UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
MÉTODOS ESTATÍSTICOS COMPUTACIONAIS

ANÁLISE DE DADOS AMOSTRAIS COMPLEXOS DA PESQUISA PROALFA DE 2010 MINAS GERAIS

Mariana Verbena Casella
Patrícia Rezende De Almeida
Orientador: Marcel de Toledo Vieira

Juiz de fora
2011

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
MÉTODOS ESTATÍSTICOS COMPUTACIONAIS

ANÁLISE DE DADOS AMOSTRAIS COMPLEXOS DA PESQUISA PROALFA 2010 DE MINAS GERAIS

Monografia apresentada por MARIANA VERBENA CASELLA e PATRÍCIA REZENDE DE ALMEIDA ao Departamento de estatística da UFJF como parte dos requisitos para obtenção do título de Especialista em Métodos Estatísticos Computacionais. Orientador: MARCEL DE TOLEDO VIEIRA (Doutor em Estatística – Universidade de Southampton)

Juiz de Fora
2011

MARIANA VERBENA CASELLA
PATRÍCIA REZENDE DE ALMEIDA

ANÁLISE DE DADOS AMOSTRAIS COMPLEXOS DA PESQUISA PROALFA 2010 DE MINAS GERAIS

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA

APROVADA EM: ___/___/___.

_____

Prof. Dr. Marcel de Toledo Vieira

_____

_____

AGRADECIMENTO

Agradecemos, primeiramente, a Deus, norteador de nossas vidas; à equipe do curso de Métodos Estatísticos Computacionais pela formação de qualidade; ao Centro de Avaliação e Políticas Públicas da Educação (CAEd), instituição vinculada à Universidade Federal de Juiz de Fora (UFJF), por nos ter fornecido os dados necessários para a execução deste trabalho; ao Iago pela inquestionável e ágil colaboração; e, em especial, ao professor e amigo Marcel que nos acolheu como orientador e sem cuja dedicação incomensurável e paciência memorável não seria possível concretizarmos essa jornada.

Agradeço à minha mãe, meu pai e meu irmão pelo amor e por serem tão especiais em minha vida. Ao meu irmão, especialmente pelos momentos de alegria, meu pai pelo apoio permanente e à minha mãe pelo incansável incentivo; à Patrícia, agradeço pela dedicação e empenho na elaboração desta monografia, pelo carinho de sempre e amizade verdadeira. Agradeço também aos meus novos amigos, essenciais nesta caminhada.

Agradeço a meu pai, já presente em outro plano espiritual. Agradeço também à meu esposo pelo amor, carinho e atenção e, sobretudo, pelo apoio e compreensão pelas horas empenhadas neste trabalho; à minha mãe pelo amor incondicional e incentivo constante à continuidade de meus estudos. Aos novos amigos e amigas, em especial àquela que esteve junto nesta empreitada, Mariana, sem cuja calma, companheirismo e empenho seria impossível prosseguir.

RESUMO

Este trabalho foi elaborado com o objetivo de avaliar a eficiência do plano amostral complexo adotado na seleção da amostra para a pesquisa do Programa de Avaliação da Alfabetização (Proalfa), o qual mensura, para as redes municipais e estaduais em Minas Gerais, o desempenho dos alunos em Língua Portuguesa. O plano amostral utilizado pelo Programa é complexo, em virtude de haver a estratificação da amostra por rede e por Superintendência Regional de Ensino, além de ser conglomerado em dois estágios, pois seleciona-se no primeiro estágio a escola e no segundo a turma de cada série a ser avaliada. Além disso, foram utilizados nesta monografia apenas os dados da pesquisa para o 4º ano do ensino fundamental, possibilitando a comparação dos resultados obtidos para o Proalfa 2010 com os dos três anos anteriores (2007, 2008 e 2009) realizados por Cunha (2010). Os cálculos realizados para tais comparações foram os de média, erro padrão, intervalo de confiança e, posteriormente, o ajuste de modelos de regressão. O trabalho realizado leva à constatação de que quando o processo de estimação desconsidera o plano amostral complexo, há a subestimação dos erros padrão e, conseqüentemente, o estreitamento dos intervalos de confiança e a redução da precisão das estimativas. A conglomeração em dois estágios e a estratificação interferem nas médias, nos intervalos de confiança e nos modelos de regressão encontrados. Uma observação a ser realizada é a de que o plano amostral do Proalfa 2010 é mais eficiente do que o dos três anos anteriores, novamente se comparado ao estudo de Cunha (2010), tendo em vista os valores de efeitos do plano amostral encontrados.

Palavras-chave: Proalfa, Plano Amostral Complexo, Média, Desvio Padrão, Intervalo de confiança, Modelos de Regressão e Efeitos do Plano Amostral.

LISTA DE FIGURAS

LISTA DE TABELAS

Pág.

SUMÁRIO

INTRODUÇÃO

A definição das políticas públicas para melhoria do sistema educacional brasileiro tem sido baseada em pesquisas estatísticas, com o fulcro de avaliar o desempenho médio dos alunos, considerando não só as características pessoais dos mesmos, como também a influência tanto de suas turmas quanto das escolas onde estão matriculados.

Dentre estes estudos, há o Sistema de Avaliação da Educação Básica (SAEB), instituído por meio da Portaria nº. 931/2005 e composto pela Avaliação Nacional do Rendimento Escolar (Anresc), divulgada como "Prova Brasil" e cuja abrangência é maior que o outro programa integrante do SAEB, que é a Avaliação Nacional da Educação Básica (Aneb), conduzida por amostragem. Cabe ressaltar que este último recebe o nome de SAEB na divulgação.

O Programa de Avaliação da Alfabetização (Proalfa), por sua vez, faz parte do Sistema Mineiro de Avaliação (SIMAVE). Ele avalia o desempenho em Língua Portuguesa dos alunos de redes estadual e municipal em fase de alfabetização. As avaliações efetuadas pelo Proalfa são realizadas de duas formas: amostral e censitária. A primeira é realizada nos 2º e 4º anos do ensino fundamental e é utilizada para gerar indicadores de alfabetização. A censitária é realizada no 3º ano e identifica o nível de alfabetização de cada aluno, possibilitando a intervenção na aprendizagem quando necessário (VIEIRA & SOUZA, 2010).

As principais justificativas para o Proalfa ser realizado por amostragem são: restrições orçamentárias dos órgãos financiadores, redução da carga de coleta para a obtenção dos dados, além da possibilidade de um nível de precisão aceitável ser alcançado. Esta pesquisa adota um plano amostral complexo, que pode ser assim classificado por envolver critérios de seleção como estratificação, conglomeração, probabilidades desiguais de seleção, entre outras características (PESSOA & NASCIMENTO SILVA, 1998).

Esta monografia tem como objetivo a avaliação da eficiência do plano amostral do Proalfa de 2010, que pode ser descrito como um Plano Amostral Conglomerado em dois estágios, com a seleção da escola no primeiro estágio e, no segundo estágio, a seleção das turmas a serem avaliadas de acordo com a série de interesse. Como a seleção foi realizada de maneira independente nas duas séries pesquisadas, a mesma escola poderá ser selecionada duas vezes. Respeitando níveis de precisão desejados e também de forma a considerar as possíveis perdas devido à ausência de aluno no dia da avaliação, à falta de interesse em

participar, dentre outros fatores, o tamanho da amostra ficou definido como 50.000 alunos para cada série, totalizando 100.000 alunos. Esta definição foi realizada conjuntamente pela equipe de amostragem, pela Secretaria de Estado da Educação (SEE) e pelo Caed/UFJF.

Em pesquisas educacionais, a estrutura hierárquica desta população e a existência de correlação entre alunos de uma mesma turma e/ou escola fazem com que as observações coletadas para integrar a amostra não sejam independentes, que constitui o pressuposto básico para aplicação de toda a inferência clássica, desde estimação de parâmetros como médias, como de suas variâncias, até o ajuste de modelos paramétricos.

Apesar da importância inquestionável das interpretações sob o ponto de vista educacional, esta monografia tem objetivos metodológicos, porquanto pretende avaliar o impacto de se ignorar o plano amostral complexo utilizado sobre as estimativas de médias, erros padrão e intervalos de confiança associados, bem como sobre os coeficientes de modelos de regressão.

Esta monografia está estruturada como a seguir: o capítulo 1 aborda o conceito de plano amostral, a distinção entre os planos ignoráveis e não ignoráveis, a descrição do plano amostral do Proalfa e de sua população alvo; o capítulo 2 descreve a metodologia utilizada na estimação de médias, erro padrão e ajuste de modelos e uma breve discussão acerca dos efeitos do plano amostral. No terceiro capítulo, são apresentados os resultados e suas interpretações sob o ponto de vista estatístico. Finalmente, a conclusão do trabalho e as referências utilizadas são apresentadas.

1 PLANO AMOSTRAL

Por questões de economicidade, de logística para coleta de dados e até mesmo de tempo de coleta, os estudiosos e agências oficiais realizam pesquisas estatísticas através da obtenção de amostras, evitando o uso de censos sempre que possível. Neste sentido, levantamentos com objetivo de inferir sobre a população alvo exigem a adoção de planos amostrais probabilísticos que garantam uma probabilidade não nula de seleção para todos os elementos da população, bem como a cada amostra possível, há uma probabilidade *p(s)* de seleção calculável. Assim, denotamos uma população de tamanho fixo definida por meio dos rótulos U= (*1,2,...., N*) e uma amostra s, com s contida no espaço U, s= (k_1, k_2, ..., k_n), cuja probabilidade de seleção é *p(s)*. De fato, é essa distribuição de probabilidades associada a cada amostra possível de ser selecionada, responsável por definir o planejamento amostral (BOLFARINE & BUSSAB, 2005).

Consoante Pessoa & Nascimento Silva (1998), diz-se que um plano amostral é informativo quando o mecanismo de seleção de unidades amostrais pode depender dos valores das variáveis de pesquisa como, por exemplo, nos estudos de caso-controle, em que, muitas vezes, algumas unidades só irão compor a amostra por conterem determinada característica, mais especificamente, ser caso ou controle. Para a realização deste trabalho, serão considerados os planos não-informativos, segundo os quais a seleção de unidades para a amostra não está relacionada com os valores da variável de interesse.

1.1 Distinção entre planos ignoráveis e não ignoráveis

Dentre os planos não-informativos, cabe ressaltar a diferença existente entre planos ignoráveis e não ignoráveis. O único que atende a todas as condições necessárias e suficientes para ser considerado ignorável é o de amostragem aleatória simples com reposição (A.A.S.C.) que, inclusive, é base para toda a inferência estatística clássica. Neste caso, os resultados obtidos diretamente a partir do modelo paramétrico, que rege a população infinita, serão os mesmos que aqueles resultantes da inferência a partir da população finita (SUGDEN & SMITH, 1984; VIEIRA, 2001).

Os planos amostrais complexos, que envolvem desde estratificação, conglomeração, seleção de unidades com probabilidades diferenciadas até mesmo a própria estrutura da

população infinita são, geralmente, não ignoráveis. Pretende-se ilustrar, através da estimação de médias populacionais e do ajuste de modelos de regressão, a necessidade de se incorporar o plano amostral do Proalfa no processo de inferência estatística, já que ele é não ignorável.

1.2 Plano amostral e população alvo do Proalfa 2010

Enquanto a população alvo é aquela para a qual se deseja inferir, a população amostrada corresponde àquela disponível para a coleta de dados que irão compor a amostra. Essa última é definida de acordo com o plano amostral, responsável por definir a forma de seleção desses dados (BOLFARINE & BUSSAB, 2005).

O Proalfa é um dos programas que integram o Sistema Mineiro de Avaliação (SIMAVE), cujo principal objetivo é mensurar o desempenho dos alunos devidamente matriculados nos 2.º, 3.º e 4.º anos do ensino fundamental de escolas da rede municipal e estadual do estado de Minas Gerais. Possui caráter censitário e amostral, porquanto a pesquisa feita com a fase II do ensino fundamental considera o universo de todos os alunos matriculados. Já para as fases I e III considera um plano amostral complexo, sendo este o alvo da inferência a ser implementada no presente trabalho (VIEIRA & SOUZA, 2010).

Desta forma, a população alvo da pesquisa é formada pelos alunos das fases I e III do ensino fundamental matriculados em escolas mineiras, municipais e estaduais. A população amostrada se baseou no banco de dados, elaborado pelo Centro de Avaliação e Políticas Públicas da Educação (CAEd), instituição vinculada à Universidade Federal de Juiz de Fora (UFJF), responsável por realizar a pesquisa para o período 2006-2010. Deste modo, a população de referência é a que se segue:

Tabela 1: Número de escolas, número de turmas e número de alunos matriculados, segundo região e dependência administrativa, para a Fase I.

