UMA ANÁLISE DA REPROVAÇÃO E DA EVASÃO NO ENSINO MÉDIO CATARINENSE USANDO MICRODADOS ADMINISTRATIVOS[1]

Max Cardoso de Resende[2]
Francis Carlo Petterini[3]

A partir de um inexplorado banco de microdados longitudinais de alunos do ensino médio da rede pública catarinense, o artigo analisa três pontos pouco abordados na literatura sobre a reprovação e a evasão no Brasil. O primeiro é que em dados administrativos a evasão é parcialmente observada por meio do atrito, e discute-se como contornar isso com uma tabulação sistemática da informação. Segundo, no caso catarinense, é possível observar a renda familiar, o que não acontece no Censo Escolar, do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep); então, essa pesquisa pode ser a primeira a modelar evasão e reprovação com dados administrativos, se controlando por essa covariável. Terceiro, faz-se uma abordagem econométrica, assumindo-se que fatores não observáveis podem afetar concomitantemente a reprovação e a evasão. Assim, estima-se que o efeito da derradeira reprovação na probabilidade de evasão, tendo-se o controle das características do aluno e da escola, estaria próximo de 35 pontos percentuais (p.p.). No entanto, esse efeito seria mais perto de 20 p.p., para estudantes que não estão em distorção idade-série, e poderia chegar a 45 p.p., para alunos que já reprovaram alguma vez antes.

Palavras-chave: reprovação; evasão; ensino médio; dados administrativos.

AN ANALYSIS OF GRADE RETENTION AND DROPOUT IN THE HIGH SCHOOLS OF SANTA CATARINA USING ADMINISTRATIVE MICRODATA

Based on an unexplored microdata database of secondary students from the public schools of Santa Catarina, the article analyzes three points that are rarely addressed in the literature of grade retention and dropout in Brazil. The first is that, in administrative data, school dropout is partially observed through of the attrition, and it is discussed how to solve this with a systematic tabulation of information. Second, in the case of Santa Catarina it is possible to observe family income, which does not happen in the School Census, so this research may be the first to model grade retention and school dropout with Brazilian administrative data controlling this covariate. Third, an econometric approach is taken assuming that unobservable factors can affect both retention and dropout. Thus, the effect of the ultimate retention on the probability of dropout is estimated, controlling the characteristics of the student and the school, it would be close to 35 p.p. However, this effect would be closer to 20 p.p. for students who are not in age-grade distortion, and it could reach 45 p.p. for students who was retained at least once.

Keywords: grade retention; dropout; secondary school; administrative microdata.

1. DOI: http://dx.doi.org/10.38116/ppp61art1

2. Pesquisador do Núcleo de Econometria Aplicada da Universidade Federal de Santa Catarina (UFSC). *E-mail*: <max.resende@ufsc.br>. Orcid: <https://orcid.org/0000-0002-0990-8192>.

3. Professor da UFSC. *E-mail*: <f.petterini@ufsc.br>. Orcid: <https://orcid.org/0000-0003-4410-0970>.

ANÁLISIS DE RETENSIÓN Y DESERCIÓN EN LA ESCUELA SECUNDARIA MEDIANTE MICRODATOS ADMINISTRATIVOS: EVIDENCIAS PARA SANTA CATARINA

Basado en un banco inexplorado de microdatos longitudinales de estudiantes de secundaria de la red pública de Santa Catarina, la investigación aborda tres puntos que rara vez se abordan en la literatura que analiza la retensión y deserción escolar. Primero, en los datos administrativos, la evasión se observa parcialmente a través de la fricción. Luego, discutimos cómo solucionar esto con una tabulación sistemática de la información. En segundo lugar, en el caso de Santa Catarina es posible observar el ingreso del hogar, que es raro en los datos administrativos en Brasil. Por lo tanto, esta investigación puede ser la primera en modelar la retensión y deserciíon con datos administrativos teniendo em cuenta esta covariable. Tercero, se adopta un enfoque econométrico asumiendo que los factores no observables pueden afectar tanto la retensión como la deserción. Así pues, se estima el efecto de retensióin sobre la probabilidad de deserción, controlando las características del estudiante y la escuela, que estaría cerca de 35 p.p. Sin embargo, este efecto se reduciría a 20 p.p. para los estudiantes que no están en distorsión de edad y podría alcanzar hasta 45 p.p. para los estudiantes que hayan repetido al menos una vez.

Palabras clave: retensión; deserción; escuela secundaria; base de datos administrativa.

JEL: A21; C35; I21.

1 INTRODUÇÃO

A estrutura atual do ensino formal no Brasil resulta de discussões consolidadas na Constituição Federal de 1988 (CF/1988), na Lei de Diretrizes e Bases (LDB) de 1996, no Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (Fundef), entre 1998 e 2006, no Fundo de Manutenção e Desenvolvimento da Educação Básica (Fundeb), desde 2007, e na Base Nacional Comum Curricular de 2017. Na esteira dessa construção, o país melhorou em muitos indicadores educacionais, destacando-se que a taxa de escolarização líquida do ensino fundamental (parcela da população entre 7 e 14 anos na escola) tem se mantido perto de 100% nas últimas duas décadas. Isso ilustra que o acesso ao ensino formal está praticamente universalizado, embora a qualidade dessa instrução seja questionável, porque os estudantes seguem tendo resultados ruins em exames de proficiência – *e.g.*, no Programa Internacional de Avaliação de Estudantes (Pisa).

De toda forma, há um problema persistente na transição entre os ensinos fundamental e médio, particularmente com os alunos da rede pública. Porque muitos deles simplesmente não se interessam em fazer matrícula no ensino secundário; dos que se matriculam, cerca de um terço reprova na primeira série, por baixa proficiência ou excesso de faltas; e boa parte dos que reprovam não voltam a frequentar a escola (Inep, 2017). Isso gera custos sociais, porque quem interrompe os estudos tem maior probabilidade de desemprego, de receber menores salários, de depender mais dos serviços de assistência social etc. (Eckstein e Wolpin, 1999; Rumberger e Lim, 2008; De Witte *et al.*, 2013). Portanto, a reprovação e a evasão no ensino médio da rede pública ainda são uma questão a ser estudada e tratada no Brasil.

Para promover essas investigações, uma literatura vem explorando pesquisas amostrais e dados administrativos. A Pesquisa Nacional por Amostra de Domicílios do Instituto Brasileiro de Geografia e Estatística (PNAD/IBGE) tem sido muito usada nesse sentido, com a qual se costuma tabular coortes de alunos e, então, analisar indicadores agregados de fluxo escolar por intermédio de modelos matemáticos como o profluxo (Fletcher e Ribeiro, 1996; Klein, 2003; Golgher e Rios-Neto, 2005). Além disso, a extinta Pesquisa Mensal de Emprego (PME) do IBGE já foi bem empregada em estudos longitudinais, por meio de análises de regressão, tendo como variável dependente indicadores de reprovação e/ou evasão do estudante (Duryea, 1998; Leon e Menezes-Filho, 2002; Souza *et al.*, 2012).

Por sua vez, os dados administrativos provêm das secretarias subnacionais de Educação e têm a vantagem de permitir o acompanhamento de muitos alunos por anos subsequentes, assim como eventuais características das escolas, dos professores e dos colegas – que são potencialmente relevantes para entender a trajetória individual do estudante. No entanto, também têm a desvantagem do difícil acesso ao pesquisador, por questões de sigilo de informações pessoais (Inep, 2009). Além disso, até mesmo quando o acesso é possível, existem vários desafios de tabulação e rastreamento longitudinal das pessoas (Lee, 2010; Oliveira e Soares, 2012; Shirasu e Arraes, 2015; Inep, 2017).

Destarte, independentemente da fonte dos dados e da estratégia empírica utilizada, no Brasil e em outros países a literatura tem consistentemente encontrado correlações entre as situações de reprovação e de evasão, e dessas condições com os indicadores de *background* familiar (Roderick, 1994; Riani e Rios-Neto, 2008; Chowdry, Crawford e Goodman, 2011). Embora pareça que um eventual mecanismo de causalidade da evasão ainda não esteja bem compreendido. Consequentemente, nota-se que ainda não existe um consenso de como tratar esse problema de forma adequada.

No contexto, a pesquisa aqui relatada obteve acesso inédito a um inexplorado banco de microdados longitudinais da Secretaria de Estado da Educação de Santa Catarina (SED/SC). Ao descrever a tabulação e a análise dessas informações, promovem-se três contribuições para essa literatura, no que tange à mitigação de vieses analíticos da evasão em face de: i) erro de medida da evasão em dados administrativos; ii) omissão de covariável relevante; e iii) elementos de simultaneidade entre a condição de reprovação e a decisão de evasão.

A primeira contribuição está na continuidade da discussão levantada por Oliveira e Soares (2012) e Inep (2017), no sentido de que a evasão e o atrito são parcialmente observáveis em dados administrativos. Isto é, se, entre anos subsequentes, o aluno é encontrado em alguma escola, o pesquisador sabe que não houve uma evasão. Caso contrário, isso pode indicar evasão; mas como a

informação que provém da escola pode apresentar erros e omissões, isso também pode ser simplesmente um atrito. Assim, por intermédio do caso de Santa Catarina, discute-se como mitigar erros de medida da evasão com o tratamento sistemático da informação bruta.

