Universidade de Brasília
IE - Departamento de Estatística
Trabalho de Conclusão de Curso

Estudo sobre o perfil socioeconômico dos cursos de graduação da Universidade de Brasília entre 2015 e 2017: uma Aplicação de Modelos Multiníveis para Dados Longitudinais

Bárbara Santiago Pedreira da Costa

Orientadora: Prof[a]. Maria Teresa Leão Costa

Brasília
17 de julho de 2018

Bárbara Santiago Pedreira da Costa

Estudo sobre o perfil socioeconômico dos cursos de graduação da Universidade de Brasília entre 2015 e 2017: uma Aplicação de Modelos Multiníveis para Dados Longitudinais

Orientadora: Prof[a]. Maria Teresa Leão Costa

Trabalho de conclusão de curso apresentado para obtenção do título de Bacharel em Estatística ao Departamento de Estatística da Universidade de Brasília.

Brasília

17 de julho de 2018

À minha amada mãe.

Aos meus ancestrais.

À todas as mulheres negras.

AGRADECIMENTOS

Quero agradecer, em especial, à minha mãe que não mediu esforços para me proporcionar o melhor do ensino público desde o início da minha vida escolar, onde tive acesso ao conteúdo formal e aprender a ser cidadã. Agradecer pelas noites não dormidas ao meu lado, pelas orações, pelos estímulos, pela estrutura, por toda a confiança no meu potencial e por não me deixar abaixar a cabeça nesse momento de conclusão de curso.

Agradeço às mulheres que foram muito importantes no meu desenvolvimento como ser humano, minhas madrinhas, minha avó Sandra, minhas tias maternas, minhas primas que me deram apoio no momento de conclusão do curso, minhas professoras da segunda e da oitava série, que me mostraram como a matemática pode ser encantadora.

Quero agradecer às minhas orientadoras, Maria Teresa e Ana Maria, que me mostraram o mundo maravilhoso da Estatística e me treinaram para encarar o mercado de trabalho com graça, leveza e muita competência. Agradecer, também, por confiarem no meu trabalho e por acreditarem nas minhas habilidades.

Também quero agradecer aos meus chefes no Ministério do Desenvolvimento Social e Combate à Fome, Jennifer, Frederico, Mariana, Gabriela e Diego que me mostraram que é possível ter um ambiente de trabalho muito produtivo, com leveza e companheirismo.

Agradecer ao Observatório da Vida Estudantil por proporcionar todas as minhas pesquisas e minhas produções acadêmicas.

Agradecer aos meus amigos que me ajudaram muito nos momentos de dificuldade e fizeram dos meus dias na academia, dias mais felizes. Além dos meus colegas de curso e de estágio que me desafiaram a lidar e superar as diferenças. Esse aprendizado é um dos mais importantes na vida profissional de uma pessoa.

Por fim, agradeço ao povo brasileiro por custear o ensino superior público e ao governo federal por proporcionar uma Universidade com mais oportunidades, professores qualificados, apoio às pesquisas e com um maior acesso aos diversos grupos de discentes, nos anos de 2003 a 2016.

'Não me julgue pelos meus sucessos,
julgue-me pelas vezes em que caí e levantei-me de novo'
Nelson Mandela

RESUMO

Em um contexto de expansão do acesso ao ensino superior em que, o Plano Nacional de Educação para 2011 a 2020 e o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni) estavam criando novas formas de ingresso, o Observatório da Vida Estudantil criou a "Pesquisa Perfil do Estudante – Etapa Registro" que visa entender o perfil dos alunos que estão ingressando na Universidade de Brasília. A partir dessa pesquisa, o presente estudo busca observar o perfil dos cursos de graduação da UnB com base nas características dos ingressantes. Estudos anteriores desenvolveram o Indicador de Status Socioeconômico do Estudantes que permitiu observar se as novas formas de ingresso estão atendendo toda a população em seus diversos extratos sociais. Uma análise de modelos multiníveis apontou uma redução do indicador ao longo do tempo. Alguns fatores associados ao comportamento do status socioeconômico médio dos cursos são o percentual de alunos que estão superando a escolaridade máxima dos pais e o percentual de ingressantes que cursaram o ensino médio total ou majoritariamente em escola pública.

Palavras Chave: Modelo Multinível; perfil socioeconômico; Observatório da Vida Estudantil; Dados Longitudinais.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Sumário

INTRODUÇÃO

Um dos direitos básicos do brasileiro é a educação. Em 2010, o governo federal propôs ao Congresso Nacional um projeto de lei com o Plano Nacional de Educação para 2011 a 2020 em que foram definidas metas e diretrizes para a educação brasileira. Número de vagas ofertadas e formas alternativas de ingresso nas universidades públicas fazem parte das inovações trazidas pelo PNE. Com essas mudanças, um maior número de pessoas passou a ser atendida pela educação superior.

Na Universidade de Brasília diversas iniciativas foram implementadas para atingir as metas estabelecidas durante este período. Novos *campi* foram construídos, novos cursos foram criados, o número de vagas dos cursos existentes foi ampliado dentro do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni). Assim como a adoção de novas formas de ingresso e do sistema de cotas sociais e raciais.

Em face a tantas transformações é importante avaliar a repercussão das medidas adotadas no perfil dos estudantes da Universidade de Brasília.

As características socioeconômicas dos ingressantes durante este período mudaram?

Em 2012, o Núcleo de Pesquisa sobre Ensino Superior em parceria com o Laboratório de População e Desenvolvimento (ambos ligados ao Centro de Estudos Avançados Multidisciplinares da UnB) começou a coletar dados dos calouros da UnB sob a coordenação do Observatório da Vida Estudantil. Esta coleta é feita semestralmente a partir do questionário da "Pesquisa Perfil do Estudantes da UnB – Etapa Registro" e obtém informações sobre características sociodemográficas, pretensões profissionais, vida pregressa entre outros aspectos dos novos estudantes.

A base de dados construída sobre o perfil sociodemográfico de ingressantes entre 2012 e 2017, tem um enorme potencial de informação que ainda não foi explorado nos estudos acadêmicos que já foram publicados até hoje ou nos que estão em andamento. Dentre as características não exploradas do banco, está o fato de os dados serem coletados repetidas vezes ao longo do tempo. Se considerarmos variáveis como raça/cor e observarmos o comportamento da quantidade de alunos negros que estão entrando na UnB, podemos buscar técnicas estatísticas mais avançadas para avaliar se essa quantidade está crescendo ao longo do tempo e, assim talvez avaliar a eficácia do sistema de cotas ou a necessidade de políticas públicas voltadas para esse público.

A Modelagem Multinível é um dos métodos que pode ser aplicado na análise de dados longitudinais, onde características são repetidas ao longo do tempo.

Esta modelagem permite avaliar se o comportamento de uma variável se altera ao longo do tempo e a influência de outras variáveis nesta variação.

Neste trabalho, o interesse está em olhar para esse banco de dados por um novo ângulo e descrever o perfil dos estudantes sob o ponto de vista dos cursos de graduação da UnB considerando os ingressantes do primeiro semestre dos anos de 2015, 2016 e 2017. Deseja-se verificar se o perfil socioeconômico dos estudantes dos diferentes cursos da UnB tem variado ao longo do tempo e quais variáveis podem explicar esta variação.

1 OBJETIVOS

1.1 Objetivo Geral

Descrever o perfil socioeconômico dos cursos de graduação da UnB no período entre 2015 e 2017.

1.2 Objetivos Específicos

- Verificar se existem variações nos perfis dos cursos segundo as características dos estudantes;
- Verificar se o perfil socioeconômico dos estudantes dos cursos mudou ao longo do tempo;
- Identificar características dos estudantes do curso que podem explicar as mudanças socioeconômicas;
- Aplicar modelos multiníveis nos dados longitudinais para analisar o perfil dos cursos de graduação.

2 REFERENCIAL TEÓRICO

Para compreender o modelo multinível para dados longitudinais e suas aplicações, é recomendável entender alguns conceitos estatísticos como o que são fatores fixos, medidas repetidas, modelos mistos e dados longitudinais.

Trata-se de medidas repetidas, a observação de um atributo de interesse em uma mesma unidade de informação, em mais de uma ocasião. Em geral, elas são usadas em estudos nos quais há um interesse em identificar a existência de um padrão dessas medidas de um indivíduo no tempo. Nesse caso, o atributo de interesse será observado em um mesmo indivíduo em mais de uma ocasião, gerando uma variabilidade individual decorrente de fatores não mensurados, inviabilizando a independência e a homocedasticidade dos dados.

Um caso particular de medidas repetidas ocorre quando as repetições são feitas, necessariamente, ao longo do tempo e são designados dados longitudinais. Então o atributo de interesse de um mesmo indivíduo será acompanhado durante um certo período. Espera-se que esse tipo de dados também possua a característica de falta de independência, principalmente entre as medidas mais próximas (consecutivas), causada em partes pelo efeito da memória.

Em estudos desse tipo, a resposta de cada indivíduo possui 3 componentes. São eles, um efeito fixo, um efeito aleatório e um erro que é devido à medição ou ao não registro de variáveis (Queiroz, 2012). Para analisar esse tipo de dados, existem três possíveis modelos: marginal, de transição e misto.

Antes de definir os tipos de modelo que podem ser usados para trabalhar com dados em coorte, é necessário definir alguns conceitos muito utilizados em experimentos, mas que também se aplicam aos estudos observacionais. São eles: fator, nível, efeito fixo e efeito aleatório.

Em experimentos, fator é a covariável em estudo e os níveis de um fator são os valores ou categorias dessa covariável. Se o nível do fator a ter seu efeito testado for definido pelo pesquisador, então esse efeito é chamado de fixo. Caso esse nível seja definido através de algum processo que não tenha a interferência do pesquisador, o efeito será denominado aleatório. A partir das definições elucidadas, os três modelos serão apresentados a seguir.