Região	Dependência Administrativa	Escolas		Turmas		Alunos	
		Total	%	Total	%	Total	%
Urbana	Estadual	1.818	21,5	3.796	24,9	87.975	32,7
	Municipal	2.368	28,0	6.363	41,8	141.765	52,6
Rural	Estadual	357	4,2	495	3,2	5.375	2,0
	Municipal	3.907	46,2	4.583	30,1	34.231	12,7
Total		8.450	100,0	15.237	100,0	269.346	100,0

Fonte: VIEIRA & SOUZA, 2010

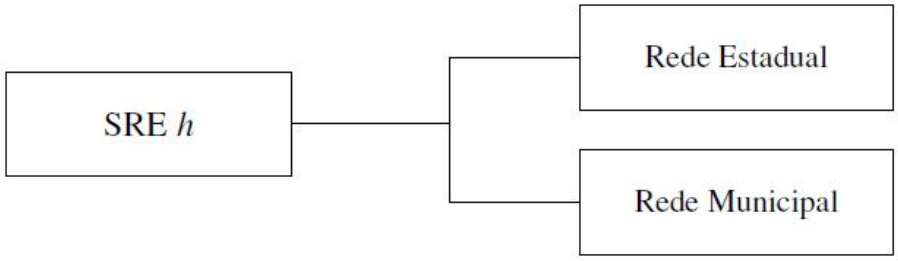
Tabela 2: Número de escolas, número de turmas e número de alunos matriculados, segundo região e dependência administrativa, para a Fase I.

Região	Dependência Administrativa	Escolas		Turmas		Alunos	
		Total	%	Total	%	Total	%
Urbana	Estadual	1.884	22,2	4.430	27,9	110.310	36,0
	Municipal	2.343	27,7	6.361	40,1	152.526	49,8
Rural	Estadual	357	4,2	500	3,2	6.524	2,1
	Municipal	3.889	45,9	4.563	28,8	37.058	12,1
Total		8.473	100,0	15.854	100,0	306.418	100,0

Fonte: VIEIRA & SOUZA, 2010

Semelhante ao PROALFA 2009, o plano amostral utilizado em 2010 considera uma amostragem aleatória estratificada por conglomerados em dois estágios, sendo as escolas as unidades primárias de amostragem (UPA's) e as turmas como as unidades secundárias de amostragem (USA's). Todos os alunos presentes no dia da avaliação que pertencem às turmas selecionadas irão compor a amostra. O primeiro fator de estratificação foi a série, havendo, também, a estratificação adicional definida a fim de publicar resultados separados para alguns domínios de interesse. Sendo assim, para cada grande estrato de interesse (série), estratificou-se segundo a área de abrangência das Superintendências Regionais de Ensino (SREs) e conforme a rede de ensino à qual a escola está vinculada, como na figura a seguir (VIEIRA & SOUZA, 2010).

Figura 1: Níveis de Estratificação

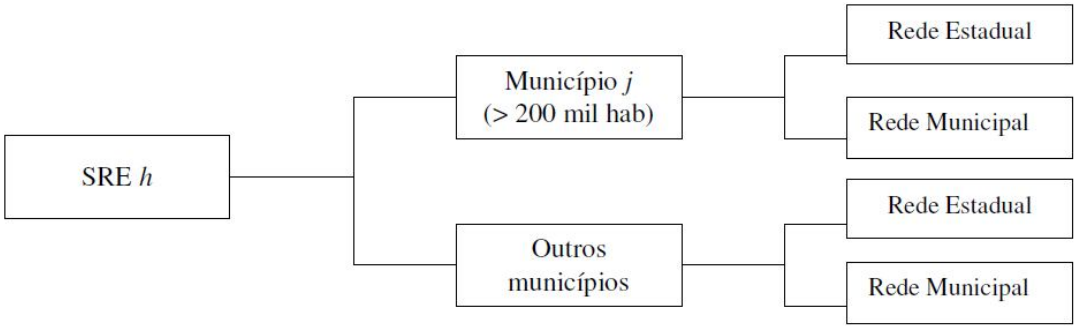


Fonte: VIEIRA & SOUZA, 2010

Outro fator de estratificação que permitiu a produção de estimativas para cada estrato foi imposto pela Secretaria Estadual de Educação de Minas Gerais (SEE-MG), de maneira que houvesse a "produção de estimativas para todos os municípios com população acima de

200 mil habitantes (estimada em 2006), garantindo ainda a existência de, pelo menos, dois municípios por pólo" (VIEIRA & SOUZA, 2010, p. 4-5), como a seguir descrito:

Figura 2: Níveis de Estratificação (SREs com municípios incluídos)



Fonte: VIEIRA & SOUZA, 2010

Assim, o total de estratos gerado foi de 130, com 50.000 alunos em cada série, que constitui uma amostra substancialmente superior às presentes no Proalfa 2007 e 2008. Adicionalmente, o tamanho total da amostra de conglomerados foi de 2.386 e 2.210 escolas a serem selecionadas nas fases I e III, respectivamente. A definição do tamanho de amostra foi feita de forma a evitar "possíveis perdas esperadas devido à ausência de alunos no dia de avaliação, recusa em participar da pesquisa e outros motivos que poderiam levar a uma redução do tamanho desejado da amostra de alunos" (VIEIRA & SOUZA, 2010, p. 6). Além disso, na alocação da amostra em cada um dos estratos de interesse, buscou-se um nível de precisão mínimo semelhante para estimação da proficiência média em Língua Portuguesa para cada um destes domínios.

Uma distinção adicional entre as pesquisas implementadas em 2009 e 2010 é que, enquanto, em 2009, dividiram-se as escolas nos dois grupos: com uma ou duas turmas da série; e com três ou mais turmas da série de interesse; em 2010, a subdivisão consistiu em: escolas com até três turmas da série, sendo selecionada apenas uma; e escolas com quatro ou mais turmas, com seleção de duas, com objetivo de elevar a precisão das estimativas através de um maior espalhamento da amostra de alunos.

No processo de seleção de escolas para cada um dos sub-estratos de tamanho, foi aplicado um método de amostragem com probabilidades proporcionais ao tamanho (p.p.t.) das escolas, que consiste na Amostragem Seqüencial de Poisson (OHLSSON, 1998). Como variável *proxy* do tamanho da escola, considerou-se o número de alunos matriculados na série

de interesse em cada escola. Este procedimento permitiu o ganho de eficiência, sem, contudo, elevar os custos de coleta. (VIEIRA & SOUZA, 2010)

2 METODOLOGIA

A aplicação dos métodos apresentados a seguir, foi realizada em dois estágios: (i) uso do software SPSS (Statistical Package for Social Sciences), versão 13.0 para implementação da abordagem estatística clássica (sem a devida consideração do plano amostral), bem como o pacote *Complex Samples*, que considera o plano amostral em suas estimativas; (ii) uso do software livre R versão 2.10.1 (2009), com o auxílio do pacote *Survey* versão 3.22-1 (LUMLEY, 2010), que também permite a consideração do plano amostral em suas estimativas.

A escolha dos pacotes estatísticos adotados foi feita tanto com base na interface gráfica, de fácil entendimento, a exemplo do SPSS, como também no adequado suporte e disponibilidade de manuais e documentação, além da importação de arquivos de diversas extensões (CARLSON, 1998).

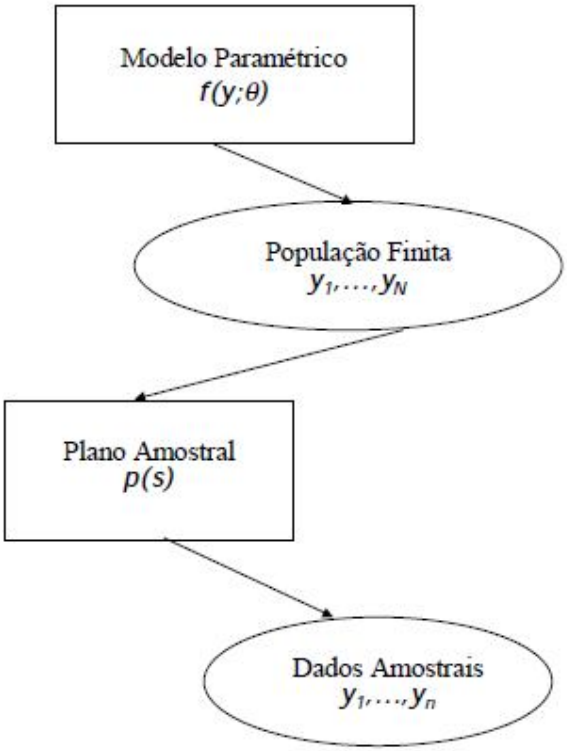
2.1 Definições básicas e modelagem

A inferência estatística se subdivide em duas correntes: (i) modelagem clássica, cujo objetivo é estimar o parâmetro indexador θ de uma superpopulação teórica a partir de uma amostra dela extraída, cujas observações são realizações de variáveis aleatórias independente e identicamente distribuídas (I.I.D.), que têm uma certa função densidade de probabilidade ou distribuição de freqüências; (ii) abordagem probabilística que, apesar de ter a inferência restrita à população finita a que se refere, tendo como parâmetros as quantidades descritivas populacionais (QDP's), tem a vantagem de ser não-paramétrica e, portanto, não supor uma distribuição paramétrica para as observações da amostra. Neste caso, o plano amostral, *p(s)*, é parte essencial da inferência, porque é através dele que são obtidas as observações amostrais (VIEIRA, 2001).

Enquanto a primeira simplesmente ignora a importância do plano amostral e suas conseqüências, a segunda não só procura considerá-lo como também, o próprio esquema de seleção, a fim de prover estimativas de totais, médias, razões. Há ainda uma terceira corrente que busca incorporar aspectos de ambas, que é a modelagem de superpopulação, em que se supõe que há uma 'superpopulação' infinita, que segue um modelo paramétrico, cujo parâmetro indexador é alvo de estimação. Porém, a mesma não é feita de maneira direta, na

medida em que a partir de um plano amostral *p(s)*, tem-se uma população finita, cujas observações são realizações das variáveis aleatórias I.I.D. Y_1, ..., Y_N. Os dados amostrais y_1, ...,y_N são usados para inferir sobre funções $g(Y_1, ..., Y_N)$, associadas ao parâmetro de interesse da superpopulação, com a aplicação de métodos usuais da primeira abordagem, ligada aos estatísticos modelistas (PESSOA & NASCIMENTO SILVA, 1998).

Figura 3: Esquema da modelagem de Superpopulação



Fonte: PESSOA & NASCIMENTO SILVA, 1998.

De acordo com Binder & Roberts (2006), a amostra final, que é utilizada para fazer inferências a respeito do parâmetro que descreve a superpopulação, é obtida em duas fases, sendo a primeira aquela em que a população finita é gerada a partir do modelo paramétrico e a segunda consiste na seleção dos elementos que irão compor a amostra, por meio do plano amostral *p(s)*, a partir desta população finita. De fato, diz-se que o plano amostral é ignorável quando a amostra é dita como diretamente proveniente deste modelo paramétrico e, portanto, a incorporação do plano amostral no processo inferencial em nada contribui para melhorar a eficiência e precisão dos estimadores.

Em situações em que não há um estimador ótimo do parâmetro θ, busca-se um estimador na classe de estimadores consistentes com o plano amostral, ou seja, que tenha "design consistency". Consistência sob a ótica de superpopulação se refere à possibilidade de tanto amostra quanto população finita manterem suas estruturas quando de seu crescimento (PFEFFERMAN, 1993).

À medida que ficava clara a necessidade de se incorporar o plano amostral ao processo de inferência, pensou-se em considerar uma variável regressora X_3 na ótica dos modelos clássicos de regressão, como se segue: $E(X_1/X_2, X_3) = \mu_1 + \beta_{1\,2.3}(X_2 - \mu_2) + \beta_{1\,3.2}(X_3 - \mu_3)$, que, todavia, levava à obtenção de estimadores inconsistentes. Na verdade, nos modelos de superpopulação, o planejamento amostral não é tido como variável causal, sendo incorporado no modelo como já explicitado, o que equivale à não existência do termo $\beta_{1\,3.2}(X_3 - \mu_3)$ (NATHAN & HOLT, 1980).