A segunda contribuição decorre do fato de que, no caso de Santa Catariana, há informações sobre a renda familiar, o que, por exemplo, não ocorre no Censo Escolar, do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep); (Oliveira e Soares, 2012; Shirasu e Arraes, 2015). Assim, esta pesquisa possivelmente é a única a estudar as probabilidades de reprovação e evasão usando microdados administrativos brasileiros, em que se consegue analisar essa covariável. Isso é relevante porque o efeito do seu viés de omissão pode ser considerável nas regressões. Nesse sentido, mostram-se evidências de que a renda está bem correlacionada com indicadores mais comuns de serem observados – *e.g.*, possuir equipamentos de tecnologia ou ser beneficiário de algum programa social. Então, evidenciam-se características que servem como *proxies* da renda para pesquisas em que isso não pode ser diretamente observado. Além disso, nota-se que o efeito da renda nas probabilidades de reprovação/evasão é menor que a influência de outras covariáveis, notadamente da distorção idade-série.

A terceira contribuição está na observação de que muitas pesquisas estimam dois modelos em sequência: no primeiro, a variável dependente é a reprovação; no segundo, a variável dependente é a evasão, sendo a reprovação uma das covariáveis a fim de medir seu efeito na evasão. Argumenta-se que isso pode gerar um viés analítico, porque as covariáveis omitidas do primeiro devem afetar a variável dependente do segundo. Alternativamente, mostra-se que as condições de reprovação e evasão podem ser estimadas simultaneamente, evitando-se um problema dessa natureza. Então, pelos dados de Santa Catarina, nota-se que o efeito de uma reprovação nas chances de evasão de um aluno em distorção idade-série, por exemplo, seria de 45 pontos percentuais (p.p.) e 35 p.p., com e sem a hipótese de simultaneidade, respectivamente.

Além desta introdução, o artigo possui mais cinco seções. A seção 2 faz uma revisão da literatura com foco no Brasil. A seção 3 apresenta os dados de Santa Catarina e discute como medir a evasão. A seção 4 aborta a simultaneidade entre a condição de reprovação e a decisão de evasão. A seção 5 mostra resultados estimados entre os modelos. Por fim, na seção 6, tem-se a conclusão.

2 REVISÃO DA LITERATURA

Nas décadas de 1980 e 1990, os trabalhos de Philip Fletcher, Sérgio Ribeiro e Rubem Klein fizeram três grandes contribuições para entender melhor o fluxo escolar no Brasil. A primeira foi mostrar que muitos indicadores oficiais estavam equivocados, porque os alunos não eram acompanhados por anos subsequentes.

Notadamente, quando se observava um elevado quantitativo de estudantes mais velhos matriculados nas primeiras séries do ensino fundamental, acreditava-se que eles enfrentavam uma dificuldade para entrar na escola na idade certa, possivelmente por causa da falta de vagas. Além disso, achava-se que a taxa de evasão nas séries iniciais era muito alta, em razão de o número de matriculados nas séries subsequentes cair abruptamente.

Todavia, quando esses autores exploraram de forma inédita os microdados das primeiras PNADs – nas quais se perguntava para as pessoas se estavam ou estiveram estudando, se reprovaram ou abandonaram a escola etc. –, ficou claro que os alunos costumavam entrar no sistema na idade certa, mas reprovavam demais. Então, evidenciou-se que as taxas de reprovação eram mais altas e as taxas de evasão eram mais baixas do que mostravam os números oficiais – pelo menos nas séries iniciais do ensino formal (Fletcher e Ribeiro, 1987; Klein e Ribeiro, 1991).

A segunda contribuição dessas pesquisas foram os modelos matemáticos – com destaque para o profluxo –, desenvolvidos para a construção e a análise de vários índices sobre o desempenho do sistema educacional, baseados nos dados transversais das PNADs, no Censo Demográfico do IBGE e em outros levantamentos predecessores das edições mais modernas do Censo Escolar. Com esses procedimentos, foi possível analisar o fluxo escolar no Brasil por intermédio de uma miríade de recortes geográficos e sociais, sem necessariamente observar longitudinalmente os alunos (Fletcher e Ribeiro, 1996; Klein, 2003; Rios-Neto e Riani, 2004; Golgher e Rios-Neto, 2005; Brasil, 2015).

A terceira contribuição vem da percepção de que as taxas de reprovação eram recorrentemente altas em muitas regiões e recortes sociais. O que levou Ribeiro (1991) a formular a hipótese de que no Brasil haveria uma "pedagogia da repetência" – o que também foi postulado em pesquisas de outros países (Roderick, 1994; Stearns *et al.*, 2007). Isto é, a reprovação seria um instrumento tradicionalmente usado para aumentar a dedicação dos alunos mediante o medo, que funcionaria se houvesse uma espécie de *background* familiar, em que o aluno seria cobrado e orientado a partir de casa. Mas se o número de estudantes com menor suporte familiar cresce – seja qual for o motivo – e a escola mantém os padrões passados de promoção escolar; então, muitos alunos, após sucessivos fracassos em avançar de uma série para a próxima, desanimariam e desistiriam da escola. Então, a existência de inflexibilidade nos limiares de proficiência para a promoção entre séries escolares poderia causar aumento nas taxas de evasão; ou talvez pudesse ser a *gota d'água* para deixar a escola, em face de precário *background* familiar.

Tendo-se como motivação essa hipótese, no fim dos anos 1990, surge uma vertente dessa literatura interessada em mensurar a probabilidade de evasão condicionada na situação de reprovação – além de outras covariáveis para controlar

os efeitos do *background* familiar, como a escolaridade dos pais e a renda. Nesse sentido, como existe um lapso temporal entre a reprovação do aluno e a verificação da evasão, era imprescindível um acompanhamento longitudinal. Assim, Duryea (1998) e Leon e Menezes-Filho (2002) foram os primeiros a explorar o fato de que as amostragens da extinta PME permitiram observar a situação de reprovação/evasão de milhares de estudantes entre anos subsequentes, bem como as características familiares. Souza *et al*. (2012) continuaram explorando a pesquisa dos anos 2000 com esse propósito, ao observarem outros milhares de estudantes e agregarem outros controles na análise, como indicadores do mercado de trabalho.

Esse novo conjunto de análises usou abordagens econométricas parecidas, ao estimar dois modelos em sequência. No primeiro, a variável dependente era a situação de reprovação – ou aprovação – do aluno, dado o vetor de covariáveis; no segundo, a variável dependente era a situação do aluno em termos de evasão – ou de permanência na escola –, sendo a reprovação um elemento do vetor de covariáveis. Nesse sentido, todos encontraram evidências de que pessoas com pais menos escolarizados e nas faixas de renda mais baixas apresentavam maiores chances de reprovação e evasão. Além disso, particularmente no caso do ensino médio, notou-se que a probabilidade média não condicionada de evasão seria algo como 20%, mas quando condicionada na reprovação na primeira série aumentaria para perto de 40%. Consequentemente, isso gerou uma evidência em favor da validade da hipótese de um mecanismo de causalidade relacionado com a pedagogia da repetência.

Complementando os trabalhos a partir de pesquisas amostrais, e investigando a situação de evasão sem usar modelos como o profluxo ou explorar a PME, Neri (2015) analisou suplementos da PNAD nos quais era perguntado aos entrevistados "qual é o principal motivo (...) (de) não frequentar a escola?", tal que a resposta "falta de interesse" foi modal. Entre diversos aspectos notados, o que chama mais atenção nessa pesquisa é a sugestão de que muitas pessoas têm uma visão limitada dos resultados da educação sobre a renda futura e outros benefícios, o que ocorreria mais frequentemente entre estudantes com características de menor suporte familiar e que eventualmente estavam em distorção idade-série – isso, reflexo de reprovações em fases de ensino anteriores.

Um dos raros trabalhos que estuda a evasão no ensino médio usando microdados administrativos é Shirasu e Arraes (2015). Os autores exploram uma base da Secretaria de Educação do Ceará, ao observarem uma coorte de 33 mil alunos que iniciaram o ensino médio em 2008, rastreando-os entre 2009 e 2011. Por meio de regressões como as de Leon e Menezes-Filho (2002) e Souza *et al.* (2012), os resultados encontrados no Ceará assemelham-se muito com os identificados nas pesquisas que exploraram a PME: os indicadores de suporte familiar são importantes para explicar as chances de reprovação e evasão, e essas duas condições são

bastante correlacionadas. Nesse caso, estimou-se que uma reprovação na primeira série do ensino médio quase dobra a probabilidade de evasão condicionada no perfil do aluno.

A explicação de haver poucos trabalhos a partir de dados administrativos começa por notar que o aluno não era o foco do antigo Censo Escolar, então as secretarias subnacionais de Educação – que alimentavam o levantamento com informações recebidas das escolas em formulários manuscritos em papel – não se obrigavam a manter registros sistemáticos e longitudinais dos estudantes. Isso muda em 2007, quando se começou a usar um *software* chamado de Educacenso, que posteriormente evoluiu para uma ferramenta *on-line* de mesmo nome, no qual as escolas passaram a diretamente digitar as informações, inclusive dos alunos. Com isso, abriu-se a possibilidade do acompanhamento dos estudantes por anos subsequentes e, consequentemente, da pesquisa da reprovação e da evasão por características das escolas, dos professores, dos colegas etc. Todavia, essas bases de dados contêm informações pessoais sigilosas e não podem ser tornadas públicas de maneira simples (Inep, 2009). Ao contrário do que pode ser o senso comum, não se trata apenas de apagar uma coluna de uma planilha eletrônica, porque esses bancos de dados são gerenciados de formas complexas, em face de seus grandes volumes de informação (Brasil, 2015; Inep, 2017).