No modelo marginal, além do foco principal que é o estudo sobre a média populacional a partir de uma amostra, é possível modelar o valor esperado em função das variáveis explicativas e especificar um modelo de associação entre as observações de cada indivíduo. Efeitos aleatórios não são levados em consideração, excluindo assim a possibilidade de, a partir desse modelo, observar informações sobre a variação

entre os indivíduos.

Como não é necessário levar em consideração a diferença entre indivíduos, essa parte da informação é negligenciada. Portanto esse tipo de modelo não capta o histórico individual de medidas, configurando uma limitação caso essa trajetória seja algo importante no estudo. Além dessa primeira limitação, esse modelo exige que os dados sejam completos e que o intervalo entre as medidas seja igual para todos os indivíduos em virtude da construção da matriz de variâncias e covariâncias.

O modelo de transição é uma extensão do modelo linear generalizado e é usado geralmente nos casos em que a distância de tempo entre as medidas são iguais. Ao contrário do modelo marginal, esse modelo pode incluir efeitos aleatórios para considerar o efeito da variação individual entre os indivíduos.

Este modelo descreve cada resposta como uma função das respostas anteriores e das variáveis explicativas. Essa estrutura de modelo é a mesma das cadeias de Markov. Mas o modelo de transição não é o foco deste trabalho e para um estudo mais avançado sobre ele, recomenda-se Diggle et al. (2002).

Esse modelo possui uma estrutura composta por efeito fixo, efeito aleatório, erro em que a variabilidade entre os indivíduos reflete uma heterogeneidade natural devido a fatores não mensurados (Diggle et al., 1994). Dentre as vantagens desse tipo de modelo estão o fato de não exigir que os dados sejam completos e nem que tempo entre as observações seja igual e incluir fatores fixos e aleatórios permitindo uma maior flexibilidade na modelagem.

Uma particularidade do modelo misto é que ele considera dois tipos de componente. O componente individual é descrito por meio de um modelo com intercepto e inclinação populacionais enquanto o componente entre indivíduos considera um intercepto variável e inclinação individual, ou seja, inclinações e/ou interceptos podem ter diferentes valores para cada indivíduo (Queiroz, 2012).

Em algumas literaturas o modelo misto é chamado de modelo de efeitos aleatórios ou modelos hierárquicos. Essas nomenclaturas estão relacionadas a casos particulares. Como o foco deste trabalho são os modelos multiníveis, eles serão definidos a seguir.

2.1 Modelo de Regressão Multinível

O modelo de regressão multinível pode ser visto como um sistema hierárquico de modelos de regressão que é construído a partir de um banco de dados que apresentam uma estrutura hierárquica, observações agrupadas em diferentes níveis. Por exemplo, estudantes agrupados em turmas, que por sua vez estão agrupados em escolas, e assim por diante. Ou ainda, pacientes que são tratados por

médicos. Na sua forma mais simples, possui dois níveis sendo o mais baixo (nível 1) composto de observações agrupadas formando o nível mais alto (nível 2). É importante perceber que as observações dentro de um agrupamento (nível 2) podem estar correlacionadas, não atendendo assim ao pressuposto de independência dos modelos de regressão clássica, daí a necessidade de utilizar modelos multiníveis.

Considerando, por exemplo, um modelo de regressão com dois níveis e duas variáveis explicativas pertencentes ao nível 1, a equação de regressão será da seguinte forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j} \cdot X_{1ij} + \beta_{2j} \cdot X_{2ij} + e_{ij} \tag{1}$$

Nesta equação de regressão,

β_{0j} é o intercepto da j-ésima unidade do nível 2;

β_{1j} é o coeficiente de regressão (inclinação) da j-ésima unidade do nível 2;

β_{2j} é o coeficiente de regressão (inclinação) da j-ésima unidade do nível 2;

e_{ij} é o erro residual.

O índice j é para as unidades do nível 2 ($j = 1...J$) e o índice i é para unidades do nível 1 ($i = 1, ..., n_j$). A diferença com o modelo de regressão usual é que assumimos que cada unidade do nível 2 tem um intercepto diferente β_{0j} e diferentes coeficientes de inclinação β_{1j} e β_{2j}. Isto é indicado nas equações 1 e 2 anexando um índice j aos coeficientes de regressão. Os erros residuais e_{ij} são assumidos como tendo uma média de zero e uma variância desconhecida.

Uma vez que o intercepto e os coeficientes de inclinação são variáveis aleatórias que variam em todas as unidades do nível 2, eles são frequentemente chamados de coeficientes aleatórios.

Em toda unidade do nível 2, os coeficientes de regressão β_j são assumidos como tendo uma distribuição normal multivariada. O próximo passo no modelo de regressão hierárquica é explicar a variação dos coeficientes de regressão β_j que introduzindo variáveis explicativas ao nível 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01j} + u_{0j} \tag{2}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \cdot Z_j + u_{1j} \tag{3}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} \cdot Z_j + u_{2j} \tag{4}$$

O coeficiente β_{0j} na equação 2 pode ser interpretado como o valor médio de Y para a j-ésima unidade do nível 2. Já os coeficientes β_{1j} e β_{2j} mostram que a relação entre Y e as variáveis explicativas do nível 1 (X_1 e X_2) depende da variável explicativa do nível 2.

Os termos u_{0j}, u_{1j} e u_{2j} nas equações 2, 3 e 4 representam termos de erro residual (aleatórios) no nível 2. Assume-se que estes resíduos possuem uma média zero e são independentes dos resíduos e_{ij} no nível individual.

O modelo multinível pode ser escrito como uma única equação de regressão complexa ao substituir as equações 2, 3 e 4 na equação 1. Reorganizando os termos, tem-se:

$$\begin{aligned} Y_{ij} = \gamma_{00} + \gamma_{10} \cdot X_{1ij} + \gamma_{20} \cdot X_{2ij} + \gamma_{01} \cdot Z_j + \gamma_{11} \cdot X_{1ij} \cdot Z_j \\ + \gamma_{21} \cdot X_{2ij} \cdot Z_j + u_{1j} \cdot Z_{1ij} + u_{2j} \cdot X_{2ij} + u_{0j} + e_{ij} \end{aligned} \tag{5}$$

Espera-se que pressuposto de independência seja violado dado que as observações do mesmo grupo tendem a ser mais parecidas entre si e as observações de grupos diferentes tendem a ser distintas. A literatura fala que o coeficiente de correlação intraclasse possui várias fórmulas diferentes para estimar a correlação e pode ser usado para expressar o grau de dependência dos dados.

Se tivermos dados hierárquicos simples, o modelo de regressão multinível pode ser usado para produzir uma estimativa da correlação intraclasse. O modelo usado para este propósito é um modelo que não contém nenhuma variável explicativa, o chamado modelo nulo ou do intercepto. O modelo nulo é derivado das equações 1 e 3 da seguinte forma. Se não houver variáveis explicativas X no nível mais baixo, a equação 1 reduz para:

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{6}$$

Da mesma forma, se não houver variáveis explicativas Z no nível mais alto, a equação 2 reduz para:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{7}$$

Encontramos o modelo de equação única, substituindo 7 em 6:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{8}$$

Poderíamos também encontrar a equação 8, removendo todos os termos que

contenham uma variável X ou Z da equação 5. O modelo de equação 8 não explica nenhuma variação em Y. Ele apenas decompõe a variância em dois componentes independentes: σ_e^2, que é a variância dos erros do nível mais baixo e_{ij}, e $\sigma_{u_0}^2$, que é a variância de os erros de nível mais alto u_{0j}. Usando este modelo, podemos definir a correlação intraclasse ρ pela equação:

$$\rho = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_{u_0}^2} \tag{9}$$

A correlação intraclasse ρ indica a proporção da variância explicada pela estrutura de agrupamento na população. A Equação 9 simplesmente demonstra que ρ é a proporção da variância do nível do grupo em comparação com a variância total.

2.2 Métodos de Estimação

O método de máxima verossimilhança é o mais usado para estimar os valores dos coeficientes de regressão, interceptos e respectivas variâncias na modelagem multinível. Neste método, o procedimento geral consiste em produzir estimativas que maximizam a probabilidade dos dados que foram observados, dado o modelo. Uma vantagem de usá-lo é que geralmente ele é robusto e produz estimativas que são assintoticamente eficientes e consistentes. Portanto, leves violações dos pressupostos, como erros não normais, poderão ser superadas sem prejuízos para a qualidade das estimativas.

O cálculo dessas estimativas, utiliza para modelos multiníveis, um caso paricular do método numérico de Newton Raphson, onde os valores iniciais são baseados em estimativas de regressão de nível único.

Na modelagem de regressão multinível são usadas duas funções de verossimilhança diferentes. Uma delas é a Máxima Verossimilhança Completa (*Full Maximum Likelihood* - FML). Nesse método, os coeficientes de regressão e os componentes de variância são incluídos na função de verossimilhança. Os coeficientes de regressão são tratados como fixos e desconhecidos quando os componentes de variância são estimados. Os graus de liberdade perdidos pela estimativa dos efeitos fixos não são levados em consideração. As estimativas para os componentes de variância são viesadas, mas geralmente o vies é pequeno. Apesar disso, a FML continua sendo usada, por dois motivos. Primeiro, os cálculos são geralmente mais fáceis que os do outro método, e segundo, como os coeficientes de regressão são incluídos na função de verossimilhança, um teste qui-quadrado baseado na verossimilhança pode ser usado para comparar dois modelos que diferem na parte fixa.

Outra função de verossimilhança é a Máxima Verossimilhança Restrita (*Res-*

tricted Maximum Likelihood - RML). Neste caso apenas os componentes de variância estão incluídos na função de verossimilhança e os coeficientes de regressão são estimados em uma segunda etapa. O método RML estima os componentes de variância após remover os efeitos fixos do modelo e possui uma propriedade em que, se os grupos são balanceados (posssuem tamanhos de grupos iguais), as suas estimativas são equivalentes às estimativas de análise de variância (ANOVA).