Nesta perspectiva, existem algumas fontes de variabilidade nos dados entre as quais: processo de medição, associado à medidas de repetitibilidade e reprodutividade, muito usadas em controle de qualidade; o mecanismo de resposta, com toda uma metodologia já desenvolvida a fim de contornar os problemas de não-resposta. Por não estarem diretamente relacionadas ao presente trabalho, estas duas fontes não serão consideradas. Aquelas que realmente explicam a incerteza subjacente à inferência aqui estudada são: o planejamento amostral, que consiste no mecanismo de seleção de unidades amostrais; e o próprio modelo de superpopulação, responsável por gerar as unidades que compõem a população finita a ser estudada. (PESSOA & NASCIMENTO SILVA, 1998).

Há que se distinguir entre duas abordagens existentes para tratar os dados provenientes de planos amostrais complexos: a agregada e a desagregada. Enquanto a primeira considera a complexidade do plano amostral e mesmo a estrutura populacional como fator complicador no cálculo de estimadores e derivação de modelos, buscando adaptar a teoria clássica inferencial, com a utilização de estimadores ponderados, entre outras ferramentas; a segunda altera os objetivos da análise estatística, sendo que as características anteriormente mencionadas servem de evidência fática de que modelos simples e procedimentos padrões, assim como pacotes estatísticos, são inadequados.

Na análise desagregada, a estrutura da população, efeitos de conglomeração, estratificação são considerados de maneira explícita a fim de explicar melhor a relação entre as variáveis (PESSOA & NASCIMENTO SILVA, 1998). Para exemplos de aplicações comparativas entre as duas modelagens, sugere-se a consulta ao trabalho de Skinner & Vieira (2004).

2.2 Médias, erros padrão e intervalos de confiança

A estimação das médias populacionais das proficiências em cada estrato h será feita de duas formas, de maneira semelhante à estratégia adotada por Cunha (2010): (i) Amostragem estratificada simples (AES) - com seleção dos elementos que irão compor o estrato através de amostragem aleatória simples, ou seja, tendo igual probabilidade de seleção e, conseqüentemente, aplicação da inferência clássica dentro de cada estrato; (ii) considerando-se o plano amostral, através do uso dos estimadores de Horvitz-Thompson, ou π- ponderados.

No primeiro caso, os parâmetros são estimados por:

$$\bar{x}_h = \sum_{i=1}^{n_h} \frac{x_{hi}}{n_h}$$

Nesta fórmula, $\bar{x}_h$ representa a média amostral da proficiência em Língua Portuguesa do *h*-ésimo estrato, x_{hi} a proficiência do *i*-ésimo aluno do *h*-ésimo estrato e n_h o número total de alunos avaliados no *h*-ésimo estrato. Para a amostragem estratificada simples – AES, que é o primeiro caso, pode-se afirmar que (BOLFARINE & BUSSAB, 2005):

$$\bar{x}_h \sim N\left(\mu_h, \frac{\sigma_h^2}{n_h}\right)$$

Na qual μ_h é a média populacional da proficiência do *h*-ésimo estrato e σ_h^2 a variância populacional da proficiência do *h*-ésimo estrato. Esta relação é válida sobretudo para amostras de tamanho suficientemente grande. O erro padrão deste estimador é dado, aproximadamente, por σ_h/n_h, podendo ser estimado por s_h/n_h, onde s_h é um estimador de σ_h:

$$s_h = \sqrt{\sum_{i=1}^{n_h} \frac{(x_{hi} - \bar{x}_h)^2}{n_h - 1}}$$

Onde μ_h é a média populacional da proficiência do *h*-ésimo estrato e σ_h^2 a variância populacional da proficiência do *h*-ésimo estrato. Esta relação é válida sobretudo para amostras de tamanho suficientemente grande. O erro padrão deste estimador é dado, aproximadamente, por σ_h/n_h, podendo ser estimado por s_h/n_h, onde s_h é um estimador de σ_h. Baseando-se nessas expressões, é possível construir os intervalos de confiança das médias das proficiências em cada estrato *h*, como se segue:

$$\left(\bar{x}_h - z_{\frac{\alpha}{2}}\frac{s_h}{\sqrt{n_h}} < \mu_h < \bar{x}_h + z_{\frac{\alpha}{2}}\frac{s_h}{\sqrt{n_h}}\right)$$

onde $z_{\alpha/2}$ provém da distribuição normal padrão. O nível de confiança (1-α), no presente trabalho, foi definido como sendo 95%. A segunda forma de estimação é a que leva em consideração o plano amostral para estimar as proficiências médias populacionais por meio dos estimadores de Horvitz-Thompson (HORVITZ & THOMPSON, 1952), como a seguir:

$$\bar{x}_h{}^{HT} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{x_{hi}}{\pi_{hi}}$$

Onde π_{hi} é a probabilidade de inclusão do *i*-ésimo aluno do *h*-ésimo estrato. O inverso desta probabilidade produz a fração ou peso amostral, que é uma forma de ponderar a estimação dos parâmetros. Estes pesos podem ser utilizados com o objetivo de permitir a consideração dos planos amostrais não ignoráveis, que podem levar ao viés de seleção, e também podem ser usados visando oferecer uma proteção no que diz respeito à especificação dos modelos paramétricos que descrevem a população infinita (KISH, 1990).

A produção de estatísticas, como médias e totais populacionais, compõem a estimação pontual. Contudo, para se ter clara visão a respeito da precisão das estimativas, é necessário adaptar o processo de estimação de variância dos próprios estimadores, viabilizando a construção de intervalos de confiança. Como não há forma direta de cálculo, existem dois possíveis métodos: Replicação, seja por Jackknife ou por Bootstrap; ou ainda a linearização de Taylor, que será aplicado nesta monografia por estar implementado nos pacotes estatísticos utilizados (VIEIRA, 2001).

O método de linearização de Taylor consiste na expansão em séries de Taylor do estimador pontual do parâmetro de interesse, em torno do verdadeiro parâmetro, considerando apenas os termos de primeira ordem. Nesta monografia, os parâmetros de interesse são os coeficientes de regressão (VIEIRA, 2001). Na estimação intervalar, os intervalos de confiança, que consideram a fração amostral, são construídos da seguinte maneira:

$$\left(\bar{x}_h{}^{HT} - z_{\frac{\alpha}{2}}EP_L\left(\bar{x}_h{}^{HT}\right) < \mu_h < \bar{x}_h{}^{HT} + z_{\frac{\alpha}{2}}EP_L\left(\bar{x}_h{}^{HT}\right)\right)$$

Onde $EP_L\ (\bar{x}_h{}^{HT})$ consiste no erro padrão de $\bar{x}_h{}^{HT}$ obtido pelo método de linearização de Taylor.

2.3 Modelos de regressão múltipla

Além da estimação de médias populacionais, buscou-se verificar o impacto de planos amostrais complexos nos estimadores de regressão. Um modelo de regressão múltipla pode ser representado por:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, i = 1,2, \dots, n$$

Onde, β_0 é o intercepto do modelo, β_j o *j*-ésimo coeficiente da regressão, x_{ij} denota a *i*-ésima observação da *j*-ésima variável explicativa, com *j* variando de 1 até *k*, em que *k* denota o número de covariáveis que compõem o modelo, ε_i constitui o resíduo do *i*-ésimo aluno, com *i* variando de 1 até *n*, sendo *n* o tamanho da amostra de alunos. Na forma matricial, o modelo é assim especificado (PINDYCK & RUBINFELD, 2004):

$$y = X\beta + \varepsilon$$

$$onde\ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} e\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Todavia, o ajuste deste modelo está condicionado ao atendimento de alguns pressupostos, quais sejam: os erros ou resíduos do modelo são independente e identicamente distribuídos – I.I.D., tendo média nula e variância constante, esta última também reconhecida como hipótese da homocedasticidade; os erros também seguem uma distribuição normal padrão de média nula e variância constante σ^2. Além disso, a hipótese da não existência de multicolinearidade se refere à independência entre as variáveis explicativas x_{ij}. Assim, as observações y_i são normalmente e independentemente distribuídas com variância constante σ^2 e média condicional dada por $\beta_0 + \Sigma\ \beta_j\ x_{ij}$, com o somatório feito para *j* variando de 1 a *k*. (PINDYCK & RUBINFELD, 2004).

Na estimação dos coeficientes do modelo, existem diversos métodos, tais como o de máxima verossimilhança, o de momentos e o de mínimos quadrados ordinários. Para o modelo clássico de regressão linear simples e múltipla, considera-se que o estimador de mínimos quadrados ordinários é o melhor, ou seja, o mais consistente e sem viés, desde que sejam respeitados os seus pressupostos e que as observações tenham igual probabilidade de seleção (KMENTA, 1988; PINDYCK & RUBINFELD, 2004).

O estimador de MQO em sua forma matricial é como:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Sendo que X' corresponde à matriz transposta de X e $(X'X)^{-1}$ é a inversa de (X'X). O erro padrão deste estimador é a raiz quadrada dos termos da diagonal da matriz dada por $(X'X)^{-1}\,\hat{\sigma}^2$, em que $\hat{\sigma}^2$ é um estimador de σ^2 (KMENTA, 1988).

Sob normalidade, o estimador de máxima verossimilhança produz resultados próximos aos de mínimos quadrados ordinários. Objetivando levar em consideração o plano amostral utilizado, que, geralmente, rompe com o pressuposto de que as observações são IID, foi proposto na literatura o estimador de máxima pseudo-verossimilhança (MPV), que considera os pesos amostrais (w_i), neste caso como o inverso da probabilidade de inclusão de cada elemento *i* na amostra (π_i), na estimação dos parâmetros do modelo. Resolvendo-se as funções de verossimilhança e de log-verossimilhança, tem-se o estimador de MPV:

$$\hat{\beta}_{MPV} = (X'WX)^{-1}X'Wy$$

$$onde\, W = diag\,[(w_1, \ldots, w_n)]$$

Conforme Pessoa & Nascimento Silva (1998), este estimador se assemelha ao de mínimos quadrados ponderados (MQP), sendo o peso definido por $w_i = \pi^{-1}$. Porém, enquanto este último é útil quando há heterocedasticidade dos resíduos e, para contornar este problema, aplica os referidos pesos; o estimador MPV busca incorporar o planejamento amostral na inferência. Além disso, "a justificativa para se adotar o MPV não se baseia na otimalidade tal como é o caso do MQO e MQP", mesmo que se mantivessem os pressupostos de IID na seleção da amostra.

Do ponto de vista da estimação intervalar, a variância assintótica deste estimador para β incorpora os pesos amostrais e as demais característica do plano amostral, sendo obtida com a aplicação do método de linearização de Taylor e cuja expressão é:

$$V_L(\hat{\beta}_{MPV}) = (X'X)^{-1}V\left(\sum_{i=1}^{n} w_i x_i \varepsilon_i\right)(X'X)^{-1}$$

Onde $V(\sum_{i=1}^{n} w_i x_i \varepsilon_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} \varepsilon_i x_i x_j{}' \varepsilon_j$ com π_{ij} é a probabilidade de inclusão conjunta dos alunos i e j. Esta variância será estimada como a seguir: (VIEIRA, 2001):

$$\hat{V}_L(\hat{\beta}_{MPV}) = (X'WX)^{-1}\hat{V}\left(\sum_{i=1}^{n} w_i x_i \varepsilon_i\right)(X'WX)^{-1}$$

Onde $V(\sum_{i=1}^{n} w_i x_i \varepsilon_i) = \sum_{i=1}^{n}\sum_{j=1}^{n}(w_i w_j - w_{ij})\,\hat{\varepsilon}_i x_i x_j{}' \hat{\varepsilon}_j$. Para maiores detalhes sobre o estimador de MPV, é interessante consultar os trabalhos de Pessoa & Nascimento Silva (1998) e Vieira (2001). Uma comparação entre os três métodos de estimação, mínimos quadrados ordinários, máxima verossimilhança e π- ponderado está presente em Holt *et al* (1980), sendo que o de máxima verossimilhança obteve resultados melhores quando as observações não têm igual probabilidade de seleção.