Destarte, até mesmo tendo acesso aos dados completos das edições mais modernas do Censo Escolar, Oliveira e Soares (2012) e Inep (2017) relatam desafios em construir uma tabulação longitudinal dos alunos, principalmente por conta de problemas relacionados com campos incompletos – *e.g.*, identificadores ausentes – ou inseridos de forma errada – ou seja, identificadores que não fazem sentido –, necessidade de deduplicação – *e.g.*, registros simultâneos em escolas diferentes – e incertezas sobre o tamanho e a natureza do atrito – isto é, estudantes são encontrados na base no ano t na 1ª série, desaparecem no ano $t+1$, e reaparecem no ano $t+2$ na 3ª série.

Nesse sentido, Oliveira e Soares (2012) merecem destaque porque descrevem um esforço inédito de montar um painel com os Censo Escolar tendo como unidade transversal o aluno, a fim de estudar a probabilidade de reprovação em face de outras covariáveis. E também porque levantam um ponto fundamental para estudar a evasão a partir de dados administrativos – apesar de não terem modelado a evasão nas suas regressões: "alunos que evadem não são (...) atrito, e sim resultado do processo educacional (...) atrito são alunos que não são encontrados pelo Censo, apesar de continuarem na escola" (*op. cit*, p. 11.). Nessa perspectiva, o atrito acontece principalmente quando o identificador do aluno é omitido ou modificado entre anos subsequentes – na mesma rede de ensino –, ou o estudante migra – para outra rede de ensino ou outro estado – e lhe é atribuído outro identificador. Além disso, também há óbitos não registrados nos sistemas e outros casos desconhecidos.

A discussão promovida por Oliveira e Soares (2012) – e posteriormente por Inep (2017) – mostra que em dados administrativos existe uma dicotomia evasão/atrito em matrículas descontinuadas, porque a situação de evasão não é perguntada aos estudantes – diferentemente das pesquisas amostrais. No contexto, como esses autores estavam preocupados em estudar a reprovação em um painel tabulado de forma inédita, o foco da discussão recaiu sobre as consequências do atrito nessas regressões. Assim, tratou-se o assunto com uma abordagem clássica, na forma usada por Ribas e Soares (2010), com o objetivo de analisar regressões que exploraram as informações da PME.

Alternativamente, nota-se que as ideias de observabilidade parcial de Poirier (1980) e Meng e Schmidt (1985) também podem ser aplicadas nessa questão, ao se definir três variáveis. Primeiro, $\tilde{A} = 1$ se aluno é observado no ano escolar t, ainda não terminou seus estudos e é visto novamente em $t + 1$ – portanto, não é um caso de atrito; e, $\tilde{A} = 0$ no caso contrário. Segundo, $E = 1$ para o caso de evasão – *i.e.*, deixar de fazer matrícula –, entre t e $t + 1$; e $E = 0$ no caso contrário. Terceiro, $F = 1$ para o caso de fazer uma matrícula não observada entre t e $t + 1$; e $F = 0$ no caso contrário. Dessa forma, $\tilde{A}$ é *totalmente observável*, e E e F são *parcialmente observáveis*. Isto é, apenas quando ocorre $\tilde{A} = 1$ se observa E e F, porque se sabe que ocorreu $E = 0$ e $F = 0$. Quando ocorre $\tilde{A} = 0$, não se sabe se é um caso de evasão ou um simples atrito ($E = 1$ e $F = 0$; ou, $E = 0$ e $F = 1$).

Poirier (2014) faz um levantamento da literatura que lidou com essa perspectiva de observabilidade parcial, em educação e outras linhas de pesquisa. Ao analisar os trabalhos listados pelo autor, nota-se a possibilidade de usar uma sequência de estratégias. Primeiro, busca-se uma forma de classificar o atrito com um tratamento sistemático da base de dados, de forma semelhante ao que foi feito por Oliveira e Soares (2012). Segundo, quando necessário, aplica-se uma estratégica econométrica mais sofisticada, com o objetivo de modelar a probabilidade de atrito não identificável. Nesse caso, costuma-se aliar uma chamada *hipótese de rotulagem*, com a finalidade de categorizar os casos parcialmente observados, e modelos de simultaneidade; notadamente, *probits* multinominais ou cópulas – que, por sua vez, guardam semelhança com alguns modelos mais sofisticados citados por Ribas e Soares (2010). No trabalho de Santa Catarina discutido a seguir, o primeiro procedimento parece ter sido suficiente para tratar a questão.

3 BASE DE DADOS

Quando a proposta do Educacenso passou a ser ventilada, a SED/SC iniciou o desenvolvimento de *software* chamado de Sistema de Gestão Educacional de Santa Catarina (Sisgesc), com o objetivo de fazer com que as escolas públicas estaduais registrassem mais informações do que as que precisavam ser enviadas ao Censo Escolar – de forma que o Sisgesc seria prioritariamente preenchido, e depois este

alimentaria automaticamente o Educacenso. Por exemplo, o sistema poderia montar os boletins bimestrais de desempenho dos alunos e, também, coletar informações do *background* familiar no ato da matrícula.

O projeto foi executado até meados de 2012, quando foi remodelado e reaplicado nos anos seguintes com outros nomes. Para a pesquisa ora relatada, obteve-se acesso inédito aos microdados dessa primeira fase. Na forma bruta, têm-se informações de cerca de 400 mil alunos que frequentaram pelo menos uma de mais de seiscentas escolas, entre 2008 e 2012, em qualquer das três séries regulares do ensino médio. Infelizmente, a base não aponta eventuais passagens pela educação de jovens e adultos (EJA), por colégios profissionalizantes e por outras formas de educação menos usuais. Além disso, os dados de 2012 estão muito incompletos; então, o último ano bem registrado é 2011.

Ao longo do processo de tabulação, manteve-se o foco em lidar com a observabilidade parcial da evasão. Isso é um pouco diferente do foco de Oliveira e Soares (2012), uma vez que esses autores buscaram mitigar os problemas de atrito, a fim de lidar com eventuais consequências para as regressões em painel que iriam operacionalizar, tendo como variável dependente exclusivamente os indicadores de reprovação.

Assim, no caso de Santa Catarina, primeiro limitou-se a análise nas coortes que iniciaram o ensino médio em 2008 e 2009, cada uma com aproximadamente 40 mil alunos após aplicar processos de deduplicação semelhantes aos descritos em Inep (2017), uma vez que muitos alunos tinham registros contemporâneos em mais de uma escola. A razão principal do foco nas coortes é que tais estudantes podem ser rastreados por pelo menos três anos subsequentes – até 2011, último ano da base completa –, e então eles podem ser classificados a fim de separar a evasão do atrito – o que é descrito a seguir. Além disso, a restrição nas coortes também se justifica para balizar características contemporâneas importantes e potencialmente não observadas, como a situação do mercado de trabalho local (Souza *et al.*, 2012).

Na sequência, rotularam-se os seguintes casos para os estudantes:

- *permanência*, quando é visto novamente em $t + 1$ na 1ª ou na 2ª série em qualquer escola (o que ocorreu em quase 70% das observações);
- *evasão*, quando não é visto novamente em $t + 1$ e $t + 2$, e também $t + 3$ no caso da coorte de 2008 (perto de 25% do total);
- *atrito*, quando não é visto novamente em $t + 1$, mas aparece em $t + 2$ na 1ª, na 2ª ou na 3ª série em qualquer escola (cerca de 5% do total); e
- *inconsistência*, quando se observa qualquer outro caso (menos de 1% do total).

O primeiro caso é intuitivo e não carece de maiores explicações. O segundo e o terceiro são idênticos aos usados por Oliveira e Soares (2012), com o objetivo de classificar os tipos de matrículas descontinuadas. O quarto caso mostra que situações sem lógica também estão registradas no sistema; por exemplo, quando o aluno é observado em t na 1ª série, em $t + 1$ na 3ª série e em $t + 2$ na 2ª série. Assim, as observações dessa última categoria foram desconsideradas.

Nessa conceituação de evasão, ainda pode haver algum erro de medida – *i.e.*, casos de estudantes que não deixaram a escola, mas foram classificados como evadidos, embora sejam inidentificáveis com a informação disponível. Todavia, ao lembrar que são alunos da rede pública no primeiro ano da etapa de ensino, parece razoável imaginar que esses eventuais casos mal classificados representam uma fração pequena do total, porque poucos estudantes devem ter ido para a rede particular, mudado para outro estado etc. Dessa forma, mesmo se o eventual erro de medida na evasão – como variável dependente – esteja correlacionado com alguma covariável a ser utilizada nas regressões, toma-se por hipótese que seriam poucos casos e que não devem gerar significativos vieses analíticos.

Também se notou conjuntos de escolas que deixavam de preencher muitos campos relevantes do sistema, inclusive se o aluno havia reprovado. Nesse sentido, percebeu-se que isso era recorrente em alguns municípios, e a explicação estaria na falta de treinamento dos chefes de expediente das secretarias, além da falta de orientações nos anos iniciais de uso do sistema. Assim, os estudantes que começaram o ensino médio nessas escolas foram desconsiderados, porque suas informações seriam inutilizadas como variáveis dependentes e/ou explicativas das condições de reprovação/evasão.

Ao cabo, 30.871 e 30.853 alunos das coortes de 2008 e 2009 são observados, respectivamente, e optou-se por trabalhar com os *cross-sections* da 1ª série por três motivos:

- porque se espera maiores probabilidades de evasão na primeira experiência na nova etapa de ensino; então, é um período que merece maior atenção por parte do pesquisador (Shirasu e Arraes, 2015; Inep, 2017);
- porque um estudante da coorte de 2009 que fosse observado no ano escolar de 2010 não poderia ser rastreado com segurança até 2012, a fim de rotulá-lo como evadido/atritado; e
- porque trabalhar com esses *cross-sections* deixarão mais clara a ideia de simultaneidade a ser discutida adiante.