Em geral, ambos os métodos produzem estimativas de parâmetros com erro padrão associado e uma deviance geral do modelo que é uma função da verossimilhança. Na prática, as diferenças entre os dois métodos são pequenas. Deste modo, a comparação entre as estimativas de FML para o modelo nulo e as estimativas RML correspondentes, evidenciam uma diferença é absolutamente trivial. Se forem encontradas diferenças não triviais, o método RML é geralmente melhor (Hox, 2010).

O método de Mínimos Quadrados Generalizados pode ser usado como substituto do método de máxima verossimilhança em casos particulares uma vez que suas estimativas são assintoticamente equivalentes. Esse método é, em tese, um procedimento de verossimilhança com apenas uma iteração que possuí a vantagem de ter um cálculo mais simples e rápido que as estimativas de FML e permitir violações de pressupostos. A desvantagem é que simulações apontam uma menor eficiência das estimativas e um problema grave de imprecisão dos erros padrão (Hox, 2010).

2.3 Teste de Significância

A estimação por máxima verossimilhança gera as estimativas de parâmetro que serão usados no teste de significância através da estatística Z. Em que:

$$Z = \frac{\hat{\beta}}{ErroPadrão(\hat{\beta})} \tag{10}$$

Dada uma hipótese nula onde o parâmetro populacional é zero, a estatística Z é usada para testar se as estimativas estão relacionadas à distribuição normal padrão através um p-valor.

Os erros padrão são assintóticos e são válidos quando as amostras são grandes.

O teste de Wald é aplicado frequentemente para testar significância. Quando o interesse está nos efeitos fixos, é mais adequado associar a estatística do teste à distribuição t com (J - p - 1) graus de liberdade. Ela é mais conservadora e reduz os riscos de ocorrer um erro tipo I (rejeitar a hipótese nula, dada que ela é verdadeira). Nos casos em que o tamanho da amostra é suficientemente grande, a

diferença entre os dois é trivial, mas quando a amostra é pequena, essa diferença pode se tornar relevante. Uma alternativa para amostras pequenas, na análise de regressão multinível, é a aproximação de Satterthwaite que estima estima o número de graus de liberdade usando as variâncias residuais e funciona melhor que o teste de Wald.

2.4 Comparação de Modelos

A *deviance* é uma estatística obtida a partir da função de verossimilhança e indica o quão bem o modelo está ajustado aos dados. Portanto, uma *deviance* pequena, em geral, mostra que o modelo se ajustou melhor do que um outro modelo que contenha uma *deviance* maior.

A estatística do teste é dada por:

$$Deviance = -2 \cdot ln(L) \tag{11}$$

em que L é o valor da função de verossimilhança.

A partir do teste da *deviance* é possível comparar os ajustes dos modelos encaixados e investigar a importância de efeitos aleatórios, comparando um modelo com esse conjunto de efeitos aleatórios e um modelo sem o conjunto. Quando um modelo específico pode ser proveniente de um mais geral, eles são denominados encaixados, aninhados ou sobrepostos. Nesse caso particular, a comparação é feita usando a *deviance*.

A *deviance* segue uma distribuição qui-quadrado com a graus de liberdade, onde a é a diferença do número de parâmetros estimados nos dois modelos.

Para modelos de regressão multinível não aninhados, a comparação pode ser feita usando o *Akaike's Information Criterion* (AIC). O AIC (Akaike, 1987) é um índice de ajuste geral calculado a partir da *deviance* d e do número de parâmetros estimados q:

$$AIC = d + 2q \tag{12}$$

No que refere a interpretação, quanto menor for o valor para cada uma das duas medidas de comparação, melhor será o ajuste do modelo.

Para finalizar, é importante salientar que o AIC é mais indicado para comparar os modelos em casos onde os dados são hierárquicos, de forma mais simples.

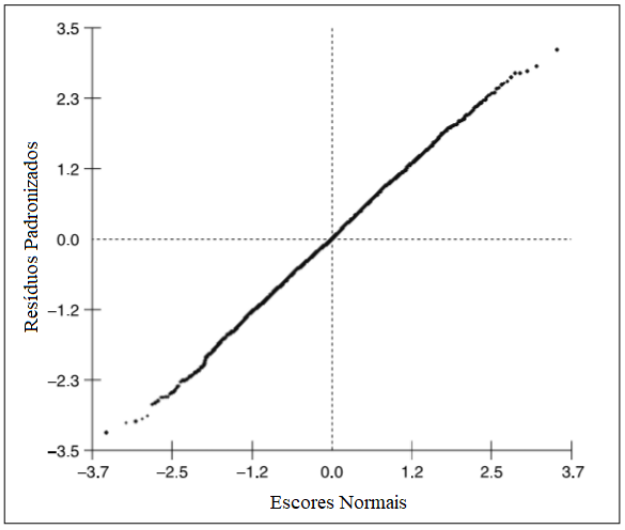
2.5 Análise de Resíduo

Em análise de regressão, a análise de resíduo é um método visual, feito a partir da análise gráfica, que valida o modelo verificando a normalidade, homocedasticidade e linearidade do modelo. Um modelo de regressão possui um resíduo para cada efeito aleatório. Caso o modelo multinível não possua interação entre os níveis, sua equação é dada por:

$$Y_{ik} = \gamma_{00} + \gamma_{10} \cdot X_{1ik} + \gamma_{20} \cdot X_{2ik} + \gamma_{01} \cdot Z_k + u_{2k} \cdot X_{2ik} + u_{0k} + e_{ik} \qquad (13)$$

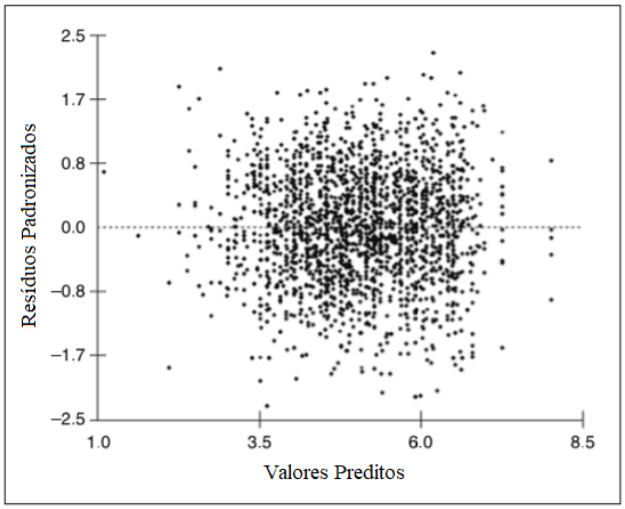
Composto por três erros, u_{2k}, u_{0k} e e_{ik}. Como nem sempre é possível identificar a presença de pontos extremos no modelo, o *boxplot* é uma ferramenta adequada para essa finalidade. Para testar a normalidade, o gráfico que relaciona os resíduos padronizados e os escores normais é utilizado. O ideal é que os pontos desse gráfico estejam bem próximos à uma linha reta diagonal crescente. Finalmente, o gráfico de dispersão dos resíduos versus os valores preditos é usado para investigar possíveis problemas e normalidade, homocedasticidade e linearidade. As figuras 1 e 2 mostram gráficos nos quais os pontos estão distribuídos aleatoriamente e indicam pressupostos não violados.

Figura 1: Gráfico de resíduos padronizado em relação a escores normais.



Fonte: Hox, 2010, pg 25.

Figura 2: Gráfico dos resíduos padronizados em relação aos valores preditos.



Fonte: Hox, 2010, pg 26.

2.6 Estratégias de Análise

A regressão multinível reduz drasticamente o número de estimativas calculando a média e a variância delas. Portanto um modelo que inicialmente teria 100 estimativas, passa a ter somente a média e a variância de todas elas, além de uma suposição de normalidade. Apesar da simplificação do modelo e da redução de variáveis explicativas, o modelo ainda é complexo. Geralmente, evita-se estimar o modelo completo porque problemas computacionais e de convergência são muito prováveis nesses casos. O ideal são modelos mais restritos que incluam apenas parâmetros significativos ou que possuam uma maior importância para a explicação do problema.

Duas estratégias podem ser aplicadas na construção do modelo. A abordagem de cima para baixo começa com o modelo que incluí o número máximo de efeitos fixos e interações possíveis para o modelo, seguida da retirada dos efeitos não significativos. Em uma segunda etapa, coloca-se todas as variâncias possíveis e em seguida, retira-se os efeitos aleatórios insignificantes. Uma desvantagem dessa abordagem são os problemas citados anteriormente para o modelo completo (complicações computacionais e problemas de convergência). A abordagem de baixo pra cima começa com um modelo simples e vai testando a significância de parâmetros adicionados aos poucos. É comum iniciar com o modelo nulo, adicionar a parte fixa e vistoriar o erro residual e a significância através das estimativas dos parâmetros e dos erros padrão. O benefício dessa estratégia é a propensão a manter os modelos

mais simples.

As estimativas dos parâmetros fixos são mais precisas que os parâmetros aleatórios. Além disso, é no nível mais baixo que estão as maiores amostras. Portanto, a construção do modelo começa com os coeficientes de regressão fixos e em seguida, adiciona-se os componentes de variância do nivel mais baixo. Numa segunda etapa, repete-se o procedimento para o segundo nível. Hox (2010) descreve abaixo os passos do mecanismo de seleção de variáveis:

- **Etapa 1:** Analisar o modelo nulo. Aquele que contém apenas o intercepto e é dado por:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{14}$$

Na equação 14, γ_{00} é a intercepto, e u_{0j} e e_{ij} são os resíduos no nível de grupo e no nível individual. Esse modelo nos proporciona um valor de referência da deviance e uma estimativa da correlação intraclasse ρ:

$$\rho = \frac{\sigma_{u_0^2}}{u_0^2 + u_e^2} \tag{15}$$

onde $\sigma_{u_0^2}$ é a variância dos resíduos de nível de grupo u_0^2 e u_e^2 é a variância dos resíduos de nível individual e_{ij}.