Nesta monografia, o nível de significância adotado é de 5%. Além disso, utilizou-se o método *backward* para construção dos modelos, nos quais se inserem todas as variáveis disponíveis, conforme as descritas no quadro a seguir:

Quadro 1 – Descrição das variáveis disponíveis no PROALFA 2010 para construção do modelo de regressão

VARIÁVEIS	DESCRIÇÃO
Proficiência	Proficiência em Língua Portuguesa
Gênero Feminino	Variável *dummy*. Gênero do aluno. Codificação: 0 – Masculino; 1 – Feminino.
Rede	Variável categórica. Rede à qual a escola pertence. Codificação: 1 – Estadual; 2 – Municipal.
Turno	Variável categórica. Turno em que as aulas são ministradas. Codificação: 1 - Manhã ; 2 - Tarde.
Idade	Idade do aluno no dia em que respondeu ao questionário. Codificação: 0 – 7 anos; 1 – 8 anos; 2 – 9 ou mais anos.
Projeto Escola Tempo Integral	Variável *dummy*. Indica se o aluno participa do projeto Escola Tempo Integral. Codificação: 0 – não; 1 – sim.

Fonte: Elaboração própria com base nos dados do CAEd

No questionário utilizado para a pesquisa, há ainda a data de nascimento informada pelo aluno, que não foi considerada no modelo por já existir a variável “Idade”. Semelhante à Cunha (2010), foram criados modelos de regressão iniciais e finais, sendo que os últimos são obtidos a partir dos primeiros, considerando apenas as variáveis estatisticamente significativas. Entretanto, nos modelos aqui construídos, o número de covariáveis foi menor, comparado às pesquisas implementadas em 2008 e 2009, estudadas por Cunha (2010).

2.4 Efeitos do plano amostral (EPA)

O erro padrão e os intervalos de confiança para os parâmetros estudados estão associados à precisão da estimação, tendo impactos sobre os testes de hipóteses e nível de confiança da inferência. Estas medidas sofrem diferentes efeitos em virtude do uso de planos amostrais complexos, em que as observações deixam de ser independentes e igualmente distribuídas (IID).

Muitos dos pacotes estatísticos fornecem estas estimativas, porém consideram estimadores que são não viesados e consistentes sob a hipótese de IID, ou seja, como se as unidades amostrais tivessem sido selecionadas por meio de amostragem aleatória simples com reposição. Com a adoção de planos complexos, os valores fornecidos nas saídas dos programas não correspondem à realidade, porque com este tipo de seleção as observações deixam de ser IID. Desta forma, há uma tendência de se subestimar o erro padrão, gerando intervalos de confiança mais estreitos e, por fim, interferindo nos resultados dos testes de hipóteses e no próprio nível de confiança do estudo, o qual pode ser inferior ao nominal (na presente monografia, 95 %).

2.4.1 Efeito do Plano Amostral de Kish:

O efeito do plano amostral (EPA) de Kish tem como objetivo a comparação de planos amostrais no estágio de planejamento da pesquisa. “O EPA de Kish é uma razão entre variâncias (de aleatorização) de um estimador, calculadas para dois planos amostrais alternativos” (PESSOA & NASCIMENTO SILVA, 1998, pag. 48)

$$EPA_{Kish}(\hat{\theta}) = \frac{V_{VERD}(\hat{\theta})}{V_{AAS}(\hat{\theta})}$$

Onde V_{VERD} é a variância induzida pelo plano amostral complexo que está sendo considerado (variância verdadeira) e V_{AAS} variância induzida pelo plano de amostragem aleatória simples. Se o resultado para o EPA_{Kish} for um valor elevado, significa que se deve considerar o plano amostral verdadeiro para calcular as estimativas, já que, sob a hipótese de AAS, subestimam-se as variâncias corretas. Valores unitários desta medida indicam que as estimativas geradas a partir do plano complexo e da AAS têm a mesma precisão.

A importância de utilizar os valores de EPA se deve ao fato de permitir a comparação, a antecipação do impacto gerado sobre a precisão dos estimadores, principalmente de variáveis relevantes, além de possibilitar o cálculo do tamanho da amostra segundo o nível de precisão desejado.

O EPA também é utilizado no cálculo do tamanho das amostras em situações em que o plano amostral é complexo. De acordo com Carlson (1998), o tamanho efetivo da amostra corresponde à quantidade de observações que devem compor a amostra complexa a fim de se ter idêntico nível de precisão àquele obtido por meio da AAS.

2.4.2 Efeito do Plano Amostral Ampliado:

Objetivando solucionar as dificuldades encontradas no EPA, foi criado o conceito do EPA Ampliado. O cálculo realizado a partir de hipótese “ingênua” (IID ou AAS) faz com que a variância das observações IID se afaste da variância do plano amostral verdadeiro. Para avaliar se o afastamento será grande ou pequeno, modifica-se a expressão de EPA para $EPA(\widehat{\theta}, \upsilon_0)$:

$$EPA\left(\widehat{\theta}, \upsilon_0\right) = \frac{V_{VERD}\left(\hat{\theta}\right)}{E_{VERD}(\upsilon_0)}$$

Onde $E_{VERD}(\upsilon_0)$ é o valor esperado do estimador da variância υ_0 do estimador $\hat{\theta}$. O resultado obtido a partir dessa fórmula indica a tendência de $\upsilon_0$subestimar ou superestimar a variância verdadeira de $\hat{\theta}$, ou seja, $V_{VERD}\left(\hat{\theta}\right)$. Quanto mais próximo de 1 (um), melhor será a especificação do modelo e/ou do plano amostral.

Ao se ignorar o plano amostral adotado para seleção de dados, e supor AAS, podem haver conseqüências como: inflacionar o EPA ao ignorar os pesos e/ou conglomeração em υ_0; e (ii) reduzir o EPA ao ignorar a estratificação em υ_0. Para que não haja análise incorreta dos dados, recomenda-se a estimação dos EPAs com o fim de verificar se há impactos relevantes

sobre a variância dos estimadores e, em última instância, sobre os resultados obtidos (PESSOA & NASCIMENTO SILVA, 1998).

O $EPA\left(\widehat{\theta},\, v_0\right)$ é mais abrangente do que o EPA de Kish por considerar o impacto sobre as variâncias causado pelo tipo de plano amostral utilizado e pela própria estrutura da população estudada. Em outros termos, o $EPA\left(\widehat{\theta},\, v_0\right)$ é capaz de avaliar o efeito do planejamento amostral e da especificação incorreta do modelo paramétrico subjacente à população infinita.

Um exemplo de investigação sobre esta medida está presente em Vieira, Salgueiro & Smith (2010), em que comprovam o maior efeito da não adoção de planos complexos e da estrutura populacional para estudos longitudinais, se cotejados com a análise transversal. Além disso, mostram que os $EPA\left(\widehat{\theta},\, v_0\right)$ relativos aos coeficientes de regressão tendem a ser de menores que aqueles das médias da variável dependente.

3 RESULTADOS E ANÁLISE

3.1 Médias

Nesta monografia, por simplificação, o problema dos dados faltantes não receberá tratamento. Sendo assim, inicialmente, a fim de tornar a base de dados Proalfa 2010, cujo total inicial de observações era de 46.282, livre dos mesmos, foram eliminados os casos com dados faltantes nas variáveis Gênero e Idade, o que resultou em 44.072 casos válidos, havendo uma perda de 4,77%. Como o objetivo maior deste trabalho é avaliar o impacto da consideração do plano amostral complexo sobre a estimação, foram calculadas para esta variável, a média, o erro padrão e o intervalo de confiança, esse último utilizando um nível de significância de 5%. Dois tipos de estimativas foram geradas: (i) considerando que o plano amostral era AAS, também conhecida como estimação clássica; e (ii) considerando as características do plano complexo verdadeiro. Em ambos os casos, inicialmente, considerou-se toda a amostra e, posteriormente, geraram-se resultados por rede e por regional.

Tabela 3: Média, erro padrão e intervalo de confiança para a variável Proficiência em Língua Portuguesa, considerando e não considerando o plano amostral

Ano	Considerando o Plano Amostral				Sem considerar o plano amostral			
	Média	Erro Padrão	IC-LI	IC-LS	Média	Erro Padrão	IC-LI	IC-LS
2010	586,254	1,089	584,118	588,389	589,975	0,401	589,189	590,760

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

A tabela 3 apresenta os valores de média, erro padrão e intervalo de confiança para a variável Proficiência considerando e não considerando o plano amostral. Ainda que a diferença entre o valor pontual da média estimada nos dois casos seja pequena, o intervalo de confiança e o erro padrão são menores quando não se considera o plano amostral do que quando se considera.

A partir da estratificação por rede, obteve-se os resultados para a rede estadual (CD_REDE1) na tabela 4, que representam 21.919 dos casos válidos, cujo total é de 44.072.

O valor encontrado para a média estimada, considerando e não considerando o plano amostral, está próximo.

Tabela 4: Média, erro padrão e intervalo de confiança para a variável Proficiência em Língua Portuguesa, considerando e não considerando o plano amostral por rede (Estadual)

Ano	Considerando o Plano Amostral				Sem considerar o plano amostral			
	Média	Erro Padrão	IC-LI	IC-LS	Média	Erro Padrão	IC-LI	IC-LS
2010	597,295	1,630	594,097	600,492	598,565	0,568	597,453	599,678

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Porém, ao levar em conta o plano amostral, o erro padrão e o intervalo de confiança são maiores do que quando este não é considerado. Comparado ao trabalho executado por Cunha (2010), o erro padrão para os dados de 2010 é menor do que os de 2007, 2008 e 2009, especialmente quando se considera o plano amostral complexo, o que pode estar indicando uma maior eficiência do plano amostral de 2010 em relação aos anos anteriores.

Tabela 5: Média, erro padrão e intervalo de confiança para a variável Proficiência em Língua Portuguesa, considerando e não considerando o plano amostral por rede (Municipal)

Ano	Considerando o Plano Amostral				Sem considerar o plano amostral			
	Média	Erro Padrão	IC-LI	IC-LS	Média	Erro Padrão	IC-LI	IC-LS
2010	579,368	1,446	576,538	582,203	581,475	0,560	580,377	582,572

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Na tabela 5, encontram-se os resultados apenas para a rede municipal (CD_REDE2) com 22.153 dos casos válidos. Na rede municipal, os resultados encontrados são semelhantes aos da rede estadual. O valor da média considerando e não considerando o plano amostral é próximo, o erro padrão e o intervalo de confiança são menores quando o plano amostral não é considerado. Novamente, comparando com o trabalho de Cunha (2010), o erro padrão foi menor na base de 2010, o que pode estar indicando uma maior eficiência deste plano amostral.

Para a realização de análises baseadas nas regiões de abrangência das secretarias regionais de ensino (SREs) existentes (definidas de acordo com a variável CD_REGIONAL)

e construção da tabela 6, foram selecionadas somente algumas regionais, pois cada rede possui 46 (quarenta e seis) regionais, perfazendo um total de 92 (noventa e duas) regionais para as duas redes (estadual e municipal). A partir da realização de testes de significância, identificou-se que todas as regionais eram significativas, porém optou-se por analisar somente as que possuíam maior erro padrão, quando considerado o plano amostral; assim as que apresentaram maiores valores foram as representadas pelos códigos 19, 24, 26, 29, 30, 51 e 53.

Tabela 6: Média, erro padrão e intervalo de confiança para a variável Proficiência em Língua Portuguesa, considerando e não considerando o plano amostral por rede (Estadual) e por regional

SRE	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	REDE 1 (Estadual)		REDE 2 (Municipal)		REDE 1 (Estadual)		REDE 2 (Municipal)	
	Média	Erro padrão	Média	Erro padrão	Média	Erro padrão	Média	Erro padrão
19	603,402	12,538	600,438	14,456	602,126	4,999	599,745	4,418
24	638,172	14,085	653,528	14,751	637,629	4,238	652,730	4,697
26	587,441	10,576	541,396	14,057	588,964	3,954	541,488	4,658
29	565,715	14,854	569,665	10,674	566,016	5,065	570,627	4,565
30	615,733	17,050	617,979	13,173	612,146	5,812	618,677	4,248
51	618,401	14,543	587,168	9,932	617,926	4,845	586,910	4,257
53	577,141	10,386	555,913	15,300	576,920	3,570	556,987	5,358

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Na tabela 6, é possível confirmar os resultados precedentes, já que não existem grandes diferenças nas estimativas pontuais para proficiência média em Língua Portuguesa. Todavia, percebe-se que os intervalos de confiança para as médias, calculados sob o plano ignorável, são menores que aqueles que adotaram o plano complexo da pesquisa Proalfa 2010 no processo de estimação. No entanto, elas não podem ser comparadas com o trabalho de Cunha (2010), pois não há como identificar se as SREs por ele consideradas são as mesmas apresentadas nesta monografia.