Dessa forma, a tabela 1 apresenta os números de reprovações e evasões observadas por coorte. Para tal, definem-se *dummies* $R = 0$ para aprovação e $R = 1$ para reprovação – por baixa proficiência ou excesso de faltas –, conforme a situação indicada no Sisgesc; e $E = 1$ para evasão e $E = 0$ para juntar as definições de *permanência* e *atrito* em classificação única de não evasão. Assim, nesse recorte, as taxas de reprovação (razão entre os reprovados e o total de casos) são de 22,9% e 23,6% para as coortes de 2008 e 2009, respectivamente. Por sua vez, as taxas de evasão (razão entre os evadidos e o total de casos) são de 26,5% e 22,1% para as coortes de 2008 e 2009, respectivamente.

TABELA 1
Números de reprovações e de evasões observadas por coorte (2008-2009)

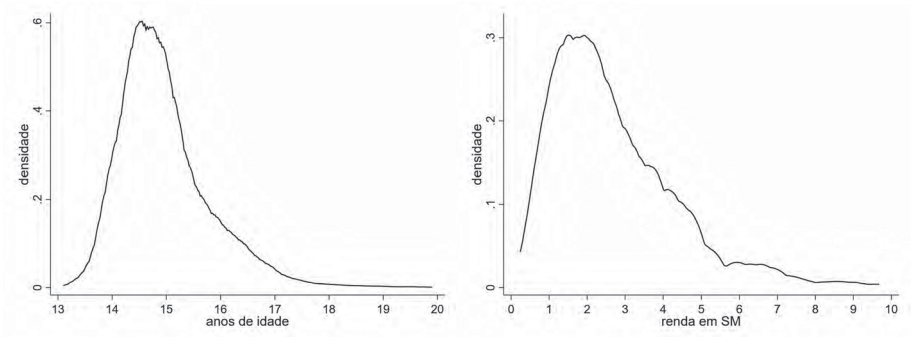
	Corte 2008		Coorte 2009	
	$R = 0$ [aprovação]	$R = 1$ [reprovação]	$R = 0$ [aprovação]	$R = 1$ [reprovação]
$E = 0$ [permanência]	19.352	3.333	20.531	3.512
$E = 1$ [evasão]	4.449	3.737	3.038	3.772

Fonte: Sisgesc.

Com os números da tabela 1, também se pode aplicar a regra de Bayes para estimar as probabilidades de evasão em face da reprovação – $P(E = 1 \mid R = 1)$ –, tal que os valores são da ordem de 52,9% e 51,8% para as coortes de 2008 e 2009, respectivamente. Da mesma forma, as probabilidades de evasão dada uma aprovação – $P(E = 1 \mid R = 0)$ – são da ordem de 18,7% e 12,9%. Tirando-se a diferença entre essas probabilidades, tem-se a noção de que o efeito marginal da reprovação nas chances de evasão seria da ordem de 34,2 p.p. e 38,9 p.p. para as coortes de 2008 e 2009, respectivamente. Complementarmente, ao considerar ambas as coortes, esse número será de 36,51 p.p. – isso servirá de *benchmark* para análises posteriores.

Quanto a covariáveis observadas na base de dados, nota-se que a distorção idade-série e a renda familiar têm correlações bem documentadas na literatura, tanto com a reprovação quanto com a evasão (Roderick, 1994; Eckstein e Wolpin, 1999; Rumberger e Lim, 2008; Chowdry, Crawford e Goodman, 2011; De Witte *et al.*, 2013). Assim, dado que no sistema há registro de ambas, o gráfico 1 apresenta histogramas alisados – *i.e.*, densidades por *kernel* – para a idade dos alunos – na data da matrícula – e a renda familiar – em salários mínimos (SMs) mensais.

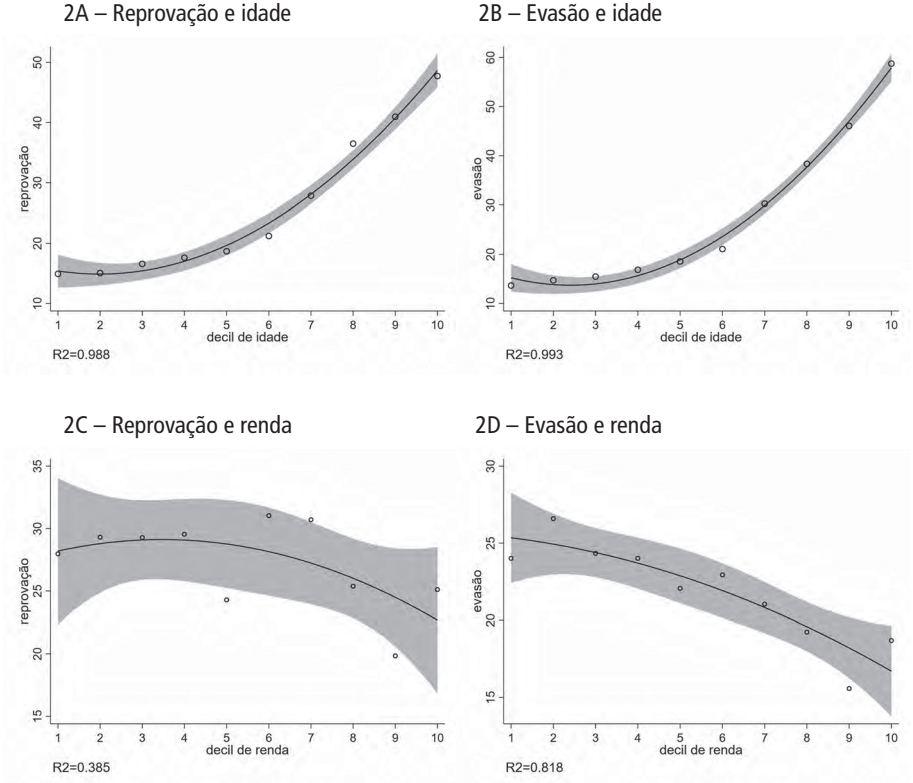
GRÁFICO 1
Histogramas alisados da idade e da renda familiar dos estudantes
1A – Idade 1B – Renda



Fonte: Sisgesc.
Obs.: Gráficos cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Como todos os estudantes analisados estão na primeira experiência no ensino médio, a idade e a distorção idade-série representam a mesma coisa: quanto maior, mais provavelmente houve alguma repetência no ensino fundamental. Assim, nota-se no gráfico 1 (1A) que a maioria tem idade próxima de 15 anos, o que é esperado para quem nunca reprovou antes. O gráfico 1 (1B) também apresenta o histograma alisado da renda, que foi perguntada aos pais ou responsáveis no ato da matrícula do aluno – em que pese se tem esse registro disso em 8.735 casos nas duas coortes (14% do total). Nota-se então que as respostas mais frequentes ficam perto de 2 SMs, o que coincide com o rendimento mensal médio de apenas um trabalhador catarinense observado no Censo Demográfico 2010. Isso sugere que a pergunta pode ter sido mal formulada, ou mal interpretada em alguns casos, e que muitas pessoas podem ter revelado sua renda individual, em vez da familiar. Portanto, embora seja animador a observação desse indicador, isso deve ser analisado com ressalvas, não apenas porque a renda foi apresentada por apenas parte da amostra, como também em razão de que isso pode não ter sido perguntado ou interpretado de forma clara.

GRÁFICO 2
Correlações das taxas de reprovação/evasão com idade/renda

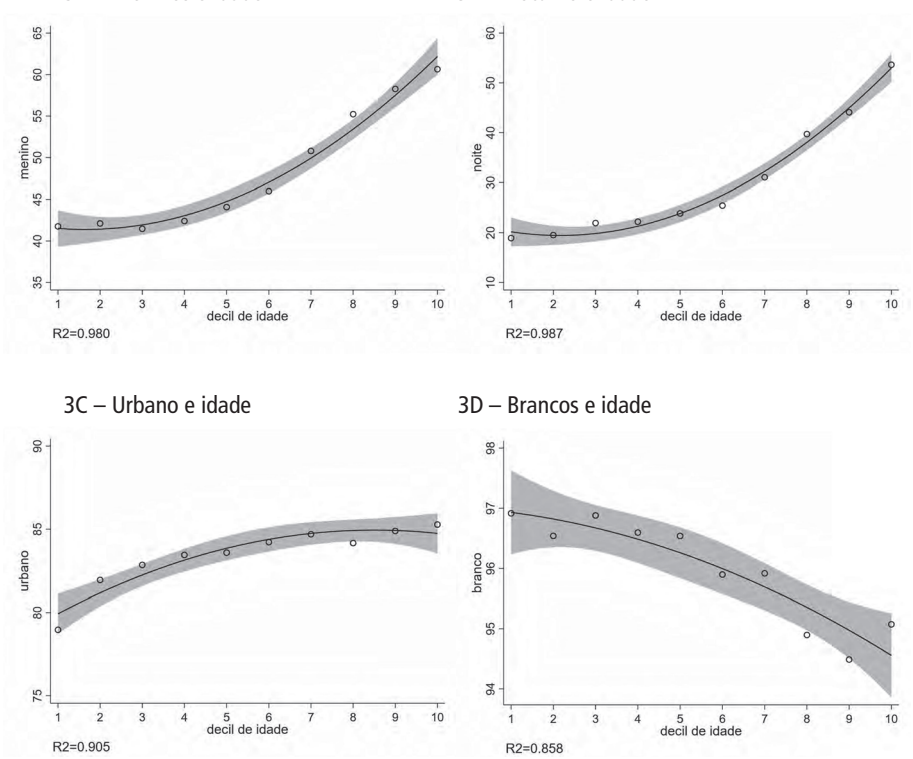


Fonte: Sisgesc.
Obs.: Gráficos cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Na sequência, o gráfico 2 busca mapear correlações da reprovação e da evasão com a idade e a renda, em que os gráficos foram construídos da seguinte maneira: faz-se a dispersão entre os decis de idade e de renda com e as taxas de reprovação e de evasão por decil, juntando as coortes; concomitantemente, plota-se uma regressão quadrática com um intervalo de 95% de confiança para o ajustamento (área acinzentada) – tal que o R^2 está abaixo de cada imagem. Dessa forma, o gráfico 2 (2A e 2B) indica que as taxas de reprovação de evasão aumentam com a distorção idade-série, e ambas ficam em torno de 50% a partir do nono decil (16,5 anos de idade). O gráfico 2 (2C) indica que não há correlação bem definida entre as taxas de reprovação e os decis de renda, mas também indica (2D) que existe algum nível de correlação inversa entre as taxas de evasão e os decis de rendas.

