
Doutora em Demografia (CEDEPLAR/UFMG). Professora Associada na UFMG.

Sobre o artigo

Recebido em 16 de novembro de 2015. Aprovado em 19 de julho de 2016.

APÊNDICE

Tabela A1 **Teste Wald para comparação de equivalência de coeficientes**

	Modelos com variáveis escolares					
	Conjunto completo de variáveis					
	ZIP		ZINB		PQL	
	Contagem	Inflado	Contagem	Inflado	Contagem	Inflado
ZIP	2505,07	188,13	6,93	5,92	119,64	610,67
Prob > chi2	0,000	0,000	0,9845	0,9937	0	0
ZINB			2513,55	171,8	121,53	573,37
Prob > chi2			0,000	0,000	0	0
PQL					886,41	290,53
Prob > chi2					0,000	0,000

	educ_ref, log_rend_pc_fam, tamanho_familia e urbano					
	ZIP		ZINB		PQL	
	Contagem	Inflado	Contagem	Inflado	Contagem	Inflado
ZIP	2260,33	39,26	6,12	3,69	27,68	92,09
Prob > chi2	0,000	0,000	0,1903	0,4489	0,000	0,000
ZINB			2323,72	37,81	28,32	89,99
Prob > chi2			0,000	0,000	0,000	0,000
PQL					311,09	31,71
Prob > chi2					0,000	0,000

	educ_ref e log_rend_pc_fam					
	ZIP		ZINB		PQL	
	Contagem	Inflado	Contagem	Inflado	Contagem	Inflado
ZIP	19,78	2,2	5,69	3,6	15,98	59,91
Prob > chi2	0,0001	0,3334	0,0582	0,16	0,0003	0,000
ZINB			38,54	0,37	16,59	58,11
Prob > chi2			0,000	0,8295	0,0002	0,000
PQL					15,32	7,12
Prob > chi2					0,0005	0,0284

(continua)

Tabela A1 **(continuação)**

	Modelos sem variáveis escolares					
	Conjunto completo de variáveis					
		ZIP		**ZINB**		**PQL**
	Contagem	**Inflado**	**Contagem**	**Inflado**	**Contagem**	**Inflado**
ZIP	2505,07	188,13	88,74	44,69	67,77	290,04
Prob > chi2	0,000	0,000	0,000	0,000	0,000	0,000
ZINB			2513,55	171,8	67,6	281,84
Prob > chi2			0,000	0,000	0,000	0,000
PQL					886,41	290,53
Prob > chi2					0,000	0,000
	educ_ref, log_rend_pc_fam, tamanho_familia e urbano					
		ZIP		**ZINB**		**PQL**
	Contagem	**Inflado**	**Contagem**	**Inflado**	**Contagem**	**Inflado**
ZIP	2260,33	39,26	56,88	17,75	11,12	85,73
Prob > chi2	0,000	0,000	0,000	0,0014	0,0252	0,000
ZINB			2323,72	37,81	12,36	79,97
Prob > chi2			0,000	0,000	0,0148	0,000
PQL					311,09	31,71
Prob > chi2					0,000	0,000
	educ_ref e log_rend_pc_fam					
		ZIP		**ZINB**		**PQL**
	Contagem	**Inflado**	**Contagem**	**Inflado**	**Contagem**	**Inflado**
ZIP	19,78	2,2	35,63	14,35	3,43	63,73
Prob > chi2	0,0001	0,3334	0,000	0,0008	0,1798	0,000
ZINB			38,54	0,37	7,43	61,78
Prob > chi2			0,000	0,8295	0,0244	0,000
PQL					15,32	7,12
Prob > chi2					0,0005	0,0284

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