3.2 Modelos de regressão

3.2.1 Estimação de modelos

Inicialmente, buscou-se ajustar um modelo linear com a proficiência em Língua Portuguesa como variável dependente, utilizando como covariáveis não só o gênero, a idade e o projeto Escola Tempo Integral, como também o Turno, já que se verificou a diferença no rendimento médio dos alunos por turno. Nestes modelos iniciais, não se considerou, todavia, a distinção por rede de escola, ou seja, se estadual ou municipal. No entanto, a variável Turno não se mostrou estatisticamente significativa, dado o nível de significância adotado – 5%, tanto nos modelos sob o pressuposto de AAS quanto naqueles cuja estimação leva em conta o plano amostral complexo.

Assim, para os modelos que desconsideram a rede à qual cada escola está vinculada, optou-se por retirar a variável turno e, por fim, proceder ao ajuste de novos modelos a fim de explicar a proficiência.

Tabela 7: Modelo estimado para explicação da proficiência em Língua Portuguesa do Proalfa 2010, excluídos Fator Turno e Rede

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	574,587	3,154	568,401	580,772	582,183	0,574	581,057	583,309
Projeto Escola Tempo Integral (Sim)	23,520	3,206	17,234	29,806	-22,676	1,377	-25,376	-19,977
Gênero Feminino	-20,055	0,985	-21,987	-18,124	19,464	0,794	17,908	21,019
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	-5,698	48,684	-101,17	89,772	1,638	31,479	-60,062	63,337
Idade (8 anos)	18,413	4,663	9,269	27,558	18,545	3,692	11,309	25,782
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Na tabela 7, a partir da comparação dos erros padrão de cada coeficiente do modelo, percebe-se que, ao se desconsiderar o plano amostral complexo da pesquisa Proalfa 2010, os mesmos são subestimados, destacando-se o caso do intercepto e da variável “Projeto Escola em Tempo Integral”.

Em ambos os modelos, a variável Idade, na categoria 7 anos, não se mostrou estatisticamente significativa, tendo em vista que os intervalos de confiança para este coeficiente contém o zero, não sendo possível provar que o mesmo seja não nulo, fato confirmado pelos p-valores. Cabe destacar, contudo, que considerar o plano amostral na estimação implicou na inversão dos sinais de coeficientes para algumas variáveis como, por exemplo, gênero e Projeto Escola em Tempo Integral. No caso desta última, o sinal negativo do coeficiente para o modelo sob o pressuposto do A.A.S. não parece coerente com o senso comum, tendo em vista que se espera o melhor rendimento dos alunos que freqüentam a escola em período integral.

A despeito do propósito deste estudo não ser o melhor ajuste de modelos de regressão, considera-se que este fato possa estar associado a dois motivos: ao ajuste de um modelo único, que ignora a diferença na proficiência como resultado do tipo de rede à qual a escola pertence; e, por fim, porque, no questionário aplicado para formação da base de dados utilizada neste trabalho, a variável “Escola Tempo Integral” possuía apenas uma opção de resposta, ou seja, se o aluno participasse da escola em tempo integral ele marcaria, se não, deixaria em branco. Com isso, não há clara diferenciação entre o que seria não-resposta e o que seria a não participação deste projeto, optando-se por considerar, nesta monografia, a não sinalização como a não participação do aluno nesta modalidade de ensino, ao invés da existência de dados faltantes.

Assim, o ajuste de modelos que consideram o efetivo plano amostral complexo utilizado para obter a amostra, e, que, portanto, rompem com os pressupostos de AAS, levou à estimação de intervalos de confiança maiores para os coeficientes das variáveis independentes dos modelos de regressão, como decorrência de erros padrão superiores. Ainda que os modelos não difiram quanto à significância dos coeficientes, já que possuem as mesmas variáveis independentes, a inversão de sinais a elas associadas pode ser explicada pela má especificação do modelo. Apesar disso, na estimação clássica do modelo, o p-valor resultante do teste “Lack of Fit Test” foi de 0,307, mostrando que não há indícios para se rejeitar a hipótese nula de que o modelo esteja com especificação completamente incorreta. Na verdade, este teste avalia se o modelo ajustado, preferencialmente com especificação mais reduzida, que contenha apenas os efeitos principais, é significativo quando comparado com a

hipótese alternativa de que há necessidade de inclusão de outras covariáveis, sobretudo as de interação. (DRAPER & SMITH, 1998).

A construção destes modelos iniciais serviu como primeiro passo deste estudo, porquanto os resultados encontrados vão de encontro às conclusões presentes em outras pesquisas na área de educação. Como exemplo, a questão do Projeto Escola Tempo Integral já mencionada, bem como o fato da proficiência média em Língua Portuguesa ser inferior no caso de alunas em, aproximadamente, 20 pontos quando controladas as demais variáveis do modelo sob plano complexo. (ALBERNAZ, FERREIRA & FRANCO, 2002; PEREIRA, 2006). Isto porque estes modelos ignoram o fator Rede, que é uma divisão natural da amostra, ou melhor, faz parte da estrutura real sob a qual a população é organizada.

Como o coeficiente da variável "Projeto Escola Tempo Integral" apresentou, sob estimação clássica, sinal negativo, que é divergente do esperado, optou-se por gerar nova série de modelos cuja especificação não a contivesse. Nestes casos, ainda não se levou em conta a estratificação natural da amostra, ou seja, a Rede à qual cada escola pertence; porém, a variável Turno foi novamente incluída, com o fulcro de verificar sua significância sem a interferência da variável Projeto Escola Tempo Integral.

Tabela 8: Modelo estimado para explicação da proficiência em Língua Portuguesa do Proalfa 2010, com variável Turno e sem Projeto Escola Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	574,805	1,832	571,213	578,398	579,248	0,705	577,867	580,629
Turno Manhã	2,240	2,477	-2,616	7,097	1,512	0,797	-0,051	3,075
Turno Tarde	-	-	-	-	-	-	-	-
Gênero Feminino	20,171	0,993	18,223	22,118	19,579	0,796	18,019	21,140
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	-6,995	45,619	-96,455	82,466	0,603	31,575	-61,284	62,489
Idade (8 anos)	19,031	4,521	10,166	27,896	19,129	3,704	11,870	26,388
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Confirmando os resultados presentes na literatura especializada, consoante exposto na Tabela 8, os erros padrão são superiores quando se considera o plano amostral complexo, resultando em intervalos de confiança maiores para os coeficientes dos modelos ajustados.

Novamente, a significância dos coeficientes não sofreu intervenção como conseqüência da adoção do plano complexo na estimação, produzindo modelos com as mesmas variáveis, especificamente, com a variável Idade (7 anos) não sendo estatisticamente significativa a um nível de 5% de significância, o que, em última instância, não fornece indícios de que haja diferença na proficiência média dos alunos de 7 anos frente à categoria de referência – 9 ou mais anos.

Mesmo com a exclusão do "Projeto Escola Tempo Integral", os testes não forneceram indícios de que a variável Turno seja estatisticamente significativa ao nível de significância adotado, tendo em vista que os intervalos de confiança para este coeficiente contêm os valores nulos. Assim, não se pode dizer que a proficiência média seja influenciada pelo turno no qual os estudantes estão matriculados.

Ainda que não considerem a covariável Rede e, de certa forma, não tenham a melhor especificação para explicar a proficiência média em Língua Portuguesa, estes modelos estão condizentes com os resultados de outros estudos, já que o sinal positivo para Gênero em ambos os modelos confirma tendência comprovada de que alunas têm melhor rendimento em Língua Portuguesa, fato que não ocorria nos modelos iniciais ajustados. Além disso, alunos com 8 anos apresentam desempenho superior aos de 9 ou mais anos de idade, fator possivelmente explicado por estarem matriculados no 4.º ano do ensino fundamental antecipadamente. Adicionalmente, nos modelos sob AAS, não houveram indícios para se rejeitar a hipótese nula e concluir que o modelo esteja incorretamente especificado, porquanto o p-valor do "Lack of Fit Test" foi de 0,255.

A fim de verificar o impacto da variável Rede sobre a estimação dos modelos, a mesma foi incorporada a dois outros modelos como covariável, como pode ser visualizado na tabela 9. Novamente, os erros padrão estimados sob o plano amostral ignorável são menores, se comparados às estimativas calculadas sob o plano complexo. Em última instância, isto leva ao estreitamento dos intervalos de confiança dos coeficientes dos modelos de regressão clássicos, ou seja, que consideram os pressupostos de AAS.

A variável Idade (7 anos) permaneceu sendo estatisticamente não significativa a um nível de 5% de significância, tanto no modelo que considera como no que desconsidera o plano do Proalfa 2010. É importante mencionar que as interações entre os fatores Rede (municipal e estadual) e as categorias da variável Idade (7 anos, 8 e 9 ou mais) também não

compuseram estes dois modelos, por não serem estatisticamente significativas. Além disso, não houve inversão dos sinais dos coeficientes destes modelos para as variáveis significativas, fato possivelmente ligado à inclusão do fator Rede como covariável. Isto também pode ter contribuído para que a especificação do modelo sob AAS seja correta, ou seja, no "Lack of Fit Test" o p-valor foi de 0,139, superior a 5% de significância, não fornecendo evidências para a rejeição da hipótese nula.

Adicionalmente, a subestimação dos erros padrão dos coeficientes quando se ignorou a existência de plano amostral complexo não influenciou na significância dos coeficientes, já que os modelos são iguais. O sinal positivo da Rede em ambos os modelos evidenciou que alunos matriculados na rede estadual possuem rendimento superior, em média, 18 pontos frente àqueles da rede municipal. Apenas com a inclusão da variável Rede, que não necessariamente captou a estrutura hierárquica da população, já foi possível obter sinais de coeficientes para gênero que se coadunam com pesquisas anteriores. No entanto, permaneceram os problemas de sinais para a variável Projeto Escola Tempo Integral.

Tabela 9: Modelo estimado para explicação da proficiência em Língua Portuguesa do Proalfa 2010, com Fator Rede, excluso Fator Turno

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	571,167	1,550	568,128	574,207	573,338	0,684	571,998	574,678
Projeto Escola Tempo Integral(Sim)	-26,747	3,154	-32,933	-20,562	-26,481	1,378	-29,182	-23,779
Gênero Feminino	19,795	0,978	17,878	21,713	19,325	0,789	17,780	20,871
Gênero Masculino	-	-	-	-	-	-	-	-
Rede Estadual	18,989	2,170	14,732	23,245	18,667	0,794	17,110	20,223
Idade (7 anos)	-7,044	45,092	-95,469	81,382	0,458	31,284	-60,859	61,775
Idade (8 anos)	16,499	4,572	7,534	25,465	16,573	3,670	9,379	23,766
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Objetivando verificar se o Turno influenciará a proficiência média em Língua Portuguesa, tendo sido inclusa a variável Rede na estimação dos modelos, nova série de modelos foi ajustada, considerando todas as variáveis disponíveis, dispostas na tabela 10.

Tabela 10: Modelo estimado para explicação da proficiência em Língua Portuguesa do Proalfa 2010, com Fator Rede e Turno, inclusa covariável Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	567,382	2,241	562,987	571,777	569,475	0,864	567,782	571,168
Projeto Escola Tempo Integral (Sim)	-26,619	3,187	-32,870	-20,369	-26,553	1,377	-29,253	-23,853
Gênero Feminino	19,827	0,978	17,910	21,745	19,365	0,788	17,820	20,910
Gênero Masculino	-	-	-	-	-	-	-	-
Rede Estadual	20,154	2,247	15,748	24,559	20,074	0,817	18,474	21,675
Turno Manhã	5,755	2,494	0,864	10,646	5,944	0,813	4,352	7,537
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	-6,458	43,630	-92,016	79,100	0,952	31,266	-60,329	62,232
Idade (8 anos)	16,655	4,504	7,824	25,487	16,829	3,668	9,640	24,019
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Os resultados corroboram o fato de que o plano amostral complexo produz erros padrões de coeficientes maiores que os obtidos com o plano ignorável, o que, por fim, leva a

intervalos de confiança de maior amplitude. Os modelos possuem a mesma estrutura; destarte, o plano amostral do Proalfa 2010 não influenciou na significância dos coeficientes, sendo que apenas a categoria 7 anos da variável Idade não foi considerada estatisticamente significativa a um nível de 5% de significância.

Os sinais dos coeficientes de Rede e Gênero vão ao encontro do entendimento existente na literatura, posto que alunos matriculados na rede estadual e do sexo feminino têm proficiência média em Língua Portuguesa superior, em média, em 39 pontos se cotejados com aqueles da rede municipal e do sexo masculino.