Nessa linha de análise, os próximos gráficos ilustram que outras quatro variáveis na base de dados apontam correlações bem definidas com a idade. O gráfico 3 (3A, 3B e 3C) mostra que os meninos, os estudantes do turno noturno e os residentes em áreas urbanas tendem a estar mais frequentemente em distorção idade-série – dado que suas proporções são maiores nos maiores decis de idade. O gráfico 3 (3D) mostra que *não brancos* estão mais frequentemente em distorção idade-série – ao frisar que a raça foi declarada pelos pais ou responsáveis no ato da matrícula.

GRÁFICO 3
Correlações das proporções de meninos, dos estudantes noturnos, dos residentes em áreas urbanas e da raça com os decis de idade

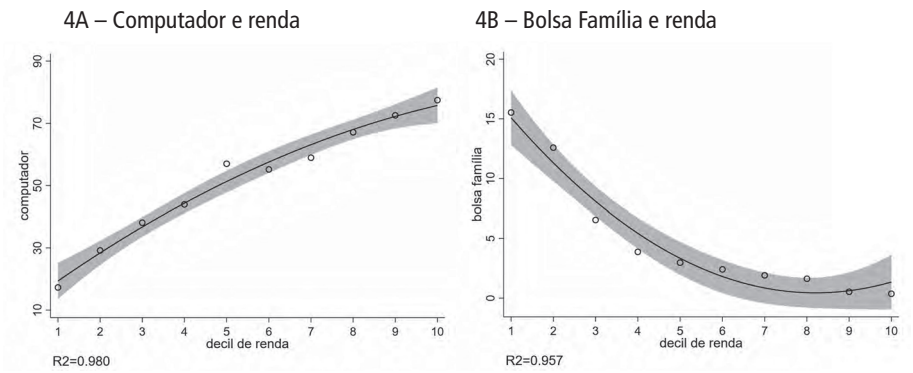


Fonte: Sisgesc.
Obs.: Gráficos cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Adiante, usando a mesma estrutura das imagens anteriores, o gráfico 4 ilustra que potenciais *proxies* para a renda familiar são os indicadores da presença de tecnologias em casa e do recebimento de auxílios governamentais. No gráfico 4 (4A), mostra-se que menos de 20% dos estudantes da faixa de renda mais baixa tinham um computador em casa – isso foi perguntado diretamente aos alunos em uma

enquete realizada ao longo do ano escolar; por sua vez, em torno de 80% daqueles de rendas mais altas tinham computador em casa. O gráfico 4 (4B) resulta de um cruzamento de dados com o Cadastro Único para Programas Sociais (CadÚnico), em que foi possível identificar quais famílias também eram beneficiárias do Bolsa Família – nota-se que na faixa de renda mais baixa esse número é perto de 15%, enquanto nas faixas de renda mais altas é difícil encontrar famílias beneficiárias do programa. Portanto, pelas relações aqui apresentadas, evidencia-se que há pertinência ao se usar como *proxy* da renda as indicações de ter computador em casa e ser beneficiário de algum programa de auxílio do governo.

GRÁFICO 4
Correlações das proporções de alunos com computador em casa e benefícios do Bolsa Família com os decis de renda



Fonte: Sisgesc e CadÚnico.
Obs.: Gráficos cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Além das características dos alunos descritas anteriormente, na base também é possível tabular algumas peculiaridades das escolas – em especial, se, concomitantemente ao ensino médio, há os ensinos infantil (notado em 82% das unidades de ensino) e fundamental (92%), e se há biblioteca (79%), laboratórios de ciência (25%) e de informática (89%) e uma quadra de esportes coberta (16%). Infelizmente, para os anos iniciais do sistema poucas informações dos professores puderam ser recuperadas. Dessa forma, a seguir, discute-se a modelagem das probabilidades de evasão e de reprovação em face das covariáveis observadas.

4 METODOLOGIA

A abordagem econométrica dessa literatura costuma usar os indicadores R e E como definido anteriormente, para a reprovação e a evasão, respectivamente. Em seguida, de forma padrão, definem-se variáveis latentes $R^*, E^* \in \mathbb{R}$, tais que $R = 1$ e $E = 1$ ocorrem quando acontece $R^* > 0$ e $E^* > 0$, respectivamente. Em sequência, as variáveis latentes são decompostas de forma semelhante ao que segue:

$$\begin{cases} R^* &= \alpha_{11} + \alpha_{12}X &+ \underbrace{\beta_1\tilde{X} + \varepsilon_1}_{u_1} \\ E^* &= \alpha_{21} + \alpha_{22}X + \alpha_{23}R &+ \underbrace{\beta_2\tilde{X} + \varepsilon_2}_{u_2} \end{cases}, \quad (1)$$

em que: α's e β's são parâmetros; X e $\tilde{X}$ representam, respectivamente, covariáveis observadas – *e.g.*, a distorção idade-série ou a renda familiar – e não observadas – ou seja, a necessidade de trabalhar, o envolvimento com drogas, a gravidez precoce ou o *bullying*; ε_j, para $j = 1$ ou 2, significa ruídos independentes entre si e as covariáveis, com média zero e variância finita; $u_j = \beta_j\tilde{X} + \varepsilon_j$, para $j = 1$ ou 2, representa termos de erro; e a presença de R na segunda equação decorre da ideia de que a reprovação precederia a decisão de evasão.

Na prática, trabalha-se com mais covariáveis que no sistema (1), mas o apresentado é suficiente para mostrar como tais formulações são usadas para estimar os efeitos da reprovação sobre a evasão, assim como os efeitos marginais em relação a determinadas covariáveis observadas (X). Nesse sentido, uma alternativa é assumir a independência entre u_1 e u_2, o que implica considerar que $\beta_1 = 0$ ou $\beta_2 = 0$, porque $\tilde{X}$ está em ambos os termos de erro. Ou seja, sob a hipótese de independência entre u_1 e u_2, assume-se automaticamente que as covariáveis não observadas que afetam a reprovação não são as mesmas que afetam a evasão.

Assumida essa hipótese, uma estratégia é considerar que u_1 e u_2 seguem distribuições normais, o que permite a estimação de modelos *probits* separados. Dessa forma, escrevem-se as probabilidades de reprovação e evasão como $P(R = 1 \mid X) = \Phi(\alpha_{11} + \alpha_{12}X)$ e $P(E = 1 \mid X, R) = \Phi(\alpha_{21} + \alpha_{22}X + \alpha_{23}R)$, respectivamente, em que Φ é a função normal cumulativa padronizada. Por exemplo, Leon e Menezes-Filho (2002), Souza *et al.* (2012) e Shirasu e Arraes (2015) seguem essencialmente essa estratégia, usando nuances particulares em cada caso, o que é comum na literatura internacional, pelo que se observa nos trabalhos citados nas revisões de Rumberger e Lim (2008) e De Witte *et al.* (2013).

Todavia, a hipótese de independência entre u_1 e u_2 – e, consequentemente, entre R e E – pode ser irrealista, porque tantas outras situações que não são observadas pelo pesquisador – representadas por $\tilde{X}$ – podem afetar ao mesmo tempo a reprovação e a evasão. Além disso, essa hipótese pode ser relaxada sem grandes dificuldades, o que permite a formulação de modelo que assume que as covariáveis não observadas podem afetar concomitantemente R e E. Para tal, basta considerar que $\alpha_{23} = 0$ e que u_1 e u_2 são correlacionados – por meio de $\tilde{X}$ – em algum nível ρ, o que implica um modelo *seemingly unrelated regressions* (SUR), tal que a decisão de evasão poderia ser vista como contemporânea ao recebimento da notícia da reprovação – que é o ponto discutido por Holm e Jaeger (2011).

Para operacionalizar modelos SUR dessa natureza, e computar efeitos marginais a partir de sistemas como (1), é possível usar os procedimentos detalhados – por exemplo, em Christofides, Stengos e Swidinsky (1997), Mullahy (2017) ou Greene (2017). Precisamente, assume-se que u_1 e u_2 seguem distribuição normal bivariada, e estimam-se os parâmetros e a correlação entre os termos de erro com um *biprobit*.