- **Etapa 2:** Analisar o modelo com todos os coeficientes de regressão das variáveis explicativas de nível inferior. Este modelo é escrito como:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} \cdot X_{pij} + u_{0j} + e_{ij} \tag{16}$$

onde o X_{pij} são as variáveis explicativas p no nível do individuo. Nesta etapa, avalia-se a contribuição de cada variável preditora de nível individual. Pode-se testar a significância de cada variável explicativa a fim de avaliar as mudanças ocorridas em termos do primeiro e segundo níveis da variância.

- **Etapa 3:** adicionar as variáveis explicativas de nível superior:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} \cdot X_{pij} + \gamma_{0q} \cdot Z_{qj} + u_{0j} + e_{ij} \tag{17}$$

onde o Z_{qj} são as variáveis explicativas no nível do grupo. Nessa etapa é possivel avaliar se as variáveis preditoras do nível mais alto explicam a variação entre grupos na variável resposta.

Os modelos construídos no segundo e terceiro passo podem ser chamados de modelos de componentes de variância. Estes desagregam a variância do intercepto em componentes de variância diversos para cada nível hierárquico. Acredita-se que o valor do intercepto é alterado entre os grupos, mas o valor dos coeficientes de regressão se mantêm inalterados.

- **Etapa 4:** Avaliar se alguns dos coeficientes de qualquer uma das variáveis explicativas tem um componente de variância significativo entre os grupos. Este modelo coeficientes aleatórios, é dado por:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} \cdot X_{pij} + \gamma_{0q} \cdot Z_{qj} + u_{pj} \cdot X_{pij} + u_{0j} + e_{ij} \tag{18}$$

  onde u_{pj} são os resíduos de nível de grupo dos coeficientes das variáveis explicativas de nível individual X_{pij}.

- **Etapa 5:** Para chegar ao modelo completo, adicionar interações de nível cruzado entre os preditores do nível mais alto e os preditores do nível mais baixo que tiveram, na etapa anterior, uma variância significativa:

$$Y_{ij} = \gamma_{00} + \gamma_{10} \cdot X_{ij} + \gamma_{01} \cdot Z_j + \gamma_{11} \cdot X_{ij} \cdot Z_j + u_{1j} \cdot X_{1ij} + u_{0j} + e_{ij} \tag{19}$$

  Observe que se houver mais de dois níveis, as etapas 3 e 4 devem ser repetidas para cada um dos níveis. E se for utilizado o método de estimação de máxima verossimilhança completa, o ajuste do modelo final de cada etapa pode ser avaliado a partir do teste da *deviance*.

Outras estratégias de análise que podem ser aplicadas dependem do tamanho da amostra. Para amostras grandes, uma divisão aleatória igualitária dos dados, possiblita que uma das partes seja usada de forma exploratória e a outra parte valide o modelo final. No caso de amostras que não permitam essa divisão, uma possiilidade é aplicar a correção de Bonferroni para testar a parte fixa individualmente em cada etapa.

2.7 Coeficiente de Determinação

A correlação múltipla quadrada também conhecida como coeficiente de determinação (denotada por R^2), é a proporção da variação total explicada pelas covariáveis. Na análise multinível, a variância explicada é um ponto complexo. Se houver coeficientes aleatórios, o conceito de variância explicada passa a ser complexo.

Além de possuir mais de uma definição, podem existir variâncias não explicadas em vários níveis. Existem maneiras alternativas para avaliar a qualidade da predição dos resultados obtidos por um modelo multinível.

Uma maneira de calcular a proporção de variância explicada para o nível mais baixo (R_1^2) é usando a diferença entre as variâncias como uma proporção da variância total do erro. Dada por:

$$R_1^2 = \frac{\sigma_{e|b}^2 - \sigma_{e|m}^2}{\sigma_{e|b}^2} \tag{20}$$

Em que:

$\sigma_{e|b}^2$ é a variância residual do menor nível para o modelo nulo, que é utilizado como modelo de referência dado que decompõe a variância total da variável resposta em dois níveis e,

$\sigma_{e|m}^2$ é a variância residual do menor nível para o modelo a ser comparado.

Analogamente, o cálculo da proporção da variância explicada R_2^2 para o segundo nível, pode ser feito a partir de:

$$R_2^2 = \frac{\sigma_{u_0|b}^2 - \sigma_{u_0|m}^2}{\sigma_{u_0|b}^2} \tag{21}$$

Em que:

$\sigma_{u_0|b}^2$ é a variância residual do maior nível para o modelo nulo, que é o modelo de referência e,

$\sigma_{u_0|m}^2$ é a variância residual do maior nível para o modelo a ser comparado.

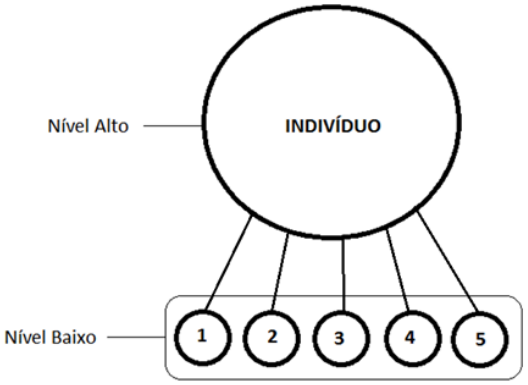
A interpretação dessas medidas é análoga à do R^2.

Em um modelo de coeficientes aleatórios que não contenha interações de nível cruzado entre os níveis com uma covariável do primeiro nível, a variação não é modelada e é análoga à variância de erro do intercepto do segundo nível.

2.8 Análise de Dados Longitudinais Multiníveis

Dados longitudinais podem ser vistos como dados que apresentam estrutura hierárquica onde as medidas repetidas de uma variável constituem um conjunto de observações do nível mais baixo agrupadas em um nível mais alto do indivíduo, como mostra a figura a seguir.

Figura 3: Modelo multinível de dois níveis



Assim podem ser analisados através de modelos multiníveis com dois níveis semelhantes ao apresentado no tópico anterior.

Quando os dados são longitudinais, existe uma correlação entre as medidas repetidas que viola o pressuposto de independência dos erros da regressão. A correlação causada pelo efeito da memória deve ser considerada no modelo através dos erros correlacionados. Além disso, é possível estudar sobre a trajetória média do grupo ou de dos indivíduos ao longo do tempo adicionando ao modelo covariáveis que variam no tempo ou covariáveis constantes ao longo do tempo.

Alguns bancos de dados com dados longitudinais possuem particularidades que podem dificultar na hora da análise. Dados faltantes e a periodicidade entre as medidas repetidas podem ser características problemáticas para algumas técnicas de análise. Se os espaços entre as coletas for grande o suficiente para que a memória não altere os dados e se as escalas usada sejam comparáveis, as particularidades não serão um problema para a modelagem multinível.

O modelo de regressão multinível para dados longitudinais também pode ser escrito como uma sequência de modelos para cada nível. No mínimo, o nível de medidas repetidas, temos:

$$Y_{ij} = \pi_{0i} + \pi_{1i} \cdot T_{ti} + \pi_{2i} \cdot X_{ti} + e_{ti} \tag{22}$$

em que os níveis mais baixos são representados pelo π, os coeficientes de nível de indivíduo, que em medidas repetidas estão no segundo nível, são indicados por β, Y_{ti} é a variável de resposta do indivíduo i medido na ocasião de medição t, T é a variável de tempo que indica a ocasião de medição e X_{ti} é uma covariável variável no tempo.

$$\pi_{0i} = \beta_{00} + \beta_{01} \cdot Z_i + u_{01} \tag{23}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} \cdot Z_i + u_{11} \tag{24}$$

$$\pi_{2i} = \beta_{20} + \beta_{21} \cdot Z_i + u_{21} \tag{25}$$

Por substituição, obtemos o modelo de equação única:

$$T_{ti} = \beta_{00}+\beta_{10}\cdot T_{ti}+\beta_{20}\cdot X_{ti}+\beta_{01}\cdot Z_i+\beta_{11}\cdot_{ti}\cdot Z_i+\beta_{21}\cdot X_{ti}\cdot Z_i+u_{1i}\cdot T_{ti}+u_{2i}\cdot X_{ti}+u_{0i}+e_{ti} \tag{26}$$

Se o ponto de interesse for testar se as médias são iguais para todas as ocasiões de medição, a análise de variância de medidas repetidas pode ser aplicada desde que a suposição de esfericidade seja satisfeita. Esfericidade significa que existem restrições complexas nas variâncias e covariâncias entre as medidas repetidas e para saber mais detalhes, veja Hox (2010, Capítulo 5).

Em determinadas situações será inadequado usar o zero para se referir ao primeiro momento de coleta. Caso isso ocorra, uma solução viável é centrar-se na média, mediana ou em um valor próximo dessas medidas de posição da variável. Portanto, é necessário ter cautela ao codificar a variável tempo. Ao gerar automaticamente a variável tempo, é preciso se atentar a fato de que a maioria dos *softwares* atribui o valor 1 à primeira ocasião tornando zero um valor inválido para a variável.

No processo de estimação da variância do segundo nível explicada pela ocasião de medição, encontrar um valor negativo é recorrente. Esse resultado impossibilita a utilização da variância do erro residual do modelo nulo como base de comparação para verificar se a inserção de variáveis explicativas melhora o modelo.

2.9 Vantagens da Análise Multinível para Dados Longitudinais

O uso de modelos multiníveis para analisar dados de medidas repetidas tem diversas vantagens. Bryk e Raudenbush (1992) apontam 5 vantagens de usar modelos multiníveis na análise de dados com medidas repetidas.

1. A modelagem dos coeficientes de regressão da ocasião de medição, gera curvas de crescimento distintas para cada indivíduo. Isso evita a perda de informações

que poderia existir se só pudéssemos usar uma curva de crescimento médio do indivíduos.

2. Essa técnica permite a análise de dados com quantidades distintas de medições e diferentes intervalos entre as ocasiões de coleta. Portanto, cada indivíduo pode ser observado em momentos diferentes.