O motivo pelo qual este modelo foi explicitado reside no fato de que, pela primeira vez, a variável Turno se mostrou estatisticamente significativa a 5% de significância, visto que seus intervalos de confiança só contêm valores positivos. Porém, a especificação do modelo parece não ser ainda a que melhor se ajusta aos dados, porque, além do coeficiente para a variável Projeto Escola Tempo Integral ser negativo em ambos os modelos, obteve-se, para o modelo que ignora o plano amostral, um p-valor de 0,001 no "Lack of Fit Test", o que significa que existem indícios para se rejeitar a hipótese nula e, destarte, concluir que, de fato, o modelo está incorretamente especificado.

Com a finalidade de verificar se a variável Turno irá permanecer significativa em modelos que consideram o fator Rede como covariável, introduziu-se nova especificação sem a presença da variável Escola Tempo Integral, cujo sinal do coeficiente não faz sentido se cotejado com estudos precedentes. Como anteriormente mencionado, isto pode ser explicado pela forma de estruturação da base de dados da pesquisa Proalfa 2010, ou ainda pela desconsideração da Rede, que, neste caso, se tornou covariável e não uma estratificação para as análises.

Conforme esperado, na tabela 11, os intervalos de confiança dos coeficientes são maiores quando se adota o plano amostral complexo como conseqüência dos maiores erros padrão, com destaque para Idade (8 anos).

Quanto à significância dos parâmetros, os modelos sob AAS e sob plano do Proalfa 2010 não diferem quanto às variáveis integrantes, sendo que apenas a Idade (7 anos) não é estatisticamente significativa a um nível de 5% de significância. O sinal da variável Gênero também é condizente com outros estudos, visto que alunas apresentam rendimento superior àquele obtido por alunos em Língua Portuguesa.

No modelo que desconsidera o plano complexo, como o nível de significância adotado foi de 5%, pode-se dizer que o modelo foi corretamente especificado, tendo em vista que o p-

valor resultante do “Lack of Fit Test” foi de 0,067, e, portanto, não rejeitando-se a hipótese nula.

Tabela 11: Modelo estimado para explicação da proficiência em Língua Portuguesa do Proalfa 2010, com Fator Rede e Turno, excluída covariável Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	565, 590	2,234	561,208	569,972	567,943	0,863	566,251	569,636
Gênero Feminino	19,957	0,987	18,020	21,893	19,497	0,792	17,945	21,048
Gênero Masculino	-	-	-	-	-	-	-	-
Rede Estadual	18,775	2,234	14,394	23,155	18,251	0,815	16,654	19,847
Turno Manhã	5,940	2,512	1,014	10,867	5,831	0,816	4,232	7,431
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	-8,094	40,922	-88,343	72,156	-0,314	31,397	-61,852	61,224
Idade (8 anos)	17,379	4,382	8,785	25,973	17,570	3,683	10,351	24,790
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Novamente, tanto a rede quanto o turno são estatisticamente significativos ao nível de significância adotado, com resultados que corroboram os do modelo anterior, porque os alunos inscritos no turno matinal da rede estadual possuem rendimento, expresso pela proficiência média, superior em, pelo menos, 23 pontos se comparado com as categorias de referência.

Além da proficiência média em Língua Portuguesa variar conforme a rede à qual a escola pertence, a variável Rede define a estrutura hierárquica da população, refletindo-se como uma estratificação natural a ser considerada na estimação. Desta forma, procedeu-se ao ajuste de modelos, com os dados da pesquisa Proalfa 2010, para a rede estadual e para a municipal.

Inicialmente, considerou-se na especificação de cada modelo todas as variáveis disponíveis, inclusive Turno, a fim de verificar sua influência e significância, consoante explicitado na tabela 12. Para a rede Estadual, enquanto a variável Idade (7 anos) não foi

significativa a 5% de significância tanto na estimação sob AAS quanto sob plano complexo, a significância da variável Turno foi afetada pela adoção do plano não ignorável, já que ela foi incluída no modelo que não considera o plano amostral, fato que não se repetiu no modelo baseado no plano da pesquisa Proalfa 2010.

Tabela 12: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Estadual de 2010, com fator Turno e Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	588,439	2,472	583,592	593,287	589,836	0,951	587,971	591,700
Projeto Escola Tempo Integral (Sim)	-24,995	3,831	-32,507	-17,483	-24,918	1,696	-28,242	-21,595
Gênero Feminino	16,900	1,529	13,901	19,898	17,448	1,123	15,248	19,648
Gênero Masculino	-	-	-	-	-	-	-	-
Turno Manhã	6,381	3,733	-0,938	13,701	6,920	1,141	4,684	9,156
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	55,169	44,205	-31,518	141,856	34,110	41,540	-47,310	115,530
Idade (8 anos)	22,724	6,270	10,429	35,019	20,974	4,768	11,628	30,320
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Quanto aos sinais dos coeficientes, apenas o do variável Escola Tempo Integral não condiz com estudos precedentes, fato possivelmente explicado pela base de dados, como já mencionado. Além disso, especificamente no caso do modelo que ignora o plano amostral, no teste de ajustamento (“Lack of Fit Test”), o p-valor foi de 0,043, inferior ao nível de significância adotado, havendo indícios para se rejeitar a hipótese nula e, conseqüentemente,

inferir que a especificação deste modelo está incorreta, devendo ser incluídas outras covariáveis, ou ainda interações entre as covariáveis já presentes, com o objetivo de captar melhor a relação entre a proficiência média em Língua Portuguesa e as demais variáveis explicativas.

Este modelo também comprovou que o plano não ignorável e com conglomeração resulta em estimativas maiores de erros padrão e, conseqüentemente, dos intervalos de confiança para cada coeficiente dos modelos de regressão.

Como não ficou comprovada a interferência da variável Turno sobre a proficiência média, ajustou-se um modelo que não a levasse em conta para a rede Estadual, o qual está disposto na tabela 13.

Tabela 13: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Estadual de 2010, excluído o Fator Turno

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	583,042	3,673	575,840	590,244	592,704	0,826	591,085	594,323
Projeto Escola Tempo Integral (Sim)	25,183	3,823	17,685	32,681	-24,959	1,697	-28,286	-21,633
Gênero Feminino	-16,885	1,528	-19,882	-13,889	17,415	1,123	15,213	19,617
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	56,335	46,164	-34,192	146,863	34,737	41,573	-46,750	116,224
Idade (8 anos)	22,516	6,397	9,972	35,061	20,667	4,772	11,314	30,020
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Neste caso, verifica-se que os erros padrão dos coeficientes do modelo são maiores quando se considera o plano amostral, o que também faz com que seus intervalos de confiança tenham maior amplitude. Isto também não teve impacto sobre a significância dos coeficientes, já que os modelos são idênticos sob AAS e sob plano complexo, ou seja, em ambos, a variável Idade (7 anos) não é estatisticamente significativa a 5% de significância. Porém, houve a inversão dos sinais dos coeficientes das variáveis “Gênero” e “Projeto Escola

Tempo Integral". No caso do modelo que ignora o plano amostral, pode-se dizer que não houve indícios de que sua especificação estivesse incorreta, em virtude do p-valor obtido no "Lack of Fit Test" superior a 5% de significância.

O sinal positivo do coeficiente da variável "Projeto Escola Tempo Integral" é mais coerente, posto que se espera melhor rendimento para alunos que freqüentam escola por um período maior de tempo. Contudo, o objetivo central desta monografia é avaliar os impactos de se ignorar o plano amostral complexo na estimação tanto de proficiência média quanto dos modelos de regressão. Assim, o foco central está sobre a magnitude de erros padrão das médias, de coeficientes dos modelos e seus intervalos de confiança, e não propriamente sobre o ajuste de modelos, que tenham necessariamente alto poder explicativo.

Ainda assim, a diferença de sinais para esta variável pode ser resultado de não ser possível distinguir claramente na base de dados a não-resposta (dados faltantes) da não participação do Projeto Escola Tempo Integral, conforme explicado anteriormente.

Ainda para a rede estadual, ajustou-se novo modelo sem a variável Projeto Escola Tempo Integral em sua estrutura, haja vista a incoerência de seu sinal. A fim de avaliar se a variável Turno, condicionada pelos demais fatores, exceto pelo tempo de permanência na escola, tem, de fato, influência sobre a proficiência média, a mesma foi mantida no modelo, como pode ser visto na tabela 14.

Tabela 14: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Estadual de 2010, com fator Turno, excluso Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	585,273	2,441	580,486	590,060	586,533	0,929	584,712	588,353
Gênero Feminino	17,288	1,554	14,241	20,334	17,736	1,128	15,525	19,947
Gênero Masculino	-	-	-	-	-	-	-	-
Turno Manhã	6,681	3,768	-0,708	14,069	6,987	1,146	4,741	9,234
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	50,055	40,517	-29,400	129,510	30,934	41,742	-50,884	112,752
Idade (8 anos)	23,381	5,951	11,711	35,051	21,824	4,791	12,434	31,215
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Cabe salientar que a variável Turno é estatisticamente significativa a 5% de significância apenas na estimação clássica, resultado que pode interferir nas políticas públicas a serem propostas. Além disso, o p-valor no teste de ajustamento ("Lack of Fit Test") foi de 0,088, superior a 5% de significância; destarte, pode-se concluir que não existem indícios para se rejeitar a hipótese de correta especificação do modelo sob AAS. Contudo, quando a estimação levou em conta o plano amostral complexo, o Turno foi irrelevante, não havendo quaisquer indicativos que ligassem a maior proficiência média em Língua Portuguesa ao turno matutino.

Novamente, a variável Idade, na categoria 7 anos, não se mostrou estatisticamente diferente de zero ao nível de significância adotado. Ademais, os sinais dos coeficientes são os mesmos para ambos os modelos, com destaque para a influência positiva que o gênero tem sobre a proficiência, resultado este que se coaduna com o senso comum e estudos precedentes, porquanto as alunas realmente têm melhor rendimento em português.

Este modelo vem corroborar o fato de que considerar a complexidade do plano amostral na estimação resulta em erros padrões de maior dimensão e, por conseqüência, intervalos de confiança de maior magnitude.

Tabela 15: Modelo final estimado para explicação da proficiência em Língua Portuguesa da rede Estadual de 2010

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	588,285	1,847	584,664	591,906	589,423	0,799	587,856	590,990
Gênero Feminino	17,275	1,553	14,231	20,320	17,703	1,129	15,491	19,916
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	51,235	42,533	-32,172	134,643	31,562	41,776	-50,323	113,447
Idade (8 anos)	23,169	6,073	11,259	35,079	21,516	4,795	12,118	30,914
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo Ca Ed.

Finalmente, gerou-se novo modelo que não contivesse nem a variável Turno, por não ser estatisticamente significativa no modelo sob plano complexo, nem a variável Projeto Escola Tempo Integral, em virtude do sinal de seu coeficiente. Na tabela 15, percebe-se que o plano não ignorável não teve influência sobre a significância dos parâmetros, haja vista que apenas a variável Idade (7 anos) não foi significativa a um nível de 5% de significância, fato ocorrido nos dois modelos. Os sinais para a variável gênero e idade (8 anos) fazem sentido, porque alunas têm, de fato, maior proficiência média em Língua Portuguesa, tal como comprovado em estudos precedentes. Ademais, alunos que estão no 4.º ano do ensino fundamental com idade de 8 anos também apresentam melhor desempenho.

Neste ponto, é interessante ressaltar que os coeficientes estimados para a variável Rede naqueles modelos em que a mesma foi covariável tiveram sua magnitude acrescida aos coeficientes do intercepto deste modelo, assim como dos anteriores para a rede estadual. Isto apenas deixa claro que a rede irá realmente influir de maneira positiva na proficiência média em Língua Portuguesa de alunos, elevando-a, em média, de 18 a 20 pontos, dependendo de cada modelo estimado.

Cabe salientar que o modelo sob AAS teve sua especificação significativa, consoante o p-valor de 0,584, obtido no “Lack of Fit Test”, o qual por ser superior a 5% de significância, leva à não rejeição da hipótese nula. Desta maneira, este modelo, ainda que com o pequeno número de covariáveis, já detém uma estrutura que é capaz de captar adequadamente o relacionamento existente entre a proficiência média em Língua Portuguesa e as demais variáveis independentes. Por fim, restou comprovado que a magnitude dos erros padrão e intervalos de confiança dos coeficientes é maior quando a estimação considera o plano complexo, que é o fim último desta monografia.