A fim de ilustrar como os resultados estimados para os efeitos marginais podem mudar com a hipótese sobre a independência entre u_1 e u_2, a tabela 2 compara alguns casos para os modelos *probit*/*biprobit* nos termos descritos anteriormente. Em todos, por simplicidade, assume-se que a covariável observada (X) é binária – Christofides, Stengos e Swidinsky (1997) mostram outros casos.

A começar pela variação da probabilidade estimada de reprovação em face de X – *e.g.*, estar ou não em distorção idade-série –, $\hat{P}(R = 1 \mid X = 1) - \hat{P}(R = 1 \mid X = 0)$, a correlação entre u_1 e u_2 não importa para a fórmula de cálculo, porque, para tal, no *biprobit* faz-se uso apenas da distribuição marginal. Todavia, é importante notar que, embora a estrutura da fórmula não mude, os valores estimados mudam, porque os estimadores do *probit* e do *biprobit* são diferentes – isso é explicitado pelos acentos circunflexos sobrescritos nos parâmetros da tabela 2.

As diferenças mais evidentes aparecem nos efeitos sobre a evasão, que, por sua vez, se dividem entre efeitos de covariável na probabilidade de evasão e de reprovação na probabilidade de evasão. O primeiro pode ser escrito como $\hat{P}(E = 1 \mid X = 1, \bar{R}) - \hat{P}(E = 1 \mid X = 0, \bar{R})$ no caso do *probit*, e leva em consideração o efeito condicionado na situação média de reprovação ($\bar{R}$). Por sua vez, esse efeito é descrito no *biprobit* simplesmente por $\hat{P}(E = 1 \mid X = 1) - \hat{P}(E = 1 \mid X = 0)$, porque se considera apenas a margem da distribuição normal bivariada.

Na última linha da tabela 2, mostra-se o outro efeito:

$$\hat{P}(E = 1 \mid \bar{X}, R = 1) - \hat{P}(E = 1 \mid \bar{X}, R = 0),$$

em que aparece a diferença mais significativa entre modelar ou não a simultaneidade entre reprovação e evasão, uma vez que entra em cena o coeficiente de correlação entre os erros (ρ). Assim, na tabela, explicita-se que, entre o *probit* e o *biprobit*, esses efeitos são computados de formas diferentes, usando-se estimadores diferentes, e então estes podem apresentar valores bem diferentes.

TABELA 2
Alguns efeitos marginais para modelos *probit/biprobit*, com e sem a hipótese simplificadora sobre a independência entre u_1 e u_2

Efeito	Descrição	Hipótese: $\rho = 0_{\text{e}}\alpha_{23} \neq 0$ (*probit*)	Hipótese: $\rho \neq 0_{\text{e}}\alpha_{23} = 0$ (*biprobit*)
Covariável na probabilidade de reprovação	$\hat{P}(R=1 \mid X=1) - \hat{P}(R=1 \mid X=0)$	$\Phi(\hat{\alpha}_{11}+\hat{\alpha}_{12}) - \Phi(\hat{\alpha}_{11})$	$\Phi(\breve{\alpha}_{11}+\breve{\alpha}_{12}) - \Phi(\breve{\alpha}_{11})$
Covariável na probabilidade de evasão	$\hat{P}(E=1 \mid X=1, \bar{R}) - \hat{P}(E=1 \mid X=0, \bar{R})$	$\Phi(\hat{\alpha}_{21}+\hat{\alpha}_{22}+\hat{\alpha}_{23}\bar{R}) - \Phi(\hat{\alpha}_{21}+\hat{\alpha}_{23}\bar{R})$	∄
	$\hat{P}(E=1 \mid X=1) - \hat{P}(E=1 \mid X=0)$	∄	$\Phi(\breve{\alpha}_{21}+\breve{\alpha}_{22}) - \Phi(\breve{\alpha}_{21})$
Reprovação na probabilidade de evasão	$\hat{P}(E=1 \mid \bar{X}, R=1) - \hat{P}(E=1 \mid \bar{X}, R=0)$	$\Phi(\hat{\alpha}_{21}+\hat{\alpha}_{22}\bar{X}+\hat{\alpha}_{23}) - \Phi(\hat{\alpha}_{21}+\hat{\alpha}_{22}\bar{X})$	$\Phi\left(\frac{\breve{\alpha}_{11}+\breve{\alpha}_{12}\bar{X}-\breve{\rho}(\breve{\alpha}_{21}+\breve{\alpha}_{22}\bar{X})}{\sqrt{1-\breve{\rho}^2}}\right) - \Phi\left(\frac{\breve{\alpha}_{21}+\breve{\alpha}_{22}\bar{X}-\breve{\rho}(\breve{\alpha}_{11}+\breve{\alpha}_{12}\bar{X})}{\sqrt{1-\breve{\rho}^2}}\right)$

Fonte: Christofides, Stengos e Swidinsky (1997); Mullahy (2017); Greene (2017).
Elaboração dos autores.
Obs.: O acento circunflexo sobrescrito nos parâmetros indica que se trata de valor estimativo para o respectivo modelo, *probit* ou *biprobit*; no caso de *biprobit*, inverteu-se o acento simplesmente para explicitar essa diferença; e a barra sobrescrita significa que se trata de um valor médio.

Para ilustrar numericamente essas diferenças, podem-se usar os números apresentados anteriormente na tabela 1, tal que $X = 0$ representaria estar na coorte de 2008 e $X = 1$, na de 2009. Dessa forma, ocorre que $X = 0{,}4999$ – resultando de [30.853]/[30.871 + 30.853] – e $\overline{R} = 0{,}2326$ – que advém de [3.333 + 3.737 + 3.512 + 3.772]/[30.871 + 30.853]. Tabulando-se apropriadamente a informação, os seguintes valores serão estimados para os parâmetros do *probit*: $\hat{\alpha}_{11} = -0{,}7421$; $\hat{\alpha}_{12} = 0{,}0231$; $\hat{\alpha}_{21} = -0{,}9195$; $\hat{\alpha}_{22} = -0{,}1759$ e $\hat{\alpha}_{23} = 1{,}0669$. Da mesma forma, os valores do *biprobit* serão: $\breve{\alpha}_{11} = -0{,}7385$; $\breve{\alpha}_{12} = 0{,}0164$; $\breve{\alpha}_{21} = -0{,}6250$; $\breve{\alpha}_{22} =$ -01484 e $\breve{\rho} = 0{,}5753$.

Dados os valores citados anteriormente, a estimativa do efeito marginal da covariável X na probabilidade de reprovação é 0,71 p.p. pelo *probit* e 0,51 p.p. pelo *biprobit*. Por sua vez, o efeito equivalente na chance de evasão será de -5,26 p.p. pelo *probit* e de -3,79 p.p. pelo *biprobit*. Por fim, as estimativas do efeito marginal da reprovação na probabilidade de evasão serão de 36,69 p.p. e 34,57 p.p., nos modelos *probit* e *biprobit*, respectivamente. Lembrando que o *benchmark* da regra de Bayes para esse valor é de 36,51 p.p.

5 RESULTADOS ECONOMÉTRICOS

Nesta seção, segue-se discutindo os resultados estimados para Santa Catarina, mas com o propósito de verificar o quão diferente o efeito da reprovação na evasão pode ser em termos de: i) modelagem – *i.e.*, *probit versus biprobit*; e ii) características individuais. O primeiro ponto busca ilustrar a relevância de considerar que as covariáveis não observadas que afetam a reprovação podem ser as mesmas que afetam a evasão. O segundo ponto procura mostrar que não se trata de efeito único, mas que os efeitos mudam em relação ao tipo de estudante.

Assim, nos exercícios, acrescentam-se as seguintes covariáveis observadas: *idade* – ou distorção idade-série –, nos termos descritos anteriormente; *renda* em SMs, quando observada; *menino*, como uma *dummy* 1 no caso de menino; *noturno*, como uma *dummy* 1, no caso de estar no turno da noite; *urbano*, como uma *dummy* 1, se reside em zona urbana; *branco*, como uma *dummy* 1 no caso de autodeclarado branco; *computador*, como uma *dummy* 1 no caso de ter computador em casa; *bolsa*, como uma *dummy* 1, se foi identificado o recebimento de Bolsa Família; *infantil*, *fundamental*, *biblioteca*, *ciências*, *informática* e *quadra* como *dummies* 1, no caso de presença dessas características na escola do estudante, conforme descrito anteriormente; e *2009* como uma *dummy* 1, para casos da coorte de 2009.

Na tabela A.1, no apêndice A, apresentam-se os parâmetros estimados para as especificações *probit*, em que as colunas (1) e (2) têm como variáveis dependentes as condições de reprovação e evasão, respectivamente, considerando todas as 61.724 observações das duas coortes. Por sua vez, nas especificações das colunas (3) e (4),

acrescenta-se como variável explicativa a renda familiar, ao custo de restringir a análise para os 8.735 casos em que isso é observado.

Os parâmetros estimados – e estatisticamente significantes – em 0,923 e 0,757 para a reprovação nas especificações (2) e (4) apontam na direção que já se evidenciou anteriormente: a reprovação aumenta a probabilidade de evasão. Nessa linha, os números apresentados na sequência corroboram que, ao aumentar a distorção idade-série, aumentam tanto as chances de reprovação quanto as de evasão; e a renda não apresenta interferência estatisticamente significante na reprovação, mas alunos de rendas mais altas tendem a ter menores chances de evasão.