3. É possível modelar as covariâncias entre a medidas repetidas explicitando uma estrutura específica para as variâncias e covariância em qualquer nível.

4. Se os dados são balanceados e estimados pelo método da máxima verossimilhança completa, a análise de testes F baseadas na variância e testes t podem ser provenientes dos resultados de regressão multinível (ver Raudenbush, 1993a). Portanto, pode-se dizer que a ANOVA para medidas repetidas é um caso particular de análise de regressão multinível.

5. Essa abordagem admite a adição de níveis mais altos para examinar o efeito de grupos familires ou sociais no comportamento dos indivíduos.

Uma vantagem não citada por Bryk e Raudenbush, é a facilidade de incluir covariáveis constantes ou que variam no tempo , no modelo, permitindo modelar o comportamento médio do grupo e dos diferentes indivíduos ao longo do tempo.

3 MATERIAIS E MÉTODOS

Os dados utilizados neste estudo, são derivados da base dados da "Pesquisa do Perfil do Estudante da Universidade de Brasília - Etapa Registro" realizada pelo Observatório da Vida Estudantil (OVE), vinculado ao Núcleo de Estudos e Pesquisas no Ensino Superior (Nesub/CEAM).

O Observatório da Vida Estudantil, visa obter informações do corpo discente da Universidade de Brasília, com a finalidade de subsidiar novas pesquisas e novas políticas estudantis. Para esse fim, desenvolveu-se um questionário com cerca de 60 questões que passou a ser documento exigido no momento do registro, a partir do primeiro semestre 2012.

A unidade de observação da base de dados preliminar, é o aluno ingressante e a do presente estudo, é o curso. A partir de uma análise inicial dos dados preliminares, tendo em vista identificar possíveis características que ajudariam a traçar os perfis dos cursos, selecionou-se um conjunto de variáveis para a construção do banco de dados a ser analisado. Para isso, calculou-se o coeficiente de variação do percentual da resposta de interesse de cada questão selecionada. Portanto, as variáveis do banco secundário são, na sua maioria, percentuais de uma determinada categoria para cada curso.

Para medir o status socioeconômico dos estudantes de cada curso, utilizou-se um Indicador de Status Socioeconômico (INDISSE) inspirado no Critério Brasil da Associação Brasileira de Empresas de Pesquisa. O indicador evidencia o status socioeconômico a partir da soma das pontuações atribuídas às variáveis relativas a quantidade de determinados bens materiais, atividades extracurriculares desenvolvidas na vida pregressa e escolaridade de mãe e pai. "É uma variável quantitativa discreta que varia de 0 a 100, sendo zero, o nível socioeconômico mais baixo e cem, o mais alto" (Costa, 2015).

A variável que mostra o percentual de ingressantes que superam a escolaridade dos pais foi construída de uma forma diferente. No banco preliminar existem as perguntas sobre escolaridade do pai e da mãe. O nível máximo de escolaridade entre pai e mãe foi observado e comparado com o grau de escolaridade do ingressante (considerando que todo ingressante tem pelo menos o nível superior incompleto).

O banco final utilizado nesta pesquisa possui 249 observações, referentes à 3 medições de cada um dos 83 cursos presenciais de graduação da Universidade de Brasília. Para a construção do referido banco, foram considerados os dados de alunos que ingressaram na UnB no primeiro semestre dos anos de 2015, 2016 e 2017, pelas modalidades PAS, SISU ou vestibular, com preenchimento válido das variáveis

Curso e Semestre.

Na modelagem multinível, foram consideradas todas as variáveis da tabela 1. A modelagem multinível foi realizada no software estatístico SAS *OnDemand* utilizando-se o PROC MIXED.

Tabela 1: Tabela das variáveis do banco final

Descrição	Tipo
Variável Resposta	
INDISSE médio	Quantitativa
Nível Alto	
Curso	Qualitativa
Nível Baixo	
Semestre	Qualitativa
Percentual de cotistas	Quantitativa
Percentual de mulheres	Quantitativa
Percentual de pretos e pardos	Quantitativa
Percentual de Pessoas que fizeram o ensino médio total ou majoritariamente em escola pública	Quantitativa
Percentual de pessoas que superam a escolaridade máxima dos pais	Quantitativa
Percentual de pessoas que ingressaram pelo PAS	Quantitativa

4 RESULTADOS

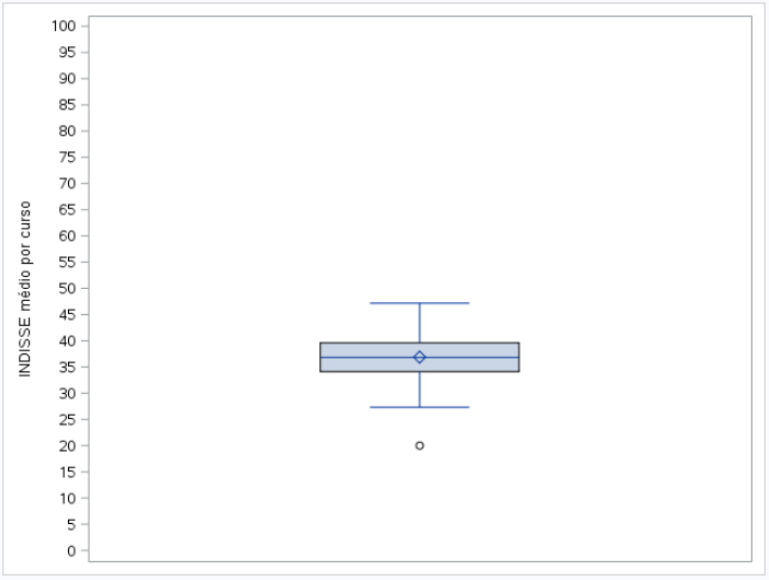
A tabela 2 apresenta as medidas das características estudadas com o objetivo de traçar o perfil dos cursos.

Tabela 2: Medidas descritivas das características dos cursos

Variável	N	Média	Desvio Padrão	Coeficiente de Variação	Mínimo	Primeiro Quartil	Mediana	Terceiro Quartil	Máximo
INDISSE médio	249	36,4	4,13	11,36	20	33,5	36,2	39,18	47,17
% de cotistas	248	50,8	9,53	18,75	16,67	45,55	51,83	56,58	76,92
% de mulheres	248	48,53	19,93	41,07	4,76	33,81	50	64,29	89,66
% de pretos e pardos	249	54,57	9,8	17,95	14,29	48,39	54,55	60,71	84,91
% de alunos oriundos do ensino médio	249	51,57	9,7	18,81	20	46,03	51,35	57,14	80
% de pessoas que superam a escolaridade dos pais	247	37,71	11,81	31,31	14,63	28,57	36,84	45	80
% de pessoas que ingressaram pelo PAS	246	48,9	11,08	22,67	5,62	46,15	50,5	55,26	68,18

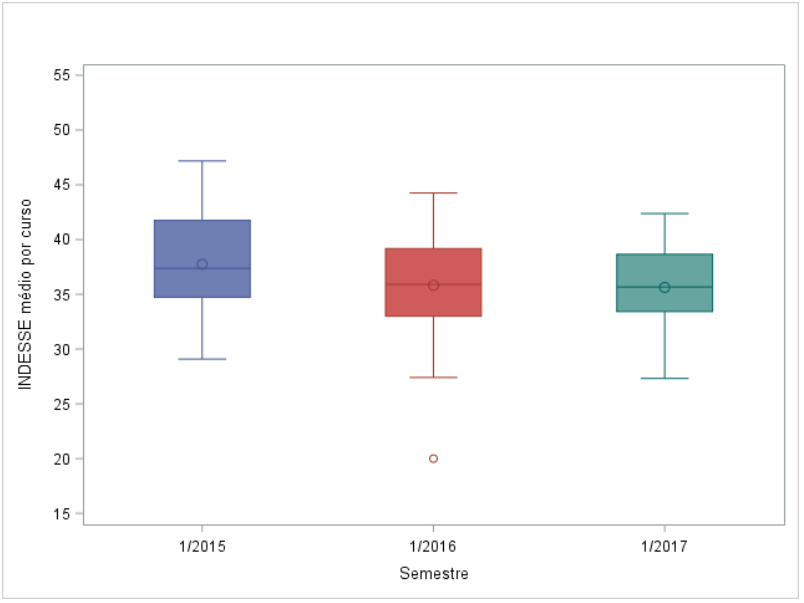
A média do Indicador de Status Socioeconômico dos Estudantes de cada curso (INDISSE médio) no período estudado foi no mínimo de 20 pontos e no máximo de 47,17. A análise da tabela 2 e do boxplot apresentado na figura 4 mostra que a mediana e a média dessa variável são bem próximas . Pode-se constatar também que 25% dos cursos tem INDISSE médio até 33,5 pontos, enquanto que para 25% dos cursos o Indicador é de 39,18 pontos ou mais. É importante destacar que, embora o INDISSE-UnB varie de 1 a 100, nenhum dos cursos citados no presente estudo, alcançou uma média de 50 pontos, a metade da pontuação máxima possível.

Figura 4: INDISSE médio por curso de Graduação da Universidade de Brasilia - 2015 a 2017



A figura 5 mostra o comportamento do INDISSE médio dos cursos de Graduação da Universidade de Brasilia - 2015 a 2017 por semestre. Nota-se uma redução da amplitude ao longo do tempo, além de uma redução da mediana.