No que diz respeito à rede municipal, também foram ajustados alguns modelos, sendo o primeiro com todas as variáveis disponíveis inclusas. Posteriormente, retirou-se o Turno, Projeto Escola Tempo Integral e ambas, respectivamente. No primeiro modelo, disposto na tabela 16, os erros padrão dos coeficientes do modelo são maiores quando se considera o plano amostral, o que também faz com que seus intervalos de confiança sejam maiores. A única exceção foi a Idade, categoria 7 anos, cujo erro padrão estimado foi maior sob plano ignorável. No modelo sob AAS, esta variável é estatisticamente nula a 5% de significância, porquanto o intervalo de confiança contém o valor nulo, o que não ocorre quando a estimação considera o plano complexo. Ademais, o sinal negativo indica que crianças com 7 anos, cursando o 4.º ano do ensino fundamental, apresentam proficiência média inferior à dos alunos com 9 ou mais anos, em média, 61 pontos.

O destaque é para a variável Idade (8 anos) que, em todos os demais modelos, era estatisticamente significativa, fato que não aconteceu nestes dois primeiros ajustes. Como já explicitado anteriormente, ainda que o sinal do parâmetro de gênero seja coerente com a realidade, o mesmo não se pode afirmar a respeito da variável Projeto Escola Tempo Integral. Os referidos modelos também diferiram quanto ao Turno, o qual só foi estatisticamente significativo ao nível de significância adotado na estimação clássica.

Vale destacar que o modelo sob AAS não foi tido como significativo ao nível de significância adotado, a partir do p-valor obtido no "Lack of Fit Test", o que pode representar mais um indicador de que a variável Turno não traz maior poder explicativo, porquanto os outros modelos, ainda que sob estimação clássica, a qual não capta a real estrutura da população e os verdadeiros pesos de cada unidade amostral, conseguiram comprovar que suas estruturas estavam corretamente especificadas.

Tabela 16: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Municipal de 2010, com fator Turno e Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	566,934	2,709	561,622	572,247	569,417	1,090	567,282	571,553
Projeto Escola Tempo Integral (Sim)	-28,586	5,176	-38,735	-18,436	-29,870	2,373	-34,521	-25,219
Gênero Feminino	21,689	1,267	19,204	24,174	21,290	1,107	19,121	23,460
Gênero Masculino	-	-	-	-	-	-	-	-
Turno Manhã	5,329	3,331	-1,204	11,861	4,962	1,158	2,693	7,232
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	-61,162	23,307	-106,86	-15,457	-43,820	47,546	-137,01	49,374
Idade (8 anos)	10,942	6,381	-1,572	23,456	10,700	5,752	-0,575	21,974
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Tabela 17: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Municipal de 2010, excluído o Fator Turno

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	563,429	5,102	553,424	573,434	572,637	0,789	571,090	574,184
Projeto Escola Tempo Integral (Sim)	28,661	5,100	18,661	38,662	-29,629	2,373	-34,281	-24,978
Gênero Feminino	-21,648	1,268	-24,135	-19,161	21,248	1,107	19,077	23,418
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	-63,028	24,624	-111,32	-14,739	-45,357	47,563	-138,58	47,871
Idade (8 anos)	10,825	6,452	-1,826	23,477	10,493	5,754	-0,785	21,772
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Dessa maneira, optou-se por ajustar novos modelos que, primeiro, desconsideraram o Turno e depois o Projeto Escola Tempo Integral. Neste modelo para a rede municipal, que ignora o Turno, disposto na tabela 17, os erros padrão dos coeficientes do modelo são maiores quando se considera o plano amostral, o que também faz com que seus intervalos de confiança sejam maiores. Há apenas uma exceção: Idade, categoria 7 anos, cujo erro padrão estimado foi maior sob plano ignorável; porém, este coeficiente não se mostrou estatisticamente significativo neste modelo a 5% de significância, porquanto o intervalo de confiança contém o valor nulo. Ao se considerar o plano complexo, este coeficiente apresenta menor erro padrão e é estatisticamente diferente de zero ao nível de significância adotado. Ademais, o sinal negativo indica que crianças com 7 anos, cursando o 4.º ano do ensino fundamental, apresentam proficiência média inferior à dos alunos com 9 ou mais anos, em média, 63 pontos.

Além disso, influenciou na significância dos coeficientes, já que os modelos são diferentes sob AAS e sob plano complexo, em virtude da 1.ª categoria da Idade (7 anos). Eles também diferem dos modelos anteriores, porque a 2.ª categoria da Idade (8 anos) também é estatisticamente nula a 5% de significância, posto que os intervalos de confiança deste coeficiente contêm o valor nulo tanto na estimação para AAS quanto considerando o plano

amostral complexo. Simultaneamente, há a inversão do sinal dos coeficientes das variáveis Gênero e Projeto Escola Tempo Integral. No último caso, este fato pode ser explicado pela forma de elaboração da base de dados. A despeito disso, o modelo que ignora o plano amostral apresentou significância estatística, posto que não foi possível rejeitar, a 5% de significância, a hipótese nula do "Lack of Fit Test".

Ainda que o foco desta pesquisa seja avaliar o impacto sobre médias, erros padrão e intervalos de confiança da adoção do plano amostral complexo em sua estimação, cabe mencionar que o sinal positivo do coeficiente de "Projeto Escola Tempo Integral" é mais condizente com a realidade, na medida em que se espera melhor rendimento para alunos que freqüentam escola por um período maior de tempo. No caso do gênero, a consideração do plano conglomerado em 2 estágios do Proalfa 2010 evidenciou que a proficiência média em Língua Portuguesa de alunas é inferior, em média, 21,6 pontos se cotejada àquela obtida pelos alunos. Além de contrariar o senso comum, também não se coaduna com os resultados provenientes das pesquisas de 2008 e 2009, tanto nos modelos iniciais quanto nos finais para a rede municipal, e até mesmo estadual, presentes no trabalho de Cunha (2010).

Ante o exposto, prosseguiu-se à análise por meio do ajuste de novo modelo, excluindo a variável Projeto Escola Tempo Integral. No entanto, o Turno foi mantido com o fim de verificar sua significância quando não controlado pelo Projeto Escola Tempo Integral.

Tabela 18: Modelo estimado para explicação da proficiência em Língua Portuguesa da rede Municipal de 2010, com fator Turno, excluso Projeto Escola Tempo Integral

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	565,166	2,682	559,906	570,426	567,933	1,087	565,803	570,064
Gênero Feminino	21,637	1,278	19,132	24,143	21,244	1,111	19,067	23,421
Gênero Masculino	-	-	-	-	-	-	-	-
Turno Manhã	5,430	3,352	-1,143	12,002	4,617	1,162	2,340	6,895
Turno Tarde	-	-	-	-	-	-	-	-
Idade (7 anos)	-59,388	23,275	-105,03	-13,746	-42,190	47,714	-135,71	51,334
Idade (8 anos)	11,756	6,362	-0,719	24,231	11,266	5,772	-0,048	22,580
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd

Conforme tabela 18, os modelos sob AAS e plano complexo diferem quanto às variáveis Idade e Turno. Quando a estimação é clássica, a proficiência média em Língua Portuguesa é afetada pelo Turno, já que o mesmo é estatisticamente significativo a 5% de significância. No entanto, o mesmo não pode ser dito da Idade, já que tanto a categoria de 7 quanto de 8 anos tiveram seus coeficientes estatisticamente nulos ao nível de significância adotado.

Não se pode deixar de mencionar o fato do modelo que ignora o plano amostral apresentar especificação considerada significativa, a partir do p-valor de 0,762, que é maior que o nível de 5% de significância, obtido no "Lack of Fit Test", não havendo indícios, a princípio, da necessidade de se incluir outras covariáveis ou as interações entre já existentes.

Da comparação entre os resultados, pode-se vislumbrar que, quando se adota o plano amostral complexo na estimação, os erros padrões resultantes têm maior dimensão e, conseqüentemente, os intervalos de confiança são maiores.

Sob o plano complexo, apenas a categoria 8 anos da variável Idade não foi estatisticamente significativa. Os sinais para gênero e Idade (7 anos) estão de acordo com o esperado; todavia, a variável Turno não se mostrou estatisticamente significativa, o que levou à estruturação de dois novos modelos, os quais não contêm o Turno e Projeto Escola Tempo Integral.

Neste modelo final para a rede municipal, disposto na tabela 19, os erros padrão são menores sob os pressupostos clássicos da estimação, o que leva ao estreitamento dos intervalos de confiança. Como os modelos não são idênticos, a significância dos parâmetros foi influenciada pela adoção do plano conglomerado em 2 estágios.

Por exemplo, a variável Idade é estatisticamente significativa a 5% de significância em suas duas categorias quando o plano amostral é considerado. Os sinais dos coeficientes condizem com a realidade, tal como o do gênero. Além disso, o modelo estimado sob AAS apresentou significância estatística, já que não foi possível rejeitar, a 5% de significância, a hipótese nula do "Lack of Fit Test".

Mesmo que o objetivo do presente estudo não seja a modelagem da proficiência média, vale ressaltar que, para a rede municipal, quando o plano não ignorável é levado em consideração, este último modelo não é, necessariamente, o melhor, já que teve uma redução da ordem de 50% em seu poder explicativo, se comparado ao modelo que continha a variável Projeto Escola Tempo Integral em sua especificação.

Tabela 19: Modelo final estimado para explicação da proficiência em Língua Portuguesa da rede Municipal de 2010

Covariáveis	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Coeficiente	Erro padrão	IC (LI)	IC (LS)	Coeficiente	Erro padrão	IC (LI)	IC (LS)
Intercepto	568,736	1,614	565,570	571,901	570,942	0,780	569,412	572,472
Gênero Feminino	21,596	1,279	19,088	24,103	21,204	1,111	19,026	23,382
Gênero Masculino	-	-	-	-	-	-	-	-
Idade (7 anos)	-61,285	24,615	-109,55	-13,014	-43,633	47,729	-137,18	49,919
Idade (8 anos)	11,639	6,429	-0,968	24,246	11,070	5,774	-0,248	22,387
Idade (9 ou mais anos)	-	-	-	-	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Por todo o exposto, foi possível depreender que a significância da variável Turno é realmente questionável, a qual só influenciou a proficiência média em Língua Portuguesa em dois modelos, que não consideravam a estratificação natural da amostra, manifestada através da Rede.

Quanto ao sinal do coeficiente de Projeto Escola Tempo Integral, o mesmo se tornava positivo e, portanto, coerente com a realidade de que a proficiência será tanto melhor quanto maior o período de permanência na escola, em alguns modelos simplesmente quando se adotava o plano complexo. Todavia, em outros modelos, também ocorreu deste parâmetro apresentar sinal negativo, resultado que poderia ser explicado não só pela forma de obtenção dos dados, tal como anteriormente mencionado, como também por não considerar a Rede.

Assim, em virtude do sinal ser incoerente com o esperado em muitos casos, uma postura simples seria não considerar esta variável na especificação, ainda que houvesse perda de poder explicativo. Outra seria promover a correção na base de dados, fato que, além de não necessariamente solucionar o problema, poderia representar altos custos de logística. Por fim, resta como sugestão para as próximas pesquisas a serem implementadas a alteração na forma de elaboração do questionário, de forma a tornar clara a não-resposta da não participação deste Projeto.

Fato inquestionável é o de que a inclusão do fator Rede seja como covariável seja em análises estratificadas realmente tornou mais adequada a especificação dos modelos. Quanto à

decisão de qual deles seria melhor: se como covariável ou como variável de estratificação, há dois possíveis caminhos.

Do ponto de vista prático e pedagógico, provavelmente, o modelo que melhor atende aos gestores de políticas públicas educacionais seja aquele que estratifica os resultados por rede, porquanto simplificará a definição de metas e programas no sentido de melhorar o sistema de ensino. Afinal de contas, fatores que sejam importantes para a rede estadual podem não ser para a municipal; esta análise em separado poderá representar, em última instância, ações melhor definidas e com maior probabilidade de sucesso.

Contudo, do ponto de vista estritamente estatístico, o modelo que melhor se ajustou aos dados e, inclusive, obteve incremento da ordem de 50% em seu poder de ajustamento, representado pelo coeficiente de determinação (R^2) ajustado, foi aquele que considerou a Rede como covariável. Ou seja, esta foi a melhor especificação estrutural dos modelos, que conseguiu captar, de fato, a relação entre a proficiência média em Língua Portuguesa e as variáveis independentes disponíveis no questionário da pesquisa Proalfa 2010.