Na sequência, nota-se que os meninos, os que estudam à noite e os que moram em áreas urbanas, apresentam maiores chances tanto de reprovação quanto de evasão; e, ao contrário, os alunos brancos têm ambas as chances reduzidas. O que, por sua vez, já foi indicado na análise descritiva dos dados, ao evidenciar-se que essas características estão correlacionadas com a distorção idade-série. Complementarmente, quem tem um computador em casa (*proxy* de renda familiar mais alta) tende a ter menores chances de evasão; e a indicação do Bolsa Família não se mostrou estatisticamente significante em nenhum dos exercícios.

Na tabela A.2, também no apêndice, apresentam-se os parâmetros estimados para as especificações *biprobit*, em que os blocos (5) e (6) têm como variáveis dependentes as condições concomitantes de reprovação e de evasão considerando todas as 61.724 observações das duas coortes e os 8.735 casos em que a renda é observada, respectivamente. A estrutura do vetor de variáveis explicativas dos *biprobits* é a mesma dos *probits* – em que pese que os coeficientes de correlação dos erros (ρ) foram estimados em 0,507 e 0,428 para as especificações (5) e (6), respectivamente.

Notadamente, tanto nos resultados dos *probits* quanto nos dos *biprobits*, todos os sinais dos parâmetros estimados em relação ao vetor de covariáveis (idade, renda, menino etc.) apontam na mesma direção. Inclusive para as covariáveis das escolas, em que apenas as indicações da presença do ensino fundamental e da quadra de esportes se mostraram estatisticamente significantes para reduzir as chances de reprovação/evasão. Sendo que uma explicação para esse efeito seria que escolas com essas características devem ter mais – e/ou melhor – infraestrutura, o que tenderia a causar um efeito benéfico para seus alunos.

Grosso modo, esses resultados não são novidade em relação ao que já foi apresentado aqui e na literatura. De fato, apenas se está controlando a influência das covariáveis, a fim de se computar efeitos marginais da probabilidade de evasão em face da reprovação. Nesse sentido, entre as variáveis explicativas que poderiam ser estudadas com mais detalhe, a fim de ilustrar os supracitados pontos i) e ii), foca-se na idade e na renda. A primeira porque, além de representar a distorção

idade-série no recorte trabalhado, também está correlacionada com as características de gênero, o turno de aula, o local da residência e a raça; e a segunda em razão de o referencial de *background* familiar é o que se dispõe, uma vez que rendas maiores devem refletir pais mais escolarizados e elementos correlatos.

Dessa forma, a tabela 3 apresenta os efeitos estimados entre decis de idade e de renda, computando-se $\hat{P}(E=1 \mid \bar{X}, R=1) - \hat{P}(E=1 \mid \bar{X}, R=0)$ da seguinte maneira: no caso da idade, usam-se os parâmetros estimados das colunas (2) da tabela A.1 e (5) da tabela A.2 para os modelos *probit* e *biprobit*, respectivamente; no caso da renda, empregam-se os parâmetros estimados das colunas (4) da tabela A.1 e (6) da tabela A.2 para os modelos *probit* e *biprobit*, respectivamente; em cada um desses casos, consideraram-se as observações do respectivo decil de idade/renda, aplicando-se a média das demais covariáveis no respectivo decil – o que é equivalente ao computo de $\bar{X}$ no exercício numérico da seção anterior.

TABELA 3
Efeitos estimados da reprovação na probabilidade de evasão, entre decis de idade e de renda, por tipo de modelo aplicado
(Em p.p.)

Decil	Idade		Renda	
	Probit	Biprobit	Probit	Biprobit
1	22,43	22,14	30,82	31,84
	(1,25)	(2,14)	(1,42)	(2,63)
2	25,59	25,36	30,80	31,70
	(1,31)	(2,23)	(1,42)	(2,67)
3	27,38	27,47	30,45	29,43
	(1,34)	(2,29)	(1,36)	(2,38)
4	28,77	29,20	30,82	30,32
	(1,36)	(2,31)	(1,38)	(2,59)
5	30,13	30,95	29,12	29,96
	(1,36)	(2,29)	(1,32)	(2,43)
6	31,58	32,78	30,72	31,79
	(1,37)	(2,32)	(1,48)	(3,35)
7	33,31	35,76	29,59	26,85
	(1,36)	(2,33)	(1,35)	(2,65)
8	34,67	38,66	29,50	27,10
	(1,34)	(2,22)	(1,37)	(2,69)
9	35,44	41,72	27,87	23,42
	(1,30)	(2,25)	(1,30)	(2,55)
10	35,85	45,88	29,67	24,85
	(1,21)	(2,00)	(1,42)	(2,99)

Elaboração dos autores.
Obs.: Desvio-padrão entre parênteses, abaixo de cada estimativa, estimado pelo método delta.

Os resultados revelados anteriormente mostram que o efeito marginal da reprovação na evasão varia tanto em função da modelagem quanto das características do aluno. Por exemplo, no décimo decil de idade, o valor é de 35,85 p.p. pelo *probit* e de 45,88 p.p. pelo *biprobit* – uma diferença de aproximadamente 10 p.p. Por sua vez, na perspectiva da renda, o efeito estimado pelo *biprobit* nos primeiros decis é um pouco acima de 30 p.p. e fica em torno de 25 p.p. nos últimos decis.

Em suma, os valores mencionados anteriormente indicam que o efeito da reprovação pode mudar bastante, ao se considerar que as covariáveis não observadas que afetam a reprovação podem ser as mesmas que afetam a evasão. Além disso, controlando-se pelas muitas covariáveis observáveis, fica claro que a reprovação teria mais efeito nos alunos que estão em distorção em idade-série – e então já estiveram retidos em alguma série do ensino fundamental. Portanto, a reprovação na 1ª série do ensino médio pode ser simplesmente a *gota d'água* para deixar a escola, e os fatores que realmente causam a evasão já se teriam se manifestado no aluno antes disso.

6 CONCLUSÃO

Ao analisar um inexplorado banco de microdados do ensino médio catarinense, o artigo buscou fazer três contribuições para a literatura. Primeiro, mostrar que uma forma de mitigar erros de medida da evasão em dados administrativos é rastreando os alunos sistematicamente por anos sequenciais – a semelhança do que foi discutido em Oliveira e Soares (2012) e Inep (2017). Segundo, evidenciar que a posse de equipamentos de tecnologia pode ser uma boa *proxy* da renda familiar do aluno, e que isso estaria mais claramente correlacionado com a condição de evasão que com a de reprovação. Terceiro, ilustrar que a reprovação e a evasão podem ser estimadas simultaneamente de forma fácil, evitando-se uma hipótese irrealista de as covariáveis não observadas de uma não afetam a outra.

As estimativas encontradas nessa pesquisa são de que o efeito da reprovação na probabilidade de evasão, que controla as características do aluno e da escola, estaria próximo de 35 p.p. No entanto, esse efeito seria mais perto de 20 p.p. para estudantes que não estão em distorção idade-série, e poderia chegar a 45 p.p. para alunos que já reprovaram alguma vez antes. O que leva a acreditar que o principal elemento causador da evasão estaria presente no aluno antes da derradeira reprovação, e que isso estaria concomitantemente atrelado com a situação de reprovação e a decisão da evasão. Talvez, como indica a pesquisa de Neri (2015), simplesmente haja uma visão limitada dos resultados da educação sobre a renda futura e outros benefícios.

Por fim, acredita-se o mecanismo de causalidade da evasão poderá ser mais bem entendido com futuras análises de outras bases de dados, com mais ampla coleta de informações dos estudantes, e eventualmente se integrando a outros sistemas – *e.g.*, CadÚnico. Mas, é claro, isso depende de que as informações sejam mais amplamente disponibilizadas para a comunidade acadêmica.

REFERÊNCIAS

BRASIL. Ministério da Educação. **Censo Escolar da Educação Básica**: Sistema Educacenso. Brasília: MEC; FNDE, 2015.

CHOWDRY, H.; CRAWFORD, C.; GOODMAN, A. The role of attitudes and behaviours in explaining socio-economic differences in attainment at age 16. **Longitudinal and Life Course Studies**, v. 2, n. 1, p. 59-76, 2011.

CHRISTOFIDES, L. N.; STENGOS, T.; SWIDINSKY, R. On the calculation of marginal effects in the bivariate probit model. **Economics Letters**, v. 54, n. 3, p. 203-208, 1997.

DE WITTE, K. *et al*. A critical review of the literature on school dropout. **Educational Research Review**, v. 10, p. 13-28, Dec. 2013.

DURYEA, S. **Children's advancement through school in Brazil**: the role of transitory shocks to household income. Washington: IDB, 1998. (Working Paper, n. 376).

ECKSTEIN, Z.; WOLPIN, K. I. Why youths drop out of high school: the impact of preferences, opportunities, and abilities. **Econometrica**, v. 67, n. 6, p. 1295-1339, 1999.

FLETCHER, P.; RIBEIRO, S. O ensino de primeiro grau no Brasil de hoje. **Em Aberto**, v. 6, n. 33, p. 1-10, 1987.

_____. Modeling education system performance with demographic data: an introduction to the profluxo model. *In*: BARRETO, E. S. S.; ZIBAS, D. M. L. (Org.). **Brazilian issues on education, gender and race**. São Paulo: FCC, 1996.

GOLGHER, A. B.; RIOS-NETO, E. L. G. **Uma comparação entre os modelos profluxo e IPC quando aplicados aos dados do sistema educacional brasileiro**. Brasília: Inep, 2005. (Texto para Discussão, n. 16).

GREENE, W. H. (Ed.). **Econometric analysis**. 8th ed. London: Pearson, 2017.