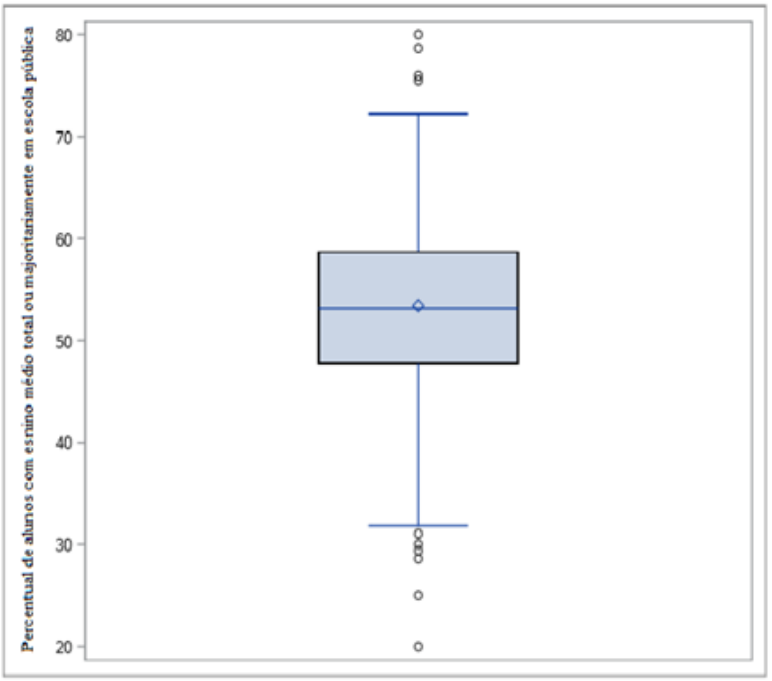
Figura 5: INDISSE médio dos cursos de Graduação da Universidade de Brasilia por semestre



O comportamento do percentual de pessoas que fizeram o ensino médio total ou majoritariamente em escola pública observado através do gráfico da figura 7, aponta uma amplitude menor do que a observada no percentual de mulheres,

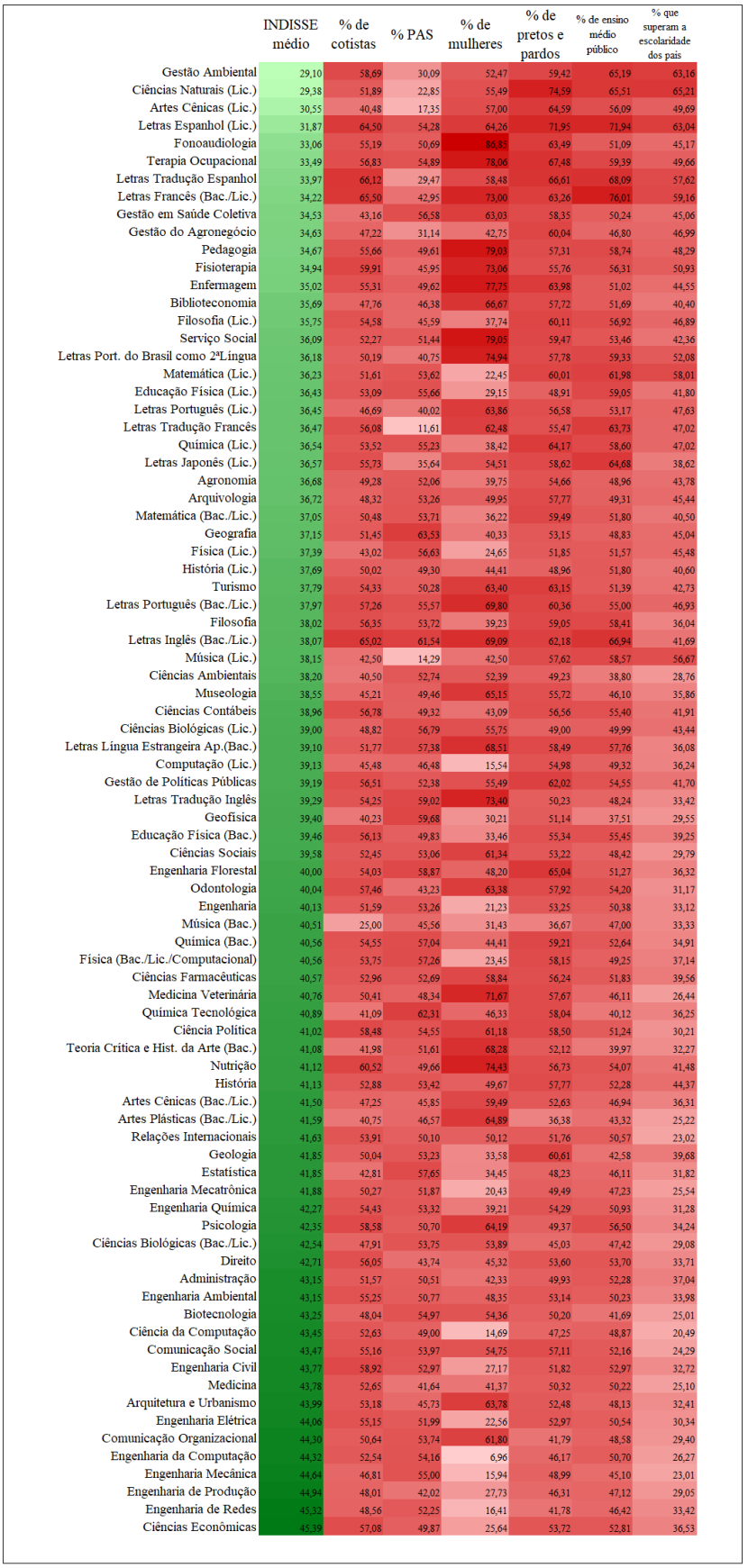
tornando a coluna que representa a variável relacionada à natureza do ensino médio mais homogênea que a coluna relacionada ao percentual de mulheres. A figura 6 mostra que o gráfico possui um maior número de pontos extremos. O coeficiente de variação do percentual de alunos oriundos de escola pública é de 18,8.

Figura 6: Percentual de ingressantes dos cursos de graduação da UnB - 2015 a 2017 que cursaram ensino médio total ou majoritariamente em escola pública



A figura 7 apresenta os cursos ordenados segundo o INDISSE- médio dos estudantes do curso variando entre 29,10 (Gestão Ambiental) a 45 ,39 (Ciências Econômicas). Nas dez primeiras linhas do gráfico da figura 7, é possível notar uma pequena mancha mais escura na variável "Percentual de mulheres" que indica a presença de dois cursos mais femininos (Fonoaudiologia e Terapia Ocupacional). Exceto o percentual de ingressantes pelo PAS, todas as outras variáveis apresentam tons mais escuros para os dez cursos com menor média de pontuação do INDISSE.