3.2.2 Comparação entre efeitos do plano amostral

Enquanto na tabela 20[1], os efeitos do plano amostral dispostos se referem aos modelos iniciais ajustados, ou seja, com todas as variáveis disponíveis, na tabela 21, as variáveis que não foram estatisticamente significativas a um nível de 5% de significância foram retiradas como, por exemplo, turno. Especificamente nos modelos para a rede municipal e estadual, os modelos finais não possuem em sua especificação final nem a variável Turno, nem o Projeto Escola Tempo Integral.

A partir dos valores de EPA, verifica-se que, em todos os modelos, considerar o plano amostral complexo subjacente à pesquisa Proalfa 2010 realmente interfere na variância dos estimadores e, conseqüentemente, nos resultados finais, fato anteriormente confirmado pelas estimativas dos erros padrão e intervalos de confiança dos coeficientes dos modelos de regressão.

[1] O Efeito do Plano Amostral considerando ou não o fator Rede utiliza toda a amostra para a inferência e estimação dos modelos, referentes às 2.ª e 3.ª colunas. O Efeito do Plano Amostral para as redes estadual e municipal considera a amostra estratificada por rede, estadual e municipal, para a geração de modelos, referentes às 4.ª e 5.ª colunas.

Tabela 20: Efeitos do Plano Amostral para Coeficientes dos Modelos Iniciais de Regressão

Variáveis	Efeito do Plano Amostral		Efeito do Plano Amostral	
	Sem Considerar Rede	Com fator Rede	Estadual	Municipal
Intercepto	7,257	8,299	5,631	8,546
Projeto Escola Tempo Integral (Sim)	5,301	5,240	3,913	6,678
Gênero Feminino	1,828	1,824	1,699	1,896
Gênero Masculino	-	-	-	-
Rede Estadual	-	8,655	-	-
Turno Manhã	11,142	11,122	10,008	11,877
Turno Tarde	-	-	-	-
Idade (7 anos)	2,868	2,832	1,689	1,969
Idade (8 anos)	2,144	2,094	2,045	2,119
Idade (9 ou mais anos)	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Tabela 21: Efeitos do Plano Amostral para Coeficientes dos Modelos Finais de Regressão

Variáveis	Efeito do Plano Amostral		Efeito do Plano Amostral	
	Sem Considerar Rede	Com fator Rede	Estadual	Municipal
Intercepto	5,137	6,617	4,543	5,831
Projeto Escola Tempo Integral (Sim)	5,270	5,146	-	-
Gênero Feminino	1,826	1,822	1,736	1,916
Gênero Masculino	-	-	-	-
Rede Estadual	-	8,397	-	-
Idade (7 anos)	2,867	2,833	1,756	2,011
Idade (8 anos)	2,157	2,141	1,895	2,123
Idade (9 ou mais anos)	-	-	-	-

Fonte: Elaboração Própria com base nos dados fornecidos pelo CAEd.

Comparando-se os modelos ajustados para a rede estadual e municipal, fica claro que nos últimos o impacto foi maior, porquanto os valores de EPA são maiores, com destaque para o EPA referente ao fator Rede Estadual, 8,397. Como esta medida também capta se a especificação dos modelos é satisfatória, foi possível perceber, através de seus valores, que aqueles que consideram a variável Rede, seja como covariável seja como variável estratificadora, são mais adequados. De forma mais explícita, as estimativas de EPA são as menores para os modelos que ignoram a Rede, quando comparadas àquelas resultantes dos modelos que tem a Rede como variável explicativa. Já os EPA's para os modelos estratificados por Rede são, em geral, superiores àqueles dos modelos com covariável Rede.

Assim, restou comprovada a importância de se considerar o plano amostral complexo subjacente à obtenção da amostra na estimação de médias, intervalos de confiança e, até mesmo, no ajuste de modelos de regressão. Finalmente, como havia pequeno número de covariáveis a ser incluído nos modelos de regressão, o grau de ajustamento dos mesmos foi reduzido. Todavia, os elevados valores de EPA obtidos evidenciam claramente a necessidade de não se ignorar a complexidade do planejamento amostral quando da estimação.

CONCLUSÃO

Nos últimos anos, houve a preocupação crescente quanto à qualidade da educação primária ofertada pelo sistema público. Nesta vertente, as políticas públicas delineadas têm sido pautadas em pesquisas estatísticas que fundamentem as decisões, metas e programas a serem implementados. Especificamente, no caso de Minas Gerais, tal como foi citado, há o Sistema Mineiro de Avaliação (SIMAVE), do qual o Proalfa 2010 é pesquisa integrante. O Proalfa busca avaliar o desempenho em Língua Portuguesa dos alunos de redes estadual e municipal em fase de alfabetização.

Mesmo que o objetivo primordial desta monografia seja de ordem metodológica, não havendo, conseqüentemente, o foco sobre a interpretação dos resultados gerados sob a ótica substantiva da Educação, é possível verificar algumas relações importantes, como por exemplo, o melhor desempenho médio de alunas, em relação aos alunos e na rede estadual, em relação à municipal. Dessa maneira, esses resultados podem ser melhor explorados em trabalhos subseqüentes, que poderão se embasar nos Boletins de Resultados do estudo para os anos de 2008 a 2010, disponíveis no sítio eletrônico do CAEd/UFJF. Nestes boletins, há também breve explicação a respeito de como são obtidos os indicadores de proficiência e as definições de letramento, entre outros.

Por meio dos resultados encontrados, foi possível mostrar que a estrutura hierárquica desta população amostrada (SRE, Redes, Escolas e turmas), tipicamente observável em pesquisas educacionais e a existência de correlação entre alunos de uma mesma turma e/ou escola implicam na ruptura dos pressupostos da inferência clássica, ou seja, as observações coletadas para integrar a amostra deixam de ser independente e identicamente distribuídas (IID).

Como decorrência deste fato, ainda que as estimativas pontuais de proficiências médias em Língua Portuguesa sejam próximas, há a subestimação de seus erros padrão e, conseqüentemente, o estreitamento dos intervalos de confiança e a redução da precisão das estimativas quando o processo de estimação desconsidera o plano amostral complexo.

O plano amostral subjacente ao Proalfa 2010 é conglomerado em 2 estágios, como explicitado no Capítulo 1, interferindo não só nas médias e intervalos de confiança, como também nos coeficientes dos modelos de regressão. Como pode ser observado, a significância de alguns parâmetros foi afetada pelo plano amostral, assim como houve a inversão de alguns

sinais de coeficientes. Especificamente quanto à variável Projeto Escola Tempo Integral, seria interessante buscar em estudos posteriores explicações para o sinal negativo em seu coeficiente ou mesmo estruturar de maneira diferente o questionário aplicado, presente na capa da avaliação.

De maneira geral, sob plano amostral complexo, os erros padrão e os intervalos de confiança a eles associados são maiores do que quando se ignora a forma de coleta de dados. Em última instância, isto se reflete sobre os resultados dos testes de hipóteses e sobre o próprio nível de confiança das análises, que pode se tornar, na realidade, inferior a seu valor nominal de 95%. Além disso, a estrutura conglomerada do plano amostral pode resultar em erros padrões maiores quando os conglomerados são homogêneos, ou seja, há pequena variação na característica de interesse da pesquisa.

Nesta monografia, contudo, os efeitos da estratificação foram suplantados por aqueles da conglomeração utilizada para obtenção da amostra, porquanto os valores de efeitos do plano amostral – EPA's – foram todos superiores à unidade, como demonstrado na Seção 3.2.2. Além disso, foi observado que o planejamento amostral subjacente à pesquisa Proalfa 2010 foi mais eficiente que os planos amostrais adotados nos anos anteriores, conclusão que se coaduna com os resultados obtidos por Cunha (2010), ainda que o mesmo apenas o tenha feito através dos erros padrões e intervalos de confiança. A análise longitudinal das pesquisas entre 2007 e 2010 poderia trazer resultados interessantes do ponto de vista pedagógico, sendo uma continuidade relevante para este estudo.

REFERÊNCIAS

ALBERNAZ, A.; FERREIRA, F.H.G.; FRANCO, C. **Qualidade e equidade no ensino fundamental brasileiro**. Pesquisa e planejamento econômico. Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada, V. 32, n. 3, p.453-476, dez.2002.

BINDER, David A..; ROBERTS Georgia R.. **Approaches for Analyzing Survey Data**: a Discussion. 2006 Joint Statistical Meetings – Section on Survey Research Methods, 2006. p. 2771-2778.

BOLFARINE, H; BUSSAB, W. O. **Elementos de Amostragem**. 1 ed. São Paulo: Blucher, 2005.

CARLSON, Barbara Lepidus. Software for statistical analysis of sample survey data: Design of Experiments and Sample Surveys. In: ARMITAGE, Peter; THEODORE, Colton, **Encyclopaedia of Biostatistics.** New Jersey: John Wiley, 1998.

CUNHA, Iago Carvalho. **Análise de Dados Amostrais Complexos da pesquisa do PROALFA de Minas Gerais**. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal de Juiz de Fora. Juiz de Fora: 2010.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3 ed. New York: Wiley-Interscience, 1998.

HOLT, D., SMITH T.M.F. e WINTER, P.D. **Regression analysis of data from complex surveys**. Journal of the Royal Statistical Society A. London: 1980. v. 143, p. 474-487.

HORVITZ, D. G; THOMPSON, D.J. **A Generalization of Sampling Without Replacement from a Finite Universe**. Journal of the American Statistical Association. Alexandria: 1952. v.47, n.260, p. 663-685, dez..

KISH, L. **Weighting - Why, When and How? A Survey for Surveys**: Proceedings of the section on survey research methods. American Statistical Association. Alexandria: 1990. pp. 121-130

KMENTA, J. **Elementos de Econometria**. 1 ed. São Paulo: Atlas, v. 2, 1988.

LUMLEY, T. **Survey**: analysis of complex survey samples. R package version 3.22-1, 2010. Disponível em: <http://www.R-project.org>.

NATHAN, G.; HOLT, D. **The effect of survey design on regression analysis**. University of Southampton: Journal of the Royal Statistical Society, 1980. v. 42, n. 3, p.377-386.

OHLSSON, E. Sequential Poisson Sampling. **Journal of Official Statistics**, 14, 1998. p. 149-162.

PEREIRA, Danielle Ramos de Miranda. **Fatores associados ao desempenho escolar nas disciplinas de matemática e de português no ensino fundamental:** uma perspectiva longitudinal. Dissertação (Doutorado em Demografia) – Universidade Federal de Minas Gerais. Belo Horizonte, 2006.

PESSOA, Djalma Galvão Carneiro; NASCIMENTO SILVA, Pedro Luis do. **Análise de Dados Amostrais Complexos.** São Paulo: Associação Brasileira de Estatística, 1998.

PFEFFERMAN, D. **The role of sampling weights when modelling survey data**. International Statistical Review. The Hague: ISI, 1993. p. 317-337.

PINDYCK, R. S.; RUBINFELD, D, L. **Econometria:** Modelos & Previsões. 4 ed. Rio de Janeiro: Elsevier, 2004.

R Development Core Team. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, 2009. Disponível em: <http://www.R-project.org>.

SKINNER, Chris J.; VIEIRA, Marcel de Toledo. **Design Effects in the Analysis of Longitudinal Survey Data:** Proceeding of the Sixth International Conference on Social Science Methodology. United Kingdom: University of Southampton, 2004.

SPSS Brasil. **Statistical Package of Social Sciences**. Disponível em: <http://www.spss.com.br>.

SUGDEN, R.A.; SMITH, T.M.F. **Ignorable and Informative Designs in Survey Sample Inference**. Southampton: Biometrika, 1984. p. 495-506.

VIEIRA, Marcel de Toledo. **Um Estudo Comparativo das Metodologias de Modelagem de Dados Amostrais Complexos**: Uma Aplicação ao SAEB 99. Dissertação (Mestrado em

Ciência da Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, 2001.

VIEIRA, Marcel de Toledo; SALGUEIRO, M. Fátima'; SMITH, Peter. W. F. **Misspecification Effects in the Analysis of Longitudinal Survey Data**. Glasgow: University of Glasgow, 2010. v. 1, p. 555-560.

VIEIRA, M. D. T., SOUZA, M. L. M.. **Plano Amostral da Pesquisa Proalfa de 2010**: Relatório Técnico. Juiz de Fora: Departamento de Estatística, UFJF, 2010.