HOLM, A.; JAEGER, M. M. Dealing with selection bias in educational transition models: the bivariate probit selection model. **Research in Social Stratification and Mobility**, v. 29, n. 3, p. 311-322, 2011.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Sigilo de informações constantes do Banco de Dados do Censo Escolar**. Brasília: Inep, 2009. (Nota Técnica Deed, n. 2).

_____. **Estimativas de fluxo escolar a partir do acompanhamento longitudinal dos registros de aluno do Censo Escolar do período 2007-2016**. Brasília: Inep, 2017. (Nota Técnica Deed, n. 8).

KLEIN, R. Produção e utilização de indicadores educacionais: metodologia de cálculo de indicadores do fluxo escolar da educação básica. **Revista Brasileira de Estudos Pedagógicos**, v. 84, n. 207/208/209, p. 107-157, 2003.

KLEIN, R.; RIBEIRO, S. C. O Censo Educacional e o modelo de fluxo: o problema da repetência. **Revista Brasileira de Estatística**, v. 1, n. 197-198, p. 5-45, 1991.

LEE, V. E. A necessidade dos dados longitudinais na identificação do efeito-escola. **Revista Brasileira de Estudos Pedagógicos**, v. 91, n. 229, p. 471-480, 2010.

LEON, F. L. L.; MENEZES-FILHO, N. A. Reprovação, avanço e evasão escolar no Brasil. **Pesquisa e Planejamento Econômico**, Brasília, v. 32, n. 3. p. 417-452, dez. 2002.

MENG, C.-L.; SCHMIDT, P. On the cost of partial observability in the bivariate probit model. **International Economic Review**, v. 26, n. 1, p. 71-85, 1985.

MULLAHY, J. Marginal effects in multivariate probit models. **Empirical Economics**, v. 52, n. 2, p. 447-461, 2017.

NERI, M. (Coord.). **Motivos da evasão escolar**. Rio de Janeiro: CPS; FGV, 2015.

OLIVEIRA, L. F. B.; SOARES, S. S. D. **Determinantes da repetência escolar no Brasil**: uma análise de painel dos censos escolares entre 2007 e 2010. Brasília: Ipea, fev. 2012. (Texto para Discussão, n. 1706).

POIRIER, D. J. Partial observability in bivariate probit models. **Journal of Econometrics**, v. 12, n. 2, p. 209-217, 1980.

_____. Identification in multivariate partial observability probit. **International Journal of Mathematical Modelling and Numerical Optimisation**, v. 5, n. 1-2, p. 45-63, 2014.

RIANI, J. L. R.; RIOS-NETO, E. L. G. Background familiar versus perfil escolar do município: qual possui maior impacto no resultado educacional dos alunos brasileiros? **Revista Brasileira de Estudos de População**, v. 25, n. 2, p. 251-269, 2008.

RIBAS, R. P.; SOARES, S. S. D. O atrito nas pesquisas longitudinais: o caso da Pesquisa Mensal de Emprego (PME/IBGE). **Estudos Econômicos**, v. 40, n. 1, p. 213-244, 2010.

RIBEIRO, S. C. A pedagogia da repetência. **Estudos Avançados**, v. 5, n. 12, p. 7-21, 1991.

RIOS-NETO, E. L. G.; RIANI, J. L. R. (Org.). **Introdução à demografia da educação**. Campinas: Abep, 2004.

RODERICK, M. Grade retention and school dropout: investigating the association. **American Educational Research Journal**, v. 31, n. 4, p. 729-759, 1994.

RUMBERGER, R. W.; LIM, S. A. **Why students drop out**: a review of 25 years of research. Santa Barbara: University of California, 2008. (California Dropout Research Project Report, n. 15).

SHIRASU, M. R.; ARRAES, R. A. Determinantes da evasão e repetência escolar no ensino médio do Ceará. **Revista Econômica do Nordeste**, v. 46, n. 4, p. 117-136, 2015.

SOUZA, A. P. *et al*. Fatores associados ao fluxo escolar no ingresso e ao longo do ensino médio no Brasil. **Pesquisa e Planejamento Econômico**, Brasília, v. 42, n. 1, p. 5-39, abr. 2012.

STEARNS, E. *et al*. Staying back and dropping out: the relationship between grade retention and school dropout. **Sociology of Education**, v. 80, n. 3, p. 210-240, 2007.

BIBLIOGRAFIA COMPLEMENTAR

HELSPER, E. J. A corresponding fields model for the links between social and digital exclusion. **Communication Theory**, v. 22, n. 4, p. 403-426, 2012.

APÊNDICE A

TABELA A.1
Parâmetros estimados para as especificações de *probit*

Covariável	Modelo por variável dependente			
	(1)	(2)	(3)	(4)
	Reprovação	Evasão	Reprovação	Evasão
Reprovação	-	0,923***	-	0,757***
		(0,035)		(0,108)
Idade	0,326***	0,363***	0,267***	0,324***
	(0,010)	(0,009)	(0,019)	(0,029)
Renda	-	-	-0,013	-0,015**
			(0,009)	(0,007)
Menino	0,286***	0,111***	0,223***	0,129***
	(0,024)	(0,014)	(0,045)	(0,032)
Noturno	0,110***	0,214***	0,132	0,281***
	(0,038)	(0,023)	(0,094)	(0,070)
Urbano	0,351***	0,182***	0,485***	0,101
	(0,050)	(0,031)	(0,103)	(0,099)
Branco	-0,207***	-0,0713**	-0,195**	-0,128**
	(0,044)	(0,035)	(0,075)	(0,064)
Computador	0,072	-0,061**	-0,103	-0,165***
	(0,054)	(0,025)	(0,081)	(0,042)
Bolsa	0,074	-0,021	0,115	-0,115
	(0,096)	(0,056)	(0,140)	(0,087)
Infantil	-0,111	0,043	-0,293	0,005
	(0,100)	(0,042)	(0,219)	(0,087)
Fundamental	-0,275*	-0,074	-0,926***	-0,210**
	(0,151)	(0,051)	(0,357)	(0,094)
Biblioteca	0,040	-0,048	0,209	-0,104
	(0,072)	(0,038)	(0,179)	(0,075)
Ciências	0,144	-0,026	0,226	0,046
	(0,125)	(0,032)	(0,243)	(0,090)
Informática	0,010	-0,032	-0,194	-0,190**
	(0,093)	(0,043)	(0,231)	(0,091)
Quadra	-0,063	-0,019	-0,479**	-0,061
	(0,102)	(0,031)	(0,208)	(0,095)
2009	-0,014	-0,218***	-0,286	-0,113
	(0,058)	(0,028)	(0,209)	(0,085)
Constante	-5,602***	-6,355***	-3,725***	-5,375***
	(0,244)	(0,159)	(0,456)	(0,481)

Elaboração dos autores.
Obs.: Desvio-padrão robusto entre parênteses, abaixo de cada estimativa: *** $p < 0,01$; ** $p < 0,05$; * $p < 0,10$.

TABELA A.2
Parâmetros estimados para as especificações de *biprobit*

Covariável	Modelo por variável dependente			
	(5) [$\rho = 0,507$]		(6) [$\rho = 0,428$]	
	Reprovação	Evasão	Reprovação	Evasão
Idade	0,323***	0,433***	0,267***	0,372***
	(0,010)	(0,009)	(0,019)	(0,028)
Renda	-	-	-0,013	-0,017**
			(0,009)	(0,008)
Menino	0,283***	0,095***	0,223***	0,176***
	(0,024)	(0,013)	(0,045)	(0,029)
Noturno	0,106***	0,230***	0,129	0,298***
	(0,039)	(0,024)	(0,095)	(0,067)
Urbano	0,351***	0,266***	0,477***	0,206**
	(0,050)	(0,032)	(0,100)	(0,096)
Branco	-0,208***	-0,130***	-0,191**	-0,170***
	(0,044)	(0,034)	(0,076)	(0,061)
Computador	0,076	-0,035	-0,098	-0,180***
	(0,054)	(0,021)	(0,061)	(0,044)
Bolsa	0,080	0,002	0,120	-0,077
	(0,097)	(0,046)	(0,140)	(0,083)
Infantil	-0,190	-0,019	-0,291	-0,072
	(0,150)	(0,040)	(0,218)	(0,088)
Fundamental	-0,274*	-0,155***	-0,929***	-0,437***
	(0,152)	(0,054)	(0,356)	(0,085)
Biblioteca	0,041	-0,032	0,219	-0,048
	(0,073)	(0,037)	(0,180)	(0,082)
Ciências	0,142	0,020	0,218	0,100
	(0,125)	(0,038)	(0,243)	(0,083)
Informática	0,012	-0,026	-0,194	-0,226**
	(0,093)	(0,052)	(0,231)	(0,112)
Quadra	-0,068	-0,037	-0,490**	-0,175**
	(0,102)	(0,040)	(0,210)	(0,088)
2009	-0,018	-0,204***	-0,289	-0,175***
	(0,059)	(0,021)	(0,211)	(0,066)
Constante	-5,557***	-7,126***	-3,729***	-5,632***
	(0,241)	(0,160)	(0,428)	(0,407)

Elaboração dos autores.
Obs.: Desvio-padrão robusto entre parênteses, abaixo de cada estimativa: *** $p < 0,01$; ** $p < 0,05$; * $p < 0,10$.

Data da submissão em: 14 mar. 2019.

Primeira decisão editorial em: 23 jan. 2020.

Última versão recebida em: 3 ago. 2020.

Aprovação final em: 3 nov. 2020.