Figura 7: *Heatmap* das variáveis por curso de graduação da UnB - 2015 a 2017



	INDISSE médio	% de cotistas	% PAS	% de mulheres	% de pretos e pardos	% de ensino médio público	% que superam a escolaridade dos pais
Gestão Ambiental	29,10	58,69	30,09	52,47	59,42	65,19	63,16
Ciências Naturais (Lic.)	29,38	51,89	22,85	55,49	74,59	65,51	65,21
Artes Cênicas (Lic.)	30,55	40,48	17,35	57,00	64,59	56,09	49,69
Letras Espanhol (Lic.)	31,87	64,50	54,28	64,26	71,95	71,94	63,04
Fonoaudiologia	33,06	55,19	50,69	86,85	63,49	51,09	45,17
Terapia Ocupacional	33,49	56,83	54,89	78,06	67,48	59,39	49,66
Letras Tradução Espanhol	33,97	66,12	29,47	58,48	66,61	68,09	57,62
Letras Francês (Bac./Lic.)	34,22	65,50	42,95	73,00	63,26	76,01	59,16
Gestão em Saúde Coletiva	34,53	43,16	56,58	63,03	58,35	50,24	45,06
Gestão do Agronegócio	34,63	47,22	31,14	42,75	60,04	46,80	46,99
Pedagogia	34,67	55,66	49,61	79,03	57,31	58,74	48,29
Fisioterapia	34,94	59,91	45,95	73,06	55,76	56,31	50,93
Enfermagem	35,02	55,31	49,62	77,75	63,98	51,02	44,55
Biblioteconomia	35,69	47,76	46,38	66,67	57,72	51,69	40,40
Filosofia (Lic.)	35,75	54,58	45,59	37,74	60,11	56,92	46,89
Serviço Social	36,09	52,27	51,44	79,05	59,47	53,46	42,36
Letras Port. do Brasil como 2ªLíngua	36,18	50,19	40,75	74,94	57,78	59,33	52,08
Matemática (Lic.)	36,23	51,61	53,62	22,45	60,01	61,98	58,01
Educação Física (Lic.)	36,43	53,09	55,66	29,15	48,91	59,05	41,80
Letras Português (Lic.)	36,45	46,69	40,02	63,86	56,58	53,17	47,63
Letras Tradução Francês	36,47	56,08	11,61	62,48	55,47	63,73	47,02
Química (Lic.)	36,54	53,52	55,23	38,42	64,17	58,60	47,02
Letras Japonês (Lic.)	36,57	55,73	35,64	54,51	58,62	64,68	38,62
Agronomia	36,68	49,28	52,06	39,75	54,66	48,96	43,78
Arquivologia	36,72	48,32	53,26	49,95	57,77	49,31	45,44
Matemática (Bac./Lic.)	37,05	50,48	53,71	36,22	59,49	51,80	40,50
Geografia	37,15	51,45	63,53	40,33	53,15	48,83	45,04
Física (Lic.)	37,39	43,02	56,63	24,65	51,85	51,57	45,48
História (Lic.)	37,69	50,02	49,30	44,41	48,96	51,80	40,60
Turismo	37,79	54,33	50,28	63,40	63,15	51,39	42,73
Letras Português (Bac./Lic.)	37,97	57,26	55,57	69,80	60,36	55,00	46,93
Filosofia	38,02	56,35	53,72	39,23	59,05	58,41	36,04
Letras Inglês (Bac./Lic.)	38,07	65,02	61,54	69,09	62,18	66,94	41,69
Música (Lic.)	38,15	42,50	14,29	42,50	57,62	58,57	56,67
Ciências Ambientais	38,20	40,50	52,74	52,39	49,23	38,80	28,76
Museologia	38,55	45,21	49,46	65,15	55,72	46,10	35,86
Ciências Contábeis	38,96	56,78	49,32	43,09	56,56	55,40	41,91
Ciências Biológicas (Lic.)	39,00	48,82	56,79	55,75	49,00	49,99	43,44
Letras Língua Estrangeira Ap.(Bac.)	39,10	51,77	57,38	68,51	58,49	57,76	36,08
Computação (Lic.)	39,13	45,48	46,48	15,54	54,98	49,32	36,24
Gestão de Políticas Públicas	39,19	56,51	52,38	55,49	62,02	54,55	41,70
Letras Tradução Inglês	39,29	54,25	59,02	73,40	50,23	48,24	33,42
Geofísica	39,40	40,23	59,68	30,21	51,14	37,51	29,55
Educação Física (Bac.)	39,46	56,13	49,83	33,46	55,34	55,45	39,25
Ciências Sociais	39,58	52,45	53,06	61,34	53,22	48,42	29,79
Engenharia Florestal	40,00	54,03	58,87	48,20	65,04	51,27	36,32
Odontologia	40,04	57,46	43,23	63,38	57,92	54,20	31,17
Engenharia	40,13	51,59	53,26	21,23	53,25	50,38	33,12
Música (Bac.)	40,51	25,00	45,56	31,43	36,67	47,00	33,33
Química (Bac.)	40,56	54,55	57,04	44,41	59,21	52,64	34,91
Física (Bac./Lic./Computacional)	40,56	53,75	57,26	23,45	58,15	49,25	37,14
Ciências Farmacêuticas	40,57	52,96	52,69	58,84	56,24	51,83	39,56
Medicina Veterinária	40,76	50,41	48,34	71,67	57,67	46,11	26,44
Química Tecnológica	40,89	41,09	62,31	46,33	58,04	40,12	36,25
Ciência Política	41,02	58,48	54,55	61,18	58,50	51,24	30,21
Teoria Crítica e Hist. da Arte (Bac.)	41,08	41,98	51,61	68,28	52,12	39,97	32,27
Nutrição	41,12	60,52	49,66	74,43	56,73	54,07	41,48
História	41,13	52,88	53,42	49,67	57,77	52,28	44,37
Artes Cênicas (Bac./Lic.)	41,50	47,25	45,85	59,49	52,63	46,94	36,31
Artes Plásticas (Bac./Lic.)	41,59	40,75	46,57	64,89	36,38	43,32	25,22
Relações Internacionais	41,63	53,91	50,10	50,12	51,76	50,57	23,02
Geologia	41,85	50,04	53,23	33,58	60,61	42,58	39,68
Estatística	41,85	42,81	57,65	34,45	48,23	46,11	31,82
Engenharia Mecatrônica	41,88	50,27	51,87	20,43	49,49	47,23	25,54
Engenharia Química	42,27	54,43	53,32	39,21	54,29	50,93	31,28
Psicologia	42,35	58,58	50,70	64,19	49,37	56,50	34,24
Ciências Biológicas (Bac./Lic.)	42,54	47,91	53,75	53,89	45,03	47,42	29,08
Direito	42,71	56,05	43,74	45,32	53,60	53,70	33,71
Administração	43,15	51,57	50,51	42,33	49,93	52,28	37,04
Engenharia Ambiental	43,15	55,25	50,77	48,35	53,14	50,23	33,98
Biotecnologia	43,25	48,04	54,97	54,36	50,20	41,69	25,01
Ciência da Computação	43,45	52,63	49,00	14,69	47,25	48,87	20,49
Comunicação Social	43,47	55,16	53,97	54,75	57,11	52,16	24,29
Engenharia Civil	43,77	58,92	52,97	27,17	51,82	52,97	32,72
Medicina	43,78	52,65	41,64	41,37	50,32	50,22	25,10
Arquitetura e Urbanismo	43,99	53,18	45,73	63,78	52,48	48,13	32,41
Engenharia Elétrica	44,06	55,15	51,99	22,56	52,97	50,54	30,34
Comunicação Organizacional	44,30	50,64	53,74	61,80	41,79	48,58	29,40
Engenharia da Computação	44,32	52,54	54,16	6,96	46,17	50,70	26,27
Engenharia Mecânica	44,64	46,81	55,00	15,94	48,99	45,10	23,01
Engenharia de Produção	44,94	48,01	42,02	27,73	46,31	47,12	29,05
Engenharia de Redes	45,32	48,56	52,25	16,41	41,78	46,42	33,42
Ciências Econômicas	45,39	57,08	49,87	25,64	53,72	52,81	36,53

Figura 8: *Heatmap* das dez maiores pontuações do INDISSE médio

	INDISSE médio	% de cotistas	% PAS	% de mulheres	% de pretos e pardos	% de ensino médio público	% que superam a escolaridade dos pais
Ciências Econômicas	45,39	57,08	49,87	25,64	53,72	52,81	36,53
Engenharia de Redes	45,32	48,56	52,25	16,41	41,78	46,42	33,42
Engenharia de Produção	44,94	48,01	42,02	27,73	46,31	47,12	29,05
Engenharia Mecânica	44,64	46,81	55,00	15,94	48,99	45,10	23,01
Engenharia da Computação	44,32	52,54	54,16	6,96	46,17	50,70	26,27
Comunicação Organizacional	44,30	50,64	53,74	61,80	41,79	48,58	29,40
Engenharia Elétrica	44,06	55,15	51,99	22,56	52,97	50,54	30,34
Arquitetura e Urbanismo	43,99	53,18	45,73	63,78	52,48	48,13	32,41
Medicina	43,78	52,65	41,64	41,37	50,32	50,22	25,10
Engenharia Civil	43,77	58,92	52,97	27,17	51,82	52,97	32,72

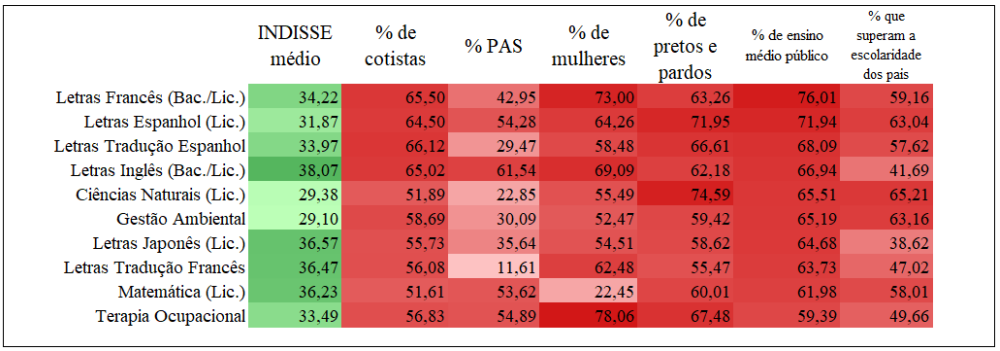
Fazendo um recorte da representação gráfica da figura 7 e observando os cursos com a maiores pontuações do INDISSE, apresentam, em geral, uma menor quantidade de mulheres e de alunos que possuem pais com escolaridade menor que ensino superior incompleto. Seis dos dez cursos com melhor status socioeconômico médio são Engenharias.

Figura 9: *Heatmap* dos dez menores percentuais de mulheres

	INDISSE médio	% de cotistas	% PAS	% de mulheres	% de pretos e pardos	% de ensino médio público	% que superam a escolaridade dos pais
Engenharia da Computação	44,32	52,54	54,16	6,96	46,17	50,70	26,27
Ciência da Computação	43,45	52,63	49,00	14,69	47,25	48,87	20,49
Computação (Lic.)	39,13	45,48	46,48	15,54	54,98	49,32	36,24
Engenharia Mecânica	44,64	46,81	55,00	15,94	48,99	45,10	23,01
Engenharia de Redes	45,32	48,56	52,25	16,41	41,78	46,42	33,42
Engenharia Mecatrônica	41,88	50,27	51,87	20,43	49,49	47,23	25,54
Engenharia	40,13	51,59	53,26	21,23	53,25	50,38	33,12
Matemática (Lic.)	36,23	51,61	53,62	22,45	60,01	61,98	58,01
Engenharia Elétrica	44,06	55,15	51,99	22,56	52,97	50,54	30,34
Física (Bac./Lic./Computacional)	40,56	53,75	57,26	23,45	58,15	49,25	37,14

Em relação aos cursos majoritariamente masculinos, os dez cursos são da área de exatas, sendo os três menores percentuais de mulheres pertencentes aos cursos de computação possuem. Os únicos dois cursos com habilitação exclusiva para licenciatura presentes nesse grupo, demonstram percentuais destoantes em relação à pontuação média do Indicado de Status Socioeconômico do Estudantes e à percentual de pais com escolaridade mais baixa que seus filhos.

Figura 10: *Heatmap* dos dez maiores percentuais de alunos oriundos de escola pública



	INDISSE médio	% de cotistas	% PAS	% de mulheres	% de pretos e pardos	% de ensino médio público	% que superam a escolaridade dos pais
Letras Francês (Bac./Lic.)	34,22	65,50	42,95	73,00	63,26	76,01	59,16
Letras Espanhol (Lic.)	31,87	64,50	54,28	64,26	71,95	71,94	63,04
Letras Tradução Espanhol	33,97	66,12	29,47	58,48	66,61	68,09	57,62
Letras Inglês (Bac./Lic.)	38,07	65,02	61,54	69,09	62,18	66,94	41,69
Ciências Naturais (Lic.)	29,38	51,89	22,85	55,49	74,59	65,51	65,21
Gestão Ambiental	29,10	58,69	30,09	52,47	59,42	65,19	63,16
Letras Japonês (Lic.)	36,57	55,73	35,64	54,51	58,62	64,68	38,62
Letras Tradução Francês	36,47	56,08	11,61	62,48	55,47	63,73	47,02
Matemática (Lic.)	36,23	51,61	53,62	22,45	60,01	61,98	58,01
Terapia Ocupacional	33,49	56,83	54,89	78,06	67,48	59,39	49,66

Entre os dez cursos com maior percentual de ingressantes que fizeram o ensino médio totalmente ou majoritariamente em escolas públicas, cinco cursos são de Letras. A pontuação média do INDISSE desses cursos alcança os menores valores observados, ao passo que os percentuais de alunos filhos de pais com baixa escolaridade são, na maioria dos casos, muito próximos do percentual máximo registrado.

Com base na análise inicial e considerando o objetivo de verificar a existência de variação dos perfis de curso ao longo do tempo, utlizou-se a modelagem multinível para dados longitudinais. Ao ajustar o modelo nulo, no qual existe somente o intercepto sem a presença de quaisquer variáveis explicativas, obteve-se que o valor do intercepto é de 36,4, o que significa que essa é a pontuação média do INDISSE-UnB para todos os cursos, em todos os períodos. A correlação intraclasse foi de 0,5198. Portanto, 52% da variância das repetições do INDISSE médio do curso pode ser explicado pelo curso, o que justifica o uso de um modelo multinível de 2 níveis.

Tabela 3: Modelo Nulo

Variáveis Explicativas	**Modelo Nulo**		
	Estimativa	**Erro Padrão**	**P-valor**
Efeito Fixo			
Intercepto	36,4	0,37	0,000
Efeito Aleatório - Nível 2			
Variância do Intercepto	8,88	1,83	0,000
Efeito Aleatório - Nível 1			
Variância do Resíduo	8,2	0,9	0,000
Correlação Intraclasse		51,98%	
AIC		1356,7	

Após executar a segunda etapa proposta pelo Hox (2010), confirmou-se a significância da variável semestre no modelo, que representa o tempo, com base na não rejeição da hipótese nula do teste t para um α=0,05.

Tabela 4: Modelo 1

Variáveis Explicativas	**Modelo 1**		
	Estimativa	**Erro Padrão**	**P-valor**
Efeito Fixo			
Intercepto	37,46	0,43	0,000
Semestre	-1,06	0,21	0,000
Efeito Aleatório - Nível 2			
Variância do Intercepto	9,25	1,82	0,000
Efeito Aleatório - Nível 1			
Variância do Resíduo	7,07	0,78	0,000
Correlação Intraclasse		56,68%	
AIC		1334,1	

A terceira etapa testou a significância da inclusão de todas as variáveis explicativas do nível mais baixo e o teste t apresentou evidências de que a hipótese de significância seria rejeitada com um α=0,05 para as variáveis referentes ao percentual de ingressantes pelo PAS e ao percentual de estudantes que se declaram pretos ou pardos.

Tabela 5: Modelo 2

Variáveis Explicativas	**Modelo 2**		
	Estimativa	**Erro Padrão**	**P-valor**
Efeito Fixo			
Intercepto	49,5451	1,51	0,000
Semestre	-1,1025	0,19	0,000
% de ingressantes pelo PAS	0,0012	0,02	0,939
% de cotistas	0,0657	0,02	0,007
% de mulheres	-0,0299	0,01	0,001
% de pretos e pardos	-0,0238	0,02	0,206
% de ingressantes oriundos de escola pública	-0,1014	0,02	0,000
% de ingressantes que superam a escolaridade dos pais	-0,1972	0,02	0,000
Efeito Aleatório - Nível 2			
Variância do intercepto	1,17	0,49	0,000
Efeito Aleatório - Nível 1			
Variância do Resíduo	4,2	0,49	0,000
Correlação Intraclasse		21,72%	
AIC		1112,2	

Tabela 6: Modelo Final

Variáveis Explicativas	**Modelo Final**		
	Estimativa	**Erro padrão**	**P-valor**
Efeito Fixo			
Intercepto	48,99	1,04	0,000
Semestre	-1,13	0,19	0,000
% de cotistas	0,07	0,02	0,001
% de mulheres	-0,03	0,01	0,001
% de ingressantes oriundos de escola pública	-0,12	0,02	0,000
% de ingressantes que superam a escolaridade dos pais	-0,2	0,02	0,000
Efeito Aleatório - Nível 2			
Variância do intercepto	1,16	0,47	0,007
Efeito Aleatório - Nível 1			
Variância do resíduo	4,24	0,48	0,000
Correlação Intraclasse		21,55%	
AIC		1118,9	

Assim o modelo final é dado por:

$$INDISSE_{jt} = 48,99 - 1,13W_{jt} + 0,0744X_{1jt} - 0,0298X_{2jt} - 0,1177X_{3jt} - 0,2048X_{4jt}$$

Em que:

W_{jt} é o semestre t, no curso j;

X_{1jt} é o percentual de cotistas no curso j, semestre t;

X_{2jt} é o percentual de mulheres no curso j, no semestre t;

X_{3jt} é o percentual ingressantes que fizeram o ensino médio total ou majoritariamente em rede pública no curso j, no semestre t;

X_{4jt} é o percentual de ingressantes que superam a escolaridade máxima dos pais no curso j, no semestre t;

j é o índice usado para indicar o curso. $j = 1, ..., 83$;

t é o índice usado para indicar o semestre. $t = 0, 1, 2$.

Portanto, as interpretações a seguir consideram que todas as demais variáveis são mantidas constantes. Cada acréscimo de um semestre, o INDISSE médio decai 1,13 pontos. Para cada acréscimo de um por cento de mulheres no curso, o INDISSE médio reduz 0,03 pontos. Para cada acréscimo de um por cento de ingressantes provenientes, total ou majoritariamente, do ensino público, o INDISSE

decresce 0,12 pontos. Para cada acréscimo de um ponto percentual de alunos que superam a escolaridade máxima dos pais, o INDISSE médio decresce 0,2 pontos. E para cada acréscimo de um por cento de cotistas no curso, o INDISSE médio aumenta 0,07 pontos.

5 CONCLUSÃO

Para analisar o perfil socioeconômico dos cursos de graduação da Universidade de Brasília no período de 2015 a 2017, foi utilizado o modelo proposto pela análise multinível. Inicialmente, verificou-se a existência de variações nos perfis dos cursos segundo as características dos ingressantes e ao longo dos semestres. Em um segundo momento, foi construído um banco para subsidiar a análise. Por fim, utilizou-se a modelagem multinível para dados longitudinais verificando assim, as características associadas ao perfil socioeconômico dos cursos.

A redução do INDISSE médio, variável que retrata o perfil socioeconômico dos cursos ao longo do tempo, está associada ao aumento do percentual de alunos que estão superando a escolaridade máxima dos pais, que pode ser um efeito da expansão do acesso ao ensino superior. Outra variável que está associada à redução do INDISSE, é o aumento do percentual de estudantes de escolas públicas, que pode ser resultado da implementação de cotas para este público. A terceira variável relacionada à diminuição do INDISSE, é a redução do percentual de cotistas que exige estudos mais profundo que justifiquem essa relação positiva. Por fim, a quarta variável é o crescimento do percentual de mulheres no curso, que pode ser justificado pela criação de vagas em cursos que possuem uma maior procura por pessoas do sexo feminino, como Fisioterapia, Terapia Ocupacional e Enfermagem. A ampliação do acesso ao ensino superior é um argumento plausível para a maioria dos resultados elencados acima, entretanto é recomendado um estudo mais profundo sobre a razão dos efeitos e da ocorrência desses fatores.

Desta forma, o aumento do acesso ao ensino superior pode explicar a mudança do perfil socioeconômico dos cursos ao longo do tempo e o crescimento da proporção de ingressantes de maior vulnerabilidade socioeconômica – situação em que os pais não tiveram acesso ao ensino superior e pessoas que estudaram em escolas públicas.

Os resultados obtidos são preliminares e estimulam futuros estudos sobre o perfil dos cursos, impactos desse perfil na demanda e na nota de corte, bem como a produção de outros trabalhos que explorem todo o potencial dos bancos de dados utilizados nessa pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

COSTA, B. S. P. **Medidas de desigualdade:** uma análise de segregação socioespacial na Área Metropolitana de Brasília. 2015.

CROWDER, M.J. & HAND, D.J. **Analysis of repeated measures**. London: Chapman & Hall, 1990, 256p.

DIGGLE, P.J. An approach to the analysis of repeated measurements. **Biometrics**, v.44, p. 959-971, 1988.

FERREIRA, W.L. **Análise de dados com medidas repetidas em experimento com a ingestão de café**. 2012. 51f. Trabalho de conclusão de curso (Dissertação) – Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Minas Gerais.2012.

FREITAS, A.R. *et al.* Alternativas de análise de dados de medidas repetidas de bovinos de corte. **Revista Brasileira de Zootecnia**, v.34, n.6, p.2233-2244, 2005 (supl.)

HOX, J. J. **Multilevel analysis**: techniques and applications. 2. ed. New York: Routledge, 2010. 382 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. New York: Prentice Hall, 2007. 773 p.

LIMA, C.G. **Análise de dados com medidas repetidas**. 12f. Universidade de São Paulo.

QUEIROZ, G. **Métodos e interpretação de modelos estatísticos de análise de medidas repetidas:** uma aplicação a ensaio clínico. 2012. 132 p. Dissertação (Mestrado em ciências) - Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro.

RAUDENBUSH, S, & BHUMIRAT, C. The distribution of resources for primary education and its consequences for educational achievement in Thailand. **International Journal of Educational Research**,17, 143-164 p.2012.

SANTOS, M.P. *et al*, **Análise de medidas repetidas para avaliar o desempenho educacional de alunos do ensino fundamental**. 5f. (Resumo expandido). Universidade Estadual da Paraíba, Paraíba.

SPYRIDES,MHC., *et al.* **Análise de dados com medidas repetidas**. In: KAC, G., SICHIERI,R., and GIGANTE, DP., orgs. *Epidemiologia nutricional* [online]. Rio de Janeiro: Editora FIOCRUZ/Atheneu,2007,pp. 245-260, ISBN 978-85-7541-320-3. Disponível em: http://books.scielo.org. Acesso em: 29 Set. 2017